Intelligent Systems, Control and Automation:
Science and Engineering

Honglei Xu
Xiangyu Wang *Editors*

# Optimization and Control Methods in Industrial Engineering and Construction

Springer

# Intelligent Systems, Control and Automation: Science and Engineering

Volume 72



For further volumes:
http://www.springer.com/series/6259

Honglei Xu · Xiangyu Wang
Editors

# Optimization and Control Methods in Industrial Engineering and Construction

Springer

*Editors*
Honglei Xu
Department of Mathematics
and Statistics
Curtin University
Perth, WA
Australia

Xiangyu Wang
School of Built Environment
Curtin University
Perth, WA
Australia

ISSN 2213-8986 ISSN 2213-8994 (electronic)
ISBN 978-94-017-8043-8 ISBN 978-94-017-8044-5 (eBook)
DOI 10.1007/978-94-017-8044-5
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013956328



Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

# Preface

> As it was stated in the Nine Chapters on the Mathematical Art (Jiu Zhang SuanShu) "Mathematics problems are able to vary to be extremely infinite, fine or unmeasurable. In spite of the much complexity, the approaches can always be discovered, not as difficultly as supposed, which involve no more than measurement, reasoning and calculation to learn common laws".
>
> —Hui Liu, The Nine Chapters on the Mathematical Art, no later than 100 BC

The field of optimization is vast with applications appearing in almost every area of science and engineering. Generally speaking, optimization is to do with minimizing or maximizing an objective function (e.g. cost, energy, profit) subject to various types of constraints that arise due to engineering requirements or physical specifications. The optimization techniques for solving optimization problems are particularly important in the aspects of engineering and science applications. There are many efficient optimization techniques available in the literature, while many new techniques continue to be developed so as to meet the needs of solving various new practical problems in areas such as industrial engineering and construction, which are motivated by the need of satisfying more stringent requirements on energy saving, environment protection, and green manufacturing and construction. The natural formulations of the corresponding optimization problems have become much more complicated. The purpose of this edited book is to gather papers which address interesting optimization and control methods and new applications of optimization methods in industrial engineering and construction. Topics include optimization and control theory, statistical measurement, monitoring, fault detection, process control, construction design and production management. This edited book could be used as a reference book for researchers and postgraduate students in science and engineering.

The book is composed of three parts. The first three chapters are devoted to the development of new optimization methods. From "Optimum Confidence Interval Analysis in Two-factor Mixed Model with a Concomitant Variable for Gauge Study" to "Economic Scheduling of CCHP Systems Considering the Tradable Green Certificates", the focus is on the new applications of optimization and control methods in industrial engineering. For the rest of the chapters, different optimization problems in construction projects are being addressed.

In "Robustness of Convergence Proofs in Numerical Methods in Unconstrained Optimization", the robustness of convergence proofs in numerical methods of unconstrained optimization is presented. It is developed based on an important principle in dynamic control system theory, where control policies are preferred to be of feedback form, rather than in an open loop manner. In "Robust Optimal Control of Continuous Linear Quadratic System Subject to Disturbances", the robust optimal control of linear quadratic system is considered. It is formulated as a minimax optimal control problem which admits a unique solution. A control parameterization scheme is developed to transform the infinite dimensional optimal control problem to one with finite dimension. It is further shown that the transformed finite-dimensional optimal control problem can be solved through semi-definite programming. In "A Linearly-Growing Conversion from the Set Splitting Problem to the Directed Hamiltonian Cycle Problem", a linearly growing conversion from the set splitting problem to the directed Hamiltonian cycle problem is discussed. A constructive procedure for such a conversion is given, and it is shown that the input size of the converted instance is a linear function of the input size of the original instance.

In "Optimum Confidence Interval Analysis in Optimum Confidence Interval Analysis in Two-Factor Mixed Model with a Concomitant Variable for Gauge Study", the efforts on optimum confidence interval analysis in two-factor mixed model for gauge study are studied. The analysis of variance is performed in the model and variabilities in the model are represented as a linear combination of variance components. Optimum confidence intervals are constructed using a modified large sample approach and a generalized inference approach is proposed to determine the variability such as repeatability, reproducibility, parts, gauge and the ratio of variability of parts to the variability of gauge. In "Optimization of Engineering Survey Monitoring Networks", the focus is on various ways of engineering survey monitoring networks, such that those used for tracking volcanic and large-scale ground movements may be optimized to improve the precision. These include the traditional method of fixing control points, the Lagrange method, free net adjustment, the g-inverse method and the singular value decomposition (SVD) approach using the pseudo-inverse. In "Distributed Fault Detection Using Consensus of Markov Chains", a fault detection procedure appropriate for use in a variety of industrial engineering contexts is proposed, where consensus among a group of agents about the state of a system is employed. Markov chains are used to model subsystem behaviours, and consensus is reached by way of an iterative method based on estimates of a mixture of the transition matrices of these chains. In "Engineering Optimization Approaches of Nonferrous Metallurgical Processes", an intelligent sequential operating method based on genetic programming is developed for solving nonferrous metallurgical processes, where optimization is being carried out while avoiding violent variation by operating the parameters in the ordered sequence. Real practical industrial data are used for carrying out the verification. In "Development of Neural Network Based Traffic Flow Predictors Using Pre-processed Data", a simple but effective training method by incorporating the mechanisms of back-propagation algorithm and the

exponential smoothing method is proposed to pre-process traffic flow data before training purposes. The pre-processing approach intends to aid the back-propagation algorithm to develop more accurate neural networks, as the pre-processed traffic flow data are more smooth and continuous than the original unprocessed traffic flow data. This approach is evaluated based on some sets of traffic flow data captured on a section of the freeway in Western Australia. Experimental results indicate that the neural networks developed based on this pre-processed data outperform those that are developed based on either original data or data which are pre-processed by the other pre-processing approaches. In "Economic Scheduling of CCHP Systems Considering the Tradable Green Certificates", tradable green certificate mechanism is introduced for the operation of CCHP system, and the impacts of tradable green certificate on the scheduling of CCHP system are studied. Then the economic dispatch model for multi-energy complementary system considering the TGC is proposed to maximize renewable energy utilization. This is a non-convex scheduling optimization problem. A global descent method is applied, which can continuously update the local optimal solutions by global descent functions. Finally, one modified IEEE 14-bus system is used to verify the performance of the proposed model and the optimization solver.

The remainder of the book relates to construction engineering optimization, more or less. Many types of optimization problems arise in construction engineering, such as sizing optimization, shape optimization, topology optimization, production optimization, contract dispatching and project management. Considering the differences in production conditions in the manufacturing industry, these problems are worth studying and complex for seeking valuable laws in optimization. First, the construction is rooted in place and conducted as on-site manufacturing. Second, every construction project is unique and a one-of-a-kind production, managed by a temporary organization, and consists of several companies. Third, highly interdependent activities have to be conducted in limited space, with multiple components, a lack of standardization and with many trades and subcontractors represented on-site. In "Optimizations in Project Scheduling: A State-of-Art Survey", a state-of-art survey of project management and scheduling is presented. This survey focuses on the new optimization formulations and new solution algorithms developed in the recent years. In "Lean and Agile Construction Project Management: As a Way of Reducing Environmental Footprint of the Construction Industry", a way of reducing the environmental footprint of the construction industry is proposed with the concept of lean and agile construction project management. It focuses on the construction project management with respect to the agility and leanness perspective and provides an in-depth analysis of the whole project life cycle phases based on lean and agile principles. Considering managing construction projects in Hong Kong, dynamic implications of industrial improvement strategies are analysed in "Managing Construction Projects in Hong Kong: Analysis of Dynamic Implications of Industrial Improvement Strategies". Based on a series of face-to-face interviews with experienced practitioners and a focus group exercise, this chapter presents the mapping of various interacting and fluctuating behaviours patterns during the site

installation stage of building services in construction projects, with the aid of a generic system dynamics model, and draws interesting conclusions about the relationships among factors in construction project management. In "Dynamic Project Management: An Application of System Dynamics in Construction Engineering and Management", system dynamics (SD) are taken into consideration for construction engineering and project management. It is expected to serve as a useful guideline for the application of SD in construction and to contribute to expanding the current body of knowledge in construction simulation. Since production control is an essential part of any complex and constrained construction project, a lean framework for production control in complex and constrained construction projects ($PC^4P$) is discussed in "A Lean Framework for Production Control in Complex and Constrained Construction Projects ($PC^4P$)", which is based on an open system-theory mindset and consists of components, connections and inputs. In "Optimization in the Development of Target Contracts", by formulating the sharing problem in optimization terms, specific quantitative results will be obtained for all the various combinations of the main variables that exist in the contractual arrangements and project delivery. Such variables include the risk attitudes of the parties (risk-neutral, risk-averse), single or multiple outcomes (cost, duration, quality), single or multiple agents (contractors, consultants), and cooperative or non-cooperative behaviour. This chapter will be particularly of interest to academics and practitioners in the discipline of the design of target contracts and project delivery. It provides an understanding of optimal sharing arrangements within projects, broader than currently available.

We take this opportunity to express our immense gratitude to Prof. Kok Lay Teo for his guidance and encouragement all the time. We would also like to acknowledge financial support from Curtin University and the Natural National Science Foundation of China (11171079). In addition, we wish to thank Nathalie Jacobs and Cynthia Feenstra from Springer for their kind cooperation and professional support. Our special thanks go to Dr. Xiaofang Chen for his technical support during this book's editing process. Finally, we would like to convey our appreciation to all contributors, authors and reviewers who made this book possible.

# Contents

# Robustness of Convergence Proofs in Numerical Methods in Unconstrained Optimization

**B. S. Goh, W. J. Leong and K. L. Teo**

**Abstract** Numerical methods to solve unconstrained optimization problems may be viewed as control systems. An important principle in dynamic control system theory is that control policies should be prescribed in a feedback manner rather than in an open loop manner. This is to ensure that the outcomes are not sensitive to small errors in the state variables. A standard proof in numerical methods in unconstrained optimization like the Zoutendijk method is, from the control theory point of view, an open loop type of analysis as it studies what happens along a total trajectory for various initial state variables. In this chapter, an example is constructed to show that the eventual outcome and convergence to a global minimum point or otherwise can be very sensitive to initial values of the state variable. Convergence of a numerical method in unconstrained optimization can also be established by using the Lyapunov function theorem. The Lyapunov function convergence theorem provides feedback type analysis and thus the outcomes are robust to small numerical errors in the initial states. It requires that the level sets of the objective function are properly nested everywhere in order to have global convergence. This means the level sets of the objective function must be topologically equivalent to concentric spherical surfaces.

## 1 Introduction

An iterative method to compute the minimum point in an unconstrained optimization problem can be viewed as a control system. Thus to achieve robust solutions it is desirable to have feedback solution rather than open loop control policies [1].

B. S. Goh (✉)
Research Institute, Curtin University Sarawak, 98009 Miri, Sarawak, Malaysia
e-mail: goh2optimum@gmail.com

W. J. Leong
Institute Mathematical Sciences, UPM, 43400 UPM Serdang, Selangor, Malaysia

K. L. Teo
Mathematics and Statistics, Curtin University, Perth, WA 6845, Australia

H. Xu and X. Wang (eds.), *Optimization and Control Methods in Industrial Engineering and Construction*, Intelligent Systems, Control and Automation: Science and Engineering 72,
DOI: 10.1007/978-94-017-8044-5_1, © Springer Science+Business Media Dordrecht 2014

A typical proof of a numerical method in optimization examines what happens along the total path of a trajectory for all admissible initial values. Thus, it is an open loop type of analysis. On the other hand, a proof of convergence of a numerical method by Lyapunov theorem in an unconstrained optimization problem examines what happens to changes in the value of the objective function relative to the level sets of the function in a typical iteration and it is re-started with numerical errors of the state variable. This is an example of feedback type control analysis and thus it is robust to numerical errors in the computation of the current position.

We shall draw on an example due to Barbashin and Krasovskii [1–3], and use Lyapunov function theory to illustrate the differences between open loop and closed loop convergence analysis of a numerical method in unconstrained optimization. It will also be demonstrated that open loop type of convergence along each trajectory for all possible initial conditions may not guarantee convergence to a global minimum point. It only establishes convergence to stationary points. What is needed is the concept of properly nested level sets of the objective function which is a key requirement for global convergence in a proof by using Lyapunov function theorem. Globally, an objective function has properly nested level sets if all the level sets are topologically equivalent to concentric spherical surfaces.

For convenience, brief reviews of Lyapunov function theorem for the global convergence of an iterative system and the Zoutendijk theorem for the convergence of a line search method in optimization will be given.

## 2 Convergence Proof by Using Lyapunov Function Theorem in Optimization

The traditional statement of the Lyapunov function theorem [1, 4–7] for a system of iterative equations is as follows: Let $x^*$ be the optimal solution in an optimization problem. It is the equilibrium point of a system of iterative equations. Let $L$ and $C$ be positive constants. The vector iterative equation is,

$$x(k+1) = F[x(k)], \quad x \in R^n, \quad k = 0, 1, 2, \ldots , \tag{1}$$

where $F(x)$ is a vector of continuous functions which does not explicitly contain the time variable $k$. It is said to be a time independent system. Thus, this analysis is not immediately applicable to time varying iterative systems like Quasi-Newton iterations in optimization. Some changes of this analysis can be made and they would then be applicable to time dependent systems.

We seek a continuous and nonnegative scalar function, $V(x)$, such that,

$$\Delta V[x(k)] = V[F(x(k))] - V[x(k)] < 0, \quad k = 0, 1, 2, \ldots \tag{2}$$

for all $x(k) \in \{x|0 \leq V(x) \leq L\} = \Omega(x, x^*, L)$ and where $x \neq x^*$. At the equilibrium point, $V(x^*) = 0$ and trivially, $\Delta V(x^*) = 0$. By definition, a *sublevel set* of the function $V(x)$is defined by $\Omega(x, x^*, L) = \{x|0 \leq V(x) \leq L\}$. Here, if $L$ is a large positive constant, it defines a large sublevel set of $V(x)$. On the other hand, a *level set* of the function $V(x)$ is the set given by $\Gamma = \{x|V(x) = C\}$. If condition (2) is satisfied, the function $\Delta V(x)$ is said to be *negative definite* in the sublevel region, $\Omega(x, x^*, L)$. It is important to differentiate between a level set and a sublevel set in a convergence analysis.

In an unconstrained optimization problem with objective function $f(x)$, the following function is a natural Lyapunov function

$$V(x) = f(x) - f(x^*). \tag{3}$$

Clearly, $V(x)$ is a merit function in optimization theory, with an additional requirement that it has a zero value at the optimal point, $x^*$. Furthermore, all the level sets of the function $V(x)$ must be *properly nested* in a sublevel set or global region, which means that they are topologically equivalent to concentric spherical surfaces. The function $V(x)$ with the required properties is called a *Lyapunov function*. The condition that the level sets of a function are properly nested can be verified easily for a function of two variables. This is done by plotting samples of the level sets of the function and by invoking the assumption that the function is continuous.

Suppose that $f(x)$ is the objective function for an unconstrained optimization problem with higher dimension. Then a sufficient condition to ensure that the level sets of a Lyapunov function are properly nested globally is that there exists a positive constants, $\gamma$, such that

$$V(x) - V(x^*) = f(x) - f(x^*) \geq \gamma ||x - x^*||, \tag{4}$$

for all $x \in R^n$, where $||.||$ is a norm. If (4) is satisfied globally, the Lyapunov function is also said to be *radially unbounded*. In (4), a *fixed point* at the point $x^*$ is used. On the other hand, a Lipschitz type condition in place of (4) for use in convergence analysis in numerical methods, would require that for all $x$ and $y$ in a finite region,

$$||\nabla f(x) - \nabla f(y) \leq \gamma ||x - y||. \tag{5}$$

Note that the inequality signs in (4) and (5) are in opposite directions. Furthermore, (4) is a condition on the objective function rather than its gradient function in (4).

**Theorem 2.1** *The equilibrium, $x^*$, of the iterative equation (1) is globally convergent if*

*(i) there exists a continuous nonnegative function $V(x)$ with $V(x^*) = 0$, such that the function change $\Delta V(x)$ in (2) is negative definite globally and*
*(ii) all the level sets of $V(x)$, are properly nested.*

**Proof** Suppose as $k \to \infty$ the function $V[x(k)] \to K_\infty \neq 0$

We maximize the function

$$W(x) = \Delta V[x(k)] = V[F(x(k))] - V[x(k)] \tag{6}$$

for all $x(k) \in \Omega(x, K_\infty, V(x(0))) = \{x | K_\infty \leq V(x) \leq V(x(0))\}$. This set, which is bounded by the two level sets of the function, $V(x)$, is a closed and bounded (i.e., compact) set because of the assumption that the level sets of $V(x)$ are properly nested. Thus, by Weierstrass's theorem for continuous functions, the maximum of $W(x) = \Delta V(x) = V[F(x)] - V(x)$ in $\Omega(x, K_\infty, V(x(0)))$ exists and it is attained in this compact set. Let the maximum value of $W(x) = -\theta$. Furthermore, $\theta$ is a nonzero positive parameter as $\Delta V(x)$ is negative definite and by assumption, $K_\infty \neq 0$.
We have

$$V[x(N)] = \sum_{0}^{N-1} \Delta V[x(k)] + V[x(0)] \leq -N\theta + V[x(0)]. \tag{7}$$

This implies that $V[x(N)] \to -\infty$ as $N \to \infty$. This is impossible as $V(x)$ is nonnegative for all values of $N$. Hence we must have $K_\infty = 0$.This shows that the equilibrium is globally asymptotically convergent.

**Corollary 2.1** *Suppose that the two conditions in Theorem 2.1 are satisfied only in a finite sublevel region, $\Omega(x, x^*, L)$. Then the convergence is valid in the finite region.*

To apply Theorem 2.1 to a numerical method in an unconstrained optimization problem, minimize $f(x)$, a natural choice of the Lyapunov function is,

$$V(x) = f(x) - f(x^*).$$

This implies that,

$$\Delta V(x) = \Delta[f(x) - f(x^*)] = \Delta f(x). \tag{8}$$

It is an important practical result, because $\Delta V(x)$ can be calculated in each step of an iterative method for an optimization problem even though the Lyapunov function $V(x)$ is not explicitly defined. *This property provides an important way to ensure that the Lyapunov function theorem is satisfied in a specific problem when a numerical method is used.*

On careful examination of (7), it is observed that the reduction of value of the Lyapunov function is finite and negative in a typical iteration. When the Lyapunov function theorem is applied to a numerical method for finding a solution of a specific problem, $\Delta V(x) = \Delta f(x)$ can be computed at each step. If numerical errors of an algorithm cause it to be positive in a particular iteration, $\Delta V(x) = \Delta f(x)$ would require re-computation until it is negative or stop—indicating failure of the numerical method.

## 3 Zoutendijk Convergence Analysis of a Numerical Method in Optimization

The Zoutendijk theorem is a set of prototype conditions which are used to establish the convergence of a numerical method for computing the minimum point of an optimization problem. It examines what happens along the *total* trajectory for a given initial state. A numerical method in unconstrained optimization may be viewed as a control system where the position is called the state vector and the steplengths and directions are control variables. For a control system, a feedback control policy is preferred over an open loop control policy [1, 8]. This is because an open loop control can be very sensitive to errors in the initial or current values of the state variables. This sensitivity of outcomes to numerical errors in the initial or current state variables will be explicitly and clearly demonstrated in an example.

For convenience, we briefly describe the application of Zoutendijk theorem to establish convergence of a line search method in unconstrained optimization of the objective function $f(x)$. Assume a line search method generates the iterative equation,

$$x(k+1) = x(k) + \alpha(k)p[x(k)], \quad x \in R^n, \quad k = 0, 1, 2, \ldots . \tag{9}$$

The key conditions required are: (i) The objective function is bounded below; and (ii) the gradient vector of the objective function satisfies the Lipschitz condition in an open subset $\Omega(x, x_0)$ of the sublevel set $\{x|f(x) \leq f(x_0)\}$. This means that for any pair of points x and y in $\Omega(x, x_0)$, there exists a positive constant $\gamma$ such that,

$$||\nabla f(y) - \nabla f(x)|| \leq \gamma ||y - x||. \tag{10}$$

Furthermore, the steplength in the iterative equation (9) is chosen to satisfy the Wolfe's conditions, namely,

$$f[x(k) + \alpha(k)p(k)] \leq f[x(k)] + c_1\alpha(k)\nabla f[x(k)]^T p(k), \tag{11}$$

$$\nabla f[x(k) + \alpha(k)p(k)]^T p(k) \geq c_2 \nabla f[x(k)]^T p(k). \tag{12}$$

The positive constants $c_1$ and $c_2$ are such that $0 < c_1 < c_2 < 1$.

Under these conditions, the Zoutendijk theorem states that,

$$\sum \cos^2\theta(k)||\nabla f[x(k)]||^2 < \infty. \tag{13}$$

Here, $\theta(k)$ is the angle between the search direction $p(k)$ and the steepest descent direction, $-\nabla f[x(k)]$. Thus, if there exists a positive constant $\sigma$ such that

$$\cos(\theta(k)) \geq \sigma > 0, \tag{14}$$

then it can be deduced that

$$\lim ||\nabla f[x(k)]|| = 0 \tag{15}$$

as $k \to \infty$. This means that the trajectory generated from an arbitrary initial point $x_0$ would converge to a stationary point.

It is important to note that condition (13) or (15) is a property of the total trajectory from an arbitrary initial point $x_0$. Thus, there is no way to predicts what will happen if there are numerical errors in the initial state vector or a current vector as the iterative method progresses. In control system terminology, this may be viewed as an open loop control policy which is sensitive to numerical errors in the state variable during the computation of successive iterations. We shall demonstrate this by a specific example in the next section. Furthermore, the convergence of the iterative method is only to a stationary point which may not be even a local minimum point. This will be shown in an example.

## 4 Analysis of a Counterexample Without Properly Nested Level Sets

We shall adapt a counterexample due to Barbashin and Krasovskii [1–3] in Lyapunov theory for a system of ordinary differential equations to a system of iterative equation equations. For a system of ordinary differential equations, without the property that all the level sets are properly nested, an objective function can be monotonic decreasing, but the trajectories may not converge to the global minimum point.

From this counterexample, it is observed that if all the level sets of the objective function are not properly nested, then the solutions can be very sensitive to errors in the values of initial variables and hence they are not robust against numerical errors.

**Example 4.1** Consider an unconstrained optimization with its objective function,

$$V(x) = f(x) = x_1^2/(1 + x_1^2) + x_2^2. \tag{16}$$

Its global minimum is at the origin. For convenience, let

$$w = (1 + x_1^2). \tag{17}$$

It does not have properly nested level sets for states in the set defined by

$$\{x|V(x) \geq b > 1\},$$

where $b$ is a constant.

Assume that the iterative equations to compute the minimum point are given by

$$x_1(k + 1) = x_1(k) - \alpha(k)[6x_1(k)/w^2(k) + 2x_2(k)], \tag{18}$$

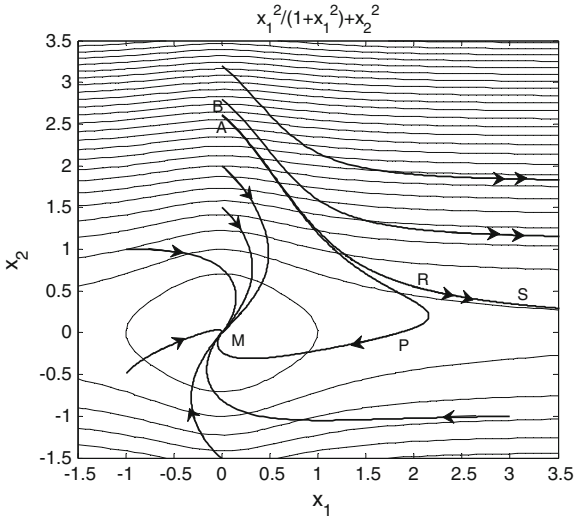



**Fig. 1** Objective function has properly nested level sets only in a sublevel set $\{x|f(x) \leq c < 1\}$. Lyapunov theorem guarantees convergence only is such sets. Zoutendijk theorem applies globally as function is monotonic decreasing everywhere. But trajectories with initial condition (0,a) with $a \geq 2.61$ converge to $(\infty, 0)$. BRS and APM show sensitivities to initial values

$$x_2(k+1) = x_2(k) - \alpha(k)[2(x_1(k) + x_2(k))/w^2(k)], \tag{19}$$

with the steplength, $\alpha(k) = 0.01$

With sufficiently small steplengths, it follows from Taylor's approximation that

$$\begin{aligned} \Delta V(x) &= V[x(k+1)] - V[x(k)] \\ &= \nabla V[x(k)]^T \Delta x(k) \\ &= -12\alpha x_1^2/w^4 - 4\alpha x_2^2/w^2 < 0. \end{aligned} \tag{20}$$

Thus, the objective function is monotonic decreasing globally except at the origin.

Apply the Zoutendijk theorem to this example, we deduce that

$$\lim \cos^2 \theta ||\nabla f(x)||^2 \to 0, \tag{21}$$

with iterations from any point, globally. But the outcomes could be the global minimum point at the origin or a stationary point at $(\infty, 0)$ or $(-\infty, 0)$.

By Lyapunov function theorem, the level sets of the objective function are only properly nested in a sublevel set, $\{x|f(x) = V(x) \leq c < 1\}$, where $c$ is a constant. Thus, by Lyapunov function theorem, we are ensured that all trajectories with initial points in this sublevel set will converge to the global minimum at the origin (0,0), as depicted in Fig. 1.

Zoutendijk theorem can be applied to all initial points but for initial points, such as (0,a) with $a \geq 2.61$, the trajectories converge to $(\infty, 0)$, rather than the global minimum point. A more important issue is that as under an open loop control policy, the trajectories can be sensitive with respect to numerical errors in the initial state vector. This is illustrated by the trajectories APM from (0,2.6) and BRS from (0,2.61) in Fig. 1. Here, a small change in initial conditions leads to entirely different outcomes. Thus, Zoutendijk theorem in a proof of convergence only provides conditions for a trajectory from a typical initial point to converge to a stationary point. More importantly, the trajectories can be very sensitive to the choice of the values of the initial state variables. This phenomena is a well known weakness of an open loop policy in control systems.

## 5 Conclusion

Numerical methods in unconstrained optimization can be viewed as control systems. It is well know that a feedback control policy is much preferred over an open control policy in control systems. Proofs of convergence of a numerical method, such as those based on Zoutendijk theorem, are in the context of open loop control policies. They examine what happens along the total path of a trajectory for different initial values. Thus the outcome could be sensitive to numerical errors of the initial values or the current state. Furthermore, Zoutendijk theorem ensures only convergence to stationary points.

On the other hand, the Lyapunov theorem proof of the convergence of a numerical method in unconstrained optimization is a feedback type of analysis. It requires that in a typical iteration the decrease in the objective function must be finite and negative. If numerical errors caused the failure of this monotonic decrease condition of the objective function, then it requires new iterations by line search or otherwise to recompute a new iteration which causes a decrease in the objective function. Thus the Lyapunov function approach has feedback properties.

Furthermore, the Lyapunov function requires that the objective function has properly nested level sets globally or in a finite sublevel set which defines an estimate of its region of convergence. With the properly nested level sets property, convergence to a minimum point is guaranteed and not just to a stationary point.

## References

1. Vincent TL, Grantham WJ (1997) Nonlinear and optimal control systems. Wiley, New York
2. Barbashin EA, Krasovskii NN (1952) On the stability of a motion in the large. Dokl Akad Nauk SSR 86:453–456
3. Hahn W (1967) Stability of motion. Springer, New York
4. Ortega JM (1973) Stability of difference equations and convergence of iterative processes. SIAM J Num Anal 10:268–282

5. LaSalle JP (1976) The stability of dynamical systems. SIAM, Philadephia
6. Kalman RE, Bertram JE (1960) Control system analysis and design via the second method of Liapunov. II. Discrete-time systems. ASME J Basic Eng 82:394–400
7. Goh BS (2010) Convergence of numerical methods in unconstrained optimization and the solution of nonlinear equations. J Optim Theory Appl 144:43–55
8. Khalil HK (2002) Nonlinear systems, 3rd edn. Prentice Hall, Englewood Cliffs

# Robust Optimal Control of Continuous Linear Quadratic System Subject to Disturbances

**Changzhi Wu, Xiangyu Wang, Kok Lay Teo and Lin Jiang**

**Abstract** In this chapter, the robust optimal control of linear quadratic system is considered. This problem is first formulated as a minimax optimal control problem. We prove that it admits a solution. Based on this result, we show that this infinite-dimensional minimax optimal control problem can be approximated by a sequence of finite-dimensional minimax optimal parameter selection problems. Furthermore, these finite-dimensional minimax optimal parameter selection problems can be transformed into semi-definite programming problems or standard minimization problems. A numerical example is presented to illustrate the developed method.

## 1 Introduction

A fundamental problem of theoretical and practical interest, that lies at the heart of control theory, is the design of controllers that yield acceptable performance for a family of plants under various types of inputs and disturbances [1]. This problem is often referred to as a robust optimal control problem. Normally, there are two kinds of criteria to achieve robust controller design. One is based on a statistical description, i.e., the criterion of the expectations of the cost and the constraints is adopted [17]. For the other one, the worst-case performance criterion is adopted [2–4, 9–12, 15, 18].

C. Wu (✉) · X. Wang
School of Built Environment, Curtin University, Perth, WA 6845, Australia
e-mail: changzhiwu@yahoo.com

X. Wang
e-mail: Xiangyu.Wang@curtin.edu.au

K. L. Teo
School of Mathematics and Statistics, Curtin University, Perth, WA 6845, Australia
e-mail: k.l.teo@curtin.edu.au

L. Jiang
School of Mathematics, Anhui Normal University, 241000 Wuhu, China

H. Xu and X. Wang (eds.), *Optimization and Control Methods in Industrial Engineering and Construction*, Intelligent Systems, Control and Automation: Science and Engineering 72,
DOI: 10.1007/978-94-017-8044-5_2, © Springer Science+Business Media Dordrecht 2014

The dynamical systems can be classified into two kinds—discrete dynamical system and continuous dynamical system. For the robust optimal control of linear discrete dynamical system with quadratic cost function, there are many results available [2–5, 9–12, 15, 18]. If disturbances lie in an ellipsoid, then it is shown in [3] that such an optimal control problem without constraints is equivalent to a semi-definite programming (SDP) problem. If the optimal control problem is subject to constraints on the state and control, it can be relaxed (see [3]) as a second-order cone programming (SOCP). If disturbances lie in a polyhedral, then such a robust optimal control problem becomes computationally highly demanding, (see [2, 12]). For other results on such robust optimal control problems, see, for example, [2, 3, 10–12, 18]. For robust optimal control governed by continuous dynamical system, a computational scheme is developed in [16]. By introducing a linear operator and resorting to its norm, the original minimax optimal control problem can be transformed into a standard optimal control problem. This method depends crucially on the special form of the cost function. If the cost function is with the terminal cost, then this method does not work.

In this chapter, we consider a class of robust optimal control problems governed by continuous dynamical systems subject to constraints on the admissible controls and the disturbances. The cost function involves not only a quadratic integral cost, but also a terminal cost expressed in the form of quadratic terminal state. Furthermore, we will use piecewise functions, rather than orthornormal basis as in [16], to approximate admissible control functions. We first show that this robust optimal control problem admits a solution. Based on this result, we show that this infinite-dimensional minimax optimal control problem can be approximated by a sequence of finite-dimensional minimax optimal parameter selection problems. Then, we show that these minimax optimal parameter selection problems can be transformed into SDPs. We also show that these minimax optimal parameter selection problems can also be transformed into standard minimization problems. Thus, gradient-based optimization methods can be applied. To illustrate our developed method, a numerical example is presented.

## 2 Problem Formulation

Consider the continuous linear dynamical system

$$\begin{aligned} &\dot{x}(t) = A(t)\,x(t) + B(t)\,u(t) + C(t)\,w(t),\ t \in [0, T], \\ &x(0) = x^0, \end{aligned} \tag{1}$$

where $T$ is the given terminal time, $x(t) \in \mathbb{R}^n$ is the state at time $t$, $x^0$ is a given initial state, $u(t) \in \mathbb{R}^m$ is the input at time $t$, $w(t) \in \mathbb{R}^r$ is the uncertainty at time $t$, and $A$, $B$ and $C$ are matrices with appropriate dimension.

Let

$$\mathscr{W} = \left\{ w \in L^2\left([0,T],\mathbb{R}^r\right) : \|w\|_{L^2}^2 = \int_0^T (w(t))^T w(t)\,dt \le \zeta^2 \right\}, \tag{2}$$

and

$$\mathscr{U} = \left\{ u \in L^2\left([0,T],\mathbb{R}^m\right) : \|u\|_{L^2}^2 = \int_0^T (u(t))^T u(t)\,dt \le \eta^2 \right\}. \tag{3}$$

A function $u$ is said to be an admissible control if $u \in \mathscr{U}$. Note that $\mathscr{W}$ and $\mathscr{U}$ are weakly closed in $L^2([0,T],\mathbb{R}^r)$ and $L^2([0,T],\mathbb{R}^m)$, respectively. For brevity, they are simply referred to as weakly closed.

Now our robust optimal control problem can be stated as follows.

*Problem (P).* Choose $(u^*, w^*) \in \mathscr{U} \times \mathscr{W}$ such that

$$J(u^*, w^*) = \min_{u\in\mathscr{U}} \max_{w\in\mathscr{W}} J(u,w) = (x(T))^T Px(T) + \int_0^T (x(t))^T Q(t)x(t) + (u(t))^T R(t)u(t)\,dt, \tag{4}$$

where $P$, $Q(t)$ and $R(t)$ are all positive definite matrices with appropriate dimensions.

To proceed, we assume that the matrices $A(t)$, $B(t)$, $C(t)$, $Q(t)$ and $R(t)$ are continuous matrix-valued functions defined on $[0,T]$.

## 3 Existence Theorem

Note that for each given $t \in [0,T]$, $P$, $Q(t)$ and $R(t)$ are all positive definite matrices. Let $F(t,\tau)$ be the $n \times n$ state transition matrix that satisfies

$$\begin{aligned} \dot{F}(t,\tau) &= A(t)F(t,\tau), \\ F(\tau,\tau) &= I, \end{aligned} \tag{5}$$

where $I$ is the identity matrix. Then, for each given $u$ and $w$, the solution of (1) can be expressed as

$$x(t|u,w) = F(t,0)x_0 + \int_0^t F(t,\tau)B(\tau)u(\tau)\,d\tau + \int_0^t F(t,\tau)C(\tau)w(\tau)\,d\tau. \tag{6}$$

Since $P$ and $Q(t)$ are positive definite matrices for each given $t \in [0,T]$, $J(u,w)$ is strictly convex with respect to $x$. From (6), it follows that $x$ is linear with respect to $w$. Thus, $J(u,w)$ is strictly convex with respect to $w$. For each given $u \in \mathscr{U}$, since $\mathscr{W}$ is a weakly sequentially compact subset of $L^2([0,T],\mathbb{R}^r)$, there exists a $w(u)$

such that

$$J\left(u, w\left(u\right)\right) = \max_{w \in \mathscr{W}} J\left(u, w\right).$$

Let

$$\mathscr{G}\left(u\right) = \int_0^T \left(u\left(t\right)\right)^T R\left(t\right) u\left(t\right) dt.$$

Note that for each given $t \in [0, T]$, $R\left(t\right)$ is positive definite, it is easily to verify that $\mathscr{G}\left(u\right)$ is a strictly convex function with respect to $u$. Now we have the following lemmas.

**Lemma 1.** *If $u_n \rightharpoonup u$ as $n \to \infty$, ($u_n \rightharpoonup u$ means that $u_n$ converges to $u$ weakly in $L^2\left([0, T], \mathbb{R}^m\right)$). Then,*

$$u \in \mathscr{U} \text{ and } \mathscr{G}\left(u\right) \leq \varliminf_{n \to \infty} \mathscr{G}\left(u_n\right). \tag{7}$$

*If $u_n \to u$ as $n \to \infty$, ($u_n \to u$ means that $u_n$ converges to $u$ in the norm of $L^2\left([0, T], \mathbb{R}^m\right)$), where $\{u_n\} \subset \mathscr{U}$, then*

$$u \in \mathscr{U} \text{ and } \lim_{n \to \infty} \mathscr{G}\left(u_n\right) = \mathscr{G}\left(u\right). \tag{8}$$

*Proof.* Suppose that $u_n \rightharpoonup u$. Clearly, $u \in \mathscr{U}$, as $\mathscr{U}$ is a weakly closed set in $L^2\left([0, T], \mathbb{R}^m\right)$. By the convexity of $\mathscr{G}\left(u\right)$, we have

$$\mathscr{G}\left(u_n\right) \geq \mathscr{G}\left(u\right) + \left\langle D\mathscr{G}\left(u\right), u_n - u\right\rangle = \mathscr{G}\left(u\right) + 2\int_0^T \left(u_n\left(t\right) - u\left(t\right)\right)^T R\left(t\right) u\left(t\right) dt. \tag{9}$$

Note that $\{u_n\} \subset \mathscr{U}$ and $R\left(\cdot\right)$ is continuous on $[0, T]$, we can show that

$$\int_0^T \left(u_n\left(t\right)\right)^T R\left(t\right) u_n\left(t\right) dt$$

is bounded uniformly with respect to $n$. Thus, $\varliminf_{n \to \infty} \mathscr{G}\left(u_n\right)$ exists. Since $R\left(\cdot\right)$ is continuous on $[0, T]$ and $u \in L^2\left([0, T], \mathbb{R}^m\right)$, it follows that $R\left(\cdot\right) u\left(\cdot\right) \in L^2\left([0, T], \mathbb{R}^{n \times m}\right)$. Thus,

$$\lim_{n \to \infty} \int_0^T \left(u_n\left(t\right)\right)^T R\left(t\right) u\left(t\right) dt = \int_0^T \left(u\left(t\right)\right)^T R\left(t\right) u\left(t\right) dt \tag{10}$$

as $u_n \rightharpoonup u$. Therefore, (7 ) holds.
Suppose that $u_n \to u$, *i.e.*,

$$\left\|u_n - u\right\|_{L^2} \to 0. \tag{11}$$

Clearly, $u \in \mathscr{U}$. Since $\{u_n\} \subset \mathscr{U}$ and $R(\cdot)$ is continuous on $[0, T]$, there exists a constant $\varkappa$ such that

$$\|R(\cdot)\, u_n(\cdot)\|_{L^2} \leq \varkappa \text{ for all } n = 1, 2, \ldots,$$

and

$$\|R(\cdot)\, u(\cdot)\|_{L^2} \leq \varkappa.$$

Thus,

$$\begin{aligned} |\mathscr{G}(u_n) - \mathscr{G}(u)| &\leq \left| \int_0^T (u_n(t) - u(t))^T R(t)\, u_n(t)\, dt \right| + \\ &\quad \left| \int_0^T (u_n(t) - u(t))^T R(t)\, u(t)\, dt \right| \\ &\leq \varkappa \|u_n - u\|_{L^2} + \left| \int_0^T (u_n(t) - u(t))^T R(t)\, u(t)\, dt \right| \qquad (12) \\ &\leq 2\varkappa \|u_n - u\|_{L^2}. \end{aligned}$$

Since $u_n \to u$, it follows that $\lim_{n\to\infty} \mathscr{G}(u_n) = \mathscr{G}(u)$. This completes the proof.

Define

$$\mathscr{F}(u, w) = (x(T|u, w))^T Px(T|u, w) + \int_0^T (x(t|u, w))^T Q(t)\, x(t|u, w)\, dt,$$

We have the following lemma.

**Lemma 2.** *Suppose that $u_n \rightharpoonup u$ and $w_n \rightharpoonup w$ as $n \to \infty$, where $\{u_n\} \subset \mathscr{U}$ and $\{w_n\} \subset \mathscr{W}$. Then,*

$$\lim_{n\to\infty} \mathscr{F}(u_n, w_n) = \mathscr{F}(u, w), \qquad (13)$$

*where $u \in \mathscr{U}$ and $w \in \mathscr{W}$.*

*Proof.* Since $\mathscr{U}$ and $\mathscr{W}$ are weakly closed, $u \in \mathscr{U}$ and $w \in \mathscr{W}$. By the continuity of $A(t)$, $F(t, \cdot)$ is continuous on $[0, t]$ for each $t \in [0, T]$ Note that

$$\begin{aligned} &|x(t|u_n, w_n) - x(t|u, w)| \\ &= \left| \int_0^t F(t, \tau)\, B(\tau)(u_n(\tau) - u(\tau))\, d\tau + \int_0^t F(t, \tau)\, C(\tau)(w_n(\tau) - w(\tau))\, d\tau \right| \\ &\leq \left| \int_0^t F(t, \tau)\, B(\tau)(u_n(\tau) - u(\tau))\, d\tau \right| + \left| \int_0^t F(t, \tau)\, C(\tau)(w_n(\tau) - w(\tau))\, d\tau \right| \\ &= \left| \int_0^T \tilde{F}(t, \tau)\, B(\tau)(u_n(\tau) - u(\tau))\, d\tau \right| + \left| \int_0^T \tilde{F}(t, \tau)\, C(\tau)(w_n(\tau) - w(\tau))\, d\tau \right|, \end{aligned}$$

where

$$\tilde{F}(t,\tau)=\begin{cases}F(t,\tau), & if\ \tau\le t,\\ 0_{n\times n} & \text{else.}\end{cases}$$

Clearly, $\tilde{F}(t,\tau)B(\tau)$ and $\tilde{F}(t,\tau)C(\tau)$ are continuous on $[0,T]$ except at the point $\tau=t$ and hence $\tilde{F}(t,\tau)B(\tau)\in L^2\left([0,T],\mathbb{R}^{n\times m}\right)$ and $\tilde{F}(t,\tau)C(\tau)\in L^2\left([0,T],\mathbb{R}^{n\times r}\right)$. Thus, for each $t\in[0,T]$, we have

$$\lim_{n\to\infty}x_n(t|u_n,w_n)=x(t|u,w). \tag{14}$$

On the other hand,

$$\begin{aligned}|x(t|u_n,w_n)|=&\left|F(t,0)x_0+\int_0^t F(t,\tau)B(\tau)u_n(\tau)d\tau+\right.\\&\left.\int_0^t F(t,\tau)C(\tau)w_n(\tau)d\tau\right|\\\le&|F(t,0)x_0|+\left|\int_0^t F(t,\tau)B(\tau)u_n(\tau)d\tau\right|+\\&\left|\int_0^t F(t,\tau)C(\tau)w_n(\tau)d\tau\right|\\\le&|F(t,0)x_0|+\left[\sum_{i=1}^m\left(\int_0^t\left((F(t,\tau)B(\tau))_i\right)^2d\tau\right)\right]^{1/2}\\&\left[\sum_{i=1}^m\int_0^T\left(u_{n,i}(\tau)\right)^2d\tau\right]^{1/2}\\&+\left[\sum_{i=1}^r\int_0^t\left((F(t,\tau)C(\tau))_i\right)^2d\tau\right]^{1/2}\\&\left[\sum_{i=1}^r\int_0^T\left(w_{n,i}(\tau)\right)^2d\tau\right]^{1/2},\end{aligned}$$

where $(F(t,\tau)B(\tau))_i$ is the $i$-th element of $F(t,\tau)B(\tau)$. By the continuity of $\int_0^t\left((F(t,\tau)B(\tau))_i\right)^2d\tau$, $\int_0^t\left((F(t,\tau)C(\tau))_i\right)^2d\tau$ and $F(t,0)x_0$, there exists a $\rho$ such that

$$\begin{aligned}\rho=\max_{i=1,\ldots,m;\,j=1,\ldots,r;\,t\in[0,T]}\Bigg\{&\int_0^t\left((F(t,\tau)B(\tau))_i\right)^2d\tau,\\&\int_0^t\left((F(t,\tau)C(\tau))_j\right)^2d\tau,|F(t,0)x_0|\Bigg\}.\end{aligned}$$

It follows that

$$|x(t|u_n, w_n)| \leq \rho + \rho^{1/2}\left(\|u_n\|_{L^2} + \|w_n\|_{L^2}\right) \leq \rho + \rho^{1/2}(\zeta + \eta), \ \forall t \in [0, T].$$

Since $Q(t)$ is continuous on $[0, T]$ and is positive definite for each $t \in [0, T]$, we have, for any $t \in [0, T]$,

$$0 \leq (x(t|u_n, w_n))^T Q(t) x(t|u_n, w_n) \leq \max_{i,j=1,\ldots,n;t\in[0,T]} \left|Q_{i,j}(t)\right| \left(\rho + \rho^{1/2}(\zeta + \eta)\right)^2.$$

Therefore, by Lebesgue Dominated Convergence Theorem (Theorem 2.6.4 in [14]), it holds that

$$\begin{aligned}
&\lim_{n\to\infty} \int_0^T (x(t|u_n, w_n))^T Q(t) x(t|u_n, w_n)\, dt \\
&= \int_0^T \lim_{n\to\infty} (x(t|u_n, w_n))^T Q(t) x(t|u_n, w_n)\, dt \\
&= \int_0^T (x(t|u, w))^T Q(t) x(t|u, w)\, dt. \qquad (15)
\end{aligned}$$

By virtue of (14) with $t = T$ and (15), we obtain

$$\lim_{n\to\infty} \mathscr{F}(u_n, w_n) = \mathscr{F}(u, w).$$

This completes the proof.

From Lemma 1 and Lemma 2, we have the following lemma.

**Lemma 3.** *If $u_n \rightharpoonup u$ and $w_n \rightharpoonup w$, where $\{u_n, w_n\} \subset \mathscr{U} \times \mathscr{W}$, then,*

$$(u, w) \in \mathscr{U} \times \mathscr{W} \text{ and } J(u, w) \leq \varliminf_{n\to\infty} J(u_n, w_n).$$

*If $u_n \to u$ and $w_n \rightharpoonup w$, where $\{u_n, w_n\} \subset \mathscr{U} \times \mathscr{W}$, then,*

$$(u, w) \in \mathscr{U} \times \mathscr{W} \text{ and } J(u, w) = \lim_{n\to\infty} J(u_n, w_n).$$

Now we have the following main theorem in this section.

**Theorem 1.** *Consider Problem $(P)$. Then, there exists a $(u^*, w^*) \in \mathscr{U} \times \mathscr{W}$ such that*

$$J(u^*, w^*) = \min_{u\in\mathscr{U}} \max_{w\in\mathscr{W}} J(u, w). \qquad (16)$$

*Proof.* Note that $L^2([0,T],\mathbb{R}^r)$ is reflexive and $\mathscr{U}$ is a compact and convex set. It follows that $\mathscr{U}$ is weakly sequentially compact. To prove (16), it suffices, by Proposition 38.12 in [19], to prove that

$$J(u) = \max_{w\in\mathscr{W}} J(u,w)$$

is weakly sequentially lower semi-continuous. That is to say, we only need to prove

$$J(u) \le \varliminf_{n\to\infty} J(u_n) \text{ when } u_n \rightharpoonup u. \tag{17}$$

Suppose that $u_n \rightharpoonup u$. From Lemma 3, we know that

$$J(u,w) \le \varliminf_{n\to\infty} J(u_n,w), \text{ for any } w\in\mathscr{W}.$$

Clearly,

$$\max_{w\in\mathscr{W}} J(u_n,w) \ge J(u_n,w).$$

It follows that

$$J(u,w) \le \varliminf_{n\to\infty} J(u_n,w) \le \varliminf_{n\to\infty} \max_{w\in\mathscr{W}} J(u_n,w), \text{ for any } w\in\mathscr{W}.$$

Thus,

$$J(u) = \max_{w\in\mathscr{W}} J(u,w) \le \varliminf_{n\to\infty} \max_{w\in\mathscr{W}} J(u_n,w) = \varliminf_{n\to\infty} J(u_n,w(u_n)) = \varliminf_{n\to\infty} J(u_n).$$

This completes the proof.

## 4 Problem Approximation

Consider a monotonically non-decreasing sequence $\{S^p\}_{p=1}^{\infty}$ of finite subsets of $[0,T]$. For each $p$, let $n_p+1$ points of $S^p$ be denoted by $t_0^p, t_1^p, \ldots, t_{n_p}^p$. These points are chosen such that

$$t_0^p = 0, t_{n_p}^p = T, \text{ and } t_{k-1}^p < t_k^p, k = 1,2,\ldots,n_p.$$

Thus, associated with each $S^P$ there is the obvious partition $\mathscr{I}^p$ of $[0,T]$ defined by

$$\mathscr{I}^p = \{I_k^p : k = 1,\ldots,n_p\},$$

where $I_k^p = [t_{k-1}^p, t_k^p)$.

We choose $S^P$ such that $\lim_{p\to\infty} S^p$ is dense in $[0, T]$, that is

$$\lim_{p\to\infty}\max_{k=1,\ldots,n_p}\left|I_k^p\right| = 0,$$

where $\left|I_k^p\right| = t_k^p - t_{k-1}^p$, the length of the $kth$ interval.

Let

$$u^p(t) = \sum_{k=1}^{n_p}\sigma^{p,k}\chi_{I_k^p}(t), \tag{18}$$

$$w^p(t) = \sum_{k=1}^{n_p}\theta^{p,k}\chi_{I_k^p}(t), \tag{19}$$

and

$$\sigma^p = [(\sigma^{p,1})^T, \ldots, (\sigma^{p,n_p})^T]^T \text{ and } \theta^p = [(\theta^{p,1})^T, \ldots, (\theta^{p,n_p})^T]^T,$$

where

$$\sigma^{p,k} = \left[\sigma_1^{p,k}, \ldots, \sigma_m^{p,k}\right]^T, \text{ and } \theta^{p,k} = \left[\theta_1^{p,k}, \ldots, \theta_r^{p,k}\right]^T,$$

$\chi_I$ denotes the indicator function of $I$ defined by

$$\chi_I(t) = \begin{cases} 1, & t \in I, \\ 0, & \text{elsewhere.} \end{cases}$$

Define

$$\Pi^p = \left\{\sigma^p \in \mathbb{R}^{mn_p} : \left(\sigma^p\right)^T U^P \sigma^p \le \eta^2\right\}, \tag{20}$$

$$\Xi^p = \left\{\theta^p \in \mathbb{R}^{rn_p} : \left(\theta^p\right)^T W^P \theta^p \le \zeta^2\right\}, \tag{21}$$

$$\mathscr{U}^p = \left\{u^p(t) = \sum_{k=1}^{n_p}\sigma^{p,k}\chi_{I_k^p}(t) : \sigma^p \in \Pi^p\right\},$$

and

$$\mathscr{W}^p = \left\{w^p(t) = \sum_{k=1}^{n_p}\theta^{p,k}\chi_{I_k^p}(t) : \theta^p \in \Xi^p\right\},$$

where

$$U^P = diag\left(\left|I_1^p\right| I_{m\times m}, \left|I_2^p\right| I_{m\times m}, \ldots, \left|I_{n_p}^p\right| I_{m\times m}\right),$$

and

$$W^P = diag\left(\left|I_1^p\right| I_{r\times r}, \left|I_2^p\right| I_{r\times r}, \ldots, \left|I_{n_p}^p\right| I_{r\times r}\right).$$

It is clear that $\mathscr{U}^p \subseteq \mathscr{U}$ and $\mathscr{W}^p \subseteq \mathscr{W}$. Furthermore, we have the following lemma.

**Lemma 4.** *For any $u \in \mathscr{U}$ and $w \in \mathscr{W}$, let*

$$u^p(t) = \sum_{j=1}^{n_p} \sigma^{p,j} \chi_{I_j^p}(t) \tag{22}$$

*and*

$$w^p(t) = \sum_{j=1}^{n_p} \theta^{p,j} \chi_{I_j^p}(t), \tag{23}$$

*where*

$$\sigma^{p,j} = \frac{1}{\left|I_j^p\right|} \int_{I_j^p} u(t)dt$$

*and*

$$\theta^{p,j} = \frac{1}{\left|I_j^p\right|} \int_{I_j^p} w(t)dt.$$

*Then, $u^p \in \mathscr{U}^p$ and $w^p \in \mathscr{W}^p$. Furthermore,*

$$u^p \to u \text{ and } w^p \to w. \tag{24}$$

*Proof.* Note that

$$\begin{aligned}
\int_0^T \left(u^p(t)\right)^T u^p(t)dt &= \int_0^T \left(\sum_{j=1}^{n_p} \sigma^{p,j} \chi_{I_j^p}(t)\right)^T \left(\sum_{j=1}^{n_p} \sigma^{p,j} \chi_{I_j^p}(t)\right) dt \\
&= \sum_{j=1}^{n_p} \int_{I_j^p} \left(\sigma^{p,j}\right)^T \sigma^{p,j} dt = \sum_{j=1}^{n_p} \frac{1}{\left|I_j^p\right|} \int_{I_j^p} u^T(t)dt \int_{I_j^p} u(t)dt \\
&\le \sum_{j=1}^{n_p} \frac{1}{\left|I_j^p\right|} \left|I_j^p\right| \int_{I_j^p} u^T(t)u(t)dt = \int_0^T u^T(t)u(t)dt. \qquad (25)
\end{aligned}$$

Thus, $u^p \in \mathscr{U}^p$. In a similar way, we can show that $w^p \in \mathscr{W}^p$. From Lemma 6.4.1 of [14], we have

$$u^p(t) \to u(t), \text{ for almost all } t \in [0, T],$$

and

$$w^p(t) \to w(t), \text{ for almost all} t \in [0, T].$$

Note that $\{u^p\} \times \{w^p\} \subset \mathscr{U} \times \mathscr{W}$ and $u \times w \in \mathscr{U} \times \mathscr{W}$. We have $\|u^p\|_{L^2}^2 \leq \eta^2$ and $\|w^p\|_{L^2}^2 \leq \zeta^2$ for all $p = 1, \ldots$, while $\|u\|_{L^2}^2 \leq \eta^2$ and $\|w\|_{L^2}^2 \leq \zeta^2$. Since $T$ is a finite number, the conclusion of the lemma follows readily.

With $u \in \mathscr{U}^p$ and $w \in \mathscr{W}^p$, the dynamical system (1) becomes

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)\sum_{k=1}^{n_p}\sigma^{p,k}\chi_{I_k^p}(t) + C(t)\sum_{k=1}^{n_p}\theta^{p,k}\chi_{I_k^p}(t), \\ x(0) &= x^0, \end{aligned} \tag{26}$$

and $J(u, w)$ becomes

$$\tilde{J}\left(\sigma^p, \theta^p\right) = (x(T))^T Px(T) + \int_0^T \left\{(x(t))^T Q(t)x(t) + \left(\sum_{k=1}^{n_p}\sigma^{p,k}\chi_{I_k^p}(t)\right)^T R(t)\left(\sum_{k=1}^{n_p}\sigma^{p,k}\chi_{I_k^p}(t)\right)\right\} dt.$$

Now we define the following minimax optimal parameter selection problem.
*Problem* $\left(P_p\right)$ : For the given dynamical system (26), choose $(\sigma^{p,*}, \theta^{p,*}) \in \Pi^p \times \Xi^p$ such that

$$\tilde{J}\left(\sigma^{p,*}, \theta^{p,*}\right) = \min_{\sigma^p \in \Pi^p} \max_{\theta^p \in \Xi^p} \tilde{J}\left(\sigma^p, \theta^p\right).$$

*Remark 1.* Following a similar argument given for the proof of Theorem 1, we can show that for Problem $\left(P_p\right)$, there exists a $(\sigma^{p,*}, \theta^{p,*}) \in \Pi^p \times \Xi^p$ such that

$$J\left(\sigma^{p,*}, \theta^{p,*}\right) = \min_{\sigma^p \in \Pi^p} \max_{\theta^p \in \Xi^p} \tilde{J}\left(\sigma^p, \theta^p\right). \tag{27}$$

**Theorem 2.** *Suppose that* $(u^*, w^*)$ *and* $(\sigma^{p,*}, \theta^{p,*})$ *are the optimal solutions of Problem* $(P)$ *and Problem* $\left(P_p\right)$*, respectively. That is,*

$$J\left(u^*, w^*\right) = \min_{u \in \mathscr{U}} \max_{w \in \mathscr{W}} J(u, w) \text{ and } \tilde{J}\left(\sigma^{p,*}, \theta^{p,*}\right) = \min_{\sigma^p \in \Pi^p} \max_{\theta^p \in \Xi^p} \tilde{J}\left(\sigma^p, \theta^p\right).$$

*Then,*

$$\lim_{p \to \infty} \tilde{J}\left(\sigma^{p,*}, \theta^{p,*}\right) = J\left(u^*, w^*\right). \tag{28}$$

*Proof.* Suppose that (28) is not true. Then, there exists an $\varepsilon_0 > 0$ and a sub-sequence $\{\sigma^{p_k,*}, \theta^{p_k,*}\}$ such that

$$\left|\tilde{J}\left(\sigma^{p_k,*},\theta^{p_k,*}\right)-J\left(u^*,w^*\right)\right|\geq\varepsilon_0. \tag{29}$$

Let $\tilde{u}^{p_k,*}(t)$ and $\tilde{w}^{p_k,*}(t)$ be the piecewise constant functions corresponding to $\sigma^{p_k,*}$ and $\theta^{p_k,*}$ given by (18) and (19), respectively. Clearly, $(\tilde{u}^{p_k,*},\tilde{w}^{p_k,*})\in\mathscr{U}\times\mathscr{W}$ for any $k$. Since $\mathscr{U}\times\mathscr{W}$ is weakly sequentially compact in $L^2\left([0,T],\mathbb{R}^m\right)\times L^2\left([0,T],\mathbb{R}^r\right)$, it follows that $\{(\tilde{u}^{p_k,*},\tilde{w}^{p_k,*})\}_{k=1}^{\infty}$ contains a subsequence, which is denoted by the same sequence, and a $(\tilde{u}^*,\tilde{w}^*)\in\mathscr{U}\times\mathscr{W}$ such that

$$\left(\tilde{u}^{p_k,*},\tilde{w}^{p_k,*}\right)\rightharpoonup\left(\tilde{u}^*,\tilde{w}^*\right). \tag{30}$$

Let $u^{*,p_k}$ and $w^{*,p_k}$ be constructed from $u^*$ and $w^*$ according to (22) and (23), respectively, as follows:

$$u^{*,p_k}(t)=\sum_{j=1}^{n_{p_k}}\sigma^{*,p_k,j}\chi_{I_j^{p_k}}(t),$$

and

$$w^{*,p_k}(t)=\sum_{j=1}^{n_{p_k}}\theta^{*,p_k,j}\chi_{I_j^{p_k}}(t),$$

where

$$\sigma^{*,p_k,j}=\frac{1}{\left|I_j^{p_k}\right|}\int_{I_j^{p_k}}u^*(t)dt$$

and

$$\theta^{*,p_k,j}=\frac{1}{\left|I_j^{p_k}\right|}\int_{I_j^{p_k}}w^*(t)dt.$$

From Lemma 4, we have

$$u^{*,p_k}\to u^*\text{ and }w^{*,p_k}\to w^*.$$

Note that $u^{*,p_k}\to u^*$ and $\tilde{w}^{p_k,*}\rightharpoonup\tilde{w}^*$, it follows from Lemma 3 that

$$J\left(u^{*,p_k},\tilde{w}^{p_k,*}\right)\to J\left(u^*,\tilde{w}^*\right)\text{ as }k\to\infty.$$

Thus, for any $\varepsilon>0$, there exists a constant $K\in\mathbb{N}$, such that

$$J\left(u^{*,p_k},\tilde{w}^{p_k,*}\right)\leq J\left(u^*,\tilde{w}^*\right)+\varepsilon,\forall k>K.$$

Hence, we have

$$\tilde{J}\left(\sigma^{p_k,*},\theta^{p_k,*}\right)=\min_{\sigma^{p_k}\in\Pi^{p_k}}\tilde{J}\left(\sigma^{p_k},\theta^{p_k,*}\right)\le\tilde{J}\left(\sigma^{*,p_k},\theta^{p_k,*}\right)=J\left(u^{*,p_k},\tilde{w}^{p_k,*}\right)$$
$$\le J\left(u^*,\tilde{w}^*\right)+\varepsilon\le\max_{w\in\mathscr{W}}J\left(u^*,w\right)+\varepsilon=J\left(u^*,w^*\right)+\varepsilon.$$

Letting $k\to\infty$, we obtain

$$\overline{\lim_{k\to\infty}}\tilde{J}\left(\sigma^{p_k,*},\theta^{p_k,*}\right)\le J\left(u^*,w^*\right)+\varepsilon.$$

Since $\varepsilon$ is arbitrary, it holds that

$$\overline{\lim_{k\to\infty}}\tilde{J}\left(\sigma^{p_k,*},\theta^{p_k,*}\right)\le J\left(u^*,w^*\right). \tag{31}$$

On the other hand, we note that $\tilde{u}^{p_k,*}(t)\rightharpoonup\tilde{u}^*$ and $w^{*,p_k}\to w^*$. Then, from Lemma 3, we have

$$J\left(u^*,w^*\right)=\min_{u\in\mathscr{U}}J\left(u,w^*\right)\le J\left(\tilde{u}^*,w^*\right)\le\underline{\lim_{k\to\infty}}\tilde{J}\left(\sigma^{p_k,*},\theta^{*,p_k}\right) \tag{32}$$
$$\le\lim_{k\to\infty}\max_{\theta^{p_k}\in\Xi^{p_k}}\tilde{J}\left(\sigma^{p_k,*},\theta^{p_k}\right)=\lim_{k\to\infty}\tilde{J}\left(\sigma^{p_k,*},\theta^{p_k,*}\right).$$

Combining (31) and (32), it holds that

$$\lim_{k\to\infty}\tilde{J}\left(\sigma^{p_k,*},\theta^{p_k,*}\right)=J\left(u^*,w^*\right). \tag{33}$$

(33) is a contradictory to (29). Thus, (28) is true. This completes the proof.

Next, we have the following theorem.

**Theorem 3.** *Suppose that $(\sigma^{p,*},\theta^{p,*})$ is the optimal solution of Problem $(P_p)$. Let $u^{p,*}(t)$ and $w^{p,*}(t)$ be the piecewise constant functions corresponding to $\sigma^{p,*}$ and $\theta^{p,*}$ given by (18) and (19), respectively. If there exists a subsequence $\{u^{p_k,*},w^{p_k,*}\}$ such that $u^{p_k,*}\to\bar{u}$ and $w^{p_k,*}\to\bar{w}$, as $k\to\infty$. Then, $(\bar{u},\bar{w})$ is also an optimal solution of Problem $(P)$.*

*Proof.* Since $u^{p_k,*}\to\bar{u}$ and $w^{p_k,*}\to\bar{w}$, as $k\to\infty$, we have

$$\lim_{k\to\infty}J\left(u^{p_k,*},w^{p_k,*}\right)=J\left(\bar{u},\bar{w}\right). \tag{34}$$

For any $u\in\mathscr{U}$, let $u^{p_k}$ be constructed from $u$ according to (22). Then, by Lemma 4, $u^{p_k}\in\mathscr{U}$ and $u^{p_k}\to u$ as $k\to\infty$. Now, as $w^{p_k,*}\to\bar{w}$, we have

$$\lim_{k\to\infty}J\left(u^{p_k},w^{p_k,*}\right)=J\left(u,\bar{w}\right). \tag{35}$$

From (34) and (35), it follows that for any $\varepsilon>0$, there exists a $K\in\mathbb{N}$ such that for any $k>K$, we have

$$J\left(u^{p_k,*}, w^{p_k,*}\right) - \varepsilon \leq J\left(\bar{u}, \bar{w}\right) \leq J\left(u^{p_k,*}, w^{p_k,*}\right) + \varepsilon.$$

Thus,

$$J\left(\bar{u}, \bar{w}\right) \leq J\left(u^{p_k,*}, w^{p_k,*}\right) + \varepsilon \leq J\left(u^{p_k}, w^{p_k,*}\right) + \varepsilon.$$

Letting $k \to \infty$, we have

$$J\left(\bar{u}, \bar{w}\right) \leq J\left(u, \bar{w}\right) + \varepsilon.$$

Since $\varepsilon$ is arbitrary, it holds that

$$J\left(\bar{u}, \bar{w}\right) \leq J\left(u, \bar{w}\right), \forall u \in \mathscr{U}. \tag{36}$$

In a similar way, we can prove that

$$J\left(\bar{u}, w\right) \leq J\left(\bar{u}, \bar{w}\right), \ \forall w \in \mathscr{W}. \tag{37}$$

Combining (36) and (37), we conclude that $(\bar{u}, \bar{w})$ is an optimal solution of Problem $(P)$.

*Remark 2.* By a close examination of Theorems 2, we see that it suggests a method for solving Problem $(P)$. First, choose an integer $p \geq 2$, select a partition of the interval $[0, T]$ and solve Problem $\left(P_p\right)$. Then, increasing $p$ and using the solution obtained from the previous step as the initial guess, solve Problem $\left(P_p\right)$ again. This process is repeated until the change of the optimal value of the cost function is within a desired tolerance.

## 5 Sub-problem Solution

From Remark 2, we know that Problem $(P)$ can be solved by solving a sequence of Problem $\left(P_p\right)$. However, Problem $\left(P_p\right)$ is still a minimax optimal parameter selection problem, and hence, it is hard to solve. In this section, we will develop a computational method to solve Problem $\left(P_p\right)$.

For each given $\sigma^p$ and $\theta^p$, the solution of (26) can be rewritten as:

$$x\left(t\right) = F\left(t, 0\right) x^0 + \tilde{F}_1\left(t\right) \sigma^p + \tilde{F}_2\left(t\right) \theta^p,$$

where $\tilde{F}_1 \in \mathbb{R}^{n \times m n_p}$ and $\tilde{F}_2 \in \mathbb{R}^{n \times r n_p}$ are, respectively, defined by

$$\tilde{F}_1\left(t\right) = \left[\int_0^t F\left(t, \tau\right) B\left(\tau\right) \chi_{I_1^p}(\tau) d\tau, \int_0^t F\left(t, \tau\right) B\left(\tau\right) \chi_{I_2^p}(\tau) d\tau, \ldots, \int_0^t F\left(t, \tau\right) B\left(\tau\right) \chi_{I_{n_p}^p}(\tau) d\tau\right]$$

and

$$\tilde{F}_2(t) = \left[\int_0^t F(t,\tau) C(\tau) \chi_{I_1^p}(\tau) d\tau, \int_0^t F(t,\tau) C(\tau) \chi_{I_2^p}(\tau) d\tau, \ldots, \int_0^t F(t,\tau) C(\tau) \chi_{I_{n_p}^p}(\tau) d\tau\right].$$

Thus, $\tilde{J}(\sigma^p, \theta^p)$ can be rewritten as

$$\tilde{J}(\sigma^p, \theta^p) = (\sigma^p)^T G_1 \sigma^p + (\theta^p)^T G_2 \theta^p + 2h_1^T \sigma^p + 2h_2^T \theta^p + 2(\sigma^p)^T G_3 \theta^p + c_0, \tag{38}$$

where

$$\begin{aligned}
G_1 &= \left(\tilde{F}_1(T)\right)^T P \tilde{F}_1(T) + \int_0^T \left(\tilde{F}_1(t)\right)^T Q(t) \tilde{F}_1(t) dt + \int_0^T \left(\bar{F}(t)\right)^T R(t) \bar{F}(t) dt, \\
\bar{F}(t) &= \left[\chi_{I_1^p}(t) I_{r\times r}, \chi_{I_2^p}(t) I_{r\times r}, \ldots, \chi_{I_{n_p}^p}(\tau) I_{r\times r}\right], \\
G_2 &= \left(\tilde{F}_2(T)\right)^T P \tilde{F}_2(T) + \int_0^T \left(\tilde{F}_2(t)\right)^T Q(t) \tilde{F}_2(t) dt, \\
G_3 &= \left(\tilde{F}_1(T)\right)^T P \tilde{F}_2(T) + \int_0^T \left(\tilde{F}_1(t)\right)^T Q(t) \tilde{F}_2(t) dt, \\
h_1 &= \left(F(T,0) x^0\right)^T P \tilde{F}_1(T) + \int_0^T \left(F(t,0) x^0\right)^T Q(t) \tilde{F}_1(t) dt, \\
h_2 &= \left(F(T,0) x^0\right)^T P \tilde{F}_2(T) + \int_0^T \left(F(t,0) x^0\right)^T Q(t) \tilde{F}_2(t) dt,
\end{aligned}$$

and

$$c_0 = \left(F(T,0) x^0\right)^T P F(T,0) x^0 + \int_0^T \left(F(t,0) x^0\right)^T Q(t) F(t,0) x^0 dt.$$

It is easy to verify that $G_1$ and $G_2$ are positive definite. Let $\tilde{\sigma}^p = G_1^{1/2}\sigma^p + G_1^{-1/2} h_1$. Then, Problem $(P_p)$ becomes Problem $(\bar{P}_p)$ with a difference of a constant in the cost, which is defined as follows:
*Problem*$(\bar{P}_p)$ : choose $(\tilde{\sigma}^{p,*}, \theta^{p,*}) \in \tilde{\Pi}^p \times \Xi^p$ such that

$$\bar{J}(\tilde{\sigma}^{p,*}, \theta^{p,*}) = \min_{\tilde{\sigma}^p \in \tilde{\Pi}^p} \max_{\theta^p \in \Xi^p} \bar{J}(\tilde{\sigma}^p, \theta^p) = (\tilde{\sigma}^p)^T \tilde{\sigma}^p + (\theta^p)^T G_2 \theta^p + 2\tilde{h}_2^T \theta^p + 2(\tilde{\sigma}^p)^T \tilde{G}_3 \theta^p, \tag{39}$$

where $\tilde{G}_3 = G_1^{-1/2} G_3$, $\tilde{h}_2 = h_2 - G_3^T G_1^{-1} h_1$ and

$$\tilde{\Pi}^p = \left\{ \tilde{\sigma}^p \in \mathbb{R}^{mn_p} : \left( G_1^{-1/2} \tilde{\sigma}^p - h_1 \right)^T U^P \left( G_1^{-1/2} \tilde{\sigma}^p - h_1 \right) \leq \eta^2 \right\}.$$

Now we have the following theorem.

**Theorem 4.** *Problem* $\left(\bar{P}_p\right)$ *is equivalent to the following SDP:*

$$\min_{\tilde{\sigma}^p, \lambda, z} z$$

*subject to*

$$\begin{bmatrix} I & \tilde{\sigma}^p & \tilde{G}_3 \\ (\tilde{\sigma}^p)^T & z - \zeta^2 \lambda & -\tilde{h}_2^T \\ \tilde{G}_3^T & -\tilde{h}_2 & \lambda W^P - G_2 + \tilde{G}_3^T \tilde{G}_3 \end{bmatrix} \succeq 0, \tag{40}$$

*and*

$$\begin{bmatrix} I & \left(U^P\right)^{1/2} G_1^{-1/2} \tilde{\sigma}^p - \left(U^P\right)^{1/2} h_1 \\ \left(\left(U^P\right)^{1/2} G_1^{-1/2} \tilde{\sigma}^p - \left(U^P\right)^{1/2} h_1\right)^T & \eta^2 \end{bmatrix} \succeq 0. \tag{41}$$

*Proof.* Problem $\left(\bar{P}_p\right)$ can be re-written as:

$$\min z$$

subject to

$$z - \left(\tilde{\sigma}^p\right)^T \tilde{\sigma}^p - \left(\theta^p\right)^T G_2 \theta^p - 2\tilde{h}_2^T \theta^p - 2\left(\tilde{\sigma}^p\right)^T \tilde{G}_3 \theta^p \geq 0, \ \forall \theta^p : \left(\theta^p\right)^T W^P \theta^p \leq \zeta^2, \tag{42}$$

$$\left( G_1^{-1/2} \tilde{\sigma}^p - h_1 \right)^T U^P \left( G_1^{-1/2} \tilde{\sigma}^p - h_1 \right) \leq \eta^2. \tag{43}$$

Clearly, (43) is equivalent to (41). For the proof of the equivalence between (42) and (40), it is referred to [3].

With the increase of the partition number $n_p$, the size of (40) becomes very large. It is well-known that solving a large SDP can be computationally expensive. For an alternative approach, we show that Problem $\left(\bar{P}_p\right)$ is equivalent to an optimiztion problem solvable by gradient-based optimization methods. We have the following theorem.

**Theorem 5.** *Problem* $\left(\bar{P}_p\right)$ *is equivalent to the following optimization problem, which is referred to as Problem* $(GP)$.

$$\min_{\lambda,\tilde{\sigma}^p} \hat{J}\left(\lambda,\tilde{\sigma}^p\right) = \zeta^2\lambda + \left(\tilde{\sigma}^p\right)^T\tilde{\sigma}^p + \left(\tilde{h}_2 + \left(\tilde{\sigma}^p\right)^T\tilde{G}_3\right)^T\left(\lambda W^P - G_2\right)^\dagger\left(\tilde{h}_2 + \left(\tilde{\sigma}^p\right)^T\tilde{G}_3\right), \tag{44}$$

*subject to the constraints*

$$\left(G_1^{-1/2}\tilde{\sigma}^p - h_1\right)^T U^P\left(G_1^{-1/2}\tilde{\sigma}^p - h_1\right) \le \eta^2, \tag{45}$$

$$\lambda \ge \lambda_0 = \max_{k=1,\ldots,n_p-1, j=1,\ldots,m} \frac{\gamma_{km+j}}{\left|I_k^p\right|}, \tag{46}$$

*and*

$$\left(I - \left(\lambda W^P - G_2\right)\left(\lambda W^P - G_2\right)^\dagger\right)\left(\tilde{h}_2 + \left(\tilde{\sigma}^p\right)^T\tilde{G}_3\right) = 0, \tag{47}$$

*where* $\left(\lambda W^P - G_2\right)^\dagger$ *is the pseudo-inverse of* $\lambda W^P - G_2$, $\gamma_i$, $i = 1,\ldots,n_p$ *are the eigenvalues of the matrix* $G_2$.

To prove this theorem, we need the following lemma.

**Lemma 5.** *(Theorem 4.3 in [6]) Consider a symmetric matrix* $M = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$. *Then,* $M \succeq 0$ *if and only if*

$$A \succeq 0,\ \left(I - AA^\dagger\right)B = 0,\ C - B^TA^\dagger B \succeq 0.$$

*Proof of Theorem 5*: From [3], we know that the (42) can be replaced by the following homogeneous inequality.

$$t^2\left(z - \left(\tilde{\sigma}^p\right)^T\tilde{\sigma}^p\right) - \left(\theta^p\right)^T G_2\theta^p - 2t\tilde{h}_2^T\theta^p - 2t\left(\tilde{\sigma}^p\right)^T\tilde{G}_3\theta^p \ge 0,$$
$$\forall\theta^p : \left(\theta^p\right)^T W^P\theta^p \le \zeta^2t^2. \tag{48}$$

(48) can be re-written as

$$\left[\left(\tilde{\sigma}^p\right)^T, t\right]\begin{bmatrix} -G_2 & -\tilde{h}_2 - \left(\tilde{\sigma}^p\right)^T\tilde{G}_3 \\ -\tilde{h}_2^T - \tilde{G}_3^T\tilde{\sigma}^p & z - \left(\tilde{\sigma}^p\right)^T\tilde{\sigma}^p \end{bmatrix}\begin{bmatrix} \tilde{\sigma}^p \\ t \end{bmatrix} \ge 0, \forall\theta^p : \left[\left(\tilde{\sigma}^p\right)^T, t\right]$$
$$\begin{bmatrix} -W^P & \\ & \zeta^2 \end{bmatrix}\begin{bmatrix} \tilde{\sigma}^p \\ t \end{bmatrix} \ge 0. \tag{49}$$

Thus, (42) is satisfied if and only if there exists a $\lambda \ge 0$ such that

$$\begin{bmatrix} \lambda W^P - G_2 & -\tilde{h}_2 - \left(\tilde{\sigma}^p\right)^T\tilde{G}_3 \\ -\tilde{h}_2^T - \tilde{G}_3^T\tilde{\sigma}^p & z - \zeta^2\lambda - \left(\tilde{\sigma}^p\right)^T\tilde{\sigma}^p \end{bmatrix} \succeq 0. \tag{50}$$

According to Lemma 5, (50) is satisfied if and only if

$$\lambda W^P - G_2 \succeq 0, \tag{51}$$

$$\left(I - \left(\lambda W^P - G_2\right)\left(\lambda W^P - G_2\right)^{\dagger}\right)\left(\tilde{h}_2 + \left(\tilde{\sigma}^p\right)^T \tilde{G}_3\right) = 0, \tag{52}$$

and

$$z - \zeta^2\lambda - \left(\tilde{\sigma}^p\right)^T \tilde{\sigma}^p - \left(\tilde{h}_2 + \left(\tilde{\sigma}^p\right)^T \tilde{G}_3\right)^T \left(\lambda W^P - G_2\right)^{\dagger} \left(\tilde{h}_2 + \left(\tilde{\sigma}^p\right)^T \tilde{G}_3\right) \geq 0. \tag{53}$$

The condition (51) leads to (46). (53) is equivalent to

$$\zeta^2\lambda + \left(\tilde{\sigma}^p\right)^T \tilde{\sigma}^p + \left(\tilde{h}_2 + \left(\tilde{\sigma}^p\right)^T \tilde{G}_3\right)^T \left(\lambda W^P - G_2\right)^{\dagger} \left(\tilde{h}_2 + \left(\tilde{\sigma}^p\right)^T \tilde{G}_3\right) \leq z. \tag{54}$$

Combining (43), (51), (52) and (54), we obtain that Problem $\left(\bar{P}_p\right)$ is equivalent to Problem $(GP)$. This completes the proof.

*Remark 3.* The constraint (47) in Problem $(GP)$ can be removed if $\lambda > \lambda_0$. This is because

$$I - \left(\lambda W^P - G_2\right)\left(\lambda W^P - G_2\right)^{-1} = 0.$$

Furthermore, we claim that there exists a $\bar{\lambda}$ such that the constraint $\lambda \geq \lambda_0$ in Problem $(GP)$ can be replaced by $\lambda_0 \leq \lambda \leq \bar{\lambda}$. The reason is as follows.

$$\begin{aligned}\lim_{\lambda\to\infty} \hat{J}\left(\lambda, \tilde{\sigma}^p\right) &= \lim_{\lambda\to\infty} \zeta^2\lambda + \left(\tilde{\sigma}^p\right)^T \tilde{\sigma}^p + \left(\tilde{h}_2 + \left(\tilde{\sigma}^p\right)^T \tilde{G}_3\right)^T \left(\lambda W^P - G_2\right)^{\dagger} \\ &\qquad \left(\tilde{h}_2 + \left(\tilde{\sigma}^p\right)^T \tilde{G}_3\right) \\ &\geq \lim_{\lambda\to\infty} \zeta^2\lambda = \infty \text{ for any } \tilde{\sigma}^p \in \tilde{\Pi}^p.\end{aligned}$$

Now we can divide the interval $\left[\lambda_0, \bar{\lambda}\right]$ as two parts $[\lambda_0, \lambda_0 + \epsilon]$ and $\left[\lambda_0 + \epsilon, \bar{\lambda}\right]$ during the process of solving Problem $(GP)$, where $\epsilon$ is a small constant. Thus, Problem $(GP)$ can be solved by solving two sub-optimization problems, *i.e.*,

$$\min_{\lambda, \tilde{\sigma}^p} \hat{J}\left(\lambda, \tilde{\sigma}^p\right) = \min\left\{\min_{\lambda_0 \leq \lambda \leq \lambda_0 + \epsilon} \min_{\tilde{\sigma}^p} \hat{J}\left(\lambda, \tilde{\sigma}^p\right), \min_{\lambda_0 + \epsilon \leq \lambda \leq \bar{\lambda}} \min_{\tilde{\sigma}^p} \hat{J}\left(\lambda, \tilde{\sigma}^p\right)\right\}.$$

For each fixed $\lambda \in [\lambda_0, \lambda_0 + \epsilon]$, $\tilde{\sigma}^p$ is obtained by only solving a convex optimization with quadratic cost function and a quadratic constraint. During the minimization process of $\min_{\lambda_0 + \epsilon \leq \lambda \leq \bar{\lambda}} \min_{\tilde{\sigma}^p} \hat{J}\left(\lambda, \tilde{\sigma}^p\right)$, $\left(\lambda W^P - G_2\right)^{\dagger}$ in (44) can be replaced by $\left(\lambda W^P - G_2\right)^{-1}$. In this case, the gradient $\frac{\partial \hat{J}(\lambda, \tilde{\sigma}^p)}{\partial \lambda}$ is easily obtained. By direct ver-

ification, we know that $\hat{J}(\lambda, \tilde{\sigma}^p)$ is convex with respect to $\lambda$ and $\tilde{\sigma}^p$, respectively. In view of this property, a bi-iterative method can be applied. First $\lambda$ is fixed and $\hat{J}(\lambda, \tilde{\sigma}^p)$ is minimized with respect to $\tilde{\sigma}^p$. Then, $\tilde{\sigma}^p$ is fixed as the one obtained in the previous step, and $\hat{J}(\lambda, \tilde{\sigma}^p)$ is minimized with respect to $\lambda$. This process is repeated until the change of the cost is within the given tolerance. The remaining question is how to solve the minimization problem $\min_{\lambda_0 \le \lambda \le \lambda_0 + \epsilon} \min_{\tilde{\sigma}^p} \hat{J}(\lambda, \tilde{\sigma}^p)$. If $\lambda \to \lambda_0$, then the matrix $\lambda W^P - G_2$ becomes singular. Thus, it is important to develop an efficient computation method for this case. It is a future research problem. Nevertheless, Theorem 5 does offer a possible way to solve Problem $\left(P_p\right)$.

## 6 Numerical Experiment

In the following computation, the computer routines are implemented in a Matlab environment and the software packages SeDuMi [13] and YALMIP [8] are used.

To illustrate our developed method, let us consider the following example, where the dynamical system is given by

$$\begin{aligned} \dot{x}_1(t) &= 2x_1 + x_2 + u_1 + w_1, \\ \dot{x}_2(t) &= x_2 + u_2 + w_2, \end{aligned} \tag{55}$$

with the given initial condition

$$x_1(0) = 1, x_2(0) = -1.$$

$$\mathscr{W}_\zeta = \left\{ w \in L^2\left([0,1], \mathbb{R}^2\right) : \|w\|_{L^2}^2 = \int_0^1 (w(t))^T w(t)\, dt \le \zeta^2 \right\},$$

and

$$\mathscr{U}_\eta = \left\{ u \in L^2\left([0,1], \mathbb{R}^2\right) : \|u\|_{L^2}^2 = \int_0^1 (u(t))^T u(t)\, dt \le \eta^2 \right\}.$$

Since

$$A = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix},$$

it is clear that

$$F(t,\tau) = \begin{bmatrix} e^{2(t-\tau)} & (t-\tau)\, e^{2(t-\tau)} \\ 0 & e^{2(t-\tau)} \end{bmatrix}.$$

Now we consider the following optimization problem

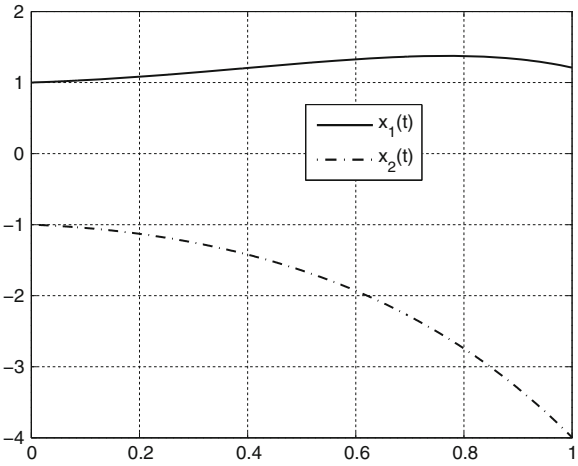



**Fig. 1** The optimal state without the disturbance

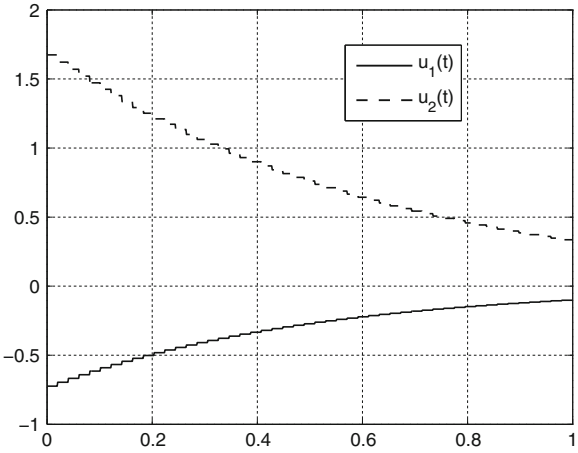



**Fig. 2** The optimal control without the disturbance

$$\min_{u\in\mathscr{U}_\eta}\max_{w\in\mathscr{W}_\zeta} J(u,w) = (x(1))^T x(1) + 0.05\int_0^1 (u(t))^T u(t)\,dt.$$

We first consider the case when there is no disturbances, *i.e.*, $w_1 = w_2 = 0$ in (55). For this case, we use MISER [7] with 50 equally spaced knots to solve it. The state and the optimal control obtained are depicted in Figs. 1, 2. The corresponding optimal cost is 17.4711.

For the case with disturbances, we let $\zeta = 0.1$ and $\eta = 1$. During the computational process, all the partitions of [0, 1] are equally spaced. Then, the computational results of Problem $(\bar{P}_p)$ with different $n_p$ are given in Table 1. From Table 1, we can

**Table 1** The computational results of problem $(\bar{P}_p)$ with different $n_p$

| | $z$ | $\lambda$ |
|---|---|---|
| $n_p = 25$ | 0.508171679707355 | 29.210423777966025 |
| $n_p = 50$ | 0.288402123372642 | 28.78547097896896 |
| $n_p = 100$ | 0.288402115236866 | 28.789910948034414 |

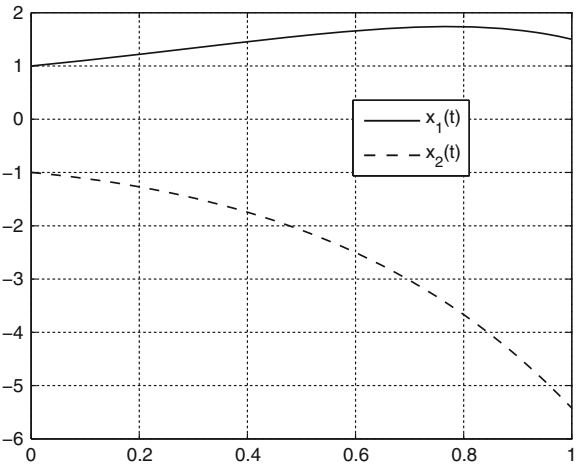



**Fig. 3** The optimal state $x(t)$ with disturbances

see that when the number of sub-intervals is doubled from 50 to 100, the change of $z$ is smaller than $10^{-6}$. Thus, we take $n_p = 50$ as the optimal solution. The corresponding optimal states without disturbances is plotted in Fig. 3. The corresponding piecewise constant optimal control elements are depicted in Figs. 4, 5. The obtained optimal cost in this case is 33.3236. Clearly, there is a significant increase of the optimal cost value when disturbances are presented.

Now let us examine how the optimal cost is changed with reference to different $\zeta$ and $\eta$. Taking $n_p = 100$ with the partition points being equally spaced, the results of Problem $(\bar{P}_p)$ corresponding to different $\zeta$ and $\eta$ are presented in Table 2. From Table 2, we can see that for the same $\zeta$, the larger the $\eta$, the smaller the $z$ is obtained. On the other hand, for the same $\eta$, the larger the $\zeta$, the larger the $z$ is obtained. From our computation, we know that a large $z$ always gives rise to a large optimal cost of Problem $(P)$. Thus, a large $\zeta$ will lead to a large cost function value of Problem $(P)$ and a large $\eta$ will lead to a small cost function value of Problem $(P)$. This is expected, because if the feasible set is enlarged, then the optimal cost is reduced.

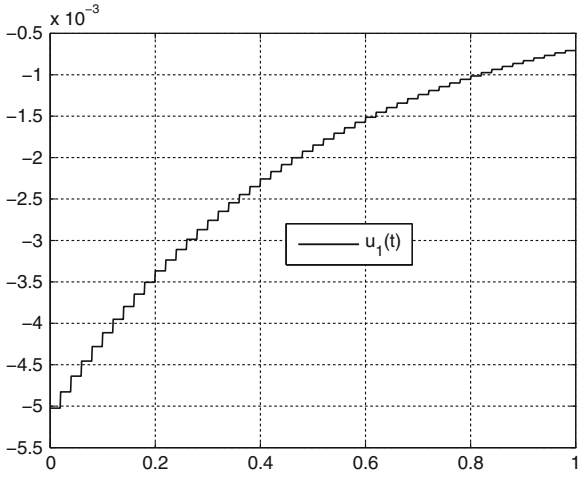



**Fig. 4** The optimal control $u_1(t)$ of Problem (P) with disturbance

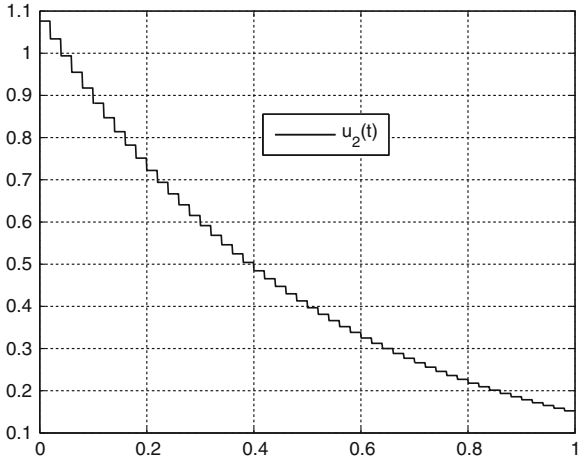



**Fig. 5** The optimal control $u_2(t)$ of Problem (P) with disturbance

**Table 2** The computational results of problem $(\bar{P}_p)$ with different $\zeta$ and $\eta$

| $\zeta$ | $\eta$ | $z$ |
|---|---|---|
| 0.1 | 1 | 0.288402115236866 |
| 0.2 | 1 | 1.151966252722664 |
| 0.5 | 1 | 7.196915158138601 |
| 0.1 | 5 | 0.184774427787928 |
| 0.2 | 5 | 0.737455469828136 |
| 0.5 | 5 | 5.829605274519559 |

## 7 Concluding Remarks and Future Research

In this chapter, we have shown that an infinite-dimensional minimax optimal control problem can be approximated by a sequence of finite-dimensional minimax optimal parameter selection problems. Furthermore, these finite dimensional minimax optimal parameter selection problems can be transformed into either SDPs or standard minimization problems. For SDP, it is easily solved by available software packages. However, an efficient method for Problem $(GP)$ is still not available since the matrix $\lambda W^P - G_2$ becomes singular at the point $\lambda = \lambda_0$. It remains an open question.

**Acknowledgments** Changzhi Wu was partially supported by NSFC 11001288, the key project of Chinese Ministry of Education 210179, and the project from Chongqing Nature Science Foundation cstc2013jjB0149 and cstc2013jcyjA1338.

## References

1. Basar T, Bernhard P (1991) $H^{\infty}$-optimal control and related minimax design problems. Birkhauser, New Jersey
2. Bemporad A, Borrelli F, Morari M (2003) Min-max control of constrained uncertain discrete-time linear systems. IEEE Trans Autom Control 48:1600–1606
3. Bertsimas D, Brown DB (2007) Constrained stochastic LQC: a tractable approach. IEEE Trans Autom Control 52:1826–1841
4. Chernousko FL (2005) Minimax control for a class of linear systems subject to disturbances. J Optim Theory Appl 127:535–548
5. Fisher J, Bhattacharya R (2009) Linear quadratic regulation of systems with stochastic parameter uncertainties. Automatica 45:2831–2841
6. Gallier J (2011) Schur complements and applications, Geom Methods Appl, Texts in Applied Mathematics 38:431–437.
7. Jennings LS, Teo KL, Fisher ME, Goh CJ (2005) MISER version 3, optimal control software, theory and user manual. Department of Mathematics, University of Western Australia.http://www.cado.uwa.edu.au/miser/OCT.html
8. Johan LG (2004) YALMIP: a toolbox for modelling and optimization in Matlab. In: IEEE international symposium on computer aided control systems design, Taipei, Taiwan, pp 284–289
9. Kostyukova O, Kostina E (2006) Robust optimal feedback for terminal linear-quadratic control problems under disturbances. Math Prog 107:131–153
10. Mayne DQ, Eschroeder WR (1997) Robust time-optimal control of constrained linear systems. Automatica 33:2103–2118
11. Rakocic SV, Kerrigan EC, Mayne DQ, Kouramas KI (2007) Optimized robust control invariance for linear discrete-time systems: theoretical foundations. Automatica 43:831–841
12. Scokaert POM, Mayne DQ (1998) Min-max feedback model predictive control for constrained linear systems. IEEE Trans Autom Control 43:1136–1142
13. Sturm JF (1998) Using SeDuMi 1.02, A Matlab toolbox for optimization over symmetric cones, Optim Methods Softw 11:625–653
14. Teo KL, Goh CJ, Wong KH (1991) A unified computational approach to optimal control problems. Longman Scientific and Technical, London
15. Vinter RB (2005) Minimax optimal control. SIAM J Control Optim 44:939–968
16. Wu CZ, Wu SY, Teo KL (2013) Min-max optimal control of linear systems with uncertainty and terminal state constraints, Automatica, 49:1809–1815

17. Yao DD, Zhang S, Zhou XY (2004) Stochastic linear-quadratic control via semidefinite programming. SIAM J Control Optim 40:801–823
18. Yoon MG, Ugrinovskii VA, Pertersen IR (2005) On the worst-case disturbance of minimax optimal control. Automatica 41:847–855
19. Zeidler E (1985) Nonlinear functional analysis and its applications III, variational methods and optimization. Springer, New York

# A Linearly-Growing Conversion from the Set Splitting Problem to the Directed Hamiltonian Cycle Problem

**Michael Haythorpe and Jerzy A. Filar**

**Abstract** We consider a direct conversion of the, classical, set splitting problem to the directed Hamiltonian cycle problem. A constructive procedure for such a conversion is given, and it is shown that the input size of the converted instance is a linear function of the input size of the original instance. A proof that the two instances are equivalent is given, and a procedure for identifying a solution to the original instance from a solution of the converted instance is also provided. We conclude with two examples of set splitting problem instances, one with solutions and one without, and display the corresponding instances of the directed Hamiltonian cycle problem, along with a solution in the first example.

## 1 Introduction

The set splitting problem (SSP) is a famous decision problem that can be simply stated: given a finite universe set $\mathscr{U}$, and a family $\mathscr{S}$ of subsets of $\mathscr{U}$, decide whether there exists a partition of $\mathscr{U}$ into two, disjoint, non-empty subsets $\mathscr{U}_1$ and $\mathscr{U}_2$ such that every subset $S^i \in \mathscr{S}$ is *split* by this partition. That is, for each subset $S^i \in \mathscr{S}$, we have $S^i \not\subset \mathscr{U}_1$ and $S^i \not\subset \mathscr{U}_2$. If such a partition exists, we call it a *solution* of the SSP instance, and say that the decision of the instance is YES. Similarly, if no such partition exists, then the decision of the instance is NO.

This problem has been studied by such distinguished mathematicians as Erdős [9] and Miller [14] since the 1930s. Since then, it has been studied by many authors in the mathematics, computer science, and engineering communities. It has acquired a theoretical interest by virtue of its relationship to hypergraph colourability problems

M. Haythorpe (✉) · J. A. Filar
Flinders University, Adelaide, Australia
e-mail: michael.haythorpe@flinders.edu.au

J. A. Filar
e-mail: jerzy.filar@flinders.edu.au

H. Xu and X. Wang (eds.), *Optimization and Control Methods in Industrial Engineering and Construction*, Intelligent Systems, Control and Automation: Science and Engineering 72,
DOI: 10.1007/978-94-017-8044-5_3, © Springer Science+Business Media Dordrecht 2014

(e.g. see Radhakrishnan and Srinivasan [15]). In addition, it has applicability in modern lines of research such as DNA computing (e.g. see Chang et al. [4]), and several recent algorithms for solving SSP have been developed (e.g. see Dehne et al. [7], Chen and Lu [5], Lokshtanov and Saurabh [13]).

SSP is known to be NP-complete [10]. One key feature of NP-complete problems is that an instance of any one NP-complete problem can be converted to an instance of any other NP-complete problem, in such a way that the two instances have the same answer, and the cardinalities of the variables sets in the second instance are polynomial functions of the size of input data for the original instance. The study of NP-complete problems originated with Cook [6], who proved that an instance of any problem in the set of NP decision problems can be converted to an equivalent instance of the boolean satisfiability problem (SAT). Therefore, SAT was the first problem proven to be NP-complete. Then, if any NP-complete problem $P_1$ can be *converted* to another problem $P_2$, the second problem $P_2$ is also proved to be NP-complete. This is because any instance of $P_2$ can be converted to an instance of SAT (via Cook's theorem), and from there converted (possibly through multiple other problems) to an instance of $P_1$.

Cook's breakthrough approach provided the template for NP-complete conversions, and subsequently it has become commonplace for problems to be converted to SAT. A recent study of this may be seen in Kugele [12]. However, there is nothing inherently special about SAT to set it apart from other fundamental NP-complete problems. Motivated by this line of thinking, in this chapter we investigate the conversion of SSP to another fundamental NP-complete problem, namely the directed Hamiltonian cycle problem (HCP). Directed HCP can be described simply: given a graph $\Gamma$ containing a set of vertices $V$, such that $|V| = N$, and a set of directed edges $E$, decide whether there exists a simple cycle of length $N$ in the graph $\Gamma$, or not. Directed HCP was one of the earliest known NP-complete problems [11] and is a classical graph theory problem which has been the subject of investigation for well over a century. Indeed, a famous instance of HCP—the so-called "Knight's tour" problem—was solved by Euler in the 1750s, and it remains an area of active research (e.g. see Eppstein [8], Borkar et al. [3], and Baniasadi et al. [2]).

Arguably, it is interesting to consider what might be called "linear orbits" of famous NP-complete problems, such as directed HCP. By this, we mean the set of other NP-complete problems which may be converted to, say, directed HCP in such a way that the input size of the resultant HCP instance is a linear function of the input size of the original problem instance. We refer to such a conversion as a *linearly-growing* conversion. Although conversions between NP-complete problems have been extensively explored since 1971, less attention has been paid to the input sizes of the resultant instances after such conversions, and yet input sizes that grow quadratically or higher are likely to produce intractable instances.

In this chapter, we provide a linearly-growing conversion procedure that accepts any instance of SSP as input, and produces an equivalent instance of directed HCP as output. The equivalence is in the sense that a Hamiltonian in the output graph instance supplies a solution to the original SSP instance, and non-Hamiltonicity in the output instance implies infeasibility of the original SSP instance.

## 2 Simplifying the SSP Instance

Consider an instance of SSP, containing the universe set $\mathscr{U}$ and the family $\mathscr{S}$ of subsets of $\mathscr{U}$. Before we begin solving the problem, we can attempt to simplify it, to obtain a smaller instance that must still have the same answer as the original. The following steps may be performed:

1. If any $S^i \in \mathscr{S}$ contains only a single entry, then the decision of the SSP instance is NO, as this set cannot be split. In this case there is no need to solve the SSP.
2. If any element $u \in \mathscr{U}$ is not contained in any $S^i \in \mathscr{S}$, then it may be removed from $\mathscr{U}$. This is because $u$ could be placed in either partition without affecting the solution, so it is inconsequential to the problem.
3. If any $S^i \in \mathscr{S}$ is equal to $\mathscr{U}$, then it may be disregarded, as any partitioning of $\mathscr{U}$ into non-empty subsets will split $S^i$.
4. If any $S^i \in \mathscr{S}$ is a subset of some other $S^j \in \mathscr{S}$, then $S^j$ may be disregarded, as any partitioning of $\mathscr{U}$ that splits $S^i$ necessarily splits $S^j$ as well.

Once the instance has been simplified in this manner, we say it is in *simple form*, and we are ready to begin converting it to an instance of directed HCP.

## 3 Algorithm for Converting an SSP Instance to an Instance of Directed HCP

For a given instance $\langle \mathscr{U}, \mathscr{S} \rangle$ of SSP we shall construct an instance $\Gamma = \langle V, E \rangle$ of HCP possessing the property that any Hamiltonian cycle corresponds, in a natural way, to a solution of the original instance of SSP. Additionally, in the case the constructed graph does not possess a Hamiltonian cycle, neither does the original instance of SSP have a solution.

The algorithm for constructing $\Gamma$ from $\langle \mathscr{U}, \mathscr{S} \rangle$ has three main steps in which three sets of vertices and edges are constructed. Collectively, these will comprise the vertex and edge sets of $\Gamma$.

Suppose that we have an instance $\langle \mathscr{U}, \mathscr{S} \rangle$ of SSP in simple form. Let $\mathscr{U} = \{1, 2, \ldots, u\}$ denote the universe set, and assume that each $S^i \in \mathscr{S}$ contains entries $s^i_j$ in ascending order. Denote by $s$ the number of subsets remaining after the simplification process, and also define $c := \sum_{i=1}^{s} |S^i|$ to be the total number of elements over all subsets $S^i$. Note that $c \geq u$, where the case $c = u$, trivially, has the answer YES. Then we may define an instance of HCP in the form of a graph $\Gamma$ containing vertices $V$ and directed edges $E$, such that the HCP instance is equivalent to the SSP instance, as follows.

The vertex set $V$ can be partitioned into three mutually disjoint subsets of vertices $V^U$, $V^S$ and $V^C$. That is, $V = V^U \cup V^S \cup V^C$. The subset $V^U$ will contain vertices corresponding to each element of the universe set $\mathscr{U}$. The subset $V^S$ will contain

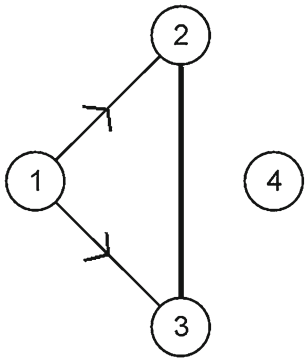



**Fig. 1** One $V^{U,i}$ component. Here, vertex $v_j^{U,i}$ is represented simply by label $j$, for the sake of neatness

vertices corresponding to each subset $S^i \in \mathscr{S}$. The subset $V^C$ will contain two additional "connecting" vertices, that will link the $V^U$ and $V^S$ parts to form a cycle.

Likewise, the edge set $E$ can be partitioned into three mutually disjoint subsets of edges $E^U$, $E^S$ and $E^C$. That is, $E = E^U \cup E^S \cup E^C$. The subset $E^U$ will contain edges whose endpoints lie entirely within $V^U$. Similarly, the subset $E^S$ will contain edges whose endpoints lie entirely within $V^S$. Finally, $E^C$ will contain many "connecting" edges, which connect vertices from any of the three partitions to other vertices, possibly in different partitions.

**The conversion algorithm**

**Step 1:**

We will first consider the vertex subset $V^U$ and edge subset $E^U$. These sets can be further partitioned into $u$ subsets, one for each element in the universe set. That is, $V^U = \bigcup_{i=1}^{u} V^{U,i}$ and $E^U = \bigcup_{i=1}^{u} E^{U,i}$. Then, for each element $i \in \mathscr{U}$, we can describe $V^{U,i}$ and $E^{U,i}$ directly:

$$V^{U,i} = \left\{v_1^{U,i}, v_2^{U,i}, v_3^{U,i}, v_4^{U,i}\right\}$$
$$E^{U,i} = \left\{\left(v_1^{U,i}, v_2^{U,i}\right), \left(v_1^{U,i}, v_3^{U,i}\right), \left(v_2^{U,i}, v_3^{U,i}\right), \left(v_3^{U,i}, v_2^{U,i}\right)\right\}$$

Note that at this stage of the construction, $v_4^{U,i}$ is an isolated vertex. Each subgraph $\langle V^{U,i}, E^{U,i}\rangle$ may be visualised as in Fig. 1. The thick bold undirected edge between vertices $v_2^{U,i}$ and $v_3^{U,i}$ represents directed edges in both directions between the two vertices.

**Step 2:**

We will next consider the vertex subset $V^S$ and edge subset $E^S$. These sets can be further partitioned into $s$ subsets, one for each subset $S^i \in \mathscr{S}$. That is, $V^S = \bigcup_{i=1}^{s} V^{S,i}$ and $E^S = \bigcup_{i=1}^{s} E^{S,i}$. Then, for each subset $S^i \in \mathscr{S}$, we must first determine $|S^i|$. For neatness, when no confusion is possible we define $k = |S^i|$, taking care to remember that the value of $k$ depends on $i$. Then, the number of vertices in $V^{S,i}$ is chosen to be $5 + 6k$, each of which will be denoted by $v_j^{S,i}$. The edge set $E^{S,i}$ is the union of the following three groups of edges:

Group I:

$$\begin{aligned}(v_1^{S,i}, v_6^{S,i}), (v_1^{S,i}, v_{6+3k}^{S,i}), (v_{5+3k}^{S,i}, v_2^{S,i}),\\ (v_{5+6k}^{S,i}, v_2^{S,i}), (v_3^{S,i}, v_4^{S,i}), (v_4^{S,i}, v_3^{S,i}), (v_4^{S,i}, v_5^{S,i}), (v_5^{S,i}, v_4^{S,i}),\end{aligned} \tag{1}$$

Group II: for all $j = 1, \ldots, k$ (for neatness, we define $a_j = 3 + 3j$, $b_j = 4 + 3j$ and $c_j = 5 + 3j$):

$$\begin{aligned}(v_{a_j}^{S,i}, v_{b_j}^{S,i}), (v_{b_j}^{S,i}, v_{a_j}^{S,i}), (v_{b_j}^{S,i}, v_{c_j}^{S,i}), (v_{c_j}^{S,i}, v_{b_j}^{S,i}),\\ (v_{a_j+3k}^{S,i}, v_{b_j+3k}^{S,i}), (v_{b_j+3k}^{S,i}, v_{a_j+3k}^{S,i}), (v_{b_j+3k}^{S,i}, v_{c_j+3k}^{S,i}), (v_{c_j+3k}^{S,i}, v_{b_j+3k}^{S,i}),\end{aligned} \tag{2}$$

Group III: for all $j = 1, \ldots, k-1$ (retaining the definitions of $a_j$, $b_j$ and $c_j$ from above):

$$\begin{aligned}(v_{c_j}^{S,i}, v_{c_j+1}^{S,i}), (v_{c_j}^{S,i}, v_{c_j+1+3k}^{S,i}), (v_{c_j+3k}^{S,i}, v_{c_j+1}^{S,i}), (v_{c_j+3k}^{S,i}, v_{c_j+1+3k}^{S,i}),\\ (v_{c_j}^{S,i}, v_3^{S,i}), (v_3^{S,i}, v_{c_j+1}^{S,i}), (v_{c_j+3k}^{S,i}, v_5^{S,i}), (v_5^{S,i}, v_{c_j+3k+1}^{S,i}).\end{aligned} \tag{3}$$

Each subgraph $\langle V^{S,i}, E^{S,i} \rangle$ has a characteristic visualisation. In Fig. 2 we display such a subgraph for the case where $k = |S^i| = 3$. The thick bold undirected edges represent directed edges in both directions between two vertices. Note that in Fig. 2, the Group I edges are the two directed edges emanating from each of vertex $v_1^{S,i}$ and $v_2^{S,i}$, as well as the undirected edges between vertices $v_3^{S,i}$, $v_4^{S,i}$ and $v_5^{S,i}$. The Group II edges are the undirected edges on the top and bottom of the figure. The Group III edges are all of the directed edges in the interior of the figure.

**Step 3:**

Finally, we consider the vertex subset $V^C$ and edge subset $E^C$. There are only two vertices in $V^C$, namely $v_1^C$ and $v_2^C$. However, there are many edges in $E^C$, and a

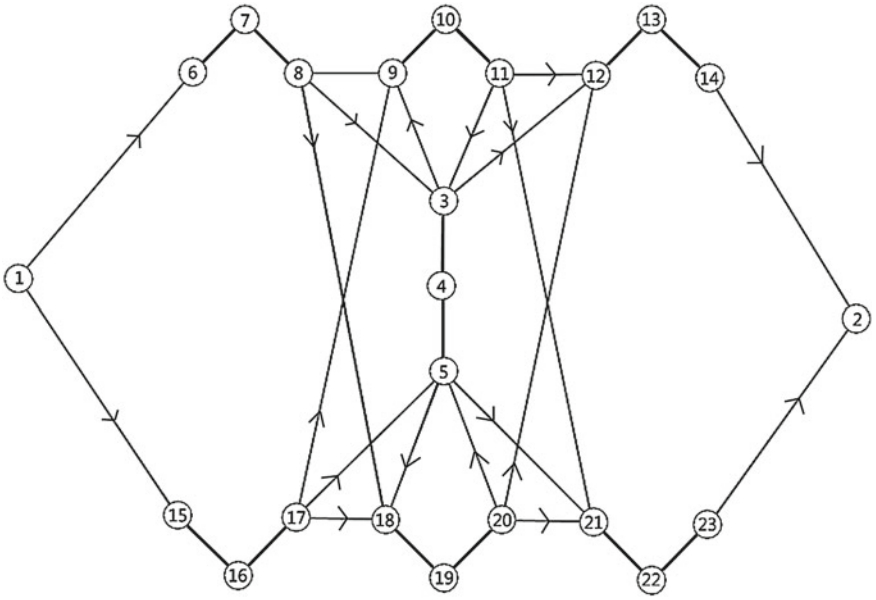



**Fig. 2** One $V^{S,i}$ component. Here, vertex $v_j^{S,i}$ is represented simply by label $j$, for the sake of neatness

procedure must be undertaken to identify them all. First, we include the following edges in $E^C$:

$$(v_4^{U,u}, v_1^C), (v_1^C, v_1^{S,1}), (v_{5+6|S^s|}^{S,s}, v_2^C), (v_2^C, v_1^{U,1}), \tag{4}$$

as well as the following edges for each $i = 1, \ldots, u-1$:

$$(v_4^{U,i}, v_1^{U,i+1}), \tag{5}$$

and also the following edges for each $j = 1, \ldots, s-1$:

$$(v_{5+6|S^j|}^{S,j}, v_1^{S,j+1}). \tag{6}$$

The edges in (4)–(6) link the various components of the graph together. Specifically, the first group of edges links the $V^U$ component to the $V^S$ component, the second group links each $V^{U,i}$ component with the $V^{U,i+1}$ component that follows it, and the third group links each $V^{S,i}$ component with the $V^{S,i+1}$ component that follows it. At this stage of construction, the graph $\Gamma$ can be visualised as in Fig. 3.

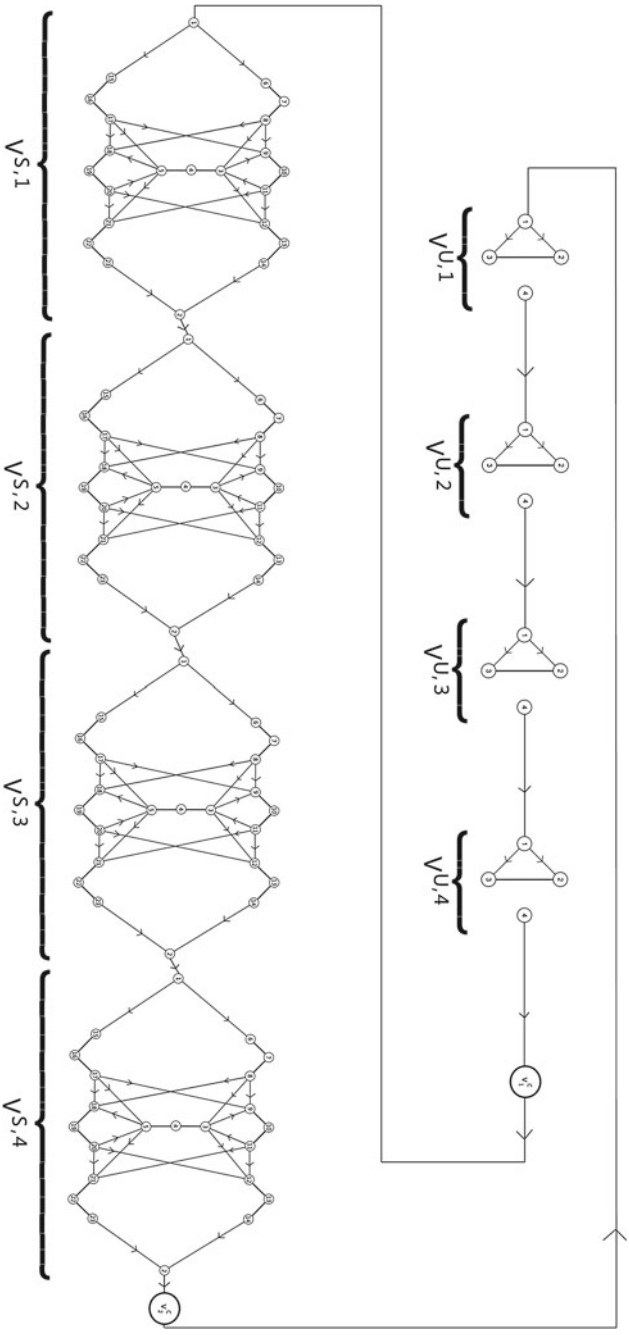



**Fig. 3** The graph $\Gamma$ at an intermediate stage of construction. Note that in this example, $u = s = 4$ and all $|S^i| = 3$

However, the above edges do not comprise all of $E^C$. Additional edges need to be added, as follows. For each $i \in \mathscr{U}$, we undertake the following procedure to insert additional edges in $E^C$. First we identify all subsets in $\mathscr{S}$ which contain the element $i$, and store their indices in a set $F^i$. We also record a set $R^i$, which contains the positions of element $i$ in each subset $S^{F^i_j}$. For example, suppose that the subsets are $S^1 = (1, 3, 6)$, $S^2 = (2, 3, 4)$, $S^3 = (2, 4, 6)$ and $S^4 = (1, 2, 5)$. Then $F^1 = (1, 4)$, and $R^1 = (1, 1)$. Similarly, $F^2 = (2, 3, 4)$, and $R^2 = (1, 1, 2)$. For the sake of neatness, when no confusion is possible, we will define $f = |F^i|$, taking care to remember that $f$ depends on the value of $i$.

For each $i \in \mathscr{U}$, we define $d_{ij} = 3 + 3R^i_j$, and $e_{ij} = 3 + 3|S^{F^i_j}| + 3R^i_j$, and insert the following edges:

$$(v_2^{U,i}, v_{2+e_{i1}}^{S,F^i_1}), (v_{e_{if}}^{S,F^i_f}, v_4^{U,i}), (v_3^{U,i}, v_{2+d_{i1}}^{S,F^i_1}), (v_{d_{if}}^{S,F^i_f}, v_4^{U,i}), \tag{7}$$

into $E^C$. Finally for $i \in \mathscr{U}$, and each $j = 1, \ldots, f - 1$ (retaining the definitions of $f$, $d_{ij}$ and $e_{ij}$ from above), we insert pairs of edges:

$$(v_{e_{ij}}^{S,F^i_j}, v_{2+e_{i,j+1}}^{S,F^i_{j+1}}), (v_{d_{ij}}^{S,F^i_j}, v_{2+d_{i,j+1}}^{S,F^i_{j+1}}), \tag{8}$$

into $E^C$. The edges in (7) and (8) have the effect of creating two paths, that each travel from one of the vertices in $V^{U,i}$ (either $v_2^{U,i}$ or $v_3^{U,i}$), through three vertices in each $V^{S,F^i_j}$, and finally return to $V^{U,i}$, specifically to the vertex $v_4^{U,i}$. Two such paths are illustrated in Fig. 4, for $i = 2$. In the example shown in Fig. 4, we assume $S^1 = (1, 2, 3)$, $S^2 = (1, 2, 4)$, $S^3 = (1, 3, 4)$ and $S^4 = (2, 3, 4)$. The completed graph would have six more paths, two for each of $i = 1$, $i = 3$ and $i = 4$, creating a connected graph. The latter have been omitted for the sake of visual clarity.

This completes the construction of $\Gamma = \langle V, E \rangle$ from $\langle \mathscr{U}, \mathscr{S} \rangle$. We shall now calculate the cardinalities of $V$ and $E$.

**Dimensionality of the constructed graph**

The final graph $\Gamma$ will contain four vertices for each $i \in \mathscr{U}$, five vertices for each $S^i \in \mathscr{S}$, six vertices for each entry $s^i_j$, and two additional vertices $v_1^C$ and $v_2^{\mathscr{C}}$. Therefore the total number of vertices in the graph is $4u + 5s + 6c + 2$.

Counting the number of edges takes a bit more work. There are four edges (counting undirected edges as two, directed, edges) in each $E^{U,i}$. So $E^U$ contributes $4u$ edges.

Then for each $E^{S,i}$, there are eight edges that will always be present (two from $v_1^{S,i}$, two going to $v_2^{S,i}$, and four between $v_3^{S,i}$, $v_4^{S,i}$ and $v_5^{S,i}$. Then for each element

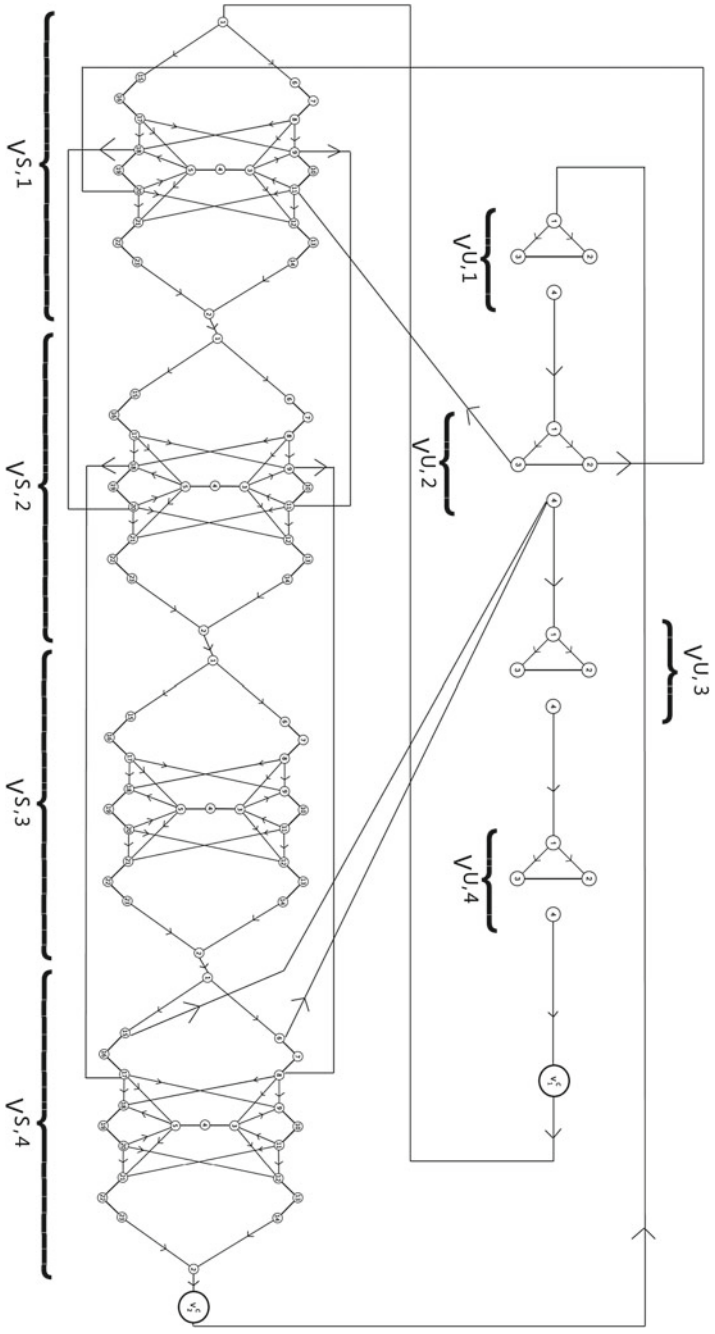



**Fig. 4** The graph $\Gamma$ after being fully constructed. Note that, for the sake of clarity, we only show the new paths corresponding to the element 2. In this example, $S^1 = (1, 2, 3)$, $S^2 = (1, 2, 4)$, $S^3 = (1, 3, 4)$ and $S^4 = (2, 3, 4)$

$s^i_j$ there are 16 additional edges, except for the final element $s^i_{|S^i|}$ for which there are eight edges. So $E^S$ contributes $8s + 16(c - s) + 8s = 16c$ edges.

Finally, for $E^C$, there are $u$ connecting edges for the $V^{U,i}$ components, and $s$ connecting edges for the $V^{S,i}$ components. There are two more connecting edges emerging from $v^C_1$ and $v^C_2$. Finally, for each element $i \in \mathscr{U}$, there are $2|F_i| + 2$ connecting edges forming the two paths, where $|F_i|$ is the number of subsets containing element $i$. So $E^C$ contributes $u + s + 2 + 2(c + u) = 3u + s + 2c + 2$.

Therefore, the total number of directed edges in the graph is $7u + s + 18c + 2$. It should be noted that both the cardinality of the vertex set $V$ and edge set $E$ are linear functions of the cardinalities of the input sets for the original problem. For this reason, we refer to the conversion described above as a *linearly-growing conversion* from SSP to directed HCP.

## 4 The Intended Interpretation of the Converted Graph

Once the conversion is complete, the graph $\Gamma$ contains many components $V^{U,i}$ and $V^{S,i}$ corresponding to elements in $\mathscr{U}$ and subsets $S^i \in \mathscr{S}$, respectively. We now consider in some detail the intended interpretation of any Hamiltonian cycle traversing those components.

For each element $\hat{u} \in \mathscr{U}$, there are four vertices in $V^{U,\hat{u}}$. As will be proved later, any Hamiltonian cycle in $\Gamma$ visits $v^{U,\hat{u}}_2$ and $v^{U,\hat{u}}_3$ in succession (in either order). If $v^{U,\hat{u}}_2$ is visited last, it corresponds to placing the element $\hat{u}$ in the first partition. If vertex $v^{U,\hat{u}}_3$ is visited last, it corresponds to placing the element $\hat{u}$ in the second partition.

For each $S^i \in \mathscr{S}$, and $k := |S^i|$, by construction, there are $5 + 6k$ vertices in $V^{S,i}$. Now, consider a particular element $s^i_j$, that is, the $j$th entry of $S^i$. Naturally, $s^i_j \in \mathscr{U}$. Then, corresponding to this element being chosen in the first partition, there are vertices $v^{S,i}_{3+3j}$, $v^{S,i}_{4+3j}$ and $v^{S,i}_{5+3j}$. Similarly, corresponding to this element being chosen in the second partition, there are vertices $v^{S,i}_{3+3k+3j}$, $v^{S,i}_{4+3k+3j}$ and $v^{S,i}_{5+3k+3j}$.

Now, suppose there is an element $\hat{u}$ that we have chosen to be in $\mathscr{U}_1$. That is, the cycle visits $v^{U,\hat{u}}_2$ after $v^{U,\hat{u}}_3$. Consider all subsets $S^i$ that contain $\hat{u}$, that is $s^i_{j_i} = \hat{u}$ for some $j_i$. Then, immediately after visiting $v^{U,\hat{u}}_2$, the cycle will begin to travel to each such component $V^{S,i}$, and in each of them it will traverse the three vertices corresponding to selecting $s^i j_i$ in $\mathscr{U}_2$ (not $\mathscr{U}_1$). In this way, only those vertices corresponding to this element being chosen in the correct partition will remain to be traversed later. Once all such vertices in all relevant $V^{S,i}$ components have been visited, the cycle returns to the $V^{U,\hat{u}}$ component in order to proceed to make a choice of partition for the next element in $\mathscr{U}$.

The process until this point will be called "stage 1". Once a choice of partition has been made for all elements in $\mathscr{U}$, the cycle travels through $v^C_1$, and on to $v^{S,1}_1$, whereby "stage 2" begins.

The intention is that, upon reaching vertex $v_1^{S,i}$ for each $i$, the cycle will recognise which partition $s_1^i$ is in, because the vertices corresponding to the alternative partition will have already been visited during stage 1. The cycle will then proceed through the three vertices corresponding to the correct choice of partition. Then, the cycle will again recognise which partition $s_2^i$ is in, traverse the three vertices corresponding to that choice, and so on. Once the vertices corresponding to all elements, in their partitions, of $S^i$ are traversed, the cycle will reach vertex $v_2^{S,i}$ and proceed to the next component $V^{S,i+1}$. Specifically, it will travel from $v_2^{S,i}$ to $v_1^{S,i+1}$.

During this process, however, the option to visit one of $v_3^{S,i}$ or $v_5^{S,i}$ will present itself each time the the three vertices for an element and partition are traversed. Specifically, after traversing through vertices corresponding to an element being in $\mathscr{U}_1$, it will be possible to visit $v_3^{S,i}$. Alternatively, after traversing through vertices corresponding to an element being in $\mathscr{U}_1$, it will be possible to visit $v_5^{S,i}$. If the cycle chooses to visit one of these two vertices, it will continue through $v_4^{S,i}$ and out to the other of the two vertices. At this point the cycle will continue to traverse the three vertices corresponding to the next element in the set, but it will be forced to choose vertices corresponding to the opposite partition to that of the previous element. For example, suppose the cycle traverses the three vertices corresponding to $s_j^i$ being chosen in the first partition, and then visits $v_3^{S,i}$. It will subsequently visit $v_4^{S,i}$ and $v_5^{S,i}$, and then proceed to visit the three vertices corresponding to $s_{j+1}^i$ being chosen in the second partition.

The above paragraph describes the essence of why the conversion works. Vertices $v_3^{S,i}$, $v_4^{S,i}$ and $v_5^{S,i}$ must be traversed at some point during stage 2, but after they are traversed, the element that is subsequently considered must be in a different partition to the previous element. This will be possible in all $V^{S,i}$ if and only if the partition choices made in stage 1 split all of the subsets. Therefore, as will be shown rigorously below), Hamiltonian cycles will only exist if it is possible to solve the original instance of SSP.

The formal proof of the validity of the above description is presented next.

## 5 Proof of Conversion

In this section we will prove that, for an instance of SSP, the conversion given in Sect. 3 produces a graph that possesses the desired properties; that is, the graph is Hamiltonian if and only if the original instance of SSP has solutions, that all Hamiltonian cycles correspond to solutions of the instance of SSP, and that all such solutions can be recovered from the Hamiltonian cycles in polynomial time. Throughout the proof we will take advantage of the structure produced the graph construction. We now outline the primary such advantages, before continuing with the main result.

While attempting to construct a Hamiltonian cycle in the converted graph, we will regularly take advantage of *forced* choices. These fall into three categories. If

a vertex is arrived at, and only one outgoing edge $e$ exists that leads to an as-of-yet unvisited vertex, the cycle will be forced to proceed along the edge $e$. In such a case, we say there is only *one remaining outgoing edge*. Alternatively, if a vertex is arrived at, and there is an outgoing edge $e$ that leads to another vertex, for which no other incoming edges from as-of-yet unvisited vertices exist, the cycle will be forced to proceed along the edge $e$. In such a case, we say there is only *one remaining incoming edge*. Finally, if a vertex is arrived at, and there is an outgoing edge $e$ that leads to an as-of-yet unvisited vertex $v$ which has degree 2, the cycle will be forced to proceed along edge $e$. By degree 2, we mean that $v$ has exactly two outgoing edges that lead to vertices $v_2$ and $v_3$, and exactly two incoming edges that also come from vertices $v_2$ and $v_3$. Note that this is a slightly non-standard, but convenient, use of the word "degree" in a directed graph.

Suppose that, during in the process of constructing a Hamiltonian cycle, we arrive at a vertex $v_1^{U,i}$. There are two possible choices of which vertex to visit next: $v_2^{U,i}$ and $v_3^{U,i}$. Whichever choice is made, we are forced to visit the other immediately afterwards, as there will only be one remaining incoming edge. At this point, regardless of which vertex we have arrived at, there will only be one remaining outgoing edge, which will lead to a vertex in one of the $V^{S,j}$ components, specifically a vertex $v_{5+3k}^{S,j}$ for some positive integers $j$ and $k$, where $5 + 3k \leq |V^{S,j}|$. We will refer to this situation as a *type 1 forced path*.

Also, suppose that, during the process of constructing a Hamiltonian cycle, we travel from a vertex in any component other than $V^{S,j}$, to a vertex $v_{5+3k}^{S,j}$ for some positive integers $j$ and $k$. Note that this vertex is adjacent to a degree 2 vertex $v_{4+3k}^{S,j}$. Since the vertex $v_{4+3k}^{S,j}$ was not visited immediately before $v_{5+3k}^{S,j}$, we are forced to visit it immediately, and then proceed along the one remaining outgoing edge to the vertex $v_{3+3k}^{S,j}$. At this point, there will also be only one remaining outgoing edge, which will either lead to a vertex of the form $v_{5+3m}^{S,l}$ for some positive integers $l$ and $m$, where $l > j$ and $5 + 3m \leq |V^{S,l}|$, or it will lead to the vertex $v_4^{U,i}$. Note that in the former case, we arrive in the same type of situation that we started with at the beginning of this paragraph. In the latter case, however, we arrive at $v_4^{U,i}$ and are then forced to proceed either to $v_1^{U,i+1}$ (if $i < u$), or to $v_1^C$ (if $i = u$). We will refer to this situation as a *type 2 forced path*.

We now pose the main result of this chapter.

**Proposition 1** *Consider an instance of SSP with universe set $\mathscr{U}$ and a family of subsets $\mathscr{S}$, and the graph $\Gamma = \langle V, E \rangle$ constructed as in Sect. 3. Then the following three properties hold:*

*(i) If no partition of $\mathscr{U}$ exists that splits all $S^i \in \mathscr{S}$, then $\Gamma$ is non-Hamiltonian.*
*(ii) If a partition of $\mathscr{U}$ exists that does split all $S^i \in \mathscr{S}$, a corresponding Hamiltonian cycle exists in $\Gamma$.*
*(iii) From any Hamiltonian cycle in $\Gamma$ we can, in polynomial time, identify a corresponding partition of $\mathscr{U}$ that constitutes a solution of the original instance of SSP.*

*Proof.* **Stage 1:**

Suppose now that we attempt to construct a Hamiltonian cycle in $\Gamma$. Since we may begin at any vertex, we choose to begin at the vertex $v_1^{U,1}$. As described above, we undergo a type 1 forced path, and eventually depart from either $v_2^{U,1}$ or $v_3^{U,1}$ and arrive at the vertex $v_{5+3j}^{S,i}$ for some $i$ and $j$. Then, since we have arrived from a component other than $V^{S,i}$, we will undergo a type 2 forced path. Then, we may (or may not) arrive at another vertex for which a type 2 forced path is applicable. Inductively, the process continues until we do not arrive at such a vertex.[1] Throughout the process, we visit all of the vertices that correspond to placing element one into a particular member of the partition. The construction is such that visiting $v_2^{U,1}$ after $v_3^{U,1}$ forces us to visit the vertices corresponding to the element one being in $\mathscr{U}_2$. Similarly, visiting $v_3^{U,1}$ after $v_3^{U,1}$ forces us to visit the vertices corresponding to the element one being in $\mathscr{U}_1$. Once all of these vertices are visited, we travel to the vertex $v_4^{U,1}$ and proceed to the vertex $v_1^{U,2}$. Note that at this point, all vertices in the $V^{U,1}$ component have been visited.

The above process is then repeated for all components $V^{U,i}$. The only choice that is made in each component is whether to traverse through $v_1^{U,i} \to v_2^{U,i} \to v_3^{U,i}$, or to traverse through $v_1^{U,i} \to v_3^{U,i} \to v_2^{U,i}$. After this choice has been made, the path that is taken—through all vertices corresponding to the element $i$ in the opposite member of the partition, in all components corresponding to the subsets in which element $i$ appears—is forced until the next component $v^{U,i+1}$ is reached. Eventually, after the entirety of $V^U$ has been traversed, vertex $v_1^C$ is reached, and we are forced to proceed to the vertex $v_1^{S,1}$.

**Stage 2:**

There are two outgoing edges from $v_1^{S,1}$, leading to vertices $v_6^{S,1}$ and $v_{6+3k}^{S,1}$ respectively, where $k := |S^1|$. However, exactly one of these must have been visited already in stage 1. If element 1 was placed in $\mathscr{U}_1$, then vertex $v_{6+3k}^{S,1}$ will have already been visited, or similarly, if element 1 was placed in $\mathscr{U}_2$, then vertex $v_6^{S,1}$ will have already been visited. So the choice at this stage is forced. Then, both vertices $v_6^{S,1}$ and $v_{6+3k}^{S,1}$ are adjacent to degree 2 vertices, so the next two steps are forced as well. This means we visit all three vertices corresponding to the element $s_1^1$ being placed in the member of the partition that was chosen during stage 1.

At this point, there are two choices. We may either continue onto the vertices corresponding to the element $s_2^1$, or we may visit one of $v_3^{S,1}$ or $v_5^{S,1}$ (depending on whether $s_1^1$ was placed in $\mathscr{U}_1$ or $\mathscr{U}_2$, respectively). If we make the former choice, we repeat the above process for the element $s_2^1$, and end up again having to choose

[1] In fact, due to the nature of the construction, if the element appears in $q$ subsets, then exactly $q$ type 2 forced paths must occur here.

whether to visit the vertices corresponding to the element $s_3^1$, or visit one of $v_3^{S,1}$ or $v_5^{S,1}$. Suppose that we make the former choice for the first $j-1$ elements of $S^1$, and then choose the latter for the $j$th element. Without loss of generality, suppose the element $s_j^1$ was chosen, during stage 1, to be in $\mathscr{U}_1$. So, after traversing the vertices corresponding to $s_j^1$, we then choose to visit $v_3^{S,1}$. Then there is an adjacent degree 2 vertex, $v_4^{S,1}$, so we are forced to travel there next, and then to $v_5^{S,1}$.

At this point, our choice is forced as the first of the vertices corresponding to choosing $s_{j+1}^1$ in $\mathscr{U}_2$ has at most one remaining incoming edge, from $v_5^{S,1}$. Therefore this choice must be made, if possible. If it is not possible (because element $s_2^1$ was chosen in $\mathscr{U}_1$ during stage 1), then the Hamiltonian cycle cannot be completed at this point. In such a case, choosing to visit vertex $v_3^{S,1}$ after traversing the vertices corresponding to the first $j$ elements was an incorrect choice. It is then clear that this choice may only be made if $s_{j+1}^1$ was chosen in $\mathscr{U}_2$ during stage 1.

An equivalent argument to that in the above paragraph can be made if element $s_j^1$ was chosen to be in $\mathscr{U}_2$. Then we can see that we may only choose to visit $v_3^{S,1}$ (or $v_5^{S,1}$) when elements $s_j^1$ and $s_{j+1}^1$ are in opposite members of the partition.

After we have visited vertices $v_3^{S,1}$, $v_4^{S,1}$ and $v_5^{S,1}$ once, they cannot be visited again, and so the remaining path through $V^{S,1}$ is forced until we finally reach $v_2^{S,1}$ and are forced to continue to $v_1^{S,2}$. The same process then continues for all components $V^{S,i}$. Finally, the vertex $v_2^C$ is visited, and we travel back to the vertex $v_1^{U,1}$ to complete the Hamiltonian cycle.

The only way in which the above process might fail to produce a Hamiltonian cycle is if there is a component $V^{S,i}$ for which we are unable to find a $j$ such that $s_j^i$ and $s_{j+1}^i$ are in opposite members of the partition. In this case, following the above argument, vertices $v_3^{S,i}$, $v_4^{S,i}$ and $v_5^{S,i}$ cannot possibly be visited in a Hamiltonian cycle. This situation arises only when all entries in $S^i$ are contained in a single member of the partition. In such a situation, the partitioning choices in stage 1 do not form a solution to the original set splitting problem. So making partition choices in stage 1 that do not solve the instance of SSP will always make it impossible to complete stage 2. Clearly then, if there is no solution to the set splitting problem, it will be impossible to find any Hamiltonian cycle. Therefore part (i) holds.

If there are solutions to the original set splitting problem, then for any such solution we can make the corresponding choices in stage 1. Then, in stage 2 there will be an opportunity in each $V^{S,i}$ component to visit the vertices $v_3^{S,i}$, $v_4^{S,i}$ and $v_5^{S,i}$, and continue afterwards. Therefore, for any such solution, a Hamiltonian cycle exists, and hence part (ii) holds.

Finally, identifying the solution to the original instance of SSP is as easy as looking at $v_2^{U,i}$ and $v_3^{U,i}$ for each $i \in \mathscr{U}$ to see which vertex was visited last on the Hamiltonian cycle. If vertex $v_2^{U,i}$ was visited last, element $i$ should be placed in $\mathscr{U}_1$. If vertex $v_3^{U,i}$ was visited last, element $i$ should be placed in $\mathscr{U}_2$. This process can obviously be performed in polynomial time, and therefore part (iii) holds.

Proposition 1 ensures that the conversion given in Sect. 3 produces a graph which is Hamiltonian if and only if the original set splitting problem has solutions, and each Hamiltonian cycle specifies one such solution. Since the number of vertices and edges in the resultant graph are linear functions of the original problem variables, this process constitutes a linearly-growing NP-complete conversion.

## 6 Examples

We now conclude with two small examples of SSP instances and their corresponding directed HCP instances. In the first example we provide a Hamiltonian cycle in the graph, and hence deduce a solution to the original SSP instance. In the second example, the corresponding graph is non-Hamiltonian, and we deduce that the original SSP instance has no solutions and hence the decision of the instance is NO.

*Example 1* Consider the SSP instance with $\mathscr{U} = \{1, 2, 3, 4\}$ and a family of subsets $\mathscr{S} = \{S^1, S^2\}$ where $S^1 = \{1, 2, 3\}$ and $S^2 = \{2, 4\}$.

Following the construction given in Sect. 3, we obtain an instance $\Gamma$ of directed HCP, which is displayed in Fig. 5. A Hamiltonian cycle in $\Gamma$ is also displayed in Fig. 5, with the edges in the Hamiltonian cycle designated by dashed or dotted edges. The dashed edges correspond to edges which are chosen in stage 1, while the dotted edges correspond to edges which are chosen in stage 2.

Consider the Hamiltonian cycle indicated in Fig. 5, and traverse it from the starting vertex $v_1^{U,1}$ as marked on the figure. Then, to determine the solution of the instance of SSP, we simply need to look at which of vertices $v_2^{U,i}$ and $v_3^{U,i}$ were visited last, for each $i$, on the Hamiltonian cycle. In Fig. 5 vertex $v_2^{U,i}$ is drawn above vertex $v_3^{U,i}$ in each case. It can be easily checked that vertices $v_2^{U,1}$, $v_3^{U,2}$, $v_3^{U,3}$ and $v_2^{U,4}$ are the last visited vertices in each case. This corresponds to a solution of the instance of SSP where elements 1 and 4 are placed in $\mathscr{U}_1$, and elements 2 and 3 are placed in $\mathscr{U}_2$. Clearly, this choice provides a splitting of $S^1$ and $S^2$.

We now check that an incorrect choice of partitioning will make it impossible to complete a Hamiltonian cycle. Suppose we assign elements 1 and 3 to $\mathscr{U}_1$, and elements 2 and 4 to $\mathscr{U}_2$. Note that this is not a solution of the original instance of SSP, as the subset $S^2$ is not split. In Fig. 6 we use dashed edges to denote the edges visited in stage 1 (corresponding to our incorrect choice of partitioning), and dotted edges to denote the edges subsequently visited in stage 2. However, the three middle vertices in the $V^{S,2}$ component are unable to be visited, and hence this choice of partitioning can not lead to a Hamiltonian cycle.

Finally, we consider an SSP instance with no solutions, and the corresponding instance of directed HCP.

*Example 2* Consider the SSP instance with $\mathscr{U} = \{1, 2, 3\}$ and a family of subsets $\mathscr{S} = \{S^1, S^2, S^3\}$ where $S^1 = \{1, 2\}$, $S^2 = \{1, 3\}$ and $S^3 = \{2, 3\}$. It is clear that

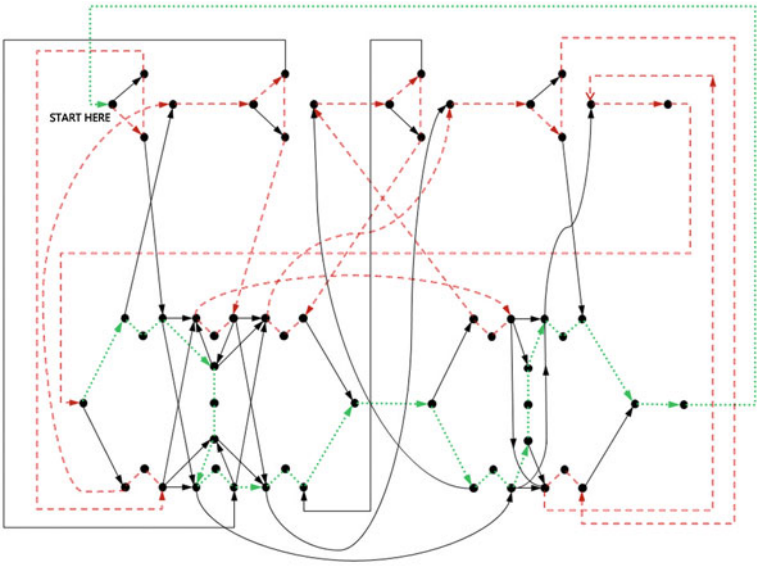



**Fig. 5** The converted graph $\Gamma$ arising from the SSP instance in Example 1. The *dashed* and *dotted edges* correspond to the Hamiltonian cycle, with the *dashed edges* being chosen in stage 1, and the *dotted edges* being chosen in stage 2

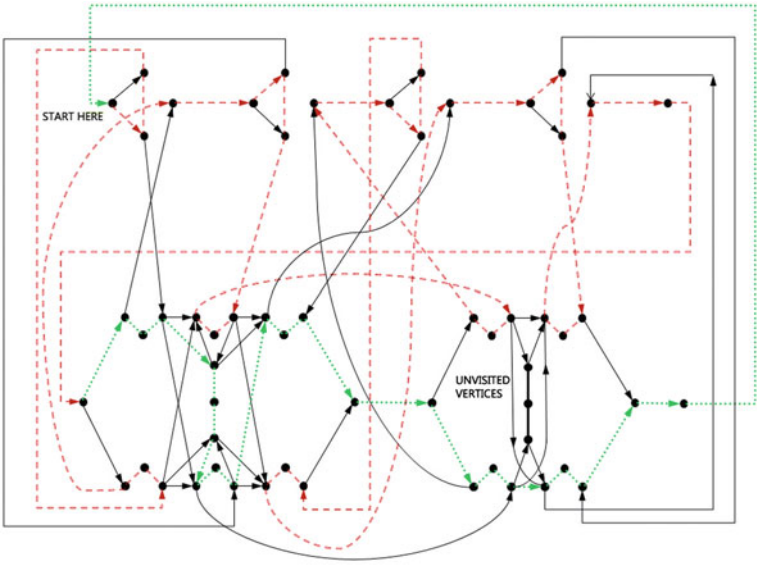



**Fig. 6** The converted graph $\Gamma$ as in Fig. 5, but with a choice of edges in stage 1 that do not correspond to a solution. The three middle vertices in the *bottom-right* component cannot be visited during stage 2

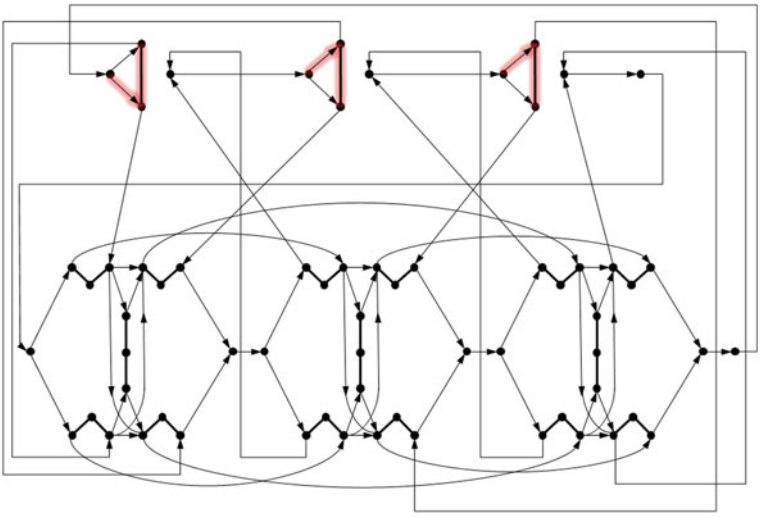


**Fig. 7** The converted graph $\Gamma_2$ arising from the SSP instance in Example 2. Since the SSP instance has decision NO, the graph $\Gamma_2$ is non-Hamiltonian

this instance has no solution since, without loss of generality, if element is placed in the first partition, elements 2 and 3 must be placed in the second partition (to split the first two subsets), and then the third subset is not split.

Following the construction given in Sect. 3, we obtain an instance $\Gamma_2$ of directed HCP, which is displayed in Fig. 7. Although it is not obvious at first glance, the HCP instance is, indeed, non-Hamiltonian. We have confirmed its non-Hamiltonicity using the Concorde TSP Solver [1].

Since $\Gamma$ is non-Hamiltonian, by Proposition 1 (i) there is no partition $(\mathscr{U}_1, \mathscr{U}_2)$ which splits all $S^i \in \mathscr{S}$. Suppose we naively tried to create such a splitting by assigning, for example, element 1 to the first partition, and elements 2 and 3 to the second partition. This would correspond to attempting to find a Hamiltonian cycle in $\Gamma_2$ containing the six highlighted edges in Fig. 7. Then it is straightforward to verify that it is impossible to complete a Hamiltonian cycle. This is a similar feature to that exhibited in Fig. 6, with the important exception that in the latter there were alternative, correct, choices of partitioning $\mathscr{U}$ which do permit Hamiltonian cycles, such as the one illustrated in Fig. 5.

**Acknowledgments** The authors gratefully acknowledge useful conversations with Dr. Richard Taylor that helped in the development of this chapter.

## References


1. Applegate D, Bixby RE, Chvátal V, Cook WJ (2013) Concorde TSP solver. http://www.tsp.gatech.edu/concorde/. Accessed 30 Apr 2013
2. Baniasadi P, Ejov V, Filar JA, Haythorpe M, Rossomakhine S (2012) Deterministic "Snakes and Ladders" Heuristic for the Hamiltonian cycle problem. Math Program Comput (submitted 2012) doi:10.1007/s12532-013-0059-2
3. Borkar VS, Ejov V, Filar JA, Nguyen GT (2012) Hamiltonian cycle problem and Markov chains. Springer, New York
4. Chang W-L, Guo M, Ho M (2004) Towards solutions of the set-splitting problem on gel-based DNA computing. Future Gener Comput Syst 20(5):875–885
5. Chen J, Lu S (2007) Improved algorithms for weighted and unweighted set splitting problems. In: COCOON, vol 4598 of, Lecture Notes in Computer Science, pp 537–547
6. Cook S (1971) The complexity of theorem proving procedures. In: Proceedings of the third annual ACM symposium on theory of computing, pp. 151–158
7. Dehne FKHA, Fellows MR, Rosamond FA (2003) An FPT algorithm for set splitting. In: WG, Lecture notes in computer science, vol 2880, pp 180–191
8. Eppstein D (2007) The traveling salesman problem for cubic graphs. J Graph Algorithms Appl 11(1):61–81
9. Erdős P, Hajnal A (1961) On a property of families of sets. Acta Math Hung 12(1–2):87–123
10. Garey MR, Johnson DS (1979) Computers and intractiability: a guide to the theory of NP-completeness. WH Freeman and Company, New York
11. Karp R (1972) Reducibility among combinatorial problems. In: Miller Raymond E, Thatcher James W (eds) Complexity of computer computations. Plenum, New York, pp 85–103
12. Kugele S (2006) Efficient solving of combinatorial problems using SAT-solvers. Ph.D Thesis, Technische Universität München
13. Lokshtanov D, Saurabh S (2009) Even faster algorithm for set splitting!. In: Parameterized and exact computation: 4th international workshop, IWPEC 2009, Copenhagen, pp. 288–299, Sept 2009
14. Miller EW (1937) On a property of families of sets. Comptes Rendus Vars 30:31–38
15. Radhakrishnan J, Srinivasan A (2000) Improved bounds and algorithms for hypergraph 2-coloring. Random Struct Algorithms 16(1):4–32

# Optimum Confidence Interval Analysis in Two-Factor Mixed Model with a Concomitant Variable for Gauge Study

**Dong Joon Park and Min Yoon**

**Abstract** Measurements often include a variety sources of variability due to measurement systems. The measurement systems study is performed to determine if the measurement procedure is appropriate for monitoring a manufacturing process. The measurement systems study here focuses on determining the amount of variability in a two-factor mixed model with a concomitant variable and no interaction in situations where two variables are correlated. The Analysis of variance is performed for the model and variabilities in the model are represented in a linear combination of variance components. Optimum confidence intervals are constructed using Modified Large Sample approach and Generalized Inference approach in order to determine the variability such as repeatability, reproducibility, parts, gauge, and the ratio of variability of parts to variability of gauge. A numerical example is provided to compute the optimum confidence intervals to investigate the variability in the model.

## 1 Introduction

We explain what gauge study is and how the confidence intervals for variabilities due to measurement system can be employed in gauge study. Literature that relates confidence intervals to gauge study is reviewed in this chapter.

### *1.1 Gauge Study*

To improve a manufacturing process, it is necessary to gain knowledge about the process. This generally requires practitioners to collect and monitor measurements

D. J. Park · M. Yoon (✉)
Department of Statistics, Pukyong National University, Busan, South Korea
e-mail: myoon@pknu.ac.kr

D. J. Park
e-mail: djpark@pknu.ac.kr

H. Xu and X. Wang (eds.), *Optimization and Control Methods in Industrial Engineering and Construction*, Intelligent Systems, Control and Automation: Science and Engineering 72,
DOI: 10.1007/978-94-017-8044-5_4, 

from the process. The action to manage parts and final product leads to proper analysis of the measurements collected in the process. The measurements are regarded to include different sources of variability due to the measurement systems. The measurements systems capability study is therefore performed to determine whether a measurement procedure is adequate for monitoring a manufacturing process. The gauge study in this chapter focuses on determining the amount of variability in the collected measurements that are due to the measurement systems.

A gauge is employed to collect replicated measurements on units by several different operators, setups, or time periods in many measurement studies. Two main elements of variability in gauge study in the manufacturing process are repeatability and reproducibility. Burdick et al. [5] define that repeatability represents the gauge variability when it is used to measure the same unit (with the same operator or setup or in the same time period). Reproducibility refers to the variability arising from different operators, setups, or time periods. Measurement systems capability studies are often referred to as gauge repeatability and reproducibility (R & R) study. The sources of variability in gauge R & R study are expressed as variance components so that confidence intervals for the variance components are employed to determine the adequacy of manufacturing process for gauge R & R study.

## *1.2 Literature Review*

Montgomery and Runger [13, 14] provided sufficient background for gauge capability and designed experiments. They compared classical R & R analysis using $\overline{X}$ and $R$ charts with ANOVA analysis by using numerical examples. Burdick and Larsen [7] proposed alternative confidence intervals on measures of variability in a balanced two-factor crossed random models. They compared a modified large sample (MLS) method and restricted maximum likelihood (REML) method. Borror et al. [3] compared MLS method with REML method for confidence intervals on gauge variability in balanced two-factor crossed random models with interaction. They concluded that REML method resulted in shorter intervals than MLS method but REML method did not maintain the stated level of confidence. Dolezal et al. [9] considered a balanced two-factor crossed random model with fixed operators and compared two confidence intervals; part as a random factor as usual and operator as a fixed factor.

Park and Burdick [17] demonstrate that Wald option as confidence limit used in PROC MIXED procedure of SAS (Statistical Analysis System) produces upper bounds that are enormously larger than usual upper bound. Burdick et al. [4] provided more recent review of the methods for conducting and analyzing measurement systems capabilities, focusing on the analysis of variance approach. Adamec and Burdick [1] proposed two confidence intervals on a discrimination ratio used to determine the adequacy of the measurement system considering a traditional two-factor model. Gong et al. [11] considered methods for constructing confidence intervals in a two-factor gauge R & R when there are unequal replicates. They provided easy and simple EXCEL and SAS code for calculating confidence intervals. The classical R &

R study gives a downward biased estimator of the variance components representing gauge reproducibility. Confidence intervals on variance components are therefore analyzed using ANOVA approach that leads to a direct and convenient method in an experimental design [14]. Burdick and Graybill [6] reviewed the research of confidence intervals in linear models and provided a short history of the research up until 1992. The book rarely dealt with linear models with factors and concomitant variables (independent variables).

Park and Burdick [15, 16] derived confidence intervals in a linear model with one factor and a concomitant variable, which can be also called as a regression model with nested error structure. Park and Burdick [17, 18] extended their works by constructing confidence intervals on the linear functions of variance components in the model with one factor and a concomitant variable that has unbalanced nested error structure. This chapter extends their work by adding one more factor to the model researched by Park and Burdick [17, 18]. We consider two factors with a concomitant variable and no interaction to construct optimum confidence intervals for the linear functions of variance components in the model.

## 2 Model Description

The concept of confidence intervals on variability (variance components) for gauge study is explained. A two-factor model with a concomitant variable and no interaction is described in detail to derive confidence intervals for the linear functions of variance components.

### *2.1 Confidence Interval Approach*

Confidence intervals are more informative than hypothesis tests to make statistical inferences for parameters in applications. We pursue to derive optimum confidence intervals in a two-factor model with a concomitant variable for gauge study. Optimum confidence intervals here mean the approximate intervals with confidence coefficients closed to or at least the stated confidence level across feasible ranges of the model restrictions. Chen et al. [8] refer to an optimal confidence interval as an interval that maximizes the coverage probability with the expected confidence. An approximate $100(1-\alpha)$ % interval for a parameter, i.e. variance $\sigma^2$ is defined as $[L; U]$ that satisfies following equation

$$P[L \leq \sigma^2 \leq U] \cong 1-\alpha \tag{1}$$

where $1-\alpha$ is a stated confidence level and $\alpha$ ranges from 0 to 1. When the equality of Eq. (1) holds, $[L; U]$ is called an exact $100(1-\alpha)$ % interval for $\sigma^2$. Although we prefer exact intervals to approximate intervals, we have to choose approximate

intervals if exact intervals do not exist. That is, an approximate interval provides actual confidence coefficients closed to or at least the stated confidence level across feasible ranges of the model restrictions. If the probability $P[L \leq \sigma^2 \leq U]$ is greater than a stated confidence level $1 - \alpha$, then an interval $[L; U]$ is called a conservative interval. If the probability is less than the stated confidence level, then an interval $[L; U]$ is called a liberal interval. In general, conservative intervals are preferred only when approximate intervals are available. The constraints to the approximate confidence intervals are that $F$-values in computing intervals should be easily determined and an investigator should know the probability of missing parameter.

Burdick and Graybill [6] showed that the confidence intervals using equal tailed $F$-values than any other $F$-values satisfy these two constraints. The $F$-values are available at the tables of the statistics references. The approximate confidence intervals are thus regarded as optimized intervals in the sense that they have actual confidence coefficients at least the stated confidence level and they use equal tailed $F$-values that are easy to determine with knowledge of prior probability of missing the parameter. The optimum confidence intervals constructed here can be employed to monitor gauge R & R study of a manufacturing process that uses in two factors with a concomitant variable and no interaction.

## 2.2 A Two-Factor Mixed Model with a Concomitant Variable and No Interaction

Two-factor crossed mixed model with a concomitant variable and no interaction for gauge R & R study employs $J$ operators randomly chosen to conduct measurements on $I$ randomly selected parts from a manufacturing process. In this R & R study each operator measures each part $K$ times. The model proper for this study is written as

$$\begin{aligned} Y_{ijk} &= \mu + \beta X_{ijk} + P_i + O_j + E_{ijk} \\ i &= 1, \cdots, I; \;\; j = 1, \cdots, J; \;\; k = 1, \cdots, K \end{aligned} \tag{2}$$

where $Y_{ijk}$ is the $k$th measurement (response) of the $i$th part measured by the $j$th operator, $\mu$ and $\beta$ are unknown constants, $X_{ijk}$ is a concomitant variable, $P_i$ is the $i$th randomly selected part, $O_j$ is the $j$th randomly chosen operator, and $E_{ijk}$ is the $k$th measurement error of the $i$th part measured by the $j$th operator. $P_i$, $O_j$, and $E_{ijk}$ are jointly independent normal random variables with zero means and variances $\sigma_P^2$, $\sigma_O^2$, and $\sigma_E^2$, respectively. That is, $P_i \sim N(0, \sigma_P^2)$, $O_j \sim N(0, \sigma_O^2)$, and $E_{ijk} \sim N(0, \sigma_E^2)$. The numbers in Eq. (2) are generally positive real numbers, since the $X_{ijk}$ and $\beta$ are fixed and $P_i$ and $O_j$ are random factors, model (2) is a mixed model. The model (2) does not have interaction effects $(PO)_{ij}$ between $P_i$ and $O_j$ factors. Interaction effects mean that differences in responses under the levels of one factor are different at two or more levels of the other factor.

Model (2) is written in matrix notation

$$\mathbf{y} = \mathbf{X}\underline{\alpha} + \mathbf{Z}_1\mathbf{p} + \mathbf{Z}_2\mathbf{o} + \mathbf{Z}_3\mathbf{e} \tag{3}$$

where $\mathbf{y}$ is a vector of measurements with $I \times J \times K$ rows and one column which means an $IJK \times 1$ vector of observations, $\mathbf{X}$ is an $IJK$ matrix of covariate values with a column of 1's in the first column and a column of $X_{ijk}$'s in the second column, $\underline{\alpha}$ is a $2 \times 1$ vector of parameters with $\mu$ and $\beta$ as elements, $\mathbf{Z}_1$ is an $IJK \times I$ design matrix with 0's and 1's, $\mathbf{Z}_2$ is an $IJK \times J$ design matrix of with 0's and 1's, $\mathbf{Z}_3$ is an an $IJK \times IJK$ design matrix with 0's and 1's, $\mathbf{p}$ is an $I \times 1$ vector of random part effects, $\mathbf{o}$ is a $J \times 1$ vector of random operator effects, and $\mathbf{e}$ is an $IJK \times 1$ vector of random measurement errors. That is, $\mathbf{Z}_1 = \bigoplus_{i=1}^{I} \mathbf{1}_{JK}$, $\mathbf{Z}_2 = \mathbf{1}_I \otimes \left(\bigoplus_{j=1}^{J} \mathbf{1}_K\right)$, and $\mathbf{Z}_3 = \bigoplus_{i=1}^{I} \bigoplus_{j=1}^{J} \bigoplus_{k=1}^{K} 1$, where $\bigoplus$ is a direct sum operator, $\bigotimes$ is a direct product operator, $\mathbf{1}_{JK}$, $\mathbf{1}_I$, and $\mathbf{1}_K$ are a $JK \times 1$, $I \times 1$, and $K \times 1$ column vector of 1's, respectively.

The variance-covariance matrix of $\mathbf{y}$ is thus

$$\mathbf{V} = V(\mathbf{y}) = \sigma_P^2 \mathbf{Z}_1\mathbf{Z}_1' + \sigma_O^2 \mathbf{Z}_2\mathbf{Z}_2' + \sigma_E^2 \mathbf{Z}_3\mathbf{Z}_3'. \tag{4}$$

In particular, the covariance of $Y_{ijk}$ and $Y_{i'j'k'}$ is obtained as

$$Cov(Y_{ijk}, Y_{i'j'k'}) = \begin{cases} 0 & \text{if } i \neq i' \\ \sigma_P^2 & \text{if } i = i',\ j \neq j' \\ \sigma_P^2 + \sigma_O^2 & \text{if } i = i',\ j = j',\ k \neq k' \\ \sigma_P^2 + \sigma_O^2 + \sigma_E^2 & \text{if } i = i',\ j = j',\ k = k'. \end{cases}$$

## 2.3 ANOVA Tables of the Model

Analysis of variance (ANOVA) method is one of the most commonly used means to derive confidence intervals on variance components in gauge study. The total variation of $Y_{ijk}$ in ANOVA table is partitioned into variations of parts, operators, and measurement errors in terms of sources of variation (SV), sums of squares (SS), degrees of freedom (DF), mean squares (MS), and expected mean squares (EMS). The partitions of the components are shown in Tables 1 and 2. The notation in Table 1 is defined as

$$\begin{aligned} SS_{PY} &= JK\Sigma_i(\overline{Y}_{i..} - \overline{Y}_{...})^2, \\ SS_{PX} &= JK\Sigma_i(\overline{X}_{i..} - \overline{X}_{...})^2, \\ SP_{PXY} &= JK\Sigma_i(\overline{X}_{i..} - \overline{X}_{...})(\overline{Y}_{i..} - \overline{Y}_{...}), \\ SS_{OY} &= IK\Sigma_j(\overline{Y}_{.j.} - \overline{Y}_{...})^2, \\ SS_{OX} &= IK\Sigma_j(\overline{X}_{.j.} - \overline{X}_{...})^2, \\ SP_{OXY} &= IK\Sigma_j(\overline{X}_{.j.} - \overline{X}_{...})(\overline{Y}_{.j.} - \overline{Y}_{...}), \end{aligned}$$

**Table 1** ANOVA for model (2)

| *SV* | Sums of squares and cross products | | | |
|---|---|---|---|---|
| | *Y* | *X* | *XY* | DF |
| Parts | $SS_{PY}$ | $SS_{PX}$ | $SP_{PXY}$ | $I-1$ |
| Operators | $SS_{OY}$ | $SS_{OX}$ | $SP_{OXY}$ | $J-1$ |
| Errors | $SS_{EY}$ | $SS_{EX}$ | $SP_{EXY}$ | $IJK-I-J+1$ |
| Total | $SST_Y$ | $SST_X$ | $SPT_{XY}$ | $IJK-1$ |

**Table 2** Adjusted ANOVA for model (2)

| SV | SS | DF | MS | EMS |
|---|---|---|---|---|
| Parts | $R_1$ | $n_1$ | $S_P^2$ | $\sigma_E^2+JK\sigma_P^2$ |
| Operators | $R_2$ | $n_2$ | $S_O^2$ | $\sigma_E^2+IK\sigma_O^2$ |
| Errors | $R_3$ | $n_3$ | $S_E^2$ | $\sigma_E^2$ |

$$
\begin{aligned}
SS_{EY} &= \Sigma_i\Sigma_j\Sigma_k(Y_{ijk}-\overline{Y}_{i..}-\overline{Y}_{.j.}+\overline{Y}_{...})^2,\\
SS_{EX} &= \Sigma_i\Sigma_j\Sigma_k(X_{ijk}-\overline{X}_{i..}-\overline{X}_{.j.}+\overline{X}_{...})^2,\\
SP_{EXY} &= \Sigma_i\Sigma_j\Sigma_k(X_{ijk}-\overline{X}_{i..}-\overline{X}_{.j.}+\overline{X}_{...})(Y_{ijk}-\overline{Y}_{i..}-\overline{Y}_{.j.}+\overline{Y}_{...}),\\
SST_Y &= \Sigma_i\Sigma_j\Sigma_k(Y_{ijk}-\overline{Y}_{...})^2,\\
SST_X &= \Sigma_i\Sigma_j\Sigma_k(X_{ijk}-\overline{X}_{...})^2, \text{ and}\\
SPT_{XY} &= \Sigma_i\Sigma_j\Sigma_k(X_{ijk}-\overline{X}_{...})(Y_{ijk}-\overline{Y}_{...}).
\end{aligned}
$$

Table 1 is modified in Table 2 that includes necessary elements to construct confidence intervals on the variance components. The mean squares in Table 2 are obtained by dividing the sums of squares by their degrees of freedom and they are defined as $S_P^2 = R_1/n_1$, $S_O^2 = R_2/n_2$, and $S_E^2 = R_3/n_3$ where

$$R_1 = JK\mathbf{y}'\mathbf{W}_1'\left(\mathbf{D}_I-\mathbf{H}_1\right)\mathbf{W}_1\mathbf{y}, \tag{5}$$

$$R_2 = IK\mathbf{y}'\mathbf{W}_2'\left(\mathbf{D}_J-\mathbf{H}_2\right)\mathbf{W}_2\mathbf{y}, \tag{6}$$

$$R_3 = \mathbf{y}'\mathbf{W}_3'\left(\mathbf{D}_{IJK}-\mathbf{H}_3\right)\mathbf{W}_3\mathbf{y}, \tag{7}$$

$$
\begin{aligned}
\mathbf{W}_1 &= \frac{1}{JK}\mathbf{Z}_1',\\
\mathbf{W}_2 &= \frac{1}{IK}\mathbf{Z}_2',\\
\mathbf{W}_3 &= Z_3' = \mathbf{D}_{IJK},\\
\mathbf{H}_1 &= \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1',\\
\mathbf{H}_2 &= \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2',
\end{aligned}
$$

$$\begin{aligned}
\mathbf{H}_3 &= \mathbf{X}_3(\mathbf{X}_3'\mathbf{X}_3)^{-}\mathbf{X}_3', \\
\mathbf{X}_1 &= \mathbf{W}_1\mathbf{X}, \\
\mathbf{X}_2 &= \mathbf{W}_2\mathbf{X}, \\
\mathbf{X}_3 &= [\mathbf{X}\ \mathbf{Z}_1\ \mathbf{Z}_2], \\
n_1 &= I-2, \\
n_2 &= J-2, \text{ and} \\
n_3 &= IJK-I-J
\end{aligned}$$

## 3 Statistical Distributions of Sums of Squares

Statistical distributions of the sums of squares are derived by using the assumptions in model (2). Theorems are provided to show that the sums of squares are chi-squared distributed and the chi-squared distributed random variables are independent.

### *3.1 Theorems of Chi-Squared Distributions*

**Theorem 1.** *Under the distributional assumptions in* (2)*,* $R_1/(\sigma_E^2 + JK\sigma_P^2)$ *is a chi-squared random variable with* $n_1$ *degrees of freedom.*

*Proof.* Multiply both sides of the Eq. (3) on the left by $(\mathbf{D}_I - \mathbf{H}_1)\mathbf{W}_1$. Note that $(\mathbf{D}_I - \mathbf{H}_1)\mathbf{W}_1\mathbf{X} = \mathbf{0}$, $\mathbf{W}_1\mathbf{Z}_1 = \mathbf{D}_I$, and $\mathbf{W}_1\mathbf{Z}_3 = \frac{1}{JK}\mathbf{Z}_1'$. It follows that

$$\mathbf{y}_1 = (\mathbf{D}_I - \mathbf{H}_1)\mathbf{p} + (\mathbf{D}_I - \mathbf{H}_1)\mathbf{W}_1\mathbf{Z}_2\mathbf{o} + \frac{1}{JK}(\mathbf{D}_I - \mathbf{H}_1)\mathbf{Z}_1'\mathbf{e}$$

where $\mathbf{y}_1 = (\mathbf{D}_I - \mathbf{H}_1)\mathbf{W}_1\mathbf{y}$. By using that $\mathbf{W}_1\mathbf{Z}_2 = \frac{1}{J}\mathbf{1}_I\mathbf{1}_J'$, $(\mathbf{D}_I - \mathbf{H}_1)\mathbf{W}_1\mathbf{Z}_2\mathbf{Z}_2'\mathbf{W}_1'$ $(\mathbf{D}_I - \mathbf{H}_1) = \mathbf{0}$, $(\mathbf{D}_I - \mathbf{H}_1)\mathbf{1}_I = \mathbf{0}$, and $\mathbf{Z}_1'\mathbf{Z}_1 = JK\mathbf{D}_I$, the variance of $\mathbf{y}_1$ is

$$V(\mathbf{y}_1) = \frac{1}{JK}\left(\sigma_E^2 + JK\sigma_P^2\right)(\mathbf{D}_I - \mathbf{H}_1).$$

The distribution of $R_1$ is determined by writing

$$\frac{R_1}{\sigma_E^2 + JK\sigma_P^2} = \mathbf{y}_1'\left[\frac{JK(\mathbf{D}_I - \mathbf{H}_1)}{\sigma_E^2 + JK\sigma_P^2}\right]\mathbf{y}_1$$

and noting

$$\left[\frac{JK(\mathbf{D}_I - \mathbf{H}_1)}{\sigma_E^2 + JK\sigma_P^2}\right] V(\mathbf{y}_1) = (\mathbf{D}_I - \mathbf{H}_1),$$

$$\frac{1}{2}E(\mathbf{y}_1)' \left[\frac{JK(\mathbf{D}_I - \mathbf{H}_1)}{\sigma_E^2 + JK\sigma_P^2}\right] E(\mathbf{y}_1) = 0, \text{ and}$$

$$rank(\mathbf{D}_I - \mathbf{H}_1) = I - 2.$$

By Theorem 7.3 in Searle [19, p. 232], $R_1/(\sigma_E^2 + JK\sigma_P^2)$ is a chi-squared random variable with $n_1$ degrees of freedom. □

**Theorem 2.** *Under the distributional assumptions in (2), $R_2/(\sigma_E^2 + IK\sigma_O^2)$ is a chi-squared random variable with $n_2$ degrees of freedom.*

*Proof.* The proof is given by showing that

$$\frac{R_2}{\sigma_E^2 + IK\sigma_O^2} = \mathbf{y}_2' \left[\frac{IK(\mathbf{D}_J - \mathbf{H}_2)}{\sigma_E^2 + IK\sigma_O^2}\right] \mathbf{y}_2,$$

$$\left[\frac{IK(\mathbf{D}_J - \mathbf{H}_2)}{\sigma_E^2 + IK\sigma_O^2}\right] V(\mathbf{y}_2) = \mathbf{D}_J - \mathbf{H}_2,$$

$$\frac{1}{2}E(\mathbf{y}_2)' \left[\frac{IK(\mathbf{D}_J - \mathbf{H}_2)}{\sigma_E^2 + IK\sigma_O^2}\right] E(\mathbf{y}_2) = 0, \text{ and}$$

$$rank(\mathbf{D}_J - \mathbf{H}_2) = J - 2.$$

where $\mathbf{y}_2 = (\mathbf{D}_J - \mathbf{H}_2)\mathbf{W}_2\mathbf{y}$. □

**Theorem 3.** *Under the distributional assumptions in (2), $R_3/\sigma_E^2$ is a chi-squared random variable with $n_3$ degrees of freedom.*

*Proof.* This theorem is proved by showing that

$$\frac{R_3}{\sigma_E^2} = \mathbf{y}_3' \left[\frac{(\mathbf{D}_{IJK} - \mathbf{H}_3)}{\sigma_E^2}\right] \mathbf{y}_3,$$

$$\left[\frac{(\mathbf{D}_{IJK} - \mathbf{H}_3)}{\sigma_E^2}\right] V(\mathbf{y}_3) = \mathbf{D}_{IJK} - \mathbf{H}_3,$$

$$\frac{1}{2}E(\mathbf{y}_3)' \left[\frac{(\mathbf{D}_{IJK} - \mathbf{H}_3)}{\sigma_E^2}\right] E(\mathbf{y}_3) = 0, \text{ and}$$

$$rank(\mathbf{D}_{IJK} - \mathbf{H}_3) = IJK - I - J.$$

where $\mathbf{y}_3 = (\mathbf{D}_{IJK} - \mathbf{H}_3)\mathbf{W}_3\mathbf{y}$. □

### 3.2 Theorem of Independency of Chi-Squared Distributions

**Theorem 4.** *Under the distributional assumptions in* (2), $R_1/(\sigma_E^2 + JK\sigma_P^2)$ *and* $R_3/\sigma_E^2$ *are independent and* $R_2/(\sigma_E^2 + IK\sigma_O^2)$ *and* $R_3/\sigma_E^2$ *are independent.*

*Proof.* Note that Eqs. (5) and (7) are quadratic forms and variance of $\mathbf{y}$ is defined in (4). It is therefore shown that

$$\mathbf{W}_1^{'}(\mathbf{D}_I - \mathbf{H}_1)\mathbf{W}_1 V(\mathbf{y})\mathbf{W}_3^{'}(\mathbf{D}_{IJK} - \mathbf{H}_3)\mathbf{W}_3 = \mathbf{0}$$

by using $\mathbf{Z}_1^{'}\mathbf{W}_3^{'}(\mathbf{D}_{IJK} - \mathbf{H}_3) = \mathbf{0}$, $\mathbf{Z}_2^{'}\mathbf{W}_3^{'}(\mathbf{D}_{IJK} - \mathbf{H}_3) = \mathbf{0}$, and $\mathbf{W}_1\mathbf{Z}_3\mathbf{Z}_3^{'}\mathbf{W}_3^{'}(\mathbf{D}_{IJK} - \mathbf{H}_3)\mathbf{W}_3 = \mathbf{0}$, Therefore, by Theorem 7.4 in Searle [19, p. 232], $R_1/(\sigma_E^2 + JK\sigma_P^2)$ and $R_3/\sigma_E^2$ are independent.

Similarly, by using Eqs. (4), (6), and (7), it can be shown that

$$\mathbf{W}_2^{'}(\mathbf{D}_J - \mathbf{H}_2)\mathbf{W}_2 V(\mathbf{y})\mathbf{W}_3^{'}(\mathbf{D}_{IJK} - \mathbf{H}_3)\mathbf{W}_3 = \mathbf{0}.$$

By Theorem 7.4 in Searle [19, p. 232], $R_2/(\sigma_E^2 + IK\sigma_O^2)$ and $R_3/\sigma_E^2$ are independent. □

Let the expected mean squares of parts, operators, and measurement errors be respectively $\theta_P$, $\theta_O$, and, $\theta_E$. Using the results of Theorems 1 to 4 and the fact that expectation of a chi-squared random variable is its degrees of freedom, the expected mean squares are written as follows:

$$E(S_P^2) = \sigma_E^2 + JK\sigma_P^2 = \theta_P, \tag{8}$$

$$E(S_O^2) = \sigma_E^2 + IK\sigma_O^2 = \theta_O, \text{ and} \tag{9}$$

$$E(S_E^2) = \sigma_E^2 = \theta_E. \tag{10}$$

## 4 Confidence Intervals on Variability

Theorems 1–4 with respect to chi-squared distributions of the sums of squares and their independency are employed to construct confidence intervals of the linear functions of variance components for gauge study. We use two methods for deriving confidence intervals: modified large sample (MLS) method and generalized inference method.

### 4.1 MLS Intervals on Variance Components

Since repeatability is the variation among repeated measurements on a single part by same operator using same measurement system, the variance component measuring repeatability is the variance of measurement errors

$$\sigma^2_{repeatability} = {\sigma_E}^2. \tag{11}$$

An exact confidence interval for the variance of repeatability is obtained by Theorem 3. Since $R_3/\sigma_E^2 \sim \chi^2_{n_3}$, an exact $100(1-\alpha)$ % two-sided confidence interval on $\sigma^2_{repeatability}$ is

$$\left[\frac{S_E^2}{F_{(\alpha/2:n_3,\infty)}}; \frac{S_E^2}{F_{(1-\alpha/2:n_3,\infty)}}\right] \tag{12}$$

where $F_{(\delta:\nu_1,\nu_2)}$ is the $F$-value for $\nu_1$ and $\nu_2$ degrees of freedom with $\delta$ area to the right of $F$ distribution as equal tailed $F$-values.

The variance of part effect $\sigma^2_{part}$ means the variance component of parts randomly chosen from a manufacturing process, i.e. $\sigma^2_{part} = \sigma^2_P$ and it is computed by (8) and (10)

$$\sigma^2_{part} = \frac{\theta_P - \theta_E}{JK}. \tag{13}$$

Ting et al. [20] proposed a general method for constructing confidence intervals on $\gamma = \sum_{q=1}^{P} c_q\theta_q - \sum_{r=1}^{Q} c_r\theta_r$ where $c_q, c_r \geq 0$ and the sign of $\gamma$ is unknown. The intervals are extended to the general case where $Q > 2$ and $P > 2$. Their two-sided confidence intervals were shown to be very close to the stated confidence level by simulation study. A confidence interval on $\sigma^2_{part}$ can be constructed by using the method of Ting et al. [20]. An approximate $100(1-\alpha)$ % two-sided confidence interval on $\sigma^2_{part}$ is

$$\frac{1}{JK}\Big[S_P^2 - S_E^2 - \left(G_1^2S_P^4 + H_3^2S_E^4 + G_{13}S_P^2S_E^2\right)^{\frac{1}{2}}; \\ S_P^2 - S_E^2 + \left(H_1^2S_P^4 + G_3^2S_E^4 + H_{13}S_P^2S_E^2\right)^{\frac{1}{2}}\Big] \tag{14}$$

where $G_i = 1 - 1/F_{(\alpha/2:n_i,\infty)};\ i = 1, 2, 3,\ H_i = 1/F_{(1-\alpha/2:n_i,\infty)} - 1;\ i = 1, 2, 3,$ $G_{13} = \left((F_1-1)^2 - G_1^2F_1^2 - H_3^2\right)/F_1,\ G_{23} = \left((F_3-1)^2 - G_2^2F_3^2 - H_3^2\right)/F_3,$ $H_{13} = \left((1-F_2)^2 - H_1^2F_2^2 - G_3^2\right)/F_2,\ H_{23} = \left((1-F_4)^2 - H_2^2F_4^2 - G_3^2\right)/F_4,$ $F_1 = F_{(\alpha/2:n_1,n_3)}, F_2 = F_{(1-\alpha/2:n_1,n_3)}, F_3 = F_{(\alpha/2:n_2,n_3)},$ and $F_4 = F_{(1-\alpha/2:n_2,n_3)}$.

Since reproducibility results from different operators, the variance component measuring reproducibility is variance of operator effect, i.e. $\sigma^2_{reproducibility} = \sigma^2_O$ and it is represented by (9) and (10)

$$\sigma^2_{reproducibility} = \frac{\theta_O - \theta_E}{IK}. \tag{15}$$

Similarly, an approximate $100(1-\alpha)$ % two-sided confidence interval on $\sigma^2_{reproducibility}$ is derived by using Ting et al. [20]

$$\frac{1}{IK}\Big[S_O^2 - S_E^2 - \Big(G_2^2 S_O^4 + H_3^2 S_E^4 + G_{23} S_O^2 S_E^2\Big)^{\frac{1}{2}}; \tag{16}$$
$$S_O^2 - S_E^2 + \Big(H_2^2 S_O^4 + G_3^2 S_E^4 + H_{23} S_O^2 S_E^2\Big)^{\frac{1}{2}}\Big].$$

Since repeatability and reproducibility consist of two major sources of variability in gauge study, the variability of gauge in a manufacturing process is obtained by sum of the two variabilities using (11) and (15), i.e. $\sigma^2_{gauge} = \sigma^2_{reproducibility} + \sigma^2_{repeatability}$. The variance of gauge is written as

$$\sigma^2_{gauge} = \frac{1}{IK}\Big(\theta_O + (IK-1)\theta_E\Big) \tag{17}$$

A confidence interval for the variance of gauge is constructed using the method proposed by Graybill and Wang [12] because the sign of $\theta_O$ and $\theta_E$ are all positive in the right side of (17). An approximate $100(1-\alpha)$ % two-sided confidence interval on $\sigma^2_{gauge}$ is

$$\frac{1}{IK}\Big[S_O^2 + (IK-1)S_E^2 - \Big\{\Big(G_2 S_O^2\Big)^2 + \Big((IK-1)G_3 S_E^2\Big)^2\Big\}^{\frac{1}{2}}; \tag{18}$$
$$S_O^2 + (IK-1)S_E^2 + \Big\{\Big(H_2 S_O^2\Big)^2 + \Big((IK-1)H_3 S_E^2\Big)^2\Big\}^{\frac{1}{2}}\Big].$$

If variation due to the measurement system is small relative to variation of the process, then the measurement system is deemed capable. This means the measurement system can be used to monitor the manufacturing process. Let the ratio of variance of parts to variance of gauge be

$$\delta_P = \frac{\sigma_P^2}{\sigma^2_{gauge}}. \tag{19}$$

A confidence interval for this ratio is derived using the method by Arteaga et al. [2]. An approximate $100(1-\alpha)$ % two-sided confidence interval on $\delta_P$ is

$$\left[\frac{I\ F_8\ S_P^4 - I\ S_P^2\ S_E^2 + I\ \{F_1 - F_8\ F_1^2\}\ S_E^2}{J(IK-1)\ S_P^2\ S_E^2 + J\ F_8\ F_5\ S_P^2\ S_O^2}; \right. \tag{20}$$
$$\left.\frac{I\ F_7\ S_P^4 - I\ S_P^2\ S_E^2 + I\ \{F_2 - F_7\ F_2^2\}\ S_E^2}{J(IK-1)\ S_P^2\ S_E^2 + J\ F_7\ F_6\ S_P^2\ S_O^2}\right]$$

where $F_5 = F_{(\alpha/2:n_1,n_2)}$, $F_6 = F_{(1-\alpha/2:n_1,n_2)}$, $F_7 = F_{(\alpha/2:\infty,n_1)}$, and $F_8 = F_{(1-\alpha/2:\infty,n_1)}$.

## 4.2 Generalized Intervals on Variance Components

Tsui and Weerahandi [21] introduced the concept on generalized inference for testing hypotheses in situations where exact methods do not exist. Their method can be applied to form approximate confidence intervals on variance components. Since an exact confidence interval on $\sigma_{repeatability}$ exists, Eq. (12) is used for determining the variability of measurement errors .

The variance of part effect $\sigma^2_{part}$ is written by (13) as follows:

$$\sigma^2_{part} = \frac{1}{JK}\left[\frac{n_1}{P^*}s_P^2 - \frac{n_3}{E^*}s_E^2\right] \tag{21}$$

where $s_P^2$ and $s_E^2$ are respectively the observed values of $S_P^2$ and $S_E^2$. $P^*$ and $E^*$ are respectively pivotal quantities of $n_1 S_P^2/\theta_P$ and $n_3 S_E^2/\theta_E$. Define $R_1$ as the solution for $\sigma^2_{part}$ in (21). The distribution of $R_1$ is completely determined by $P^*$ and $E^*$ using Monte Carlo methods. An approximate $100(1-\alpha)$ % two-sided confidence interval on $\sigma^2_{part}$ is

$$\left[R_{1_{\alpha/2}};\, R_{1_{1-\alpha/2}}\right] \tag{22}$$

where $R_{1_{\alpha/2}}$ and $R_{1_{1-\alpha/2}}$ are the percentage of $\alpha/2$ and $1-\alpha/2$ of the distribution $R_1$, respectively.

Similarly, we can form confidence intervals on $\sigma^2_{reproducibility}$ by (15) as

$$\sigma^2_{reproducibility} = \frac{1}{IK}\left[\frac{n_2}{O^*}s_O^2 - \frac{n_3}{E^*}s_E^2\right] \tag{23}$$

where $s_o^2$ is the observed value of $S_O^2$. $O^*$ is a pivotal quantity of $n_2 S_O^2/\theta_O^2$. Define $R_2$ as the solution for $\sigma^2_{reproducibility}$. The distribution of $R_2$ is completely determined by $O^*$ and $E^*$ using Monte Carlo methods. An approximate $100(1-\alpha)$ % two-sided confidence interval on $\sigma^*_{reproducibility}$ is

$$\left[R_{2_{\alpha/2}};\, R_{2_{1-\alpha/2}}\right] \tag{24}$$

where $R_{2_{\alpha/2}}$ and $R_{2_{1-\alpha/2}}$ are the percentage of $\alpha/2$ and $1-\alpha/2$ of the distribution $R_2$, respectively.

The variance of gauge is written by (17) as follows:

$$\sigma^2_{gauge} = \frac{1}{IK}\left[\frac{n_2}{O^*}s_O^2 + (IK-1)\,\frac{n_3}{E^*}s_E^2\right]. \tag{25}$$

The distribution of $R_3$ is completely determined by $O^*$ and $E^*$ using Monte Carlo methods. An approximate $100(1-\alpha)$ % two-sided confidence interval on $\sigma^2_{gauge}$ is

$$\left[R_{3_{\alpha/2}};\, R_{3_{1-\alpha/2}}\right] \tag{26}$$

where $R_{3_{\alpha/2}}$ and $R_{3_{1-\alpha/2}}$ are the percentage of $\alpha/2$ and $1-\alpha/2$ of the distribution $R_3$, respectively.

The ratio of variance of parts to variance of gauge $\delta_P$ is written using (21) and (25) as follows:

$$\begin{aligned}\delta_P &= \frac{I\left[\theta_P - \theta_E\right]}{J\left[\theta_O + (IK-1)\,\theta_E^2\right]} \\ &= \frac{I\left[\frac{n_1}{P^*}s_P^2 - \frac{n_3}{E^*}s_E^2\right]}{J\left[\frac{n_2}{O^*}s_O^2 + (IK-1)\,\frac{n_3}{E^*}s_E^2\right]}\end{aligned} \tag{27}$$

The distribution of $R_4$ is completely determined by $P^*$, $O^*$, and $E^*$ using Monte Carlo methods. An approximate $100(1-\alpha)$ % two-sided confidence interval on $\delta_P$ is

$$\left[R_{4_{\alpha/2}};\, R_{4_{1-\alpha/2}}\right] \tag{28}$$

where $R_{4_{\alpha/2}}$ and $R_{4_{1-\alpha/2}}$ are the percentage of $\alpha/2$ and $1-\alpha/2$ of the distribution $R_4$, respectively.

## 5 Numerical Example

Semiconductor manufacturing uses a variety of chemical and physical processes. Most semiconductor manufacturers buy their raw material in the form of round silicon wafers of a specified resistivity. Bare wafers are processed to produce functional devices such as microprocessors, memories, micro-controllers, simple logic chips, etc, in very complex fabrication procedure. However, the entire fabrication procedure can be divided into five broad categories: lithography, etch, diffusion, thin film deposition, and ion implant. In the process of etching, the areas not protected by resist in a wafer are removed by use of a partially ionized reactive gas. During the process the etch rate in the wafer is generally affected by temperature in an acid bath.

The data set published by Drain [10] includes temperature (T) in degrees Celsius and the etch rate (E) in Angstroms per second. Table 3 presents total 24 measurements assuming that four parts ($I = 4$), three operators ($J = 3$), and two measurements ($K = 2$) are selected and that temperature (T) and etch rate (E) in the Table 3 are respectively regarded as $X_{ijk}$ and $Y_{ijk}$ in model (2) for gauge R & R study.

**Table 3** Sample data of temperatures and etch rates in etching process

| Part | Operator 1 | | Operator 2 | | Operator 3 | |
|---|---|---|---|---|---|---|
| | T | E | T | E | T | E |
| 1 | 22 | 38.8 | 22 | 39.2 | 26 | 45.8 |
| | 26 | 45.3 | 28 | 49.4 | 26 | 39.1 |
| 2 | 22 | 38.9 | 24 | 43.1 | 20 | 39.3 |
| | 20 | 38.2 | 30 | 51.6 | 22 | 40.8 |
| 3 | 20 | 39.8 | 24 | 43.2 | 28 | 48.5 |
| | 24 | 43.0 | 24 | 44.8 | 28 | 47.8 |
| 4 | 20 | 38.3 | 26 | 47.6 | 30 | 52.8 |
| | 26 | 44.8 | 24 | 44.1 | 30 | 52.0 |

**Table 4** Adjusted ANOVA

| SV | SS | DF | MS | EMS |
|---|---|---|---|---|
| Parts | $R_1 = 19.530$ | $n_1 = 2$ | $S_P^2 = 9.765$ | $\sigma_E^2 + JK\sigma_P^2$ |
| Operators | $R_2 = 4.140$ | $n_2 = 1$ | $S_O^2 = 4.140$ | $\sigma_E^2 + IK\sigma_O^2$ |
| Errors | $R_3 = 47.687$ | $n_3 = 17$ | $S_E^2 = 2.805$ | $\sigma_E^2$ |

**Table 5** The estimates of variances

| Variance | $\sigma^2_{part}$ | $\sigma^2_{reproducibility}$ | $\sigma^2_{repeatability}$ | $\sigma^2_{gauge}$ | $\delta_P$ |
|---|---|---|---|---|---|
| Estimate | 1.160 | 0.167 | 2.805 | 2.972 | 0.390 |

**Table 6** 95 % confidence intervals for the variances

| Variance | MLS intervals | | | Generalized intervals | | |
|---|---|---|---|---|---|---|
| | LB | UB | LE | LB | UB | LE |
| $\sigma^2_{part}$ | 0.000 | 31.225 | 31.225 | 0.005 | 31.440 | 31.435 |
| $\sigma^2_{reproducibility}$ | 0.000 | 131.231 | 131.231 | 0.000 | 131.274 | 131.580 |
| $\sigma^2_{repeatability}$ | 1.729 | 5.499 | 3.771 | 1.730 | 5.506 | 3.776 |
| $\sigma^2_{gauge}$ | 2.449 | 134.074 | 131.625 | 2.078 | 131.274 | 133.054 |
| $\delta_P$ | 0.001 | 10.422 | 10.421 | 0.000 | 7.316 | 7.316 |

Table 4 is obtained from the definitions explained in Sect. 2.3 using the sample data set of Table 3. The estimates of variances of parts, reproducibilities, repeatability, gauge, and the ratio of variance of parts to variance of gauge in Table 5 are computed by (11), (13), (15), (17), and (19). Table 6 reports the 95 % confidence intervals for variances of parts, reproducibilities, repeatability, gauge, and the ratio of variance of parts to variance of gauge calculated by confidence interval formulas in this chapter. LB, UB, and LE in Table 6 respectively represent lower limit, upper limit, and length of confidence intervals. Negative values of lower limits are set to

zero because variance cannot be negative. The confidence interval for variance of repeatability is an exact interval and other intervals are approximate.

MLS and generalized intervals calculated in Table 6 are very similar especially for $\sigma^2_{part}$, $\sigma^2_{reproducibility}$, $\sigma^2_{repeatability}$, and $\sigma^2_{gauge}$. However, the confidence interval for $\delta_P$ is slightly different. In this case we can choose narrower confidence interval as long as it maintains the stated confidence level.

We proposed an exact confidence interval for the variance of repeatability and approximate confidence intervals for the variances of parts, reproducibility, gauge, and the ratio of parts to the variance of gauge in two-factor mixed model with concomitant variable and no interaction for gauge R & R study. In a manufacturing process that the model can be applied the confidence intervals proposed here can be used for monitoring whether the variabilities are appropriate for the process.

## References


1. Adamec E, Burdick RK (2003) Confidence intervals for a discrimination ratio in a gauge R & R study with three random factors. Qual Eng 15(3):383–389
2. Arteaga C, Jeyaratnam S, Graybill FA (1982) Confidence intervals for proportions of total variance in the two-way cross component of variance model. Commun Stat Theory Methods 11:1643–1658
3. Borror CM, Montgomery DC, Runger GC (1997) Confidence intervals for variance components from gauge capability studies. Qual Reliab Eng Int 13:361–369
4. Burdick RK, Borror CM, Montgomery DC (2003) A review of methods for measurement system capability analysis. J Qual Tech 35(4):342–354
5. Burdick RK, Borror CM, Montgomery DC (2005) Design and analysis of gauge R & R studies. SIAM, Philadelphia
6. Burdick RK, Graybill FA (1992) Confidence intervals on variance components. Marcel Dekker, New York
7. Burdick RK, Larsen GA (1997) Confidence intervals on measures of variability in R & R studies. J Qual Tech 29(3):261–273
8. Chen HC, Li H-L, Wen M-J (2008) Optimal confidence intervals for the largest mean of correlated normal populations and its application to stock fund evaluation. Comput Stat Data Anal 52:4801–4813
9. Dolezal KK, Burdick RK, Birch NJ (1998) Analysis of a two-factor R & R study with fixed operators. J Qual Tech 30(2):163–170
10. Drain D (1996) Statistical methods for industrial process control. International Thomson Publishing, New York
11. Gong L, Burdick RK, Quiroz J (2005) Confidence intervals for unbalanced two-factor gauge R & R studies. Qual Reliab Eng Int 21:727–741
12. Graybill FA, Wang C-M (1980) Confidence intervals on nonnegative linear combinations of variance components. J Am Stat Assoc 75(372):869–873
13. Montgomery DC, Runger GC (1993) Gauge capability and designed experiments. Part I: basic methods. Qual Eng 6(1):115–135
14. Montgomery DC, Runger GC (1993) Gauge capability and designed experiments. Part II: experimental design models and variance component estimation. Qual Eng 6(2):289–305
15. Park DJ, Burdick RK (1993) Confidence intervals on the among group variance component in a simple linear regression model with nested error structure. Commun Stat Theory Methods 22:3435–3452

16. Park DJ, Burdick RK (1994) Confidence intervals on the regression coefficient in a simple linear regression model with nested error structure. Commun Stat Simul Comput 23:43–58
17. Park DJ, Burdick RK (2003) Performance of confidence intervals in regression model with unbalanced one-fold nested error structures. Commun Stat Simul Comput 32:717–732
18. Park DJ, Burdick RK (2004) Confidence intervals on total variance in a regression model with unbalanced one-fold nested error structure. Commun Stat Theory Methods 33:2735–2744
19. Searle SR (1987) Linear models for unbalanced data. Wiley, New York
20. Ting N, Burdick RK, Graybill FA, Jeyaratnam S, Lu T-FC (1990) Confidence intervals on linear combinations of variance components. J Stat Comput Simul 35:135–143
21. Tsui K, Weerahandi S (1989) Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. J Amer Stat Assoc 84(381):602–607

# Optimization of Engineering Survey Monitoring Networks

**Willie Tan**

**Abstract** This chapter considers the various ways in which engineering survey monitoring networks, such that those used for tracking volcanic and large-scale ground movements, may be optimized to improve the precision. These include the traditional method of fixing control points, the Lagrange method, free net adjustment, the g-inverse method, and the Singular Value Decomposition (SVD) approach using the pseudo-inverse. A major characteristic of such inverse problem networks is that the system is rank deficient. This deficiency is solved using either exterior (i.e. a priori) or inner constraints. The former requires additional resources to provide the control points. In contrast, inner constraints methods do not require the imposition of external control and offer higher precision because the network geometry is preserved.

## 1 Introduction

The purpose of this chapter is to outline the various ways in which engineering survey monitoring networks are optimized. Such networks are used in monitoring deformation in large structures as well as soil movements.

The functional form is given by

$$\mathbf{y} = f(\mathbf{x}, \boldsymbol{\beta})$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations (typical angles and distances, including GPS observations), $f(.)$ is a function, $\mathbf{x}$ is a $k \times 1$ vector of variables, and $\boldsymbol{\beta}$ is a $k \times 1$ vector of parameters that are typically absolute or updated adjustments to coordinates. The nonlinear equation is often linearized using a Taylor series approximation about $\mathbf{x}_0$

W. Tan (✉)
Department of Building, School of Design and Environment, National University of Singapore, Kent Ridge Crescent, Singapore 117566, Singapore
e-mail: bdgtanw@nus.edu.sg

H. Xu and X. Wang (eds.), *Optimization and Control Methods in Industrial Engineering and Construction*, Intelligent Systems, Control and Automation: Science and Engineering 72,
DOI: 10.1007/978-94-017-8044-5_5, © Springer Science+Business Media Dordrecht 2014

so that

$$f(\mathbf{x}) = f(\mathbf{x}_0) + (\nabla f)^{\mathrm{T}}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2!}(\mathbf{x} - \mathbf{x}_0)^{\mathrm{T}}\mathbf{H}(\mathbf{x} - \mathbf{x}_0) + R.$$

Here $\nabla f$ is the gradient vector and $(.)^{\mathrm{T}}$ denotes the transpose. The Hessian matrix ($\mathbf{H}$) and remainder ($R$) are often ignored so that the model is linear. Setting $\mathbf{y} = f(\mathbf{x}) - f(\mathbf{x}_0)$, $\mathbf{X} = (\nabla f)^{\mathrm{T}}(\mathbf{x}_0)$ (the $n \times k$ design matrix), and $\boldsymbol{\beta} = (\mathbf{x} - \mathbf{x}_0)$ gives the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is the $n \times 1$ vector of error terms, and $\mathbf{I}$ is the identity matrix of order $n$. The estimated model is

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{2}$$

where $\mathbf{b}$ is the least squares estimator and $\mathbf{e}$ is the residual vector. Pre-multiplying both sides in by $\mathbf{X}^{\mathrm{T}}$ and using the orthogonal condition $\mathbf{X}^{\mathrm{T}}\mathbf{e} = \mathbf{0}$ gives the normal equations

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{N}^{-1}\mathbf{f} \tag{3}$$

where $\mathbf{N} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ and $\mathbf{f} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$. These equations are solved iteratively using $\mathbf{x}_0$ as the intimal guess, which is the standard Newton's method. For 2-D monitoring networks, the initial guess is computed using distance and angle observations, which provides a very good starting point. However, in 3-D photogrammetric or survey networks, determining $\mathbf{x}_0$ is more problematic, and more iterations are required.

Combining Eqs. (1) and (3),

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}.$$

Thus $\mathrm{E}(\mathbf{b}) = \boldsymbol{\beta}$, that is, the least squares estimator is unbiased. Taking variances on both sides,

$$\mathrm{Var}(\mathbf{b}) = \mathrm{Var}(\boldsymbol{\beta} + (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}. \tag{4}$$

For any given $\sigma$, one form of optimization is to use Eq. (4) to design a network (i.e. select $\mathbf{X}$) such that Trace[Var($\mathbf{b}$)] is minimum, that is, the sum of the elements of the main diagonal of Var($\mathbf{b}$) is minimized. Generally, for a trilateral or triangulation network, this implies selecting well-conditioned triangles.

In the weighted least squares case, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{W})$ where $\mathbf{W}$ is a known diagonal weight matrix of heteroscedastic variances. This may be transformed into the standard unweighted linear model by considering a transformation $\mathbf{T}$ on Eq. (1) so that

$$\mathbf{T}\mathbf{y} = \mathbf{T}\mathbf{X}\boldsymbol{\beta} + \mathbf{T}\boldsymbol{\varepsilon},$$

that is,

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$$

where $\mathbf{y}^* = \mathbf{T}\mathbf{y}$, $\mathbf{X}^* = \mathbf{T}\mathbf{X}$ and $\boldsymbol{\varepsilon}^* = \mathbf{T}\boldsymbol{\varepsilon}$. Then the normal equations become

$$\mathbf{b}^* = (\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*)^{-1}\mathbf{X}^{*\mathrm{T}}\mathbf{y}^*,$$

and

$$\mathrm{Var}(\mathbf{b}^*) = \sigma^2(\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*)^{-1}.$$

Since $\mathbf{W}$ is a diagonal matrix, $\mathbf{T}$ can be selected such that

$$\mathrm{Var}(\mathbf{T}\boldsymbol{\varepsilon}) = \mathbf{T}\,\mathrm{Var}(\boldsymbol{\varepsilon})\mathbf{T}^{\mathrm{T}} = \sigma^2(\mathbf{T}\mathbf{W}\mathbf{T}^{\mathrm{T}}) = \sigma^2\mathbf{I}.$$

In brief, the weighted least squares case can easily be transformed into the unweighted case without loss of generality. Henceforth, we need only deal with the unweighted case.

For the normal equations in Eq. (3), $\mathbf{N}$ is rank deficient, which means that $\mathbf{N}^{-1}$ does not exist. This is because a datum for the survey network has not been established and, as we shall see, the choice of datum affects network geometry and hence its precision.

## 2 Overcoming Rank Deficiency: Use of Exterior Constraints

In this section, four methods of imposing external constraints on the network to remove the rank deficiency are discussed.

### *2.1 Direct Substitution*

This is the oldest method. Here, control points with known coordinates are substituted into the normal equations to remove the rank deficiency. The originator of this method is unknown, but it can be found in standard textbooks such as [4, 12, 14, 19]. The substitution results in a slightly smaller system of equations. If there are $q$ constraints, then the normal equations have $q$ fewer equations.

If the number of constraints is just enough to remove the rank deficiency in $\mathbf{N}$, then it is called a minimally adjusted network. Otherwise, the network is over-constrained, which may be carried out so that the adjusted coordinates are consistent with the higher-order network. To the extent that the imposed constraints may be biased, the least square estimator will also be biased.

## 2.2 Imposition of Linear Constraints

Equivalently, the $q$ constraints required to overcome the rank deficiency may be written as

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{d} \tag{5}$$

where $\mathbf{R}$ is a $q \times k$ matrix of independent constraints and $\mathbf{d}$ is a $q \times 1$ vector of constants. These constraints are then added to Eq. (1) to obtain the augmented system

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{d} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{R} \end{bmatrix} \beta + \begin{bmatrix} \varepsilon \\ \mathbf{0} \end{bmatrix}.$$

The resulting normal equations are

$$[\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R}]\mathbf{b} = \mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{R}^{\mathrm{T}}\mathbf{d}. \tag{6}$$

The matrix on the left hand side is now of full rank and be inverted to find the least squares estimator $\mathbf{b}$.

A variant of this approach is to start with the normal equations in Eq. (3) so that

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{b} = \mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

Pre-multiplying the constraint equations $\mathbf{Rb} = \mathbf{d}$ on both sides by $\mathbf{R}^{\mathrm{T}}$ gives

$$\mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{R}^{\mathrm{T}}\mathbf{d}.$$

These equations may now be compatibly added, which gives Eq. (6) as before. Further,

$$\begin{aligned} \mathrm{Var}(\mathbf{b}) &= \mathrm{Var}[(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{R}^{\mathrm{T}}\mathbf{d})] \\ &= \mathrm{Var}[(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}] \\ &= \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}. \end{aligned} \tag{7}$$

This expression is difficult to evaluate, making it uncertain whether additional constraints can improve network precision.

## 2.3 Lagrange Method

The third classical approach is to use the method of Lagrange multipliers. The Lagrangean is

$$L = (\mathbf{y} - \mathbf{Xb})^{\mathrm{T}}(\mathbf{y} - \mathbf{Xb}) + 2\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Rb} - \mathbf{d}).$$

Here $\boldsymbol{\lambda}$ is a vector of Lagrange multipliers. The first-order conditions are

$$\partial L/\partial \mathbf{b} = -2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b} + 2\mathbf{R}^{\mathrm{T}}\boldsymbol{\lambda} = \mathbf{0};$$
$$\partial L/\partial \boldsymbol{\lambda} = \mathbf{R}\mathbf{b} - \mathbf{d} = \mathbf{0}.$$

These conditions may be compactly written as a [8] block system

$$\begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{X} & \mathbf{R}^{\mathrm{T}} \\ \mathbf{R} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{y} \\ \mathbf{d} \end{bmatrix}. \tag{8}$$

The matrix on the left-hand side is invertible, and the system may be solved for the least squares estimator $\mathbf{b}$. The vector $\boldsymbol{\lambda}$ is not required as part of the solution and hence need not be computed. This approach is seldom used for large networks because it involves solving a larger system of equations.

## 3 Use of Inner Constraints

Classical methods using external constraints are less precise than modern methods [11]. This is because the use of a few control points tends to "distort" the network geometry, resulting in large standard errors. The modern approaches overcome this defect by using inner constraints, g-inverse or minimum norm solutions that do not rely on externally imposed control points.

### *3.1 Free Net Adjustment*

To avoid deforming the network, [10] suggested that a larger subset of points or all the points in the network should be used to define the datum, rather than just a few control points. For the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, consider small changes so that

$$\Delta\mathbf{y} = \mathbf{X}\Delta\boldsymbol{\beta} + \Delta\boldsymbol{\varepsilon}.$$

If $\Delta\boldsymbol{\varepsilon} = \mathbf{0}$, then $\Delta\mathbf{y} = \mathbf{X}\Delta\boldsymbol{\beta}$. We then consider changes in $\Delta\boldsymbol{\beta}$ such that $\Delta\mathbf{y}$ remains unchanged (i.e. equals $\mathbf{0}$). Such changes are implied by imposing inner constraints so that there is no overall rotation, translation, or scale changes in the network. Then

$$\Delta\boldsymbol{\beta} = \mathbf{G}\Delta\mathbf{t}$$

where, for a 2-D survey network [1, 5, 7],

$$\mathbf{G}\Delta\mathbf{t} = \begin{bmatrix} 1 & 0 & -y_1 \\ 0 & 1 & x_1 \\ 1 & 0 & -y_2 \\ 0 & 1 & x_2 \\ \vdots & & \\ 1 & 0 & -y_m \\ 0 & 1 & x_m \end{bmatrix} \begin{bmatrix} \Delta t_x \\ \Delta t_y \\ \Delta\phi \end{bmatrix}.$$

Here $\mathbf{G}$ is matrix representing the linear transformation, $\Delta t_x$, $\Delta t_y$ represent a small translation of each point and $\Delta\varphi$ represents a small rotation of each point with respect to the centroid. Hence,

$$\Delta\mathbf{y} = \mathbf{X}\Delta\boldsymbol{\beta} = \mathbf{X}\mathbf{G}\Delta\mathbf{t} = \mathbf{0}.$$

Since $\Delta\mathbf{t}$ is non-zero, we have

$$\mathbf{X}\mathbf{G} = \mathbf{0}. \tag{9}$$

In other words, the columns of $\mathbf{G}$ span the null space of $\mathbf{X}$. The derivation of $\mathbf{G}$ is not always so easy [13]. Tan [17] provided a simple derivation of the $\mathbf{G}$ matrix for a 3-D network such as those found in photogrammetry.

The above derivation uses geometrical constraints, that is, the network must not have overall translation, rotation, or scale changes with respect to the centroid. An alternative derivation is purely algebraic and uses the rank-nullity theorem. Recall from Eq. (6) that the normal equations for the augmented system are

$$[\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R}]\mathbf{b} = \mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{R}^{\mathrm{T}}\mathbf{d}.$$

we select $\mathbf{R}$ such that

$$\mathbf{X}\mathbf{R}^{\mathrm{T}} = \mathbf{0}, \tag{10}$$

that is, that columns of $\mathbf{R}^{\mathrm{T}}$ span the null space of $\mathbf{X}$. The reason for this choice is that the left-hand side of Equation becomes invertible. To see this, we solve for its null space, that is,

$$[\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R}]\mathbf{g} = \mathbf{0}$$

for some vector $\mathbf{g}$. Pre-multiplying both sides by $\mathbf{R}$ gives

$$\mathbf{R}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{g} + \mathbf{R}\mathbf{R}^{\mathrm{T}}\mathbf{R}\mathbf{g} = \mathbf{0}.$$

Because $\mathbf{X}\mathbf{R}^{\mathrm{T}} = \mathbf{0}$, its transpose is also zero. Further, $\mathbf{R}\mathbf{R}^{\mathrm{T}}\mathbf{R} \neq \mathbf{0}$ so $\mathbf{g} = \mathbf{0}$. In other words, the null space of $\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R}$ is the null vector. Thus the nullity of $\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R}$ is zero and, by the rank-nullity theorem, $\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R}$ is of full rank and is therefore invertible. By comparing Eqs. (9) and (10), it can further be seen that $\mathbf{G} = \mathbf{R}^{\mathrm{T}}$. This links the geometric and algebraic relations.

In summary, free net adjustment imposes the conditions that there should not be overall rotational, translational, and scale changes in the network with respect to the centroid. These constraints are based on approximate coordinates and, to the extent that these coordinates are biased, the least squares estimator is also biased. However, unlike traditional approaches, the constraints are based on the centroid (also called a fictitious station) and use all the points or a large subset of points to avoid deforming the network.

### *3.2 The Singular Value Decomposition (SVD) Approach*

The SVD approach does not use the null space (or equivalently, inner geometrical constraints based on the centroid) to remove the rank deficiency. Instead, it solves for the minimum norm solution from an infinite number of solutions e.g. [2, 9, 18].

The SVD of $\mathbf{X}$ is given by

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$$

where $\mathbf{U}$ is an $n \times n$ orthogonal matrix, $\mathbf{D}$ is a $n \times k$ diagonal matrix of $r \leq k$ singular values $\sigma_1, \ldots, \sigma_r$ (not to be confused with the common variance of observations, $\sigma^2$), and $\mathbf{V}$ is a $k \times k$ orthogonal matrix (i.e. $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$). Premultiplying both sides by $\mathbf{X}^{\mathrm{T}}$ gives

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = (\mathbf{V}\mathbf{D}\mathbf{U}^{\mathrm{T}})\mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}} = \mathbf{V}\mathbf{D}^2\mathbf{V}^{\mathrm{T}}.$$

The normal equations in Eq. (3) may now be written as

$$\mathbf{V}\mathbf{D}^2\mathbf{V}^{\mathrm{T}}\mathbf{b} = \mathbf{V}\mathbf{D}\mathbf{U}^{\mathrm{T}}\mathbf{y}.$$

Pre-multiplying by $\mathbf{V}^{\mathrm{T}}$ leads to

$$\mathbf{D}^2\mathbf{V}^{\mathrm{T}}\mathbf{b} = \mathbf{D}\mathbf{U}^{\mathrm{T}}\mathbf{y},$$

that is,

$$\mathbf{D}^2\mathbf{h} = \mathbf{D}\mathbf{c} \tag{11}$$

where

$$\mathbf{h} = \mathbf{V}^{\mathrm{T}}\mathbf{b} \tag{12}$$

and

$$\mathbf{c} = \mathbf{U}^{\mathrm{T}}\mathbf{y}.$$

Because $\mathbf{U}^{\mathrm{T}}$, $\mathbf{y}$, $\mathbf{D}$, $\mathbf{V}^{\mathrm{T}}$ and $\mathbf{c}$ are known, Eq. (11) may be solved for $\mathbf{h}$ and the result is used in Eq. (12) to solve for $\mathbf{b}$. Further, from Eq. (4),

$$\mathrm{Var}(\mathbf{b}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^2(\mathbf{V}\mathbf{D}^2\mathbf{V}^{\mathrm{T}})^{-1} = \sigma^2\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^{\mathrm{T}}.$$

To obtain the minimum norm solution, we use Eq. (12) to minimize

$$\mathbf{b}^{\mathrm{T}}\mathbf{b} = (\mathbf{V}\mathbf{h})^{\mathrm{T}}\mathbf{V}\mathbf{h} = \mathbf{h}^{\mathrm{T}}\mathbf{h}.$$

From Eq. (11), if the $i$th element of $\mathbf{h}$ is $h_i$ and that of $\mathbf{c}$ is $c_i$, then

$$h_i = c_i/\sigma_i, \qquad i = 1, \ldots, r.$$

Thus,

$$\mathbf{b}^{\mathrm{T}}\mathbf{b} = \sqrt{\left\{\left(h_1^2 \ldots + h_r^2\right) + \left(h_{r+1^2} \ldots + h_k^2\right)\right\}}.$$

Letting $h_{r+1} = \ldots = h_k = 0$ gives the minimum norm least squares solution.

A neater way of presenting the above steps is to define

$$\mathbf{D}^{+} = \begin{bmatrix} 1/\sigma_1 & \cdots & 0 & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1/\sigma_r & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

so that Eq. (12) may be written as

$$\mathbf{b} = \mathbf{V}\mathbf{D}^{+}\mathbf{U}^{\mathrm{T}}\mathbf{y} = \mathbf{X}^{+}\mathbf{y}. \tag{13}$$

The matrix $\mathbf{X}^{+} = \mathbf{V}\mathbf{D}^{+}\mathbf{U}^{\mathrm{T}}$ is called the pseudo-inverse of $\mathbf{X}$. The pseudo-inverse approach is elegant because it does not impose external constraints on the network. Once the SVD of $\mathbf{X}$ is calculated, it is a simple matter of computing $\mathbf{b}$ and Var($\mathbf{b}$). The latter is obtained by taking the variance in Eq. (13), that is,

$$\mathrm{Var}(\mathbf{b}) = \sigma^2\mathbf{V}(\mathbf{D}^{+})^{\mathrm{T}}\mathbf{D}^{+}\mathbf{V}^{\mathrm{T}}. \tag{14}$$

The main drawback is that computing the SVD is computationally more intensive. However, the minimum norm solution is attractive because the mean square error (MSE) is given by

$$\begin{aligned} \mathrm{MSE}(\mathbf{b}) &= E[(\mathbf{b} - \boldsymbol{\beta})^{\mathrm{T}}(\mathbf{b} - \boldsymbol{\beta})] \\ &= \mathrm{Trace}[\mathrm{Var}(\mathbf{b})] + [\mathrm{bias}(\mathbf{b})]^{\mathrm{T}}[\mathrm{bias}(\mathbf{b})]. \end{aligned}$$

Here E[.] is the expectations operator. It can be seen that Var($\mathbf{b}$) has minimum trace if $\mathbf{b}$ is unbiased.

### 3.3 The g-Inverse Method

The generalized inverse (g-inverse) technique in solving rank-deficient systems is well known [15, 16]. Its main disadvantage is that it solves for a particular solution out of an infinite number of solutions because, unlike the regular inverse, the g-inverse is not unqiue. This particular solution may not have desirable properties. For this reason, it is better to use a variant of the g-inverse technique to find the minimum norm solution, with optimizes the trace of Var($\mathbf{b}$) as shown earlier.

From Eq. (3), the system of normal equations is

$$\mathbf{N}\mathbf{b} = \mathbf{f}.$$

If a g-inverse of $\mathbf{N}$ is defined as

$$\mathbf{N}\mathbf{N}^{-}\mathbf{N} = \mathbf{N},$$

then post-multiplying both sides by $\mathbf{b}$ gives

$$\mathbf{N}\mathbf{N}^{-}\mathbf{N}\mathbf{b} = \mathbf{N}\mathbf{b}.$$

Comparing this equation with the normal equations, we have

$$\mathbf{N}\mathbf{N}^{-}\mathbf{f} = \mathbf{f}.$$

That is,

$$\mathbf{b} = \mathbf{N}^{-}\mathbf{f}.$$

The solution is not unique because

$$\mathbf{b} = \mathbf{N}^{-}\mathbf{f} + (\mathbf{N}^{-}\mathbf{N} - \mathbf{I})\mathbf{w}$$

where $\mathbf{w}$ is an arbitrary vector is also a solution to the normal equations, that is,

$$\mathbf{N}[\mathbf{N}^{-}\mathbf{f} + (\mathbf{N}^{-}\mathbf{N} - \mathbf{I})\mathbf{w}] = \mathbf{N}\mathbf{N}^{-}\mathbf{f} + \mathbf{N}(\mathbf{N}^{-}\mathbf{N} - \mathbf{I})\mathbf{w} = \mathbf{f}.$$

Two further properties of g-inverses are important, namely,

(a) $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}$ is a g-inverse of $\mathbf{X}$; and
(b) $\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}$ is invariant to the choice of $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}$.

One approach to overcome this lack of uniqueness is to use estimable functions [3], that is, if a solution to the normal equations $\mathbf{b}$ depends on particular g-inverse, it may be desirable to estimate a linear function $\mathbf{m}^{\mathrm{T}}\boldsymbol{\beta}$ rather than $\boldsymbol{\beta}$ itself. Now,

**Table 1** Data for simple loop level network

| From | To | Observed height difference (m) | Approximate height (m) |
|---|---|---|---|
| $A$ | $B$ | $+2.1$ | $H_A = 10.0$ |
| $B$ | $C$ | $-1.0$ | $H_B = 12.1$ |
| $C$ | $A$ | $-0.9$ | $H_C = 11.1$ |
| | | | Total $= 33.2$ |
| | | | Mean height $= 11.067$ |

$$\begin{aligned} E[\mathbf{m}^{\mathrm{T}}\mathbf{b}] &= \mathbf{m}^{\mathrm{T}} E[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{y}] \\ &= \mathbf{c}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} \quad \text{where } \mathbf{m}^{\mathrm{T}} = \mathbf{c}^{\mathrm{T}}\mathbf{X} \\ &= \mathbf{c}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} \qquad\qquad\qquad \text{using Property (a)} \\ &= \mathbf{m}^{\mathrm{T}}\boldsymbol{\beta} \text{ (unbiased).} \end{aligned}$$

Further,

$$\mathbf{m}^{\mathrm{T}}\mathbf{b} = \mathbf{c}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{y},$$

which, by Property (b) above, is invariant to the choice of $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}$. Thus $\mathbf{m}^{\mathrm{T}}\mathbf{b}$ is unbiased and invariant to the choice of g-inverse, that is, $\mathbf{m}^{\mathrm{T}}\boldsymbol{\beta}$ is said to be estimable.

In engineering survey monitoring networks, it is clear from geometry that, for instance, height differences are estimable even if the absolute heights are not estimable for lack of a a datum. However, estimable functions (i.e. linear functions of coordinates) are difficult to interpret geometrically for triangulation, trilateration or photogrammetric networks. Hence, it is still desirable to compute the minimum norm solution and not use the estimable function approach.

The basic idea is to use a two-step procedure. We start from the rank-deficient normal equations. Instead of imposing additional constraints like the classical and free net approaches, we first delete the redundant rows of $\mathbf{N}$ and $\mathbf{f}$ so that $\mathbf{N}$ is of full rank. In the second step, the underdetermined system of equations is then solved for the minimum norm solution.

Let the reduced normal equations be

$$\mathbf{N}_0\mathbf{b} = \mathbf{f}_0. \tag{15}$$

If the rank deficiency is $q$, $\mathbf{N}_0$ is a $(k-q) \times k$ matrix of full row rank and $\mathbf{f}_0$ is the corresponding $(k-q) \times 1$ vector. We minimize $\mathbf{b}^{\mathrm{T}}\mathbf{b}$ subject to the constraint that $\mathbf{N}_0\mathbf{b} = \mathbf{f}_0$. The Lagrangean is

$$L = \mathbf{b}^{\mathrm{T}}\mathbf{b} + \boldsymbol{\lambda}^{\mathrm{T}}(\mathbf{f}_0 - \mathbf{N}_0\mathbf{b}).$$

The first-order condition is

$$\partial L/\partial \mathbf{b} = 2\mathbf{b} - \mathbf{N}_0^{\mathrm{T}}\boldsymbol{\lambda} = \mathbf{0}.$$

Pre-multiplying both sides by $\mathbf{N}_0$ gives

$$2\mathbf{N}_0\mathbf{b} = \mathbf{N}_0\mathbf{N}_0^{\mathrm{T}}\boldsymbol{\lambda}.$$

Thus,

$$\boldsymbol{\lambda} = 2(\mathbf{N}_0\mathbf{N}_0^{\mathrm{T}})^{-1}\mathbf{f}_0.$$

Substituting $\boldsymbol{\lambda}$ into into the first-order condition, we have

$$\mathbf{b} = \mathbf{N}_0^{\mathrm{T}}(\mathbf{N}_0\mathbf{N}_0^{\mathrm{T}})^{-1}\mathbf{f}_0. \tag{16}$$

Because $\mathbf{N}_0$ is of full rank row rank, the matrix $\mathbf{N}_0\mathbf{N}_0^{\mathrm{T}}$ is invertible. Thus, the solution $\mathbf{b}$ is relatively easy to compute because, unlike the SVD approach, the eigenvalues and eigenvectors are not required. Further,

$$\begin{aligned}\mathrm{Var}(\mathbf{b}) &= \mathrm{Var}[\mathbf{N}_0^{\mathrm{T}}(\mathbf{N}_0\mathbf{N}_0^{\mathrm{T}})^{-1}\mathbf{f}_0] \\ &= \mathrm{Var}[\mathbf{N}_0^{\mathrm{T}}(\mathbf{N}_0\mathbf{N}_0^{\mathrm{T}})^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{y}] \\ &= \sigma^2\mathbf{N}_0^{\mathrm{T}}(\mathbf{N}_0\mathbf{N}_0^{\mathrm{T}})^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0(\mathbf{N}_0\mathbf{N}_0^{\mathrm{T}})^{-1}\mathbf{N}_0.\end{aligned}$$

Here $\mathbf{X}_0$ is obtained by deleting the corresponding columns of $\mathbf{X}$. Again, the multiplication is relatively straightforward, and the solution is optimized to minimum norm.

## 4 Applications

In this section, a simple example is used to illustrate all the techniques. Consider a clockwise loop level network from point A to B, B to C, and then back to A (Table 1). Let $H_A$ be the height (reduced level) of point A, which is arbitrarily set to 10.0 m and observed height differences are then used to compute the approximate heights as shown in the last column in the table. The mean height is required in free net adjustment. The example has been kept simple to illustrate the computational steps.

### 4.1 Classical Methods

The observations equations $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ are

$$\begin{bmatrix} 2.1 \\ -1 \\ -0.9 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} H_A \\ H_B \\ H_C \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix}.$$

From Eq. (3), the rank-deficient normal equations $\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$ are

$$\begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \mathbf{b} = \begin{bmatrix} -3 \\ 3.1 \\ -0.1 \end{bmatrix}. \tag{17}$$

The rank deficiency is one, implying that only one constraint equation is required to solve the system. This may be introduced in the traditional way by fixing $H_A = 10$, that is,

$$\mathbf{R}\boldsymbol{\beta} = [100]\boldsymbol{\beta} = [10]$$

and solving Eq. (6). Alternatively, we can use Lagrange's method and solve the augmented system in Eq. (8) for the estimator $\mathbf{b}$.

### 4.2 Free Net Adjustment

In free net adjustment we impose the constraint that the mean height $(H_A + H_B + H_C)/3$ is 11.067, that is,

$$\mathbf{R}\boldsymbol{\beta} = [111]\boldsymbol{\beta} = [33.2].$$

From Eq. (6), the least square estimator is

$$\begin{aligned} \mathbf{b} &= [\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R}]^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{R}^{\mathrm{T}}\mathbf{d}) \\ &= \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/3 \end{bmatrix} \begin{bmatrix} 30.2 \\ 36.3 \\ 33.1 \end{bmatrix} = \begin{bmatrix} 10.067 \\ 12.100 \\ 11.033 \end{bmatrix}. \end{aligned}$$

Note that the adjusted value of $H_A$ is 10.067 m and not 10.0 m because the mean height is based on approximate heights. If desired, $H_A$ may be reset to 10.0 m by subtracting 0.067 m and consequently $H_B = 12.03$ m and $H_C = 10.97$ m. Further, from Eq. (7),

$$\mathrm{Var}(\mathbf{b}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1} = (\sigma^2/9)\begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}.$$

### 4.3 The SVD Approach

In the SVD approach, we first compute the eigenvalues and eigenvectors of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, which give matrices $\mathbf{D}$ and $\mathbf{V}$ respectively. Then $\mathbf{U}$ is computed column by column using $\mathbf{XV} = \mathbf{UD}$ so that

$$\mathbf{X} = \mathbf{UDV}^{\mathrm{T}} = \begin{bmatrix} 1/\sqrt{6} & \sqrt{3}/\sqrt{6} & u_{13} \\ -1/\sqrt{6} & \sqrt{3}/\sqrt{6} & u_{23} \\ 2/\sqrt{6} & 0 & u_{33} \end{bmatrix}\begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & \sqrt{3} & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} -1/\sqrt{2} & 0 & 1/\sqrt{2} \\ -1/\sqrt{6} & 2/\sqrt{6} & -1/\sqrt{6} \\ 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \end{bmatrix}.$$

Thus, from Eq. (13),

$$\mathbf{b} = \mathbf{VD}^{+}\mathbf{U}^{\mathrm{T}}\mathbf{y} = \begin{bmatrix} -1 \\ 1.03 \\ -0.03 \end{bmatrix}.$$

This is the minimum norm solution. The absolute heights are unknown and arbitrary in the SVD approach; only the height differences matter. Thus if 11.0 m is added to all three heights, then $H_A = 10.0\,\mathrm{m}$, $H_B = 12.03\,\mathrm{m}$, and $H_C = 10.97\,\mathrm{m}$, the same answer as that of free net adjustment. From Eq. (14),

$$\mathrm{Var}(\mathbf{b}) = \sigma^2\mathbf{V}(\mathbf{D}^{+})^{\mathrm{T}}\mathbf{D}^{+}\mathbf{V}^{\mathrm{T}} = (\sigma^2/9)\begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}.$$

This is identical to the result using free net adjustment.

### 4.4 The g-Inverse Method

In the two-step g-inverse procedure, we remove the third redundant equation in Eq. (17) so that $\mathbf{N}_0\mathbf{b} = \mathbf{f}_0$ is given by

$$\begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \end{bmatrix}\mathbf{b} = \begin{bmatrix} -3 \\ 3.1 \end{bmatrix}.$$

Similarly, the third column of $\mathbf{X}$ may be removed so that

$$\mathbf{X}_0 = \begin{bmatrix} -1 & 1 \\ 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Then, straightforward multiplication gives

$$\mathbf{b} = \mathbf{N}_0^{\mathrm{T}}(\mathbf{N}_0\mathbf{N}_0^{\mathrm{T}})^{-1}\mathbf{f}_0 = \begin{bmatrix} -1 \\ 1.03 \\ -0.03 \end{bmatrix}.$$

The result is identical to that of the SVD and free net adjustment approaches. Further, it is a simple matter to compute

$$\mathrm{Var}(\mathbf{b}) = \sigma^2\mathbf{N}_0^{\mathrm{T}}(\mathbf{N}_0\mathbf{N}_0^{\mathrm{T}})^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0(\mathbf{N}_0\mathbf{N}_0^{\mathrm{T}})^{-1}\mathbf{N}_0 = (\sigma^2/9)\begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}.$$

Again, the result is identical to the SVD and free net adjustment techniques. Unlike free net adjustment, there is no requirement to fix the height of the centroid. The two-step procedure is also computationally simpler than the SVD approach because there is no need to compute the eigenvalues and eigenvectors of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.

## Conclusion

This chapter has discussed the various ways in which engineering survey monitoring networks may be optimized. Unlike many other optimization problems, the issue here is rank deficiency rather than convergence through iterations or other search techniques. The classical methods are well known but they do not optimize the network precision. The modern approach began with [10], the originator of free net adjustment. The SVD approach is well known in many fields (e.g. [6]), as is the g-inverse method. In this chapter, it is suggested that the g-inverse technique may be modified to compute the minimum norm estimator.

## Appendix: Notation

| | |
|---|---|
| $\mathbf{y}$ | $n \times 1$ vector of observations with $i$th element $y_i$ |
| $\mathbf{X}$ | $n \times k$ design matrix with $i$th column $[\mathbf{x}_i]$ |
| $\boldsymbol{\beta}$ | $k \times 1$ vector of parameters |
| $\boldsymbol{\varepsilon}$ | $n \times 1$ vector of error terms with $i$th element $\varepsilon_i$ |
| $\sigma^2$ | $\mathrm{Var}(\varepsilon_i)$ |

| | |
|---|---|
| $\mathbf{I}$ | $n \times n$ identity matrix |
| $N(.)$ | Normal distribution |
| $\mathbf{b}$ | Least squares estimator of $\boldsymbol{\beta}$ with $i$th element $b_i$ |
| $\mathbf{N}$ | $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ |
| E(.) | Expectation |
| Var(.) | Variance |
| $\mathbf{f}$ | $\mathbf{X}^{\mathrm{T}}\mathbf{y}$ |
| $\mathbf{W}$ | $n \times n$ known diagonal weight matrix |
| $\mathbf{T}$ | $n \times n$ transformation matrix |
| $\mathbf{y}^*$ | $\mathbf{Ty}$ |
| $\mathbf{X}^*$ | $\mathbf{TX}$ |
| $\boldsymbol{\varepsilon}^*$ | $\mathbf{T}\boldsymbol{\varepsilon}$ |
| $\mathbf{b}^*$ | Weighted least squares estimator |
| $\mathbf{R}$ | $q \times k$ coefficient matrix of constraints |
| $\mathbf{d}$ | $q \times 1$ vector of constants |
| $L$ | Lagrangean function |
| $\boldsymbol{\lambda}$ | Vector of Lagrange multipliers |
| $\mathbf{g}$ | Arbitrary vector |
| $\mathbf{U}$ | $n \times n$ orthogonal matrix |
| $\mathbf{D}$ | $n \times k$ diagonal matrix of $r \leq k$ singular values $\sigma_1, \ldots, \sigma_r$ |
| $\lambda_i$ | Eigenvalue of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, $\lambda_i = \sigma_i^2$ |
| $\sigma_i$ | Singular values |
| $\mathbf{V}$ | $k \times k$ orthogonal matrix |
| $\mathbf{h}$ | $\mathbf{V}^{\mathrm{T}}\mathbf{b}$, with element $h_i$ |
| $\mathbf{c}$ | $\mathbf{U}^{\mathrm{T}}\mathbf{y}$, with element $c_i$; also denotes an arbitrary vector |
| $\mathbf{D}^+$ | Matrix whose main diagonal consists of the inverses of singular values |
| $\mathbf{X}^+$ | Pseudo-inverse of $\mathbf{X}$ |
| MSE(.) | Mean square error |
| Tr(.) | Trace |
| $\mathbf{N}^-$ | g-inverse of $\mathbf{N}$ |
| $\mathbf{w}$ | Arbitrary vector |
| $\mathbf{m}^{\mathrm{T}}\boldsymbol{\beta}$ | Linear function of $\boldsymbol{\beta}$ |
| $\mathbf{N}_0$ | Reduced matrix in normal equations |
| $\mathbf{f}_0$ | Reduced vector in normal equations |
| $\mathbf{X}_0$ | Reduced design matrix |
| $H_A$ | Height of point A; similarly for $H_B$ |

## References

1. Blaha G (1971) Inner adjustment constraints with emphasis on range observations. Report No. 148, Department of Geodetic Science, Ohio State University
2. Bjorck A (1996) Numerical methods for least squares problems. Siam, Philadelphia

3. Bose R (1949) Least squares aspects of the analysis of variance. Institute of Statistics, North Carolina University, Chapel Hill
4. Caspary W (1987) Concepts of network and deformation analysis. School of Geomatic Engineering, UNSW, Sydney
5. Cooper M (1987) Control surveys in civil engineering. William Collins, London
6. Galub G, van Loan C (1996) Matrix computations. Johns Hopkins Press, Baltimore
7. Granshaw S (1980) Bundle adjustment methods in engineering photogrammetry. Photogram Rec 10:181–207
8. Helmert F (1955) Adjustment computation by the method of least squares. Aeronautical Chart and Information Center, St Louis
9. Lawson C, Hanson R (1974) Solving least squares problems. Prentice-Hall, New Jersey
10. Meissl P (1962) Die innere genauigkeit eines punkthaufens. Oz Vermessungswesen 50:159–65
11. Meissl P (1982) Least squares adjustment: a modern approach. Geodatische Institute der Technischen Universitat Graz, Graz
12. Mikhail E, Gracie G (1981) Analysis and adjustment of survey measurements. Van Nostrand Reinhold, New York
13. Papo H, Perelmuter A (1982) Free net analysis in close range photogrammetry. Photogram Eng Remote Sens 48(4):571–576
14. Rainsford H (1958) Survey adjustments and least squares. Constable, London
15. Rao C (1973) Linear statistical inference and its applications. Wiley, New York
16. Rao C, Mitra S (1971) Generalized inverse of matrices and its applications. Wiley, New York
17. Tan W (2005) Inner constraints for 3-D survey networks. J Spat Sci 50(1):91–94
18. Tan W (2009) The structure of leveling networks. J Spat Sci 54(1):37–43
19. Wolf P, Ghilani C (1997) Adjustment computations. Wiley, New York

# Distributed Fault Detection Using Consensus of Markov Chains

**Dejan P. Jovanović and Philip K. Pollett**

**Abstract** We propose a fault detection procedure appropriate for use in a variety of industrial engineering contexts, which employs consensus among a group of agents about the state of a system. Markov chains are used to model subsystem behaviour, and consensus is reached by way of an iterative method based on estimates of a mixture of the transition matrices of these chains. To deal with the case where system states cannot be observed directly, we extended the procedure to accommodate Hidden Markov Models.

**Keywords** Fault detection · Consensus algorithm · Mixtures of Markov chains · The EM algorithm · Hidden Markov Model (HMM) · Multi-agent systems

## 1 Introduction

With the development of new types of engineering systems, such robotic networks and networks of unmanned aerial vehicles (UAVs), and initiatives to modernize already established engineering systems such as "smart" power grids, the problem of fault detection is becoming more import, because early detection of deviations in system characteristics from the norm leads to increased reliability and maintainability [1]. Early detection is achieved by building local behavioural models and establishing information exchange protocols to estimate local characteristics for decentralized decision making.

Theoretical background on fault detection is part of the wider milieu of procedures for recognizing unwanted behaviour in monitored systems. Fault diagnosis generally

D. P. Jovanović (✉) · P. K. Pollett
Department of Mathematics, University of Queensland, Brisbane QLD 4072, Australia
e-mail: dejan.jovanovic@uqconnect.edu.au

P. K. Pollett
e-mail: pkp@maths.uq.edu.au

H. Xu and X. Wang (eds.), *Optimization and Control Methods in Industrial Engineering and Construction*, Intelligent Systems, Control and Automation: Science and Engineering 72,
DOI: 10.1007/978-94-017-8044-5_6, 

comprises three sequential steps (known as the fundamental tasks of fault diagnosis [2]). The first, *fault detection*, is to decide whether the characteristics of the system in question are outside permissible limits. The second, *fault identification*, is to determine which subsystems contain a fault of a particular type and the time when it occurred. Finally, *fault analysis* provides insight into the time-varying characteristics of the fault and the scale of disturbance that occurred.

We focus here primarily on fault detection. In order to identify change, we need an adequate reference model for system features. This is the backbone of the *model-based* fault detection approach [2–6] adopted here. Three common methods used in feature generation within the context of model-based fault detection are parameter estimation [7, 8], state estimation [3, 9] and parity (consistency) checking [5]. Once estimated, the present state is compared with that of nominal (normal) system behaviour and a residual is generated that measures any change. In model-based fault detection the full set of residuals is used for decision making, and a change in their mean and/or covariance signals a fault [6].

The choice of model depends on the problem at hand. However, it is often impractical to build a model for the entire system. This is particularly true for distributed systems, where there are many interrelated and interconnected parts. A natural approach is to simplify the task by first decomposing the system into a number of subsystems, which would usually be spatially separated and assumed to evolve independently. The decomposition may be deterministic [10] or probabilistic [11]. Attached to each subsystem is a set of independent local observations and a set of local parametric models that describe different working conditions. We will used the term "agent" as an abstraction that integrates these two components. A group decision is accomplished through interaction between neighbouring agents, and one of our goals is determine the conditions for logical (behavioural) consensus [12] among group members.

A state-space model is a common option for many practical problems in fault detection [9], and a Kalman estimator [13] is frequently used to estimate the mean and covariance of the state. However, the latter approach suffers from a lack of robustness to noise and uncertainties [6]. Our approach is different. Instead of generating residuals, we estimate the probability distribution of the state from the given observations. This is compared with corresponding distribution in the normal mode. The "distance" between these distributions is measured in order to decide on the presence of a fault. We assume that under any given operating mode, be it normal or faulty, the state of any subsystem can be described faithfully by a discrete-state Markov chain [14]. Normal operating conditions are described by a single model and there are a number of models indicating a fault. We assume each agent can accurately detect faults in its own subsystem.

Achieving consensus by arriving at a common distribution representing belief among agents is an idea that goes back to the early sixties, when Stone [15] introduced pooling of opinions to determine a group decision. Stone assumed that each opinion was modelled by a continuous probability distribution, and the opinion pool as a mixture of distributions. However, he considered only equal mixture weights. DeGroot [16] extended this idea in two ways: first, by introducing an iterative

approach whereby each agent revises its own subjective probability distribution based on changes in other group members opinions, and, second, by allowing the possibility of distinct weights and specifying convergence conditions for the iteration. The convergence conditions were generalized further by Chatterjee and Seneta [17]. We follow DeGroot's approach, but extend this using Markov chains rather than continuous distributions. In our method, agents reach consensus about the transition matrices that govern changes in subsystem state within each operating mode. At each iteration agent $i$ revises its own transition matrix by taking a weighted sum of the other agents' matrices and its own. The weights are selected optimally using an Expectation-Maximization (EM) framework. As we shall see, our method extends easily to the case of hidden Markov chains, thus allowing for states that may not be directly observable.

Once consensus is achieved, fault diagnosis can commence. To decide whether there is a fault, each agent compares its local distribution with a consensus distribution, being the stationary distributions of the local and the consensus transition matrices, respectively. If the "distance" between these distributions is greater than 0, a fault is recorded. Once it has been established that there are faulty subsystems, the next step is to determine which subsystems are faulty and what type of fault is present. There are two scenarios. In the first, a subsystem identifies its own fault and checks if there are group members affected by the same fault (the agent compares its own local model with other local models in the group by measuring the distance between corresponding stationary distributions). In the second, if the subsystem has no fault then the faulty subsystems can be detected in the same manner by again comparing stationary distributions. A range of different distance values indicates multiple faults.

We note that a consensus algorithm has been used in the work of Franco et al. [18], Ferrari et al. [19] and Stanković et al. [20]. It differs from ours in that first and second order moments only were used, rather than the entire distribution. We note also that Petri Nets have been used extensively in fault detection in distributed systems [21]. Particularly interesting is the work of Benveniste et al. [22], who introduced a probabilistic extension of Petri Nets for distributed and concurrent systems (components that evolve independently) and applied this to fault detection. However, they did not consider the problem of fault detection within a consensus framework. Our work is motivated by the need to detect faults in electrical power systems. Accordingly, we mention also the work of Kato et al. [23] and Garza et al. [24]. In [23] a multi-agent approach was suggested to locate and isolate fault zones. In [24] this problem was considered within a probabilistic framework using dynamic Bayesian networks, rather than the present consensus framework.

The rest of this chapter is organized as follows. The problem is formulated in Sect. 2. This is followed in Sect. 3 with a derivation of the likelihood function for our mixture of Markov chains and the EM procedure for selecting optimal weights. Section 4 contains an extension to the case of unobserved Markov chains. Finally, a fault diagnosis scheme and simulation results are presented in Sects. 5 and 6, respectively.

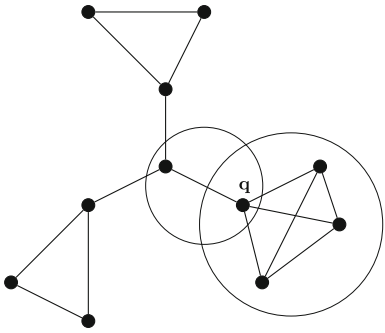



**Fig. 1** Communication graph of an agent network

## 2 Problem Formulation

It is assumed that each agent has local observations of the state of its subsystem, and that the local model for changes in state is represented by a transition probability matrix. Furthermore, it is assumed that agents can exchange information over a computer network; specifically, each agent can know other agents' transition matrices. The idea is to modify the transition matrices within a group of agents in such a way that, under certain conditions, those of all agents in the group converge to a common transition matrix.

The underlying distributed system is represented by undirected graph $G = (V, E)$ whose vertices $V = \{1, \ldots, n\}$ represent agents (measurement and monitoring points) and whose edges $E$ ($E \subseteq V \times V$) represent communication links. It is assumed that $G$ itself is connected and composed of one of more complete subgraphs (cliques), each corresponding to a group of agents trying to achieve consensus. We assume that the cliques are known in advance (we do not consider a problem of finding them [25]). An example of one such graph is given in Fig. 1, where a 2-vertex clique and a 4-vertex clique have been circled. The neighbourhood of an agent $q$ is defined as the set of agents $\mathcal{N}_q \subseteq V$ such that $\mathcal{N}_q \triangleq \{p \in V | (q, p) \in E\}$. So a given agent can potentially belong to more than one group. This is illustrated in Fig. 1; notice that vertex $q$ belongs to two cliques. We will assume that if an agent is a member of more than one group, it will engage independently in achieving consensus within those groups; it will not share information among the groups. Additionally, we will suppose that communication links between group members are completely reliable with no latency.

Suppose that there are $K$ agents all of whom have the same set of subsystem states $S = \{1, 2, \ldots, N\}$. Starting from iteration $\tau = 1$, agent $i$ updates its transition matrix by taking a weighted sum of the other agents' matrices and its own. Let $\mathbf{P}_i^{(\tau)}$ be the transition matrix of agent $i$ at iteration $\tau$ ($\tau \geq 1$). Then,

$$\mathbf{P}_i^{(\tau)} = \sum_{j=1}^{K} \psi_{ij} \mathbf{P}_j^{(\tau-1)}, \qquad i = 1, \ldots, K, \tag{1}$$

where $\Psi = [\psi_{ij}]$ is a $K \times K$ is ergodic stochastic matrix with strictly positive entries; $\psi_{ij}$ is a "consensus rating" assigned by agent $i$ to agent $j$ to rate the influence of agent $j$ on agent $i$. The transition matrices $\mathbf{P}_j^{(0)}$ at iteration $\tau = 0$ are the initially pooled transition matrices.

The updating procedure at iteration $\tau$ is represented for all agents by

$$\begin{bmatrix} \mathbf{P}_1^{(\tau)} \\ \mathbf{P}_2^{(\tau)} \\ \vdots \\ \mathbf{P}_K^{(\tau)} \end{bmatrix} = \begin{bmatrix} \psi_{11}I & \psi_{12}I & \dots \psi_{1K}I \\ \psi_{21}I & \psi_{22}I & \dots \psi_{2K}I \\ \vdots & \vdots & \vdots \\ \psi_{K1}I & \psi_{K2}I & \dots \psi_{KK}I \end{bmatrix} \begin{bmatrix} \mathbf{P}_1^{(\tau-1)} \\ \mathbf{P}_2^{(\tau-1)} \\ \vdots \\ \mathbf{P}_K^{(\tau-1)} \end{bmatrix} \tag{2}$$

where $I$ is the $N \times N$ identity matrix. Defining the group transition matrix at iteration $\tau$ to be the block matrix $\mathcal{P}^{(\tau)} = [\mathbf{P}_1^{(\tau)} | \mathbf{P}_2^{(\tau)} | \dots | \mathbf{P}_K^{(\tau)}]^\top$, where $\top$ denotes transpose, Eq. (2) is expressed compactly as $\mathcal{P}^{(\tau)} = [\Psi \otimes I]\mathcal{P}^{(\tau-1)}$, or equivalently

$$\mathcal{P}^{(\tau)} = [\Psi \otimes I]^{(\tau)}\mathcal{P}^{(0)}, \tag{3}$$

where $\otimes$ is the Kronecker product and $\mathcal{P}^{(0)} = [\mathbf{P}_1^{(0)} | \mathbf{P}_2^{(0)} | \dots | \mathbf{P}_K^{(0)}]^\top$ is the block matrix made up of the initial transition matrices participating in the algorithm. Convergence of (2) is assured under the condition that $\Psi \otimes I$ is a contraction, that is, $\|\Psi \otimes I\| \leq 1$. We exploit the following properties of the Kronecker product [26, 27]:

**Lemma: 1** If $A$ is an $m_A \times n_A$ matrix and $B$ is an $m_B \times n_B$ matrix, then, for any $p$-norm $\|\cdot\|$, $\|A \otimes B\| = \|A\|\|B\|$.

**Lemma: 2** If $A$ and $B$ are square matrices, then $(A \otimes B)^n = A^n \otimes B^n$.

Since $\|\Psi\|_\infty = 1$ and $\|I\|_\infty = 1$, applying Lemma 1 to $\Psi \otimes I$ shows that $\|\Psi \otimes I\| \leq 1$. Furthermore, applying Lemma 2 to the group transition matrix $\mathcal{P}^{(\tau)}$, given by (3), we obtain

$$\mathcal{P}^{(\tau)} = [\Psi^\tau \otimes I]\mathcal{P}^{(0)}. \tag{4}$$

As $\tau$ goes to infinity $\mathcal{P}^{(\tau)}$ approaches to the group consensus matrix $\mathcal{P}_c$ given by

$$\mathcal{P}_c = \begin{bmatrix} \mathbf{P}_c \\ \mathbf{P}_c \\ \vdots \\ \mathbf{P}_c \end{bmatrix} = \begin{bmatrix} \pi_{\Psi_1}I & \pi_{\Psi_2}I \dots \pi_{\Psi_K}I \\ \pi_{\Psi_1}I & \pi_{\Psi_2}I \dots \pi_{\Psi_K}I \\ \vdots & \vdots & \vdots \\ \pi_{\Psi_1}I & \pi_{\Psi_2}I \dots \pi_{\Psi_K}I \end{bmatrix} \begin{bmatrix} \mathbf{P}_1^{(0)} \\ \mathbf{P}_2^{(0)} \\ \vdots \\ \mathbf{P}_K^{(0)} \end{bmatrix} \tag{5}$$

where $\pi_\Psi = [\pi_{\Psi_1} \pi_{\Psi_2} \dots \pi_{\Psi_K}]$ is the limiting distribution of the stochastic matrix $\Psi$. Notice that, for the iterative procedure (4) to converge, the weights must be chosen so that $\Psi$ is ergodic [17]. It remains to specify how to estimate entries of $\Psi$ in the first

iteration of the algorithm. We note that since the right-hand side of (1) is a mixture of transition matrices [28, 29], the weights $\psi_{ij}$ of agent $i$ can be interpreted as the distribution of a latent variable.

In the next section we derive the likelihood function for our Markov chain mixture, which is used in the subsequent EM framework to estimate our consensus ratings. Once the optimal ratings are estimated and $\Psi$ is formed, its stationary distribution (denoted by $\pi_\Psi$) can be evaluated [30]. From that and the initially pooled transition matrices $\mathcal{P}^{(0)}$, an estimated consensus transition matrix $\mathbf{P}_c$ and corresponding stationary distribution $\pi_c$ can be determined.

## 3 Estimation of Optimal Consensus Ratings

To estimate concensus we must first determine a likelihood function for the linear combination of transition matrices involved in the consensus scheme for the $i$th agent in (2) for $\tau = 1$. For simplicity let us write

$$\mathbf{P}_i = \sum_{j=1}^{K} \psi_j \mathbf{P}_j. \tag{6}$$

The iteration indices have been omitted, and $\psi_{ij}$ has been replaced by $\psi_j$. In a group of agents it is assumed that each has observed its own state sequence and corresponding transition matrix. When a particular agent $i$ revises its own transition matrix, it invites the other agents to transmit theirs. Agent $i$ then adapts its own transition matrix based on the information received. We will explain how the consensus weights depend on the state sequences and the corresponding transition probabilities of each of the agents in the group. We follow an approach of Anderson and Goodman [31], but extended this to Markov chain mixtures.

It is assumed that the state sequence $\{X_t,\ 0 \le t \le T\}$ of agent $i$ is governed by $\mathbf{P}_i$ in (6). Transitions in this sequence are obtained as a mixture of sequences $X_t^k$, $k = 1, \ldots, K$, of all $K$ agents in the group. Since each agent $k$ is weighted by some value $\psi_j$, the probability of a particular transition at time $t$, from state $x_{i-1}$ to state $x_i$, can be modelled as the product of two probabilities: the probability of a transition from one state to another and the probability that the transition itself is caused by agent $k$. Consequently, the probability of the state sequence $x_0, \ldots, x_T$ is

$$\psi_{(x_0x_1)_k} p_{(x_0x_1)_k} \psi_{(x_1x_2)_k} p_{(x_1x_2)_k} \cdots \psi_{(x_{T-1}x_T)_k} p_{(x_{T-1}x_T)_k}. \tag{7}$$

Expression (7) can be further extended by introducing a random process $(Z_t)$ to model random selection of the source $k$ of a particular transition from $x_{i-1}$ to $x_i$. Since this transition at time $t$ can come from only one source, an indicator $I_{\{Z_t=k\}}$ of this source is introduced. In that case, the weight $\psi_{(x_{i-1}x_i)_k}$ can be interpreted as the probability that a particular transition probability $p_{(x_{i-1}x_i)_k}$ comes from agent $k$,

denoted as $P(Z_t = k)$. Thus, for each transition from $x_{i-1}$ to $x_i$, expression (7) is modified to obtain

$$\prod_{k=1}^{K}\{\psi_{(x_0x_1)_k}\, p_{(x_0x_1)_k}\}^{I_{\{Z_1=k\}}} \prod_{k=1}^{K}\{\psi_{(x_1x_2)_k}\, p_{(x_1x_2)_k}\}^{I_{\{Z_2=k\}}} \cdots \prod_{k=1}^{K}\{\psi_{(x_{T-1}x_T)_k}\, p_{(x_{T-1}x_T)_k}\}^{I_{\{Z_T=k\}}}. \tag{8}$$

The next step towards calculating (7) requires counting the number of transitions, from $x_{i-1}$ to $x_i$ for agent $k$ until time $t$ on the entire sequence, as follows:

$$\prod_{x_0x_1}\prod_{k=1}^{K}\{\psi_{(x_0x_1)_k}\, p_{(x_0x_1)_k}\}^{I_{\{Z_1=k\}}N_1(x_0x_1)_k} \prod_{x_1x_2}\prod_{k=1}^{K}\{\psi_{(x_1x_2)_k}\, p_{(x_1x_2)_k}\}^{I_{\{Z_2=k\}}N_2(x_1x_2)_k} \cdots \prod_{x_{T-1}x_T}\prod_{k=1}^{K}\{\psi_{(x_{T-1}x_T)_k}\, p_{(x_{T-1}x_T)_k}\}^{I_{\{Z_T=k\}}N_T(x_{T-1}x_T)_k}, \tag{9}$$

or, more compactly,

$$\prod_{t=1}^{T}\prod_{i,j=1}^{N}\prod_{k=1}^{K}\{P(Z_t = k)P(X_t = j|X_{t-1} = i, Z_t = k)\}^{I_{\{Z_t=k\}}N_t(X_{t-1}=i, X_t=j)_k} \tag{10}$$

To simplify notation, $P(Z_t = k)$ will be denoted as $\psi_k$, the transition probability $P(X_t = j|X_{t-1} = i, Z_t = k)$ will be given in the shortened form $P_t(i, j)_k$ and the number of transitions in a state sequence $X_t^k$ by time $t$ for a particular agent $k$ by $N_t(i, j)_k$: (10) becomes

$$\prod_{t=1}^{T}\prod_{i,j=1}^{N}\prod_{k=1}^{K}\{\psi_k P_t(i, j)_k\}^{I_{\{Z_t=k\}}N_t(i,j)_k}. \tag{11}$$

It is apparent from (11) that the random variable $Z_t$ is not directly observable. However, this incomplete-data problem can be converted to a complete-data problem; if the problem is extended to find the likelihood of the sequence $\{(X_t, Z_t), 0 \le t \le T\}$ instead, it opens up the possibility of using the EM framework [32].

As previously discussed, expression (11) is a likelihood function of the complete-data vector whose logarithm is given by

$$\log L(\boldsymbol{\Psi}; \boldsymbol{X}, \boldsymbol{Z}) = \sum_{t=1}^{T}\sum_{i,j=1}^{N}\sum_{k=1}^{K} I_{\{Z_t=k\}}N_t(i, j)_k\{\log P_t(i, j)_k + \log \psi_k\}. \tag{12}$$

The EM algorithm is a two-step iterative procedure. In the first step, called the E-step, the $Q$ function is calculated, which is the mathematical expectation of (12) given observations $\{X_t, 0 \le t \le T\}$: $Q(\psi_k|\psi_k^{(\iota)}) = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\Psi}}\{\log L(\boldsymbol{\Psi}; \boldsymbol{X}, \boldsymbol{Z})\}$, where $\psi_k^{(\iota)}$ is a set of parameter estimated in previous iteration $\iota$. The $Q$ function evaluation is reduced to computing the mathematical expectations of indicator functions, because the transition probabilities $P_t(i, j)_k$, counts $N_t(i, j)_k$ and initial mixing proportions $\psi_k^{(\iota)}$ are known in advance. By Bayes' Theorem

$$P(A|B \cap C) = \frac{P(B|A \cap C)P(A|C)}{P(B|C)}. \tag{13}$$

Furthermore, assuming that $X_t$ depends only on $X_{t-1}$ and $I_{\{Z_t=k\}}$, as well as presuming that $I_{\{Z_t=k\}}$ and $X_{t-1}$ are independent, the mathematical expectation of the indicator function is given as follows:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\Psi}}\{I_{\{Z_t=k\}}|\mathbf{X}; \boldsymbol{\Psi}\} &= \frac{P(X_t=j|X_{t-1}=i, I_{\{Z_t=k\}}=1)P(I_{\{Z_t=k\}}=1)}{P(X_t=j|X_{t-1}=j)} \\ &= \frac{P(X_t=j|X_{t-1}=i, I_{\{Z_t=k\}}=1)P(I_{\{Z_t=k\}}=1)}{\sum_{h=1}^{K} P(X_t=j|X_{t-1}=i, I_{\{Z_t=h\}}=1)P(I_{\{Z_t=h\}}=1)} \\ &= \frac{\psi_k^{(\iota)} P_t(i, j)_k}{\sum_{h=1}^{K} \psi_h^{(\iota)} P_t(i, j)_h} = \varphi_t^{(\iota)}(i, j)_k. \end{aligned} \tag{14}$$

Finally an expression for the $Q$ function is given by

$$Q(\psi_k|\psi_k^{(\iota)}) = \sum_{t=1}^{T}\sum_{i,j=1}^{N}\sum_{k=1}^{K} \varphi_t^{(\iota)}(i, j)_k N_t(i, j)_k\{\log P_t(i, j)_k + \log \psi_k\}. \tag{15}$$

In the second step of the EM algorithm, called the M-step, previously assumed parameters are optimized based on the expectation of the log likelihood: $\psi_k^{(\iota+1)} = \arg\max_{\psi_k} Q(\psi_k|\psi_k^{(\iota)})$. Introducing a Lagrange multiplier $\mu$ into (15) for the constraint $\sum_{k=1}^{K} \psi_k = 1$, we obtain

$$\begin{aligned} Q(\psi_k|\psi_k^{(\iota)}) = &\sum_{t=1}^{T}\sum_{i,j=1}^{N}\sum_{k=1}^{K} \varphi_t^{(\iota)}(i, j)_k N_t(i, j)_k\{\log P_t(i, j)_k + \log \psi_k\} \\ &-\mu\left(\sum_{k=1}^{K} \psi_k - 1\right). \end{aligned} \tag{16}$$

After taking the derivative of $Q$ with respect to $\psi_k$ we get

$$\frac{\partial Q(\psi_k|\psi_k^{(\iota)})}{\partial \psi_k} = \frac{\sum_{t=1}^{T}\sum_{i,j=1}^{N} \varphi_t^{(\iota)}(i, j)_k N_t(i, j)_k}{\psi_k} - \mu = 0. \tag{17}$$

Rearranging (17) and using the constraint we obtain

$$\mu = \sum_{h=1}^{K} \sum_{t=1}^{T} \sum_{i,j=1}^{N} \varphi_t^{(\iota)}(i,j)_h N_t(i,j)_h.$$

Finally, the updated equation for $\psi_k$ is given by

$$\psi_k = \frac{\sum_{t=1}^{T} \sum_{i,j=1}^{N} \varphi_t^{(\iota)}(i,j)_k N_t(i,j)_k}{\sum_{h=1}^{K} \sum_{t=1}^{T} \sum_{i,j=1}^{N} \varphi_t^{(\iota)}(i,j)_h N_t(i,j)_h}. \tag{18}$$

By altering the E-step and M-step, in each iteration $\iota$, the function $Q(\psi_k|\psi_k^{(\iota)})$ is calculated and the parameters are optimized. From an implementation point of view there are two ways to halt the procedure. The first is when the difference in the value of $Q$ is below some threshold $\Theta$ assumed in advance, that is, $Q(\psi_k|\psi_k^{(\iota)}) - Q(\psi_k|\psi_k^{(\iota-1)}) \leq \Theta$. The second is to specify in advance the total number of iterations $\Upsilon$ and stop when $\iota \geq \Upsilon$.

Taking into consideration the E-step and M-step used to estimate optimal ratings of the stochastic matrix $\Psi$, it is apparent that there are two specific kinds of information each agent in the group requires. Firstly, by (14) it follows that each agent in the group has to know the other agents' transition probabilities. In other words, information on the models perceived by the group members are supposed to be shared among the group. Secondly, even more interesting conclusions can be drawn from (18). In order to rate other group members, agent $i$ relies on information regarding a number of transitions, $N_t^k, k = 1, \ldots, K$, of pooled state sequences $X_t^k$. As we will see shortly, a major problem in applying our algorithm is related to the inability to observe these state sequences directly.

Before proceeding, recall briefly the notation introduced of this section. The index $i$, denoting the agent that revises its distribution, is omitted: $X_t^k$ is shorthand for the state sequence $X_t^{ik}$ that models the influence of agent $k$ on agent $i$, and $N_t^k$ is short for $N_t^{ik}$, the number of transitions. Which notation we will use depends on the context.

## 4 Extension to Unobservable Case

It is clearly unrealistic to assume that each state transition is an observable discrete event, because of measurement noise or because the observed signal is generated by one of multiple sources randomly switched by an unobservable Markov chain, as depicted in Fig. 2. It is therefore desirable to extend our algorithm to the case where the states are hidden.

In order to adapt our algorithm to the context of Hidden Markov Models (HMMs) [33, 34], it will be necessary to establish a link with the solutions of the three essential problems connected with HMMs described by Rabiner [33]. The basic elements of

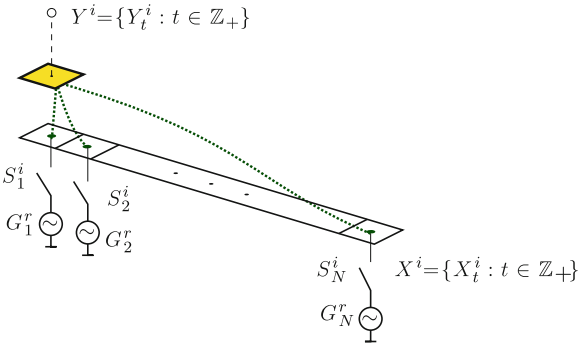



**Fig. 2** Local hidden Markov model

a HMM are the number of states $N$; the number of distinct observation symbols (discrete observations) $O = \{o_1, o_2, \ldots, o_M\}$, or the number of mixture components (continuous observations) $M$; and the state transition probabilities $\boldsymbol{A} = [\alpha_{ij}]_{N\times N}$, where

$$\alpha_{ij} \equiv P(X_t = j | X_{t-1} = i), \qquad 1 \le i, j \le N. \tag{19}$$

The next element is an observation probability defined as a matrix $\boldsymbol{B} = [\beta_j(y_t)]_{N\times 1}$, where $\beta_j(y_t)$ is a probability of a particular observation. For discrete observations,

$$\beta_j(y_t) \equiv P(Y_t = o_m | X_t = j), \qquad 1 \le j \le N,\ 1 \le m \le M. \tag{20}$$

The final element is the initial state distribution $\boldsymbol{\Pi} = [\pi_k]_{N\times 1}$. We use the following compact notation:

$$\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Pi}), \tag{21}$$

where $N$ and $M$ are defined implicitly.

The three basic problems identified by Rabiner [33] are as follows.

Problem 1 : The efficient computation of the probability $P(Y|\lambda)$ of the observed sequence $Y = Y_1, Y_2, \ldots, Y_T$, given the model $\lambda$.
Problem 2 : Estimation of the hidden state sequence $X = X_1, X_2, \ldots, X_T$, given the observation sequence $Y = Y_1, Y_2, \ldots, Y_T$ and the model $\lambda$.
Problem 3 : Estimation of the parameters of model $\lambda$ that maximize $P(Y|\lambda)$.

To solve Problem 1 the forward-backward procedure [33] is applied, which is a recursive method for efficient calculation of the probability of the observed sequence given the model $\lambda$. To maximize the probability of a hidden state sequence for given observations, the Viterbi algorithm [35] is used to address Problem 2. Finally, Problem 3 is addressed applying the Baum-Welch algorithm [36].

The connection with our algorithm will now be explained. As given previously, each agent $i$ in the group takes observations $Y_t^i = \{y_1^i, y_2^i, \ldots, y_T^i\}$, independently

allowing the application of a consensus algorithm. In addition, it is assumed that each of these observations is generated by a Markov switching model. In other words, a particular observation at time $t$ comes from one of the underlying signal generators $G_i^r$, $r = 1, \ldots, R$, $i = 1, \ldots, N$, where $R$ is a number of working regimes and $N$ is the number of signal generators. We do not assume anything about these hypothetical generators. Indeed we are only interested in the sequence of their activations/deactivations, modelled by an associated Markov chain $X^i = \{X_t^i,\ 0 \le t \le T\}$ defined on a state space $S = \{1, 2, \ldots, N\}$ with a corresponding transition probability matrix $\boldsymbol{A}_r$. It is very important to note that that an unobservable Markov chain transition matrix $\boldsymbol{A}_r$ will be used to form the initial pool of Markov chain transition matrices in (5), that is, if agent $i$ is in the working regime $r$ then $\boldsymbol{A}_r \equiv P_i^{(0)}$. Note that the number of generators $N$ defines the size of $S$ and that number is a same for all working regimes. Consequently, all transition probability matrices are the same size.

In the context of the algorithm described in Sects. 2 and 3, it is necessary for each operating mode $r$ to determine a transition probability matrix $\boldsymbol{A}_r$ of an unobservable Markov chain $X^i$ of corresponding hypothetical switches $S^i = \{S_1^i, \ldots, S_N^i\}$, $i = 1, \ldots, |V|$, and similarly for $\boldsymbol{B}_r$ and the initial state distribution $\boldsymbol{\Pi}_r$. With this in mind, it is apparent that the Baum-Welch algorithm provides a means of estimating different working regimes and designing a bank of competing models $\Lambda = \{\lambda(\Theta_1) \ldots \lambda(\Theta_R)\}$. These models describe possible working conditions of each subsystem, where $\lambda(\Theta_r)$ is model (21) and $\Theta_r = (\boldsymbol{A}_r, \boldsymbol{B}_r, \boldsymbol{\Pi}_r)$ is the vector of the parameters of model $r$. The currently active model of a particular agent $i$ is determined by the forward-backward procedure, selecting the most likely model $\lambda_i = \lambda(\Theta_r)$.

Finally, as per the proposed consensus scheme, to estimate a rating $\psi_{ik}$ using (18), agent $i$ needs to estimate the number of transitions, $N_t^k$, in the underlying state sequence $X_t^k$. First, $X_t^{ik}$ must include in it information about a model of the agent $k$, whose rating is being assessed. Second, local observations $Y_t^i$ of the agent $i$ are used to revise its distribution. Thus, the Viterbi algorithm logically connects these two aspects, all with the aim of estimating the state sequences $X_t^{ik}$. Of course after $X_t^{ik}$ is estimated, $N_t^k$ is easily determined. So, to estimate the optimal ratings $\psi_{ik}$, the Viterbi algorithm is applied to estimate the hidden sequences $X_t^{ik} = \{x_1^{ik}, x_2^{ik}, \ldots, x_T^{ik}\}$ from local models $\lambda_k$ of the all agents in the group and local observations $Y_t^i = \{y_1^i, y_2^i, \ldots, y_T^i\}$.

To help explain an application of the algorithm to real-time systems, a sequence diagram is given in Fig. 3, the steps summarized as follows:

Step 0: At time $t$ each agent $i$ from the group of $K$ agents has collected $T$ observations $Y_t^i = \{y_1^i, y_2^i, \ldots, y_T^i\}$ within the sliding window. Additionally every agent in the group possesses a bank of competing models $\Lambda = \{\lambda(\Theta_1) \ldots \lambda(\Theta_R)\}$;

Step 1: By applying the forward-backward recursive procedure to a bank $\Lambda$, given a set of observations $Y_t^i$, a currently active model $\lambda_i = \lambda(\Theta_r)$ of agent $i$, with a set of parameters $\Theta_r = (A_r, B_r, \Pi_r)$, is determined. Note that initial tran-

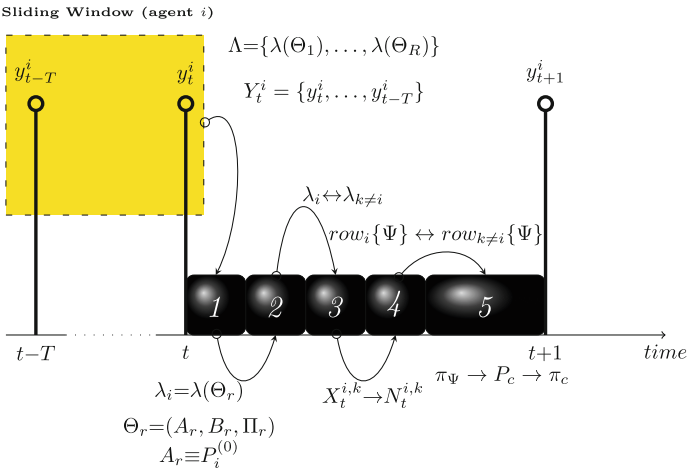



**Fig. 3** An algorithm state transition diagram

sition matrices $P_i^{(0)}$ of consensus algorithm and model transition matrices $A_r$ the same ($P_i^{(0)}\equiv A_r$);

Step 2: All agents in the group exchange, over a computer network, currently active models $\lambda_i \leftrightarrow \lambda_{k\neq i}$, $k = 1,\ldots,K$;

Step 3: By the means of the Viterbi algorithm every agent $i$ estimates the unobservable sequences $X^{ik}$ and corresponding transitions $N_t^{ik}$ from $Y_t^i$ and $\lambda_k$, $k = 1,\ldots,K$. Using the EM algorithm an agent $i$ estimates a row $i$ of a stochastic matrix $\Psi$, $\text{row}_i\{\Psi\} = (\psi_{i1},\ldots,\psi_{iK})$;

Step 4: The group of agents exchange over the computer network rows of the stochastic matrix $\Psi$, $\text{row}_i\{\Psi\} \leftrightarrow \text{row}_{k\neq i}\{\Psi\}$, allowing each agent to form the matrix $\Psi$;

Step 5: Once these transactions are complete, each agent computes a stationary distribution [30] $\pi_\Psi = (\pi_{\psi_1},\ldots,\pi_{\psi_K})$ required to estimate a consensus transition matrix $\mathbf{P}_c$, from which the stationary distribution $\pi_c$ is computed. After a stationary distribution $\pi_c$ is estimated, a fault detection scheme is applied, as explained in the next section.

## 5 Fault Detection Scheme

As explained in Sect. 1, an important aspect of our fault detection scheme is a distance measurement between the stationary distribution $\pi_c$ of the consensus transition matrix and the stationary distribution $\pi_{P^{(0)}}$ of the agent's initial transition matrix. To measure the distance between the two stationary distributions, $\pi_1$ and $\pi_2$, we will use the $L_2$-norm:

**Table 1** The fault table (Dictionary)

| $\delta(\pi_{A_i}, \pi_{A_j})$ | $\pi_{\mathbf{A}_1}$ | $\pi_{\mathbf{A}_2}$ | … | $\pi_{\mathbf{A}_R}$ |
|---|---|---|---|---|
| $\pi_{\mathbf{A}_1}$ | 0 | $\delta_{12}$ | … | $\delta_{1R}$ |
| $\pi_{\mathbf{A}_2}$ | $\delta_{12}$ | 0 | … | $\delta_{2R}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $\pi_{\mathbf{A}_R}$ | $\delta_{1R}$ | $\delta_{2R}$ | … | 0 |

$$\delta(\pi_1, \pi_2) = \sum_{i=1}^{N} \left( \sum_{j=1}^{i} \pi_1(j) - \sum_{j=1}^{i} \pi_2(j) \right)^2, \tag{22}$$

where $N$ is a number of discrete states. Once evaluated, these distances are stored to form a fault table (or dictionary). Because the table is symmetric, the total number of different values, as a function of the number of operation modes $n$, is $\frac{1}{2}n(n-1)$. As noted earlier, every fault diagnosis scheme consists of a series of tasks, the first of which, *fault detection*, is implemented as Step 5 (above). Here we measure the distance between the stationary distribution of the consensus transition matrix and the stationary distribution of the agent's local transition matrix (model). This step allows us to identify deviations from normal operation.

**Algorithm 1** Fault Diagnosis - agent $i$

**Require:** $\pi_c, \pi_{P_1^{(0)}}, \ldots, \pi_{P_K^{(0)}}$, The Fault Table
1: **if** $\delta(\pi_c, \pi_{P_i^{(0)}}) > 0$ **then**
2: **for all** $j \in K \backslash \{i\}$ **do**
3: **if** $\delta(\pi_{P_i^{(0)}}, \pi_{P_j^{(0)}}) > 0$ **then**
4: Search the fault table to identify the model $r$ associated with agent $j$, $F_{vec}(j) \leftarrow r$
5: **end if**
6: **end for**
7: **return** Fault Vector - $F_{vec}$
8: **end if**

Once a fault is detected, the second task, *fault identification*, begins. Multiple faults in monitored subsystems can be identified, because all agents have the fault table that summarizes information about different working regimes and each has information about the currently active modes of all other agents in the group. It has become apparent that a potential pitfall of this approach, which is particularly evident when the number of models is large, lies in the need to retrieve the fault table. Consideration of this issue will be part of future research (Table 1).

A realization of the first and second tasks is summarized in **Algorithm 1**, which is executed for all agents $i$. In the third fundamental task, *fault analysis*, we analyse

**Table 2** Normal working regime $\lambda_1$

| $\lambda_1$ | $\boldsymbol{A}_1$ | | $\boldsymbol{B}_1$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $s_1$ | $s_2$ | $o_1$ | $o_2$ | $o_3$ | $o_4$ | $o_5$ | $o_6$ | $o_7$ |
| $s_1$ | 0.68 | 0.32 | 0.25 | 0.7 | 0.025 | 0.0063 | 0.0063 | 0.0063 | 0.0063 |
| $s_2$ | 0.2 | 0.8 | 0.7 | 0.25 | 0.025 | 0.0063 | 0.0063 | 0.0063 | 0.0063 |

**Table 3** Faulty working regime $\lambda_2$

| $\lambda_2$ | $\boldsymbol{A}_2$ | | $\boldsymbol{B}_2$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $s_1$ | $s_2$ | $o_1$ | $o_2$ | $o_3$ | $o_4$ | $o_5$ | $o_6$ | $o_7$ |
| $s_1$ | 0.4 | 0.6 | 0.0125 | 0.05 | 0.1 | 0.65 | 0.15 | 0.025 | 0.0125 |
| $s_2$ | 0.75 | 0.25 | 0.0063 | 0.025 | 0.65 | 0.15 | 0.15 | 0.0125 | 0.0063 |

**Table 4** Faulty working regime $\lambda_3$

| $\lambda_3$ | $\boldsymbol{A}_3$ | | $\boldsymbol{B}_3$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $s_1$ | $s_2$ | $o_1$ | $o_2$ | $o_3$ | $o_4$ | $o_5$ | $o_6$ | $o_7$ |
| $s_1$ | 0.3 | 0.7 | 0.0125 | 0.0125 | 0.0125 | 0.0125 | 0.0125 | 0.25 | 0.7 |
| $s_2$ | 0.71 | 0.29 | 0.0125 | 0.0125 | 0.0125 | 0.0125 | 0.05 | 0.7 | 0.2 |

the time-varying characteristics of the fault(s), which are connected to model observations.

## 6 Simulation Results

We illustrate the method using a hypothetical system composed of three interconnected subsystems ($K = 3$). The corresponding agents are labelled 1, 2 and 3. We supposed that there are three possible subsystem working regimes, each modelled by a HMM with two states ($N = 2$), an unobservable Markov chain and a discrete observations set $O = \{1, 2, \ldots, 7\}$ ($M = 7$).

The HMM labelled $\lambda_1$ models normal operation, while those labelled $\lambda_2$ and $\lambda_3$ represent faults. The parameters of these models are given in Tables 2, 3 and 4, respectively.

Together, they form a bank of competing models $\Lambda = \{\lambda(\Theta_1), \lambda(\Theta_1), \lambda(\Theta_3)\}$. The dynamics are rather simple and emulate dynamical changes in signal amplitude. Notwithstanding this, these simplified models serve to illustrate collective decision-making in a group of agents. Figure 4 depicts agents' local observations. Agents collect observations by means of a sliding window that contains, at time $t$, the last $T$ samples of the monitored signal. For each agent $i$ in the group, a fault diagnosis

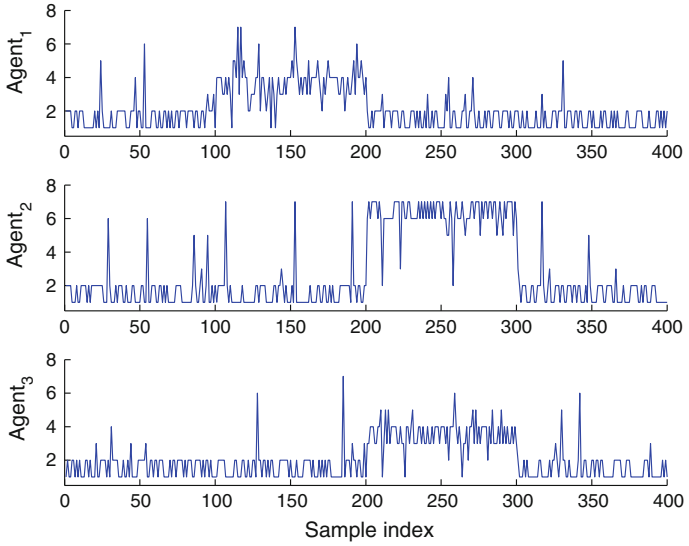



**Fig. 4** Agents' observations

scheme is applied to local observations $Y_t^i = \{y_1^i, y_2^i, ..., y_T^i\}$, $i = 1, 2, 3$ (in our case, the window length is $T = 7$).

We analyse the test case depicted in Fig. 4. The transition of agent 1 from mode $\lambda(\Theta_1)$ to $\lambda(\Theta_2)$ occurs at $t = 101$. At $t = 201$, agent 1 returns to mode $\lambda(\Theta_1)$, while agents 2 and 3 transition to modes $\lambda(\Theta_3)$ and $\lambda(\Theta_2)$, respectively. At $t = 301$, all agents are once again in mode $\lambda(\Theta_1)$.

In Step 1 of the state transition diagram (Fig. 3), the forward-backward recursive procedure applies for estimating a currently active model using the bank of models $\Lambda$ and the agent's local observations $Y_t^i$. The changes of model indexes are depicted in Fig. 5. The dashed lines represent true index values, while the solid lines are estimates obtained from the forward-backward procedure. Notice that there is some delay in identifying these changes. Agents' decision delay $\Delta$, as a function of a sliding window length, is given in Table 5.

Next, in Step 2, the agents mutually exchange information on current working regime models $\lambda_k = \lambda(\Theta_r)$, meaning that agent $k$ has identified a model $r$.

In Step 3 the Viterbi algorithm is used to estimate the state sequences $X_t^{ik}$, $i, k = 1, 2, 3$, given observations $Y_t^i$ and model parameters $\lambda_k = \lambda(\Theta_r)$ in the group . Once all agents in the group estimate their own set of sequences $X_t^{ik}$ and count number of transitions $N_t^{ik}$ they will estimate rows, $\text{row}_i\{\Psi\}$ $i = 1, 2, 3$, of the consensus matrix $\Psi$ using the EM procedure. Numerical examples of optimal ratings estimation of $\Psi$ will be given for three different cases: $t = 50$, 150 and 250. Simulation results are presented in Figs. 6, 7 and 8, which assume that parameters of the EM algorithm are initialized randomly. These figures demonstrate how agents' rates evolve with the EM iterations. As described before, each agent in the group

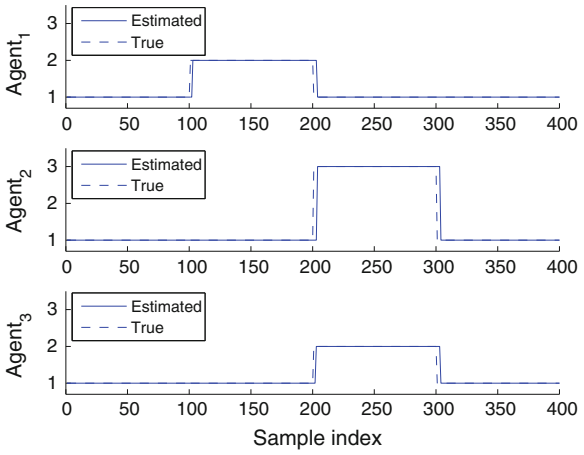



**Fig. 5** Agents' model indexes

**Table 5** Decision delay $\Delta$ as a function of a sliding window length $T$

| $T$ | 5 | 7 | 9 | 11 |
|---|---|---|---|---|
| $\Delta$ | 2 | 3 | 4 | 6 |

rates itself and other agents. Agent 1's rates, $\psi_{11}$, $\psi_{21}$ and $\psi_{31}$, are a measure of the influence of agent 1 working regime model on the group. These values form the first column of the stochastic matrix $\Psi \equiv \Psi^{(0)}$. As we showed in Sect. 2, the first column of $\Psi^{(\tau)}$ in the limit as $\tau \to \infty$ will be the stationary value $\pi_{\psi_1}$. Similarly, for agents 2 and 3, their estimated rates, $\psi_{12}$, $\psi_{22}$, $\psi_{32}$, and $\psi_{13}$, $\psi_{23}$, $\psi_{33}$, converge to stationary values $\pi_{\psi_2}$ and $\pi_{\psi_3}$, respectively. Stochastic matrices $\Psi$, for $t = 150$ and 250, and their stationary distributions, are summarized in Tables 6 and 7.

Another important conclusion can be drawn from Figs. 6, 7 and 8. It concerns the stopping criteria of the EM algorithm, which combines a fixed number of iterations, $\Upsilon = 10$, with a threshold value, $\Theta = 10^{-4}$. Selection of these parameters is an essential part of the practical implementation of our algorithm to provide real-time response.

In Step 4, previously estimated rows of the matrix $\Psi$, $\text{row}_i\{\Psi\} = (\psi_{i1}, \ldots, \psi_{iK})$ $i = 1, \ldots, K$, are exchanged among group members to form a stochastic matrix $\Psi$.

Finally, in Step 5, once each agent in the group has its own stochastic matrix $\Psi$, with stationary distribution $\pi_\psi$, a fault diagnosis scheme is launched. As we have already seen $\pi_\psi$ allows the computation of a consensus transition matrix $\mathbf{P}_c$ as a weighted sum of the unobservable Markov chains of all agents in the group; see (5). Computing the stationary distribution $\pi_c$ of a consensus transition matrix $\mathbf{P}_c$ is the starting point of **Algorithm 1**. A practical implementation of line 1 of the algorithm is slightly modified in that the condition $\delta(\pi_c, \pi_{P_i^{(0)}}) > 0$ is tested as $\log \delta(\pi_c, \pi_{P_i^{(0)}}) < \kappa$, for suitable $\kappa$, assumed here to be $\kappa = -20$. In Fig. 9 the

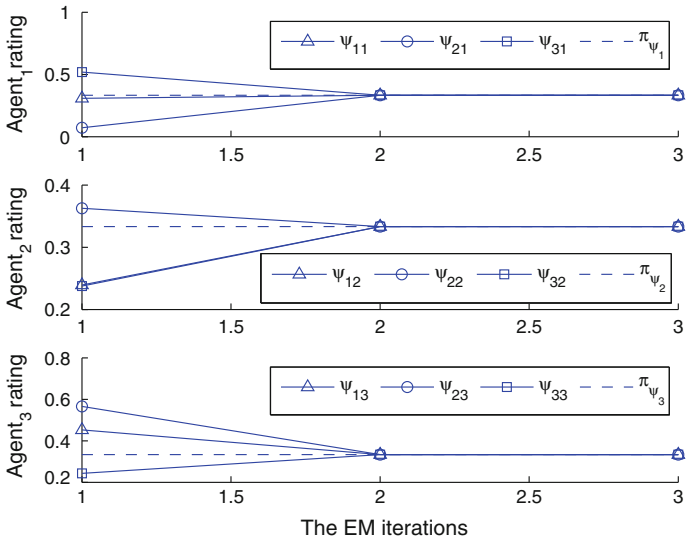



**Fig. 6** Consensus matrix estimation at $t = 50$

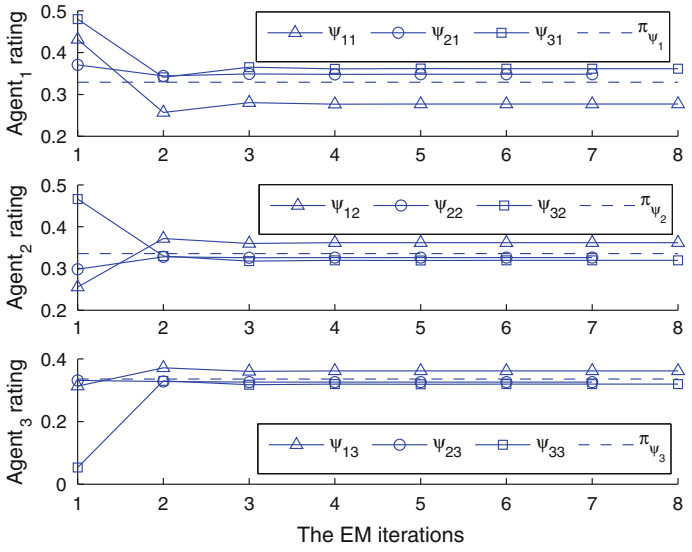



**Fig. 7** Consensus matrix estimation at $t = 150$

logarithm of the distance measure of all agents the group is given, showing how it changes over time.

In view of Fig. 9 notice first that in the interval $t \in [103, 303]$ all three agents assessed the value $\log \delta(\pi_c, \pi_{P_i^{(0)}})$, $i = 1, 2, 3$, as being less than the threshold

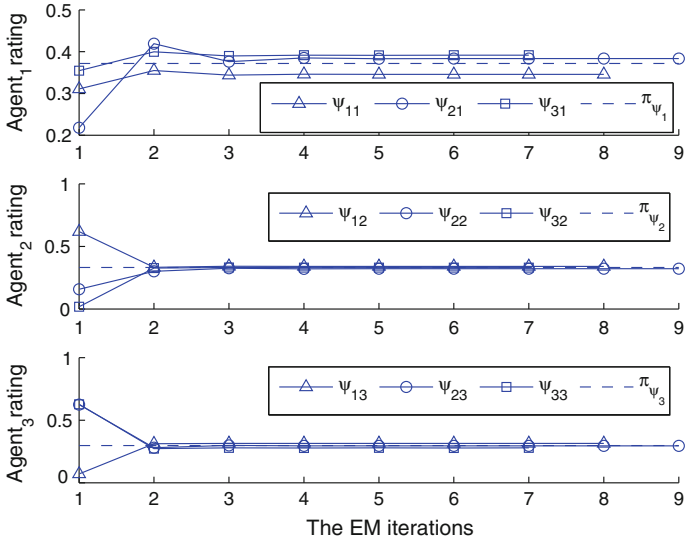



**Fig. 8** Consensus matrix estimation at $t = 250$

**Table 6** Initial consensus matrix for $t = 150$

| | | | |
|---|---|---|---|
| $\Psi^{(0)}_{t=150}$ | 0.2772 | 0.3614 | 0.3614 |
| | 0.3482 | 0.3259 | 0.3259 |
| | 0.3618 | 0.3191 | 0.3191 |
| $\pi_{\Psi_{t=150}}$ | 0.3294 | 0.3353 | 0.3353 |

**Table 7** Initial consensus matrix for $t = 250$

| | | | |
|---|---|---|---|
| $\Psi^{(0)}_{t=250}$ | 0.3457 | 0.3403 | 0.3140 |
| | 0.3837 | 0.3213 | 0.2950 |
| | 0.3916 | 0.3302 | 0.2782 |
| $\pi_{\Psi_{t=250}}$ | 0.3719 | 0.3310 | 0.2971 |

$\kappa$, indicating faults in the system. Secondly, it is evident that agent 2 has slightly higher fluctuations on the interval $t = [204, 303]$ than other agents. Before we give a detailed explanation of the reasons for this behaviour, let us look at both the value of the stationary distributions $\pi_{\boldsymbol{A}_i}$ of models $\lambda(\Theta_i)$, $i = 1, 2, 3$ (Table 8) and the consensus matrices $\pi_c$ at times $t = 50, 150, 250, 350$ (Table 9). We see that that $\pi_{\boldsymbol{A}_3}$ and $\pi_c^{t=250}$ have similar values. From (5) it follows that consensus transition matrix cannot be equal to any particular initially pooled transition matrix, except in the trivial case that all transition matrices are the same. However, in any non-trivial case, a problem of similarity between unobservable Markov chains arises, indicating the importance of measuring it in the training phase. As proposed by Rabiner [33] the concept of model distance can be used for this purpose. This problem will be the subject of future research.

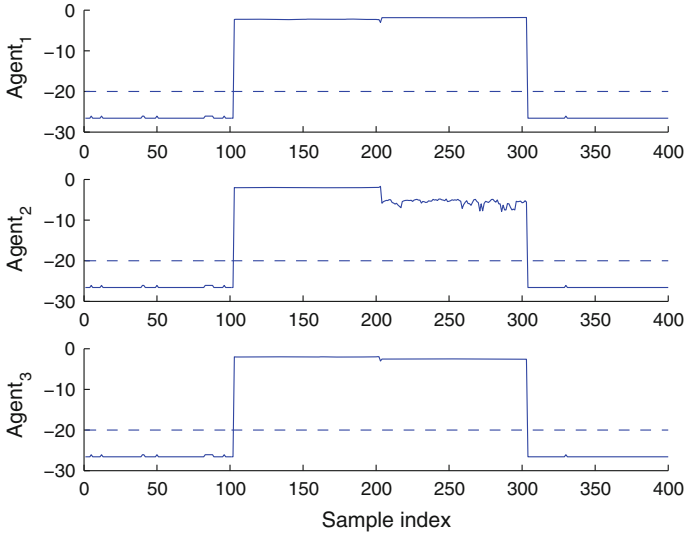



**Fig. 9** Distribution distance

**Table 8** Model stationary distributions

| | | |
|---|---|---|
| $\pi_{A_1}$ | 0.3846 | 0.6154 |
| $\pi_{A_2}$ | 0.5556 | 0.4444 |
| $\pi_{A_3}$ | 0.5035 | 0.4965 |

**Table 9** Consensus stationary distributions at times $t = 50, 150, 250, 350$

| | | |
|---|---|---|
| $\pi_c^{t=50}$ | 0.3846 | 0.6154 |
| $\pi_c^{t=150}$ | 0.4804 | 0.5196 |
| $\pi_c^{t=250}$ | 0.5015 | 0.4985 |
| $\pi_c^{t=350}$ | 0.3846 | 0.6154 |

Figure 10 illustrates how the group members perceive group behaviour and how they achieve behavioural consensus; exchanging model parameters among group members and by applying local observations to these models leads to a common perception of group behaviour.

## 7 Discussion

We have proposed a fault detection scheme for distributed systems in which subsystems are represented by agents. Operating modes of subsystems are modelled by Markov chains. The agents form groups whose common (consensus) transition matrix is estimated. Change in consensus within the group is monitored and, once a change is detected, the distances between the stationary distributions of operating modes are estimated in order to identify the new condition of the system. Future

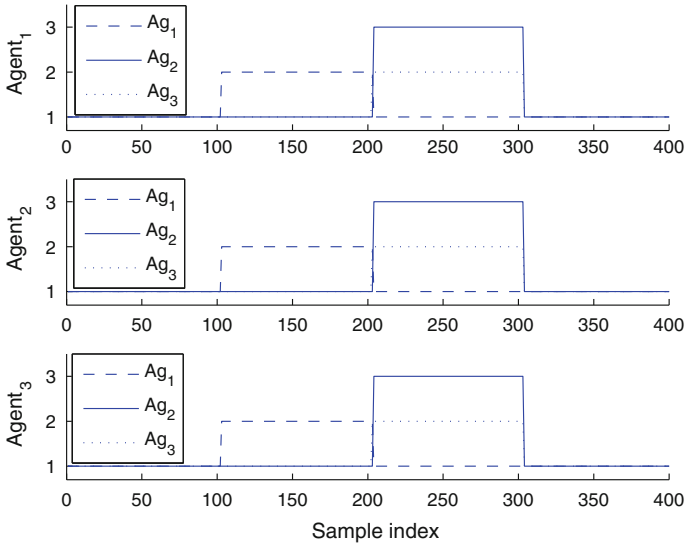



**Fig. 10** Group decision

work will include the practical implementation of our algorithm to fault diagnosis in power systems. To be fully applicable it will be necessary to extend our approach to accommodate continuous observation schemes.

**Acknowledgments** This work was supported by the Australian Research Council Centre of Excellence for Mathematics and Statistics of Complex Systems (MASCOS). We are grateful to Ross McVinish for helpful comments.

## References

1. Leitch RD (1995) Reliability analysis for engineers : an introduction. Oxford University Press, New York
2. Patton R, Frank P, Clark R (eds) (2000) Issues of fault diagnosis for dynamic systems. Springer, London
3. Ding SX (2008) Model-based fault diagnosis techniques: design schemes, algorithms, and tools. Springer-Verlag Berlin Heidelberg, Berlin
4. Russell LHCE, Braatz RD (2000) Data-driven methods for fault detection and diagnosis in chemical processes. Springer, New York
5. Gertler J (1998) Fault detection and diagnosis in engineering systems. Marcel Dekker Inc, New York
6. Simani S, Fantuzzi C, Patton R (2003) Model-based fault diagnosis in dynamic systems using identification techniques. Springer, London
7. Isermann R (1995) "Model base fault detection and diagnosis methods", In: Proceedings of the American control conference, vol 3. pp 1605–1609 June 1995
8. Isermann R (1984) Process fault detection based on modeling and estimation methodsa survey, Automatica 20(4):387–404

9. Isermann R (2005) Model-based fault-detection and diagnosis—status and applications, Ann Rev Control 29(1):71–85
10. Siljak D (1991) Decentralized control of complex systems. Academic Press, Boston
11. Fabre E, Pigourier V (2002) "Monitoring distributed systems with distributed algorithms", in Decision and Control, 2002. Proceedings of the 41st IEEE conference on, vol 1. pp 411–416
12. Setola R, De Porcellinis S (2009) "Complex networks and critical infrastructures", in Modelling, Estimation and Control of Networked Complex Systems, ser. Understanding Complex Systems. In: Chiuso A, Fortuna L, Frasca M, Rizzo A, Schenato L, Zampieri S (eds) vol 50. Springer, Berlin, pp 91–106
13. Kalman RE, Bucy RS (1961) New results in linear filtering and prediction theory. J Basic Eng 83(1):95–108
14. Norris JR (1997) Markov chains. Cambridge University Press, Cambridge
15. Stone M (1961) The opinion pool. Ann Math Stat 32(4):1339–1342
16. DeGroot MH (1974) Reaching a consensus. J Am Stat Assoc 69(345):118–121
17. Chatterjee S, Seneta E (1977) Towards consensus: Some convergence theorems on repeated averaging. J Appl Probab 14(1):89–97
18. Franco E, Olfati-Saber R, Parisini T, Polycarpou M (2006) Distributed fault diagnosis using sensor networks and consensus-based filters, pp 386–391
19. Ferrari R, Parisini T, Polycarpou M (2007) Distributed fault diagnosis with overlapping decompositions and consensus filters. In: American control conference, ACC '07, pp 693–698
20. Stankovic S, Ilic N, Stankovic M, Johansson K (2011) Distributed change detection based on a consensus algorithm. IEEE Trans Signal Process 59(12):5686–5697
21. Aghasaryan A, Fabre E, Benveniste A, Boubour R, Jard C (1998) Fault detection and diagnosis in distributed systems: An approach by partially stochastic petri nets. Discrete Event Dyn Syst 8:203–231
22. Benveniste A, Fabre E, Haar S (2003) Markov nets: probabilistic models for distributed and concurrent systems. IEEE Trans Autom Control 48(11):1936–1950
23. Kato T, Kanamori H, Suzuoki Y, Funabashi T (2005) "Multi-agent based control and protection of power distributed system—protection scheme with simplified information utilization-", in Intelligent systems application to power systems. In: Proceedings of the 13th international conference on, pp 49–54
24. Garza L, Cantu F, Acevedo S (2002) "Integration of fault detection and diagnosis in a probabilistic logic framework", in advances in Artificial Intelligence, IBERAMIA 2002, In: Garijo F, Riquelme J, Toro M (eds) vol 2527. Springer, Berlin Heidelberg, pp 265–274
25. Bron C, Kerbosch J (1973) Algorithm 457: finding all cliques of an undirected graph. Commun ACM 16:575–577
26. Horn RA, Johnson CR (1991) Topics in matrix analysis. Cambridge University Press, Cambridge
27. Langville AN, Stewart WJ (2004) The kronecker product and stochastic automata networks. J Comput Appl Math 167(2):429–447
28. Batu T, Guha S, Kannan S (2002) Inferring mixtures of markov chains. National Institute of Standards and Technology, Tech Rep
29. Frydman H (2003) Estimation in the Mixture of Markov chains, SSRN eLibrary
30. Stewart WJ (1994) Introduction to the numerical solution of Markov chains. Princeton University Press, Princeton
31. Anderson TW, Goodman LA (1957) Statistical inference about markov chains. Ann Math Stat 28(1):89–110
32. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. J Roy Stat Soc Ser B (Methodol) 39(1):1–38
33. Rabiner L (1989) A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 2:257–286
34. Cappé O, Moulines E (2005) Inference in hidden Markov models. Springer, Ryden
35. Viterbi A (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inf Theory 13(2):260–269
36. Baum LE, Petrie T (1966) Statistical inference for probabilistic functions of finite state markov chains. Ann Math Stat 37(6):1554–1563

# Engineering Optimization Approaches of Nonferrous Metallurgical Processes

**Xiaofang Chen and Honglei Xu**

**Abstract** The engineering optimization approaches arising in nonferrous metallurgical processes are developed to deal with the challenges in current nonferrous metallurgical industry including resource shortage, energy crisis and environmental pollution. The great difficulties in engineering optimization for nonferrous metallurgical process operation lie in variety of mineral resources, complexity of reactions, strong coupling and measurement disadvantages. Some engineering optimization approaches are discussed, including operational-pattern optimization, satisfactory optimization with soft constraints adjustment and multi-objective intelligent satisfactory optimization. As an engineering optimization case, an intelligent sequential operating method for a practical Imperial Smelting Process is illustrated. Considering the complex operating optimization for the Imperial Smelting Process, with the operating stability concerned, an intelligent sequential operating strategy is proposed on the basis of genetic programming (GP) adaptively designed, implemented as a multi-step state transferring procedure. The individuals in GP are constructed as a chain linked by a few relation operators of time sequence for a facilitated evolution with compact individuals. The optimal solution gained by evolution is a sequential operating program of process control, which not only ensures the tendency to optimization but also avoids violent variation by operating the parameters in ordered sequences. Industrial application data are given as verifications.

**Keywords** Engineering optimization · Nonferrous metallurgical processes · Sequential operating · Genetic programming · Imperial smelting furnace

X. Chen · H. Xu (✉)
School of Information Science and Engineering, Central South University, Changsha, China
e-mail: h.xu@curtin.edu.au

X. Chen
e-mail: xiaofangchen@csu.edu.cn

H. Xu
Department of Mathematics and Statistics, Curtin University, Perth, Australia

H. Xu and X. Wang (eds.), *Optimization and Control Methods in Industrial Engineering and Construction*, Intelligent Systems, Control and Automation: Science and Engineering 72,
DOI: 10.1007/978-94-017-8044-5_7, © Springer Science+Business Media Dordrecht 2014

## 1 Introduction

Nonferrous metals play basic roles in modern industry, economy development and national defense. However, the development of the nonferrous metallurgical industry is challenged by resource shortage, energy crisis and environmental pollution. Although the technical level of nonferrous smelting equipment has been greatly improved, the process operation is roughly instructed with faulty performance in yields, costs, energy consumption, environmental pollution and mineral recovery rate. The engineering optimization approaches are developed to deal with some restrictions in nonferrous metallurgical processes resulted from the difficulties of a variety of mineral resources, complexity of reactions, strong coupling and measurement hardness [1].

The operation level in nonferrous metallurgical production is mainly restricted by two problems for optimization [2]: (1) The imprecise process model. There are a variety of physical and chemical reactions in the process involving not only coexistence of gas, liquid and solid but also complex material and energy transfer and conversion. It is hard to build precise global mechanical models. In addition, the uncertainty contained in the closed reaction devices and long, coupled processes, and the nonlinearity between process parameters and production objects are difficult to be perfectly described and estimated in the models. So it is inevitable to utilize existing imprecise models for optimization and design smart operating strategy to avoid mismatched model to bring severe problems in production. (2) The heavy burden of process fluctuation. The key reason for fluctuation, which arises the heavy loss of quality, yields and energy, is the changeful component of the raw ore for metallurgy as well as the heavily variable process environment like some uncontrollable parameters. Furthermore, some characteristics of the nonferrous metallurgical process, such as long process, multi-process, and strong coupling, enhance the tendency of larger amplitude and longer period of process fluctuation in process operation. Therefore, the optimal process state is extremely difficult to achieve and maintain by operation. These practical problems in process operation hinder the full capability of the existing advanced process equipment and the effect of the basic automation systems and instrumentation. In many nonferrous metallurgical processes, the existing manual operation based on operators' experiences shows subjectivity and blindness in production.

Traditional optimization methods for process parameter optimization based on gradient include steepest descent method, conjugate gradient method, Newton method, Marquardt method and sequential quadratic programming algorithm, et al. However, it is convinced that the performance of these traditional optimization algorithms are heavily depending on the accuracy of the process models, which is hard to be guaranteed in nonferrous metallurgical process. From the point of engineering optimization in nonferrous metallurgical process, it is of more importance to approximate and maintain the good or satisfactory state than to set the parameters in perfect configuration according to idealized rigorous mathematic model, since the uncertain or imprecise model and process fluctuation are long-standing.

The rest parts of the chapter are arranged as following. In Sect. 2, the content and intention of engineering optimization are presented. Then a few tested engineering optimization approaches are discussed in Sect. 3. In Sect. 4, an engineering optimization case for an Imperial Smelting Process is illustrated and application data are provided as verification. Finally some conclusions and research expectations are drawn in Sect. 5.

## 2 Engineering Optimization of Nonferrous Metallurgical Processes

In a larger sense, engineering optimization of nonferrous metallurgical process includes three levels of issues: design optimization, simulation optimization and operation optimization. Design optimization is to optimize process devices, process flow and reactor structure. Simulation optimization aims to achieve the optimal parameter settings of process and the change rules of process under different working conditions through simulation calculation, process simulation and experiments. For the operation optimization, it takes technical requirements, product quality, economic indicators, and environmental indicators as the optimization objectives. When working conditions change, it provides guidance for process operation, control and scheduling to maintain optimal running and the stability of the process.

Currently, in the nonferrous metallurgical production process, the setting points tracking control has been achieved by basic automation in low level. However, it is still difficult to achieve operation optimization of the whole process taking the objectives of energy saving and reducing material consumption into account. The operation optimization faces the complexity arising by the process characteristics, such as the long process with multi-procedure, the multiple objectives and multiple models for optimization problem, as well as the uncertainties in the process. Take the alumina sintering production process as an example. The alumina sintering production process includes seven procedures as blending, sintering, dissolution, desiliconisation, decomposition, roasting and evaporation. In practice, the fluctuations of the process often spend several work shifts or even weeks of manual adjustment to restore balance again, which is a prominent case of running optimization problem.

To solve the complex engineering optimization problems, many intelligent optimization methods are proposed, including heuristic reasoning methods (like expert reasoning method), data-driven methods (like case reasoning and operation mode optimization), random search methods (like neural networks, simulated annealing algorithm, evolutionary algorithms, particle swarm optimization, state transition algorithm [3]). These intelligent optimization methods usually have good flexibility and good performance in searching the global optimal solutions without special requirement of the mathematical model. Some have been successfully applied to solve optimization problems in non-ferrous metallurgical process [4]. In [5], using the experts experiences of dealing with optimization problems, the optimal solution

is derived through heuristic knowledge reasoning. In [6], NN learning, of which the objective function is taken as the optimization goal of the network, is adopted to train the optimization process. Swarm intelligence algorithms, such as improved genetic algorithms or Particle Swarm Optimization algorithms, are used to solve optimization problems with no special requirements for the optimization model, like applied in [7, 8]. Chai et al. [9–12] studied the optimization problems in the processes of fused magnesia smelting, flotation, grinding circuits of mineral processing and shaft furnace roasting respectively, and the optimization control methods, such as case reasoning, rule reasoning, neural network, fuzzy self-correction method, and the multi-model hybrid control method are proposed according to the different process characteristics. Based on these results, online optimization settings are implemented [13]. However, at present, there is no generally applicable optimization framework for nonferrous metallurgical processes [14]. Here, several optimization methods in view of different process characteristics are presented for non-ferrous metallurgical process.

## 3 Typical Engineering Optimization Approaches

### *3.1 Process Optimization Based on Operational Pattern*

Actually, a series of goals, such as enhancing production efficiency, saving energy, and reducing pollutant emission for complex industrial process, are achieved through specified optimal operations. Blind operation will lead to process fluctuation, by which not only the quantity and quality of products will be reduced, but also energy consumption, material consumption and pollutant emissions will increase.

The operating parameters of complex nonferrous metallurgical process are highly coupling and conflicting with each other. The overall optimization of the entire production process is very complex and difficult to balance. Furthermore, there are many operation variables which need to be decided at the same time. The system input conditions and the operating parameters need to be decided, which actually construct an operational pattern. In most of the processes, the operational patterns are obtained through long-term production practice. The operators can make operation decisions by memorizing and exploring these operational patterns. Nevertheless, human operation mode is subjective, rough, hard to remember, and difficult to update.

In practical industrial production process, a massive number of data are transmitted to the data server through distributed control systems and industrial networks. They are then restored in various forms. These massive data implicate rich information on relationship between the operation rules and the process parameters. Thus, process input conditions and controllable operating parameters are used to form an operational pattern of non-ferrous metallurgical process. By using the operational patterns, a data-driven operational pattern optimization method is proposed [15]. The core idea of the method is as follows. Firstly, the relationships among input condi-

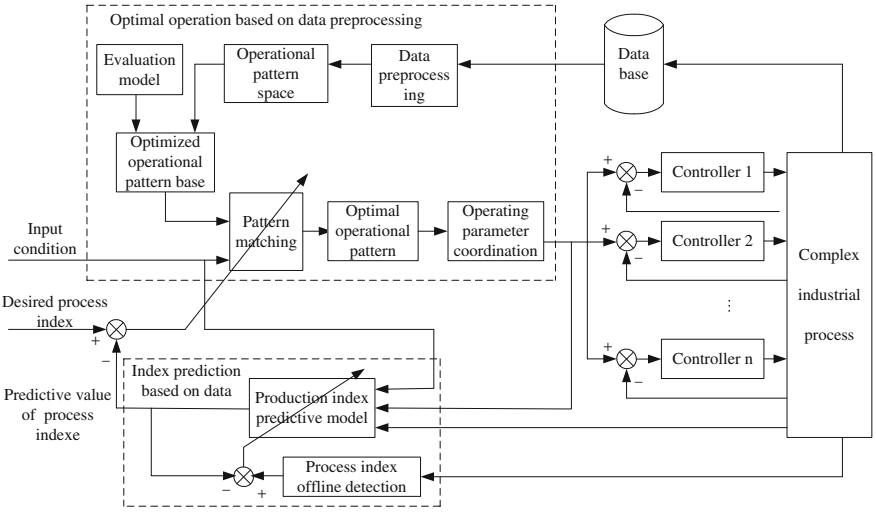



**Fig. 1** Frame of data-driven operational pattern optimization

tions, state parameters, operation parameters and process indexes are extracted from the industrial operating data. They are then used to establish optimized operational pattern base. Finally, according to the current operating condition and state, an optimal pattern which is the best match for the current operating condition is found, such that the process indicators tend to be optimal. The frame of data-driven operational pattern optimization is shown in Fig. 1. It mainly includes data preprocessing, data-based index prediction, optimized operational pattern base and operating parameters optimization.

This method is applied to develop the optimal control of a copper flash smelting process [15]. The results obtained show that the matte quality and quantity are improved, and the running conditions are stabilized.

## *3.2 Satisfactory Optimization Method Based on Soft Constraint Adjustment*

The complexities of the nonferrous metallurgical process optimization problem make it hard to solve. But the specified process may present some engineering features which may offer good conditions for solving engineering optimization problems. For real non-ferrous metallurgical processes, the constraints, which may conflict with other constraints, are too complex to guarantee optimal solution. Actually, most of these constraints are arising from the production experiences and not necessary to meet every boundary constraint. Therefore, the working conditions are integrated with the constraints to construct a satisfactory degree function based on

soft constraints adjustment. By adjusting the constraint domain appropriately, the computation efficiency is enhanced and the solution quality is improved. Then, an optimization method based on soft constraints adjustment is proposed for the nonferrous metallurgical production process with soft boundary constraints and conflicting constraints.

Consider the blending process of a copper flash smelting process. Since the boundary constraints are not very strict, the boundary value can be modified by transforming the constraints into boundary adjustment objective functions according to the priority order.

Suppose the constraints as follows.

$$A_{\min}^{(p)} \leq A^{(p)} \cdot X \leq A_{\max}^{(p)} \tag{1}$$

where $p$ is the priority level of the adjustment of the constraint conditions, which is set according to the actual situation. The larger the value is, the more likely the adjustment will be accepted. For example, we introduce $p$ groups of logical variables $\delta_{\min}^{(i)}$ and $\delta_{\max}^{(i)}$, $i = 1, 2, \ldots, p$, and intermediate variables $\varepsilon_{\min}^{(i)}$ and $\varepsilon_{\max}^{(i)}$, $i = 1, 2, \ldots, p$.

Then the constraint can be converted to following equation:

$$\begin{aligned} A_{\min}^{(i)}(1 - \delta_{\min}^{(i)}) + \delta_{\min}^{(i)} \cdot \varepsilon_{\min}^{(i)} &\leq A^{(i)} \cdot X \\ &\leq A_{\max}^{(i)}(1 - \delta_{\max}^{(i)}) + \delta_{\max}^{(i)} \cdot \varepsilon_{\max}^{(i)} \end{aligned} \tag{2}$$

If and only if when $\delta_{\min}^{(i)} \cdot \delta_{\max}^{(i)} = 0$, the constraint with the $i$th priority level is satisfied. The constraints are sorted in ascending order according to their priority levels. Then, the minimum and maximum values of the adjusted constraints are obtained. Each constraint boundary is updated sequentially according to the order of priority levels until all constraints updating is finished.

Based on the concept of the soft constraints adjustment method, a satisfactory optimization method [16] of burden process of copper flash smelting has been applied to a large copper smelting enterprise. It enhances the batching control accuracy, stabilizes ingredient quality and reduces production costs.

### *3.3 Satisfactory Optimization Method Through Uncertain Decentralization*

The non-ferrous metallurgical process is a long procedure flow with distributed uncertainties and is characterized with diversity, fuzziness and object conflicting. For this, an uncertain decentralization-based optimization method for nonferrous metallurgical process is proposed in engineering. In this method, the optimization problem is firstly divided into several subordinate optimization problems by introducing intermediate target variables. Then, some intelligent methods are adopted to adjust

the intermediate optimization target values to coordinate these sub-optimization-systems, such that the final product quality can meet the strictly specified requirements. In the uncertain decentralization based optimization method, the uncertainties are handled separately and the impacts of the uncertainties are gradually weakened.

In a bauxite raw slurry blending process [17], the materials of bauxite, adjustment ore, limestone, alkali powder, anthracite, carbonation decomposed spent liquid, and silicon slag are mixed together to get the qualified raw slurry with required ingredients for sintering process. Due to the unstable composition of ore sources supplied to the process, large fluctuations of bauxite composition and difficulties of on-line detection, the information of raw materials bears significant uncertainty. To improve the quality of raw slurry flown into the tank, an optimization system is established. To solve this optimization problem, a two-stage intelligent optimization method, is developed to realize the optimization of raw material proportioning and re-mixing operation. The real-time adjustment laws in each stage is reasoned and yielded using the uncertain decentralization based optimization method. The application of this two-stage intelligent optimization method shows that the quality of the raw slurry is greatly improved.

For the nonferrous metallurgical processes, the optimal solution of the optimization problems is difficult to obtain or too costly to solve because of inaccuracies of the model, ambiguity of constraints, and conflicting multi-objectives. An intelligent optimization method based on multi-objective satisfaction optimization is proposed for such engineering optimization problems. In which, a satisfactory function is derived to evaluate the satisfaction degree of decision makers about the performance indicators. A comprehensive satisfactory function is proposed to estimate the requirement of the decision makers about multi-objective coordination. Based on the satisfactory function and the comprehensive satisfactory function, a satisfactory optimization model is established. Since, this model combines all the objectives and constraints satisfactory functions, thus it can be solved by maximizing the overall satisfactory function. Therefore, the results obtained will achieve better performance on actual process indexes evaluation.

In the process of raw material bauxite blending and converting for alumina production, the optimization objective is to supply clinker kiln with acceptable raw slurry make full use of all converting tanks. This optimization problem can be described as: find a combination of tanks to minimize the error between the real raw slurry indexes sent to the clinker kilns and the setting indexes, constrained by the indexes of remaining tanks and the numbers of the tanks are selected [18]. Clearly, this optimization problem is a combinatorial optimization problem. It is solved by using the intelligent optimization method based on multi-objective satisfactory optimization method. The best combination of the tanks shows that the converting times are successfully reduced from 3 to 2 times. So that the process operation is simplified and the energy consumption is reduced.

# 4 Intelligent Sequential Operating Method and its Application in Imperial Smelting Furnace

## *4.1 ISP Process Description*

Imperial Smelting Process (ISP) is a typical complex metallurgical process firstly patented by Imperial Smelting Company. The outstanding feature of ISP is to smelt lead and zinc in the closed furnace (Imperial Smelting Furnace, ISF) with less cost and more metallurgical complexity. Lead and zinc are simultaneously and continuously smelted in a furnace in a series of complicated chemical reactions with little process details known.

At the beginning a preparative sintering process is performed for smelting material preparation, the agglomerate. In this process the zinc-lead ore is sintered and desulfurated under certain burning conditions and agglomerate is put out with certain rigidity, size and permeability.

The smelting process is a bit like blast furnace in iron smelting process as shown in Fig. 2, the agglomerate is blended with coke in certain ratio and put into ISF in two bells. Three blast furnaces are producing and blasting air of more than 800 °C for ISF in turn. Most hot air is put into the bottom of the furnace and the rest, as secondary air, is sent into the upper part near the surface of reacting material. With the redox reaction taking place, the sulfates of the metals are reduced under the function of carbon in coke. Zinc, of low boil point, is gasified and passes through the throat of furnace as an entrance to a condenser. Here plenty of tiny drops of liquid lead are sprayed to absorb zinc gas of high temperature. After passing through a series of devices for segregating, the mixture of liquid zinc and lead is separated. Zinc is lighter and accumulates in the upper layer while lead is heavier and accumulates in the lower layer. After abstracting, refining and purification successively we obtain raw zinc of 98 % and pure zinc of 99.995 %. Because the lead is heavy and fusible, the liquid lead accumulates in the bottom of furnace and is discharged at intervals together with floating residue. In an electric heating fore well, the residue is separated from lead and the raw lead of 98 % is refined including a little noble metal like gold or silver. The exhaust gas released from furnace throat is put into a washing tower for dust removal.

In ISP process, the inputs are coke, agglomerate and air and the outputs are zinc, lead, residue and exhaust gas. The operations and control commands are material feed batches, the temperature of airflow, primary airflow, secondary airflow as well as the interval of discharging residue. The reactions in the production are happening under the conditions of high temperature, high pressure and hermetic state. The above operations and commands can exert remarkable influence on the smelting process but few ready laws can be conformed to so they are performed empirically.

The target of ISP optimal control is to maintain a best metallurgical balance and stable reaction conditions for a less cost caused by production variety with a guarantee of satisfactory zinc direct recovery rate, which is the rate of zinc transformation from compounds in the raw mine to pure metal.

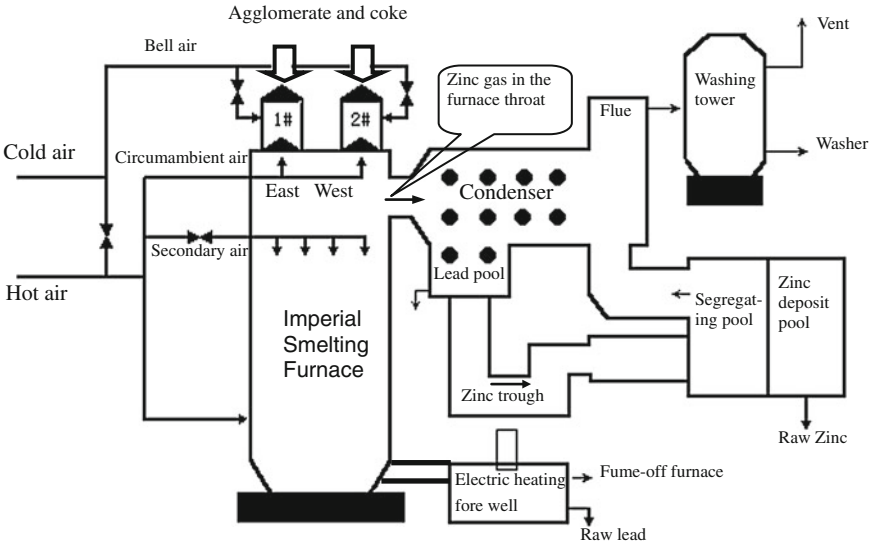



**Fig. 2** Technical sketch of Imperial Smelting Process

## *4.2 Sequential Operating Strategy*

In ISP metallurgical processes the conditions are changing continuously and the data measured is insufficient, lag or inaccurate due to heat, dust or acute reaction. The operators must be cautious to restrain from unstable state, but manual operation seldom responses well for subjectivity and faulty knowledge.

Here we take the plant as a sequential transition procedure and the target of optimal control is to operate the system parameters fluently and stably toward a relatively good or satisfactory state under certain conditions in next a few steps. The original idea of sequential operating strategy is proposed based on adaptively designed swarm optimization method.

Considering both the format of sequential configuration scheme and proper expression for search algorithm, we designed a length-flexible chain structure for evolving individuals which is composed of operating vectors as leaf nodes connected by temporal relation symbols as link nodes. Based on single objective models already achieved or learned in processes, a multi-step comprehensive evaluation algorithm is introduced for complete judgment of control program in optimizing procedure. The optimal solution gained by optimization is a sequential control program with each step corresponding to a configuration vector of control parameters. The solution can be either directly sent to controller or taken as instructions for control operation of engineers at the console.

The output of the module is sent to a coordinator with confirmation or intervention to control operations from human operators. The coordinator produces a certain setting points of control parameters and sends them to controllers or DCS. After

final detection the output of the process as the practical value of objects is compared with the predictive value generated by models, the error is used for model adaptive modification.

## *4.3 Adaptively Designed Genetic Programming*

### 4.3.1 Genetic Programming

A programming problem for optimization is generally described as follows. Let $\{(x_i, y_i) : x_i \in X, y_i \in Y, i \in I\}$ denote the set of given input-output pairs in which $X, Y$ are subsets on finite dimensional spaces and $I$ is the target set. If $C(X)$ is the total of continuous functions of $X$ and $F \subset C(X)$ and $\rho$ is the distance defined on product domain $\Pi_{i \in I} Y$, then the programming problem is to find a function $f^* \in F$ to satisfy Eq. (3) for arbitrary $f \in F$.

$$\rho(\{f^*(x_i)\}, \{y_i\}) \leq \rho(\{f(x_i)\}, \{y_i\}) \tag{3}$$

The other form is that for any $\varepsilon > 0$ we have

$$\rho(\{f^*(x_i)\}, \{y_i\}) \leq \varepsilon \tag{4}$$

So we have the optimization expression like Eq. (5).

$$\min_{f \in F} \rho(\{f^*(x_i)\}, \{y_i\}) \tag{5}$$

Genetic Programming (GP) was proposed by Koza in 1992 [19] and developed rapidly as a significant branch of Evolution Algorithm. In Genetic Programming, the individual chromosome is a layered structure composed of functions and terminals. The set of functions includes operators, mathematical functions, condition expressions and so on. The set of terminals includes variables and constants. The common structure of GP individuals is tree type composed of elements in function set as arithmetic operators or sub-functions and elements in terminal set as variables and constants. This kind of layered architecture with branches is apparently flexible to represent diverse and complex objects in solution space. But on the other hand, this layered architecture arouses a complexity of possible combinations of elements defined at the beginning. The searching space will expand exponentially as the number of layers increases and the searching process is handicapped in these cases.

Koza [20] pointed out that Genetic Programming addresses one of the central goals of computer science, namely automatic programming. The goal of automatic programming is to create, in an automated way, a computer program that enables a computer to solve a problem. In an extensive sense, GP promotes an available expectation to create a scheme of best solution automatically for a certain issue as long

as the preconditions are satisfied. A flexible and active structure is the most attractive feature of Genetic Programming(GP) which makes it possible to restructure control program from independent individual operations. This feature distinguished from other EA methods overcomes the restriction of solution expression and leads to many fruits like in parameter identification [22], speech modeling [23] and knowledge mining [24].

There are two preconditions for a successful GP optimizing approach [21]. The first precondition is that the solution for the problem can be expressed completely by defined functions and terminals. Second, any solution for the problem can be evaluated by its performance, i.e. a proper fitness principle should be provided. The two preconditions of GP application mentioned above are settled in following ways for process optimal control. We define the set of parameter configuration vectors $\boldsymbol{C}_{1\times n}$ as the terminal set where $n$ is the number of controllable parameters. The function is the symbol of temporal relation of operation commands. In this way any optimal control program is structured as a sequential control schema of multi-step operations. To solve the evaluation problem, a multi-step comprehensive evaluation algorithm is introduced for fitness comparison of different control programs.

### 4.3.2 Sequential Operation Evaluation

The precondition of sequential comprehensive evaluation algorithm is that a reliable and proper model of the relationship between process control variables and single production target. The mission fulfilled by the algorithm is to evaluate the total performance of a sequence of process states within a foreseeable future if the inputs and control variables are decided. A group of models for single objective prediction are also built in advance and produces indexes such as production, quality, cost and other important process objects.

Suppose $P = \{\rho_1, \ldots, \rho_n\}$ is the assessment index set of process performance and $n$ is the number of indexes. Eq. (6) is the formula of synthetic evaluation for a process state.

$$E = \sum_n \alpha_i f_i(\rho_i) \tag{6}$$

where $A = \{\alpha_1, \ldots, \alpha_n\}$ is the weight set of indexes as a reflection of influence of each index in the process, $f_i \in [0, 1]$ is the normalization function of the $i$th index. The weight set $A$ is computed by variation coefficient. To each group of index $\rho_1, \ldots, \rho_n$, let

$$\bar{\rho} = \frac{1}{n}\sum_{i=1}^{n} \rho_i \tag{7}$$

and

$$\sigma_k = \sqrt{(\frac{1}{n-1}\sum_{i=1}^{n}(\rho_i - \bar{\rho})^2)}, \qquad k = 1, \ldots, n \tag{8}$$

Then we have

$$v_k = \frac{\sigma_k}{|\bar{\rho}|}, \qquad k = 1, \ldots, n \tag{9}$$

where $v_k (k = 1, \ldots, n)$ is the variation coefficient of group $\rho_1, \ldots, \rho_n$ and the weight is

$$\alpha_j = \frac{v_j}{\sum_{i=1}^{n} v_i}, \qquad j = 1, \ldots, n \tag{10}$$

Based on (7–10), Eq. (6) can be solved for synthetic evaluation E for one state of process.

Let matrix $\Gamma_{k \times n} = [P_1; P_2; \ldots; P_k]$ denote a state transferring procedure of a temporal sequence, the line vector $P_i = \{\rho_{i1}, \ldots, \rho_{in}\}$ is the index vector of the $i$th state $(i = 1, \ldots, k)$ and the column vector $\{\rho_{1j}; \ldots; \rho_{kj}\}$ is the $k$-dimension temporal series of the $j$th index $(j = 1, \ldots, n)$. For the $i$th state, we compute the synthetic evaluation $E_i (i = 1, \ldots, k)$ by Eq. (6)–(10). We construct a $k$-dimensional temporal factor sequence $\{\beta_1, \ldots, \beta_k\}$ and let $E_\Gamma$ denote synthetic evaluation of a sequence of transferring states, then

$$E_\Gamma = \sum_k E_i \beta i \tag{11}$$

For an assessment of industrial process the temporal factor sequence $\{\beta_i, \ldots, \beta_k\}$ is decreasing as time goes on. So a closer state will give a larger influence to $E_\Gamma$, but the value of $\beta_i$ is related to the response cycle and time-varying characteristics of plant.

Thus far we have discussed the synthetic evaluation of a sequence of transferring states. The second thought concerned by engineers is the stability of process state transferring. A Euclidian distance of indexes between neighbor states is introduced as a measurement of the transferring volatility and the cost of stability for a potentially process improvement.

$$D_i = \sqrt{\sum_{j=1}^{n} [f_j(\rho_{ij}) - f_j(\rho_{i+1,j})]^2} \quad i = 1, \ldots, k-1 \tag{12}$$

the index is normalized to eliminate the impact of different ranges and units to distance computation.

So the evaluation to the stability of a sequence of transferring states is

$$D_\Gamma = \sum_{i=0}^{k-1} D_i \beta_{i+1} \tag{13}$$

where $D_0$ is for transferring stability from the present state to the first step.

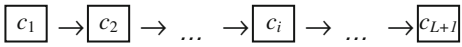



**Fig. 3** The chain structure of individuals in GP

As a synthesis of state evaluation and stability evaluation, the fitness function of multi-step synthesized evaluation is defined as

$$J = E_\Gamma - \theta \cdot D_\Gamma \tag{14}$$

where $\theta$ is a punishment coefficient for process state fluctuation.

Via the function of comprehensive evaluation, a total assessment is abstracted from a series of independent state objects, in which the degree of state changing is involved.

### 4.3.3 Implementation of Intelligent Sequential Operation

The individuals in GP evolution for process optimal control are various control schemes of parameter vectors in temporal sequences, which brings a difference from common layered structure of GP individuals. The chain structure is shown in Fig. 3, in which the node element is the control vector $\boldsymbol{c} \in \boldsymbol{C}_{1\times n}$ and the link element "→" is the temporal relation symbol indicating that the left command should be performed before the right one.

The definitions of terminals and functions satisfy the two requirements of initial definition of GP elements, sufficiency and closure. Sufficiency means all possible solutions can be expressed as a certain combination of terminals and functions. The solutions of process optimal control are a sequence of parameter settings for controller of uncertain temporal length. The sequence is determined by temporal relation symbols and the content of each control command is determined by control vectors. So the expression is sufficient. The property of closure guarantees the validity of combinations, i.e. the combinations of functions and terminals are valid in solution space. The form of combination is defined as chain structure of terminals linked by functions alternately. Any combination in this way, even only a unique terminal in the chain, makes practical as well as doubtless sense as control schemes regardless of the length of steps. So the expression is also close.

The implementation steps are expatiated as follows.

*step 1:* Randomly produce an original population $G_0$ of $s$ size with chain structure individuals. The maximum length of the operation chain $L$, i.e. the maximum number of → links, should be decided to be a small integer at the beginning for conformable individual size. First, a shorter maximum individual brings a much less possibility of combination and is necessary to save searching time. Second, for a slow and complex process with severe disturbance it is insignificant to consider control operation long after present states for the length $L$ is the number of control steps in the future.

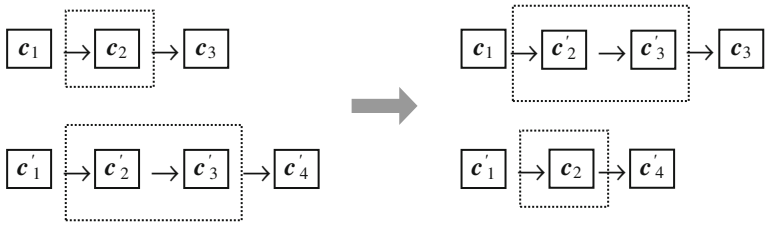



**Fig. 4** Crossover operation of chain individuals

*step 2:* The nodes in an individual are control vectors corresponding to process parameters. Let control vectors and process conditions be the inputs of single objective models and the estimations $\rho_1, \ldots, \rho_n$ are acquired. Then compute the multi-step synthetic evaluation $J$ of this individual by the algorithm introduced in Sect. 4. In this way we get $J_i (i = 1, \ldots, s)$ for all individuals in $G_0$.

*step 3:* Make a proper choice of evolution parameters such as selection probability $P_s$, mutation probability $P_m$, crossover probability $P_c$ and inversion probability $P_i$, then according to fitness $J_i (i = 1, \ldots, s)$ carry out evolution operations in $G_0$ until new individuals are produced and we have

$$P_s + P_m + P_c + P_i = 1 \tag{15}$$

In Crossover, choose pairs of individuals from $G_0$ with probability $P_c$, randomly decide the crossover positions and sectors and exchange the sectors at the positions (shown in Fig. 4).

*step 4:* Add one to the number of evolving generations $t$ and regard the new $s$ individuals as a new generation.

*step 5:* If $t$ reaches the maximum generations predetermined or the evolving process satisfies the convergence conditions, the evolution terminates and the best individual is the optimal control program returned by GP searching; else loop to *step 2*.

The optimal program of a sequence of operation acquired by GP will be sent to controllers step by step for a stable optimizing process. However, in applications like path programming and program self-organizing, GP method sometimes encounters the problem of space over expansion or code over increasing which leads to difficulties for convergence in evolution. In fact, the variety of function elements and the binary tree structure of individuals sometimes cause a dimensionality curse in evolution operations of crossover and mutation. The individuals constructed as a chain structure and designed for multi-step optimal control for complex process are not only easy for GP searching but also convenient for implementation with simple functions and links.

## 4.4 Implementation

To accomplish the optimal control target, we must determine the control variables and the assessment objects. Some of input parameters are not controllable but predetermined, like the composition of agglomerate, which is determined by sintering process before smelting. By analysis of mechanism and data relativity, the control vector is selected as {primary airflow, airflow temperature, secondary airflow, feed batches, coke ratio}. The first two variables of primary airflow determine the reduction atmosphere and the reduction reactions. The secondary airflow has an effect on the temperature in the top and a high temperature can prevent the reoxidization of zinc. The feed batches decide how much and how fast the material is added into the furnace. The coke ratio is the proportion of coke in material and coke is the only reducer in all reactions in the furnace. All these variables are controllable and measurable online.

The assessment objective vector is defined as {permeability, zinc yield, zinc in residue}. The permeability is derived from and defined by aerodynamic CARMAN formula

$$K = Q^{1.7}/(P_B^2 - P_T^2) \tag{16}$$

where $K$ is the permeability, $Q$ is the primary airflow, $P_B$ is the pressure in the bottom and $P_T$ is the pressure in the top. In the furnace $P_T$ is a relatively small constant. The permeability reflects the degree of fluency of chemical reacting and the situation of whole process conditions. The zinc yield is the quantity objective of main product. The zinc in residue is to judge how much zinc is wasted and how thoroughly the reduction is performed.

There are also some predetermined and fixed process conditions like compositions in agglomerates, the feeding method and discharging interval. They are treated as measurable data for models.

Based on field research with technicians of ISP production, three single objective prediction models for permeability, zinc yield and zinc in residue had been built for application here. The GP optimal control strategy proposed is applied in ISP process control. The control interval between steps is decided to be 9 min according to data sampling interval and the reaction cycle. Another reason for the interval selection is the interval of feeding is around 9 min so that we can decide the next feeding parameters. The maximum length of chain individuals is defined to be three because the furnace is a large temporal inertia plant and it is unnecessary to consider too many control steps in the future.

According to the sequential operating idea, the optimal control strategy in ISP process is implemented as following. The size of GP population, maximum evolving generations, the parameters of genetic operator $P_s$, $P_m$, $P_i$ and $P_c$ are determined by simulation test and verified in practice. After evolving, it returns no more than three control vectors in temporal order. The first vector physically means a best operation under present conditions subject to stabilization constraints. By the function of coordinator, human operators have a chance to decide whether to accept the operation or

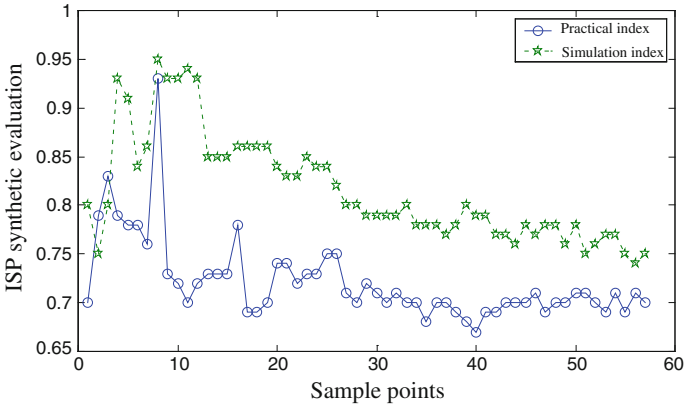



**Fig. 5** ISP simulated optimal control

not. In the next step, the system will recompute operations or not depending on the degree of conditions' change of the time.

As a comparison, a simulated optimization based on practical data is tested and shown in Fig. 5, in which the circles linked by solid lines are synthetic evaluations of practical process states and the stars linked by dashed lines are synthetic evaluations of optimized states by simulation. To compare the points of states, the synthetic evaluation, together with all evaluations mentioned in this section, is computed by Eq. (6) and does not involve the concept of temporal sequence and state stability. The latter evaluations are obviously better and less fluctuating than practical ones.

The algorithm involving this optimal control strategy was included in the ISP Monitoring & Control System and the result of optimal control commands was shown on screen as instructions for process engineers. The test running of the system also proved the effectiveness of optimal control program for instructions and the superiority to previous empirical control.

In fact, the online control procedure is a rolling optimization and the states of process are less fluctuating than shown in Fig. 5 because whole furnace conditions tend to be stable. The simulated optimization and practical application to ISP optimal control have provided evidences for the advantages of the operating strategy.

With attention to the difficulties in engineering optimization of the nonferrous metallurgical processes, the intelligent operating method is applied to the case of ISF. Firstly, the method makes an effort to approximate the optimal balance point by operating sequence in multi steps so as to avoid likely risks of fluctuation of process objects. Secondly, based on models of the relationships between parameters and single objective a synthetic evaluation for total assessment of a sequence of process states is proposed with a punishment of state fluctuation. Finally an adaptive design for sequential operating optimization algorithms is proposed based on GP algorithm, including the chain structure of individuals and the flexible length of individuals for sequential operating.

## 5 Conclusions

In most nonferrous metallurgical processes, it is hard to implement optimal operation because of imprecise model and process fluctuation. Based on engineering practice of nonferrous metallurgical process control, engineering optimization problems are considered and practical industrial applications are studied and analyzed. From the view of engineering optimization, the key focus in production is how to maintain the good balance and response properly to variations of all kinds of conditions. This chapter concentrates the focus by discussing the idea of engineering optimization and some effective approaches. Then we take the ISF process as an example to reveal the implementation of the method of intelligent sequential operation. Honestly, the idea of engineering optimization is far from rigorous methodology and many problems are challengeable in optimal operation for nonferrous metallurgical processes, such like operation optimization with multi-model presentation, operation optimization with implicit objective function, trajectory optimization of working point migration in long process with minimum energy consumption, pattern optimization of under-operation conditions. Considering the strategic needs of energy saving and pollutant emission reduction, these challenges of operation optimization problems will be explored oriented to the green production of nonferrous metallurgical process.

**Acknowledgments** The work mentioned in this chapter is supported by Nature Science Foundation of China (61104078), Foundation of State Education Ministry grant of China (20100162120019) and the Science and Technology Program of Hunan Province grant (2011CK3066).

## References

1. Hodouin D (2011) Methods for automatic control, observation, and optimization in mineral processing plants. J Process Control 21(2):211–225
2. Gui WH, Yang CH (2010) Intelligent modeling, control and optimization of complex nonferrous metallurgical process. Science Press, Beijing (in Chinese)
3. Zhou XJ, Yang CH, Gui WH (2012) State transition algorithm. J Ind Manag Opt 8(4):1039–1056
4. Chai QQ, Yang CH, Teo KL et al (2012) Optimal control of an industrial-scale evaporation process: sodium aluminate solution. Control Eng Pract 20(6):618–628
5. Liau LCK, Yang TCK, Tsai MT (2004) Expert system of a crude oil distillation unit for process optimization using neural networks. Expert Syst Appl 26(2):247–255
6. Nascimento CAO, Giudici R, Guardani R (2002) Neural network based approach for optimization of industrial chemical processes. Comput Chem Eng 4(9–10):2303–2314
7. Liu P, Su JH, Dong QM et al (2005) Optimization of aging treating treatment in lead fram copper alloy by intelligent technique. Mater Lett 59(26):3337–3342
8. Dorrah HT, El-Garhy AM, El-Shimy ME (2011) PSO-BELBIC scheme for two-coupled distillation column process. J Adv Res 2(1):73–83
9. Kong WJ, Chai TY, Yang SX, Ding JL (2013) A hybrid evolutionary multiobjective optimization strategy for the dynamic power supply problem in magnesia grain manufacturing. Appl Soft Comput 13(5):2960–2969
10. Chai TY, Geng ZX, Yue H, Wang H, Su C-Y (2009) A hybrid intelligent optimal control method for complex flotation process. Int J Syst Sci 40(9):945–960

11. Zhou P, Chai TY, Wang H (2009) Intelligent optimal-setting control for grinding circuits of mineral processing process. IEEE Trans Autom Sci Eng 6(4):730–743
12. Chai TY, Ding JL, Wu FH (2011) Hybrid intelligent control for optimal operation of shaft furnace roasting process. Control Eng Pract 19(3):264–275
13. Wang ZJ, Wu QD, Chai TY (2004) Optimal-setting control for complicated industrial processes and its application study. Control Eng Pract 12(1):65–74
14. Chai TY (2009) Challenges of optimal control for plant-wide production processes in terms of control and optimization theories. Acta Automatica Sinica 35(6):641–649( in Chinese)
15. Gui WH, Yang CH, Li YG et al (2009) Data-driven operational-pattern optimization for copper flash smelting process. Acta Automatica Sinica 35(6):717–724( in Chinese)
16. Yang CH, Wang X-L, Tao J et al (2008) Modeling and intelligent optimization algorithm for burden process of copper flash smelting. J Syst Simul 20(8):2152–2155( in Chinese)
17. Yang CH, Gui WH, Kong LS, Wang YL (2009) Modeling and optimal-setting control of blending process in a metallurgical industry. Comput Chem Eng 33(7):1289–1297
18. Yang CH, Gui WH, Kong LS, Wang YL (2009) A two-stage intelligent optimization system for the raw slurry preparing process of alumina sintering production. Eng Appl Artif Intell 22(4–5):796–805
19. Koza JR (1992) Genetic programming: on the programming of computers by means of nature selection. MIT Press, Cambridge
20. Koza JR (1998) Genetic programming. In: Williams JG, Kent A (eds) Encyclopedia of computer science and technology. Marcel-Dekker, New York
21. Koza JR (2003) Advances in evolutionary computing: theory and applications. Springer, Berlin
22. Gusel Leo, Brezocnik Miran (2011) Application of genetic programming for modelling of material characteristics. Expert Syst Appl 38(12):15014–15019
23. Wu XJ, Yang ZZ (2013) Nonlinear speech coding model based on genetic programming. In: Applied soft computing. Corrected proof, Available online 14 March 2013 (In Press)
24. Malo Pekka, Siitari Pyry, Sinha Ankur (2013) Automated query learning with wikipedia and genetic programming. Artif Intell 194(1):86–110

# Development of Neural Network Based Traffic Flow Predictors Using Pre-processed Data

**Kit Yan Chan and Cedric K. F. Yiu**

**Abstract** Neural networks have commonly been applied for traffic flow predictions. Generally, the past traffic flow data captured by on-road detector stations, is used to train the neural networks. However, recently research mostly focuses on development of innovative neural networks, while it lacks development of mechanisms on pre-processing traffic flow data priors on tr aining in order to obtain more accurate neural networks. In this chapter, a simple but effective training method is proposed by incorporating the mechanisms of back-propagation algorithm and the exponential smoothing method, which is proposed to pre-process traffic flow data before training purposes. The pre-processing approach intends to aid the back-propagation algorithm to develop more accurate neural networks, as the pre-processed traffic flow data is more smooth and continuous than the original unprocessed traffic flow data. This approach was evaluated based on some sets of traffic flow data captured on a section of the freeway in Western Australia. Experimental results indicate that the neural networks developed based on this pre-processed data outperform those that are developed based on either original data or data which is preprocessed by the other pre-processing approaches.

**Keywords** Intelligent traffic management · Traffic flow predictions · Neural network · Data processing · Time-series forecasting · Data cleansing

K. Y. Chan (✉)
Department of Electrical and Computer Engineering, Curtin University, Bentley, WA, Australia
e-mail: kit.chan@curtin.edu.au

C. K. F. Yiu
Department of Applied Mathematics, The Hong Kong Polytechnic University,
Hong Kong, People's Republic of China
e-mail: macyiu@polyu.edu.hk

H. Xu and X. Wang (eds.), *Optimization and Control Methods in Industrial Engineering and Construction*, Intelligent Systems, Control and Automation: Science and Engineering 72,
DOI: 10.1007/978-94-017-8044-5_8, 

## 1 Introduction

Traffic flow prediction is essential for intelligent traffic management systems, which mainly intend to reduce traffic congestion and improve mobility of traffic flow [6]. It has a horizon of only a few minutes, in order to support proactive dynamic traffic control on anticipating traffic congestion. Although traffic flow predictors with reasonable accuracies can be generated by classical statistical methods such as filtering methods [10], autoregressive moving average approaches [12] and k-nearest-neighbor methods [3], they cannot capture highly nonlinear characteristics in traffic flow data. Recently, artificial neural networks (NNs) have commonly been applied for development of traffic flow predictors [4, 7], which can effectively address nonlinear characteristics on traffic flow data.

However, solely using NNs may not achieve the best generalization capability for traffic flow prediction. Development of innovative NN approaches is commonly involved on incorporation with other intelligent methods. For example, Takagi-Sugeno fuzzy NNs [9] have been used for traffic flow prediction [11, 13, 17] by using the mechanisms of both fuzzy logic and NNs. Also, a new NN method is proposed to predict traffic flow conditions based on forecasting outputs from both NNs and Kalman filters [14]. While all these improved NN approaches outperform the pure NN approach on traffic flow prediction, more NN parameters needed to be determined for these improved NNs than those on the pure NN models. Also, more computational power and space is needed when implementing innovative NN models than is needed by the pure NN models. Hence, those NN models are not suitable to be tuned adaptively compared with the pure NN models, as stronger processors are needed for those improved NN models.

This chapter presents a hybrid method incorporating the mechanisms of exponential smoothing method and back-propagation algorithm (BP) namely EXP-BP. It is computational effective and efficient because the simple three-layers-neural-network is only implemented on traffic flow prediction. This approach intends to predict more accuracy traffic flow conditions.

We are motivated by the observations that the characteristics of traffic flow data are highly lumpy. Training error of NNs can be made to zero value by fitting all lumpiness in the traffic flow data. However, having a zero training error may cause an overtrained NN, which degrades the traffic flow prediction on unseen data. If lumpiness on original data is filtered before using for training, the accuracies of the traffic flow predictions generated by the NNs could be increased [18]. Also, this method has been used on short-term electric load forecasting, where more accurate predictions can be achieved than those obtained by only using the BP algorithm [5]. In EXP-BP, lumpiness in traffic flow data is removed by the exponential smoothing method [8, 16] before using for developing NNs. After filtering out lumpiness based on exponential smoothing, EXP-BP uses the BP algorithm to generate the NNs based on the pre-processed traffic flow data. The resulting NNs intend to fit the traffic flow characteristics when the lumpiness is removed. Using traffic flow data captured on a

section of the freeway in Western Australia, experimental results show that NNs with better prediction in future traffic flow conditions can be obtained by the EXP-BP.

## 2 Pro-processing Traffic Flow Data

### *2.1 Original Traffic Flow Data*

The traffic flow predictor was developed using traffic flow data captured from the n detector stations ($D_1, D_2, \ldots D_n$) installed along the freeway. The $i$th detector station, $D_i$ s, captures two traffic flow measures namely, the average speed $s_i(t)$ of vehicles at $D_i$ and the average headway distance $h_i(t)$ between two consecutive vehicles passing through $D_i$ from time t to time $t + T_s$ with the sample time $T_s$. When $s_i(t)$ is near the speed limit of the freeway and $h_i(t)$ is high, traffic flow at the location of $D_i$ can be supposed to be smooth. Otherwise, traffic congestion may occur on this section of the freeway.

Here each detection station $D_i$ captures the vehicle speed by using the on-road sensor which is installed with two inductive loop detectors of which both of them are separated by a small distance. The inductive loop detector is assembled with a metallic coil in a big loop size, and the metallic coil is buried beneath a lane on a particular section of the road. These two inductive loop detectors are linked with the detection station, which supplies electric current to the loops and processes the measures captured by the inductive loop detectors. Those captured measures are used to determine if a vehicle is passing through. In order to estimate the vehicle speed, first the time taken for the vehicle to travel between the two inductive loop detectors is captured. Second, the vehicle speeds are estimated based on the time difference between the two captions and the distance between the two inductive loop detectors. Finally, the estimated vehicle speeds are transferred to the proactive traffic control center to forecast future traffic flow conditions and perform traffic controls.

Based on the following $N_D$ pieces of captured past traffic flow data illustrated in Eq. (1), the traffic flow predictor can be developed to forecast traffic flow conditions with $p$ sample time ahead at the location of $D_L$.

$$d(i) = [\theta(i), \phi(i)] \quad \text{with } i = 1, 2, \ldots N_D, \tag{1}$$

where the $\theta(i)$ is the $i$th future traffic flow data, which is the average speed of vehicles collected from the $L$th detection station at the time $(t(i) + mT_s)$; $\theta(i)$ is given by

$$\theta(i) = s_L(t(i) + mT_s);$$

the past traffic flow data, $\varphi(i)$, collected from the $n$ detection stations, is given as:

$$\phi(i) = \begin{bmatrix} h_1(t(i) - T_s), h_1(t(i) - 2T_s), \ldots, h_1(t(i) - pT_s), \\ h_2(t(i) - T_s), h_2(t(i) - 2T_s), \ldots, h_2(t(i) - pT_s), \ldots, \\ h_n(t(i) - T_s), h_n(t(i) - 2T_s), \ldots, h_n(t(i) - pT_s), \\ s_1(t(i) - T_s), s_1(t(i) - 2T_s), \ldots, s_1(t(i) - pT_s), \\ s_2(t(i) - T_s), s_2(t(i) - 2T_s), \ldots, s_2(t(i) - pT_s), \ldots, \\ s_n(t(i) + T_s), s_n(t(i) - 2T_s), \ldots, s_n(t(i) - pT_s) \end{bmatrix}; \quad (2)$$

$h_j(t(i) - kT_s)$ and $s_j(t(i) - kT_s)$ are the average headway distance between cars and the average speed of cars collected by $D_j$ respectively at time $(t(i) - kT_s)$ with $j = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, p$.

Based on $d(i)$ with $i = 1, 2, \ldots N_D$, the accuracy of the traffic flow predictor can be evaluated based on the mean absolute relative error ($e_{MARE}$), which is formulated as:

$$e_{MARE} = \frac{1}{N_D} \frac{\sum_{i=1}^{N_D} \left| \theta(i) - \hat{\theta}(i) \right|}{\theta(i)}, \quad (3)$$

where $\theta(i)$ is the $i$th true collected future traffic flow data; $\hat{\theta}(i)$ is the $i$th predicted future traffic flow data which is given as

$$\hat{\theta}(i) = \hat{s}_L(t(i) + mT_s); \quad (4)$$

and $\hat{s}_L(t(i) + mT_s)$ is determined based on traffic flow predictor discussed in Sect. 2.3.

## 2.2 Pre-processing Data Using Exponent Smoothing

The goodness-of-fit of the traffic flow predictor increases when the training error, $e_{MARE}$, decreases. If $e_{MARE} = 0$, the traffic flow predictor can fit wholly all the collected traffic flow data, and also all the characteristics of the collected traffic flow data can be completely included too. For example, we consider the traffic flow data regarding the average speeds of vehicles collected from a detector station installed on Mitchell Freeway before the on-ramp of Reid Highway (illustrated in Fig. 1). These traffic flow data (shown in Fig. 2) was collected over the 2-h peak traffic periods (7.30–9.30 am) on the five working days from 15 to 19 December 2008. The speed of each vehicle passing through the detector station was recorded. The sampling time was 60 s (or 1 min).

The traffic flow predictor can be obtained by fitting all the captured traffic flow data that is lumpy. However, these lumpy characteristics are not useful for prediction of future traffic flow conditions, and training with these characteristics may overtrain the traffic flow predictor. An overtrained traffic flow predictor can achieve a very

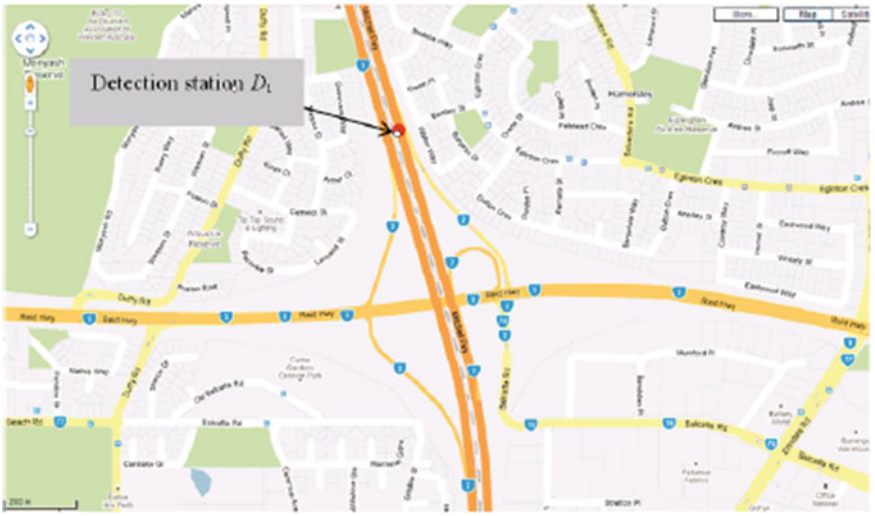



**Fig. 1** Location of the detection station, $D_1$

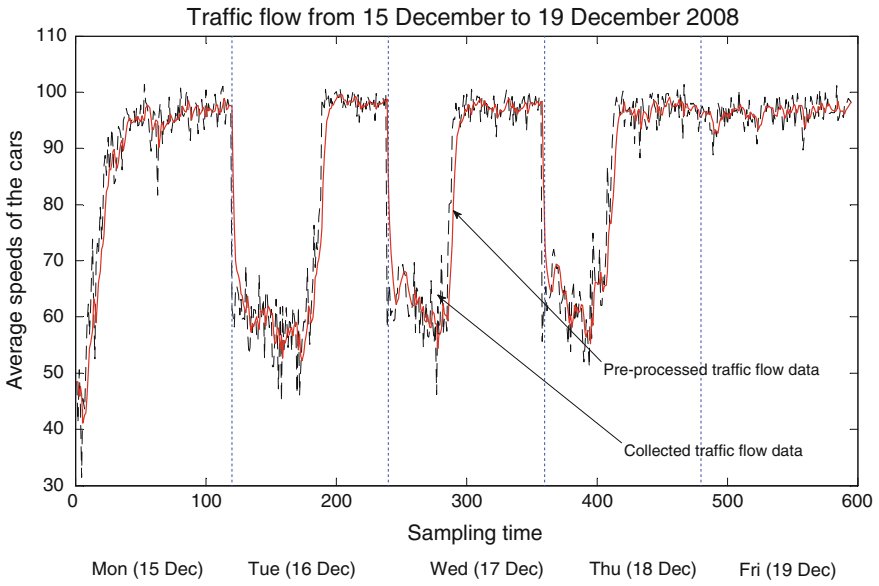



**Fig. 2** Traffic flow data collected from 15th to 19th December 2008

small $e_{MARE}$ (or even zero $e_{MARE}$), but good generalization capability with respect to traffic flow predictions cannot be produced on untrained patterns.

To avoid developing an overtrained traffic flow predictor, this lumpy characteristic is necessary to be filtered from the original traffic flow data before developing the traffic flow predictor. Therefore, here the exponential smoothing method, namely EXP, is used to remove the lumpiness on the traffic flow data [2].

In the EXP, the $i$th pre-processed traffic flow data $\theta'(i)$ is generated using the $(i-1)$th pre-processed traffic flow data, $\theta'(i-1)$, and the $(i-1)$th original captured traffic flow data, $\theta(i-1)$, while the error $\left(\theta(i-1)-\theta'(i-1)\right)$, is integrated for the pre-processing. Hence, the $i$th pre-processed traffic flow data, $\theta'(i)$, is given as:

$$\theta'(i) = \theta'(i-1) + \alpha\left(\theta(i-1) - \theta'(i-1)\right) \tag{5}$$

where $\alpha$ is the smoothing constant within the range, $0 < \alpha \leq 1$, and $i \geq 3$. The first and the second pre-processed traffic flow data are defined as $\theta'(1) = \theta(1)$, and

$$\theta'(2) = \frac{1}{3}\sum_{i=1}^{3}\theta(i) \tag{6}$$

respectively.

When $\alpha$ is large, the pre-processed traffic flow data $\theta'(i)$ varies quickly, and it contains more lumpy characteristics of the traffic flow data. When $\alpha$ is small, $\theta'(i)$ varies slowly, and it contains less lumpy characteristics of the traffic flow data. Here grid search with increments of $(0.8/N_G)$ of the parameter range between $\alpha = 0.1$ and $\alpha = 0.9$ is used in order to estimate the appropriate $\alpha$, where $N_G$ is the searching grid size. An appropriate $\alpha$ is chosen in order to minimize the sum of squares for the residuals which is given as:

$$R^2(\alpha) = \sum_{i=1}^{N_D}\left(\theta'(i)\big|_{\alpha} - \theta(i)\right)^2 \tag{7}$$

Figure 2 also shows the pre-processed traffic flow data used by the EXP. It shows that the pre-processed traffic data seeks to filter out the lumpiness due to irregular variation on the captured traffic flow data. It could reduce the forecasting performance of the traffic flow predictor. Therefore, the traffic flow predictor, which has better forecasting performance, is more likely to be developed, when the pre-processed traffic flow data excluding lumpiness is used.

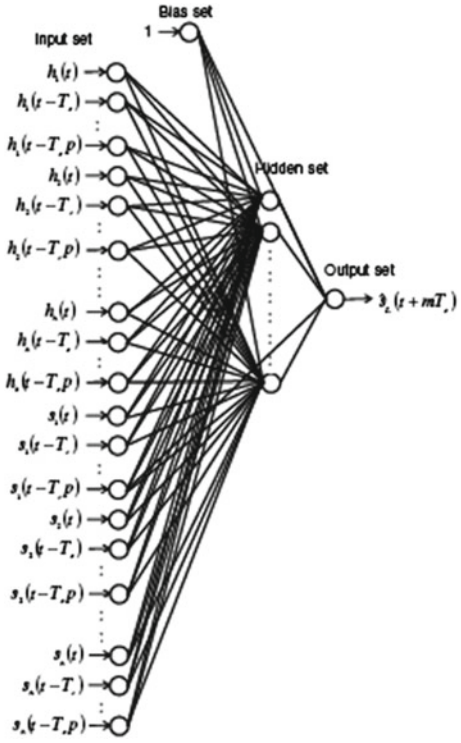



**Fig. 3** The configuration of the simple three-layers-neural-network namely NN for traffic flow predictions

## 2.3 *Development of Traffic Flow Predictors Using EXP-BP Algorithm*

Here the commonly used simple three-layers-neural-network namely NN, which has commonly been used on developing traffic flow predictor [4, 7], is used, where the NN consists of three layers and four sets, input set, bias set, hidden set and output set, and is shown in Fig. 3. The NN is formulated as follows:

$$\hat{s}_L\left(t+mT_s\right)=\sum_{k=1}^{n}\sum_{j=1}^{M}\left[\beta_{j,k}^{h}\Psi\left(\gamma_{0,j,k}^{h}+\sum_{i=1}^{p}\gamma_{i,j,k}^{h}h_k\left(t-iT_s\right)\right)+\beta_{j,k}^{s}\Psi\left(\gamma_{0,j,k}^{s}+\sum_{i=1}^{p}\gamma_{i,j,k}^{s}s_k\left(t-iT_s\right)\right)\right]+\alpha_0 \tag{8}$$

where $M$ is the number of nodes in the hidden set, and $\alpha_0$, $\beta_{i,k}^{h}$, $\beta_{i,k}^{s}$, $\gamma_{0,j,k}^{h}$, $\gamma_{0,j,k}^{s}$, $\gamma_{i,j,k}^{s}$ and $\gamma_{i,j,k}^{h}$ are the NN parameters; $\alpha_0$ denotes the bias of the output set; $\beta_{i,k}^{h}$ and $\beta_{i,k}^{s}$ denote the weights on the connections from the hidden set to the output set; $\gamma_{0,j,k}^{h}$ and $\gamma_{0,j,k}^{s}$ denote the biases of the hidden set; $\gamma_{i,j,k}^{h}$ and $\gamma_{i,j,k}^{s}$ denote the weights

on the connections from the input set to the hidden set; and $\Psi$ (.) is the activation function of the hidden set in which sigmoid and hyperbolic tangent functions are the two commonly used functions.

A training method, namely hybrid exponential smoothing and back-propagation algorithm (EXP-BP), is proposed to determine the NN parameter, where EXP-BP uses the traffic flow data pre-processed by EXP to train the NN and the back-propagation algorithm (BP), a commonly used method [6], is used to determine optimal NN parameters for short-term traffic flow forecasting. In the EXP-BP, the BP is based on the least mean squares algorithm to determine the NN parameters by minimizing the mean absolute relative error $e_{MARE}$. The EXP-BP starts by randomly generating the first two initial guesses of NN parameters $w(0)$ and $w(1)$ at the 0th and the 1-st iterations, where

$$w(0) = \begin{bmatrix} \beta^h_{j,k}(0), \beta^s_{j,k}(0), \gamma^h_{0,j,k}(0), \gamma^s_{0,j,k}(0), \\ \gamma^h_{i,j,k}(0), \gamma^s_{i,j,k}(0) \end{bmatrix} \tag{9}$$

and

$$w(1) = \begin{bmatrix} \beta^h_{j,k}(1), \beta^s_{j,k}(1), \gamma^h_{0,j,k}(1), \gamma^s_{0,j,k}(1), \\ \gamma^h_{i,j,k}(1), \gamma^s_{i,j,k}(1) \end{bmatrix} \tag{10}$$

with $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, m$, and $k = 1, 2, \ldots, n$ respectively.

Then the BP algorithm then changes the NN parameters at the $(l + 1)$th iteration based on the following formulation:

$$w(l+1) = w(l) - \eta_1 \frac{\partial(e_{MARE}(l))}{\partial(w(l))} + \eta_2[w(l) - w(l-1)] \tag{11}$$

where $w(l)$ and $w(l - 1)$ are the NN parameters at the $l$th iteration and the $(l + 1)$th iteration respectively; $e_{MARE}(l)$ is the mean absolute relative error, which is determined based on Eq. (3) with respect to the NN parameters $w(l)$; $\eta_1$ is the learning rate (0.01–1.0); and $\eta_2$ is the momentum coefficient (usually 0.9). The accuracy of the NN for traffic flow forecasting is evaluated based on the mean absolute relative error ($e_{MARE}$), formulated in (3). The EXP-BP algorithm keeps updating the NN parameters, until the pre-defined number of iterations is reached or $e_{MARE}(l)$ is smaller than the pre-defined satisfactory value.

## 3 Evaluations of Traffic Flow Predictors

In this section, the performance of using EXP-BP algorithm in developing NNs for traffic flow predictions is evaluated using the traffic flow data collected from a section of the freeway in Western Australia, Australia. Comparisons between the EXP-BP

**Table 1** Details for the four traffic flow data sets used in this research

| Dates of data collections | Data collected from the intersection of Reid Highway |
|---|---|
| Week 38 in 2008 (15 Sep. 2008–19 Sep. 2008) | Reid-2008-38 |
| Week 41 in 2008 (6 Oct. 2008–10 Oct. 2008) | Reid-2008-41 |
| Week 52 in 2008 (22 Dec. 2008–24 Dec. 2008) | Reid-2008-52 |
| Week 02 in 2009 (5 Jan. 2008–9 Jan. 2009) | Reid-2009-02 |


**Fig. 4** The detection stations, $D_1$ and $D_2$, located in Reid Highway, Western Australia

and the other algorithms, which involve mechanisms for pre-processing traffic flow data in order to avoiding development of overtrained NNs, are carried out.

## *3.1 Collected Traffic Flow Data Sets*

Here the NNs were developed based on four traffic flow data sets illustrated by Table 1, in which the dates and the locations for collecting the traffic flow data are shown. These traffic flow data was collected from weeks 38, 41 and 52 in 2008, and week 2 in 2009. These four traffic flow data sets (namely Reid-2008-38, Reid-2008-41, Reid-2008-52, and Reid-2009-02) were captured by the two detection stations located at the Reid Highway and Mitchell Freeway intersection, Western Australia (Fig. 4) in which the two detection stations were located near the on-ramp and off-ramp in order to capture the traffic flow data.

### 3.2 Comparison with Other Algorithms

To evaluate the effectiveness of the EXP-BP algorithm, the following three algorithms have also been used and the results obtained have been used for comparing with those obtained by the EXP-BP algorithm:

1. **Standard BP algorithm, namely S-BP**: It is similar to EXP-BP but no data pre-processing approach is used.
2. **The hybrid simple moving and BP algorithm namely SM-BP**: It uses the approach of simple moving method to remove the lumpiness on traffic flow data before generating the NN models by using the BP algorithm. The mechanisms of SM-BP are similar to EXP-BP, but SM-BP uses the simple moving method to remove lumpiness on traffic flow data. In the SM-BP, the $i$th pre-processed traffic flow data, $\theta'(i)$ is produced by using the previous four traffic flow data as:

$$\theta'(i) = \frac{1}{4}\sum_{i=1}^{5}(\theta(i-1) + \theta(i-2) + \theta(i-3) + \theta(i-4)) \tag{12}$$

   with $i$>4, where $\theta'(1) = \theta(1), \theta'(2) = \theta(2), \theta'(3) = \theta(3)$ and $\theta'(4) = \theta(4)$.
3. **The hybrid weight moving and BP algorithm, namely WM-BP**: It uses the approach of weight moving method to remove lumpiness in the traffic flow data. The mechanisms of WM-BP are similar to EXP-BP, but WM-BP uses the approach of weighted moving method to remove the lumpiness in traffic flow data before using the BP algorithm to develop the NN models. In the WM-BP, the $i$th pre-processed traffic flow data, $\theta'(i)$, is produced by using the previous four traffic flow data as:

$$\theta'(i) = \frac{1}{10}\sum_{i=1}^{5}(4\cdot\theta(i-1) + 3\cdot\theta(i-2) + 2\cdot\theta(i-3) + \theta(i-4)) \tag{13}$$

   with $i$>4, where $\theta'(1) = \theta(1), \theta'(2) = \theta(2), \theta'(3) = \theta(3)$ and $\theta'(4) = \theta(4)$.

The parameters applied in the four algorithms, EXP-BP, S-BP, SM-BP, and WM-BP are given as following: the number of hidden nodes used on the NN models is $\log_2(480) \approx 9$, where the number of pieces of traffic flow data used for training, $N_D$, is 480. The value of $\log_2(N_D)$ is the number of hidden nodes recommended by Wanas et al. [15]. The termination iteration is pre-defined as 100. Hence, we stopped training the NN models, if the termination iteration is reached, or $e_{MARE}$ is smaller than 0.01.

All the four algorithms, EXP-BP, S-BP, SM-BP, and WM-BP, were run for 30 times with different initial NN parameters, and we recorded the results for the 30 runs. Table 2 shows the mean training errors and variances regarding training errors among the 30 runs of the four BP algorithms, which were used on developing $NN_2^{\text{Reid}}$ and $NN_6^{\text{Reid}}$ for the four traffic flow data sets (namely Week 38-2008, Week 41-2008,

**Table 2** Training error obtained for Reid Highway based on EXP-BP, SM-BP, M-BP, and S-BP

| | | | EXP-BP | SM-BP | WM-BP | S-BP |
|---|---|---|---|---|---|---|
| $NN_2^{Reid}$ | Reid-2008-38 | Mean error | 9.7543 | 9.5445 | 9.2090 | 6.5009 |
| | | Vari. of errors | 0.2619 | 0.3650 | 0.4200 | 1.3381 |
| | | Rank | 4 | 3 | 2 | 1 |
| | Reid-2008-41 | Mean error | 6.1899 | 6.2172 | 5.7726 | 4.5109 |
| | | Vari. of errors | 0.1420 | 0.1730 | 0.2127 | 1.1024 |
| | | Rank | 3 | 4 | 2 | 1 |
| | Reid-2008-51 | Mean error | 6.1558 | 5.9347 | 5.8585 | 4.3318 |
| | | Vari. of errors | 0.3079 | 0.1972 | 0.3535 | 1.3017 |
| | | Rank | 4 | 3 | 2 | 1 |
| | Reid-2009-02 | Mean error | 4.1706 | 4.1334 | 4.5578 | 2.8366 |
| | | Vari. of errors | 0.0389 | 0.0512 | 0.0606 | 0.0841 |
| | | Rank | 3 | 2 | 4 | 1 |
| $NN_6^{Reid}$ | Reid-2008-38 | Mean error | 11.9142 | 11.1502 | 11.6467 | 8.9893 |
| | | Vari. of errors | 1.9150 | 0.7736 | 1.9364 | 4.6056 |
| | | Rank | 4 | 2 | 3 | 1 |
| | Reid-2008-41 | Mean error | 7.3997 | 7.1603 | 7.0732 | 5.5208 |
| | | Vari. of errors | 0.8808 | 1.4714 | 1.4714 | 6.3988 |
| | | Rank | 4 | 3 | 2 | 1 |
| | Reid-2008-51 | Mean error | 6.5627 | 6.6012 | 6.3010 | 4.1315 |
| | | Vari. of errors | 0.3372 | 0.8577 | 0.7166 | 0.5966 |
| | | Rank | 3 | 4 | 2 | 1 |
| | Reid-2009-02 | Mean error | 5.0080 | 5.3499 | 4.7681 | 4.3961 |
| | | Vari. of errors | 0.3159 | 0.5061 | 0.5165 | 4.3884 |
| | | Rank | 3 | 4 | 2 | 1 |
| *Average of mean errors* | | | 8.5477 | 8.3125 | 8.1825 | 6.2098 |
| *Average of variances of mean errors* | | | 0.5456 | 0.7100 | 0.8187 | 2.5486 |
| *Average of ranks of mean error* | | | 3.5 | 3.125 | 2.375 | 1 |

Week 51-2008, and Week 02-2009). It also shows the ranks of mean training errors regarding the four tested BP algorithms.

The results show that averages of mean training errors obtained by the S-BP, of which lumpiness in traffic flow data was not removed, are less than those of the training errors obtained by the other three tested algorithms EXP-BP, SM-BP and WM-BP, which removed lumpiness in traffic flow data before generating the NN models. We can also see that the EXP-BP obtained the largest average of mean training errors and the highest average rank of mean training errors compared with the other tested three algorithms. Therefore, we can conclude that EXP-BP algorithm has the poorest capability in fitting the original collected traffic flow data.

Then, we evaluated the generalization capability of the NN models developed by the four tested algorithms, where we used the other sets of test data which was not used for training the NN models. Table 3 shows the mean test errors and variances of test errors with respect to the 30 runs. We can find that the NN models developed

**Table 3** Testing error obtained for Reid Highway based on EXP-BP, SM-BP, WM-BP, and S-BP

| | | | EXP-BP | SM-BP | WM-BP | S-BP |
|---|---|---|---|---|---|---|
| $NN_2^{\text{Reid}}$ | Reid-2008-38 | Mean error | 10.9876 | 11.2729 | 15.9876 | 12.7144 |
| | | Vari. of errors | 49.5302 | 69.3616 | 67.8201 | 25.2742 |
| | | Rank | 1 | 2 | 4 | 3 |
| | Reid-2008-41 | Mean error | 10.5322 | 15.1073 | 15.2113 | 17.0728 |
| | | Vari. of errors | 16.8746 | 73.8550 | 54.4759 | 80.0044 |
| | | Rank | 1 | 2 | 3 | 4 |
| | Reid-2008-51 | Mean error | 4.6879 | 4.5160 | 5.0011 | 6.9143 |
| | | Vari. of errors | 1.7023 | 2.3104 | 2.3971 | 7.9011 |
| | | Rank | 2 | 1 | 3 | 4 |
| | Reid-2009-02 | Mean error | 2.1034 | 2.2154 | 2.0086 | 3.0448 |
| | | Vari. of errors | 0.0662 | 0.0395 | 0.0860 | 0.1594 |
| | | Rank | 2 | 3 | 1 | 4 |
| $NN_6^{\text{Reid}}$ | Reid-2008-38 | Mean error | 12.5178 | 11.5339 | 11.7540 | 14.8721 |
| | | Vari. of errors | 46.0617 | 39.5638 | 33.2793 | 48.6260 |
| | | Rank | 3 | 1 | 2 | 4 |
| | Reid-2008-41 | Mean error | 11.3453 | 12.0229 | 12.6142 | 16.3917 |
| | | Vari. of errors | 5.4937 | 12.3180 | 8.4307 | 13.1221 |
| | | Rank | 1 | 2 | 3 | 4 |
| | Reid-2008-51 | Mean error | 6.0566 | 6.6409 | 6.8802 | 8.6046 |
| | | Vari. of errors | 11.6981 | 10.2548 | 12.1908 | 72.2594 |
| | | Rank | 1 | 2 | 3 | 4 |
| | Reid-2009-02 | Mean error | 2.5190 | 2.5503 | 2.9752 | 4.2524 |
| | | Vari. of errors | 0.1046 | 0.3586 | 0.2757 | 1.7328 |
| | | Rank | 1 | 2 | 3 | 4 |
| *Average of mean errors* | | | 8.3850 | 9.2063 | 10.8500 | 11.3180 |
| *Average of variances of mean errors* | | | 19.4885 | 37.2550 | 33.0223 | 27.9226 |
| *Average of ranks of mean errors* | | | 1.5 | 1.875 | 2.75 | 3.875 |

by EXP-BP can obtain the smallest average of mean test errors and the best average of mean ranks compared with those obtained by the other three tested algorithms. Also, we can find that the average variances of test errors obtained by NN models, which were developed by EXP-BP, are the smallest.

Based on these experimental results regarding the test data, we can conclude that EXP-BP algorithm can find the NN models with the best generalization capability when compared with the NN models obtained by the other three tested BP algorithms (S-BP, SM-BP and WM-BP). Also, these experimental results show that EXP-BP algorithm can generally develop more robust NN models than the other three algorithms, because the average of variances of mean errors is the smallest compared with the other three tested algorithms.

Therefore, the test errors obtained by the EXP-BP algorithm are generally the smallest among all in generating the $NN_2^{\text{Reid}}$ and $NN_6^{\text{Reid}}$ models for predicting the traffic flow conditions in Reid Highway, Western Australia, and the training errors

obtained by the EXP-BP algorithm are the highest among all NN models developed by the other three tested algorithms. This indicates that the EXP-BP algorithm avoids generating overtrained NN models, and it can generate NN models which have better generalization capability than the other three tested algorithms.

## 4 Conclusion and Further Works

Based on our observation, when lumpiness of the traffic flow data is included in training the traffic flow predictors, small training errors for the traffic flow predictors can be obtained by fitting tightly with all the lumpiness, but it may create the poor generalization capability with respect to traffic flow predictions, as the lumpiness may not be helpful in generating a more accuracy traffic flow predictor. This chapter intends to present a training method namely EXP-BP algorithm to generate traffic flow predictors. For the presented EXP-BP algorithm, the approach of exponential smoothing is first employed to filter lumpiness from original traffic flow data before using the BP algorithm for training the NN models. Experimental results show that training errors of the traffic flow predictors obtained by EXP-BP are higher than those obtained by the other tested algorithms, but test errors are smaller than those obtained by the other three tested algorithms. Therefore, traffic flow predictors with better generalization capabilities for traffic flow predictions can be obtained by using the EXP-BP algorithm. The following works are currently underway:

1. Further evaluation of the effectiveness of the EXP-BP algorithm is carried out by comparing with other forecasting methods like fuzzy neural networks and statistical regressions. Also incorporation with the other existing approaches [1, 15] on avoiding overtraining will also be carried out.
2. The filtering approach will be further evaluated by applying on different traffic flow conditions including smoothing traffic conditions, traffic congestions and traffic contingences.
3. The mechanism in adjusting the exponential smoothing parameter is investigated in order to control the filtering strength on handling traffic flow data with different levels of lumpiness.
4. Apart from the currently filtering approach, the approach on identifying significant neural network nodes is investigated in order to create the neural networks with more significant structures. By doing so, overtrained neural networks are less likely to be generated.

**Acknowledgments** The first author wishes to thank Main Roads, Western Australia for providing the traffic flow data for this research. He also wishes to express his sincere thanks to Tharam Dillon, Elizabeth Chang and Jaipal Singh for many useful suggestion and valuable discussions. He would also like to further acknowledge their very practical comments for this research.

## References


1. Amari S, Murata N, Muller K, Finke M, Yang HH (1997) Asymptotic statistical theory of overtraining and cross validation. IEEE Trans Neural Networks 8(5):958–996
2. Brown RG, Meyer RF (1961) The fundamental theorem of exponential smoothing. Oper Res 9:673–685
3. Davis GA, Nihan NL (1991) Nonparametric regression and short-term freeway traffic forecasting. J Transp Eng 177(2):178–188
4. Dia H (2001) An object-oriented neural network approach to short-term traffic forecasting. Eur J Oper Res 131:253–261
5. Dillon TS, Sestito S, Leung S (1991) Short term load forecasting using an adaptive neural network. Electr Power Energy Syst 13(4):186–192
6. Innamaa S (2006) Effect of monitoring system structure on short-term prediction of highway travel time. Transp Planning Technol 29(2):125–140
7. Ledoux C (1997) An urban traffic flow model integrating neural network. Transp Res 5C:287–300
8. Lilien GL, Kotler P (1983) Marketing decision making: a model building approach. Harper and Row Publishers, New York
9. Mastorocostas PA, Theocharis JB (2003) An orthogonal least-squares method for recurrent fuzzy-neural modeling. Fuzzy Sets Syst 140:285–300
10. Okutani I, Stephanedes YJ (1984) Dynamic prediction of traffic volume through Kalman filtering theory. Transp Res Part B Methodol 18(1):1–11
11. Quek C, Pasquier M, Lim BBS (2006) POP-TRAFFIC: a novel fuzzy neural approach to road traffic analysis and prediction. IEEE Trans Intell Transp Syst 7(2):133–146
12. Smith BL, Williams BM, Oswald RK (2002) Comparision of parametric and nonparametric models for traffic flow forecasting. Transp Res Part C 19:303–321
13. Stathopoulos A, Dimitriou L, Tskeris T (2008) Fuzzy modeling approach for combined forecasting of urban traffic flow. Comput Aided Civ Infrastruct Eng 23:521–535
14. Tan MC, Wong SC, Xu JM, Guan ZR, Zhang P (2009) An aggregation approach to short term traffic flow prediction. IEEE Trans Intell Transp syst 10(1):60–69
15. Wanas N, Auda G, Kamel MS, Karray F (1998) On the optimal number of hidden nodes in a neural network. IEEE Can Conf Electr Comput Eng 2:918–921
16. Yaffee RA, McGee M (2000) Introduction to time series analysis and forecasting. Academic, San Diego
17. Yin H, Wong SC, Xu J, Wong CK (2002) Urban traffic flow prediction using a fuzzy-neural approach. Transp Res Part C 10:85–98
18. Zhang GP, Kline DM (2007) Quarterly time series forecasting with neural networks. IEEE Trans Neural Networks 18(6):1800–1814

# Economic Scheduling of CCHP Systems Considering the Tradable Green Certificates

**Hongming Yang, Dangqiang Zhang, Ke Meng, Mingyong Lai and Zhao Yang Dong**

**Abstract** Due to the fossil fuel crisis contributed by the explosive growth in energy demand, combined cooling heating and power (CCHP) systems which can jointly supply electricity and hot/cold have become the mainstream of energy generation technology. In this chapter, tradable green certificate mechanism is firstly introduced to operation of CCHP system, and the impacts of tradable green certificate on the scheduling of CCHP system are studied. And then, based on the probability distribution of wind speed and solar radiation intensity as well as the copula join function, the joint probability distribution of maximum available output of multiple solar and wind farms is built. The economic dispatch model for multi-energy complementary system considering the TGC was proposed to maximize renewable energy utilization. This model aims at minimizing total system cost whist fulfilling the constraints of power system stable operation and hot/cold water pipes safe operation. After that, in order to address the non-convex scheduling optimization problem, global descent method is applied, which can continuously update the local optimal solutions by global descent

H. Yang (✉) · D. Zhang
School of Electrical Engineering and Information, Changsha University of Science and Technology, Changsha 410114, China
e-mail: yhm5218@hotmail.com

D. Zhang
e-mail: qqzdq@hotmail.com

K. Meng · Z. Y. Dong
Centre for Intelligent Electricity Networks, The University of Newcastle, NSW 2308, Australia
e-mail: ke.meng@newcastle.edu.au

Z. Y. Dong
e-mail: joe.dong@newcastle.edu.au

M. Lai
Key Laboratory of Logistics Information and Simulation Technology, Human University, Changsha 410082, China
e-mail: laimingyong0731@hotmail.com

H. Xu and X. Wang (eds.), *Optimization and Control Methods in Industrial Engineering and Construction*, Intelligent Systems, Control and Automation: Science and Engineering 72,
DOI: 10.1007/978-94-017-8044-5_9, © Springer Science+Business Media Dordrecht 2014

function, and find global optimal solution. Finally, one modified IEEE 14-bus system is used to verify the performance of the proposed model and optimization solver.

**Keywords** Combined cooling heating and power system · Copula function · Economic scheduling · Global descent method · Renewable energy · Tradable green certificate

## 1 Introduction

Nowadays, due to the fossil fuel crisis contributed by the explosive growth in energy demand, energy saving solutions has attracted widespread concerns. In order to reduce energy consumption, combined cooling heating and power (CCHP) systems have been widely deployed around the world. CCHP units can generate electricity whilst recovering thermal energy that normally would be wasted in an electricity generator, and then uses it to produce steam, hot water, space heating, or cooling [1]. By using a CCHP system, the fuel that would otherwise be used to produce heat or steam in a separate unit is saved. The most significant benefits of CCHP systems can be summarized as, (a) supply base-load power meanwhile cover cooling heating requirements during hot and cold seasons; (b) offer efficient and economical solution for emission reduction; and (c) reduce operating cost and life-cycle cost. In order to assist with the management of CCHP systems, extensive researches have been conducted to formulate operational strategies. A hybrid electric-thermal load operational strategy was proposed in [2], which was a good alternative to the operation of CCHP system since it can yield good reductions of primary energy consumption, operational cost, and carbon dioxide emissions. An efficient algorithm was proposed to optimize the operation of a CCHP gas-motor-based system. The results indicate that optimal operation of CCHP system under the objective function of investment on power plant and equipment can be controlled [3]. In [4], a novel optimal operational strategy depending on an integrated performance criterion (IPC) was proposed, which can divide the whole operating space of CCHP system into several regions. Then the operating point of the CCHP system is located in a corresponding operating mode region to achieve improved IPC. In [5], uncertainties in CCHP system, such as the thermal load, natural gas and electricity prices, and engine performance were characterized in a representative steady-state model. Moreover, some optimization methods have been developed for the scheduling of CCHP systems. In [6], a multi-objective approach based on evolutionary programming was used to solve the economic operation of combined heating and power (CHP) systems under emission reductions. Genetic algorithm was proposed as an optimal operation strategy for a CHP system to improve its competitiveness in electricity market in [7]. In [8], particle swarm optimization was applied to schedule the optimal operation of CHP systems, which was formulated as a mixed-integer nonlinear optimization problem.

However, the conventional CCHP systems are all built on fossil fuels, including coal, oil, and natural gas, which are the main source of greenhouse gases (GHGs).

The increasing environmental challenges have forced power generation enterprises to modify their system operation routines to reduce GHGs emissions by exploiting clean energy [9]. Along with the introduction of various emission reduction schemes, increasing number of renewable energy projects have been constructed around the world [10]. Meanwhile, to promote the sustainable development of renewable energy, a series of environmental policies have been introduced, such as government subsidies under quota system [11] and tradable green certificates (TGC) policy [12, 13]. However, due to the stochastic characteristics, renewable energy brings great challenges to CCHP system economic scheduling problems. Economic scheduling aims to allocate power generation to match load demand and minimize total operational cost while satisfying all the power units and system constraints [14]. Better scheduling strategies normally can provide effective solutions to improve the current situation of system operation and reduce carbon emissions dramatically. After the introduction of tradable green certificates, how to coordinately schedule the CCHP units and renewable facilities to produce both electricity power and heat/cold energy while satisfying all the determined and probabilistic constraints becomes more complicated.

While the answer to energy crisis certainly includes alternative energy, a strategy for coordinately scheduling these generating units is required. Moreover, in order to accommodate the revised scheduling strategy, more efficient solvers should be developed. The motivation of this is to take full advantage of various performance criteria and to improve the performance of multi-energy CCHP systems. This chapter is organized as follows, after introduction section a mathematical model of multi-energy CCHP system is proposed, and the impacts of tradable green certificate on the CCHP unit scheduling are studied. Then, an economic scheduling model is proposed, which considers the uncertainties of renewable energy and tradable green certificate price, and the constraints of secure operation of electricity network and pipe network. After that, global descent method (GDM) is applied to solve the optimization problem and one modified IEEE 14 bus system is used to verify the performance of the proposed model and optimization solver. Conclusions and further developments are discussed in the last section.

## 2 Mathematical Model of Multi-Energy CCHP System Considering Green Certificates Trading

### *2.1 Models of Multi-Energy CCHP Systems*

Multi-energy CCHP systems are composed of coal, wind, solar, and other primary energy sources, which can produce electricity power and hot/cold energy simultaneously. The input-output model is shown in Fig. 1. The produced electricity power and hot/cold energy can be connected to system through electricity network and hot

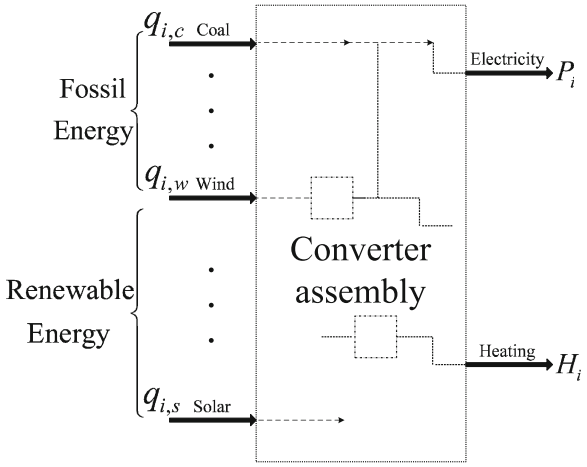



**Fig. 1** Multi-energy CCHP system

(cold) water pipe network, in order to meet the electrical load and demand for heating and cooling.

In Fig. 1, in an economic scheduling period, the exchange relationship of conventional fossil energy can be expressed as,

$$P_{i,fo} = \frac{1}{\omega_d}\eta_{i,fo,e}\upsilon_{i,fo,e}q_{i,fo}, \tag{1}$$

$$H_{i,fo} = \eta_{i,fo,h}\upsilon_{i,fo,h}q_{i,fo}, \tag{2}$$

where $i \in \{1, 2, \cdots N\}$ denotes the $i$ th CCHP unit; $N$ is the total number of CCHP units in the system; the subscript $fo$ represents the fossil fuel; $\omega_d$ is time interval [h]; $q_{i,fo}$ is the amount of consumed fossil fuel [kg] by CCHP unit during time interval $\omega_d$; $\eta$ is conversion efficiency; $P_{i,fo}$ and $H_{i,fo}$ represent the produced electricity power [MW] and heat/cold energy [KJ] by CCHP unit; $\upsilon_{i,fo,e}$ and $\upsilon_{i,fo,h}$ represent the ratio of fossil fuel used for producing electricity power and heat/cold energy, called scheduling factor. For $\forall i$, $fo$,

$$\upsilon_{i,fo,e} + \upsilon_{i,fo,h} = 1. \tag{3}$$

The exchange relationship between maximum renewable power generation and primary energy can be expressed as,

$$P_{i,re}^{\max} = f_{i,re,P}\left(\upsilon_{i,re,e}q_{i,re}\right), \tag{4}$$

$$H_{i,re}^{\max} = f_{i,re,H}\left(\upsilon_{i,re,h}q_{i,re}\right), \tag{5}$$

where subscript $re$ represents renewable energy; $P_{i,re}^{\max}$ and $H_{i,re}^{\max}$ are the maximum produced renewable electricity power [MW] and heat/cold energy [KJ] by CCHP unit; $q_{i,re}$ is the consumed renewable energy source, such as wind, solar, and etc.; $f_{i,re,P}(\cdot)$ and $f_{i,re,H}(\cdot)$ is function of maximum renewable electricity power and heat/cold energy and primary energy; $\upsilon_{i,re,e}$ and $\upsilon_{i,re,h}$ is scheduling factor. For $\forall i, re$,

$$\upsilon_{i,re,e} + \upsilon_{i,re,h} = 1. \tag{6}$$

## 2.2 Models of Solar-CCHP Unit and Wind-CCHP Unit

### 2.2.1 Solar-CCHP Model

The maximum electricity power and heat/cold energy produced by solar power can be expressed as [15–18],

$$P_{i,s}^{\max} = f_{i,s,P}\left(\upsilon_{i,s,e} q_{i,s}\right) = k_{i,se}\upsilon_{i,s,e} q_{i,s}, \tag{7}$$

$$H_{i,s}^{\max} = f_{i,s,H}\left(\upsilon_{i,s,e} q_{i,s}\right) = k_{i,sh}\upsilon_{i,s,h} q_{i,s}, \tag{8}$$

where $k_{i,se} = A_{i,se}\eta_{i,se}[\mathrm{MW}/(\mathrm{W}/\mathrm{m}^2)]$ and $k_{i,sh} = A_{i,sh}\eta_{i,sh}[\mathrm{kJ}/(\mathrm{W}/\mathrm{m}^2)]$ are linear coefficients; $A_{i,se}$ and $A_{i,sh}$ are areas of PV array [$\mathrm{m}^2$], which are used to produce electricity power and heat/cold energy, respectively; $\eta_{i,se}$ and $\eta_{i,sh}$ is the conversion efficiency; $q_{i,s}$ is the solar radiation intensity [$\mathrm{kJ/m}^2 \cdot \mathrm{s}$].

An extensive of research has shown that Beta distribution is the most accurate density function that can be used to describe solar radiation intensity frequency curve [19, 20]. The empirical and theoretical distributions of solar radiation intensity percentage (solar radiation intensity divided by maximum solar radiation intensity) are shown in Fig. 2.

Therefore, from Eqs. (7), (8), the probability density function (PDF) of maximum electricity power and heat/cold energy generation by solar power can be calculated as,

$$\begin{aligned}\rho_{i,se}\left(\frac{P_{i,s}^{\max}}{P_{i,s,r}}\right) &= \rho_{i,sh}\left(\frac{H_{i,s}^{\max}}{H_{i,s,r}}\right) = \rho_{i,s}\left(\frac{q_{i,s}}{q_{i,s,r}}\right) \\ &= \frac{1}{B(\alpha_i,\beta_i)}\left(\frac{q_{i,s}}{q_{i,s,r}}\right)^{\alpha_i-1}\left(1-\frac{q_{i,s}}{q_{i,s,r}}\right)^{\beta_i-1},\end{aligned} \tag{9}$$

where $\rho_{i,se}(\cdot)$, $\rho_{i,sh}(\cdot)$, and $\rho_{i,I}(\cdot)$ are PDFs of electricity power and heat/cold energy generated by solar CCHP, and PDF of solar radiation intensity, respectively; $P_{i,s,r}$ [MW] and $H_{i,s,r}$ [kJ] are the rated electricity power and heat/cold energy; $q_{i,s,r}$ is the rated solar radiation intensity; $\alpha_i$ and $\beta_i$ are coefficients which can be estimated by maximum likelihood method; $B(\alpha_i, \beta_i)$ is normalized coefficients, which can be represented as,

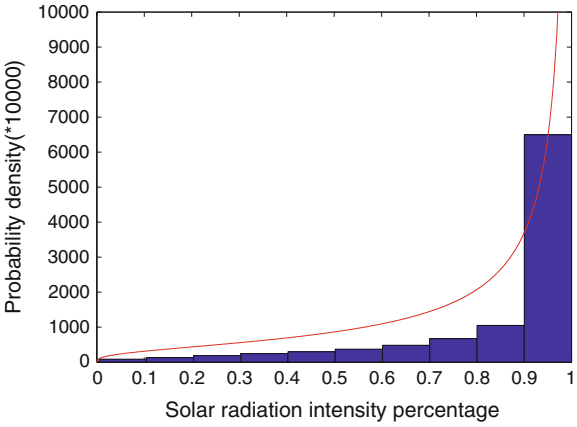



**Fig. 2** Empirical and theoretical distribution of solar radiation intensity

$$B\left(\alpha_i, \beta_i\right) = \int_0^1 x^{\alpha_i - 1} \left(1 - x\right)^{\beta_i - 1} dx = \frac{\Gamma\left(\alpha_i\right) \Gamma\left(\beta_i\right)}{\Gamma\left(\alpha_i + \beta_i\right)}. \tag{10}$$

Because of similar weather conditions, the maximum available power generation of different solar farms may show certain correlation. To model this correlation, the Copula function can be employed. Assume that the marginal distributions of maximum available outputs of N solar farms are $F_1\left(P_{1,s}^{\max}\right), \cdots, F\left(P_{N,s}^{\max}\right)$ respectively, and then there exists a Copula function $C_s\left(\cdot\right)$ such that the joint distribution $\Re_s\left(\cdot\right)$ of multiple solar farms outputs can be expressed as [21, 22]:

$$\Re_s\left(P_{1,s}^{\max}, \cdots, P_{N,s}^{\max}\right) = C_s\left(F_1\left(P_{1,s}^{\max}\right), \cdots, F_N\left(P_{N,s}^{\max}\right)\right). \tag{11}$$

There exist several different types of Copula functions. Since the Gumbel-Copula function is unsymmetrical and upper fat-tailed, which well matches the characteristics of solar power correlation [23]; it is employed to model the joint distribution of maximum available outputs of multiple solar farms:

$$\begin{aligned} &\Re_s\left(P_{1,s}^{\max}, \cdots, P_{N,s}^{\max}\right) = \\ &\exp -\left[\left(-InF_1\left(P_{1,s}^{\max}\right)\right)^{\varsigma_s} + \cdots + \left(-InF_N\left(P_{1,s}^{\max}\right)\right)^{\varsigma_s}\right]^{\frac{1}{\varsigma_s}} \end{aligned} \tag{12}$$

where $\varsigma_s$ is the parameter of the Copula function and can be estimated using the MLE method.

### 2.2.2 Wind-CCHP Model

Due to the low conversion efficiency from wind energy into heat/cold energy, in this chapter, only the electricity power generated by wind CCHP is considered. The maximum output power of the wind turbine can be expressed as a function of wind speed,

$$P_{i,w}^{\max} = f_{i,w,P}\left(\upsilon_{i,w,e} q_{i,w}\right) = \begin{cases} 0, & q_{i,w} < q_{i,w,in}, q_{i,w} > q_{i,w,out}, \\ k_{i,w}\upsilon_{i,w,e} q_{i,w} + d_{i,w}, & q_{i,w,in} \le q_{i,w} \le q_{i,w,out}, \\ P_{i,w,r}, & q_{i,w,r} \le q_{i,w} \le q_{i,w,out}. \end{cases} \tag{13}$$

where $q_{i,w}$ is wind speed [m/s]; $q_{i,w,in}, q_{i,w,r}, q_{i,w,out}$ are cut-in, rated, and cut-out wind speeds [m/s]; $\upsilon_{i,w,e}$ is scheduling factor, for $\forall i$, $\upsilon_{i,w,e} = 1$, $P_{i,w,r}$ is the rated electricity power of wind unit [MW]; $k_{i,w}$ and $d_{i,w}$ are the liner coefficients with satisfying,

$$k_{i,w} = \frac{P_{i,w,r}}{q_{i,w,r} - q_{i,w,in}}, \tag{14}$$

$$d_{i,w} = -\frac{q_{i,w,in} P_{i,w,r}}{q_{i,w,r} - q_{i,w,in}}. \tag{15}$$

An extensive review of various PDFS of wind speed was provided in [24, 25], and comparisons were made. The results indicated that Rayleigh or Weibull distribution is the widely accepted model. In this, it is assumed that the wind speed follows Rayleigh distribution, the PDF is,

$$\rho_{i,w}\left(q_{i,w}\right) = \frac{q_{i,w}}{\lambda_i^2} \exp\left(-\frac{q_{i,w}^2}{2\lambda_i^2}\right), \tag{16}$$

where $\rho_{i,w}(\cdot)$ is the PDF of wind speed; $\lambda_i$ is distribution factor, which meets the following relationship [26],

$$\lambda_i = q_{i,w,f}/\sqrt{\pi/2}. \tag{17}$$

The empirical and theoretical distributions of wind speed are shown in Fig. 3. It shows that the probability distribution of wind speed can be described as Rayleigh distribution.

According to Eq. (13), three portions of wind power output can be analyzed and the corresponding PDF can be calculated based on the wind turbine power output curve and wind speed PDF, respectively.

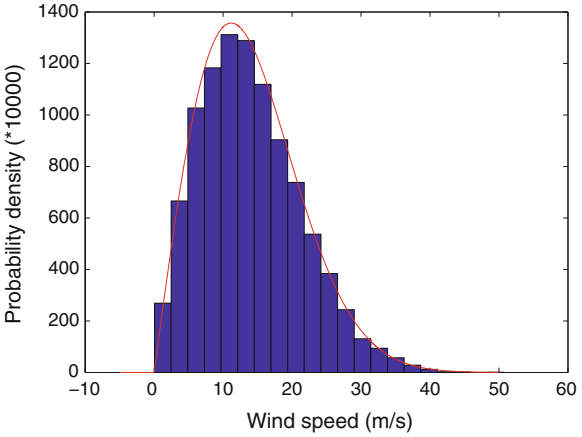



**Fig. 3** Empirical and theoretical distribution of wind speed

(1) For $0 < P_{i,w}^{\max} < P_{i,w,r}$,

$$\rho_{i,w}\left(P_{i,w}^{\max}\right) = \frac{P_{i,w}^{\max} - d_{i,w}}{\lambda_i^2 k_{i,w}^2} \exp\left[\frac{\left(P_{i,w}^{\max} - d_{i,w}\right)^2}{-2\lambda_i^2 k_{i,w}^2}\right]; \tag{18}$$

(2) For $P_{i,w}^{\max} = 0$,

$$\rho_{i,w}\left(P_{i,w}^{\max}\right) = \left[1 - \exp\left(-\frac{q_{i,w,in}^2}{2\lambda_i^2}\right) + \exp\left(-\frac{q_{i,w,out}^2}{2\lambda_i^2}\right)\right]\delta\left(P_{i,w}^{\max}\right); \tag{19}$$

(3) For $P_{i,w}^{\max} = P_{i,w,r}$,

$$\rho_{i,w}\left(P_{i,w}^{\max}\right) = \left[\exp\left(-\frac{q_{i,w,r}^2}{2\lambda_i^2}\right) - \exp\left(-\frac{q_{i,w,out}^2}{2\lambda_i^2}\right)\right]\delta\left(P_{i,w}^{\max} - P_{i,w,r}\right), \tag{20}$$

where $\delta(\cdot)$ is Dirac Delta function.

Similarly, the joint distribution of maximum available outputs of multiple wind turbines is:

$$\begin{aligned}&\Re_w\left(P_{1,w}^{\max}, \cdots, P_{N,w}^{\max}\right) = C_w\left(F_1\left(P_{1,w}^{\max}\right), \cdots F_N\left(P_{N,w}^{\max}\right)\right) = \\ &\exp\left\{-\left[\left(-InF_1\left(P_{1,w}^{\max}\right)\right)^{\varsigma_w} + \cdots + \left(-InF_N\left(P_{1,w}^{\max}\right)\right)^{\varsigma_w}\right]^{\frac{1}{\varsigma_w}}\right\}\end{aligned} \tag{21}$$

where $\varsigma_w$ is the parameter of the Copula function.

### 2.2.3 Impacts of Green Certificates on CCHP Systems

Green certificates, issued by governments, represent a certain amount of renewable energy products that has been produced. When trading green certificates in the markets, the holders may sell the surplus green certificates, and the buyers can purchase green certificates to meet their quota requirements. In this chapter, we assume that there are sufficient green certificates in the market, which can be freely traded, not constrained by the number of transactions.

The benefits from green certificates trading can be calculated as,

$$\pi_{tgc,i} = \varpi_{tgc} \cdot Q_{tgc,i}, \tag{22}$$

where $\varpi_{tgc}$ is the trading price, and $Q_{tgc,i}$ is the volume of green certificates traded in one clearing interval. Assume that the equal levels of electricity and heat/cold are obtained through economic dispatch in one clearing period, and then the volume of traded certificates can be represented as:

$$\begin{aligned} Q_{tgc,i} = &\left(\omega_T \sum_{re} \kappa_{re,e} P_{i,re} + \omega_T \sum_{re} \kappa_{re,h} H_{i,re}\right) - \\ &\gamma_e \left(\omega_T \sum_{f0} P_{i,fo} + \omega_T \sum_{re} P_{i,re}\right) - \gamma_h \left(\omega_T \sum_{fo} H_{i,fo} + \omega_T \sum_{re} H_{i,re}\right), \end{aligned} \tag{23}$$

where, $\omega_T$ is the clearing interval of green certificates [h]; $\gamma_e$ and $\gamma_h$ are the renewable energy quotas of electricity power and heat/cold energy, respectively; $\kappa_{re,e}$ and $\kappa_{re,h}$ are the number of green certificates obtained in producing one unit of electricity power and heat/cold energy by renewables; $Q_{tgc,i} > 0$ means selling and $Q_{tgc,c} < 0$ means purchasing.

Suppose the green certificate market is oligopolistic competition, the linear relationship between the transaction prices and demand for green certificates is [27],

$$\varpi_{tgc} = K_0 + K_1 \left[ Q_{tgc,0} - \sum_i Q_{tgc,i} \right], \tag{24}$$

where $K_0$ and $K_1$ are constants; $Q_{tgc,0}$ is the number of green certificates purchased by other system (department).

Due to the uncertainties in producing electricity power and heat/cold energy by renewables in different districts, in order to meet the requirements of renewable energy quotas and customer demand, $Q_{tgc,0}$ shows uncertain patterns. Here we assume that, $Q_{tgc,0}$ follows the normal distribution with mean $\mu_0$ and variance $\sigma_0$. The PDF is shown as follows,

$$\rho_0\left(Q_{tgc,0}\right)=\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left[-\frac{\left(Q_{tgc,o}-\mu_0\right)^2}{2\sigma_0^2}\right]. \tag{25}$$

## 3 Economic Scheduling Model of Multi-Energy CCHP Systems Considering Green Certificates Trading

In the multi-energy CCHP system economic scheduling model, the decision variables are coal-fired electricity generation, coal-fired heat/cold energy generation, electricity generated by solar, heat/cold energy generated by solar, and electricity generated by wind, considering the impacts of green certificate trading mechanism. The objective is to minimize the total system cost and maximize the penetration of renewable energy under the constraints of power network and heat/cold water pipe networks security operation. Assume that the external environmental factors, e.g. load and weather condition, are the same during the studied economic dispatch interval, based on the DC power flow, the economic dispatch model for multi-energy CCHP system is discussed in the following subsections.

### *3.1 Objectives*

The objective function of the economic scheduling model is composed of production costs of fossil generators and renewable machines $\Pi_1$, overestimate/under-estimate costs of renewable energy $\Pi_2$, and the costs of green certificates $\Pi_3$,

$$\underset{P_{i,re},P_{i,fo},H_{i,re},H_{i,fo}}{Min} \quad E\left[\Pi\right]=E\left[\Pi_1\right]+E\left[\Pi_2\right]+E\left[\Pi_3\right], \tag{26}$$

where $E\left[\cdot\right]$ is the expectation; $\Pi$ is the total system cost. the decision variables $P_{i,fo}$, $H_{i,fo}$ and $P_{i,re}$, $H_{i,re}$, are the output of thermal units and renewable machines.

#### 3.1.1 Production Costs

Production costs, including fuel costs $\Pi_{11}$, operation and maintenance costs $\Pi_{12}$. In this chapter, start-up and shunt-down costs of conventional units are ignored, and the fuel cost of renewable energy is set as zero. $\Pi_{11}$ and $\Pi_{12}$ can be calculated as follows,

$$\Pi_{11}=\omega_T\sum_i\sum_{fo}\left[\begin{array}{l}a_{i,fo}\left(\frac{H_{i,fo}+\vartheta P_{i,fo}}{\eta_{i,fo,h}\upsilon_{i,fo,h}+\vartheta\eta_{i,fo,e}\upsilon_{i,fo,e}}\right)+\\ b_{i,fo}\left(\frac{H_{i,fo}+\vartheta P_{i,fo}}{\eta_{i,fo,h}\upsilon_{i,fo,h}+\vartheta\eta_{i,fo,e}\upsilon_{i,fo,e}}\right)^2\end{array}\right], \tag{27}$$

$$\Pi_{12} = \omega_T \sum_i \left[ \sum_{fo} \left( c_{i,fo,e} P_{i,fo} + c_{i,fo,h} H_{i,fo} \right) + \sum_{re} \left( c_{i,re,e} P_{i,re} + c_{i,re,h} P_{i,rh} \right) \right], \tag{28}$$

where $\vartheta$ ($\vartheta = 3.6\,\mathrm{kJ/MWh}$) is the electrothermal equivalent coefficient; $a_{i,fo}$ and $b_{i,fo}$ are the coefficients of fossil fuel cost; $c_{i,fo}$ and $c_{i,re}$ are the coefficients of operation and maintenance cost of generating units.

### 3.1.2 Overestimate/Underestimate Costs

Due to the uncertainty of renewable energy, the predictions normally have some errors. The underestimation and overestimation penalty cost are introduced to maximize the penetration of renewable energy. The underestimation situation occurs if the actual generated power is greater than the predicted (i.e. the available wind power has not been fully utilized), thus the system operator should compensate for the surplus power cost. On the other hand, if the actual power is smaller than the scheduled power (i.e. the available wind power is insufficient), the operator needs to purchase power from an alternate source and pay the overestimation penalty cost. These two penalty costs are assumed as follows,

$$\Pi_2 = \omega_T \sum_i \sum_{re} \begin{pmatrix} c^u_{i,re,e} \left[ P^{\max}_{i,re} - P_{i,re} \right]^+ + c^o_{i,re,e} \left[ P_{i,re} - P^{\max}_{i,re} \right]^+ + \\ c^u_{i,re,h} \left[ H^{\max}_{i,re} - H_{i,re} \right]^+ + c^o_{i,re,h} \left[ H_{i,re} - H^{\max}_{i,re} \right]^+ \end{pmatrix}, \tag{29}$$

where $c^u_{i,re}$ and $c^o_{i,re}$ are underestimate and overestimate coefficients, and the operator $[x]^+ = \max\{x, 0\}$.

When considering a single solar farm and a single wind farm or multiple solar farms and wind farms with independent probabilistic characteristics, based on the probability density functions of maximum electricity and heat/cold energy generation by solar power and electricity generation by wind power, the analytical expressions of penalty cost expectation can be obtained:

- The underestimate penalty cost expectation of electricity power generated by solar is,

$$E\left( c^u_{i,se} \left[ P^{\max}_{i,s} - P_{i,s} \right]^+ \right) = \frac{P_{i,s,r} c^u_{i,se}}{B(\alpha_i, \beta_i)} \left[ \begin{array}{l} B(\alpha_i + 1, \beta_i) - B_{\wp_i}(\alpha_i + 1, \beta_i) - \\ \frac{P_{i,s}}{P_{i,s,r}} \left[ B(\alpha_i, \beta_i) - B_{\wp_i}(\alpha_i, \beta_i) \right] \end{array} \right], \tag{30}$$

where $B_x(\alpha_i, \beta_i) = \int_0^x y^{\alpha_i - 1} (1 - y)^{\beta_i - 1} \, dy$ is the incomplete beta function; $\wp_i = \frac{P_{i,s}}{P_{i,s,r}}$ is the electricity power percentage generated by solar CCHP.

- The overestimate penalty cost expectation of electricity power generated by solar is,

$$E\left(c_{i,se}^{o}\left[P_{i,s}-P_{i,s}^{\max}\right]^{+}\right)=\frac{P_{i,s,r}c_{i,se}^{o}}{B\left(\alpha_i,\beta_i\right)}\left[\frac{P_{i,s}}{P_{i,s,r}}B_{\wp_i}\left(\alpha_i,\beta_i\right)-B_{\wp_i}\left(\alpha_i+1,\beta_i\right)\right]; \tag{31}$$

- The underestimate penalty cost expectation of heat/cold energy generated by solar is,

$$E\left(c_{i,sh}^{u}\left[H_{i,s}^{\max}-H_{i,s}\right]^{+}\right)=\frac{H_{i,s,r}c_{i,sh}^{u}}{B\left(\alpha_i,\beta_i\right)}\left[\begin{array}{l}B\left(\alpha_i+1,\beta_i\right)-B_{\ell_i}\left(\alpha_i+1,\beta_i\right)\\-\frac{H_{i,s}}{H_{i,s,r}}\left[B\left(\alpha_i,\beta_i\right)-B_{\ell_i}\left(\alpha_i,\beta_i\right)\right]\end{array}\right]; \tag{32}$$

where $\ell_i=\frac{H_{i,s}}{H_{i,s,r}}$ is the heat/cold energy percentage generated by solar CCHP.

- The overestimate penalty cost expectation of heat/cold energy generated by solar is,

$$E\left(c_{i,sh}^{o}\left[H_{i,s}-H_{i,s}^{\max}\right]^{+}\right)=\frac{H_{i,s,r}c_{i,sh}^{o}}{B\left(\alpha_i,\beta_i\right)}\left[\frac{H_{i,s}}{H_{i,s,r}}B_{\ell_i}\left(\alpha_i,\beta_i\right)-B_{\ell_i}\left(\alpha_i+1,\beta_i\right)\right]; \tag{33}$$

- The underestimate penalty cost expectation of electricity power generated by wind is,

$$\begin{aligned}E\left(c_{i,we}^{u}\left[P_{i,w}^{\max}-P_{i,w}\right]^{+}\right)=&\\ &c_{i,we}^{u}\left(P_{i,w,r}-P_{i,w}\right)\left[\exp\left(-\frac{q_{i,w,r}^{2}}{2\lambda_i^{2}}\right)-\exp\left(-\frac{q_{i,w,out}^{2}}{2\lambda_i^{2}}\right)\right]\\ &-c_{i,we}^{u}\left(P_{i,w,r}-P_{i,w}\right)\left[\exp\left(\frac{\left(P_{i,w,r}-d_{i,we}\right)^{2}}{2\lambda_i^{2}k_{i,we}^{2}}\right)\right]\\ &+\frac{\sqrt{2\pi}\lambda_i c_{i,we}^{u}k_{i,we}}{2}\left[erf\left(\frac{P_{i,w,r}-d_{i,we}}{\sqrt{2}\lambda_i k_{i,we}}\right)-erf\left(\frac{P_{i,w}-d_{i,we}}{\sqrt{2}\lambda_i k_{i,we}}\right)\right],\end{aligned} \tag{34}$$

where $erf\left(\cdot\right)$ is Gaussian error function.

- The overestimate penalty cost expectation of electricity power generated by wind is,

$$\begin{aligned}E\left(c_{i,we}^{o}\left[P_{i,w}-P_{i,w}^{\max}\right]^{+}\right)=&\\ &c_{i,we}^{o}P_{i,w}\left[1-\exp\left(-\frac{q_{i,w,in}^{2}}{2\lambda_i^{2}}\right)+\exp\left(-\frac{q_{i,w,out}^{2}}{2\lambda_i^{2}}\right)\right]\\ &-\frac{\sqrt{2\pi}\lambda_i c_{i,we}^{o}k_{i,we}}{2}\left[erf\left(\frac{P_{i,w}-d_{i,we}}{\sqrt{2}\lambda_i k_{i,we}}\right)-erf\left(\frac{-d_{i,we}}{\sqrt{2}\lambda_i k_{i,we}}\right)\right]\\ &+c_{i,we}^{o}P_{i,w}\exp\left(-\frac{d_{i,we}^{2}}{2\lambda_i^{2}k_{i,we}^{2}}\right).\end{aligned} \tag{35}$$

When considering the probability correlation of renewable energy at different nodes subjecting to similar weather conditions, the mathematical form of Copula

function becomes complex and the expectations of overestimation and underestimation penalty costs cannot be expressed analytically. Therefore, the SAA method (29) [28] is applied to transform the stochastic optimization model into a computable deterministic nonlinear programming problem:

$$E\left[\Pi_2\right]=\frac{\omega_T}{M}\sum_i\sum_{re}\sum_{k=1}^{M}\left(\begin{array}{l}c_{i,re,e}^{u}\left[P_{i,re,k}^{\max}-P_{i,re}\right]^{+}+c_{i,re,e}^{o}\left[P_{i,re}-P_{i,re,k}^{\max}\right]^{+}+\\ c_{i,re,h}^{u}\left[H_{i,re,k}^{\max}-H_{i,re,j}\right]^{+}+c_{i,re,h}^{o}\left[H_{i,re}-H_{i,re,k}^{\max}\right]^{+}\end{array}\right), \tag{36}$$

where, $P_{i,re,k}^{\max}H_{i,re,k}^{\max}$ $(i=1,\ldots,N;k=1,\ldots,M)$ is the sampled maximum renewable power generation from Copula joint probability distribution by Monte Carlo simuation; $M$ is Monte Carlo sampling number.

### 3.1.3 Green Certificates Trading Costs

Based on the benefits obtained from green certificates trading, the green certificate transaction costs can be obtained as,

$$\Pi_3=-\sum_i\pi_{tgc,i}=-\sum_i\varpi_{tgc}\cdot Q_{tgc,i}. \tag{37}$$

The trading cost expectation of green certificates is,

$$\begin{aligned}E\left(\sum_i\varpi_{tgc}\cdot Q_{tgc,i}\right)&=\int_{-\infty}^{+\infty}\left[K_0+K_1\left(Q_{tgc,0}-\sum_i Q_{tgc,i}\right)\right]\left(\sum_i Q_{tgc,i}\right)dQ_{tgc,0}\\&=\left[K_0+K_1\left(\mu_0-\sum_i Q_{tgc,i}\right)\right]\left(\sum_i Q_{tgc,i}\right).\end{aligned} \tag{38}$$

## *3.2 Constraints*

According to the actual requirements of secure and economic operation of multi-energy CCHP systems, the system constraints should be considered, including the balance of supply and demand, power network constraints, heat/cold water pipe network constraints, and CCHP unit limits.

(1) The balance of supply and demand

$$\sum_i\left(\sum_{fo}P_{i,fo}+\sum_{re}P_{i,re}\right)=\sum_i L_{i,e}, \tag{39}$$

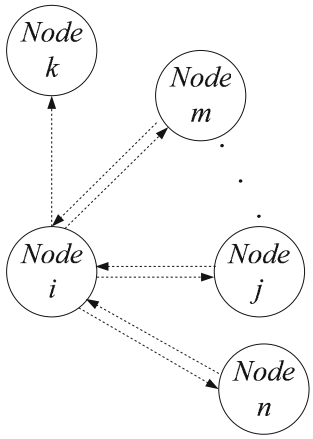



**Fig. 4** Radial hot/cold water pipes network

$$\sum_{i}\left(\sum_{fo}H_{i,fo}+\sum_{re}H_{i,re}\right)=\sum_{i}L_{i,h}, \tag{40}$$

where $L_{i,e}$ [MW] and $L_{i,h}$ [kJ] are the electricity power load and heat/cold demand at $i$ th bus.

(2) Electricity network security constraints

$$-P_{l}^{\max}\leq\sum_{i}\chi_{il}\left(\sum_{fo}P_{i,fo}+\sum_{re}P_{i,re}-L_{i,e}\right)\leq P_{l}^{\max}, \tag{41}$$

where $l=1,2,\cdots,L$ is the transmission line; $L$ is the total number of transmission lines; $\chi_{li}$ is the sensitivity coefficients of injected power at bus $i$ with respect to transmitted power of line $l$; $P_{l}^{\max}$ is the upper limit of transmitted power.

(3) Hot/Cold water pipe network secure operation constraints

In order to simplify this problem, we suppose the hot/cold water pipe network are radial, as shown in Fig. 4. There are two types of transmission situations in hot/cold water pipes, which are bidirectional and unidirectional. For example, there is bidirectional hot/cold transmission between nodes $i$ and $j$; and while unidirectional transmission is occurred if the load node $k$ can only consume heat/cold energy.

According to thermodynamic equation, the fluid temperature increase caused by absorbing heat in unit time can be computed by,

$$\sum_{fo}H_{i,fo}+\sum_{re}H_{i,re}-L_{i,h}=C\rho AV_{i}\,(T_{i}-T_{i0}), \tag{42}$$

where $T_{i0}$ is temperature of water in pipe network before heating or cooling; $T_{i}$ is temperature of water in pipe network after heating or cooling; $C$ is the specific heat

of water; $\rho$ is density; $A$ is the pipeline area; $V_i$ is flow rate.

$$-V_i^{\max} \le \frac{\sum_{fo} H_{i,fo} + \sum_{re} H_{i,re} - L_{i,h}}{C\rho A\,(T_i - T_{i0})} = V_i \le V_i^{\max}, \forall i. \tag{43}$$

(4) Constraints of CCHP units

$$\forall i,\ fo, re,$$

$$\underline{P_{i,fo}} \le P_{i,fo} \le \overline{P_{i,fo}}, \tag{44}$$

$$\underline{P_{i,re}} \le P_{i,re} \le \overline{P_{i,re}}, \tag{45}$$

$$\underline{H_{i,fo}} \le H_{i,fo} \le \overline{H_{i,fo}}, \tag{46}$$

$$\underline{H_{i,re}} \le H_{i,re} \le \overline{H_{i,re}}, \tag{47}$$

$$0 \le \upsilon_{i.fo,e} \le 1, \tag{48}$$

$$0 \le \upsilon_{i.re,e} \le 1, \tag{49}$$

where $\underline{P_{i,fo}}$, $\overline{P_{i,fo}}$ and $\underline{P_{i,re}}$, $\overline{P_{i,re}}$ are the upper and lower electricity power limits of fossil-based and renewable CCHP units; $\underline{H_{i,fo}}$, $\overline{H_{i,fo}}$ and $\underline{H_{i,re}}$, $\overline{H_{i,re}}$ are the upper and lower heat/cold energy limits of fossil-based and renewable CCHP units.

## 4 Global Descent Method Based Economic Scheduling Solution

The economic scheduling model discussed above is a non-convex optimization problem; therefore, some traditional optimization methods (e.g. interior point method) can only get local optimal solutions. In order to obtain the global optimal solution, a global descent method is introduced. For simplicity, the optimization problem can be written as follows,

$$\min_x \quad f(x), \tag{50}$$

$$s.t. \quad h(x) = 0, \underline{g} \le g(x) \le \bar{g}, \tag{51}$$

where $x$ is a vector composed of decision variables $P_{i,fo}$, $H_{i,fo}$ and $P_{i,re}$, $H_{i,re}$; $f(x) = E[\Pi]$ is objective function; $h(x)$ is equality constraint; $g(x)$ is inequality constraint.

In this, the nonlinear primal-dual interior point algorithm is applied to find the local optima $\hat{x}^{(k)}$ of this problem, and then the global descent function is introduced, which can be represented as,

$$G_{\varsigma,\theta,\hat{x}^{(k)}}(x) = A_\varsigma\left[f(x) - f\left(\hat{x}^{(k)}\right)\right] - \theta\left\|x - \hat{x}^{(k)}\right\|, \tag{52}$$

where $\theta > 0, 0 < \varsigma < 1$; $A_\varsigma(y) = yV_\varsigma(y)$; and

$$V_\varsigma\left[f(x) - f\left(\hat{x}^{(k)}\right)\right] = \varsigma\left[(1-r)\left(\frac{\varsigma - r\varsigma}{1 - r\varsigma}\right)^{\frac{\left[f(x) - f\left(\hat{x}^{(k)}\right)\right]}{\tau}} + r\right], \tag{53}$$

where $0 < r < 1$; $\tau > 0$ is an enough small positive number, which meets the following inequality,

$$0 < \tau < \min\left\{\left|f^* - f^{**}\right| : f^*, f^{**} \in f^*_{all}, f^* \neq f^{**}\right\}, \tag{54}$$

where $f^*_{all}$ is local optimal solutions, $f^*$ and $f^{**}$ are the two elements of $f^*_{all}$.

For $\forall x_{cur} \in X \backslash N_\varepsilon\left(\hat{x}^{(k)}\right)$, if the following criterion is met,

$$\left\|\nabla G_{\varsigma,\theta,\hat{x}^{(k)}}(x_{cur})\right\| < k_{small}, \tag{55}$$

or

$$\left(x_{cur} - \hat{x}^{(k)}\right)^T \nabla G_{\varsigma,\theta,\hat{x}^{(k)}}(x_{cur}) \geq 0, \tag{56}$$

where $X$ is the feasible region of decision variable $x$; $N_\varepsilon\left(\hat{x}^{(k)}\right)$ is the $\varepsilon$ neighbor region of $\hat{x}^{(k)}$; and $k_{small}$ is an enough small positive number. And then, by adjusting $\varsigma$, make the inequality meet the following criterion,

$$\begin{cases} \left\|\nabla G_{\varsigma,\theta,\hat{x}^{(k)}}(x_{cur})\right\| \geq k_{small} \\ \left(x_{cur} - \hat{x}^{(k)}\right)^T \nabla G_{\varsigma,\theta,\hat{x}^{(k)}}(x_{cur}) < 0 \end{cases} \tag{57}$$

Based on the global descent function, the adjustment of decision variables can be obtained,

$$\Delta x = -step\nabla G_{\varsigma,\theta,\hat{x}^{(k)}}(x_{cur}). \tag{58}$$

Therefore, the updated decision variable is

$$x^{(j)} = x^{(j-1)} - step\nabla G_{\varsigma,\theta,x^*}(x_{cur}). \tag{59}$$

If new decision variable $x^{(j)}$ meets,

$$f\left(x^{(j)}\right) < f\left(\hat{x}^{(k)}\right), \tag{60}$$

then $x^{(j)}$ is the transitional point of optimization problem. Stating from this point, the nonlinear primal-dual interior point algorithm will be used to search new local optimal solution $\hat{x}^{(k+1)}$, and repeat the above searching process until any $x^{(j)}$ meets,

**Table 1** Parameters of CCHP systems

| Bus | | a | b | c | $c^u_{re}$ | $c^o_{re}$ | $\underline{P}/\underline{H}$ | $\overline{P}/\overline{H}$ | $\eta$ | $\upsilon_{.,e}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Electricity | 20 | 0.043 | 50 | – | – | 0 | 200 | 0.5 | 1 |
| | Heat/Cool | – | – | – | – | – | – | – | – | 0 |
| 2 | Electricity | – | – | 500 | 150 | 150 | 0 | 80 | – | 0.5 |
| | Heat/Cool | – | – | 300 | 150 | 150 | 0 | 60 | – | 0.5 |
| 3 | Electricity | – | – | 200 | 70 | 70 | 0 | 100 | – | 1 |
| | Heat/Cool | – | – | – | – | – | – | – | – | 0 |
| 6 | Electricity | – | – | – | – | – | – | – | – | 0 |
| | Heat/Cool | 13 | 0.012 | 40 | – | – | 0 | 100 | 0.3 | 1 |
| 8 | Electricity | 30 | 0.031 | 45 | – | – | 0 | 150 | 0.7 | 0.6 |
| | Heat/Cool | 30 | 0.031 | 45 | – | – | 0 | 120 | 0.5 | 0.4 |

*Note* When generating electricity, the units of $a, b, c, c^u_{re}, c^o_{re}$ are \$/WM, \$/WM$^2$, \$/WM, \$/WM, \$/WM; when generating heat/cold energy, the units of $a, b, c, c^u_{re}, c^o_{re}$ are \$/kJ, \$/kJ$_2$, \$/kJ, \$/kJ, \$/kJ.

**Table 2** The electricity and heat/cold energy demand

| Bus | Electricity load | Heat/Cold demand | Bus | Electricity load | Heat/Cold demand |
|---|---|---|---|---|---|
| 2 | 21.7 | 10.8 | 10 | 9 | 3.3 |
| 3 | 94.2 | 42.6 | 11 | 3.5 | 2.1 |
| 4 | 47.8 | 25.9 | 12 | 6.1 | 4.6 |
| 5 | 7.6 | 5.2 | 13 | 13.5 | 8 |
| 6 | 11.2 | 8.5 | 14 | 14.9 | 9.7 |
| 9 | 29.5 | 19.3 | – | – | – |

$$f\left(x^{(j)}\right) > f\left(\hat{x}^{(k)}\right). \tag{61}$$

and then the global optimal solution $\hat{x}^{(k)}$ is obtained [29].

## 5 Case Study

In this chapter, one modified IEEE 14-bus system is applied to verify the proposed economic scheduling model and the optimization solver. The benchmark system consists of five generators and eleven loads. Fossil-based CCHP units are located at bus 1, 6, and 8; and a solar-CCHP unit and a wind-CCHP unit are constructed at bus 2 and bus 3, respectively. The fuel cost coefficients, generator limits, locations of generators are shown in Table 1, and the demand at each bus is shown in Table 2.

Take the green certificate trading is Australia as an example [30], suppose $\omega_T = 24$ h. The green certificates issued by government to electricity power and heat/cold energy generated by solar are 5.0/MW and 3.0/KJ, and to electricity power generated by wind is 2.0/MW; the quotas for electricity and heat/cold energy generated by renewable energy

**Table 3** Results Comparisons by DGM and NPDIP

| Bus | GDM | | NPDIP | |
|---|---|---|---|---|
| | Electricity power | Heat/Cold energy | Electricity power | Heat/Cold energy |
| 1 | 63.1129 | – | 68.2132 | – |
| 2 | 31.0091 | 15.5045 | 39.7723 | 19.8861 |
| 3 | 70.1523 | – | 79.2566 | – |
| 6 | – | 79.6011 | – | 85.6577 |
| 8 | 64.8781 | 30.8943 | 59.7580 | 28.4562 |
| Cost ($) | 68,599 | | 72,678 | |

are 1.5/MW and 1.0/KJ. Suppose the green certificates required by other system or department is distributed normally with mean $u = 100$ and variance $\sigma = 13$, and constants $K_0 = 50$, $K_1 = 0.1$.

According to the solar radiation intensity data from LiNuo Solar Energy Group, the parameters of PV array are $k_{se} = 0.8$, $k_{sh} = 0.4$. The two parameters of beta distribution estimated by maximum likelihood method is $\alpha = 1.3535$, $\beta = 0.2253$. For typical wind turbine, the cut-in, rated, and cut-out wind speeds are $v_{in} = 5$, $v_r = 15$, and $v_{out} = 25$, therefore according to Eqs. (10), (11), $k_w = 10$, $d_w = -50$. Suppose the forecasted wind speed used in the study is 14 m/s, and then $\lambda = 11.1735$.

## 5.1 Performance of Global Descent Method

Suppose $\omega_d = 1$ h, the following Table 3 summarizes and compares the economic scheduling results by GDM and NPDIP.

From the results above, we can find that the total system costs optimized by DGM and NPDIP methods are $68,599 and $72,678, respectively. The final solution solved by DGM is much better than the result obtained by NPDIP.

## 5.2 Impacts of Tradable Green Certificates

Suppose the price of green certificate fluctuates in the range of $40–$70/MW, the impacts of price fluctuation on electricity and heat/cold energy generated by CCHP units as well as total system cost are shown in Figs. 5 and 6.

As can be seen from Figs. 5 and 6, when the price of green certificate is low, the solar CCHP unit generates no power. Along with the increase of green certificate price, when it is higher than $50/MW, the output of solar CCHP unit increases rapidly, and total system cost climbs linearly; after it reaches $56/MW, solar CCHP unit generates at its rated power, but total system cost reduces linearly and the output of wind CCHP unit increases. Coal-fired electricity power generation at bus 1 and coal-fired hot/cold energy generation at bus 6 reduce gradually. Therefore, according to the actual situation,

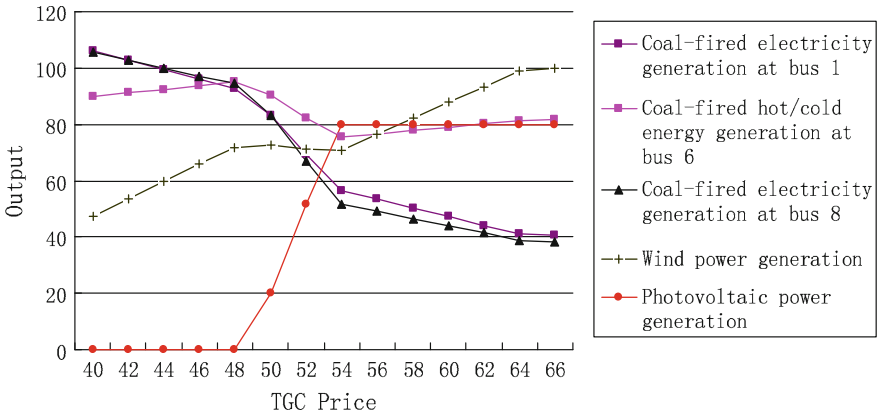



**Fig. 5** Impacts on output of CCHP units

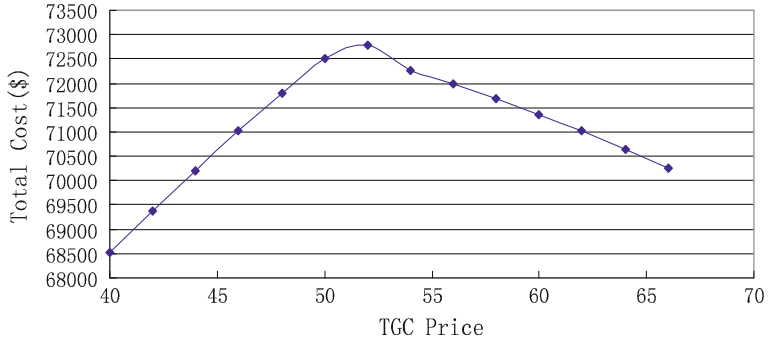



**Fig. 6** Impacts on total system cost

by selecting suitable green certificate price, the total system cost can be reduced, and the utilization of renewable energy can be promoted.

## *5.3 Impacts of Renewable Energy Generation Quotas*

Suppose $(\gamma_e, \gamma_h)$ is the quotas for electricity power and heat/cold energy generation by renewable energy, the impacts of quotas on the outputs of renewable machines as well as total system cost are shown in Table 4.

As can be seen from Table 4, along with the renewable quota increases, the wind turbine output shows small variations, but the output of PV array increase largely from 45.8350 under the quota of (1.0,1.0) to 80.0000 under the quota of (2.0, 1.0), and the total system cost increases to $79,199 from $65,906. Therefore, according to the actual situation, choosing suitable renewable quota is conducive to the utilization of renewable

**Table 4** The impacts of renewable quotas on economic scheduling results

| $(\gamma_e, \gamma_h)$ | Wind power generation | Solar power | Solar heating/cooling | Total cost ($) |
|---|---|---|---|---|
| (1.0, 1.0) | 71.5518 | 45.8350 | 22.9175 | 65,906 |
| (1.5, 1.0) | 71.2566 | 64.7723 | 32.3861 | 72,678 |
| (2.0, 1.0) | 77.8187 | 80.0000 | 40.0000 | 79,199 |
| (1.5, 0.5) | 71.4091 | 54.5396 | 27.2698 | 68,995 |
| (1.5, 1.5) | 70.7793 | 80.0000 | 40.0000 | 76,002 |

energy and also benefits to the optimal allocation of resources and total system cost reduction.

## 6 Conclusions

In this chapter, tradable green certificate policy is introduced to stimulate the utilization of renewable energy, and then mathematical models of multi-energy CCHP units considering green certificate impacts are analyzed. After that, the economic scheduling model for multi-energy CCHP system is proposed. Based on the PDFs of electricity power and heat/cold energy generated by solar, the maximum output of wind turbine, and the amount of green certificates required by other systems (departments), the objective function is developed, and the global descent method is used to solve this optimization problem. Finally, one modified IEEE 14 bus system is used to verify the performance of the proposed scheduling model and optimization solver. The following conclusions can be drawn from the case study,

(1) The global descent method is an effective solver for the economic scheduling of multi-energy CCHP system;
(2) By selecting suitable green certificate price and renewable quotas, the optimal schedule can be determined and the flexibility can be improved;
(3) Green certificate trading policy can stimulate the development and utilization of renewable energy, and is conducive to the optimal allocation of resources and system cost reduction.

However, there are still other uncertainties in the economic scheduling problem of actual multi-energy CCHP system. How to incorporate other uncertain factors (such as electricity and hot/cold demand) in the scheduling model is one of the key research directions in the further.

**Acknowledgments** This work was supported by the National Natural Science Foundation of China (Outstanding Youth Project 70925006, Key Project 71331001, General Project 71071025).

## References


1. Combined heat and power—effective energy solutions for a sustainable future (2011). Oak Ridge National Laboratory, Oak Ridge
2. Mago PJ, Chamra LM (2009) Analysis and optimization of CCHP systems based on energy, economical, and environmental considerations. Energy Build 41:1099–1106
3. MacGregor PR, Puttgen HB (1989) Optimum scheduling procedure for cogenerating small power producing facilities. IEEE Trans Power Syst 4(3):957–964
4. Yanagi M, Uekusa T, Yamada J, Shinjo K (1999) Optimal design of cogeneration systems by using Hamiltonian algorithm. In: Proceedings of the building simulation, vol 2, p 699–706
5. Fang F, Wang QH, Yang S (2012) A novel optimal operational strategy for the CCHP system based on two operating modes. IEEE Trans Power Syst 27(2):032–1041
6. Tsay MT (2002) The operation strategy of cogeneration systems using a muliobjective approach. IEEE Trans Power Syst 3:1653–1657
7. Huang SH, Chen BK, Chu WC, Lee WJ (2004) Optimal operation strategy for cogeneration power plants. In: Proceedings of the IEEE transactions industry applications conference 3:2057–2062
8. Tsukada T, Tamura T, Kitagawa S, Fukuyama Y (2003) Optimal operational planning for cogeneration system using particle swarm optimization. In: Proceedingsof the IEEE swarm intelligence Symposium p 138–143
9. Clean energy future—appendix A: carbon pricing mechanism. Available www.cleanenergyfuture.gov.au. Accessed 1 July 2012
10. Dong ZY, Wong KP, Meng K, Luo FJ, Yao F, Zhao JH (2010) Wind power impact on system operations and planning. IEEE PES Gen Meet Minneapolis, USA
11. Tamas MM, Bade Shrestha SO, Zhou HZ (2010) Feed-in tariff and tradable green certificate in oligopoly. Eng Policy 38:4040–4047
12. Hasani-Marzooni M, Hosseini SH (2012) Dynamic interactions of TGC and electricity markets to promote wind capacity investment. IEEE Syst J 6(1):46–57
13. Linares P, Santos FJ, Ventosa M, Lapiedra L (2008) Incorporating oligopoly, CO2 emissions trading and green certificates into a power generation expansion model. Automatica 44:1608–1620
14. Yao F, Dong ZY, Meng K, Xu Z, Iu H, Wong KP (2012) Quantum-inspired particle swarm optimization for power system operations considering wind power uncertainty and carbon tax in Australia. IEEE Trans Industr Inf 8(4):880–888
15. Daut M, Irwanto YM, Irwan N, Gomesh R, Ahmad NS (2011) Potential of solar radiation and wind speed for photovoltaic and wind power hybrid generation. In: 5th international power engineering and optimization conference (PEOCO 2011), Perlis, Northern Malaysia, p 148–153
16. Attaviriyanupap P, Tokuhara K, Itaya N, Marmiroli M, Tsukamoto Y, Kojima Y (2011) Estimation of photovoltaic power generation output based on solar irradiation and frequency classification. In: Proceedings of the IEEE PES Innovative Smart Grid Technologis Asia (ISGT 2011) p. 1–7
17. Abbas A, Marie-Joseph I, Linguet L, Clergeot H (2006) Improvement of individual solar heater efficiency. Environ Identities Mediterr Area 2006:133–138
18. Uchida Kousuke, Senjyu Tomonobu, Urasaki Naomitsu, Yona Atsushi (2009) Installation effect by solar heater system using solar radiation forecasting. IEEE Transm Distrib Asia 2009:1–4
19. Youcef Ettoumi F, Mefti A, Adane A, Bouroubi MY (2001) Statistical analysis of solar measurements in Algeria using beta distributions. Renewable Energy 26 (2002), p 47–67
20. Soubdhan T, Emilion R, Calif R (2009) Classification of daily solar radiation distributions using a mixture of Dirichlet distributions. Sol Energy 83(7):1056–1063
21. Hofert M (2008) Sampling archimedean copulas. Comput Stat Data Anal 52(12):5163–5174
22. Yu L, Voit EO (2006) Construction of bivariate s-distributions with copulas. Comput Stat Data Anal 51(3):1822–1839

23. Br N (2005) An introduction to copulas. New York, Springer, p 7–48
24. Celik ANA (2003) Astatistical analysis of wind power density based on the Weibull and Rayleigh models at the southern region of Turkey. Renewbale, Energy p 593–604
25. Hetzer J, Yu DC, Bhattarai K (2008) An economic dispatch model incorporating wind power. IEEE Trans Energy Convers 23(2):603–611
26. Zhao J, Wen F, Zhao YD, Yusheng X, Kit PW (2012) Optimal dispatch of electric vehicles and wind power using enhanced particle swarm optimization. IEEE Trans Ind Inf 8(4):889–899
27. Pedro L, Francisco JS, Mariano V, Luis L (2008) Incorporating oligopoly, CO2 emissions trading and green certificates into a power generation expansion model. Automatica 44(6):1608–1620
28. Jirutitijaroen P, Singh C (2008) Reliability constrained multi-area adequacy planning using stochastic programming with sample-average approximations. IEEE Trans Power Syst 23(2):504–513
29. Chi-kong N, Duan L, Lian-sheng Z (2010) Global descent method for global optimization. Soc Ind Appl Math 20(6):3161–3184
30. Australian REC and STC prices (Renewable Energy Certificates) (2011). http://localpower.net.au/recs.htm. Accessed Sept 2011

# Optimizations in Project Scheduling: A State-of-Art Survey

**Changzhi Wu, Xiangyu Wang and Lin Jiang**

**Abstract** Project scheduling is concerned with an optimal allocation of resources to activities realized over time. To survive in today's competitive environment, efficient scheduling for project development becomes more and more important. The classical project scheduling is based on the critical path method (CPM) in which resources required are assumed unlimited. This is however impractical. To overcome CPM's drawback, several techniques and optimizations have been proposed in project scheduling literature. In this chapter, we will present a state-of-art survey on project scheduling from the optimization point of view. In particularly, we will focus on the advancements of optimization formulations and solutions on project scheduling in the recent years.

## 1 Introduction

A project is informally defined as a unique undertaking, composed of a set of precedence related tasks that have to be executed using diverse and mostly limited company resources. Project scheduling consists of deciding when tasks should be started and finished, and how many resources should be allocated to them.

For more than two decades the industry has been going through an intense period of introspection as a result of its poor performance and low productivity. In the 1990s numerous reports recommended actions that need to be undertaken to address the industry's prevailing problems [1]. Since the commencement of the global finical

C. Wu (✉) · X. Wang
School of Built Environment, Curtin University, Perth, WA 6845, Australia
e-mail: changzhiwu@yahoo.com

X. Wang
e-mail: Xiangyu.Wang@curtin.edu.au

L. Jiang
School of Mathematics, Anhui Normal University, Wuhu 241000, China

H. Xu and X. Wang (eds.), *Optimization and Control Methods in Industrial Engineering and Construction*, Intelligent Systems, Control and Automation: Science and Engineering 72,
DOI: 10.1007/978-94-017-8044-5_10, © Springer Science+Business Media Dordrecht 2014

crisis in 2008, construction industries worldwide have been subjected to significant reductions in private and public sector investment. Hitherto, issues associated with poor project performance and productivity remains a pervasive problem [2]. In Australia, the Queensland Department of Main Roads reported that 10 % of projects with a contract value greater than AU$1 m experienced an overrun of over 10 % [3]. Blake Waldron found that less than 48 % of Australian infrastructure projects surveyed were delivered on time, on budget and to the required quality [4]. A survey exploring the completion of construction projects in Saudi Arabia showed that 76 % of project contractors experienced delays of 10–30 % of the projected duration [5].

To improve the productivity and reduce the delay, various efforts have been applied to improve scheduling in construction projects. The earliest work on this issue dates back to 1950s. In the late of 1950s, the critical path method (CPM) was developed as a result of the increasing complexity of construction projects. Since then, CPM has been regarded as one of basic project scheduling and control tools for supporting project managers to ensure project completion on time and on budget. In the Critical Path Method, each activity is listed, linked to another activity, and assigned durations. Interdependency of an activity is added as either predecessors or successors to another activity. Moreover, the duration of the activities are entered. Based on the dependency and duration of the activities, the longest path is defined as the most critical path. In CPM, resource limitation is not considered and an activity can always start as long as all its predecessors are completed. This, however, is not practical, as resources are not unlimited and the availability of resources would affect resource allocation and project scheduling. On the other hand, CPM does not allow interruption of an activity or overlap between two connected activities, which is also unpractical [6]. In practice, an activity could be temporarily interrupted due to short-term transfer (e.g., one or two days) of resource(s) to a more important or urgent activity. Slight overlap between two connected activities might happen to repetitive construction projects, such as road pavement projects, etc. Furthermore, based on resource availability, the duration of an activity might vary, which results in various execution modes [6]. To overcome CPM major limitations, several techniques and optimizations have been proposed in project management and scheduling literature and usually can be classified in four categories: resource-constrained scheduling [7], time cost trade-off, resource levelling and resource allocation.

For the predefined durations and demands on each of the given renewable resources, the objective of resource-constrained project scheduling is to determine the sequence of project activities and to allocate available resources to project activities in an attempt to optimize a given objective function such as minimizing project makespan. The objective of time-cost trade off analysis is to find a time/cost trade-off curve showing the relationship between activity duration and cost. The objective of resource levelling is to smooth day-to-day resource demand. The objective of resource allocation is to allocate limited resources to activities so as to optimize a certain goal such as cost minimization. In the past decades, there are many optimization-based methods developed to solve these classes of project scheduling problems. Although many survey papers available in this area [8–10], there are still

many significant advances not included. The aim of this paper is two-fold. On one hand, it services as a tutorial for both practical engineers and beginners in this area. On the other hand, it will offer a comprehensive survey of the optimization methods developed in the recent results.

## 2 Variants of Project Scheduling Problems

### *2.1 Resource-Constrained Project Scheduling (RCPS)*

To formulate a project scheduling as an optimization problem, we first need to model a project in mathematics. In general, there are two ways to represent the project topology, i.e., activity-on-arc (AOA) network and activity-on-node (AON) network. Since almost all the available literatures on project scheduling are based on AON network representation, we here only introduce AON network.

Consider a project with $J$ activities which are labelled as $j = 1, \cdots, J$. Let the activity 0 represent by convention the start of the schedule and the activity $J + 1$ represent the end of the schedule. Denote $V = \{0, 1, \cdots, J, J + 1\}$. The precedence relations are represented by a set $E$ of index pairs such that $(i, j) \in E$ means that the execution of the activity $i$ must precede that of the activity $j$. Then, the AON network is a grapy $G = (V, E)$. Let $R = \{1, \cdots, I\}$ be the set of the renewable resources. The availability of the resource $i$ is $B_i, i = 1, \cdots, I$ and their durations are $p_i, i = 1, \cdots, I$, respectively. Since the activities 0 and $J + 1$ are dummy activities, define $b_{0i} = 0$ and $b_{j+1,i} = 0$ for all $i \in R$. A standard resource-constrained project scheduling can be stated as: choose a schedule such that the makespan is minimized while the resource constraints are satisfied. In general, there are two different methods to formulate it as an optimization problem: discrete time formulation (DTF) and continuous time formulation (CTF).

In discrete time formulations, the time horizon $T$ is partitioned into $T$ time slots $[t-1, t], t = 1, \cdots, T$. Then, the time-indexed variables $x_{jt}$ is introduced to indicate whether the $j$ activity starts at time $t$, i.e.,

$$x_{jt} = \begin{cases} 1 \text{ if the activity j starts at time t} \\ 0 \text{ otherwise} \end{cases}$$

The RCSP can now be formulated as the following optimization problem

$$\min \sum_{t=0}^{T} t x_{J+1,t}, \tag{1}$$

$$\text{s.t.} \sum_{t=0}^{T} x_{jt} = 1, \forall j \in V, \tag{2}$$

$$\sum_{t=0}^{T} t x_{jt} - \sum_{t=0}^{T} t x_{it} \geq p_i, \forall (i, j) \in E, \tag{3}$$

$$\sum_{j=1}^{J} b_{jk} \sum_{\tau=\max\{0, t-p_i+1\}}^{t} x_{j\tau} \leq B_k, t = 1, \cdots, T, k = 1, \cdots, I, \tag{4}$$

$$x_{jt} \in \{0, 1\}, \forall j \in A \text{ and } t \in \{0, 1, \cdots, T\}. \tag{5}$$

In the above optimization problem, (2) and (5) impose non-preemption of the project activities, (3) is simple translation of precedence relations and (4) is the resource constraints. This technique has been extended to solve the multi-mode resource constrained project scheduling in which the variables $x_{jt}$ has one more subscript to indicate which mode is used. More details for the multi-mode case can be referred to [11, 12]. It is clear that the optimization problem defined by (1–5) is an integer optimization problem in which the only variables are $x_{jt}$.

Different from DTF, CTF involves three types of decision variables: (i) the starting time continuous variables $S_i$ for each activity $i \in V$, (ii) Sequential binary variables $x_{ij}, (i, j) \in V \times V$ to indicate whether the activity $j$ starts before the activity $i$ or not, (iii) Continuous flow variables $f_{ijk}$ which are the quantities of resource $k$ that is transferred from the activity $i$ to the activity $j$, $(i, j, k) \in V \times V \times R$. Then, a standard optimization problem for the RCPS based on continuous time formulations can be posed as:

$$\min \quad S_{J+1} \tag{6}$$

$$\text{s.t.} \quad x_{ij} + x_{ji} \leq 1, \forall i, j \in V, i < j, \tag{7}$$

$$x_{ij} + x_{jh} - x_{ih} \leq 1, \forall i, j, h \in V, \tag{8}$$

$$S_j - S_i \geq -M_{ij} + (p_i + M_{ij}) x_{ij}, \forall i, j \in V, \tag{9}$$

$$f_{ijk} \leq \min\{b_{ik}, b_{jk}\} x_{ij}, \forall i, j \in V, k \in R, \tag{10}$$

$$\sum_{j=0}^{J+1} f_{ijk} = b_{ik}, \forall i \in V, k \in R, \tag{11}$$

$$\sum_{i=0}^{J+1} f_{ijk} = b_{jk}, \forall j \in V, k \in R, \tag{12}$$

$$f_{J+1,0,k} = B_k, \forall k \in R, \tag{13}$$

$$x_{ij} \in 0, 1, S_i \geq 0, f_{ijk} \geq 0, \forall i, j \in V, k \in R, \tag{14}$$

where $M_{ij} = ES_i - LS_j$, $ES_i$ is the date before the activity $i$ will not start and $LS_i$ is the date before which it should start.

The constraint (7) states that for two distinct activities, either $i$ precedes $j$, or $j$ precedes $j$ or $i$ and $j$ are processed in parallel. The constraint (7) together with (8)

ensures that no cycles can occur in the sequencing decisions. The constraint (9) links the variables $S_i$ and $x_{ij}$. The constraint (10) lines the variables $f_{ijk}$ with the variables $x_{ij}$. The constraints (11)–(13) are resource flow conservation constraints. In [13], this technique has been extended to solve resource-constrained project scheduling with scarce resources and generalized precedence relations. In [14], the facility layout problem concept is introduced to formulate a project scheduling as an optimization problem which is highly similar to CTF.

## *2.2 Time-Cost Trade-Off*

In scheduling of construction projects, the project duration can be compressed (crashed) by expediting some of its activities in several ways including: increasing crew size above the normal level, working overtime, or using alternative construction methods. The objective of time-cost trade-off problem is to identify the set (or sets) of time-cost alternatives that will provide the optimal schedule. The time-cost relationship of a project activity can be either continuous or discrete. Accordingly the time-cost trade-off problem (TCTP) can be categorized as continuous time-cost trade-off problem (CTCTP) and discrete time-cost trade-off problem (DTCTP). In the literature, almost all studies on this topic are concentrated on DTCTP [15] since the activity is discrete. So here we only review the discrete DTCTP.

Up to now, three versions of the DTCTP have been studied in the literatures: the deadline problem (DTCTP-D), the budget problem (DTCTP-B) and the efficiency problem (DTCTP-E). In DTCTP-D, given a set of time/cost pairs (modes) and a project deadline, each activity is assigned to one of the possible modes in such a way that the total cost is minimized. Conversely, the budget problem minimizes the project duration while not exceeding a given budget. On the other hand, DTCTP-E is the problem of constructing efficient time/cost solutions over the set of feasible project durations [16]. To formulate them as optimization problems, let AON network be $G = (V, E)$, $\theta$ be a specific path of $G$ and $\Theta$ be the set of all possible pathes of $G$. For each $\theta \in \Theta$, denote $t(\theta)$ and $c(\theta)$ as the makespan and cost, respectively. For the given deadline $d$ and budget $b$, DTCTP-D can be posed as the following optimization problem:

$$\min_{\theta \in \Theta} \{c(\theta) : t(\theta) \leq d\}. \tag{15}$$

Similarly, DTCTP-B can be posed as

$$\min_{\theta \in \Theta} \{t(\theta) : c(\theta) \leq b\}. \tag{16}$$

Different from (15) and (16) which are single-objective optimization problems, DTCTP-E is a multi-objective optimization problem which can be stated as:

$$\min_{\theta \in \Theta} \{t(\theta), c(\theta)\}. \tag{17}$$

To solve the optimization problems (15), (16) and (17), we need to transform the feasible set $\Theta$ from a graph to analytical expressions. This task can be achieved by applying the techniques DTF and CTF introduced for RCSP to formulate (15), (16) and (17) as either integer optimization problems or mixed-integer optimization problems. To meet the requirements of real-life scheduling problems, the above TCTPs have been further extended. For example, in [15], time-switch constraints, work continuity constraints, and net present value optimization are included. In [17], TCTP including generalized precedence relationship constraints between project activities, project duration constraints, logical constraints and a budget constraint is formulated as a mixed integer nonlinear optimization problem. In [18], DTCTP with multi-mode case is studied.

## 2.3 Resource Levelling

Resource levelling arises whenever it is expedient to reduce the fluctuations in patterns of resource utilizations over time, while maintaining compliance with a prescribed project completion time. In particular, in cases where even slight variations in resource needs represent financial burden or heightened risks of accidents, a resource levelling approach helps to schedule the project activities such that the resource utilization will be as smooth as possible over the entire planning horizon.

For a given AON network $G = (V, E)$, a standard resource levelling problem can be posed as the following optimization problem [19]:

$$\min \sum_{t=1}^{T} (u_t - \bar{u}_t)^2, \tag{18}$$

$$\text{s.t. } f_i \leq f_j - d_i, \forall (i, j) \in E, \tag{19}$$

$$\sum_{i \in V} u_{it} \leq U. \tag{20}$$

Here the objective function (18) expresses the minimization of the sum of the squared deviations of the resource requirements around the average resource requirement for each time period. In [20], the cost (18) is generalized in integral form including both continuous- and discrete-time case.

## 2.4 Resource Allocation

Resource allocation is to distribute the limited resources to various projects reasonably so as to optimize a certain objective. According to the nature of the distributed resources, it can be categorized as discrete resource allocation and continuous

resource allocation. In the construction industry, most of the materials are discrete. So here we only survey the discrete resource allocation.

A standard discrete allocation problem can be formulated as [21, 22]:

$$\max \sum_{i=1}^{N}\sum_{j=0}^{M} e_{ij}X_{ij}, \tag{21}$$

$$\min \sum_{i=1}^{N}\sum_{j=0}^{M} c_{ij}X_{ij}, \tag{22}$$

$$\text{s.t.} \sum_{i=1}^{N}\sum_{j=0}^{M} jX_{ij} \le M, \tag{23}$$

$$\sum_{j=0}^{M} X_{ij} = 1, \forall i \in \{1, \cdots, N\}, \tag{24}$$

$$X_{i,j} \in \{0, 1\}, \forall i, j. \tag{25}$$

Here the objective function (21) is to maximize the total efficiencies for all the jobs, and the objective function (22) is to minimize the total costs for all the workers. Constraint (23) ensures that the workers cannot be assigned more than the total numbers of workers. Constraint (24) ensures that each job $i$ can be assigned to workers only once.

In [23], the above problem is extended to the time-independent and multi-modal case. In [24], the nonlinear resource allocation is studied which is formulated as a nonlinear integer optimization problem. More variants on the resource allocation can refer to [25].

## *2.5 Integrated Models*

The integrated models is to consider more than one category above. For example, in [26], a multi-mode resource-constrained discrete time-cost trade off problem is studied. The objective is either to minimize the cost subject to constraints including precedence relationship or to minimize project duration subject to budget constraints. In [18], a multi-mode resource constrained discrete time-cost-resource is considered in which resource allocation and levelling problem are taken into account simultaneously. This problem is formulated as a multi-objective optimization problem in which three objectives: minimum duration, minimum cost and minimum resources moment deviation on the project makespan, are included.

## 3 Optimization Methods Solving Project Scheduling Problems

In the past several decades, numerous methods are developed to solve the project scheduling problem including the exact and heuristic or meta-heuristic approaches.

### *3.1 Exact Methods*

The exact procedures in the literature for the project scheduling are the branch and bound algorithms [13, 27–29]. This kind of methods heavily relies on the lower bound computation. In the literature, two classes of lower bounds are well known: constructive and destructive lower bounds. The first class is computed through relaxation methods [13], such as the Lagrangian relaxation. The second class is obtained by means of iterated binary search based routine [30]. In the recent, this class of methods has been further extended to solve variants of project scheduling problems.

In [13], a project scheduling problem with generalized precedence relationships, resource constraints and makespan minimization objective is studied. This problem is first formulated as a mixed integer optimization problem. Then, the formulated optimization problem is solved by branch and bound. During the branching process, a lower bound based on Lagrangian relaxation is adopted to prune the tree of the corresponding AON network.

In [27], a resource-constrained project scheduling problem to minimize the total weighted resource tardiness penalty cost is studied in which the renewable resources cannot be used before the ready date, but are permitted to be used after their due dates by paying the associated penalty cost. Then, a depth-first branch and bound algorithm is applied to solve it in which the lower bound is adopted from the concept of constructing partial schedules [31]. This method has been further extended to solve the resource-constrained project scheduling problem with minimizing the weighted late work criterion [28].

In [29], an alternative-technologies project scheduling problem is studied in which the expected net present value is maximized. This problem is first formulated as a nonlinear integer optimization problem and then solved by a branch and bound algorithm. The bounding strategy is destructive which is based on three dominance rules to fathom unvalued nodes. The first and the second dominance rules prevent duplication of schedules while the third one fathoms low quality schedules.

In [32], a branch and bound algorithm is introduced to solve a resource levelling problem in a machine environment. The upper bound is obtained by a heuristic approach in which exhaustive neighborhood search is used with tabu list control while the lower bound is computed through the Lagrangian relaxation. In [13], a branch and bound algorithm is introduced to solve resource levelling problems and overload problems which are formulated as nonlinear mixed integer optimization problems. To achieve fast computation, linearizing the corresponding objective functions and

improving the quality of the resulting formulations are discussed. Then, the available branch and bound algorithms for linear integer programming is applied.

## *3.2 Meta-Heuristic Methods*

Although the exact methods can always find a global optimal schedule for the given objectives, they are deficient to deal with large scale problems [33]. Instead of searching the global optimal solutions, the researchers and practitioners have tried to design efficient methods with the goal of producing optimal or near optimal solutions. Then, meta-heuristic optimization methods come into play. Meta-heuristics are general purpose high level search frameworks that can be applied to any optimization problem with the use of appropriate local problem dependent solution procedures. Examples of meta-heuristics include simulated annealing (SA), ant colony optimization (ACO), evolutionary algorithm (EA), genetic algorithm (GA), particle swarm optimization (PSO), shuffled frog-leaping (SFL) and bee algorithm (BA). There are a wide variety of meta-heuristics and a number of properties along which to classify them. One of the fundamental classification of the meta-heuristic is Single-Solution Based Meta-heuristics and Population-Based Meta-heuristics. Single-solution based meta-heuristics are included tabu search and simulated annealing in which only one solution is maintained at each cycle of the algorithm. The aim of these methods is to find a new solution with better quality from the current solution iteratively. Population-based meta-heuristics are included genetic algorithm, ant colony optimization, and particle swarm optimization in which a set of solutions are maintained at each cycle of the algorithm. These approaches solve the RCPSP by employing an initial population of individuals each of which representing a candidate schedule for the project. Then, they evolve the initial population by successively applying a set of operators on the old solutions to transform them in to new solutions [34].

Regardless of the metaheuristic chosen for a study, there are common issues that need to be addressed [8]: solution representation, generation of initial solution(s), evaluation function, generation of neighborhood solutions, handling constraint violations and stopping criteria. In the following, we will review the latest meta-heuristic optimization methods in project scheduling.

### 3.2.1 Single-Solution Based Meta-Heuristics

In [35], a simulated annealing algorithm is developed for the resource-constrained project scheduling problems with single and multi-mode versions. There the solution is represented by activity list in which a precedence-feasible solution is represented by an ordered list of activities. Each activity in the list appears in a position after all its predecessors. The generation of neighbor is based on a cyclical shift of all the activities between the old and new positions for a randomly selected activity from the current feasible list. The stopping criterion is chosen as the process time.

This method has been further extended to solve the multi-mode resource-constrained project scheduling problems with discounted cash flows in [36] in which four payment models have been examined: lump-sum payment at the completion of the project, payments at activities' completion times, payments at equal time intervals and progress payments. A Tabu search method is also developed to solve it in [36] and compared with SA. The numerical results show that each method may outperform the other one for different models. In [37], the Tabu search method is further extended to solve a project scheduling problem with the multi-mode and schedule-dependent setup times where a schedule-dependent setup time is defined as a setup time dependent on the assignment of resources to activities over time, when resources are, e.g., placed in different locations. In [38], a simulated annealing method is employed to solve the resource-constrained project scheduling problem with a due date for each activity in which the objective is chosen to minimize the net present value of the earliness-tardiness penalty costs and the neighborhood solutions are generated in a similar way in [35, 36]. In [39, 40], the simulated and Tabu search are introduced to solve the multi-mode capital-constrained project payment scheduling problem, where the objective is to assign activity modes and payments so as to maximize the net present value (NPV) of the contractor under the constraint of capital availability. In [41], experimental evaluations of five variants of simulated algorithms for time-cost trade off project scheduling problems are studied in which they have in common the way to construct the new solutions, the initial temperature and the cooling schedule, whereas with differ cycle size, the number of iterations in each cycle, and the stopping criterion. The best identified variant generated local optima there very close to point estimate of the global optimum but with the slowest time satisfactory efficiency.

### 3.2.2 Population-Based Meta-Heuristics

Comparing with the single-solution based meta-heuristic methods, population-based meta-heuristics are more popular in solving the project scheduling problems. The most studied population-based methods are related to Evolutionary Computation (EC) and Swarm Intelligence (SI).

Montoya-Torres et al. [42] developed a genetic algorithm to solve resource constrained project scheduling problems. The notable advantage of this algorithm is that the chromosomes is represented by a multi-array object-oriented model. The numerical simulations show that this algorithm outperforms some other genetic algorithms for the project with 60 activities. Hartmann [11] developed a genetic algorithm for the case with multiple execution modes for each activity and makespan minimization as objective. The scheduling problem is formulated as an optimization problem based on the representation of the project as an acyclic AON network. The genetic encoding is based on a precedence feasible list of activities and a mode assignment. The novelty of this genetic algorithm is that a local search based on the definition of a multi-mode left shift is employed to improve the schedules found by the basic genetic algorithm. In stead of local search method introduced in [11], a magnet-based

crossover operator that can preserve up to two contiguous parts from the receiver and one contiguous part from the donator genotype is employed in [43]. Numerical simulations of four sets of project scheduling problems with 30, 60, 90, and 120 activities are compared with some other genetic algorithms in literature. Peteghem and Vanhoucke [44] developed a genetic algorithm to the preemptive multi-mode resource constrained project scheduling which allows activities to be preempted at any time instance and restarted later on at no additional cost. The consideration of preemption is very effective to improve the optimal project makespan in the presence of resource vacations and temporary resource unavailability and that the makespan improvement is dependent on the parameters that impact resource utilization [45]. Long and Ohsato [46] developed a genetic algorithm for scheduling repetitive construction projects with several objectives such as the project duration, the project cost, or both of them. The method deals with constraints of precedence relationships between activities, and constraints of resource work continuity. Ghoddousi et al. [18] employed an elitist non-dominated sorting genetic algorithm in [47] to solve multi-mode resource-constrained project scheduling problem (MRCPSP) while considering discrete time-cost trade-off problem (DTCTP) and also resource allocation and resource levelling problem simultaneously which is formulated as a multi-objective integer optimization problem.

Lorenzoni et al. [48] studied a scheduling problem of attending ships within agreed time limits at a port under the condition of the first come first served order. This problem is formulated as a mathematical model of a multi-mode resource-constrained scheduling problem and then is solved by a differential evolution algorithm which has been further discussed in [49]. Zamni [50] proposed a evolution algorithm to solve time-cost trade off of multi-mode resource constrained project scheduling problems in which activities are subject to finish-start precedence constraints under renewable limited resources. In optimizing time-cost performance, the procedure treats the cost as a non-renewable resource whose limit can affect the duration of the project and balances cost versus time through the notion of priority-rank.

Wang and Fang [51] proposed an estimation of distribution algorithm to solve the multi-mode resource-constrained project scheduling problem. The individuals are encoded based on the activity-mode list and decoded by the multi-mode serial schedule generation scheme. To improve the searching quality, a multi-mode forward backward iteration and a multi-mode permutation based local search method are incorporated to enhance the exploitation ability. Different from a univariate probabilistic model in [51, 52] employed ensemble probabilistic models by combining the univariate probabilistic model with the bi-variate probabilistic model which learns different population characteristics in the estimation of distribution algorithm. Wang et al. [53] proposed a Pareto-based estimation of distribution algorithm to solve the multi-objective flexible job-shop problem in which the fitness evaluation based on Pareto optimality was employed and a probability model was built with the Pareto superior individuals for estimating the probability distribution of the solution space. To avoid premature convergence and enhance local exploitation, the population was divided into two sub-populations at certain generations according to a splitting criterion, and different operators were designed for the two sub-populations to gener-

ate the promising neighbor individuals. To enhance the exploitation ability, a local search strategy based on critical path was introduced. The Taguchi method was used to investigated the influence of parameters.

Mahdi Mobini et al. [54] developed an enhanced scatter search algorithm for the resource constrained project scheduling problem. The activity list representation method was used as the encoding scheme. Activity lists in the initial population were decoded to the solutions using both serial and parallel schedule generation scheme. The two-point crossover, the path relinking, and the permutation-based operator were used to generate new solutions from existing solutions in the reference set. Mario [55] introduced the scatter search algorithm to solve the single-mode resource-constrained project scheduling problem with discounted cash flows based on the two assumptions: renewable resources with a constant availability and no activity preemption, and all activity cash flows occurred at predefined time points during execution of the corresponding activity.

The first literature to introduce ant colony optimization algorithm (ACO) to RCPS is [56]. A single ant corresponds to one application of the serial SGS. The eligible activity to be scheduled next is selected using a weighted evaluation of the latest start time priority rule and pheromones which representing the learning effect of previous ants. The additional features include separate ants for forward and backward scheduling, the changing rate of pheromone evaporation over the ant generations, and the restricted influence of the elitist solution by forgetting it at regular intervals. ACO has been further extended to solve the RCPS with multi-modes in [57] in which both renewable and nonrenewable resources are considered. To guide the solution search consisting both activity sequencing and mode selection, two levels of pheromones are introduced: one is to probabilistically select an activity $j$ at the place $i$ in the activity-list, and the other is to probabilistically select the execution mode for this activity. Yin and Wang [24] developed an ACO to solve nonlinear resource allocation problem. Adaptive resource bounds which were dynamically calculated to reduce the search space were incorporated to ensure all resource constraints were simultaneously satisfied. Its convergence was analyzed using node transition entropy to validate that the quality solution obtained was due to the consensus knowledge possessed by the whole ant colony instead of by the wandering of a lucky ant. A modified ACO was proposed to solve a multi-objective resource allocation problem in [58]. The modifications included a special heuristic information, pheromone updating rule, and selection probability equation. Xiong and Kuang [59] proposed an ACO to solve the time-cost trade-off problems in which the modified adaptive weight approach developed in [60] was adopted to guide the solution to the Pareto-front. Mokhtari et al. [61] developed an ACO for stochastic DTCTP which was aimed to improve the project completion probability by a predefined deadline on program evaluation and review technique networks. In [61], the activities were subjected to a discrete cost function and assumed to be normally distributed. Then, the model was formulated as a nonlinear mathematical 0-1 programming problem, where the mean and variance of the activity durations were decision variables and the objective function was to maximize the project completion probability.

Chen et al. [62] developed an improved particle swam optimization (PSO) approach to solve RCSP. During the implementation of PSO, delay local search and bidirectional scheduling were introduced. The delay local search enabled some activities delayed and altering the decided start processing time, and being capable of escaping from local minimum. The bidirectional scheduling rule which combined forward and backward scheduling to expand the searching area in the solution space for obtaining potential optimal solution. The critical path was applied to speed up convergence of the algorithm in which the critical path was used to determine the heuristic value which played a key role in the state transition rule. To enhance the efficiency and effectiveness of PSO, the justification technique in [63] was incorporated in PSO in [64] which was called as justification particle swarm optimization (JPSO). In addition to the justification technique in [64], two other designed mechanisms were integrated. One was the mapping technique for enhancing the exploitation efficiency of justification, and the other was the adjusting ratio of communication topology of PSO for trade-off between exploration and exploitation. PSO to solve RCSP was further improved in [65] by integrating a double justification skill and a move operator for the particles, in association with rank-priority-based representation, greedy random search, and serial scheduling scheme, to execute the intelligent updating process of the swarms to search for better solutions. The computational experiments validated the significant improvements over the other PSO-based algorithms through testing the benchmarks with 30, 60 and 120 activities. Zhang and Li [66] applied PSO to solve TCTP. Numerical simulations for TCTP with 18 activities have been conducted and compared with some other heuristic methods.

Fang and Wang [67] developed an effective shuffled frog-leaping algorithm (SFLA) for the multi-mode resource-constrained project scheduling problem. In the SFLA, the virtual frogs were encoded as the extended multi-mode activity list and decoded by the multi-mode serial schedule generation scheme. Initially, the mode assignment lists of the population were generated randomly, and the activity lists of the population were generated by the regret-based sampling method and the latest finish time priority rule. Then, virtual frogs were partitioned into several memeplexes that evolved simultaneously by performing the simplified two-point crossover. The combined local search including permutation-based local search and forward-Cbackward improvement was further performed in each memeplex to enhance the exploitation ability. Virtual frogs were periodically shuffled and reorganized into new memeplexes to maintain the diversity of each memeplex. This method has been further adapted to solve RCPS with the multi-modes in [12]. Ashuri and Tavakolan [68] proposed a Shuffled-Frog Leaping model to solve complex Time-Cost-Resource optimization problems in construction project planning which was formulated as a multi-objective optimization problem in which total project duration, total project cost, and total variation of resource allocation were minimized simultaneously.

Bee algorithms for resource constrained project scheduling problem was investigated in [34]. Three variants of bee algorithms methods were developed: bee algorithm (BA), artificial bee colony (ABC), and bee swarm optimization (BSO). The parallel-SGS and serial-SGS were introduced to construct active schedules. The different variants made use of different types of bees to provide appropriate level of

exploration over search space while maintaining exploitation of good solutions. A constraint handling method was introduced to resolve the infeasible solutions in which the activity which violated the constraints was replaced by the next activity with smaller priority. Huang and Lin [69] proposed a bee colony optimization algorithm to solve open shop scheduling problems. To shorten the time-cost for obtaining a possible bad solution, an idle-time-based filtering scheme was introduced based on the processes of the Forward Pass of bee's foraging behaviors. During the course of bee colony foraging, the idle-time of partial scheduling was regarded as the reciprocal of profitability. When the profitability of a partial foraging route of a bee was smaller than the average profitability accepted by the colony, this scheduler would automatically stop searching the foraging trip of this particular bee.

## 4 Conclusion

This paper has surveyed the optimization modelling and solution techniques for the project scheduling problems. We first presented five mathematical models to formulate a project scheduling problems as optimization problems. Then, the solution algorithms, including exact algorithms and meta-heuristic algorithms are surveyed in which we mainly focus on the latest three years publications in archival journals since the survey of previous results are available in other survey papers.

It is well-known that the project scheduling problems are NP hard. Most of results on the exact methods published in the recent years are to extend the branch-and-bound methods to solve new variants of project scheduling problems rather than algorithms development. Furthermore, the efficiency of the exact methods is heavily dependent on the bounding strategy which is problem-specific. Different project scheduling problems should adopt different bounding strategies. Despite all efforts applied to solve the instances of the library, up to now only the group with 30 activities has been reported completely solved to optimality. How to develop new exact efficient algorithms to solve large-scaled project scheduling problems are still waiting for study.

Comparing with the exact methods, the meta-heuristic methods seem more promising. In fact, if the problems that cannot be easily solved by conventional exact methods, meta-heuristic methods are good choices in the optimization field. Each meta-heuristic method has its own tunable parameters. The performances of the meta-heuristic methods are heavily dependent on the choice of these parameters. However, there is still lack of a unified framework to tune them. Thus, almost all the meta-heuristic methods are problem-specific which reduce their applicability.

**Acknowledgments** Changzhi Wu was partially supported by NSFC 11001288, the key project of Chinese Ministry of Education 210179, and the project from Chongqing Nature Science Foundation cstc2013jjB0149 and cstc2013jcyjA1338.

## References


1. Egan J (1998) Rethinking construction, the report of the construction task force to the Deputy Prime Minister John Prescott, scope for improving the quality and efficiency of UK construction. HMSO, London
2. Love PED, Simpson I, Hill A, Standing C (2013) From justification to evaluation: building information modeling for asset owners, Automation in Construction (in press)
3. Queensland Department of Main Roads (2005) Roads implementation program 2004–2005 to 2008–2009. www.mainroads.qld.gov.au. Accessed 28 June 2011
4. Waldron B (2011) Scope for Improvement 2011-Project Risk Getting the Right Balance and Outcomes, 27th July 2011. (Available at: http://careers.blakedawson.com, access 15th November 2011)
5. Assaf S, Al-Hejji S (2006) Causes of delay in large construction projects. Int J Proj Manag 24:349–357
6. Galloway PD, Fasce PE (2006) Survey of the construction industry relative to the use of CPM scheduling for construction projects. J Constr Eng Manage 132:697–711
7. Horbach A (2010) A boolean satisfiability approach to the resource-constrained project scheduling problem. Ann Oper Res 181:89–107
8. Liao TW, Egbelu PJ, Sarker BR, Leu SS (2011) Metaheuristics for project and construction management–a state-of-the-art review. Autom Constr 20:491–505
9. Kolisch R, Padman R (2001) An integrated survey of deterministic project scheduling. Omega 29:249–272
10. Harmann S, Briskorn D (2010) A survey of variants and extensions of the resource-constrained project scheduling problem. Eur J Oper Res 207:1–14
11. Hartmann S (2001) Project scheduling with multiple modes: a genetic algorithm. Ann Oper Res 102:111–135
12. Wang L, Fang C (2011) An effective shuffled frog-leaping algorithm for multi-mode resource-constrained project scheduling problem. Inf Sci 181:4804–4822
13. Bianco L, Caramia M (2012) An exact algorithm to minimize the makespan in project scheduling with scarce resources and generalized precedence relations. Eur J Oper Res 219:73–85
14. Jia Q, Seo Y (2013) Solving resource-constrained project scheduling problems: conceptual validation of FLP formulation and efficient permutation-based ABC computation. Comput Oper Res 40:2037–2050
15. Son J, Hong T, Lee S (2013) A mixed (continuous + discrete) time-cost trade-off model considering four different relationships with lag time. KSCE J Civil Eng 17:281–291
16. Hazir O, Haouari M, Erel E (2010) Discrete time/cost trade-off problem: a decomposition-based solution algorithm for the budget version. Comput Oper Res 37:649–655
17. Klansek U, Psunder M (2012) MINLP optimization model for the nonlinear discrete time-cost trade-off problem. Adv Eng Softw 48:6–16
18. Ghoddousi P, Eshtehardian E, Jooybanpuur S, Javanmardi A (2013) Multi-mode resource-constrained discrete time-cost-resource optimization in project scheduling using non-dominated sorting genetic algorithm. Autom Constr 30:216–227
19. Koulinas GK, Anagnostopoulos KP (2013) A new tabu search-based hyper-heuristic algorithm for solving construction leveling problmes with limited resource availablitis. Autom Constr 31:169–175
20. Rieck J, Zimmermann J, Gather T (2012) Mixed-integer linear programming for resource leveling problem. Eur J Oper Res 221:27–37
21. Lin C, Gen M (2007) Multiobjective resource allocation problem by multistage decision-based hybrid genetic algorithm. Appl Math Comput 187:574–583
22. Fan K, You W, Li YY (2013) An effective modified binary particle swarm optimization (mBPSO) algorithm for multi-objective resource allocation problem (MORAP). Appl Math Comput 221:257–267
23. Baradaran S, Ghomi SMTF, Ranjbar M, Hashemin SS (2012) Multi-mode renewable resource-constrained allocation in PERT networks. Appl Soft Comput 12:82–90

24. Yin PY, Wang JY (2006) Ant colony optimization for the nonlinear resource allocation problem. Appl Math Comput 174:1438–1453
25. Patriksson M (2008) A survey on the continuous nonlinear resource allocation problem. Eur J Oper Res 185:1–46
26. Peng W, Wang C (2009) A multi-mode resource constrained discrete time-cost tradeoff problem and its genetic algorithm based solution. Int J Proj Manage 27:600–609
27. Ranjbar M, Khalilzadeh M, Kianfar F, Etminani K (2012) An optimal procedure for minimizing total weighted resource tardiness penalty costs in the resource-constrained project scheduling problem. Comput Ind Eng 62:264–270
28. Ranjbar M, Hosseinabadi S, Abasian F (2013) Minimizing total weighted late work in the resource-constrained project scheduling problem. Appl Math Model (in press)
29. Ranjbar M, Davari M (2013) An exact method for scheduling of the alternative technologies in R&D projects. Comput Oper Res 40:395–405
30. Klein R, Scholl A (1999) Computign lower bounds by destructive improvement: an application to resource-constrained project scheduling. Eur J Oper Res 112:322–346
31. Stinson JP, Davis EW, Khumawala BM (1978) Multiple resource-constrained scheduling using branch and bound. AIIE Trans 10:23–40
32. Drotos M, Kis T (2011) Resource leveling in a machine environment. Eur J Oper Res 212:12–21
33. Kone O, Artigues C, Lopez P, Mongeau M (2011) Event-based MILP models for resource-constrained project scheduling problems. Comput Oper Res 38:3–13
34. Ziarati K, Akbaria R, Zeighamib V (2011) On the performance of bee algorithms for resource-constrained project scheduling problem. Appl Soft Comput 11:3720–3733
35. Bouleimen K, Lecocq H (2003) A new efficient simulated annealing algorithm for the resource-constrained project scheduling problem and its multiple mode version. Eur J Oper Res 149:268–281
36. Mika M, Waligora G, Weglarz J (2005) Simulated annealing and tabu search for multi-mode resource-constrained project scheduling with positive discounted cash flows and different payment models. Eur J Oper Res 164:639–668
37. Mika M, Waligora G, Weglarz J (2008) Tabu search for multi-mode resource-constrained project scheduling with schedule-dependent setup times. Eur J Oper Res 187:1238–1250
38. Khoshjahan Y, Najafi AA, Nadjafi BA (2013) Resource constrained project scheduling problem with discounted earliness tardiness penalties: mathematical modeling and solving procedure. Comput Ind Eng (in press)
39. He Z, Wang N, Jia T, Xu Y (2009) Simulated annealing and tabu search for multi-mode project payment scheduling. Eur J Oper Res 198:688–696
40. He Z, Liu R, Jia T (2012) Metaheuristic for multi-mode capital-constrained project payment scheduling. Eur J Oper Res 223:605–613
41. Anagnostopoulos KP, Kotsikas L (2010) Experimental evaluation of simulated annealing algorithms for the time-cost trade-off problem. Appl Math Comput 217:260–270
42. Montoya-Torres JR, utierrez-Franco E, Pirachican-Mayorga C (2010) Project scheduling with limited resources using a genetic algorithm. Int J Project Manage 28:619–628
43. Zamani R (2013) A competitive magnet-based genetic algorithm for solving the resource-constrained project scheduling problem. Eur J Oper Res 229:552–559
44. Peteghem V, Vanhoucke M (2010) A genetic algorithm for the preemptive and non-preemptive multi-mode resource-constrained project scheduling problem. Eur J Oper Res 201:409–441
45. Buddhakulsomsiri J, Kim D (2006) Properties of multi-mode resource-constrained project scheduling problems with resource vacations and activity splitting. Eur J Oper Res 175:279–295
46. Long LD, Ohsato A (2009) A genetic algorithm-based method for scheduling repetitive construction projects. Autom Constr 18:499–511
47. Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comput 6:182–197
48. Lorenzoni LL, Ahonen H, Alvarenga AG (2006) A multi-mode resource-constrained scheduling problem in the context of port operations. Comput Ind Eng 50:55–65

49. Damaka N, Jarbouia B, Siarryb P, Loukila T (2009) Differential evolution for solving multi-mode resource-constrained project scheduling problems. Comput Oper Res 36:2653–2659
50. Zamani R (2013) An evolutionary search procedure for optimizing timeCcost performance of projects under multiple renewable resource constraints. Comput Ind Eng (in press)
51. Wang L, Fang C (2012) An effective estimation of distribution algorithm for the multi-mode resource-constrained project scheduling problem. Comput Oper Res 39:449–460
52. Chen SH, Chen MC (2013) Addressing the advantages of using ensemble probabilistic models in estimation of distribution algorithms for scheduling problems. Int J Prod Econ 141:24–33
53. Wang L, Wang S, Liu M (2013) A Pareto-based estimation of distribution algorithm for the multi-objective flexible job-shop scheduling problem. Int J Prod Res 51:3574–3592
54. Mahdi Mobini MD, Rabban M, Amalnik MS, Razmi J, Rahimi-Vahed AR (2009) Using an enhanced scatter search algorithm for a resource-constrained project scheduling problem. Soft Comput 13:597–610
55. Mario V (2010) A scatter search heuristic for maximizing the net present value of a resource-constrained project with fixed activity cash flows. Int J Prod Res 48:1983–2001
56. Merkle D, Middendorf M, Schmeck H (2002) Ant colony optimization for resource-constrained project scheduling. IEEE Trans Evol Comput 6:333–346
57. Li H, Zhang H (2013) Ant colony optimization-based multi-mode scheduling under renewable and nonrenewable resource constraints. Autom Constr 35:431–438(online publishing)
58. Chaharsooghi SK, Kermani AHM (2008) An effective ant colony optimization algorithm (ACO) for multi-objective resource allocation problem (MORAP). Appl Math Comput 200:167–177
59. Xiong Y, Kuang Y (2008) Applying an ant colony optimization algorithm-based multiobjective approach for time-cost trade-off. J Constr Eng Manage 134:153–156
60. Zheng DXM, Ng ST, Kumaraswamy MM (2004) Applying a genetic algorithm-based multi-objective approach for time-cost optimization. J Constr Eng Manage 130:168–176
61. Mokhtari H, Baradaran Kazemzadeh R, Salmasnia A (2011) Time-cost tradeoff analysis in project management: an ant system approach. IEEE Trans Eng Manage 58:36–43
62. Chen RM, Wu CL, Wang CM, Lo ST (2010) Using novel particle swarm optimization scheme to solve resource-constrained scheduling problem in PSPLIB. Expert Sys Appl 37:1899–1910
63. Valls V, Ballest F, Quintanilla S (2005) Justification and RCPSP: a technique that pays. Eur J Oper Res 165:375–386
64. Chen RM (2011) Particle swarm optimization with justification and designed mechanisms for resource-constrained project scheduling problem. Expert Sys Appl 38:7102–7111
65. Jia Q, Seo Y (2013) An improved particle swarm optimization for the resource-constrained project scheduling problem. Int J Adv Manuf Technol 67:2627–2638
66. Zhang H, Li H (2010) Multi-objective particle swarm optimization for construction time-cost trade off problems. Constr Manage Econ 28:75–88
67. Fang C, Wang L (2012) An effective shuffled frog-leaping algorithm for resource-constrained project scheduling problem. Comput Oper Res 39:890–901
68. Ashuri B, Tavakolan M (2013) A shuffled frog-leaping model for solving time-cost-resource optimization (TCRO) problems in construction project planning. J Comput Civil Eng (in press)
69. Huang YM, Lin JC (2011) A new bee colony optimization algorithm with idle-time-based filtering scheme for open shop-scheduling problems. Expert Sys Appl 38:5438–5447

# Lean and Agile Construction Project Management: As a Way of Reducing Environmental Footprint of the Construction Industry

**Begum Sertyesilisik**

**Abstract** Construction industry effects the environment through its outputs and its process (i.e. causing $CO_2$ emissions, exploitation of raw materials, energy consumption). There is need to reduce its environmental footprint of the construction industry with the help of efficient and effective construction project management, where possible benchmarking with management principles and applications in manufacturing industry. Such a key concept originated and adapted from manufacturing industry is lean and agile construction which can contribute to the reduction of environmental footprint of the construction industry, enabling especially reduction in waste, increasing value added activities. For this reason, this chapter focuses on the construction project management with respect to the agility and leanness perspective. It provides an indepth analysis of the whole project life cycle phases based on lean and agile principles.

**Keywords** Agile · Lean · Green · Construction

## 1 Introduction

The construction industry effects the environment not only through its outputs but also through the construction process and its inputs. Buildings consume approximately 45–50 % of energy and 50 % of water resources [26, p. 369]. They also cause consumption of raw materials affecting the natural resources. The footprint of the construction industry needs to be reduced. A 'footprint' can be defined as "a quantitative measurement describing the appropriation of natural resources by humans [27] as well as how human activities can impose different types of burdens

B. Sertyesilisik (✉)
Liverpool John Moores University, Liverpool, UK
e-mail: B.Sertyesilisik@ljmu.ac.uk

Istanbul Technical University, Istanbul, Turkey

H. Xu and X. Wang (eds.), *Optimization and Control Methods in Industrial Engineering and Construction*, Intelligent Systems, Control and Automation: Science and Engineering 72,
DOI: 10.1007/978-94-017-8044-5_11, 

**Table 1** Footprint family (after [16])

| | |
|---|---|
| Environmental footprints | Carbon footprint |
| | Water footprint |
| | Energy footprint |
| | Emission footprint |
| | Nitrogen footprint |
| | Land footprint |
| | Biodiversity footprint |
| | Phosphorus footprint |
| | Fishing-grounds footprint |
| | Human footprint |
| | Waste footprint |
| Social footprints | Social footprint |
| | Human rights footprint |
| | Corruption footprint |
| | Poverty footprint |
| | Online social footprint |
| | Job footprint |
| | Work environmental footprint |
| | Food-to-energy footprint |
| | Health footprint |
| Economic footprints | Financial footprint |
| | Economic footprint |
| Combined environmental, social and/or economic footprints | Exergy footprint |
| | Chemical footprint |
| Composite footprints | Ecological footprint |
| | Sustainable process index |
| | Sustainable environmental performance indicator |

and impacts on global sustainability [48]" [16, p. 10]. The construction industry has environmental, social, economic, combined environmental, social and/or economic footprints as well as composite footprints. The main and sub-categories of the footprints have been listed in the Table 1 which has been summarised based on [16]. These footprints establish the footprint family which is defined by [21] as "a suite of accounting tools characterized by a consumption-based perspective able to track human pressure on the surrounding environment."

The global pollution attributable to buildings include [13]: deterioration of air quality (23 %); emission of green house gases (50 %); pollution of drinking water (40 %); depletion of ozone (50 %); and causing landfill waste (50 %). The quantity of generated waste depends on the construction processes used, the structure type being constructed and the technology used to build the structure [43]. This reveals that lean and agile construction project management can contribute to the sustainability and to the reduction of the construction industry's footprint. This chapter focuses on the construction project management with respect to the agility and leanness perspective.

## 2 Basic Concepts

*Lean* is giving the customers what they want, and delivering it instantly, with no waste [29]. Lean focuses on continually increasing the proportion of value-added activity of business through ongoing waste elimination [15, p. 45]. The essence of the lean is [34, pp. 10–11]: elimination of all waste; continuous improvement (kaizen); manufacturing at a rate equal to true customer demand, just when the customer wants it (JIT); and ensuring quality by sensing defects and stopping production until the causes are found and corrected (jidoka). Lean comprises of five main principles [34, pp. 8–17]:

- Specifying value in terms of the final customer
- Identifying and mapping the value stream; that is, how value is being created for the customer
- Making the value flow without interruption
- Letting the customer pull value from the production process
- Pursuing perfection through continuous improvement.

*Lean production* is a philosophy or strategy to minimize defects and to improve company performance [53]. Similarly, Carroll [15, p. xxxi] defines 'lean production' as "the philosophy and practice of eliminating all waste in all production processes continuously". The basic of lean production is to design a production system so that it delivers a custom product instantly without intermediate inventories [28, p. 88]. Lean production enables the flow of value creating work steps while eliminating non-value steps [42, p. 90]. Ohno identified seven categories of waste [34, p. 37]:

- Waste of overproduction
- Waste of time on hand (waiting)
- Waste in transportation
- Waste of processing itself
- Waste of stock on hand (inventory)
- Waste of movement
- Waste of making defective parts.

In addition to these seven categories of waste, Liker added waste of human potential and creativity, and Imai added waste of time [34, pp. 37–38]. Regarding the lean production process, Vais et al. [53] emphasized the following:

> the essence of the lean production process is to identify and eliminate anywaste, which consumes resources, but creates no value: transport, inventory, motion, waiting, overproduction and defects, in short "do more with less and less"—less resources, less effort, less equipment, less time, while becoming closer to providing what costumers actually want. To become Lean and Green one has to focus on the waste of materials and energy.

*Lean construction* "is a concurrent and continuous improvement to the construction project by reducing waste of resources and at the same time able to increase productivity and secure a better health and safety as well as greener environment" [42]. In other words, lean methods are introduced to the construction industry under

the name of lean construction which aims to eliminate re-work, provide continuous work flow and minimize mistakes [12, p. 182]. Sertyesilisik et al.'s [44, p. 173] findings on the main causes of waste in the construction industry revealed the importance of management in waste reduction. The lean construction system sees production as a flow of material, information, equipment, and labor from raw material to the product with the aim of reducing or eliminating non value-adding activities and increasing efficiency of value-adding activities [1, p. 190].

Koskela [35] identified the following lean construction for lean construction implementation: reducing the share of non value-adding activities; reducing variability; reducing cycle time, minimising the number of steps, parts and linkages; increasing output value through systematic consideration of customer requirements; increasing output flexibility; increasing process transparency; balancing flow improvement with conversion improvement and benchmarking; focusing control on the complete process; building continuous improvement into the process. According to [58], the lean construction principles are: specifying value from the customers' view, identifying the value stream, make the value-creating flow, achieving customer pull at the right time and pursue perfection for continuous improvement. Lim [41] and Bashir et al. [9] agreed with the principles stated by Womack and Jones (1996). The Lean Enterprise Institute [40] identified the lean construction principles loop starting with identification of value followed by successive principles of mapping the value stream; creation of flow; and establishment of "pull" and seek perfection.

*Agility* and *leanness* are two concepts which complement each other for an effective supply chain. The total supply chain contains both leanness and agility. [40, p. 108] described 'leanness' and 'agility' as follows:

> *Leanness* means developing a value stream to eliminate all waste, including time, and to ensure a *level* schedule. *Agility*, on the other hand, means using market knowledge and a virtual corporation to exploit profitable opportunities in a *volatile* market place…Agility and leanness are closely related to the total supply chain strategy and the positioning of the decoupling point [40, p. 108].

Leanness and agility has been compared in the Table 2 based on Naylor et al.'s [40] research.

## 3 Need for Lean

Companies are becomming lean so that they can:

- survive among lean competitors [15, p. 45]
- attain high levels of efficiency [3, p. 169]
- attain competitiveness [3, p. 169]
- reduce construction cost with the help of precise materials usage and decreasing the amount of waste generated [35]
- shorten construction period with the help of smoothly and accurately performed works [35]

**Table 2** Comparison of leaness and agility (adapted from [40, pp. 109–112]

| | Leanness | Agility |
|---|---|---|
| *Characteristics* | | |
| Essential characteristics | 'Use of market knowledge' | 'Use of market knowledge' |
| | 'Virtual corporatiom/value stream/integrated supply chain' | 'Virtual corporatiom/value stream/integrated supply chain' |
| | 'Lead time compression' | 'Lead time compression' |
| | 'Eliminate muda' | 'Rapid reconfiguration' |
| | 'Smooth demand/level scheduling' | 'Robustness' |
| Desirable characteristics | 'Rapid reconfiguration' | 'Eliminate muda' |
| Arbitrary characteristics | 'Robustness' | 'Smooth demand/level scheduling' |
| *Metrics* | | |
| Key metrics | 'Lead time' | Lead time, |
| | 'Costs', and | Service, and |
| | 'Quality' | Quality |
| Secondary metrics | 'Service' | Costs |
| Suitability | Where demand variability and product variability is low | Where demand variability and product variability is high |
| Place in the supply chain | Upstream of the decoupling point | Downstream from the decoupling point |

- gain a strategic advantage [15, p. 45]
- meet customer expectations [15, p. 45]
- attain speed of response [3, p. 169]
- respond quickly to opportunities and threats [15, p. 45]
- attain flexibility in production systems [3, p. 169]
- become more environmentally friendly, energy efficient and to generate less waste [42, p. 90]
- eliminate waste (non-value adding activities in production process [18] (Abdek-Razek et al. 2007) [3, p. 169]
- enhance construction process to cater with the client's need and to the needs of environment and communities [35, p. 91]
- attain more synergistic, sustainable and greener future due to the enabled better value of future construction project [35]
- improve the construction process continuously [28]
- minimize the combined effects of dependence and variation [28]
- deliver instantly complex projects in an uncertain environment [31]
- balance the need for stability as the world around a project and its technology are subject to change [31]
- create value [31]

- offer a highly diversified range of products, at the lowest cost, with highlevels of productivity, speed of delivery, minimum stock levels and optimum quality [3, p. 169]
- provide volumes of a variety of products at a relatively low cost by using resources of multi–skilled workers at all levels of organization and highly flexible, increasingly automated machines ([33] as cited by Abdel-Razek et al. 2007, p. 190)
- obtain a product that is adapted to actual demand using the minimum amount of resources and therefore minimising the cost, with the appropriate quality and very high speed of response [3, p. 170]
- increase in safety with the help of the lean principles forcing the firms to define new approaches to design and manage the whole organization and consequently the safety management system [23, p. 96]
- support the development of team work and to increase willingness to shift burdens along supply chains [28]
- attain better labor and cost performance in construction (Abdel-Razek et al. 2007, p. 189).

Lean construction performs well in complex, uncertain and fast projects (Salem et al. 2005). On the other hand, the stakeholders tend to perceive 'lean' as unsuitable due to their belief that construction and manufacturing industries are different (i.e. uniqueness of the construction projects) [28, 35, 36].

## 4 Successful Transformation

Arbos [3, p. 170] emphasised that "implementation of the processes adapted to lean production must be accompanied by the design and management of each of the aspects involved to allow maximum elimination of waste and introduction of the required degree of flexibility." Lean construction requires detailed planning and control of onsite activities and logistic processes [12, p. 182].

Senior management's engagement in the appropriate business processes as well as explicit, clear and consistent guidelines affect the success of the lean implementation [34, p. 303]. An integrated planning/managing process including business, marketing, supply chain, and manufacturing leaders needs to be established [34, pp. 303–304].

Involvement of the stakeholders in the lean construction process enhance the lean transformation [42]. Howell [28] recommends the followings for successful lean transformation in the construction:

- understanding the production 'physics' including the effects of dependence and variation along the supply chain
- applying pull techniques
- establishment of clear set of objectives for the delivery process
- concurrent design of product and process, and

- applying production control throughout the life of the product from design to delivery [28].

Other success factors for lean transformation include [15, pp. 18–21]:

- fundamental adjustments in thinking, attitudes, and behaviours
- management commitment to becomming lean
- involvement and participation by all employees
- a team approach
- a long-run view
- assignment of responsibilities and tasks to teams on a project basis
- lean assignments incorporated into each employee's job on a daily basis
- education and training at all levels
- an organization wide lean process perspective
- application of lean principles, tools, and practices can to all processes
- adopting the lean cultural and transformational principles.

To be lean in production requires [15, pp. 15–16]:

- elimination of waste in all its forms (i.e. operator time; materials; techniques; inventory; waiting time)
- implementation of lean tools (i.e. poke-yoke; TAKT time; kanbans)
- implementation of the lean human resource practices, and
- application of lean operational practices (i.e. pay for performance wage systems; payment for ability; developing and rewarding multiple skill operators).

Howell [28] sees partnering as a tool for lean construction in accordance with the following statements:

"Partnering is a solution to the failure of central control to manage production in conditions of high uncertainty and complexity… Partnering … provides the opportunity for collaborative redesign of the planning system to support close coordination and reliable work flow… Partnering relationships coupled with lean thinking make rapid implementation possible. Where Partnering is about building trust, lean is about building reliability" [28].

The cross-functional processes needed for lean improvement and for attaining lean linkage and flow between all organization processes, include [15, pp. 21–22]:

- Lean quality management
- Lean maintenance
- Lean new product introduction
- Lean design and engineering
- Lean accounting.

These cross-functional processes, their requirements and advantages have been summarized in the Table 3 based on [15, pp. 21–69].

**Table 3** Cross-functional processes, their requirements and advantages (adapted from [15, pp. 21–69])

| Cross-functional processes | |
|---|---|
| Lean quality management | Quality is a central tenet of the organization's lean business policies and strategies. Quality management practices are implemented across the supply chain to continually improve the standard and acceptable level of conformance to quality standards. Quality applies to inputs, throughputs, and outputs of all processes in the supply chain<br>Requirements for lean quality management include:<br>• commitment and involvement of top management in achieving quality objectives<br>• a permanent, organized company wide effort to continuously improve product process quality (i.e. training and measurements)<br>• consideration of process owners and operators as process experts<br>• adaptation of cultural principles of lean |
| Lean maintenance | Lean maintenance is the implementation of the lean principles with empowered staff using lean diagnostic tools to continuously improve the maintenance of their own process and process enablers<br>Requirements for lean maintenance include:<br>• schedulling of all maintenance activities<br>• assignment of responsibility to the operations team for the leaning of the maintenance process<br>Advantages of lean maintenance include:<br>• elimination of production problems<br>• consideration of maintenance processes in equipment purchase, and new product and process design decisions to ensure that maintenance problems do not contribute to inefficient and ineffective operations<br>• standardization of the maintenance processes<br>• addressing maintenance problems as soon as possible, through preventive maintenance and operator inspections<br>• keeping maintenance and performance records on all major pieces of equipment |
| Lean new product introduction | Requirements for lean new product introduction include:<br>• assignment of personnel to a new development project for a product<br>• enabling a career ladder based on new product development, with performance evaluations<br>• assignment of the new product development manager<br>Advantages of lean new product introduction include:<br>• increase in communication, and productivity,<br>• better teamwork,<br>• reduction of costly reengineering,<br>• limiting off-target product development, and<br>• quality improvement |

Continued

**Table 3** Continued

| Cross-functional processes | |
|---|---|
| Lean design and engineering | Lean design and engineering is based on concurrently linked lean design and engineering process<br>Requirements for lean design and engineering include:<br>• examination of all from a team-based lean perspective<br>• monitoring of the design and engineering processes by top management from a lean perspective<br>• assignment of the lean performance engineering team the design and engineering processes for the project<br>Advantages of lean design and engineering include:<br>• the detection of usual errors based on faulty assumptions or misinformation about downstream capabilities<br>• linking design and engineering to other process areas effectively so that poor communication, lack of information transfer, and uncoordinated schedules are minimized or avoided<br>• the detection of product specifications that cannot be met by production processes,<br>• elimination of product features that are not required by customers,<br>• the avoidance of the corresponding time delays to incorporate those specifications or features |
| Lean accounting | Lean accounting is actively concerned with understanding the value-creating processes and proactively working to enable lean processes in customer relationship management, new product introduction, and the other critical processes in the enterprise<br>Requirements for lean accounting include:<br>• utilisation of "transactions matrix" to match transactions currently being performed to the processes that report, consume, or measure these transactions<br>• assigning no value to the inventory<br>Advantages of lean accounting include:<br>• enhancing internal accounting controls while replacing "control by transactions" with operational controls and visual management reporting<br>• knowing which transactions need to be eliminated supporting the enterprise to get lean while maintaining appropriate financial controls<br>• facilitation of the cultural loyalty by the accounting function in a lean enterprise as the lean performance cultural principle is "loyalty to people enables continuous improvement" |

## 5 Lean Tools and Construction Phases

The lean transformation requires lean tools to be effectively used. Lean tools include [34, pp. 8–15]: value stream mapping (VSM); takt time; kaizen; 5S; jidoka; single minute exchange of dies (SMED); poka-yoke; five whys; standard work; total productive maintenance (TPM); cellular manufacturing; heijunka; just-in-time (pull); and kanban. Based on the researches of [2] and [25], [42, p. 93] listed the main concepts of lean construction as: just-in-time (JIT); total quality management (TQM); business process reengineering (BPR); concurrent engineering (CE) and last planner system (LPS); teamwork and value based management (VBM) and OHSAS 18001. Similarly, Abdel-Razek, et al. (2007, pp. 189–190), based on [35], Ballard and Howell [4], Ballard and Howell [6], Ballard and Howell [7] Tommelein [48] and Thomas et al. [46] stated that the lean construction tools include: practice of JIT; usage of pull-driven scheduling; reduction of variability in labor productivity; improvement of flow reliability; elimination of waste, and simplification of the operation; and benchmarking. Most of these concepts are interconnected and can improve performance while minimising construction waste. These tools and practices of lean production needs to be utilised adopting the lean cultural and transformational principles (Carroll 2008, p. 18). The main lean tools have been briefly summarized in the follwoing paragraphs:

- *Value stream mapping (VSM)*: Based on Toyota's material and information flow diagrams, VSM is a method of visually depicting the processes in terms of the physical flow of material and how that creates value for the customer [34, pp. 11–57]. It enables understanding of where waste is created in the process and what might be improved to reduce or eliminate it [34]. A VSM consists of three main components [34, p. 58]:
  - *Material flow* showing the flow of material as it progresses from raw materials to finished goods including all inventories along the flow.
  - *Information flow* governing what is to be made and when it is to be made.
  - *Time line* showing the value-add time and the effect of waste.
- VSM can [34, p. 89]:
  - bring a higher level of clarity on how the process is currently performing
  - highlight areas where improvement should be made
  - provide a template for documenting the ideal future state.
- *Takt time*: Expressed as a "pounds per minute" rate, takt is the time interval at which each item is produced to meet customer demand [34].
- *Kaizen*: Kaizen means continuous improvement. It refers to continuous incremental improvements by working with people and creating ownership of changes at Gemba [53, p. 51]. In the Kaizen events, hands-on operational experience is incorporated into proposed solutions. This is enabled by empowering the employees actually doing the work to remove waste as well as to design and implement more effective processes. Kaizen relies on dedicated team activities. Kaizen events

are completed usually within a few days. As the process includes implementing and testing the improvement results are seen quickly. Continuous improvement of procedures, equipment and processes can help elimination of waste [42]. Kaizen requires the organization to move towards a culture of continuous incremental improvement. This cultural change can be supported through kaizen workshops and kaizen training as they can change people's attitude [11]. [34, p. 160] emphasized the importance of planning and determination of the scope in the kaizen success.

- *5S*: 5S comes from the japanese words for the five specific steps: seiri, seiton, seiso, seiketsu, and shitsuke (in English known as: sort, set, shine, standardize, and sustain). It is a five-step process for workplace organization, housekeeping, cleanliness, and standardized work [53, p. 51]. 5S enhances productivity, quality, speed, and safety of the work [11].
- *Jidoka*: According to Jidoka, everything must stop at the first sign of quality problems so that the problem can be corrected before production resumes, to limit the waste being produced [34, p. 12].
- *Single minute exchange of dies (SMED)*: SMED is a process for systematically analysing all tasks to be performed in a product changeover with the aim of simplification and acceleration of the process [34, p. 13].
- *Poka-yoke*: Poka-yoke is a set of techniques for mistake proofing, used both to prevent defective products from being produced and to prevent production equipment from being set up incorrectly [34, p. 13]. Poka-yoke includes designing things so that they can be put together only one way via sensors or color coding [34, p. 13].
- *Five whys*: Five whys is a way of asking 'why' as many times as needed so that the root causes of the problem as well as the causes of waste in value stream map are identified [34].
- *Standard work*: Standard work aims to find an optimum way of performing the work so that performance is optimized and variability is reduced [34].
- *Total productive maintenance (TPM)*: [11] states the aim of TPM as ensuring that machines in production lines perform required tasks without interruption. TPM enables improvement of the equipment performance as well as of the operation and maintenance to prevent equipment downtime [34]. TPM enhances the 'lean effort' as it helps reducing waste associated with poor equipment performance (i.e. over-production, inventory, defects, transportation, and waiting) [34]. For TPM to be successful, the main responsibility needs to be assigned to the operators who work with the equipment [34].
- *Cellular manufacturing*: Cellular manufacturing is arrangement of machinery layout in a way that operations in a tight sequence can be performed so that transportation and waiting time is minimized [11]. This can often lead to shorter changeovers, higher quality, reduced variability, increased throughput, and better flow [34, p. 14].
- *Heijunka*: Heijunka is production levelling practice to stabilize the operations and to reduce variability in resource utilization [34].
- *Just-in-time (Pull)*: JIT refers to the set of principles, tools, and techniques that enable a company to make what is needed only when it is needed and in the

exact quantity needed. JIT avoids overproduction and reduces inventories to the minimum required for smooth flow. JIT can also help waste reduction [35].

- *Kanban*: Kanban means visible sign. It describes a mechanism for visually signalling what is needed-that is, what must be produced to replenish materials pulled by the customer, which may be the final customer or the next step in the process [34, p. 15].
- *3R*: 3R is "reducing, reusing and recycling" to become lean and green [11].
- *Total quality management (TQM)*: TQM supports organization-wide focus on quality [50]. It leads to improvements in quality of product, environment and occupational health and safety practices [11].
- *Business process reengineering (BPR)*: BPR focuses upon improvement through rapid gains in organizational performance starting from scratch in designing the foundation business development (Small et al. 2011 as cited by [42]).
- *Concurrent engineering (CE)*: CE focuses on product design incorporated with the constraints of subsequent phases into the conceptual phase with increased change control towards the end of the design process [35].
- *Last Planner System (LPS)*: LPS supports achievement of lean goals (i.e. reducing waste and uncertainty, increasing productivity) by making a planning mutual attempt and by increasing the reliability of team members' commitments ([45] as cited from [42]).
- *Teamwork*: Teamwork is established when staff who posses complementary skills and who are committed to the common purpose collaborate holding themselves mutually accountable for the achievement of the goal participating in the teamwork ([18] as cited from [42]).
- *Benchmarking*: Benchmarking is "a process of continuously measuring and comparing an organization's business process against business leaders anywhere in the world to identify, measure, compare, perform gap analysis, adapt and implement new ideas so that the organization can take action for the improvement its performance" [19, 35, 38, 42, 43]". (cited from [1, pp. 189–190]).
- *Value based management (VBM)*: VBM considers product value for the customers as well as for the workers and project participants ([10] as cited [42]).
- *OHSAS 18001*: OHSAS 18001 supports companies in managing their occupational health and safety risks [41] enabling support for the lean construction through improved work environment, and productivity as well as increased job satisfaction [42].

Lean principles, tools, and practices can be used where appropriate in the construction process [15, p. 18]. The majority of the researchers suggested pre-construction and construction are the best time to synergise the lean construction concepts [42, p. 95]. Their early introduction to the process enables to meet the lean targets easier as it is easier to influence the efficiency and effectiveness of the process. Each lean tool is used in the suitable construction phases. Table 4 provides information on lean tools and their usage according to the construction phases as adapted from [42]. Possible suitable phases of lean tools which have not been classified in the previous studies have been classified in Table 5. The material and pre-cast manufacturing companies

**Table 4** Lean tools and construction phases (adapted from [42]: 93)

| | Pre-construction | Construction | Post-construction (Usage) | Author |
|---|---|---|---|---|
| JIT | | X | | Salem et al. [47]; Koskela [35] |
| TQM | X | X | | Salem et al. (2011); George and Jones [22] |
| | X | X | X | Summers [51] |
| BPR | X | X | | Salem et al. (2011); George and Jones [22] |
| CE | X | X | | Koskela [35] |
| LPS | | X | | Seppanen et al. [45]; Salem et al. [47] |
| Teamwork | | X | | Excellence [18] |
| VBM | | X | | Bertelsen [10] |
| | | X | X | Koskela [35] |
| OHSAS 18001 | | X | | Mohd Yunus [39] |

**Table 5** Other lean tools and the construction phases

| | Pre-construction | Construction | Post-construction (Usage) |
|---|---|---|---|
| VSM | X | | |
| Kaizen | X | X | |
| 5S | X | X | |
| Total productive maintenance | X | X | |
| Five whys | X | X | |
| Heijunka | X | X | |
| Kanban | | X | |
| 3R | X | X | X |
| Benchmarking | X | X | X |

in construction industry can also benefit from these lean tools. Especially, among the lean tools takt time, jidoka, SMED, poke-yoke, standard work, and cellular manufacturing are more suitable for mass production companies (the material and pre-cast manufacturing companies) then the construction companies working project-based.

The usage of lean tools can enhance the lean construction process. For example, 5S can support safety performance of the construction. In the lean construction project management, partnering can be a successful contracting strategy as [28] mentioned. Lean construction requires detailed planning and control of onsite activities and logistic processes [12, p. 182]. Planning (defining criteria for success and producing strategies for achieving objectives) and control (causing events to conform to plan and triggering learning and replanning) are two sides of a coin that keeps revolving throughout a project [28].

- *Planning*: Planning needs to cope with uncertainties and changes so that waste is decreased. Direct workers are recommended to be involved in the planning and exe-

cution of experiments in work methods design [4, 5]. The planning system needs to be extended down to execution, first addressing the making of assignments (goal setting, division of labor), then acquisition and management of shared resources, and finally the design of work methods [5]. Look ahead Planning under lean is the progressive reduction of uncertainty to assure constraint free assignments so that variation in workflow is reduced enabling reduction in time and cost [31]. Koskela [35], Koskela and Huovila [37], and [28] criticise current forms of production and project management focus on activities due to their ignorance of flow and value considerations. For this reason, as [28]), [5], Ballard and Howell (2001, p. 10) emphasized, planning needs to focus on stabilizing the work environment as well as on reducing in flow variation.

– Stabilizing the work environment: Lean works to isolate the crew from variation in supply by providing an adequate backlog or tries to maintain excess capacity in the crew so they can speed up or slow as conditions dictate [28]. [8, p. 10] critised traditional planning as follows:

> As currently designed, planning systems do not shield direct production from upstream variation and uncertainty resulting in longer durations and higher costs. To reduce project duration and costs, direct production must first be shielded by the introduction of a near-term, commitment level of planning, with explicit plan quality characteristics. In addition, processes must be installed to identify workable backlog, to match labor to work flow into backlog, and to measure and improve the match of *did* and *will*.

– Moving upstream to reduce in flow variation: "Flow variation can be reduced by stabilizing all functions through which work flows from concept to completion. Better understanding uncertainty, suppliers and customers can eliminate the causes and so reduce variation in shared processes. In addition planning systems must be redesigned to include a level for adjusting *should*, so operations can better match *should* with *will*" [30]
– Improving downstream performance.

- *Control*: Lean construction is based on the simultaneous consideration of product and process development through implementation of the lean concept and production control throughout the project life [42, p. 91]. In lean construction, production control is applied throughout the project life [28]. As waste is defined by the performance criteria for the production system and failure to meet client's requirements [42, p. 91], the quality of construction will affect the waste generated. This reveals the importance of 'control' in lean construction.Lean production control needs to assure that assignments are performed by the crews as planned [31]. Variances need to be identified to assess performance against false standards as well as plan quality needs to be proactively controlled to understand whether or not the labor has matched with the work to be done, whether or not the planned productivity is reasonable and whether or not the right work is performed in the right amount [5]. The control and monitoring process of project and supply chain activities can be enhanced with the help of building information modeling (BIM) and geographic information systems (GIS) as they can provide visual representation. BIM can be

integrated with GIS to track the supply chain status and to provide warning signals to ensure the delivery of materials [32, p. 241]. Regular meetings, covering management, design, user liaison, cost and progress are also important for the control ([20, p. 189]). Obstacles (i.e. lack of materials) need to be removed so that the works can be achieved as planned [4]. For this reason, assignments need to be selected from workable backlog so that missing materials begins to disappear and replaced by items more within local control via coordinating interdependent work activities and shifting priorities enabling elimination of obstacles' root causes [4]. Shared resources need to be coordinated [4]. Furthermore, plan, do, check, and act cycle needs to be pursued. Inspection and testing need to be carefully carried out. The planning and inspection systems need to be linked [4].

In the lean construction project 'design' and 'construction' phases are supported by the lean principles and tools so that the need for variation is decreased, the quality and safety levels are improved, and value is generated.

- *Design*: Design process needs to be able to respond to the advances in technology and to cope with demands of uncertainty and speed [28]. Koskela et al. emphasized difficulty in determining the optimal order of tasks, especially in the early phases of a design project, as the optimal order may depend on a design decision to be made. The design process can be enhanced with the help of BIM and ICT [12]. The end-users' and contractors' involvement in the design process can support leanness of the process as this can help to reduce the need for variation in the construction process. The design process needs to cover the requirements for green buildings. For this reason, the factors which need to be considered throughout the design process include (Green building solutions website):
  - usage of energy-efficient materials
  - creation of products and systems that leave a light footprint on the environment over the full life-cycle
  - consideration of sustainable green design as a process
  - evaluation of manufactured products in terms of waste disposal
  - evaluation of influences of a product's impact on society
  - determination of the environmental impact of a product in terms of energy consumption at each state of a product's life cycle.
- *Construction*: Planning and controlling activities continue throughout the construction phase. The construction phase can be enhanced via visual management. Visual management facilitates effective, purposeful communication between managers, and staff [34]. Visual management's requirements include [34, p. 148]:
  - keeping the process areas clean and well organized through the use of tools like 5S
  - enabling visually display the schedule
  - usage of andons to make it obvious when there is unscheduled stop in the process
  - displaying metrics to see performance and trends in quality, reliability, and other important parameters

– usage of signs, labels, and color coding.

Construction phase is important for the achievement of sustainability and decreasing of execive usage of the materials and reducing the amount of waste generated through [17]: a well structured waste management plan; agreements between contractors and sub-contractors to determine who is responsible for waste on-site; usage of waste companies that offer free waste skip disposal or usage of colour coded bins to segregate waste for recycling; and appointment of a designated waste manager to deal with the delivery and storage of materials.

Furthermore, in the lean construction project management, there is need for identification of the appropriate place of the decoupling point as leaness and agility principles are used effectively in accordance with supply chain strategy best suits the projects needs.

## 6 Summary

Construction project management based on lean and agile principles can enhance the reduction of the construction industry's footprint due to the effective and efficient usage of resources, reduction in waste, variation and rework, and increased quality and value. All these effects can also contribute to the construction companies' competitiveness. For this reason, lean tools and principles are recommended to be applied to the construction project management process which needs to be enhanced by agility principles based on the supply chain strategy.

## References

1. Abdel-Razek R, Elshakour HA, Abdel-Hamid M (2007) Labour productivity: benchmarking and variability in egyptian projects. Int J Project Manage 25(2):189–197
2. Alinaitwe HM (2009) Prioritising lean construction barriers in Uganda's construction industry. J Constr Developing Countries 14:15–30
3. Arbos LC (2002) Design of a rapid response and high efficiency service by lean production principles: Methodology and evaluation of variability of performance. Int J Prod Econ 80:169–183
4. Ballard G, Howell G (1994) Implementing lean construction: improving downstream performance. In: Proceedings of the second annual conference of the international group for lean construction, Santiago, Chile
5. Ballard G, Howell G (1994) Implementing lean construction: stabilizing work flow. In: Proceedings of the second annual conference of the international group for lean construction, Santiago, Chile
6. Ballard G, Howell G (1995)"Toward construction JIT" Proceedings of Conference-Association of Researchers in Construction Management, Sheffield, UK
7. Ballard G, Howell GA (1998) Shielding production: an essential step in production control. J Constr Eng Manage 124(1), 11–17, ASCE, Reston, VA, Accessed 26 Aug 2011 from http://www.cce.ufl.edu

8. Ballard G, Koskela L, Howell G, Zabelle T (2001) "Production System Design: Work Structuring Revisited", LCI White Paper 11
9. Bashir AM, Suresh S, Proverbs D, Gameson R (2011) A critical, theoretical, review of the impacts of lean construction tools in reducing accidents on construction sites. In: Egbu C, Lou ECW (eds) Proceedings of 27th Annual ARCOM Conference, Bristol, UK, Association of Researchers in Construction Management, 5–7 Sept 2011, pp 249–258
10. Bertelsen, S. (2004). Lean construction: where are we and how to proceed. http://www.kth.se. Accessed 26 Aug 2011
11. Bicheno J (2004) The new Lean toolbox towards fast, flexible flow. Picsie Books, Buckingham
12. Breit M, Vogel M, Haubi F, Marki F, Raps M (2008) 4D design and simulation technologies and process design patterns to support lean construction methods. Tsinghua Sci Technol 13(S1):179–184
13. Brown MT, Bardi E (2001) Handbook of energy evaluation. A compendium of data for energy computation issued in a series of folios. Folio no 3: energy of ecosystems. Center for Environmental Policy, Environmental Engineering Sciences, University of Florida, Gainesville. Available at http://www.emergysystems.org/folios.php
14. Burton TT (2011) Accelerating lean six sigma results how to achieve improvement excelence in the new economy, J. Ross Publishing, Hardcover
15. Carroll BJ (2008) Lean performance ERP project management implementing the virtual lean enterprise, 2nd edn, Auerbach publications Taylor & Francis Group, Routledge
16. Cucek L, Klemes JJ, Kravanja Z (2012) A Review of Footprint analysis tools for monitoring impacts on sustainability. J Cleaner Prod 34:9–20
17. Dobson DW, Sourani A, Sertyesilisik B, Tunstall A (2013) Sustainable construction: analysis of its costs and benefits. Am J Civil Eng Archit 1(2):32–38
18. Excellence C (2004) Effective teamwork: a best practice guide for the construction industry Constructing Excellence, MIT Press, Cambridge, pp 1–20
19. Fisher D (1995) Benchmarking in construction industry. J Manage Eng 11(1):50–57
20. Gabriel E (1997) The lean approach to project management. Int J Proj Manage 15(4):205–209
21. Galli A, Weinzettel J, Cranston G, Ercin E (2013) A footprint family extended MRIO model to support Europe's transition to a one planet economy. Sci Total Environ, 461–462:813–818
22. George JM, Jones GR (2008) Understanding and managing organizational behavior. Pearson International Edition
23. Gnoni MG, Andriulo S, Maggio G, Nardone P (2013) "Lean occupational" safety: an application for a near-miss management system design. Saf Sci 53:96–104
24. Green Building Solutions website "Environmental Issues in Construction" http://www.greenbuildingsolutions.org/Main-Menu/What-is-Green-Construction/Environmental-Issues-in-Construction
25. Harris F, McCaffer R (1997) Modern construction management. Blackwell Science, London
26. Hawken P, Lovins E, Lovins H (1999) Natural capitalism—creating the next industrial revolution. Little Brown and Co, New York
27. Hoekstra AY (2008) Water neutral: reducing and offsetting the impacts of water footprints. UNESCO-IHE, Delft, the Netherlands. Value of Water Research Report Series No. 28
28. Howell GA (1999) What is lean construction?. In: proceedings of IGLC-7. University of California, Berkeley, CA, USA, 26–28 Jul 1999
29. Howell G (2001) Introducing lean construction: reforming project management. Report Presented to the Construction User Round Table (CURT), Lean Construction Institute
30. Howell G, Ballard G (1994) Implementing lean construction: reducing inflow variation presented at the 2nd annual conference on lean construction at Catolica Universidad de Chile Santiago, Chile September
31. Howell GA, Koskela L (2000) Reforming project management: the role of lean construction. In: 8th annual conference of the international group for lean construction, Brighton, UK, 17–19 Jul 2000
32. Irizarry J, Karan EP, Jalaei F (2013) Integrating BIM and GIS to improve the visual monitoring of construction supply chain management. Autom Constr 31:241–254

33. Jeong H (2003) Distributed planning and coordination to support lean construction. PhD thesis, University of California, Berkeley
34. King PL (2009) Lean for the process industries dealing with complexity CRC Press Taylor and Francis Group, New York (A productivity press book)
35. Koskela L (1992) Application of the new production philosophy to construction Tech. Report No. 72, CIFE, Stanford University, CA
36. Koskela L, Ballard G, Tanhuanpää VP (1997) Towards lean design, management. In: IGLC-5 proceedings
37. Koskela L, Huovila P (1997) "On Foundations of Concurrent Engineering." Proc. Concurrent Engineering in Construction CEC'97. In: Anumba C, Evbuomwan N (eds). Paper presented at the 1st Intl. Conf., London, 3–4 July (1997) The Institution of Structural Engineers, London, pp 22–32
38. Madigan D (1997) Benchmark Method Version 3, No. ACL /DLV/96/015, Agile Construction Initiative, University of Bath
39. Mohd Yunus NM (2006) Implementation of OHSAS 18001:1999: The experienced of construction companies in Malaysia. Universiti Teknologi MARA Shah Alam, Malaysia
40. Lean Enterprise Institute (2009) Principles of lean. http://www.lean.org. Accessed 25 Aug 2012
41. Lim VLJ (2008) Lean construction: knowledge and barriers in implementing into Malaysia construction industry. http://eprints.utm.my
42. Marhani MA, Jaapar A, Bari NAA (2012) Lean Construction: Towards enhancing sustainable construction in Malaysia. Procedia Soc Behav Sci 68:87–98
43. Naylor JB, Naim MM, Berry D (1999) Leagility: integrating the lean and agile manufacturing paradigms in the total supply chain. Int J Prod Econ 62:107–118
44. OHSAS 18001 (2012) Overview of OHSAS 18001. http://www.vts.net.my
45. Olomolaiye PO (1998) Construction productivity management. Addison Wesley Longman Limited, Edinburgh Gate
46. Osman I, Abdel-Razek RH (1996) Measuring for competitiveness: the role of benchmarking. In: Proceedings of the Cairo first international conference on concrete structures, Cairo University, Cairo, vol 1, 2–4 Jan 1996, pp 5–12
47. Salem O, Solomon J, Genaidy A, Minkarah I (2006) Lean construction: from theory to implementation, Manag Eng 22(4):168–175
48. Sertyesilisik B, Remiszewski B, Al-Khaddar R (2012) Sustainable waste management in the UK construction industry. Int J Constr Proj Manage 4(2):173–188
49. Seppanen O, Ballard G, Pesonen S (2010) The combination of last planner system and location based management system. http://www.lean.org
50. Small MH, Yasin MM, Alavi J (2011) Assessing the implementation and effectiveness of process management initiatives at technologically consistent firms. Bus Process Manag J 17(1):6–20. DOI 10.1108/14637151111105553
51. Summers DC (2005) Quality Management, Creating and Sustaining Organizational Effectiveness Upper Saddle River, New Jersey: PEARSON Prentice Hall
52. Thomas HR, Michael JH, Zavrski I (2002) Reducing variability to improve performance as a lean construction principle. J Constr Eng Manage 128(2):144–154
53. Tommelein ID (1998) Pull-driven scheduling for pipe spool installation: simulation of lean construction technique. J Constr Eng Manage 124(4):279–288
54. UNEP/SETAC (2009) Life cycle management: how business uses it to decrease footprint, create opportunities and make value chains more sustainable. www.unep.fr
55. Vadera S, Woolas P, Flint C, Pearson I, Hodge M, Jordan W, Davies M (2008) Strategy for sustainable construction. Available: http://www.berr.gov.uk/files/file46535.pdf
56. Vais A, Miron V, Pedersen M, Folke J (2006) "Lean and Green" at a Romanian secondary tissue paper and board mill–putting theory into practice. Resour Conser Recycl 46:44–74
57. Womack JP, JonesDT, RoosD (1990) Themachine that changed theworld.RawsonAssociates, 567 New York
58. Womack JP, Jones DT (1996) Lean thinking:banish waste and create wealth in your corporation. simon & schuster

# Managing Construction Projects in Hong Kong: Analysis of Dynamic Implications of Industrial Improvement Strategies

**Sammy K. M. Wan and Mohan M. Kumaraswamy**

**Abstract** Boosting performance levels is one of the critical concerns of increasingly demanding construction industry clients. Focusing on building services installations in this chapter, poor practices in the site installation stage increase non-value-adding defective and demolition works, as well as consequential rework, affecting the overall project performance. Based on a series of face-to-face interviews with experienced practitioners and a focus group exercise, this chapter presents the mapping of various interacting and fluctuating behaviours patterns during the site installation stage of building services in construction projects, with the aid of a generic system dynamics model. Through a real case project for initializing the model, several scenarios were examined to test the behaviour patterns and characteristics of various influential improvement strategies. Drawing on long established industrial engineering principles in the manufacturing industry, some particularly useful concepts have been selected and modified in this chapter for addressing the causes of the identified production shortcomings. This chapter concludes that attention should be paid to prerequisite conditions and readiness of downstream processes prior to on-site installation, improvement of workmanship during installation and integration of self, successive and cross check mechanisms so as to avert the downward spiral of typical vicious cycles that have contributed to poor project performance.

S. K. M. Wan(✉)
Institute of Industrial Engineers (Hong Kong), Hong Kong, Hong Kong
e-mail: sammy.wan@iiehk.org

M. M. Kumaraswamy
Department of Civil Engineering,The University of Hong Kong, Pok Fu Lam, Hong Kong
e-mail: mohan@hku.hk

H. Xu and X. Wang (eds.), *Optimization and Control Methods in Industrial Engineering and Construction*, Intelligent Systems, Control and Automation: Science and Engineering 72,
DOI: 10.1007/978-94-017-8044-5_12, 

## 1 Introduction

The construction industry has often been criticized in high-powered review reports, for the relatively poor performance in operations and product quality, and also for other production shortcomings [1, 2]. In a Brazilian study, Formoso et al. [3] found that 'poor design and specification', 'lack of planning and control', 'poor quality of team work' and 'lack of integration between design and production' are the main reasons of producing defective products. Ekanayake and Ofori [4] found that 'errors by labourers', 'damage to work done due to subsequent trades' and 'improper planning' were the most significant factors. In fact, remedies include calls for more stringent standards and better quality workmanship in site installations. This is particularly important for the complex building services subsector of the industry that involves the coordination of multiple specialist contractors [5]. Coordination may be defined as the process of managing interdependencies between activities [6]. Effective coordination of building services processes avoids downstream field conflicts between building systems [7]. Otherwise, the difficulties of coordination increase as more conflicting tasks are performed concurrently with more interactions downstream [8], also descending into a downward spiral of defects, interferences, clashes and rework.

Current findings from an on-going study suggest that the existing approaches to managing building services processes at 'site installation' stage should be re-examined and revised. To cope with this, the various interacting and fluctuating behaviours during 'site installation' stage of this subsector should be understood in terms of major feedback loops and indeed, the understanding of such dynamic behaviour is helpful to unleash the full potential of operational improvements [9]. Wasteful defective work, demolition and/or rework may arise as indicated by a few reinforcing loops. In order to avert the downward disaster spiral of problematic loops, the subsector should be sensitized to identify and rectify 'problems' early including probable conflicts and/or uncertainties arising out of dimensional tolerances, unclear instructions, drawing uncertainties, services routings, etc. Through self, successive and cross checks by site operatives, fabrication mistakes or errors, clashes and upstream hidden changes may be discovered and fed back interdependently in 'real-time' without passing them to the downstream work team. With a higher level of coordination, it is possible to avoid or significantly reduce the impact of project changes [10]. Also, the site supervisor could now be involved with more process adjustments than traditional passive inspections and reactive corrections on site.

Through in-depth and field-based analysis, several feedback loops of the cause-and-effect relationship were identified and a generic system dynamics model was then developed, after brainstorming in an industry focus group. By using this generic model, the simulation exercises of the model and the impacts on construction performance triggered by various influential improvement strategies are highlighted and analyzed in this chapter.

## 2 Literature Review

In Hong Kong, excessive on-site fabrications contribute to defects and demolition and replacement or rework may be required for any defective or substandard work [11]. Meanwhile, the lack of knowledge of other building services trades might add to fabrication errors and potential service clashes in particular of complex routing assemblies rather than anticipation at source [12]. Rework and errors will generate new works, more rework and errors, with these also triggering more problematic behaviours that often stretch through the project duration [13].

Advocacy of quality construction for site installation activities has been urged recently in Hong Kong [1, 14]. Kumaraswamy et al. [15] advocated that construction project failures could be minimized if quality is closely and properly monitored and controlled, while Love and Li [16] claimed that real-time checking, reporting and auditing are necessary to effectively improve the quality standard of the organization. It seems that previous researchers have already presented a convincing argument that instant feedback is valuable for improving quality performance. Wan et al. [17] suggested that mistake proofing self-checks and successive checks could be applied in the construction processes to ensure that all the work-in-processes are free from defects. Indeed, a defect may not be easy to detect in a huge project but it may accumulate or 'iterate' and 'snowball' into a huge problem. The zero defect target is one of the key drivers for 'lean production', where wastes such as scrap, rework and damage can also be largely reduced as contended [18]. Meanwhile, multi-skilled deployment/utilization of tradesmen are valuable in identifying potential service clashes arising from interdependent subsector trades. Ballard et al. [19] indicated that multi-skilling of work teams is especially important so that the teams can perform more than one specialist task to maintain reliable work flows. All these suggest that improvement strategies are available for managing and improving building services works.

However, project managers may be reluctant to implement improvement strategies and sometimes, they tend to look for short term results and make their decisions mostly based on their intuition and common sense [20, 21]. System dynamics is often deployed to represent and improve the basis and effectiveness of decision-making processes in general. Sterman [22] has long advocated that system dynamics models are widely used in project management, including the business strategy and policy assessment. More recently, it has become a popular technique for dealing with dynamic complexities in construction project management [23–26]. It is justifiable to use system dynamics modeling for managing building services projects having considered the extremely complex and dynamic nature of the subsector involving multiple feedback processes and interdependent relationship [27]. As illustrated by Sterman [28] that a system dynamics modeling approach is powerful in providing analytic solutions for both complex and nonlinear systems, the use of this kind of modeling is well suited to capture 'tightly coupled' relationship and interdependencies of the subsector so that the causal impact of changes arising from the improvement strategies may be traced throughout the system.

Capelo and Dias [29] indicated that managers have to access the behaviour of critical components and acquire a shaper understanding of the impact of different decisions in order to drive improvement. In order to examine the effectiveness of various improvement strategies and formulate an effective set of practically implementable and effective strategies, model simulation with real base case studies and different scenarios is useful for managerial decisions [10, 30], on effectively attacking the critical production shortcomings and improving the overall performance of the subsector. To cope with this, various interacting and fluctuating behaviours in the construction processes in this sub-sector should be understood in terms of major feedback loops and indeed, understanding of such dynamic behaviours is helpful to unleash the full potential of operational improvements [9].

All the above literatures indicate that the production shortcomings of the building services subsector can be reduced through process improvements. To address this systematically, it was felt necessary to investigate the underlying interdependencies between the less tangible factors and to highlight the key dynamic features. As very few contributions have been made to this field, in particular of the complex, uncertain and labour intensive building services subsector of the construction industry in 'building-intensive' locations such as Hong Kong, this study thus addresses this research gap in an academic sense, while also developing useful insights that could inform the industry about improvement strategies to identify and rectify 'problems' early and promote simulation of various strategies at the outset of a project. All these are helpful for alleviating the critical production shortcomings at the site installation stage in this subsector.

## 3 Methodology of Study

In the initial stage study, it was found from a pilot postal questionnaire survey and a series of structured interviews that 'defective work', 'poor coordination of processes or trades', 'rework or variation works' and 'design changes and/or errors' were critical causes of 'production wastes' during the 'site installation' stage of the building services works [27]. The study was then discussed in a specially convened focus group of 12 senior and mid-level construction practitioners (managerial practitioners with practical work experience of not less than 10 years in the industry and who were not involved in any interviews of this on-going research study) including the first author. It was decided that further analysis was necessary to investigate causes and improvement strategies.

We then focused on the 'site installation' stage and investigated 'in-depth' the aforesaid critical causes and solicited experience-based suggestions on how to address these causes and thereby introduce suitable improvement strategies. The questions for the interviews were carefully planned and worded in the same focus group. Voting exercises were conducted for working out the questions but the authors did not take part in any voting to prevent any conflict or interruptions to the final result. A majority vote (i.e. more than 50 % of the votes) from the group members

**Table 1** Profile of the managerial and frontline interviewees

| No. | Trade of building services | Managerial | | Frontline | |
|---|---|---|---|---|---|
| | | Number | Avg. Exp. (Yr) | Number | Avg. Exp. (Yr) |
| 1 | Electrical system | 3 | 16 | 1 | 10 |
| 2 | Fire services engineering | 1 | 14 | 1 | 11 |
| 3 | Heating, ventilation and air-conditioning | 2 | 11.5 | 3 | 12.7 |
| 4 | Plumbing and drainage | 2 | 17 | 3 | 14.3 |
| 5 | Sewage and water treatment | 1 | 15 | 1 | 11 |
| 6 | Lift and escalator | 2 | 14 | 2 | 14.5 |
| 7 | Building automation and security system | 2 | 17 | 3 | 12.7 |
| 8 | Building services package | 3 | 18.3 | 4 | 13.5 |
| 9 | Building services consultant | 2 | 17.5 | 1 | 15 |
| 10 | Building services supervision in main contractor | 2 | 10 | 1 | 11 |

*Note* 'Building services package': involves not less that two trades of building services. 'Avg. Exp.': average experience

was adopted in finalizing the questions. This tailor-made questionnaire, with a 1–5 Likert scale, was developed with a total of 22 items of in-depth analysis of circumstances and proposed practices and techniques, in the form of numerous statements, at 'site installation' stage. Face-to-face semi-structured interviews were carefully planned and conducted with two groups of interviewees (i.e. Group A—managerial level and Group B—frontline level). The interviewees were practitioners who possessed practical work experience of not less than ten years in the subsector, as shown in their profile summary in Table 1.

The interviews were helpful for collecting information and expert opinions in order to probe 'deeper' into the reasons behind the critical causes of 'production wastes' through numerous dialogues involving closed and open-ended questions and eliciting experience-based opinions, and to determine whether or not these may be alleviated by the implementation of suitable improvement strategies. It is believed that the practices and techniques adopted as proposed by the practitioners can reduce the critical 'production wastes' in this subsector, and thus the proposed contextual variables were included in the study. Twenty interviewees each from Group A (managerial) and B (frontline) were interviewed respectively, and asked to indicate their agreement or disagreement with the statements provided against each variable, using five point Likert scales, where five indicates strong agreement and one indicates strong disagreement.

The survey study results were subsequently discussed in the focus group. It was decided that system dynamic modeling would possibly simulate impacts of various improvement strategies and/or project settings for the project managers. This was justifiable as most projects are built around traditional budgeting, which bear little relation to the progress in implementing any additional strategies. Indeed, some researchers suggested that the project managers may be reluctant to invest in intangible assets and they tend to look for short term results and make their decisions mostly

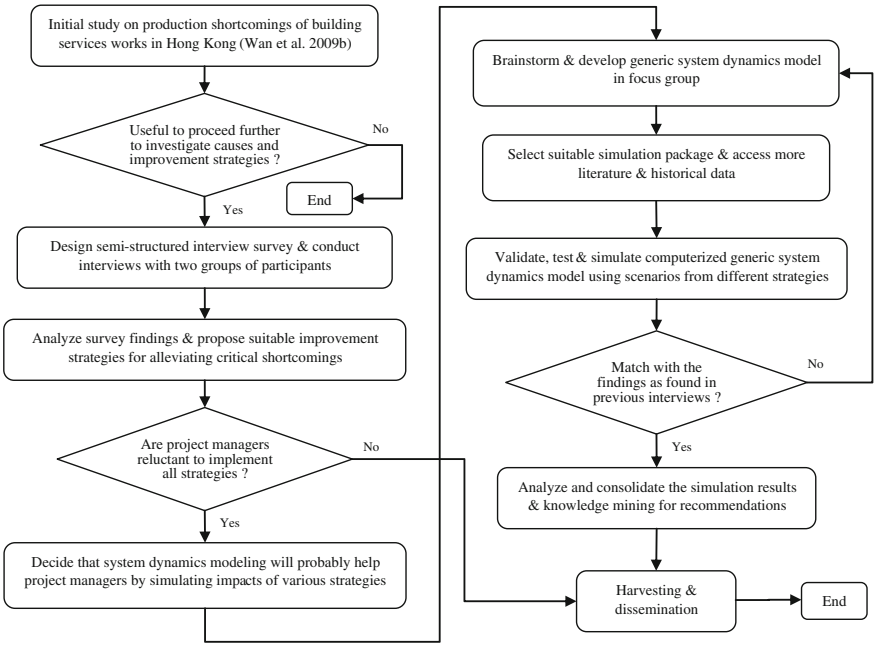



**Fig. 1** Research methodology decision map

based on their intuition and common sense [20, 21]. To cope with this, various interacting and fluctuating behaviors of this sub-sector in terms of major feedback loops were probed. The members of this focus group were then encouraged to express their understanding and perceptions of the dynamic behaviour of the cause-and-effect relationship and underlying interdependencies between the less tangible factors. The data collected was examined by using within-case and cross-case analysis approaches [31, 32]. Focus group members first conducted the within-case analysis so that the unique patterns originated from their within-case experience, while the cross-case analysis was then undertaken to unveil similarities and differences among the different perceptions. Through in-depth discussions involving closed and open-ended questions that elicited experience-based opinions during the semi-structured sessions in the group, several conceptual causal loop diagrams and feedback loops representing the dynamic behaviour of the cause-and-effect relationship were brainstormed and conceptualized in a generic system dynamics model. The generic model as developed above was then tested in order to validate the soundness and usefulness. Through a real case project for initializing the model, several scenarios were examined to test the behaviour patterns and characteristics as highlighted in the interview study. Arising out of the simulation exercises, various influential improvement strategies were recommended in the chapter. Figure 1 sketches the research flow and decision processes of the overall study on which this chapter is based, particularly in the selection of

research methodologies and implementation arrangements for better coherence and relevance of this study.

## 4 Analysis of Interview Study Findings Following Relevant Discussions in Focus Group Exercise

Different trades of the building services were targeted for the interviews, to increase the probability of extracting accurate and objective information while allowing data to be collected from most, if not all the trades, thereby providing greater reliability. Cronbach's Alpha (α) and Spearman-Brown Prophecy (ρ) are calculated to evaluate the reliability, including internal consistency, of the questions. The α scores and ρ scores for 22 items in the 'site installation' stage for Group A and B are 0.80 and 0.91 and 0.87 and 0.97 respectively. In the analysis, 0.7 was used as the acceptable reliability coefficient or threshold of the Cronbach's Alpha value. Since all scores were considerably higher than the 'acceptance level' in particular of the Group B, the high degree of reliability exhibited at this stage prompted further analysis.

The 'Relative Significance Index' (or RSI) technique [33, 34] was used to evaluate the degree of agreement on the items in the questionnaire.

$$\text{Relative Significance Index (RSI)} = \frac{\sum_{i=1}^{n} W_i}{N \times W_H} \quad \text{where } (0 \leq \text{RSI} \leq 1)$$

where $W_i$ is the score of each statement as rated by the interviewees ranging from 1 to 5 (where '1' is the lowest and '5' is the highest rating), N is the total number of interviewees, and WH is the highest score (i.e. 5) adopted in the study.

The numerical score of each item was transformed to RSI in order to assess the relative significance. On the whole, the RSI values are generally and comparably rated highly by Group A. This indicates that most of the frontline participants of Group B interviewed may have reached conclusions guided by their site experiences and may be more conservative in rating the survey. As depicted in Table 2, the S22, S13 and S15 were the three most highly rated in RSI as perceived by Group A while S19, S02 and S05 were the three most highly rated by Group B.

It was found that the two groups expressed different perceptions in the interview study. The managerial participants in Group A strongly agreed that 'instant clarification of design intents reduce rework/demolition' (S22) and generally believed that 'variations to installed works arise from design changes and/or errors' (S03). This seems to suggest, as was also supported by the focus group, that it is often too late if the conflict is resolved during installation [7], since that may result in demolition and replacement, or rework on early stage installed works. Meanwhile, changes or errors in design cause variations to installed work and in some cases, rework or variation induced work to demolish existing installations is one of the key de-motivation

factors of the site supervisors [11]. On the other hand, it was found that 'early capture and rectification of fabrication errors are essential' (S13) and 'instantaneous feedback to rectify problem is superior to passive inspections' (S15) were perceived as highly significant by the managerial participants. This indicates, as also highlighted as important by focus group, that early identification, rectification and prevention of fabrication errors, mistakes and service clashes are important in reducing demolition or rework. For this, the process flow should be arranged to allow instant feedback and monitoring of interdependent works. Love and Li [16] claimed that real-time checking is necessary to effectively improve the quality standard of the organization. It seems that the findings aligned with the argument that instant feedback is valuable for improving quality performance.

The frontline participants of Group B highly rated that 'conflicts with other services occur frequently on site' (S02) and 'poor workmanship contributes to defects/demolition' (S05). Most of the interviewees expressed their dissatisfaction about poor working conditions arranged by the main contractor or other upstream parties because of limited workspaces and poor work arrangement. Untidy and congested working conditions, in particular of false ceilings, horizontal and vertical ducts and/or pipes and conduit works, could contribute to service clashes and even defective works. In the meantime, incompetent and inexperienced gangers and workers may sometimes be employed to keep the project on track and this can lead to a dilution of the average workmanship levels. Additionally, some interviewees opined that the quality of skilled workers is inconsistent and no multi-skilling approaches are adopted. This point relates to the findings of Hawkins in the 'Building Services Research and Information Association (BSRIA) Technical Note TN14/97' that there is no multi-skilled utilization of tradesmen in the UK monitored projects [12]. Lack of knowledge of other building services trades might add to fabrication errors and potential service clashes in particular of complex routing assemblies rather than anticipation at source. On the other hand, the frontline participants of Group B strongly agreed that 'reduce interference or conflict prior to installation at same areas is crucial' (S19). Indeed, some of them might clarify intended routes with other trades only when the services are later found to clash. This is consistent with the findings of researchers that coordination of specialist contractors is a challenging task and the risk of interference problems is highest, as most of them will be performing work concurrently and competing for site resources [5, 35]

Another statistical test was conducted by estimating a wider t-distribution in order to test the statistical difference between sample means of two groups. At the 'site installation' stage as shown in Table 2, it was observed that the t-value of the S08, S15 and S22 exceeds the critical value for the 40 cases in two groups (i.e. 3.018 for 38 degrees of freedom) at 99.5 % confidence interval. In view of this, the obtained sample statistical difference between two groups in the populations was probably a result of sampling error, and Group A may not tend generally to rate higher or lower in a particular item than Group B. However, a few items as specified before were more significant statistically in the positive direction at 99.5 % confidence interval. The Wilcoxon-Mann-Whitney non-parametric test was also conducted, having considered the comparatively small sample size being available and the probable limitations

**Table 2** Responses on the factors, practices and measures at the 'site installation' stage among two groups (sample size = 40)

| ID | Descriptions | Managerial | | | Frontline | | | t-value |
|---|---|---|---|---|---|---|---|---|
| | | RSI | SD | Rank | RSI | SD | Rank | |
| $S_{01}$ | Excessive on-site fabrications occur on site | 0.65 | 0.85 | 21 | 0.66 | 0.57 | 15 | −0.21 |
| $S_{02}$ | Poor workmanship contributes to defects/demolition | 0.77 | 0.75 | 8 | 0.71 | 0.94 | 2 | 1.09 |
| $S_{03}$ | Variations to installed works arise from design changes and/or errors | 0.80 | 0.79 | 4 | 0.68 | 0.82 | 9 | 2.29 |
| $S_{04}$ | Demolition of installed works is one of the key de-motivation factors | 0.72 | 0.60 | 17 | 0.67 | 0.75 | 11 | 1.14 |
| $S_{05}$ | Conflicts with other services occur frequently on site | 0.76 | 0.77 | 10 | 0.70 | 0.76 | 3 | 1.21 |
| $S_{06}$ | Custom fabricated and installed components are required on site | 0.72 | 0.68 | 17 | 0.65 | 0.64 | 17 | 1.63 |
| $S_{07}$ | Team spirit is weak because of tiers of subcontracting mechanism | 0.63 | 0.75 | 22 | 0.58 | 0.72 | 21 | 1.05 |
| $S_{08}$ | Idling occurs when one trade is ready but others are still working | 0.78 | 0.55 | 7 | 0.65 | 0.72 | 17 | 3.13[a] |
| $S_{09}$ | Poor working conditions are arranged by main contractor | 0.77 | 0.75 | 8 | 0.67 | 0.99 | 11 | 1.76 |
| $S_{10}$ | Untidy and congested workplaces contribute to poor workmanship | 0.74 | 0.80 | 12 | 0.70 | 0.89 | 3 | 0.73 |
| $S_{11}$ | Quality of skilled workers is not consistent | 0.74 | 0.86 | 12 | 0.62 | 0.85 | 20 | 2.15 |
| $S_{12}$ | Lack of knowledge of other trades contributes to clashes and/or errors | 0.69 | 0.76 | 19 | 0.67 | 0.67 | 11 | 0.43 |
| $S_{13}$ | Early capture and rectification of fabrication errors are essential | 0.81 | 0.51 | 2 | 0.70 | 0.76 | 3 | 2.62 |
| $S_{14}$ | Validation of each component that is passed from the upstream process reduces defective works | 0.79 | 0.51 | 5 | 0.69 | 0.60 | 6 | 2.75 |
| $S_{15}$ | Instantaneous feedback to rectify problem is superior to passive inspection mechanism | 0.81 | 0.51 | 2 | 0.68 | 0.75 | 9 | 3.11[a] |
| $S_{16}$ | Multi-skilling of work teams is important | 0.74 | 0.73 | 12 | 0.67 | 0.88 | 11 | 1.34 |
| $S_{17}$ | Construction workers registration scheme is helpful to maintain skill levels | 0.66 | 0.73 | 20 | 0.57 | 0.75 | 22 | 1.88 |
| $S_{18}$ | Workers assigned to conduct simple checks to facilitate defect detection | 0.76 | 0.70 | 10 | 0.69 | 0.83 | 6 | 1.41 |
| $S_{19}$ | Reduce interference or conflict prior to installation at same areas is crucial | 0.79 | 0.60 | 5 | 0.74 | 0.57 | 1 | 1.31 |
| $S_{20}$ | Assembling begins only after resources available and uncertainties resolved mostly | 0.74 | 0.66 | 12 | 0.66 | 0.47 | 15 | 2.16 |
| $S_{21}$ | Exchange of actual work schedules among trades and crews is important | 0.73 | 0.67 | 16 | 0.63 | 0.75 | 19 | 2.17 |
| $S_{22}$ | Instant clarification of design intents reduce rework/demolition | 0.85 | 0.72 | 1 | 0.69 | 0.83 | 6 | 3.19[a] |

*Note* Standard deviation (SD), Relative significance index (RSI), rank and *t*-value are tabulated above

[a] Refers to items that exceed the critical *t*-value at 99.5 % confidence interval

in following normal distributions. It was observed that the respective p-values of the S08, S15 and S22 were still unreliable ranging between 0.002677 and 0.01143. It is understandable that the managerial participants in Group A might rate a little bit different, particularly for the items, 'idling occurs when one trade is ready but others are still working' (S08), 'instantaneous feedback to rectify problem is superior to passive inspection' (S15), and 'instant clarification of design intents reduce rework/demolition' (S22), as these are more or less correlate to the 'management-controllable' factors. Poor management of these items may cause downstream operational disturbances as construction work comprises a sequence of trades [36]. Indeed, this becomes more important for this subsector that is composed of preceding, subsequent and interdependent trades and crews in a complex supply chain. The interview results were interpreted based on the relatively limited sample size of interviewees, but supported by the intensive focus group exercise. To increase the reliability of the interview study further, more data from the industry and further in-depth interviews can be targeted in a future study.

## 5 Development of Generic System Dynamics Model

### *5.1 Formulating, Brainstorming and Validating with an Industry Focus Group*

In our analysis, we learnt that despite the general differences in perceptions between the two groups of participants, the interview study reviews reveals that coordination at 'site installation' stage is important but it can be more challenging for the building services subsector, especially on complex or services-intensive buildings and fast-tracked projects [7]. While compiling the above findings, it was considered that different characteristics and behaviours may appear but may deviate widely, given the different nature of projects. In order to investigate the behaviour patterns as highlighted in the survey and investigate the effect of several improvement strategies, a generic system dynamics model was developed and simulation exercises were conducted under different scenarios. In this generic model, variables were connected by arrows denoting the causal influences between variables for systemically identifying, analysing and communicating a feedback loop structure [37]. Each causal link was assigned a polarity to indicate how a dependent variable was impacted when an independent variable changes. In this stock-and-flow diagram as illustrated, positive or reinforcing loops tended to reinforce or amplify whatever was already occurring while negative (or balancing) loops counteracted and opposed change.

Initial assessments indicated that this approach may be used for improving practices in the building services subsector in Hong Kong and similar jurisdictions with high building densities. Through the research exercises and efforts with the focus group, several conceptual causal loop diagrams and feedback loops representing the dynamic behaviour of the cause-and-effect relationship were brainstormed as shown

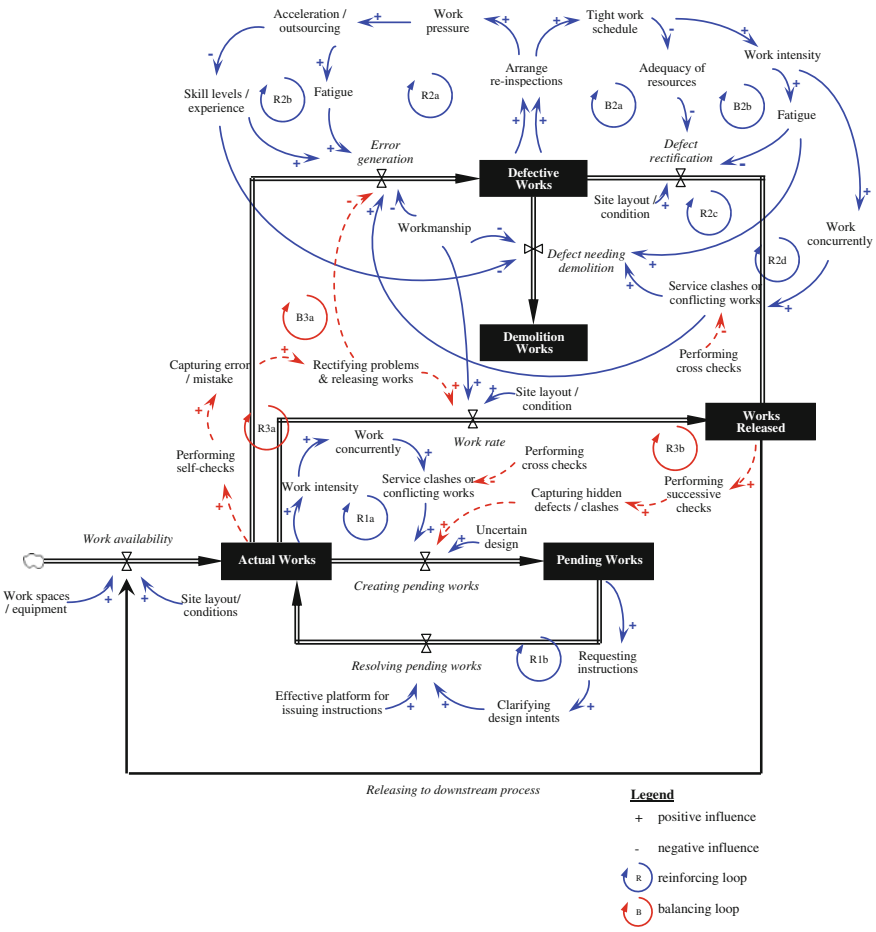



**Fig. 2** Generic system dynamics model

in Fig. 2. A few key reinforcing loops were identified by the focus group that may produce 'vicious circles' and ultimately, 'snowball' the undesirable consequences of particular situations. The process of conceptualization in the focus group also involved identification of stock and flow diagrams where stock represents accumulated quantities whereas flow controls the changing rate of quantity going into/out of stock [38]. All constituents of the entire causal loop diagram were only accepted as valid if there was 90 % agreement within the focus group.

Conceptualization and development of conceptual causal loop diagrams.

Prior to commencing actual works on site, the building services trades should have completed numerous off-site preparation tasks and coordinated with the principal contractor in order to reduce interference or conflict at the same areas as highlighted by the frontline participants in the survey. As shown in Fig. 2, the works for execution

are determined by the flow of 'Work availability' and the key concept behind this flow accounts for the reality that the amount of planned works moving to 'Actual Works' is constrained by numerous factors such as concurrent works, on-site work spaces, shared equipment and resources, site layout and readiness of upstream works.

Some of the uncertain or conflicting works and design drawings which may need to be clarified are constraints in 'Creating pending works', and the degree to which design uncertainties, service clashes, etc., are identified in the actual works. These works are accumulated in the stock of 'Pending Works' and it may be determined to return these to 'Actual Works' if the client representative or consultant engineer issues clear information/instruction to clarify the problems on hand through the flow of 'Resolving pending works'. As rated highly by the frontline participants in the survey study that conflicts with other services occur frequently on site. Higher work intensity and inadequate work spaces cause more service clashes or conflicting works where specialist contractors always work concurrently amidst tight and complex configurations of pipework, crossover, branching, cables and so on. Even though the services may be shown to be neatly connected in a shop drawing, they may in reality require a significant amount of extra space or variation works to other services in order to achieve the designated layout. This generates a reinforcing loop (R1a) with more service clashes and/or conflicting works in turn generating more pending works, and higher productivity demands from tighter schedules.

Project managers usually issue more 'requests for instructions' to consulting engineers for clarifying and dealing with the increasing number of pending works through another reinforcing loop (R1b). If readily available, the use of efficient instruction channels or platforms (e.g. through advanced information and communication technology tools) can facilitate faster processing of up-to-date information, instructions and change orders in project works [39, 40]. This reflects the situation as highlighted by the managerial participants of the interview study that instant clarification of design intents is important to reduce any rework or demolition.

Fabrication errors usually arise from inadequate site storage conditions, skill levels and relevant experience, workmanship, etc., at the flow of 'Error generation'. This leads to some completed works being accumulated into the stock of 'Defective Works'. In practice, on-site fabrication and pipework and conduit work support systems are prepared manually at the site and the poor workmanship may be attributed to special site constraints that are peculiar to the highly interdependent building services subsector. Also, the likelihood for errors increases as tasks like building services installations are performed concurrently with more interactions [8]. As the number of rectifications and re-inspections arising from defective works increases, project managers expect delays and/or losses causing more pressure on output and productivity. This is further aggravated by the generally tight schedules and greater site constraints in Hong Kong and similarly crowded cities. This in turn may lead to attempts at acceleration and/or outsourcing. Thomas [41] pointed out that acceleration is a possible option in response to the idling of resources and disruption of work sequence and/or schedule. But this kind of reactive 'fire-fighting' attitude increases workloads under short-term pressure. Demands to meet increasingly impossible milestones may raise productivity in the short term but reduce workmanship levels, and induce fatigue

in the longer term that also increases errors and decreases work quality as well as productivity, resulting in a reinforcing loop (R2a). On the other hand, incompetent and inexperienced gangers and workers with less familiarity with the project may be employed to maintain the installation milestones, stimulate work progress and keep the project on track. However, this time-focused management will lead to more poorly coordinated works, more potential fabrication errors and service clashes, as represented by another reinforcing loop (R2b). It seems that the subsector should be sensitized to identify and rectify 'problems' early as supported by the positive statement found in the interview study that early capture and rectification of fabrication errors are essential.

Defective works can be dealt with in one of two ways. The works may either be rectified and sent to the stock of 'Works Released' again, or moved to the stock of 'Demolition Works' depending on the defect condition, workmanship, worker's fatigue, etc. If rectifying the defect is likely to consume too much time and resources, the project manager may decide to demolish it instead. Even if some site operatives may identify the problems and intend to rectify the defects, availability of resources and worker capacities (including morale) may constrain their ability to communicate feedback among interdependent trades and/or crews. Missed milestones in tight schedule as a result of prior problems will limit the availability of resources for other tasks [42]. Two counteracting loops (B2a and B2b) may build up and this could ultimately result in a decreasing rate of 'Defect rectification'. Conversely, fatigue and reduced morale can create a sense of 'hopelessness' that increases errors and reduce productivity [13]. This aggravates the effect of 'Defect needing demolition' at a reinforcing loop (R2c). Also, higher work intensity, poor allocation of work sequence or processes and inadequate workspaces can generate more service clashes and/or conflicting works where specialist contractors usually work concurrently amidst tight and complex configurations of pipework, crossover, branching, cables and so on. This leads to a reinforcing loop (R2d) with service clashes and/or conflicting works, which generate even more non-value-adding demolitions instead of rectifications. In particular, when on-site installations are performed concurrently, the 'ripple effect' of the feedback loops is amplified, hence a part of completed works flow to 'Demolition Works'. This phenomenon is more critical in Hong Kong type 'building-intensive' scenarios, where specialist contractors are competed to deliver tight and complex configurations within short time frames.

It is understandable that early capture and rectification of fabrication errors or mistakes are necessary at site installation stage and if possible, instantaneous feedback by skillful workers to rectify problem is superior to passive inspection mechanism in order to prevent the disaster spiral in feedback loops as triggered from the stock of 'Defective works'. Previous researchers have already presented convincing arguments that instant feedback is valuable for improving quality performance [15, 16]. This explains why the participants in the interview study echoed instantaneous feedback is superior to reactive defect rectification. It seems that there are rooms for the concept of mistake proofing self-checks and successive checks to be adapted to ensure that all the work-in-processes are free from defects [17]. It is therefore suggested to combine the traditional flow of 'Do' and 'Check' processes, as this

facilitates the immediate capture of any fabrication error and/or mistake throughout the processes before they become defects. As suggested in Fig. 2, another reinforcing loop (R3a) represented by dotted arrows was created and the skillful workers will now have a greater responsibility and authority for capturing any fabrication error or mistake from the stock of 'Actual Works' via the self check mechanism before the works become defects. This ultimately increases the 'Work rate' towards the stock of 'Works Released'. On the other hand, the downstream assemblies in the next process will validate each in-process work from the upstream that is passed to them via successive checks. If there is any probable conflict or hidden defect arising out of dimensional tolerances, service routings or unclear instructions, corresponding work is captured and flows to the stock of 'Pending Works', in case the upstream work team cannot rectify the problematic work easily. This prevents probable interdependent assembly conflicts or hidden defects becoming more defects or even conflict assemblies occur in downstream process, as conceptualized through a reinforcing loop (R3b) in Fig. 2. As stated previously, a multi-skilled deployment/utilization of tradesmen can identify potential service clashes arising from interdependent subsector trades, for example in concurrent ceiling and routing assemblies such as pipeworks, ductworks, risers, cabling, etc. This supports that a cross check as a daily job routine can help anticipate or capture the service clashes and ultimately break the chain of the 'vicious cycles' at R1a and R2d before defects materialize.

## *5.2 Testing and Validation of Generic Model*

The generic model as developed above was then tested in order to validate the soundness and usefulness of the model [28]. Based on the aforesaid elements, the model replicating the stocks and flows was developed using one of the leading system dynamics simulation software packages, iThinkTM. The validation process was conducted with the help of the same focus group, by checking the adequacy and appropriateness through interviews with these experts and by checking with archival materials and databases from members. Judgmental estimation of values on a scale of 0–100 % representing very low to very high value of the specified variables were carefully adopted for comparisons instead of inputting formulas at this stage of validation and simulation. This means that if the perception rating of 'site layout/conditions' is very good, say 95 %, the degree of works flowing through 'work availability' to the stock of 'Actual Works' will be large. As in the system dynamics approaches adopted by Ogunlana et al. [30], the model was designed to indicate the trends of the dynamic interactions and the model behavior rather than the exact parameter values and simulation outputs.

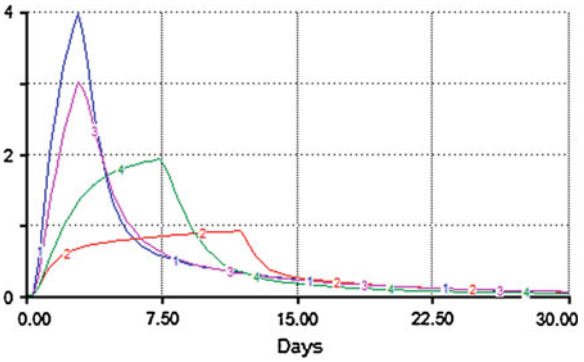
During the structural validation tests, extreme condition tests on the major changing rates of inflow parameters were performed in order to assess whether the model responds reasonably when subjected to extreme values on the flows. The tests were conducted assuming the parameter values of all flows were set at 0.5 and the duration of simulation was 30 days. The extreme conditions included zero work availability,

no error generation, very few defects needing demolition and zero defect rectification. As the model responds reasonably and is consistent with the knowledge of the real situations, the focus group concluded that in essence, the model is structurally valid.

### *5.3 Simulations and Implications of Model*

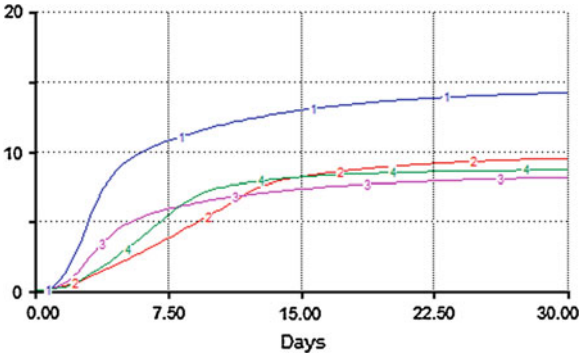
It is decided to initialize the model with a real case project in Hong Kong. For the base case, the model was simulated based on a project in Hong Kong that involved demolition of the existing hotel apartments and construction of new atrium apartments with 11-storey guest floors, one level sky lobby and restaurant, entrance lobby, plant rooms and circulation areas in a luxury hotel tower. The building services contract was awarded to two specialist contractors and the contract sum was around HK$ 142 million with a 25-month span on construction. In this project, installation of the pipework and ductwork systems was undertaken by a specialist contractor nominated by the client under the supervision of one principal contractor. Generally speaking, the actual building services activities were aggregated up to hundreds of activities. For the case studies, the model was simulated based on a period of 30 days after the installation works had been commenced for more than 5 months. The work activities were allocated at the stock of 'Planned Works' (not shown in Fig. 2) of the model for simulation.

Project managers of the principal contractor and specialist contractor of this contract were invited to join the focus group in order to determine the input parameters of variables of the base run. Detailed semi-structured interviews were conducted and interviewees were asked to estimate the level/degree of relevant variables based on their knowledge, historical data and experience of the project. In this complicated fast-tracked project, it was obvious that site layout and conditions were of poor quality, extremely congested and untidy, in particular of plant rooms. The shop drawings produced by the consultant were poorly coordinated and contained unclear information relating to actual routing of the combined services. Workers often commenced installation works even though the relevant shop drawings were not yet approved. Delays were observed because of the incorrect setting out information and changing design intent, and these caused a few clashes amongst various building services trades passing through the same work area. Engineers sometimes had to decide the actual routing after site observations and clarification about intended routes with other trades occurred only when the services clashed. On-site cutting and threading of water pipes were observed and aluminum working platforms were provided for assembling pipeworks. Pipework components such as gaskets, nuts, bolts, washers, hangers, etc., were not palletized and stored tidily on site. The work relationship between building services trades was not close and it was evident that some trades tried to work faster based on their specialities but with probable conflicts. When questioned, the engineers of those trades stated that they intended to work fast to prevent being trapped in the critical path of late work. On the other hand, they could escape




**Case 1 –** Base case
**Case 2 –** Work space/equipment and site layout/condition = 0.99
**Case 3 –** Workmanship and skill levels/relevant experience = 0.99
**Case 4 –** Self, successive and cross checks = 0.99

**Fig. 3** Effects of different scenarios on defective works

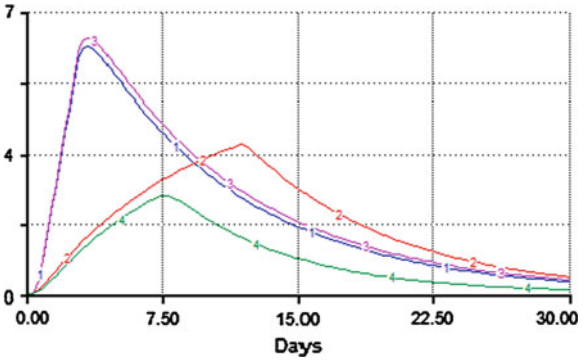



**Case 1 –** Base case
**Case 2 –** Work space/equipment and site layout/condition = 0.99
**Case 3 –** Workmanship and skill levels/relevant experience = 0.99
**Case 4 –** Self, successive and cross checks = 0.99

**Fig. 4** Effects of different scenarios on demolition works

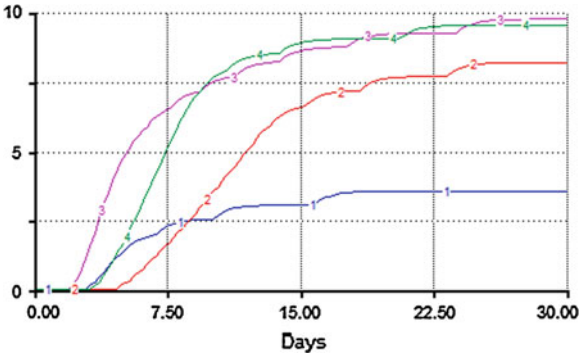
the responsibility of extra scope of work of route deviations that would probably have resulted from preceding trades being clashed.

The Base Case was initialized and simulated through this real case project in Hong Kong, as well as to test the robust behaviour of the model and assess the ability of the model to reproduce the behaviour arising from other scenarios. In this chapter, only a few scenarios that produced distinct behaviours were shown in the comparison figures and discussed for analysis. X-axis on Figs. 3, 4, 5 and 6 is the time period of simulation and the y-axis is the stock values or units in defective works, demolition works, pending works and released at different scenarios. Based on the diverse simulation exercises for the Base Case and different scenarios as shown in Figs. 3, 4,




**Case 1 –** Base case
**Case 2 –** Work space/equipment and site layout/condition = 0.99
**Case 3 –** Workmanship and skill levels/relevant experience = 0.99
**Case 4 –** Self, successive and cross checks = 0.99

**Fig. 5** Effects of different scenarios on pending works




**Case 1 –** Base case
**Case 2 –** Work space/equipment and site layout/condition = 0.99
**Case 3 –** Workmanship and skill levels/relevant experience = 0.99
**Case 4 –** Self, successive and cross checks = 0.99

**Fig. 6** Effects of different scenarios on released works

5 and 6, it is obvious that behaviour of the Base Case is relatively similar to the real case project and this confirms that the model is behaviourally valid. As illustrated in the Base Case and Case 2 scenario, poor site layout/condition and inadequate work spaces/equipment prior to actual works contribute to more defective and demolitions works, and ultimately affect the output works. The poor situation could become more significant when designated installation space is inadequate to accommodate the different interdependent services such as pipes and cables [43]. With extremely poor distribution/adequacy in work spaces, pending works arising out of uncertain and/or conflicting works will suddenly be increased to an abnormal high value resulting in more demolition works and less work outputs. As illustrated previously in this

chapter, poor allocation of site layout/conditions and inadequate work spaces can cause more services clashes and/or conflicting works. The model responded sensitively in the Case 2, for adequate work spaces/equipment and excellent site layout/condition, by reducing large amount of defective, pending and demolition works in the comparison figures. This seems to echo positively with the frontline participants in Group B of the previous interview study that conflicts with other trades occur frequently on site and reduction in interference or conflict prior to installation at same areas is crucial.

Untidy and congested working conditions in this Base Case contribute to poor workmanship and even defective works. Additionally, workmanship, skill levels and work experience of trademen influence work quality and likely to affect the amount of defective works and also demolition works. Excessive on-site fabrications contribute to defects and demolition and replacement or work may be required for any defective or substandard work. The lack of knowledge of other building services trades might add to fabrication errors and potential service clashes in particular of complex routing assemblies rather than anticipation at source. This finding is similar to that in UK projects as observed and analyzed in 'BSRIA Technical Note TN14/97' [12]. As shown in the simulation exercises for Base Case and Case 3 scenario, improvement of workmanship could reduce the defective and demolition works (Figs. 3 and 4) resulting in more released works (Fig. 6), although it seems that there was no positive impact on the pending works (Fig. 5).

As illustrated in the simulation exercises, reducing work clashes and/or conflicting works may indeed lead to less demolitions, replacement and/or reworks, in particular of false ceiling works, duct works, pipe works or conduit works [11]. By installing the dotted arrows as shown in the generic model, the number of service clashes and/or conflicting works reduced largely through cross checks while the reinforcing loop of capturing fabrication errors/mistakes via self checks propelled successfully to reduce largely the defective, pending and demolition works as shown in the comparison figures for Case 4 scenario. This is superior to passive and excessive 'requests for inspections' from designated building services inspectors at the site that hold all in-process activities until approvals are granted. As the successive checks could effectively identify upstream hidden mistakes or errors before releasing to downstream processes, non-value-adding output works reduced significantly. The simulation results favourably support the results of interview study that early capture and rectification of fabrication errors are important and instantaneous feedback to rectify problem is superior to passive or reactive inspections. It is clear from the comparative plots that the adoption of a combination of the aforesaid improvement strategies can give the optimal results for work quality and productivity. With this favourable conditions and higher level of coordination, it is possible to avoid or significantly reduce the impact of project changes [10]. The model responded with high sensitively and all the aforesaid behaviours and consequences evidently reflect real life scenarios in the building services subsector of the construction industry. Although this research needs to probe further, so as to refine the model, those initial findings are useful in helping project managers to target appropriate improvement strategies and conditions when planning building services works in a way such that

non-value-adding stocks can be reduced and productivity increased. This is particular important since construction works basically comprise a sequence of trades causing downstream operational disturbances [36], if improvement strategies are not established at the outset of the project.

The focus group experts examined the simulation results and unanimously agreed that the model is of value not only in simulating the existing dynamic behaviors at the 'site installation' stage of the building services projects, but also in assisting the project managers in improving their project outcomes by means of better planned adjustments/management of the relevant key variables. The simulation results demonstrate how understanding/applying desirable project settings and cultivating self, successive and cross checks could significantly reduce the non-value-adding pending works and demolition works and boost work productivity for 'building-intensive' site works. Although the model was developed with a number of simplifying assumptions, the usefulness of the model is mainly in the insights that it provides to project managers to help improve project outcomes by means of better-planned focus upon, and hence careful management of, relevant key variables in order to pre-empt production shortcomings when commencing site installation of the building services works.

## 6 Conclusions

This chapter has shown that the building services subsector should be sensitized to reduce conflicts prior to on-site installations and identify and to rectify 'problems' early through self, successive and cross checks. Otherwise, the detrimental impact of pending, defective and demolition works may 'snowball' the undesirable consequences including depressed work productivity on the project. A series of face-to-face semi-structured interviews led to the conceptualization of several conceptual causal loop diagrams and feedback loops representing the dynamic behaviour of the cause-and-effect relationships. These were then represented in a generic system dynamics model. Based on a real case project in Hong Kong and diverse simulation exercises, the authors concluded that (a) certain prerequisite conditions and readiness of downstream processes helps reduce interference or conflicts in actual works at the same areas; (b) excessive on-site fabrications contribute to defects and demolition and it is necessary to improve workmanship, as this can greatly reduce large amount of defective and demolition works; (c) cross and successive checks can break the chain of the 'vicious circles' of service clashes or conflicting works and identify upstream hidden mistakes or errors respectively, while self checks capture fabrication errors/mistakes early, hence reducing undesirable defective, pending and demolition works.

Although the current generic model has established the potential for greatly reducing non-value-adding works and uplifting work productivity, further research efforts are still needed to refine its detailed representation and operationalization through different cause-and-effect situations, also accessing more real case projects in various regions. This chapter is of value not only in facilitating more understanding regarding

the system dynamics behaviours of the 'site installation' stage of the building services works, but also in assisting project managers to formulate relevant improvement strategies that can address the identified shortcomings in the industry.

**Acknowledgments** The authors are grateful to the industry practitioners who contributed their time and knowledge to the study. The support of Grant HKU7138/05E of the Hong Kong Research Grants Council, Hong Kong Federation of Electrical and Mechanical Contractors and Analogue Group of Companies are also gratefully acknowledged.

## References

1. Egan J (1998) Re-thinking construction. Report of the construction task force, Department of the Environment, Transport and the Regions, London
2. Construction Industry Review Committee (CIRC) (2001) Construct for excellence. Report of the construction industry review committee, Construction Industry Review Committee, Hong Kong
3. Formoso CT, Soibelman L, Cesare CD et al (2002) Material waste in building industry: main causes and prevention. J Constr Eng Manage 128(4):316–325
4. Ekanayake EL, Ofori G (2004) Building waste assessment score: design-based tool. Build Environ 39:851–861
5. Tommelein ID, Ballard G (1997) Coordinating specialists. Technical report 97–8, CEM program, University of California, Berkeley
6. Malone TW, Crownstone K (1994) The interdisciplinary study of coordination. Comput Sur 26(1):87–119
7. Riley DR, Varadan P, James JS et al (2005) Benefit-cost metrics for design coordination of mechanical, electrical, and plumbing systems in multistory buildings. J Constr Eng Manage 131(8):877–889
8. Williams T, Eden C, Ackermann F et al (1995) Vicious circles of parallelism. Int J Proj Manage 13(3):151–155
9. Sterman JD, Kofman F, Repenning NP (1997) Unanticipated side effects of successful quality programme: exploring a paradox of organizational improvement. Manage Sci 43(4):503–521
10. Park M, Peña-Mora F (2003) Dynamic change management for construction: introducing the change cycle into model-based project management. Sys Dyn Rev 19(3):213–242
11. Wan SKM, Kumaraswamy MM, Liu DTC (2009a) Contributors to construction debris from electrical and mechanical work in Hong Kong infrastructure projects. J Constr Eng Manage 135(7):637–646
12. Hawkins G (1997) Improving M&E site productivity. Technical note TN 14/97. BSRIA, Bracknell
13. Lyneis JM, Ford DN (2007) System dynamics applied to project management: a survey, assessment, and directions for future research. Sys Dyn Rev 23(2/3):157–189
14. Tam VWY, Tam CM, Chan JKW et al (2006) Cutting construction wastes by prefabrication. Int J Constr Manage 2006:15–25
15. Kumaraswamy MM, Ng ST, Ugwu OO et al (2004) Empowering collaborative decisions in complex construction project scenarios. Eng Constr Arch Manage 11(2):133–142
16. Love PED, Li H (2000) Overcoming the problems associated with quality certification. Constr Manage Econom 18(2):139–149
17. Wan SKM, Liu DTC, Kumaraswamy MM (2006) Reduction of accidents in construction industry through mistake-proofing skills to promote lean production. In: Proceedings of 22nd annual conference of the APOSHO, Bangkok, 2006
18. Womack JP, Jones DT (1996) Lean thinking: banish waste and create wealth in your corporation. Simon & Schuster, New York

19. Ballard G, Tommelein ID, Koskela L et al (2002) Lean construction tools and techniques. In: Best R, De Valence G (eds) Design and construction: building in value. Butterworth-Heinemann, Boston, pp 227–254
20. Lantelme EMV, Formoso CT (2000) Improving performance through measurement: the application of lean production and organizational learning principles. In: Proceedings of 8th annual conference of the IGLC-8, Brighton, 2000
21. Norreklit H (2000) The balance on the balanced scorecard—a critical analysis of some of its assumptions. Manage Acc Res 11:65–88
22. Sterman J (1992) System dynamics modelling for project management. Sloan School of Management, Massachusetts Institute of Technology, Cambridge
23. Rodrigues A, Bowers J (1996) The role of system dynamics in project management. Int J Proj Manage 14(2):213–220
24. Ng W, Khor E, Lee J (1998) Simulation modelling and management of large basement construction project. J Comput Civ Eng 12(2):101–110
25. Peña-Mora F, Li M (2001) Dynamic planning and control methodology for design build fast-track construction projects. J Constr Eng Manage 127(6):445–456
26. Hao JL, Hill MJ, Shen LY (2008) Managing construction waste on-site through system dynamics modelling: the case of Hong Kong. Eng Constr Arch Manage 15(2):103–113
27. Wan SKM, Kumaraswamy MM, Liu DTC (2009b) Industrial management approaches for improving building services works in Hong Kong. In: Kazi AS, Hannus M, Boudjabeur S (eds) Open building manufacturing: key technologies, applications and industrial cases. ManuBuild, South Yorkshire, pp 85–102
28. Sterman J (2000) Business dynamics: system thinking and modelling for a complex world. McGraw-Hill, New York
29. Capelo C, Dias JF (2009) A system dynamic-based simulation experiment for testing mental model and performance effects of using the balanced scorecard. Sys Dyn Rev 25(1):1–34
30. Ogunlana SO, Li H, Sukhera FA (2003) System dynamics approach to exploring performance enhancement in a construction organization. J Constr Eng Manage 129(5):528–536
31. Eisenhardt KM (1989) Building theories from case study research. Acad Manage 14(4):532–550
32. Prasertrungruang T, Hadikusumo BHW (2008) System dynamics modelling of machine downtime for small to medium highway contractors. Eng Constr Arch Manage 15(6):540–561
33. Enshassi A, Mayer PE, Mohamed S et al (2007) Perception of construction managers towards safety in Palestine. Int J Constr Manage 7(2):41–51
34. Tam CM, Deng ZM, Zeng SX et al (2000) Quest for continuous quality improvement for public housing construction in Hong Kong. Constr Manage Econom 18(4):437–446
35. Riley DR, Horman MJ (2001) Effects of design coordination on project uncertainty. In: Proceedings of 9th annual conference of IGLC-9, Singapore, 2001
36. Karim K, Marosszeky M, Davis S (2006) Managing subcontractors supply chain for quality in construction. Eng Constr Arch Manage 13(1):27–42
37. Richardson GP (1986) Problems with causal-loop diagrams. Sys Dyn Rev 2(2):158–170
38. Park M (2005) Model-based dynamic resource management for construction projects. Autom Constr 14(5):585–598
39. Sarshar M, Tanyer AM, Aouad G et al (2002) A vision for construction IT 2005–2010: two case studies. Eng Constr Arch Manage 9(2):152–160
40. Lam PTI, Wong FWH, Tse KTC (2010) Effectiveness of ICT for construction information exchange among multidisciplinary project teams. J Comput Civ Eng 24(4):365–376
41. Thomas HR (2000) Schedule acceleration, work flow and labour productivity. J Constr Eng Manage 123(2):181–188
42. Bayer S, Gann D (2006) Balancing work: bidding strategies and workload dynamics in a project-based professional service organisation. Sys Dyn Rev 22(3):185–211
43. Wan SKM, Kumaraswamy MM (2009) Industrial management approaches for improving material control in building services works. Eng Const Arch Manage 16(3):208–223

# Dynamic Project Management: An Application of System Dynamics in Construction Engineering and Management

**Sangwon Han, SangHyun Lee and Moonseo Park**

**Abstract** Computer simulation is one of the most widely utilized tools for operational research in construction engineering and management. Although discrete event simulation (DES) has been extensively utilized in construction, system dynamics (SD) has received relatively little attention despite its great potential to address dynamic complexity in construction projects, which are inherently complex, dynamic and involve multiple feedback processes and non-linear relationships. This chapter introduces dynamic project management (DPM), an SD-based new construction project modeling approach, which has been successfully applied to deal with dynamic complexities in diverse infrastructure and building projects. Particularly, this chapter introduces three major theoretical foundations of DPM: a holistic approach, a system structure-oriented approach, and the incorporation of control time delays. This chapter is expected to serve as a useful guideline for the application of SD in construction and to contribute to expanding the current body of knowledge in construction simulation.

**Keywords** Simulation · Discrete event simulation · System dynamics · Control theory

S. Han (✉)
Department of Architectural Engineering, University of Seoul, Seoul, Korea
e-mail: swhan@uos.ac.kr

S. Lee
Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI, US
e-mail: shdpm@umich.edu

M. Park
Department of Architecture, Seoul National University, Seoul, Korea
e-mail: mspark@snu.ac.kr

H. Xu and X. Wang (eds.), *Optimization and Control Methods in Industrial Engineering and Construction*, Intelligent Systems, Control and Automation: Science and Engineering 72,
DOI: 10.1007/978-94-017-8044-5_13, © Springer Science+Business Media Dordrecht 2014

## 1 Introduction

Computer simulation is the process of designing a mathematical-logical model of a real system and experimenting with this model on a computer [1]. It enables testing without disruption of ongoing operations and committing physical resources, testing hypotheses for feasibility study, compressing or expanding time for closer observation, gaining insight about complex systems, identifying system bottlenecks and providing answers for "what if" scenarios [2]. The availability of special-purpose simulation languages, massive computational capabilities at a decreasing cost per operation, and advances in simulation methodologies have made simulation one of the most widely used and accepted tools in operations research [3].
Computer simulation has also been extensively utilized in construction over the past decades. While discrete event simulation (DES) has predominantly been used in the history of construction simulation development, system dynamics (SD) has received relative little attention despite its great potential to address dynamic complexity in construction. Therefore, research has recently focused on applying SD in the area of construction engineering and management. In order to explore the applicability of SD in construction engineering and management, this chapter introduces dynamic project management (DPM), an SD-based new construction project modeling approach. The findings from this chapter are anticipated to expand the current body of knowledge in construction simulation and provide valuable lessons to construction researchers and practitioners seeking to develop SD models.
This chapter first examines differences between discrete and continuous simulation and then explains the dominance of DES in the history of construction simulation development. Next, the capabilities of SD are explored and the applicability of control theory and construction management examined. Lastly, three major theoretical foundations of DPM are provided and conclusions are drawn focusing on its opportunities, benefits, and further improvement in the area of construction engineering and management.

## 2 Discrete Versus Continuous Simulation

Simulation models can be largely classified as either discrete or continuous based on the timing of state change. Few systems in practice are wholly discrete or continuous, but since one type of change predominates for most systems, it is usually possible to classify a given system as either discrete or continuous [4]. It is possible to model the same system with DES or a continuous simulation CS; [1] and the choice between these options is a function of the characteristics of the system and the objective of the simulation [3].

## 2.1 Discrete Event Simulation (DES)

DES is the modeling of systems in which the state variables change only at a discrete set of points in time [3]. The aim of a DES model is to reproduce system activities that the entities engage in and thereby learn something about the behavior and performance potential of the system [1]. That is, an artificial history of the system is generated based on the model assumptions, and observations are collected to be analyzed and to estimate the true system performance measures [3].
A typical example showing the concept of DES is a bank teller model that mimics a teller at a bank processing customers' transactions. The central purposes of this type of model are to forecast (a) the average time a customer spends at the bank, and (b) the proportion of the time that the teller is idle. In this model, each customer arrives at the bank at a random time (i.e., event time). On arrival, if the teller is busy (i.e., serving a customer who arrived at the bank earlier), the customer joins a queue and waits until the teller is idle (i.e., finished serving the previous customer). Then, the customer is served for an uncertain duration of time and finally leaves the bank.
In this model, the system states (e.g., status of the teller and number of waiting customers) are changed only when a customer arrives at the bank or departs the bank (i.e., event time). For example, on the arrival of a customer, if the teller is idle, the status of the teller is changed to 'busy' and the teller starts serving the customer. Otherwise, the customer waits in the queue and the number of waiting customers is increased by one. On the departure of a customer, if the queue is empty, the status of the teller is changed to 'idle'. Otherwise, the teller continues serving the next customer and the number of waiting customer is decreased by one. Since DES assumes the system states remain constant between event times, a complete portrayal of the system state can be obtained by advancing simulation time from one event to the next [1].

## 2.2 Continuous Simulation (CS)

In CS, changes in the state of a system occur continuously over time [3]. As discussed above, DES focuses on a distinct individual entity (e.g., customer) in a system and keeps track of the time taken for each entity (e.g., waiting time or service time of each customer). On the other hand, CS regards an entity as a continuous quantity (e.g., water) flowing through a system and focuses on the rate of change in the entity during the specified time unit [5]. Thus, while system state variables are determined by the sequence and timing of random events in DES, CS is usually constructed by defining mathematical equations for a set of the system state variables. Differential equations are frequently used in describing the system state variables in CS due to their effectiveness in representing the rate of change over time [1]. For example, the current state of a variable ($S(t_2)$) can be derived from its previous state ($S(t_1)$) and the rate of change over the specified time duration as shown in Eq. (1).

$$S(t_2) = S(t_1) + \int_{t1}^{t2} \left(\frac{dS}{dt}\right) dt \quad (1)$$

As expressed in Eq. (1), the system state variables are updated at finely-sliced time steps of equal duration in CS but at random event times in DES. For example, when modeling a situation where 1.25 h is taken to produce a unit, DES updates the cumulative number of production by '*one unit after* 1.25 h '(i.e., entity-based system update) while CS updates by '0.8 *units after* 1 h' if the time step is 1 h (i.e., time-based system update). Accordingly, in this case, DES assumes that there has been no progress during the first one hour while CS assumes that 0.8 units of progress have been made. From the point of view of '*production planning*', DES estimation looks more realistic and valid, whereas from the viewpoint of '*progress monitoring*', CS calculation can be more informative than DES estimation. However, it should be noted that both DES and CS assume an absence of any progress during the first half hour (when the time step is 1 h). For these reasons, when more accurate simulation results are required, DES tends to further divide an activity (e.g., production) into several sub-activities (e.g., cutting, assembling, bolting, painting and packing) while CS tends to adopt a smaller time step (e.g., 0.5 or 0.25 h).

These differences imply that DES is more efficient for point estimation (e.g., calculation of exact timing of unit production) but CS is more effective for pattern estimation (e.g., projection of progress behavior over time). Of course, it is possible for CS to detect more accurate time positions (e.g., 1.25 h) by decreasing the time step (e.g., 0.25 h). However, achieving greater accuracy in CS by using smaller time steps incurs a cost in terms of increased computational time and effort [6].

## 3 Construction Simulation

Construction simulation is the science of developing and experimenting with computer-based representations of construction systems to understand their underlying behavior [7]. Simulation has been widely applied as an effective planning and performance improvement tool in the construction management area by virtue of its advanced capabilities to analyze complexity and uncertainty [8].

Examining the history of construction simulation, it is clear that the prevalent approach for construction simulation has traditionally been DES [7, 9, 10]. The dominance of DES in the construction simulation area is primarily attributed to its advanced capabilities providing operational details that are not readily provided by network-based approaches (e.g., CPM/PERT) [11]. Current construction management approaches including CPM/PERT are conceptually rooted in the idea of decomposition [12] where it is generally hypothesized that the complexity of a project can be reduced by subdividing the project into manageable smaller activities [13]. Consequently, the general direction of these approaches is in deconstructing further into even smaller fragments of a construction project and searching for explanations at the lowest possible level [14].

These decomposition-based approaches can provide detailed information regarding 'what to build' at an activity level by subdividing a project (it is not unusual for a modern construction project to include thousands of activities), but are limited in representing 'how to build' at an operational level. For this matter, DES is an effective complementary tool that can deal with operational details (e.g., resource status). For example, the CPM/PERT generally represents earthmoving as a single activity, whereas DES zooms into its internal operational logistics and analyzes complex interactions between work tasks (e.g., load, haul, dump) and resource assignment (e.g., pushers, scrapers).

As such, utilization of DES enables management of several complex problems including bottle-neck analysis, sensitivity analysis, resource balance analysis, productivity improvement, process optimization, and so forth [15]. Because of the advanced problem solving capabilities of DES under the popularity of decomposition-based approaches, the construction management discipline has encouraged a narrow, partial view of a project, concentrating on the detailed planning of individual discrete activities and operational details [16–19].

However, due to its narrow focus and partial view, DES can sometimes provide unrealistic estimations because operational performance is significantly affected by the project contexts (e.g., schedule urgency) that are determined by other concurrent operations [9]. Thus, there is a strong need to apply simulation to high-level strategic decision making beyond construction operations [15]. Based on the analysis of 3,500 projects, [18] reported that lack of strategic analysis is a major reason for the failure of many projects. Considering the complex interrelationships between processes, subcontractors, resources, etc., in a construction project, the use of simulation for high-level strategic decision making requires a holistic approach because appropriate policies cannot be made without a complete understanding of the whole project structure [20]. For this reason, it is difficult to use DES models (based on reductionism) for high-level decision making [21]. To address this deficiency, several researchers have proposed SD as a complementary tool to DES in the strategic decision making process.

## 4 System Dynamics (SD)

SD is a methodology used to understand how systems change over time. The idea of SD originally stems from a servomechanism for automatic machine control. The concept of the servomechanism evolved during and after World War II and has been used in many engineering occasions [22]. The servomechanism is an acting machine to control the operation of a larger machine by virtue of feedback [23] and its entire science has been known as control theory. A good example is a thermostat that receives temperature information and can raise or lower the temperature operating a heater or cooler. Beyond its application to engineering, this concept is fundamental to all life and human endeavor: a person senses that he may fall, corrects his balance, and thereby is able to stand erect; a profitable industry attracts competitors until the

profit margin is reduced to equilibrium with other economic forces, and competitors cease to enter the field; the competitive need for a new product leads to research and development expenditure that produces technological change [22]. Though the majority of its application has been to 'hard' systems such as a mechanical control system, which provides more controllable environment, it can be also applied to 'soft' systems such as a management control system because it is also a fundamentally feedback-driven system [6]. Consequently, significant research efforts have been directed at understanding social systems since the late 1950s. These efforts have proceeded under the term SD, which is an approach to understand the behavior of complex systems over time using computer simulation.

By virtue of feedback structure analysis, SD can provide analytic solutions for complex and non-linear systems [6]. Hence, SD is well suited to dealing with the dynamic complexity in construction projects, which are inherently complex and dynamic, involving multiple feedback processes and non-linear relationships [24]. However, as previously discussed, DES has dominated the history of construction simulation and SD has received relatively little attention in construction despite its great potential. In order to fully explore the applicability and utilize the benefit of construction simulation, SD needs to be further investigated. To address this need, this chapter introduces DPM, which has been successfully applied to diverse infrastructure and building projects [20, 25, 26]. Particularly, this chapter focuses on the theoretical foundation of DPM by applying the original concept of control theory to construction engineering and management, instead of repeating the successful applications of DPM, which have been well reported [20, 25, 26].

## 5 Control Theory and Construction Management

Control theory has played a vital role in the advance of engineering and science [27]. Control theory aims to produce the desired or optimal outcome of the system, and its main mechanism is feedback control. Feedback represents that the output of a system is passed back to its input. In control theory, feedback is used as follows: (1) the output of the system is compared to the desired state, initially set as a reference; (2) control actions are taken to reduce this gap if any; and (3) this process is iterated until the desired state is realized to control the system. Figure 1 illustrates a simplified feedback control. Specifically, there is a plant, the object or system to be controlled, which is a combination of components that act together and perform a certain objective [27]. The plant is working with its reference (i.e., desired state), and its output (i.e., actual state) is monitored through a sensor so that the gap between reference and feedback signals (i.e., error, A in Fig. 1) can be captured. If there is any gap between them, a controller takes some control actions to reduce this gap. In addition, there can be external disturbances, which tend to adversely affect the output of a system [27]. These control actions and disturbances (B in Fig. 1) act as another input for the plant, and the sensor monitors the corresponding output. The feedback process is reiterated toward the goal where the actual state meets the desired state during the entire life

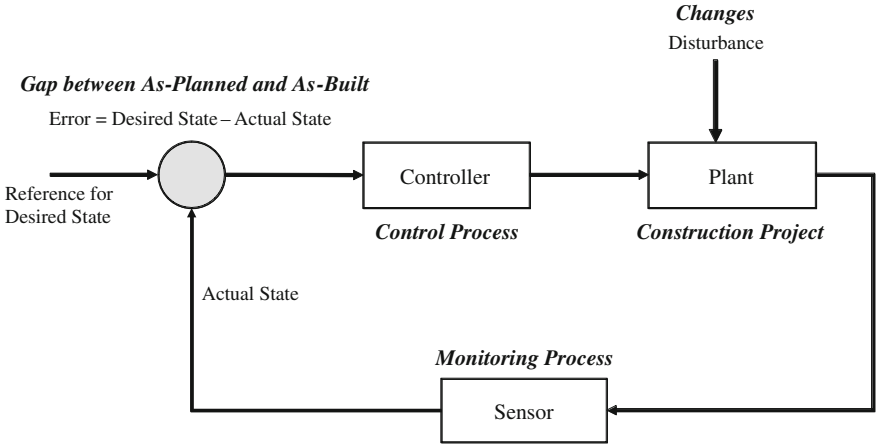



**Fig. 1** Control theory: Feedback control system

cycle of this system. The objective of the feedback control system is to design the system which produces the desired state despite error and disturbance.

Close investigation enables an analogy to be drawn between the feedback control system and project management in construction, as seen in Fig. 1 (the italics in Fig. 1 represent control theory's analogy to project management). For example, the plant in control theory can correspond to a construction project. The project is composed of many components, such as subcontractors, activities, resources, and equipment, which are all linked to each other. The project is implemented to produce its objective, as-planned performance (i.e., the desired state). The output of the project, as-built performance (e.g., the actual state), is monitored through a monitoring process (e.g., quality management process) analogous to the sensor in control theory. If there is any discrepancy between as-planned performance and as-built performance, control actions, analogous to controller in control theory, are taken to reduce this discrepancy. Additionally, there can be changes from outside the project analogous to the external disturbances in the control system, such as the owner's change request or a regulation change, which will disrupt project performance. These control actions and changes act as another input for the project, and the feedback processes will be reiterated during the life cycle of the project until the desired state is met.

The analogy from control theory implies that the dynamics is a major driver that renders the project management difficult, and the feedback can greatly intensify it. In construction, as-built performance is usually different from as-planned performance, so that significant efforts have to be made to reduce this gap. However, sensing as-built performance accurately is not an easy task. For example, the frequent manual collection of as-built performance requires a lot of effort from field crews and, further, is based on their subjective estimation. In addition, the taken control actions are not always appropriate and can even worsen the situation because they may be based on wrong as-built information. Furthermore, the decisions are often made under limited

time, budget, and resources. As a result, the gap can be increased, leading to chains of problems. The feedback that aims at stabilizing the project may actually intensify such dynamics.

In dealing with such situations, control theory provides valuable lessons to manage such dynamics. First, construction should be understood and managed as a whole including a sensor and a controller. Traditional management approaches have often focused on the project itself, particularly its operations. However, control theory shows that the project is so dynamic and feedback-driven that it cannot produce the desired state without the deliberate use of the sensor and the controller. Furthermore, construction is usually executed in an open environment and is therefore vulnerable to uncertainties, such as weather and differing site conditions. In addition, there are many change orders in the project, which also make it difficult to achieve the desired state. With respect to these issues, control theory suggests that the well-designed and implemented system with the help of a sensor and controller can stabilize the system despite such disturbance. Thus, there is a strong need for an approach that can take into account not only the project, the sensor, and the controller, but also, and more importantly, their interactions. In this way, the dynamics of the project can be better understood and controlled.

## 6 Theoretical Foundations of Dynamic Project Management (DPM)

Adopting SD as an implementation mechanism, DPM is proposed as a new method to manage dynamic complexities in construction projects. Its underlying philosophy is that construction is a system, with the parts working in coordination, which changes over time. Stemming from this philosophy, DPM focuses on the following characteristics to better understand and manage construction projects as a theoretical foundation: a holistic and a structure-oriented approach, and understanding of prevalent time delays. The following sections will investigate each of these foundations in detail.

### *6.1 Holistic Approach*

As discussed earlier, the design of a construction system including a sensor and a controller is essential to achieve the desired state of construction performance. Thus, a holistic approach that simultaneously considers the project, the sensor, and the controller is one of the core foundations taken by DPM. This assumes that the actual output of a project is different from the desired output so that the sensor and the controller should be deliberately designed as the system core. In this regard, change should be also considered as a natural part of construction. Usually, change is con-

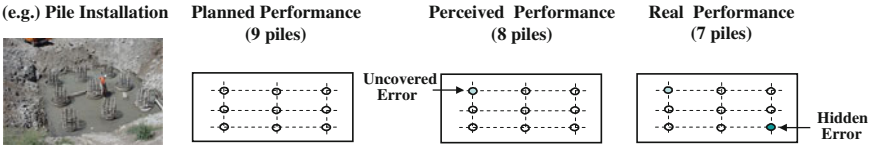



**Fig. 2** Planned, perceived, and real performance at pile installation example

sidered as 'out of control' because it often occurs from the outside project. However, an analogy from control theory reveals that the project can be stabilized based on the quality of the feedback control system despite outside disturbances. Designing a project indifferent to change may not be achievable. However, minimizing the impact of change is possible with the help of a well-designed sensor and controller, and this is worthwhile considering the overwhelming negative impact of changes. Thus, the project, sensor, and controller should be designed and implemented in an integrated fashion to cope with the dynamics of project management.

To this end, the use of automatic data capture (ADC) and computer simulation technologies has a great potential; the former for real-time monitoring and the latter for decision making support. For example, real-time performance information obtained from ADC can be input to computer simulation for diverse what-if scenarios of possible corrective actions (e.g., resource allocation strategies). In this way, the project, the sensor, and the controller can be integrated so that the performance gap can be addressed promptly and effectively.

## *6.2 System Structure-Oriented Approach*

An event is the particular happening at a point of the system's behavior and this dynamic behavior arises from the system structure [6]. System structure is identified as the interactions of two types of feedback: positive (or self-reinforcing) and negative (or self-balancing). A positive loop tends to reinforce or amplify whatever is happening in the system and a negative loop counteracts and opposes change [6]. Understanding the interactions between these feedback processes can greatly contribute to the management of dynamic behavior. It is also very useful to devise corrective actions analyzing their possible consequences.

Suppose we are installing nine piles for foundation, as illustrated in Fig. 2. If performance is measured by the number of completed piles, completing nine piles is the planned performance. However, if we find a pile with strength failure during a quality management process, eight piles are actually completed (i.e., perceived performance) and, thus, the gap between planned and perceived performance is one pile. In this case, managers take corrective actions, usually accompanied by an increase of scope, in an attempt to reduce this gap. For example, we may need to remove the existing erroneous pile and install a new one, which assigns additional scope. This

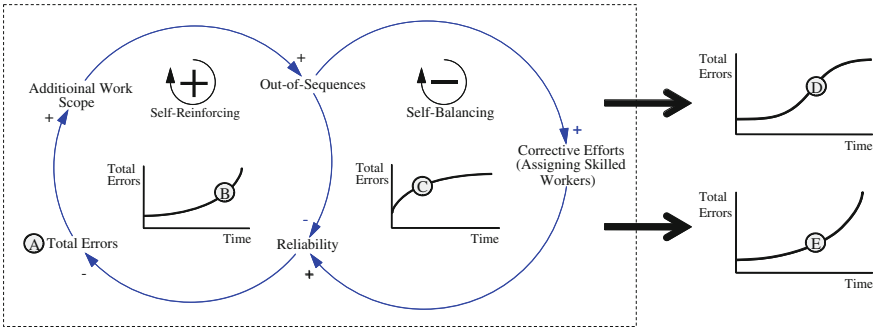



**Fig. 3** Interaction of feedbacks and resultant behavior

scope increase creates feedbacks that result in complex dynamics. As illustrated in Fig. 3, the scope increase caused by errors (i.e., Total Errors, A in Fig. 3) can disrupt a series of intended construction sequences. In this pile installation example, we may need more additional resources such as material, equipment, and man-hours, to remove and install a new pile putting on hold other succeeding activities. Further, if contingent resources are limited, procuring them may generate more issues, such as resource shortage in the succeeding activities. This will deteriorate overall project reliability, the degree to which performed tasks have been done correctly [20], leading to more Total Errors (A in Fig. 3). This will create a self-reinforcing feedback that amplifies out-of-sequence and can generate exponential growth behavior (B in Fig. 3) [6]. On the other hand, if serious out-of-sequence is experienced, a project manager can take corrective actions to rectify it. For example, more skillful workers can be assigned in order not to repeat the same error while increasing production rate. This aims at improving reliability and can reduce error. As such, the degree of out-of-sequence will be alleviated and a self-balancing feedback will be generated, which counteracts out-of-sequence and generates goal seeking behavior (C in Fig. 3) [6]. Finally, the interaction of these two different feedbacks generates complex dynamics. For example, based on the effect of corrective actions, different behavior of Total Errors (A in Fig. 3) can be observed. If assigning skilled workers is very efficient in this case, the S-shaped behavior (D in Fig. 3) can be achieved, thereby offsetting the effect of the self-reinforcing feedback. Otherwise, Total Errors continue to undergo exponential growth (E in Fig. 3).

In this regard, DPM emphasizes understanding of the system structure in order to effectively control the resultant project dynamics, thereby suggesting a system structure-oriented approach. In other words, since behavior arises from system structure, the behavior of particular interest can be controlled by changing the system structure. As such, a clear understanding of system structure is required, particularly when corrective actions are considered.

### *6.3 Incorporation of Time Delay*

Another characteristic of DPM is its appreciation of time delay. Dynamics behavior originating from the system structure can be more complex when time delay is outstanding. One of the most prominent forms of time delay is latency [28]. Continuing with the previous pile installation example, suppose that one of the piles is erroneous and has not yet been identified. In this case, even though eight piles are apparently completed, actually achieved performance is the completion of seven piles because the hidden one will be addressed at a later stage of the project, (i.e., latency) [28]. In this case, succeeding tasks, such as installing a column on this pile installation, may be already completed. If the hidden error is discovered after installing the column, the column may need to be removed before the erroneous pile, followed by the installation of a new pile and column. This creates increased additional work compared to the previous case and consequently makes the project more complex, which intensifies feedbacks. In addition, this latency involves a lot of waste, which can't be captured by only the increased work scope. For example, a lot of time can be used for request for information (RFI) to correct this erroneous pile and to decide what steps should be taken. Further, an additional quality management process should be taken in order to ensure its quality. Resource allocation should be rearranged to deal with such sudden and emergent work. In the worst case, a derivative activity can occur if the subcontractor for the piling activity has already been withdrawn. Thus, this 'invisible' effort used to address latency should be captured and minimized because it will eventually consume significant time and cost. In an effort to overcome this issue, DPM suggests that value, the ratio of the project requirements to the operational efforts [29], should be monitored and managed through the life cycle of the project. Value measures the operational efficiency by showing the extent the efforts contributed to the project requirements. In order to increase value, the method of minimizing the efforts should be investigated since the project requirements are almost constant. Particularly, the efforts caused by the performance gap and change like those in the pile installing example should be minimized because they do not add any value. DPM suggests a method for capturing and representing such efforts, but it is beyond the scope in this chapter. The interested readers can find it in [30].

On the other hand, time delay can also take place in the system structure and affect the intended impact of corrective actions. Continuing with the example in Fig. 3, suppose that skilled workers need to be shifted to the pile installation activity in order to deal with the scope increase caused by erroneous piles, but cannot be due to their shortage. In this case, another system structure that will represent other options such as hiring new skilled workers can be added. This option can be effective in producing the intended behavior, such as an S-shape curve in Total Error (A in Fig. 3). However, if it is not the case (e.g., due to the difficulty in hiring qualified workers or the excessive time taken for this hiring process), another change to the structure needs to be undertaken until the intended behavior is obtained. This iterative process will eventually lead to appropriate corrective action design.

## 7 Conclusions

Computer simulation has been utilized as an effective planning and analysis tool in the construction engineering and management area over the past decades. DES has dominated the development of construction simulation by virtue of operational details that are not readily provided by traditional network-based approaches. Under the pursuit of management at the lowest possible level encouraged by the decomposition-based approaches, DES has been primarily applied to address operational issues, by taking a narrow focus and partial view of a project. However, DES can sometimes provide unrealistic estimations, particularly when operational settings are significantly altered by other related operations. SD has great potential to address this limitation; however, it has received little attention in the construction engineering and management area.

In order to examine the opportunities and benefits of SD modeling, this chapter introduced DPM, which has been successfully applied to diverse infrastructure and building construction projects. Particularly, this chapter provided the following three theoretical foundations of DPM: a holistic approach, including planning and control functions of project management; a system structure-oriented modeling approach that enables deeper understanding of dynamic behavior and devising effective corrective actions; and the incorporation of control time delay that can make dynamic behavior more complex.

With these three theoretical foundations, DPM can successfully deal with dynamic complexities in construction projects that are not easily addressed by network-based approaches or DES, such as iterative cycles caused by errors and changes. However, as an SD-based approach, DPM inherits some of the weaknesses of SD modeling such as the lack of operational details or limitations in representing heterogeneous type of entities flowing into a stock. Therefore, the authors have been working on a hybrid simulation combining DES and SD as the next generation of DPM. This development will be reported in the authors' subsequent papers.

## References

1. Pritsker AB, O'Reilly JJ (1999) Simulation with visual SLAM and AWESIM. Wiley, New York
2. Pegden C, Shannon R, Sadowsky R (1995) Introduction to simulation using SIMAN. McGraw Hill, OH
3. Banks J, Carson II JS, Nelson BL, Nicol DM (2000) Discrete-event system simulation. Englewood Cliffs, Prentice-Hall, New Jersey
4. Law AM, Kelton DM (1999) Simulation modeling and analysis. McGraw-Hill, New York
5. Brailsford SC, Hilton NA (2000) A comparison of discrete event simulation and system dynamics for modelling healthcare systems. In: Proceedings of operational research applied to health services, 18–39
6. Sterman JD (2000) Business dynamics: systems thinking and modeling for a complex world. McGraw-Hill, Boston

7. AbouRizk SM (2010) Role of simulation in construction engineering and management. J Constr Eng Manage 136(10):1140–1153
8. Lee S, Han S, Peña-Mora F (2009) Integrating construction operation and context in large-scale construction using hybrid computer simulation. J Comput Civil Eng 23(2):76–83
9. Peña-Mora F, Han S, Lee S, Park M (2008) Strategic-operational construction management: hybrid system dynamics and discrete event approach. J Constr Eng Manage 134(9):701–710
10. Walsh KD, Hershauer JC, Walsh TA, Tommelein ID, Sawhney A (2002) Lead time reduction via pre-positioning of inventory in an industrial construction supply chain. Winter Simul Conf 2002:1737–1744
11. Williams T (2002) Modelling complex projects. Wiley, New York
12. Howell GA (1999) What is lean construction. In: Proceeding of the 7th annual conference of international group for lean construction, Berkeley, CA
13. Garcia-Fornieles JM, Fan IS, Perez A, Wainwright C, Sehdev K (2003) A work breakdown structure that integrates different views in aircraft modification projects. Concurr Eng Res Appl 11(1):47–54
14. Koskela LJ, Kagioglou M (2005) On the metaphysics of production. In: Proceeding of the 13th annual conference of international group for lean construction. Sydney, Australia
15. AbouRizk SM, Halpin DW, Lutz JD (1992) State of the art in construction simulation. In: Proceedings of the 1992 winter simulation conference, pp 1271–1277
16. Ondash SC, Maloney S, Huerta J (1988) Large project simulation: a power tool for project management analysis. In: Proceedings of the 1988 winter simulation conference los angeles. pp 231–239
17. Rodrigues A, Bowers J (1996) The role of system dynamics in project management. Int J Proj Manage 14(4):213–220
18. Morris PWG, Hough GH (1987) The anatomy of major projects: a study of the reality of project management. John Wiley & Sons Ltd, Chichester
19. Davidson FP, Huot J (1991) Large scale projects - management trends for major projects. Cost Eng 33(2):15–23
20. Lee S, Peña-Mora F, Park M (2005) Quality and change management model for large scale concurrent design and construction projects. J Constr Eng Manage 131(8):890–902
21. Martin R, Raffo D (2001) Application of a hybrid process simulation model to a software development project. J Syst Softw 59:237–246
22. Forrester J (1961) Industrial dynamics. Pegasus Communications, Waltham
23. Ford A (1999) Modeling the environment. Island Press, Washington
24. Sterman JD (1992) System dynamics modeling for project management. Working paper, MIT, Cambridge
25. Park M, Peña-Mora F (2003) Dynamic change management for construction: introducing the change cycle into model-based project management. Syst Dyn Rev 19(3):213–242
26. Han S, Love PED, Peña-Mora F (2011) A system dynamics model for assessing the impacts of design errors in construction projects. Mathematical and Computer Modelling advance online publication 22 June, doi:10.1016/j.mcm.2011.06.039
27. Ogata K (2002) Modern control engineering. Prentice Hall, Upper Saddle River
28. Lee S, Peña-Mora F, Park M (2006) Reliability and stability buffering approach: focusing on the issues of errors and changes in concurrent design and construction projects. J Constr Eng Manage 132(5):452–464
29. Thiry M (1997) Value management practice. Project Management Institute, Atlanta
30. Han S, Lee S, Peña-Mora F (2012) Identification and quantification of non-value-adding effort from errors and changes in design and construction projects. J Constr Eng Manage 138(1):98–109

# A Lean Framework for Production Control in Complex and Constrained Construction Projects (PC$^4$P)

**Søren Lindhard and Søren Wandahl**

Production conditions in construction are different than in the manufacturing industry. First of all, construction is rooted in place and conducted as on-site manufacturing. Secondly, every construction project is unique and a one-of-a-kind production, managed by a temporary organization consisting of several companies. Thirdly, highly interdependent activities have to be conducted at limited space, with multiple components, a lack of standardization, and with many trades and subcontractors represented on site. This interrelation results in a production where different contractors perform interacting and overlapping activities. This increases uncertainty and make the construction process very difficult to plan and control. Additionally, a lot of unpredictable factors (e.g. climate conditions) affects control, and makes construction even more complex.

Production control is an essential part of any complex and constrained construction project. Even though several tools have attempted to add structure and to create order, to the complex, dynamic, and uncertain context in which constructions is conducted, none has yet fully succeeded in providing a robust production control system. With outset in the lean tool Last Planner System of Production Control, a robust construction production control framework has been developed.

Søren Lindhard (✉)
Department of Mechanical and Manufacturing Engineering, Aalborg University, Aalborg, Denmark
e-mail: lindhard@m-tech.aau.dk

Søren Wandahl
Department of Engineering, Aarhus University, Aarhus, Denmark
e-mail: swa@eng.au.dk

H. Xu and X. Wang (eds.), *Optimization and Control Methods in Industrial Engineering and Construction*, Intelligent Systems, Control and Automation: Science and Engineering 72,
DOI: 10.1007/978-94-017-8044-5_14, 

## 1 The New PC$^4$P Framework for Onsite Scheduling

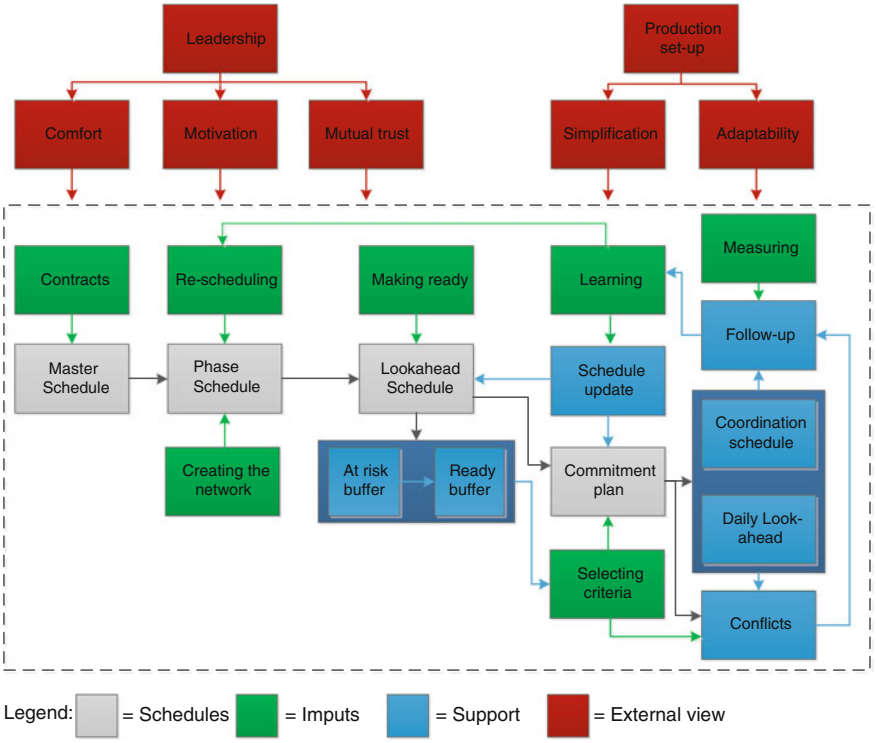



**Fig. 1** The PC$^4$P framework in its whole. With the outset in the basic PC$^4$P framework elaborating models will be presented in the following

The framework of the PC$^4$P control system is showed in Fig. 1. The PC$^4$P system is based on an open system-theory mindset and consists of components, connections, and inputs. The framework consists in total of four key schedules (marked with gray): the Master Schedule, the Phase Schedule, the Look-ahead Schedule, and the Commitment Plan. Moreover, the inputs to create the schedules are sketched (marked with green) together with support activities (marked with blue), which often is creating a link between schedules. Finally, the external control parameters (marked with red) are sketched.

The interplay between the applied elements in the PC$^4$P framework are making the system complete. The system is not stronger than the people who use it; therefore, it is crucial that all project participants understands the system and applies it correctly. Every applied element serves its unique purpose, and it is therefore critical if central parts of the production control system are omitted. Still the PC$^4$P framework should not be followed blindly, but the degree of formalization and the level of depth and

considerations putt into the schedule should suite the contextual conditions of the construction project, but still the system should be applied as a whole.

In the PC$^4$P framework the production characteristics of construction are handled through three main elements: Optimizing the sequence to production inflow, establishing inflow control, and continuous improving performance. The three elements aim towards a complete utilization of capabilities and possibilities in the production system. The sequence is important to the production workflow at site. Through a deliberate selection and ordering of activities, interdependencies are handled and production resources are allocated carefully to ensure a constant work output. Before an activity enters the work plans a making ready process is ensuring the soundness of the activity. Inflow control is established by checking-up on soundness which, by securing that only sound activities enter the Commitment Plans, minimizes the risk of interruptions and conflicts in the workflow. Continuous improvement of performance is a central part of the PC$^4$P framework to ensure that the capabilities and possibilities in the production system are completely utilized. The external factors are included in the framework because they create the world wherein the control system perfects utilization and is thus very important to performance. In the reaming pages in this chapter the developed system for production control is explained in detail.

## 2 The External View

The external environment is the outer context wherein the control framework functions and is thus having a huge impact on the performance of the control system [11]. Due to the dynamic nature, the external environment are constantly interacting with and influencing on the system. The external view consists of multiple external factors which by affecting behavior and processes affect production control at site [5]. Therefore, in the attempt to fully utilize the capabilities in the production system it is crucial important to include the external view in the framework. The production set-up is affecting processes at site; two parameters have been identified as crucial parameters: Simplification and adaptability. Management and leadership on-site are affecting behavior, and thus application of the control system. Three parameters have been identified as crucial parameters: Comfort, motivation, and mutual trust.

### *2.1 Simplification and Adaptability*

The complexity of the construction process is very much affected by decisions taken outside the boundaries of the control system. Two different but interrelated parameters have been identified as important to increase production control: Simplification and adaptability. The relationship can be viewed at Figs. 1 and 2.

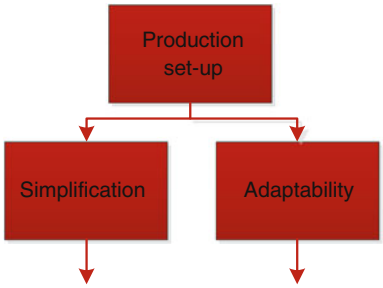



**Fig. 2** The production set-up is affecting the complexity and adaptability of the production on-site

Complexity is making construction projects difficult to control. Simplification is an approach to increase control by reducing project complexity. Moreover, simplification will, according to the Lean philosophy, reduce waste. Koskela [8] elaborate *the very complexity of a product or process increases the costs beyond the sum of the costs of individual parts or steps.* Limiting tasks and trades on site reduce interdependencies and increases process transparency [13]. Limiting task and trades on-site can be achieved by increasing prefabrication, preassembly and modularization. Moving production from site to a factory-like state makes it possible to streamline the production, to increase productivity, to improve output quality, and to reduce project lead time. Because complexity is moved outside the boundaries of the construction site, the remaining work on-site is simplified which decreases the needs of specialized craftsmen and the need of different trades to be present at site. Less specialization enables work crews to span diverse work activities which thus increase the flexibility and adaptability of the crews.

The negative effect of reduced task on-site is decreased process adaptability. The ability to respond on unforeseen events is increasingly depending on the supplier's flexibility in delivery. Flexibility in delivery is important especially in relation to the ability to make changes in orders and in deliveries. Damaged deliveries needs to be replaced quickly to avoid on-site delay while delay on-site makes is necessary to hold up deliveries to avoid inexpedient storing.

| *Simplification* | *Adaptability* |
|---|---|
| *Pros* | *Pros* |
| *Simplifies the process* | *Improves response time* |
| *Reduces interdependencies* | *Reduces waste* |
| *Increases process transparency* | |
| *Increased workforce adaptability* | |
| *Reduced lead time* | |
| *Reduces waste* | |
| *Cons* | *Cons* |
| *Decreased process adaptability* | *Increased cost* |
| *Increased supplier dependency* | |

Due to the complex, dynamic, and uncertain context wherein on-site construction is conducted changes is an inevitable fact. Adaptability is the ability to convert the production from one task to another. Thus, increased adaptability is enhancing the ability to respond to changes and unforeseen events and is thereby reducing waste in the adjustment process [13]. Adaptability can be achieved by improving flexibility in the production for instance by applying buffers or allowing crews to change work task and to adjust their work hours.

## 2.2 Comfort, Motivation, and Mutual Trust

According to the Lean-philosophy the capacity of the production system is equal the sum of work and waste [20]. Transformations are driven by the workforce present on site, and depending on their skill and motivation, both regarding output quality and quantity. Improving skills adds knowledge and expands the capacity of the production system while an improved motivation secures improved exploitation of capabilities inside the production system [10]. Lindhard and Wandahl [10] found that the theorem could be generalized and conclude that *Waste is to not fully utilize of the capabilities and possibilities in the production system.* The theorem expands Lean's existing seven types of muda (waste); see Suzaki [23] or Ohno [20], and defines the 8th source to muda.

*In Lean waste consist of:Muda (waste of resources)*

*Non value adding activities is creating waste. In Lean 7 types of waste is identified:*
*-Waste of overproduction*
*-Waste of stock on hand*
*-Waste of transportation*
*-Waste of making defective products*
*-Waste of processing itself*
*-Waste of movement*
*-Waste of time on hand*

*Mura (overburden)*

*Unreasonable demands on employees or processes, for instance high rates of work or unfamiliar work to which they are not qualified for.*

*Muri (unevenness)*

*Variation in work output within the production system. Muri is per se not waste but instead it leads to Mura and Muda.*

To fully exploit the capabilities of the workforce comfort, motivation, and mutual trust needs to be established. Comfort, motivation, and mutual trust are interrelated parameters; thus, improved comfort leads to improved motivation and increased mutual trust. Therefore, by increasing the comfort of project participants increased accountability, communication, and productivity is gained [21, 22]. Improved accountability produces dedication and in interaction with increased communication, the likelihood of observing the scheduled commitments are raises which lead to increased schedule robustness [10, 13]. The leadership of management on-site has to guide and support the construction process in order to foster comfort, motivation and mutual trust between all project participants. The relationship can be

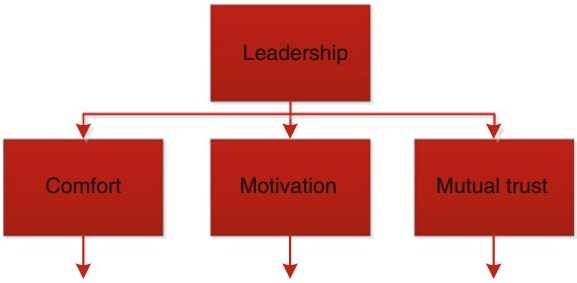



**Fig. 3** Leadership affecting the comfort, motivation and mutual trust between the project participants

viewed at Fig. 3. Leadership consists of multiple parameters among them: personality, charisma, and ethical values. As a leader your behavior is an important element in ensuring job-satisfaction and comfort. Moreover, the leader's behavior is contagious; thus, as a role model, being a good example is important.

## 3 Master Schedule

The Master Schedule, which can be viewed at Fig. 4, serves as guidance for the more detailed schedule. Therefore, at the Master Scheduling level the focus should be on creating overview to the upcoming construction process. Creation of the Master Schedule is based on contract restrictions and is thus outside site-management scope. Inputs to the Master Schedule are estimated durations adjusted to fit the contract set deadlines and milestones. It is important that the deadlines set by the contract is realistic, both a too tight and too slack time frame is undesirable.

When negotiating the time boundaries set by the contract it is important that the completion deadline is realistic; both a too tight and too slack time frame is undesirable [17]. A too tight time frame will be inflexible and thus unable to absorb variability in production and moreover induce a risk of overburdening (mura). A too slack time frame does on the other hand entail unexploited or wasted time deteriorated by the industry tendency to work best under pressure [17]. If possible the contract deadlines should be made flexible to encourage to increased collaboration and negotiation between contractor and client to create win/win situations and to move the construction industry away from contract bonded projects and bring both productivity and value creation up [17].

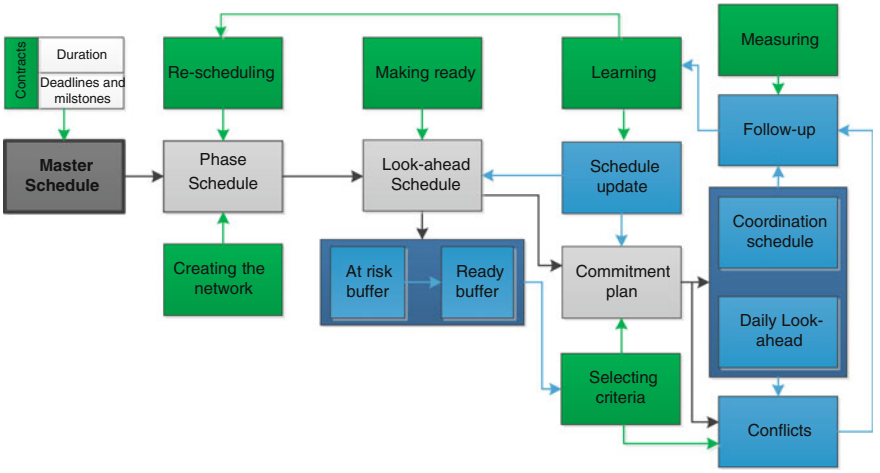



**Fig. 4** The master schedule creates the outer boundaries wherein the construction project is completed

## 4 Phase Schedule

At the Phase Schedule level the primary tasks is to create the overall network of activities, this is showed at Fig. 5. A Post-It session is used to identify durations, interrelationships, and handoffs wherefrom the overall connections is drawn. Afterwards the Critical Path Method (CPM) is applied to identify the critical path and possible float during the process. The PC$^4$P control system does not seek to complete the project as fast as possible, but instead to exploit the given time limits to increase the schedule robustness. Therefore, to minimize the risk of delay, float is, if possible, incorporated on the critical path.

To identify and consider all critical elements in the network, constraints which affect the work flow are incorporated as selection criteria. Thus, expanding the selection criteria improves not only the schedules quality but also helps in revealing and avoiding possible conflicts to evolve. Six constraints are identified as relevant to the sequencing and the selection of activities. The relevant constraints include: manning machinery and material which comprise the needed resources and working conditions, climate, and safety which affect the pace of the work [14].

*Post-It session:*

*The sequence is constructed by letting the involved companies order their activities on Post-It notes. It is important to include relation and connection to both previous and following activities on the notes. Afterwards the notes are put onto a wall and collaborative structured to achieve the best possible sequence (Ballard 2000; Ballard and Howell 2003).*

Normally, the demands from the six constraints are conflicting with each other; therefore, no optimal schedule exists. Thus, the refined network is a result of estimating and prioritizing between the constraints. Thus, it is the site-managers respon-

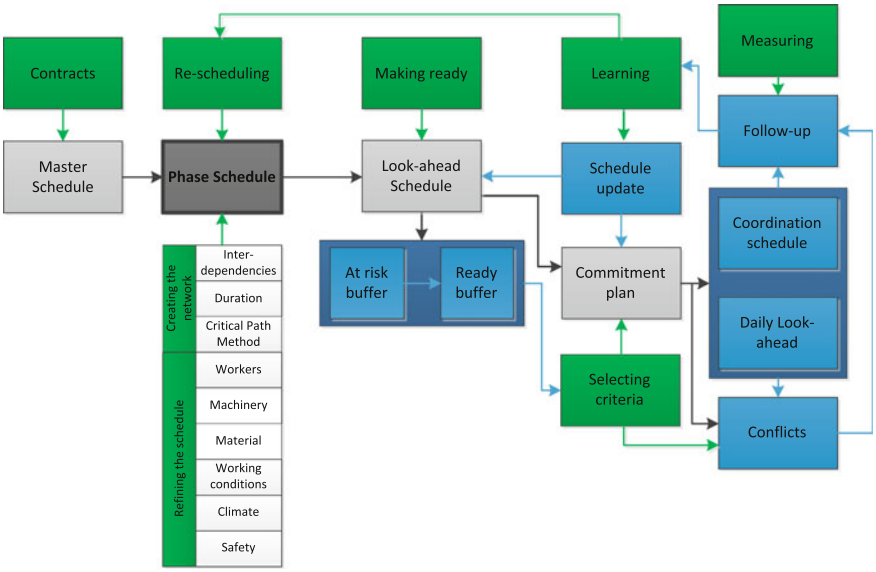



**Fig. 5** At the phase schedule the network of activities which defines the sequences is determined

sibility to ensure that the best possible sequence is achieved. It is important to state that incorporation of the six constraints into the schedule is a process which takes place after the Post-It session. The six constraints are described in the following.

*Critical Path Method:*

*Based on interrelationships and durations the earliest and latest that an activity can start and finish without extending the deadline is calculated. Free float is emerging if an activity can be delayed without affecting the subsequent tasks while possible delay in relation without affecting the project completion data is named total float. The longest path is determining the project duration and is called the critical path. Thus delay on the critical path delays the entire construction project.*

## *4.1 Workers*

The manning level on-site has an impact on labor performance [4]. It is important to avoid fluctuation in manning, especially within each trade, because it creates unevenness (muri) in the production which leads to waste (mura and muda). Moreover, by keeping a steady manning within the trades, extremes in the manning is avoided which eliminate the risk of overmanning; which decreases productivity [4].

Changing orders due to changes in schedules and plans decrease labor efficiency [3, 19], and should be minimized. When orders on site are changing the manning should ideally remain unaffected. Heighten the manning accelerates the work output

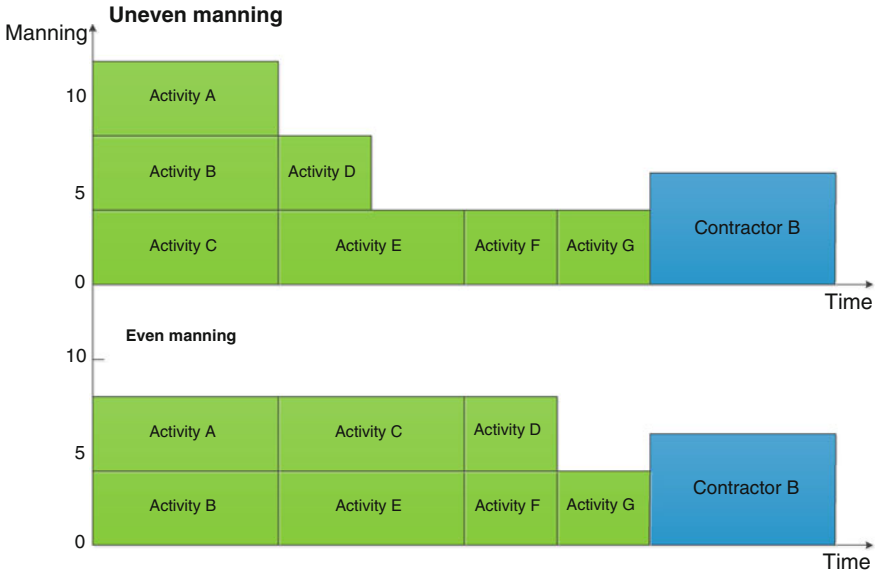



**Fig. 6** Example; adjusting the manning. Contractor A (marked with *green*) is secured an even manning by exploding the float and thereby moving activities

but reduces labor productivity [4], while lowering the manning decreases the work output but creates delay [12]. Finally, to improve output quantity and quality comfort of the individual craftsman should be secured.

To calculate the manning, the needed workforce to each activity first has to be estimated. Afterwards, the manning is summarized, for instance, from a Gant-diagram or a cyclogram, into a stacked column chart, an example can be seen at Fig. 6. Rearranging the sequence can be necessary to create a steady manning.

*Workers:*

*Define the needed workforce to each activity and calculate the manning throughout the construction project. Aim towards a steady manning. Moreover, to improve output quantity and quality initiatives to secure comfort of the individual craftsman should be implemented.*

## *4.2 Machinery*

The importance of considering utilization of required equipment and machinery is important mainly from an economical perspective. By compiling activities in relation to needed machinery, the utilization rates will increase and necessary presents will be restricted and the rental cost will be reduced [12]. The utilization rate can is calculated, in a Gant-diagram or cyclogram, by linking the needed machinery

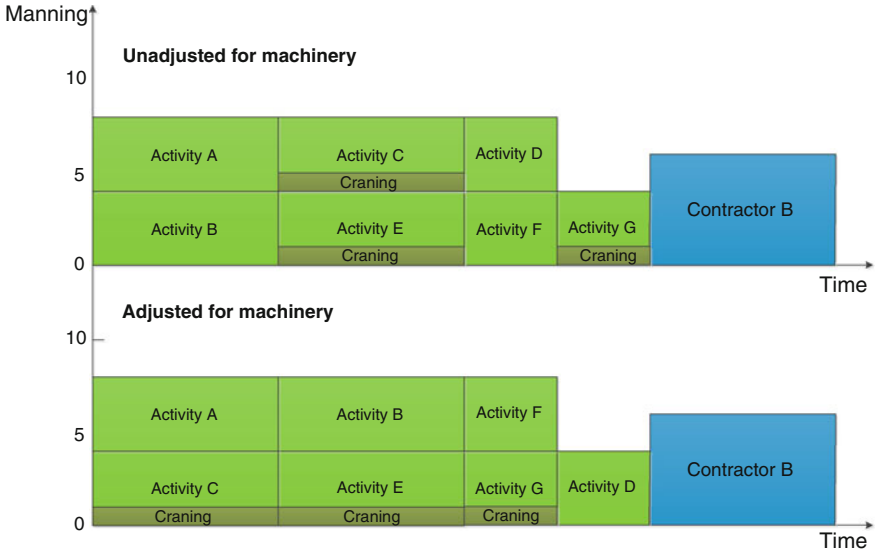



**Fig. 7** Example; adjusting usage of machinery. Contractor A (marked with *green*) is secured an even flow in equipment and machinery by moving activities within the limits of float and interdependencies

to each work activity. Afterwards, activities are rearranging to increase utilization and to avoid conflicts due to double usage, which easily is spotted. An example can be viewed at Fig. 7. It is important to notice that increased utilization rates of shared equipment increase the interdependencies and necessity of well-functioning machinery. Therefore, a small buffer between handoffs should be incorporated to absorb small variations in duration and thereby avoid an infectious delay. Finally, to minimize the risk and effect of critical breakdowns, maintaining of machinery should be considered and an emergency procedure should be completed [12].

*Machinery:*

*Link shared material and equipment to each activity. Group the activities to improve the utilization rates. Create a back-up plan to minimize the effect of breakdowns.*

## *4.3 Material*

A construction project consists of thousands of different and often unique materials or components which al, in time, has to present to complete the scheduled work activities. Material delivered in accordance with the just-in-time principle have an

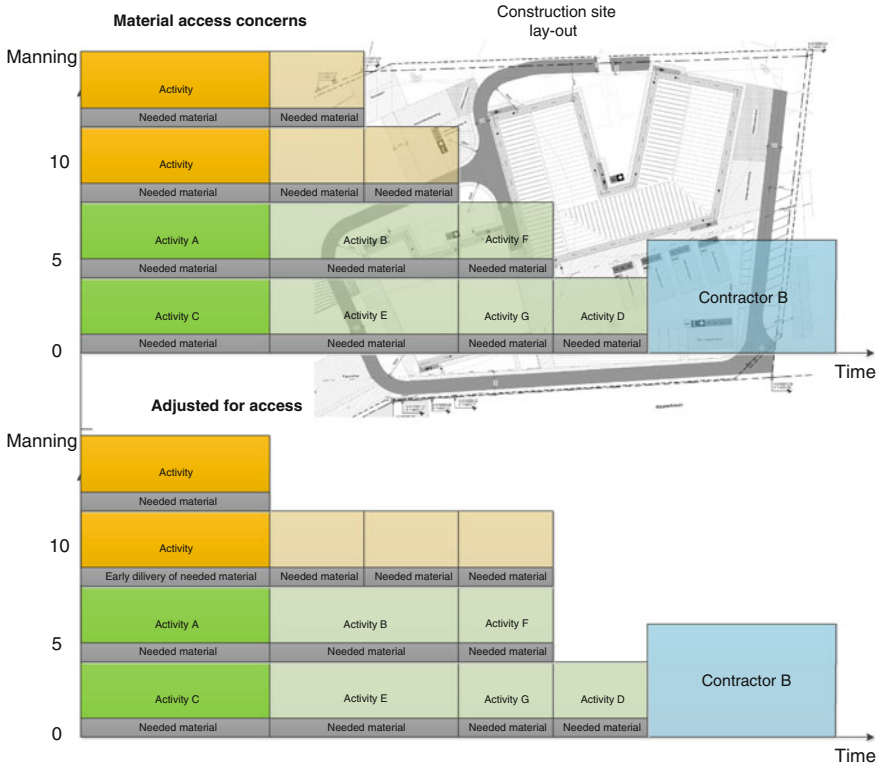



**Fig. 8** Example; adjusting for material access. Material demands are compared to the lay-out of the construction site to spot critical bottlenecks. Bottlenecks are avoided by moving either delivery or the whole activity. At the figure *green* is representing contractor A, *blue* is representing contractor B, and *orange* is representing contractor C

increased risk of non-presence at activity start, while material delivered to early have to be stored which increases cost [12]. To increase the flexibility of material deliveries, materials should not be pushed to the site by fixed material deliveries but should instead be pulled to site, thus delivered when needed. An approach to simplify the material flow would be by ordering materials in units containing all materials needed in a predefined room. To reduce the likelihood of dwindling or damaged materials, storing of materials has to be done carefully. To create an overview of the material flow, the material needed for each work activity is defined and stored in a material log. Afterwards the material flow is drawn in a cyclogram, or in a BIM-model and compared to the site lay-out, an example can be viewed at Fig. 8. The capacity of the access roads is estimated, to identify possible bottlenecks and to identify and consider relevant logistic issues. It is important to notice that the

site lay-out is dynamic and changes as construction progresses. Often capacity of the access roads is limited due to for instance periodic or time-bound restrictions. If capacity problems or bottlenecks are identified the material flow is adjusted either by controlling material deliveries or in extreme cases by rearranging the order of the activities.

*Material:*

*Define needed material to each work activity, and consider relevant logistic issues in relation to the material flow.*

## *4.4 Working Conditions*

Working conditions is important to ensure comfort of the craftsmen on-site. The constraint contains parameters to ensure working comfort and relevant location or space issues. Space issues includes access to work place, mutual interruptions and delays caused by shared work areas, etc. while Working comfort includes temperature, lighting, noise, working postures, working procedures, working base, etc. Working comfort is much related to traditional working environment issues, which is a part of safety. But where working environment is focusing on the safety and the health of the workforce working comfort is focusing on output and quality. Therefore, working comfort includes initiatives which go beyond the safety guidelines. Thus, working comfort is secured by identifying and controlling all relevant parameters to improve the working conditions.

To handle and optimize space issues, working areas and space requirements to every activity are defined. Afterwards, space usage is linked to the schedule to ensure that space is available; this can be achieved by applying Location Based Scheduling or by using BIM, an example to a cyclogram is showed at Fig. 9. Appling a visual approach is often an advantage because craftsmen often are very visual-minded. Finally, if the space and working conditions considerations has revealed critical elements in the sequence necessary rearrangements are made.

*Working location and comfort:*

*Define the working area and space requirements to each activity. Ensure that space is available by linking usage to the schedule. Identify all elements which affect working comfort and seek to improve the conditions.*

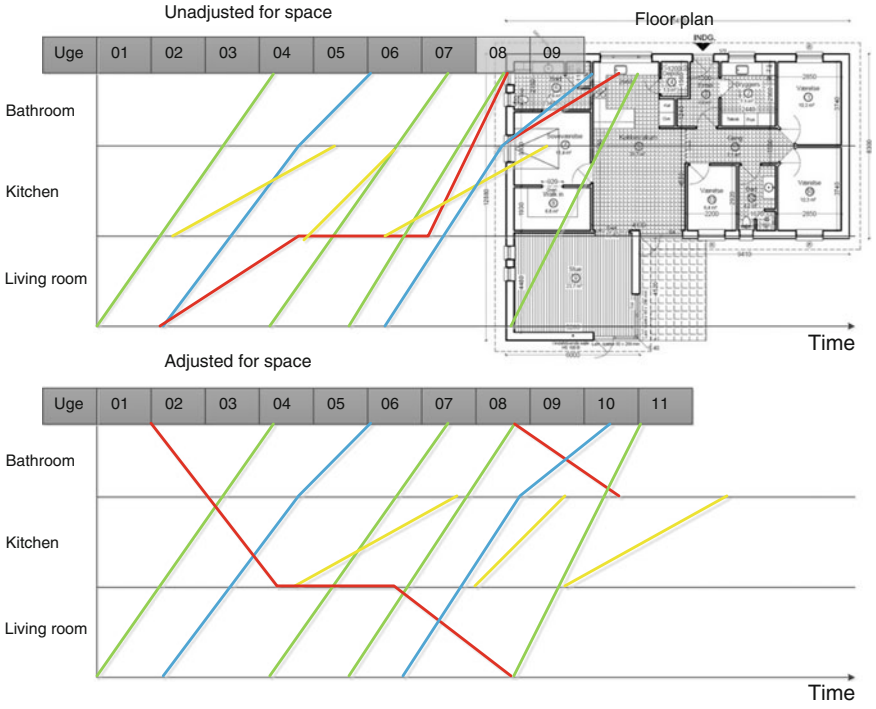



**Fig. 9** Example; adjusting for space usage. Applying a cyclogram and compare it to the floor plan to identify insufficient space and adjusting sub-sequences. The different collars in the cyclogram represent the work of different contractors

## *4.5 Climate*

*Climate precautions:*

*Relevant precautions to consider includes: Covering, heating, snow removal, water protection etc.*

Every construction project is surrounded by its unique, complex, and changing external climate. The external climate does by a number of parameters such as temperature, wind, moisture, rain, snow, waves, and visibility affect the work conducted at site. Even though the climate itself is unmanageable the impact of the climate can be reduced. Despite most climate-parameters to some extent are following the season, huge variation in actual climate conditions is making the impact especially at long term close to impossible to forecast. Climate precautions are most often very expensive to install and should therefore only be applied when necessary. Some climate precautions can be implemented at short term, but most climate precautions have to be installed at a long term basis, due to a time-consuming installation and high installation costs. Long term precautions are problematic because installation

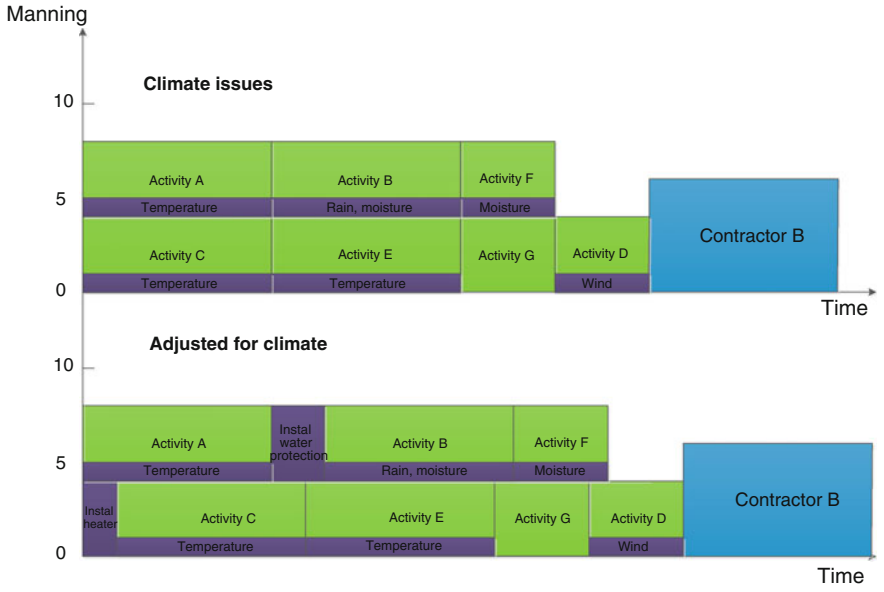



**Fig. 10** Example; incorporating climate conditions. Contractor A (*marked with green*) is handling climate conditions by noting and incorporating relevant climate precautions into the schedule. The notes serve as a reminder to climate concerns

has to take place before the climate impact and thus the possible effect is known. The unpredictability and complexity of climate impact preclude traditional cost-benefit thinking when considering climate precaution, instead decisions is based on a risk assessments [12]. Therefore, the first step to incorporate climate into the schedules is to identify risks. When the risk profile containing critical scenarios is determined, the second step is to consider the effect and cost of possible precautions. The third step is to compare risks to the effect and costs of the individual precaution and to determine if the precaution is selected or not. Based on the selection of climate precautions a climate log is formulated containing a set of thought through actions to handle the different risks if an incident occurs. Moreover, the installment of the climate-precaution is incorporated in the sequence; an example can be seen at Fig. 10.

*Climate:*

*Identify critical climate parameters, consider possible precautions, and make a plan of actions to different critical scenarios.*

## 4.6 Safety

*Safety precautions:*
*Relevant precautions to consider includes: safety distance, fall protection, covering of unsafe areas, access roads, etc.*

Safety of the workforce is crucial. No activities should be conducted if it induces any unnecessary safety-risks. Thus if safety is not considered, inappropriate solutions might end up stopping the production. Therefore, to hinder accidents or hazards in evolving and to avoid interruptions in the work flow safety risk needs to be identified. Based on the revealed risks, safety precautions need to be considered, selected and implemented. Besides direct safety fulfillments cf. the national "Health and Safety at Work Act", other preventive precautions include: safety inspections, safety trainings, hazards planning, alcohol screening etc. [6] and could be combined with a increased safety awareness in an attempt to hinder problems in developing. Finally, the sequence is adjusted in relation to the effects of the safety precautions, an example can be viewed at Fig. 11.

*Safety:*
*Identify necessary safety precautions to the individual activity and plan for implementation.*

## 4.7 Re-scheduling

In the search for continuous improvement the Phase Schedule has, at selected repetitive tasks, to be re-thought. By returning to the scheduling phase process, positive and negative experiences can be discussed, and overlooked sub-activities and problems can be incorporated into the schedule. Thus, the re-scheduling of the Phase Schedule does create an opportunity for mutual-learning and to incorporate that learning into the schedule when the processes is repeated. Re-scheduling is moreover relevant if the basis of the schedule (e.g. durations, interdependencies, constraints, critical path, or the construction design) changes; therefore, these predefined conditions needs to be continuously monitored [11]. To simplify the re-scheduling process, the sequence is divided into key phases. The placement of the re-scheduling process is a balance between time, spend learning, and time left in which the changes become effective. Normally the re-scheduling is paced after an initial test round, (for instance a month into the phase), or half way through the phase. Moreover, to get the full potential out of the re-scheduling process, it should be combined with traditional waste reduction. If the detail-level is increased while re-thinking the sequence waste could be identified and removed.


**Fig. 11** Example; incorporating safety. Contractor A (marked with *green*) is securing a safe working environment by noting and incorporating relevant safety precautions into the schedule. The notes serve as a reminder to safety concerns

## 5 Look-Ahead Schedule

The Look-ahead plan, which can be viewed at Fig. 12, contains a making ready for conduction principle, called the making ready window. It basically implies that project manager focuses on future activities and ensures that they will be conductible. The window is sliding forwards as construction progresses to ensure that activities are sound when entering the Commitment Plan. The length of the window depends on how time-consuming the making ready process is, but will normally be a six weeks window. When an activity enters the Look-ahead window a making ready process is launched. During the making ready process all constraints are removed to ensure that the activity can start and finish on schedule [11, 15]. If one of the constraints is not removed the task cannot be conducted. To avoid unfulfilled preconditions to be overlooked Lindhard and Wandahl [15] categorized the preconditions into nine main categories: (1) Known conditions, (2) construction design and management, (3) connecting works, (4) workforce, (5) equipment and machinery, (6) components and materials, (7) working location and comfort, (8) climate, and (9) safety.

The pace of the making ready process needs to be kept high, in order to continuously feed the Commitment Plans with ready activities [16]. The key rule when making the schedule is that activities if possible should be fit to capacity and not capacity to activities [16]. Thus, lowing the manning will fit capacity to activities

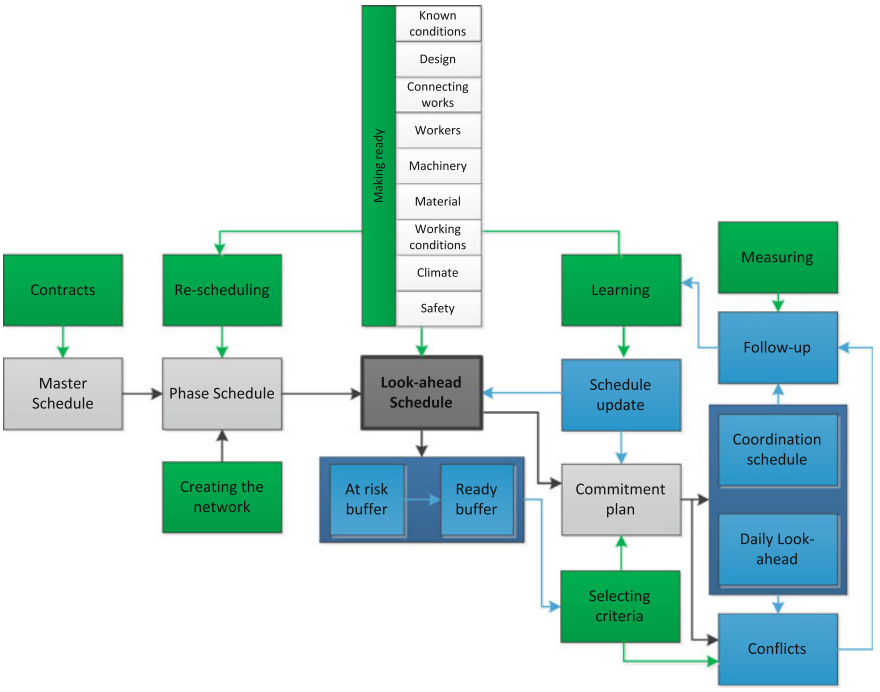



**Fig. 12** At the look-ahead schedule activities are made ready to ensure increased schedule reliability

and thereby slow down the production resulting in delay and waste, cf. not exploiting the capability in the construction system was earlier mentioned as the 8th source to waste.

It is important that the making ready process is pursuing optimal fulfillment of the precondition. Optimal fulfillment will reduce the risk of varying soundness in the preconditions resulting in reduced risk of non-ready activities in the schedule; moreover, the output ratio is improved resulting in increased productivity. Thus, the presents and the quality of the fulfillment of every precondition are important [11]. Communication and collaboration between contractors and site management is an important part of the making ready process and increases both schedule quality and conflict awareness [11].

When all preconditions have been fulfilled the activity is now regarded ready and moved to a buffer of ready work. At risk activities, i.e. activities where soundness is not yet achieved but instead is based on an anticipated fulfillment, are buffered separately in an at risk buffer until the activity enters the Commitment Plan or the risk is eliminated [12]. If the risk is eliminated the activity is moved to the buffer containing ready work, cf. the arrow on Fig. 12. Most often at risk activities occurs due to late or just-in-time deliveries of materials.

*A guiding list of possible constraints:*

*1 Known working conditions*

*a) Unknown working conditions which cause changes to plans*

*Be aware of:*

*Asbestos, rot or other unexpected conditions such incorrect or outdated drawings which create a misleading apprehension of the existing structure or soil conditions.*

*2 Construction design and management*

*a) Sufficient and correct plans, drafts, and specifications need to be present.*

*Be aware of:*

*- Drawings with wrong measurements*

*-Outdated drawings*

*-Missing drawings and clarification of -project details*

*-Missing approval of project design or details*

*b) Legal aspects*

*Be aware of:*

*-Government authorizations*

*-Building laws and Eurocodes*

*-Contracts and agreements*

*c) Communication, coordination, collaboration and individual mistakes*

*Be aware of:*

*-Misconceptions and oblivions due to high working pressure or lacking skills/experience.*

*d) Adjustments in the schedule*

*Be aware of:*

*Changes made to optimize the sequence*

*To keep the conducted schedule realistic and ensure that it can be followed.*

*Changes in soundness of activities can force the schedule to be changed.*

*A complex and changing environment forces the schedule to be rethought.*

*Incorrect time estimates*

*3 Connecting works*

*a) Completion status of connected activities*

*Be aware of:*

*Previous activities not completed according to schedule*

*The increased risk of including "at risk activities" in the commitment plan*

*b) Rework in previous activities causing delay*

*Rework caused by insufficient quality*

*Rework caused by damages caused to completed work.*

*4 Workers*

*a) Workers need to be present*

*Be aware of:*

*Illness in the workforce*

*Unexpected or overlooked vacation*

*Contractors not keeping their commitments by not showing up*

*b) Workers need to be qualified*

*Be aware of:*

*Changes in the workforce can result in lower output quality and quantity.*

*5 Equipment and machinery*

*a) Correct equipment and machinery is present*

*Be aware of:*

*Delayed equipment*

*Equipment used by other contractors*

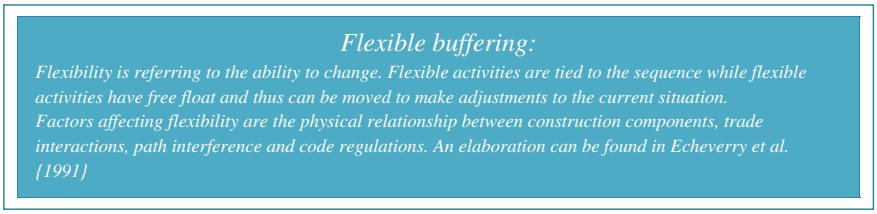
*Incorrect equipment which is not fitting the work task*

*Breakdowns in equipment*

Buffering creates a link between the Look-ahead schedule and the Weekly Work Plans, where ready or at risk activities are selected from the buffers to fill the work plans with sound work activities [11]. On-site construction is complex and unpredictable resulting in unforeseen events and errors affecting the soundness of activities. According to Lindhard and Wandahl [18], *Every precondition is a variable and composes a possible obstruction for a given assignment to be fulfilled.* Buffering increases process adaptability and thereby minimizes the effect of "error" by absorbing undesired variability to maintain a constant workflow [11].

Because of the related cost and complexity buffering should be limited to comprise next week's work. Moreover, if possible the existing buffer should be replaced with flexible buffer activities, cf. [2]. Bertelsen elaborates [1] *Many projects activities are not inter-dependent and may be executed in any sequence or even simultaneously without any effect on the overall result.* Flexible activities are not tied into the

sequence and can therefore be stored in the buffer until needed. Thus, by using a flexible buffer, activities can handle variation without affecting the future sequence [16].


## 6 Commitment Plans

Production control is founded on commitments between project participants. The quality of the schedule is depending on the quality of the settled commitments [12]. At the point when an activity enters the Commitment Plan a binding commitment is made. The Commitment Plan can be viewed at Fig. 13. To secure high quality commitments, the site-manager needs to be prepared. Preparation includes insight to the construction stage and its impact on sequencing, critical path, and the other selection characteristics. Thus, the site-mangers needs to draw lines back to the initial plans to understand the effects of possible changes. If these lines are not drawn there is actually no reason to conduct Phase Scheduling. Despite the site-manger on

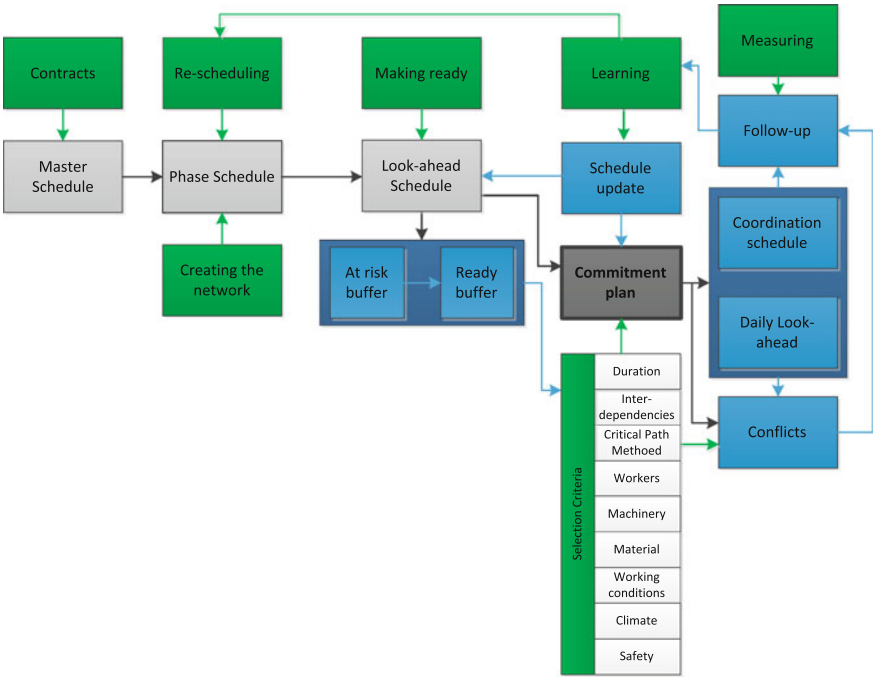


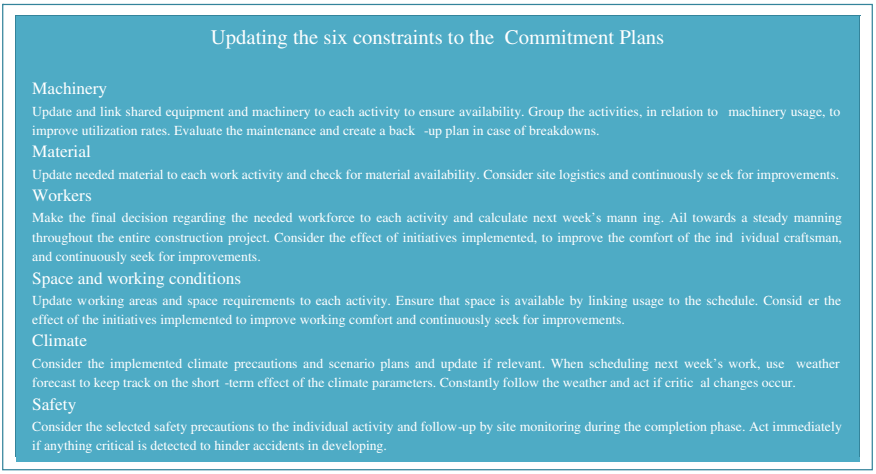

**Fig. 13** The commitment plan is the final level of scheduling and is setting the scene for the actual production at site. The commitment plan is supported by a coordination schedule to organize and daily look-ahead to keep track of progress at site

beforehand know the process he/she want, the site-manger still has to be open for changes and for details he/she might have overlooked. By allowing the craftsmen to influence the process, ownership to the schedule is ensured.

In the search for improved schedule quality the commitments have to be settled in mutual agreement and with the best possible information on hand [12]. To procure the information the schedule has to be updated to reflect the construction sites current situation. Based on the completion stage of the individual activity adjustments in the schedule has to be made to avoid any upcoming conflicts in handoffs.

Since the fulfillment of a precondition can change, a health check of the buffer should be implemented [18]. Thus, the health check does minimize the likelihood of non-ready activities entering the Commitment Plan [13]. By detecting changes on beforehand adjustments can be made to avoid conflicts between handoffs and to increase schedule quality and reliability [12, 18].

Finally, the six preconditions which are linked to the schedule at the Phase Scheduling level need to be reincorporated to the schedule. This is achieved by systematically following the six preconditions and continuously update and integrate the results into the schedule.




With this information on hand the scheduling proceeds. Selection of activities to the Commitment Plan is based on the characteristics of the individual activity. Thus, duration, interdependencies, float, critical path, and the six relevant constraints are considered, i.e. the same parameters which were applied to refine the network chart at the Phase Scheduling level. At the Commitment Plan level the capabilities in the production system are known, thus all adjustments in the schedule are a coordination of pre-defined parameters.

A systematical integration of the procured information increases the quality of the settled commitments. Increased commitment quality decrease the likelihood for changes in the schedule. Lowering the risk of changes in the schedule makes the schedule trustworthy and reliable and most importantly binding for all project participants [9]. If the plan is continually changed it loses its credibility and in worst case execution is separated from planning [7].

Making a Commitment Plan can be very time-consuming. Meetings should be limited to avoid long sessions of inactivity which is resulting in decreased concentrations followed by low scheduling quality and slow progress. It is the site-manger who is responsible for successfully organizing of scheduling meetings. To avoid inefficient and time-consuming meetings it is important to consider task relevance and detail level for discussions in plenum. If possible, inactivity could be reduced by dividing the meetings in relation to physical relationships (e.g. outdoor and indoor work).

## *6.1 Coordination Schedule*

To support communication and coordination on-site a Coordination Schedule is created. The Coordination Schedule supports the Commitment Plan by clarifying and fostering communication on-site. By structuring the needs of and clarifying the lines of communication, coordination onsite is simplified. The schedule contains interrela-

tionships and bonds between the activities. Moreover, it contains a list of relationships and the needs for coordination, together with the one responsible. Thus, applying a Coordination Schedule is supporting a decentralization of responsibility and force and supports the subcontractors to communicate.

### *6.2 Daily Look-Ahead*

Interruptions and conflicts in scheduled activities are making it necessary to focus on creating soundness. Daily Look-ahead is implemented as an element to identify and handle sudden conflicts. Conflicts are identified by every morning letting the foremen and if necessary the site-manger briefly check-up on the soundness of the scheduled activities. Especially the status of at risk activities are important to check-up, for instance by checking the presents of workforce and materials. The Daily Look-ahead which is an attempt to spot emerging conflicts as fast as possible is supported by a general soundness awareness. Early conflict identification release time to make adjustments to avoid interruptions and stops in the workflow. Identified non-ready activities are replaced with ready activities from the buffer where characteristics of the individual activity once again are decisive.

Communication and collaboration are important to secure an optimal handling of arisen conflicts [14]. It takes teamwork to work around the changes to find and exploit new possibilities and to optimize the process. Furthermore, communication and collaboration between the project participants are essential to avoid misunderstandings when implementing the changes [11, 14].

### *6.3 Follow-Up*

Halfway through the week, the site-manager walks a round on site to follow up on the Commitment Plan, if anything critical is observed it is in a mutual agreement settled how to intervene to ensure that the activity can finish on schedule. If the site-manager realizes that an activity is being completed ahead of schedule he/she needs to communicate and coordinate changes with subsequent subcontractors to secure that the gap is exploited. Thus, it is just as important to ensure that the following activity is starting on time as it is that the current activity is finishing on time. Besides the wasted time-gap, rushing to finish a deadline just to discover that time is waste can be demotivating. Gaps can be exploited by; Step (a) ensuring that the crew finishing an activity before expected can continue their work. To ensure that the crew can continue their work their activity needs to be grouped and is thus an extra argument for keeping an even manning. Step (b) ensuring that any connecting activities are able to start as fast as possible. Interdependencies between the multiple trades on site makes it difficult to adjust the sequence because the next trade is often occupied elsewhere, not-aware of the gap, or simply not ready to start the conduction of the following activity.

After completing the work-week corresponding to the Commitment Plan, output quantity and quality are controlled. The measurements are used as inputs to the mentioned schedule update.

## Measuring Performance by Calculating Man-Hours (MPH)

The consumption of man-hours to each work activity is calculated and compared to the hours completed in the schedules. Thus, the comparison will result in a direct measurable size of possible delay. A man-hour status can be calculated by summing positive and negative variation in output. Calculating performance per activity makes it possible to follow the activities. To gain extra insight to the performance on site, the calculation can be expanded to include material usage, utilization of machinery, space and workforce etc. Repeated activities should normally be conducted faster and with less effort. Hence, it is a sign of conflict if an activity starts to consume extra resources for instance man-hours. It is important to note that most conflict can be revealed and handled through communication. Talk to the subcontractors, they will probably know the problem, and be able to provide you with increased insight. In any case, communication is important in order to reveal and eliminate root-causes to deviations.

*MPH Calculation*

$MPH = d_2 \cdot m_2 - d_1 \cdot m_1 \cdot cs;$

$d_1 = scheduled\ duration; d_2 = actual\ duration$

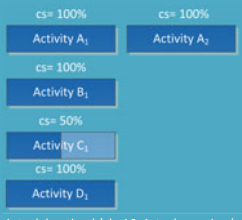
$m_1 = scheduled\ manning; m_2 = actual\ manning$

$cs = completion\ stage$

If the result is positive the activity was finished behind deadline while a negative result is revealing that the activity was finished ahead of deadline.

Examples:

Scheduled duration ($d_1$) =10; Scheduled manning ($m_1$) =3




Actual duration ($d_2$) =16; Actual manning ($m_2$) =2

$MPH\ A_1 = d_2 \cdot m_2 - d_1 \cdot m_1 \cdot cs = 16 \cdot 2 - 10 \cdot 3 \cdot 1 = +2$

Actual duration ($d_2$) =10; Actual manning ($m_2$) =4

$MPH\ B_1 = d_2 \cdot m_2 - d_1 \cdot m_1 \cdot cs = 10 \cdot 4 - 10 \cdot 3 \cdot 1 = +10$

Duration:($d_2 = d_1$);Actual manning ($m_2$) =2

$$MPH\ C_1 = d_2 \cdot m_2 - d_1 \cdot m_1 \cdot cs = 10 \cdot (2 - 3 \cdot 0{,}5) = +5$$

$MPH\ D_1 = d_2 \cdot m_2 - d_1 \cdot m_1 \cdot cs = 10 \cdot (2 - 3 \cdot 1) = -10$

$$Total = \sum MPH = MPH\ A_1 + MPH\ B_1 + MPH\ C_1 + MPH\ D_1$$

Actual duration ($d_2$) =14; Actual manning ($m_2$) =2

$MPH\ A_2 = d_2 \cdot m_2 - d_1 \cdot m_1 \cdot cs = 14 \cdot 2 - 10 \cdot 3 \cdot 1 = -2;$

$Improvement: MPH\ A_2 - MPH\ A_1 = (-2) - 2 = 4$

If a not scheduled activity is completed the result is inversed

Scheduled: $d_1; m_1 = 0$; Actual duration ($d_2$) =14; Actual manning ($m_2$) =2

$$MPH\ E_1 = -(d_2 \cdot m_2 - d_1 \cdot m_1 \cdot cs) = -(14 \cdot 2 - 0 \cdot 0 \cdot 1) = -28$$

### Measuring Quality

Output quality is important. A clear picture of performance is only achieved when the effect of poor quality and defects are deducted from the initial performance. Rework can be used as an indicator for unacceptable quality, and hours spend on rework can be added to the MPH calculation. If a clarify picture of quality is requested, a quality control check should be implemented in a handover process between handoffs. The quality control check could be undertaken by either site management or the successive work crew.

### Learning

Continuously improvement is a central part of Lean Construction. In the $PC^4P$ system continuous improvement is secured through a learning process. The Re-scheduling process provides feedback at the Phase Scheduling level while the Learning process accumulates on-site experiences and serves as feedback to the Look-ahead Schedule and the Commitment Plans. The Learning process is a part of the site manager's rounds on-site, where conversations with the men on-site are essential. By discussing the current progress both positive and negative experiences are gathered. The lessons learned helps in adjusting the upcoming process and to continuous improve. Negative experiences, which surfaces as conflicts and non-completions are reduced by tracking down root-causes to avoid repetitions. Moreover, understanding the triggers can help in predicting future conflicts [14]. Positive experiences, which surfaces as genius solutions and ideas, are preserved through reflection and discussions to understand and accumulate the experiences.

## References

1. Bertelsen S (2003) Complexity—construction in a new perspective. In: Proceedings for the 11th annual conference of the international group for lean construction, Virginia
2. Echeverry D, Ibbs CW, Kim S (1991) Sequencing knowledge for construction scheduling, J Constr Eng Manag-Asce 117(1):118–130
3. Hanna AS, Russell JS, Gotzion TW, Nordheim EV (1999) Impact of change orders on labor efficiency for mechanical construction. J Constr Eng Manage 125(3):176–184
4. Hanna AS, Taylor CS, Sullivan KT (2005) Impact of extended overtime on construction labor productivity. J Constr Eng Manage 131(6):734–739
5. Hartley J (2004) Case study research. Essential guide to qualitative methods in organizational research. Cassell, Catherine; Symon, Gil, SAGE Publications Inc, CA, pp 323–333
6. Howell G, Ballard G, Abdelhamid TS, Mitropoulos P (2002) Working near the edge: a new approach to construction safety. In: Proceedings for the 10th annual conference of the international group for lean construction, Gramado
7. Koskela L, Howell G (2001) Reforming project management: the role of planning, execution and controlling. In: Proceeding for the 9th annual conference of the international group for lean, construction, Singapore, 6–8 Aug 2001

8. Koskela L (1992) Application of the new production philosophy to construction. Standford University, Standford
9. Lindhard S (2013) Exploring the last planner system in the search for excellence, special report no. 95, Aalborg University
10. Lindhard S, Wandahl S (2012) Adding production value through application of value based scheduling. In: Proceeding of COBRA 2012-RICS international research conference, Las Vegas
11. Lindhard S, Wandahl S (2013a) Improving onsite scheduling—looking into the limits of last planner system. Built Hum Environ Rev 138
12. Lindhard S, Wandahl S (2013b) Looking for improvements in last planner system: defining selection criteria. In: Proceedings of ICCREM 2013 international conference on construction and real estate management. ASCE Press, USA
13. Lindhard S, Wandahl S (2013c) On the road to improved scheduling: reducing the effects of variations in duration. In: Proceedings of the 2013 international conference on construction and real estate mamagement. ASCE Press, USA
14. Lindhard S, Wandahl S (2012a) Designing for second generation value—future proofing constructions. In: Kashiwagi D, Sullivan K (eds) Proceedings of the construction, building and real estate conference, Arizona State University, Tempe
15. Lindhard S, Wandahl S (2012b) Improving the making ready process—exploring the preconditions to work tasks in construction. In: Proceedings for the 20th annual conference of the international group for lean construction, Montezuma Publishing, San Diego, 17–22 July 2012
16. Lindhard S, Wandahl S (2012c) On the road to improved scheduling—fitting activities to capacity. In: Kashiwagi D, Sullivan K (eds) Proceedings of the construction, building and real estate conference, Arizona State University, Tempe
17. Lindhard S, Wandahl S (2012d) The robust schedule—a link to improved workflow. In: Tommelein ID, Pasquire CL (eds) Proceedings IGLC-20: Conference of the international group for lean construction, Montezuma Publishing, San Diego, 17–22 July 2012
18. Lindhard S, Wandahl S (2011) Handling soundness and quality to improve relaibility in LPS—a case study of an offshore construction site in Denmark. In: Ruddock L, Chynoweth P (eds) COBRA 2011, Salford, 11–12 Sept 2011
19. Moselhi O, Leonard C, Fazio P (1991) Impact of change orders on construction productivity. Can J Civ Eng 18(3):484–492
20. Ohno T, (1988) Toyota production system: beyond large-scale production. Productivity Press, Boca Raton
21. Olomolaiye PO (1988) An evaluation of bricklayers' motivation and productivity, Loughborough University
22. Singh J (1996) Health, comfort and productivity in the indoor environment, indoor and Built Environ, 5(1):22–33
23. Suzaki K (1987) The new manufacturing challenge: techniques for continuous improvement. Free Press, London

# Optimization in the Development of Target Contracts

**S. Mahdi Hosseinian and David G. Carmichael**

**Abstract** Target contracts are commonly used in construction and related project industries. However, research to date has largely been qualitative, and there is no universal agreement on how any sharing of project outcomes should be allocated between contracting parties. This chapter demonstrates that by formulating the sharing problem in optimization terms, specific quantitative results can be obtained for all the various combinations of the main variables that exist in the contractual arrangements and project delivery. Such variables include the risk attitudes of the parties (risk-neutral, risk-averse), single or multiple outcomes (cost, duration, quality), single or multiple agents (contractors, consultants), and cooperative or non-cooperative behaviour. The chapter gives original, newly derived results for optimal outcome sharing arrangements. The chapter will be of interest to academics and practitioners interested in the design of target contracts and project delivery. It provides an understanding of optimal sharing arrangements within projects, broader than currently available.

## 1 Introduction

When an owner (principal) engages a contractor (agent), the contractor performs effort (at cost) that leads to some project outcome which is observable both to the contractor and owner [58]. The outcome is not only dependant on the contractor's effort but is also affected by events which are outside of the contractor's influence. That is, there exists outcome uncertainty. A detailed review of uncertainties in construction projects can be found in Barnes [13], Rahman and Kumaraswamy [60] and El-Sayegh [34]. The contractor's effort cannot be fully monitored by the owner. That

S. M. Hosseinian · D. G. Carmichael (✉)
School of Civil and Environmental Engineering, The University of New South Wales, Sydney, NSW, Australia
e-mail: D.Carmichael@unsw.edu.au

H. Xu and X. Wang (eds.), *Optimization and Control Methods in Industrial Engineering and Construction*, Intelligent Systems, Control and Automation: Science and Engineering 72,
DOI: 10.1007/978-94-017-8044-5_15, © Springer Science+Business Media Dordrecht 2014

is, there exists information asymmetry [50, 59]. Due to the existence of outcome uncertainty and asymmetric information, an opportunist contractor may act in its own interests instead of the owner's interest [58]. Because effort is at cost to the contractor, the contractor may not give the effort that the owner desires [33]. This may lead to conflict between the contractor and the owner [14, 18, 39], and it may affect the success of the project work [37, 46, 53].

One way of addressing this is for the owner to provide an incentive to the contractor based on the contractor's actual performance, as measured by project outcome expressed relative to a target [33, 45, 67, 75]. Such incentive approaches are exampled by a cost reimbursable contract, with an outcome sharing arrangement or formula, based on a target [22]. The contract aligns the contractor's interests with those of the owner, but at the price of transferring risk to the contractor. Eisenhardt [33] points out that the outcome uncertainty introduces risk, which must be borne by someone. Eisenhardt [33] argues that, as uncertainty increases, it becomes increasingly expensive to shift risk to the contractor. The trade off between incentive and risk in determining the sharing arrangement is central to the design of a contract with an outcome sharing arrangement [73].

In the construction and related project industries literature, although the notion of outcome sharing is well established, research to date has largely been qualitative, and there is no universal agreement on how any sharing of project outcomes should be allocated between contracting parties [11, 72]. This chapter demonstrates that by formulating the sharing problem in optimization terms, specific quantitative results can be obtained for all the various combinations of the main variables that exist in the contractual arrangements and project delivery. Such variables include the risk attitudes of the parties (risk-neutral, risk-averse), single or multiple outcomes (cost, duration, quality), single or multiple agents (contractors, consultants), and cooperative or non-cooperative behaviour. The chapter gives original results for optimal outcome sharing arrangements. The chapter will be of interest to academics and practitioners interested in the design of target contracts and project delivery. It provides an understanding of optimal sharing arrangements within projects, broader than currently available.

The optimization results presented here may be used by contracting parties in the design of their contracts, or as optimal benchmarks by which contracts designed differently may be compared.

The order of presentation in this chapter is as follows. Firstly, a literature review is given. The sharing problem is then established in optimization terms. This is followed by a discussion on the method of solution to the optimization problem for different cases including risk-neutral/averse parties, single/multiple agents and cooperative/non-cooperative behaviour. Finally, the optimization results are provided for all cases.

Reference in the following is to construction projects for definiteness, but the development applies equally to related project types.

## *1.1 Notation, Symbols, Terminology*

The principal notation, symbols and terminology used in this chapter are:

| | |
|---|---|
| Ac | Actual (final) cost of the work (excluding the contractor's fee) |
| At | Actual time |
| b | Constant coefficient ($/effort$^2$) |
| C | Agent's cost of its effort ($) |
| CE | Certainty equivalent |
| e | Agent's effort; action (such as hours worked, or generally paying attention to the owner's interests) [52]. |
| E[] | Expected value |
| F | Fixed component of the agent's fee |
| Fee | Agent's fee |
| **I** | An identity matrix |
| k | Constant; coefficient ($/effort) |
| m | An estimate of the costs to the owner for late completion (similar to the value of liquidated damages) |
| MinFee | Minimum fee |
| n | Sharing ratio—the proportion going to the contractor; takes values in the range 0 to 1 |
| **P** | A covariance matrix |
| **q** | An unitary column vector |
| r | Level of risk aversion ($\$^{-1}$) |
| RP | Risk premium ($) |
| Tc | Target cost estimate of work (excluding the agent's fee) |
| Tt | Target time (schedule) estimate |
| U | Utility ($) |
| x | Outcome ($) |
| $\varepsilon$ | That beyond the agent's influence (noise) |
| $\lambda$ | Lagrange multiplier |
| $\rho$ | Correlation coefficient between the actual cost and actual time uncertainty, taking a value between $-1$ and 1 |
| $\sigma^2$ | Outcome variance |

Outcome—This refers to something that the owner cares about (related to cost, time and quality), measured in monetary units and expressed relative to a benchmark or target that is desired by the owner. An outcome might, for example, be expressed with respect to: cost underruns/overruns relative to a target cost; late completion cost or early completion saving relative to a target duration; monetary value of quality of work done compared with a target level of quality.

Risk—Exposure to the chance of occurrence of events adversely or favourably affecting the project as a consequence of uncertainty [4, 24].

Risk-averse—The term applies to individuals who avoid risk, are afraid of risk, or are sensitive to risk [27].

Risk-neutral—The term refers to those who do not care about risk, and can disregard risk associated with different alternatives in decision making [27, 49].

## 2 Literature Review

The intent of using an outcome sharing arrangement in target contracts is to bring the contracting parties into closer co-operation. The sharing of savings and losses provides a strong motivational factor for the contracting parties to work together, rather than in a confrontational or adversarial fashion, desirably leading to a successful project [22, 23, 38, 51, 70]. A growing dissatisfaction among owners with payment types such as lump sum has created a call for more outcome sharing contracts in construction projects [54].

However, concerns remain about selecting the proper sharing arrangement in target contracts. Because project work can involve large risk [51], choosing an appropriate sharing arrangement becomes essential in achieving project goals [34, 51, 60]. If the sharing arrangement is judged inappropriate, then performance may be reduced [20]. Disagreements, claims, and disputes eventually distort relationships among the parties, and these can be influenced by inappropriate sharing arrangements [61].

A number of studies have looked at the outcome sharing arrangement in construction and related project industries. The following provides a review of these studies.

Barnes [13] outlines the basic principles which should govern sharing arrangements in construction contracts. For example, he suggests that owners should be allocated all outcomes that are predominantly outside the contractor's influence. Barnes [13] argues that the owner and the contractor achieve substantial benefits if outcome is properly shared between them. Abu-Hijlen and Ibbs [3] argue that the contractor's share of any underrun should be equal to or greater than the contractor's share of any overrun. Ward et al. [73] point out that project risk management might be improved if project outcomes are appropriately shared between the contracting parties. They argue that the willingness of contracting parties to take on risk is an important consideration in the sharing arrangement.

Al-Subhi Al-Harbi [5] identifies a lack of literature giving guidance on the setting of sharing arrangements in contracts. He uses utility theory to explain how owners and contractors determine a best sharing arrangement. He arbitrarily assigns numerical values to the utility functions of the owner and the contractor, and then calculates their utility levels for different combinations of outcome sharing and hypothetical cost distributions. He argues that the owner's and the contractor's attitudes toward risk and the distribution of the final project cost affect the sharing arrangement. To simplify the negotiation process between the owner and the contractor to reach an agreement about the outcome share to the contractor, he recommends two things. Firstly, the

owner and the contractor should calculate the anticipated standard deviation of the actual project cost. If it is highly probable that the project cost will be higher than initially predicted, the target cost must be raised. Secondly, the contracting parties should discuss their attitudes towards risk to help in understanding each other's position.

Based on a survey conducted on a sample of highway contracts with incentive/disincentive provisions, Arditi and Yasamis [9] suggest that when using multiple incentive schemes, care should be taken not to overemphasize a particular incentive, as this might cause an imbalance in the contractor's priorities and therefore harm the owner's interests.

Perry and Barnes [56] discuss that contractors are motivated to increase the value of the fee and decrease the value of the target. This motivation increases if the proportion of outcome sharing to the contractor is low. They argue that a low proportion of outcome sharing to the contractor decreases the contractor's motivation to put effort into reducing the actual cost. Perry and Barnes [56] suggest avoiding a proportion of outcome sharing to the contractor less than 0.5. They stress some factors that play a significant role in a successful sharing arrangement. These factors are precise and clear definitions of actual cost and fee, realistic tenders backed up by comprehensive estimates, realistic estimates by the owner, and reliable and fair methods of target adjustment.

Broome and Perry [21] describe some factors that influence a sharing arrangement. These factors are the owner's and contractor's goals in the project, constraints (such as time and cost), project risks, ability of the parties to manage project risks, and the relative financial strengths of the parties to the contract. Broome and Perry [21] conclude that there is need for research on the interaction of risk and the selection of a sharing ratio.

McGeorge and Palmer [57] suggest that the sharing be allocated 50 % to the owner and 50 % to the other parties to the project (divided in proportion to each of the other parties' contribution). Ward and Chapman [72] argue that contractors could nominate a sharing value as the part of their tender. Sappington [66] talks of an iterative approach between the owner and the contractor to establishing the contractor's fee.

A model for outcome sharing, provided by Ross [64] and discussed by Sakal [65] and Love et al. [54], suggests that any cost overrun/underrun be shared 50:50, while underrun sharing is adjusted up or down based on performance in non-cost areas (such as schedule and quality). Love et al. [54] argue that a cost sharing ratio equal to 0.5 underpins the equality of the owner and contractor relationship.

Badenfelt [11] identifies limitations in the literature on how to select a sharing ratio in a target cost contract. Based on interviews with eight construction owners and eight contractors, Badenfelt [11] identifies three factors influencing the choice of the proportion of outcome sharing to the contractor, namely perception of fairness, knowledge of target cost contracts and long-term relationships. To develop a proper sharing arrangement, Badenfelt [11] recommends that the contracting parties collect trustworthy data about each other's skills, reputation and target cost.

Despite an extensive body of publications on target contracts, a review of the literature reveals that limited in-depth analysis has been undertaken in order to provide practitioners with a proper outcome sharing arrangement to adopt. Few studies have been conducted examining the influence of factors that affect the choice of the sharing arrangement [11, 46]. Hughes et al. [46] claim that there is a clear need to explore appropriate sharing arrangements. Badenfelt [11] argues that the current practice of choosing the proportion of outcome sharing to the contractor is rather arbitrary, and not based on scientifically sound evidence or mathematical calculation. No literature appears to have focused on the optimal form of sharing arrangements in construction contracts. In such light, this chapter gives optimal sharing arrangements in target contracts.

## 3 Outline of the Optimization Problem

In the simplest optimization problem, a sole contractor (agent) is engaged by the owner (principal), and the owner is concerned with only one project outcome. Extensions to this simplest problem involve multiple outcomes (cost, duration, quality) or/and multiple agents (contractors, consultants) in different delivery configurations.

In all cases, the owner is not able to fully monitor the effort of the contractors/consultants, but the owner is able to measure the outcome of the effort. The owner desires an optimal outcome sharing contract that maximizes the owner's expected utility, while ensuring that the contractors/consultants agree to the contractual arrangement and the contractors/consultants select an effort level acceptable to the owner; this defines the objective function and the constraints to the optimization problems.

Consider, firstly, the simplest optimization problem mentioned above.

### *3.1 Underlying Basis*

The underlying basis for the optimization problems is presented here in terms of: (I) the outcome, (II) the contract, and (III) the parties' utilities.

*I. Outcome.* The outcome, denoted by x and measured in monetary units, is assumed to depend on the contractor's effort, denoted by e, and events which are outside of the contractor's influence, allowed for through a noise term, $\varepsilon$, representing the uncertainty in the model,

$$x = x(e, \varepsilon). \tag{1}$$

Although the contractor's skill influences the outcome, it is assumed here that all suitable contractors have equivalent skills. Suitability could be ensured, for example, through pre-qualification of contractors, or proper and thorough tender evaluation, against both price and non-price criteria. This is generally considered recommended practice in order to ensure minimum good standards amongst all potential contractors.

However, it is acknowledged that prequalification and thorough tender evaluation practices may not be the case in some countries and on some projects.

*II. The contract.* The contractor's fee, denoted Fee, is taken as being dependent on the outcome,

$$\text{Fee} = \text{Fee}\,(x(e, \varepsilon)) \tag{2}$$

*III. Utility.* The owner receives the outcome, x, but has to pay the contractor's fee. Therefore the owner's utility (or payoff), in monetary units and denoted $U_o$, is a function of the difference between the outcome received and the fee paid,

$$U_o = U_o\,[x(e, \varepsilon) - \text{Fee}(x)] \tag{3a}$$

The contractor receives a fee, but in doing so expends effort, e, at cost in order to produce any outcome. Let C(e) be the dollar amount necessary to pay the contractor for inducing a particular effort level, e. Let the effort $e = 0$ be the effort the contractor would select without any incentive; that is $C(0) = 0$. The contractor's utility (or payoff), in monetary units and denoted $U_c$, is a function of the difference between the fee received and the cost of the effort. The contractor's utility is assumed to be separable into the utility of the fee received, and the cost of the effort [39],

$$U_c\,[\text{Fee} - C] = U_c\,[\text{Fee}(x)] - C(e) \tag{3b}$$

Equation (3b) implies that more effort increases the contractor's utility, but at cost to the contractor.

Following Holmstrom and Milgrom [40, 41], Feltham and Xie [36], and Banker and Thevaranjan [12], the contractor's cost function C(e) is assumed to increase with e at an increasing rate. The simplest functional form that meets this requirement can be written as:

$$C(e) = \frac{b}{2}e^2 \tag{4}$$

Here b is a constant coefficient reflecting the influence of contractor effort on cost; it converts units of effort$^2$ to monetary units.

## *3.2 Optimization Components*

The owner wishes to design an optimal outcome sharing contract, that is one that maximizes its expected utility,

$$\underset{\text{Fee}}{\text{Max}}\,\{E\,[U_o\,[x(e, \varepsilon) - \text{Fee}(x)]]\} \tag{5}$$

subject to constraints. E[ ] denotes expected value.

One constraint occurs because the contractor only agrees to the contractual arrangement if its expected utility exceeds a minimum utility (MinFee).

$$E\left[U_c\left[Fee(x)\right]\right] - C\left(e\right) \geq MinFee \tag{6}$$

This minimum utility might be interpreted as the utility of the contractor in its next best work opportunity, and it reflects the bargaining power of the contractor [52].

A second constraint occurs because the contractor selects the effort level that maximizes its expected utility, and so in order to motivate the contractor to choose an effort level that is in the owner's interests, the contract needs to maximize the contractor's expected utility.

$$\underset{e}{Max}\left\{E\left[U_c\left[Fee(x)\right]\right] - C\left(e\right)\right\} \tag{7}$$

Expression (7) represents the link between the fee offered by the owner and the effort selected by the contractor.

The first constraint (6) is called the individual rationality (IR) constraint [49]. The second constraint (7) is called the incentive compatibility (IC) constraint; this constraint reflects the restriction that the owner can observe the contractor's outcome but not its effort [49].

The owner's problem can thus be expressed as a constrained maximization problem in which the owner selects the contractor fee to:

Maximize the owner's expected utility (Expression 5)
subject to:
The contractor's individual rationality constraint (Expression 6)
The contractor's incentive compatibility constraint (Expression 7)

## 4 Extended and Specialized Optimization Problems

The formulation given in the previous section may be extended and specialized for differing project situations.

### *4.1 Cooperative Contracting: Owner and Contractor*

Firstly, consider the situation where the contractor selects its effort level cooperatively without need for any incentive [52]. This removes the need for the incentive compatibility constraint, and gives what might be called a first-best solution to the optimization problem [52]. The optimization problem reduces to expressions (5) and (6). The contractor's fee is selected to maximize the owner's expected utility subject

to providing the contractor with its minimum expected utility (MinFee) in order to motivate the contractor to accept the contract.

At the optimum, it can be demonstrated that the owner does not need to pay the contractor more than the minimum fee that it needs to agree to the contractual arrangement. (The owner's utility monotonically increases with Fee.) Thus expression (6) should hold as an equality.

Introducing a Lagrange multiplier, $\lambda$, the optimization problem can be expressed as,

$$\underset{\text{Fee},\lambda}{\text{Max}} \left\{ E\left[U_o\left[x(e,\varepsilon) - \text{Fee}(x)\right]\right] + \lambda\left(E\left[U_c\left[\text{Fee}\right]\right] - C(e) - \text{MinFee}\right)\right\} \quad (8)$$

Expression (8) interprets the optimization problem as maximizing a weighted combination of the expected utilities of the owner and contractor.

Such contractor behaviour may occur in cooperative contracting, as exampled by alliances, where the contractor is assumed to behave as the owner would like; the contractor cooperatively puts in effort in the owner's interests. There are several reasons that may motivate the contractor to act this way: the prospect of future business with the owner [21, 26, 62]; moral sensitivity may prevent the contractor from providing less effort than a previously agreed level of effort [2, 58, 69]; the existence of trust between the contracting parties perhaps due to previous experiences or an existing long-term relationship [2, 10, 11, 28, 30, 35, 47, 74]; and maintaining a professional reputation [26, 67].

Appendix A gives the solution to expression (8).

## 4.2 *Cooperative Contracting: Owner, Contractor and Design Consultant*

The two-party (owner and contractor) cooperative case can be extended to include three parties—owner, contractor and design consultant—and to embrace different delivery methods.

Consider a contractor and a design consultant engaged as agents to undertake construction and design, respectively, for an owner. The contractor and design consultant here are indexed with subscripts i = 1, 2, respectively. The outcome of the collective design and construction work, measured in monetary units and denoted by x, is assumed to depend on the contractor's effort, denoted by $e_1$, the design consultant's effort, denoted by $e_2$, and events which are outside the contractor's and consultant's control, allowed for through a noise term, $\varepsilon$, representing uncertainty.

Let the contractor's fee, denoted by $\text{Fee}_1$, and the design consultant's fee, denoted by $\text{Fee}_2$, be functions of the outcome,

$$\text{Fee}_i = \text{Fee}_i\left[x(e_1, e_2, \varepsilon)\right] \qquad i = 1, 2 \quad (9)$$

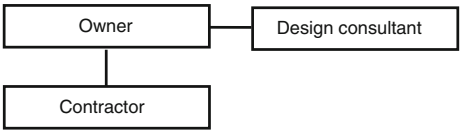



**Fig. 1** Traditional delivery method showing contractual links [22]

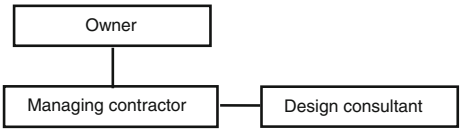



**Fig. 2** Managing contractor delivery method showing contractual links [22]

The following addresses the optimum form of Eq. (9) for two different delivery methods, representing two different owner-contractor-design consultant relationships. Common appellations for these delivery methods are 'traditional delivery', and 'managing contractor delivery' [22]. Though there is not universal agreement on this terminology, it is adopted here; as well, the word 'method' is dropped in the following usage to avoid repetition, but it is implied whenever these delivery methods are referred to. The contractual links in traditional delivery and managing contractor delivery are shown respectively in Figs. 1 and 2; other project participants, such as subcontractors and suppliers are not shown. In traditional delivery, the owner contracts separately with the design consultant and the contractor. A form of this delivery, preferred by some owners, is where the contractor is involved early in the project in order that the contractor may have input to the design; this might be referred to by some as early contractor involvement delivery. In managing contractor delivery, the contractor is engaged by the owner to manage the design consultant and to construct the work.

A constrained maximization problem is solved for each delivery method, to give the optimum outcome sharing; the owner's expected utility (expected saving) is maximized through selection of the fees of the contractor and design consultant acting as agents, while ensuring that the agents agree to the contractual arrangements (interpreted as constraints). There is no incentive compatibility constraint present because in cooperative contracting the efforts are assumed to be selected cooperatively by the agents, rather than selfishly.

### 4.2.1 Traditional Delivery Method

The owner wishes to maximize its expected utility, given by,

$$\underset{\text{Fee}_1,\text{Fee}_2}{\text{Max}} \left\{ \text{E}\left[ \text{U}_\text{o}\left[ \text{x} - \sum_{\text{j}=1}^{2} \text{Fee}_\text{j} \right] \right] \right\} \tag{10}$$

subject to individual rationality constraints,

$$E\left[U_i\left[Fee_i\right]\right] - C_i(e_i) \geq MinFee_i \qquad i = 1, 2 \tag{11}$$

At the optimal, expression (11) should hold as an equality. Using Lagrange multipliers, $\lambda_i$, i = 1, 2, the optimization problem may be expressed as,

$$\underset{Fee_1, Fee_2, \lambda_1, \lambda_2}{Max} \left\{ E\left[ U_O\left[ x - \sum_{j=1}^{2} Fee_j \right] \right] + \sum_{i=1}^{2} \lambda_i \left( E\left[U_i\left[Fee_i\right]\right] - C_i\left(e_i\right) - MinFee_i \right) \right\} \tag{12}$$

Expression (12) gives the optimization problem as maximizing a weighted combination of the expected utilities of the owner and the agents.

Appendix B gives the solution to expression (12).

### 4.2.2 Managing Contractor Delivery Method

In managing contractor delivery, the owner's utility (or payoff) is a function of the difference between the outcome received and the fee paid to the contractor,

$$U_o = U_o\left[x - Fee_1\right] \tag{13}$$

The contractor's utility (or payoff) is a function of the difference between the fee received, and the cost of the contractor's effort and the fee paid to the design consultant,

$$U_1\left[Fee_1 - Fee_2 - C_1\right] = U_1\left[Fee_1 - Fee_2\right] - C_1(e_1) \tag{14a}$$

The design consultant also receives its fee, but at the cost of its effort. The consultant's utility (or payoff) is given by,

$$U_2\left[Fee_2 - C_2\right] = U_2\left[Fee_2\right] - C_2(e_2) \tag{14b}$$

The owner's optimization problem, in designing the contractor's contract, is one that maximizes the owner's expected utility,

$$\underset{Fee_1}{Max}\left\{E\left[U_o\left[x - Fee_1\right]\right]\right\} \tag{15}$$

subject to the individual rationality constraint,

$$E\left[U_1\left[Fee_1 - Fee_2\right]\right] - C_1(e_1) \geq MinFee_1 \tag{16}$$

The contractor, in designing the consultant's contract, needs to provide the consultant with its minimum expected utility, $MinFee_2$, in order to motivate the design consultant to accept the contract,

$$E\left[U_2\left[Fee_2\right]\right] - C_2(e_2) \geq MinFee_2 \tag{17}$$

At the optimal, expressions (16) and (17) hold as equalities. Using Lagrange multipliers, $\lambda_1$ and $\lambda_2$, the optimization problem can be expressed as,

$$\begin{aligned}\underset{Fee_1,Fee_2}{Max}\ & E\left[U_o\left[x - Fee_1\right]\right] \\ & + \lambda_1\left(E\left[U_1\left[Fee_1 - Fee_2\right]\right] - C_1(e_1) - Min\,Fee_1\right) \\ & + \lambda_2\left(E\left[U_2\left[Fee_2\right]\right] - C_2(e_2) - Min\,Fee_2\right)\end{aligned} \tag{18}$$

This owner's problem may be interpreted as maximizing a weighted combination of the expected utilities of the owner, contractor and design consultant.

Appendix C gives the solution to expression (18).

### *4.3 Non-cooperative Contracting*

For the non-cooperative contracting cases, the incentive compatibility constraint is reinstated to the optimization problem. Its purpose is to motivate the contractor to behave in the owner's interest. However, by incorporating this constraint the optimization problem loses tractability. Tractability can be restored by using an alternative formulation, referred to as Linear contacting—Normal distribution (LN). The LN approach uses the simplifying assumptions that the fee is a linear function of outcome and a project's equivalent monetary outcome is normally distributed. Consider each assumption in turn.

*Fee as a linear function of outcome.* This assumption implies,

$$Fee = F + nx \tag{19}$$

Here F is a fixed component of the fee-tendered or agreed—and n is a sharing ratio distributing the outcome between the owner and the contractor, and is defined as the proportion going to the contractor; it takes values in the range 0 to 1.

*Examples.* In contracts based on a target cost, the outcome may be interpreted as a cost underrun/overrun. Using Eq. (19), the contractor's fee is calculated according to

$$Fee = F + n(Tc - Ac) \tag{20}$$

where Tc is a target cost estimate of the work (excluding contractor's fee); and Ac is the actual (final) cost of the work (excluding contractor's fee). Such arrangements are variously known as cost plus incentive fee contracts [70], target cost contracts [22, 56] or financial incentives [11], and belong to the category of payment types called prime cost or cost reimbursable. Equation (20) incorporates the special cases of a variable fee only ($F = 0$), a fixed fee only ($n = 0$), the non-target cost case ($Tc = 0$), and any combination of these, as well as in-between cases. The contractor's fee goes up or down depending on the actual cost; for cost overruns, the fee is adjusted down (disincentive); for actual cost less than estimated, the fee is adjusted up (incentive). Upper and lower limits or caps can be additionally placed on the fee.

Based on a target time (duration), the outcome may be interpreted as the monetary value of time underrun/overrun. Using Eq. (19) the contractor's fee is calculated according to,

$$\text{Fee} = \text{F} + \text{n(Tt-At)m} \tag{21}$$

where Tt is a target estimate of completion time (work duration); At is actual completion time (duration); and m converts time units to cost units, for example late completion in contracts is reflected in the amount inserted in the contract for liquidated damages. Upper and lower limits can be additionally placed on the fee.

As an implementation issue, care needs to be exercised in the choice of a target cost and target duration. Too high a target cost estimate may be easy for the contractor to achieve; too low an estimate may be hard to achieve. A similar comment applies to duration estimates [11, 22, 23]. Love et al. [54] suggest that the targets need to be established by engaging all contracting parties. This results in the non-owner parties having 'ownership' of the project and in so doing provides an incentive to achieve the desired project outcomes. Carmichael [22] suggests that the target cost and target duration estimates can be agreed by the parties, or established by a third independent party. Hughes et al. [46] stress the need for an open and transparent relationship, necessary to avoid target costs being set too high. Bower et al. [19] point out that the target should be the best estimate mutually agreed by the contracting parties.

Although a nonlinear payment may lead to a better outcome for the owner, the linear class of payment is studied for the following reasons. Firstly, simulation-based research has shown that the difference in results arising from the use of a linear payment assumption is relatively small [15, 63]. Secondly, it is straightforward to implement managerially. Finally, the choice of a linear payment assumption is consistent with that of Holmstrom and Milgrom [40]; they show that linear payments may indeed be the optimal form where agents continuously influence effort and observe outcome.

*A project's equivalent monetary outcome is normally distributed.* A project's final cost is comprised of the sum of many component costs. The central limit theorem then gives that the distribution of the final cost will approach a normal distribution independently of the distributions describing the component costs [7, 16, 25]. Barnes [13] argues for the use of a normal distribution for construction activities and estimates.

### 4.3.1 Single-Contractor, Single-Outcome

This section first introduces the case where there is one contractor engaged by the owner, and the owner is concerned with only one project outcome. This is later extended to more complicated cases including multiple agents, and then multiple outcomes. The section uses the LN approach to achieve a tractable solution to the optimization problem. For convenience, it is assumed that the outcome varies linearly with effort, e, giving,

$$\text{x} = \text{ke} + \varepsilon \tag{22}$$

where the constant coefficient k converts units of effort to monetary units, and represents the effectiveness of the contractor's effort towards the outcome. $\varepsilon$ is assumed to be normally distributed with a mean of zero and variance $\sigma^2$ in accordance with the LN approach [40]. $\sigma^2$ then is the variance in the outcome. The linearity assumption is not critical; rather it simplifies the mathematical manipulation. Coughlan and Sen [29] demonstrate that this linearity assumption does not involve much loss of generality.

Consider the case where both parties are risk-averse. The owner wishes to maximize its expected utility through choice of the contractor's fee. The contractor also wishes to maximize its expected utility through choice of effort, e. However, it is difficult to find the exact expected values of the owner and contractor utilities, and so the certainty equivalence concept is used as a work-around [27]. There is a certainty equivalent corresponding to any given expected utility. Maximizing expected utility is equivalent to maximizing its certainty equivalent [27].

For the owner, a certainty equivalent is a saving that is equivalent in the owner's mind to a corresponding situation that involves uncertainty, and equals expected saving minus its risk premium, $RP_o$. This is given by,

$$CE_o = E[x\text{-}Fee] - RP_o \tag{23}$$

where the subscript O refers to the owner. For the contractor, a certainty equivalent is a fee that is the same in the contractor's mind to a corresponding situation that involves uncertainty, and equals expected fee minus the cost of its effort, C(e), and minus its risk premium, $RP_c$. This is given by,

$$CE_c = E[Fee] - C(e) - RP_c \tag{24}$$

where the subscript C refers to the contractor. In obtaining the risk premium, a suitable approximation is provided by Pratt (1964) and also discussed in Clemen and Reilly [27] and Kirkwood [48] and for the owner and contractor are, respectively, given by,

$$RP_o = \frac{1}{2}(1-n)^2 r_o \sigma^2 \tag{25}$$

$$RP_c = \frac{1}{2} n^2 r_c \sigma^2 \tag{26}$$

Accordingly, the optimization problem can be expressed as,

$$\underset{n,F}{Max} \left\{ (1-n)ke - F - \frac{1}{2}(1-n)^2 r_o \sigma^2 \right\} \tag{27}$$

subject to

$$F + nke - \frac{b}{2}e^2 - \frac{1}{2} n^2 r_c \sigma^2 \geq MinFee \tag{28}$$

$$\underset{e}{\text{Max}} \left\{ F + nke - \frac{b}{2}e^2 - \frac{1}{2}n^2 r_c \sigma^2 \right\} \tag{29}$$

Appendix D gives the solution to this optimization problem.

### 4.3.2 Consortium of Contractors

The previous section discussed the sharing problem in contracts with a single contractor. This section extends that to a consortium of contractors (multiple agents). The sharing problem with a consortium of contractors exists in two components. The first involves sharing the project outcome between the owner and consortium. The second involves distributing the consortium's share of the project outcome among the contractors within the consortium. For simplicity the following only focuses on outcome sharing with a consortium of two contractors, each is indexed i, but the results are applicable to more than two contractors.

Based on the LN approach, let the outcome and the contractors' fees respectively be,

$$x = \sum_{i=1}^{2} k_i e_i + \varepsilon \tag{30}$$

$$Fee_i = F_i + n_i x \qquad i = 1, 2 \tag{31}$$

The other parameters of the optimization problem remain as defined previously, with the addition of the subscript i referring to contractor i within the consortium. Assume that the owner is risk-neutral. The optimization problem for contracts with a consortium of two risk-neutral contractors becomes,

$$\underset{n_1, n_2, F_1, F_2}{\text{Max}} \left\{ \left(1 - \sum_{i=1}^{2} n_i\right)\left(\sum_{i=1}^{2} k_i e_i\right) - \sum_{i=1}^{2} F_i \right\} \tag{32}$$

subject to

$$F_i + n_i \sum_{j=1}^{2} k_j e_j - \frac{b}{2}e_i^2 \geq MinFee_i \qquad i = 1, 2 \tag{33}$$

$$\underset{e_i}{\text{Max}} \left\{ F_i + n_i \sum_{j=1}^{2} k_j e_j - \frac{b}{2}e_i^2 \right\} \qquad i = 1, 2 \tag{34}$$

For contracts with a consortium of two risk-averse contractors, expressions (33) and (34) become,

$$F_i + n_i \sum_{j=1}^{2} k_j e_j - \frac{b}{2}e_i^2 - \frac{1}{2}n_i^2 r_i \sigma^2 \geq MinFee_i \qquad i = 1, 2 \tag{35}$$

$$\underset{e_i}{\text{Max}} \left\{ F_i + n_i \sum_{j=1}^{2} k_j e_j - \frac{b}{2} e_i^2 - \frac{1}{2} n_i^2 r_i \sigma^2 \right\} \qquad i = 1, 2 \tag{36}$$

The solutions to these optimization problems are given in Appendices E and F.

### 4.3.3 Multiple Outcomes, Single Contractor

Consider now the sharing problem involving multiple outcomes, where the owner is concerned with more than one of cost, time (duration), quality or safety.

*Example.* In contracts with a target cost and a target time, the outcomes may be interpreted, respectively, as cost overrun/underrun and the monetary value of time overrun/underrun. A multiple outcome sharing arrangement based on cost and time performance may have a fee calculated according to,

$$\text{Fee} = F + n_c(Tc - Ac) + n_t(Tt - At)m \tag{37}$$

where $n_c$ is the sharing ratio associated with cost, and $n_t$ is the sharing ratio associated with time. Other symbols are as defined in Eqs. (20) and (21).

Consider, for example, the situation in which the owner cares about $\mu$ different issues (cost, time/duration, quality, …). In order for each to be interpreted as an outcome in the sense defined above, they are first expressed in monetary units. Let the contractor's contribution to each associated outcome i, i = 1, 2, …, $\mu$, be

$$\mathbf{x} = \mathbf{K}\mathbf{e} + \boldsymbol{\varepsilon} \tag{38}$$

where the symbols have the same meanings as previously; $\mathbf{x}$, $\mathbf{e}$ and $\boldsymbol{\varepsilon}$ are column vectors of order $\mu$ and $\mathbf{K}$ is a $\mu \times \mu$ matrix. Assume that the owner is risk-neutral. Consistent with the earlier development, let the fee and contractor's cost function respectively be,

$$\text{Fee} = F + \mathbf{n}^T\mathbf{x} \tag{39}$$

$$C = \frac{1}{2}\mathbf{e}^T\mathbf{B}\mathbf{e} \tag{40}$$

where the superscript T denotes the transpose of the matrix; Fee, F and C are scalars; $\mathbf{n}$ is a column vector of order $\mu$ and $\mathbf{B}$ is a $\mu \times \mu$ matrix. For a risk-averse contractor and a risk-neutral owner, the optimization problem can be expressed as,

$$\underset{\mathbf{n}}{\text{Max}} \left\{ \mathbf{q}^T\mathbf{K}\mathbf{e} - F - \mathbf{n}^T\mathbf{K}\mathbf{e} \right\} \tag{41}$$

subject to,

$$F + \mathbf{n}^T\mathbf{K}\mathbf{e} - \frac{1}{2}\mathbf{e}^T\mathbf{B}\mathbf{e} - \frac{1}{2}r\mathbf{n}^T\mathbf{P}\mathbf{n} \geq \text{MinFee} \tag{42}$$

$$\underset{\mathbf{e}}{\text{Max}} \left\{ F + \mathbf{n}^T\mathbf{K}\mathbf{e} - \frac{1}{2}\mathbf{e}^T\mathbf{B}\mathbf{e} - \frac{1}{2}r\mathbf{n}^T\mathbf{P}\mathbf{n} \right\} \tag{43}$$

where **P** is a covariance matrix and **q** is an unitary column vector.

Appendix G gives the solution to this optimization problem. The results show the influence on sharing arrangements of: different levels of uncertainty; different levels of contractor's effort effectiveness towards the outcomes; different levels of contractor's effort cost for producing the outcomes; and outcome uncertainty correlation.

## 5 Optimization Results

The following provides an outline and discussion on the solutions to the above optimization problems.

### *5.1 Cooperative Owner-Contractor*

Appendix A demonstrates that the optimal outcome sharing arrangement for the cooperative owner-contractor case determines the contractor's fee according to,

$$\text{Fee} = F + nx \tag{44}$$

where F is a constant (fixed) and,

$$n = \frac{r_o}{r_o + r_c} \tag{45}$$

here $r_o$ and $r_c$ are the owner's level and the contractor's level of risk aversion respectively; and x is the monetary value of the project outcome.

Equation (45) gives a constant value for the parameter n. This is because the owner and contractor levels of risk aversion are fixed. Thus, according to Eqs. (44) and (45), the optimal sharing arrangement, in cooperative contracting, is linear in the outcome. This result answers a common question over whether the sharing should be linear or nonlinear with respect to outcome [31, 32]. ANAO [6] and Hauck et al. [38] report successes using linear relationships.

For the extension covering multiple outcomes i, i = 1, 2, …, such as cost, time, quality or safety, it can be shown that the form of the optimal sharing arrangement is,

$$\text{Fee}_i = F_i + \frac{r_o}{r_o + r_c} x_i \qquad i = 1, 2, \ldots \tag{46}$$

and,

$$\text{Fee} = \text{F} + \text{n} \sum_{\text{i}=1}^{2} \text{x}_\text{i} \tag{47}$$

Equations (46) and (47) show that the optimal sharing arrangement with multiple outcomes is linear in the outcomes. The conclusions are similar to that obtained for the single outcome sharing case.

This theoretical development shows that the optimal outcome sharing is affected by owner and contractor risk attitudes. For a method of measuring levels of risk aversion see Clemen and Reilly [27]. As the contractor becomes more risk-averse, it prefers to receive/bear a lower share of the outcome underrun/overrun. Conversely, as the owner becomes more risk-averse, it prefers the contractor to receive/bear a higher share of the outcome underrun/overrun. This implies that in cooperative contracting, the optimal outcome sharing allocation to a contractor decreases when the contractor's level of risk aversion increases, and increases when the owner's level of risk aversion increases.

The theoretical allocation (Eq. 45) also indicates that when the contractor becomes risk-neutral ($r_c$ becomes very small), the contractor prefers to carry all risk connected with any underrun/overrun. Similarly, when the owner becomes considerably risk-averse ($r_o$ becomes very large) the owner prefers to translate all risk associated with any underrun/overrun to the contractor. Accordingly, a cooperative contracting approach with a risk-neutral contractor or with a very risk-averse owner needs to translate all risk associated with any outcome to the contractor.

In contrast, when the owner becomes risk-neutral ($r_o$ becomes very small), the owner prefers to carry all risk associated with any underrun/overrun. Similarly, when the contractor becomes very risk-averse ($r_c$ becomes very large), the contractor prefers to avoid all risk associated with any underrun/overrun. This implies that in cooperative contracting with a risk-neutral owner or a very risk-averse contractor, the sharing arrangement needs to translate all risk associated with any underrun/overrun to the owner.

The theoretical result for the cooperative owner-contractor case gives that the level of uncertainty in the outcome of the work has no influence on the optimal sharing of outcome. However, Eisenhardt [33] argues that outcome uncertainty has an influence on translating risk to the agent. He believes that for work with a low level of uncertainty in the outcome, the outcome-based contract is more attractive, compared to work with a high level of uncertainty in the outcome, where a behaviour-based contract is considered more attractive. The reason for these differing viewpoints between the present results and that of Eisenhardt is that the argument of Eisenhardt is based on the existence of a conflict of interest between the owner and the contractor. By contrast, the theoretical result given here is based on cooperative behaviour where the parties work as an integrated team to meet project goals [54, 55, 64]. Cooperation is based on principles of faith and trust, as well as an open-book accounting on costs [1]. The parties cooperatively work to achieve agreed outcomes regardless of the level of uncertainty in the outcomes.

## 5.2 Cooperative Owner-Contractor-Design Consultant

### 5.2.1 Traditional Delivery

Appendix B, for cooperative traditional delivery, shows that the optimal outcome sharing is obtained by setting the contractor's and design consultant's fees according to,

$$\text{Fee}_i = F_i + n_i x \qquad i = 1, 2 \tag{48}$$

where $F_i$ is a constant and

$$n_1 = \frac{1}{1 + r_1/r_o + r_1/r_2} \tag{49}$$

$$n_2 = \frac{1}{1 + r_2/r_o + r_2/r_1} \tag{50}$$

here $r_o$, $r_1$ and $r_2$ are the levels of risk aversion of the owner, the contractor and the design consultant, respectively.

### 5.2.2 Managing Contractor Delivery

Appendix C, for cooperative managing contractor delivery, shows that the optimal sharing is obtained by setting the contractor's and design consultant's fees according to Eq. (48), where

$$n_1 = \frac{1}{1 + r_1 r_2/(r_1 + r_2)\, r_o} \tag{51}$$

$$n_2 = \frac{1}{1 + r_2/r_o + r_2/r_1} \tag{52}$$

Equations (49) – (52) give constant values for the parameters $n_1$ and $n_2$ for fixed levels of risk aversion of the owner, contractor and design consultant. Thus, according to Eq. (48), the optimal sharing arrangement in both traditional and managing contractor delivery methods is linear in the outcome. The conclusion is similar to that obtained earlier for the single contractor (and no design consultant) case.

Table 1 summarizes the results of Eqs. (49) – (52).

Equations (49) and (51) demonstrate that a contractor with a low level of risk aversion prefers to receive a high share of outcome in both delivery methods. Similar relationships exist between the level of risk aversion of the design consultant and its share of outcome (Eqs. 50 and 52). Accordingly, in cooperative contracting with traditional or managing contractor delivery, the optimal outcome sharing allocation to each agent (contractor or design consultant) increases with decreasing agent level of risk aversion.

**Table 1** Optimal outcome share to contractor and design consultant in traditional delivery and managing contractor delivery

| Delivery method | Outcome share to | |
|---|---|---|
| | Contractor ($n_1$) | Design consultant ($n_2$) |
| Traditional | $\frac{1}{1+r_1/r_o+r_1/r_2}$ | $\frac{1}{1+r_2/r_o+r_2/r_1}$ |
| Managing contractor | $\frac{1}{1+r_1r_2/(r_1+r_2)r_o}$ | |

Equation (50) indicates that, with a high risk-averse owner or a high risk-averse contractor, the design consultant's share of outcome should be greater than that with a low risk-averse owner or low risk-averse contractor. This argument applies to both traditional and managing contractor delivery methods. Accordingly, in both delivery methods, the optimal outcome sharing allocation to a design consultant increases when the owner or contractor level of risk aversion increases.

Based on Eqs. (49), in traditional delivery with a high risk-averse owner or design consultant, the outcome share to the contractor should be greater than that with a lower risk-averse owner or design consultant. This implies that, in cooperative contracting with traditional delivery, the optimal outcome sharing allocation to a contractor increases when the owner or design consultant level of risk aversion increases.

Equation (51) demonstrates that in managing contractor delivery, regardless of the owner's level of risk aversion, the contractor which engages a high risk-averse design consultant prefers to receive a lower share of outcome from the owner's contract. The argument is that in managing contractor delivery, the design consultant is a member of a 'team' (with the contractor taking the lead role) contracting to the owner, and an increase in the design consultant risk aversion leads to an increase in this contracting team's risk aversion. As discussed earlier, a high risk-averse contractor prefers to receive a lower share of the underrun/overrun. Conversely, based on Eq. (51), and regardless of the design consultant's level of risk aversion, the contractor which is engaged by a high risk-averse owner should receive a higher share of outcome from the contract with the owner. Accordingly, in cooperative contracting with managing contractor delivery, the optimal outcome sharing allocation to a contractor decreases when the design consultant's level of risk aversion increases, and increases when the owner's level of risk aversion increases.

Equations (50) and (52) present the same value for the outcome share to the design consultant in both traditional and managing contractor delivery. That is, in cooperative contracting, the optimal outcome sharing allocation to a design consultant is the same in both delivery methods. The argument is that the design consultant is not the lead player in both project delivery methods; therefore there is no motivation for it to receive a higher proportion of outcome in one delivery method compared to the other.

In order to compare the outcome share to the contractor in the traditional and the managing contractor delivery methods, let $G = r_1/r_o + r_1/r_2$ and $H = r_1r_2/(r_1 + r_2)\, r_o$. Then, according to Table 1, the outcome shares to a contractor in traditional and managing contractor delivery are respectively given by,

$$n_{T1} = \frac{1}{1+G} \tag{53}$$

$$n_{MC1} = \frac{1}{1+H} \tag{54}$$

where the subscripts T and MC refer to traditional and managing contractor delivery, respectively.

If it is assumed that $n_{MC1} > n_{T1}$, then $H < G$. Thus,

$$r_1 r_2/(r_1 + r_2) r_o < r_1/r_o + r_1/r_2 \tag{55}$$

Simplifying expression (55) leads to,

$$r_1 r_2 + r_o(r_1 + r_2) > 0 \tag{56}$$

Because the level of risk aversion is a positive number, the above expression is valid. Consequently, the assumption of $n_{MC1} > n_{T1}$ is correct. This implies that a contractor in managing contractor delivery should receive a greater proportion of the outcome share compared to traditional delivery. Accordingly, in cooperative contracting with managing contractor delivery, the optimal outcome sharing allocation to a contractor is greater than that of traditional delivery. The argument is that because the owner in the managing contractor delivery method only enters into one contractual relationship (with the contractor), it prefers to receive a lower proportion of the outcome than traditional delivery where the owner enters into two contractual relationships (with the contractor and with the design consultant); because the design consultant prefers to receive the same amount of outcome in both project delivery methods, the contractor in the managing contractor delivery method should receive a higher proportion of the outcome compared to that of the traditional delivery method. In managing contractor delivery, the contractor is the lead player, and so it accepts a higher proportion of the outcome share in this delivery method compared to that of traditional delivery.

## *5.3 Non-cooperative Parties*

### 5.3.1 Risk-Neutral Contractor

For a risk-neutral contractor and owner, Eq. (D6) in Appendix D shows that for non-cooperative behaviour, the optimal sharing ratio, n = 1. This means that, at the optimum and expressed relative to the target, any favourable or adverse monetary outcome associated with both the contractor's effort and events beyond the contractor's influence should respectively be wholly received by or wholly borne by a risk-neutral contractor. This implies that in non-cooperative contracting, a risk-neutral contractor

wishes to receive all potential monetary underruns (expressed relative to a target) associated with its effort and events beyond its influence, while accepting all potential monetary overruns. Similarly, for the time incentive contract, at the optimum, a risk-neutral contractor should be given all potential savings (expressed relative to a target) to the owner due to time underruns, while bearing all potential cost to the owner related to time overruns. The conclusion is the same as that obtained earlier for the cooperative owner-contractor case.

Appendix D also establishes the optimum sharing in the case where the owner is risk-averse while the contractor remains risk-neutral. There it is shown that the optimal sharing ratio is 1. This conclusion is the same as that for the risk-neutral owner case. Accordingly, a risk-averse or risk-neutral owner wishes a risk-neutral contractor to bear all potential monetary overruns (expressed relative to a target) associated with the contractor's effort and events beyond the contractor's influence, while accepting that the contractor receive all potential monetary underruns.

### 5.3.2 Risk-Neutral Owner

For the defined risk assumptions on the contractor (risk-averse ranging to risk-neutral) and the owner (risk-neutral), Appendix D demonstrates that, for non-cooperative behaviour, the optimum sharing ratio and fixed fee of Eq. (19) are, respectively, obtained by,

$$\mathrm{n} = \frac{1}{1 + \mathrm{r_c}\sigma^2 \mathrm{b}/\mathrm{k}^2} \tag{57}$$

$$\mathrm{F} = \mathrm{MinFee} + \frac{1}{2}\left(\mathrm{r_c}\sigma^2 - \frac{\mathrm{k}^2}{\mathrm{b}}\right) n^2 \tag{58}$$

where $\mathrm{r_c}$ is the level of contractor risk aversion; $\sigma^2$ is variance of the outcome; b is a coefficient reflecting the influence of contractor effort on cost; k is a coefficient reflecting the effectiveness of the contractor's effort; and MinFee is the minimum fee required by the contractor to motivate the contractor to agree to the contractual arrangement.

The cost of the risk borne by the contractor (risk premium) is directly affected by the contractor risk attitude and level of uncertainty. Based on Eqs. (57) and (58), the sharing ratio needs to increase and the fixed fee needs to reduce as a contractor becomes less risk-averse. As the level of risk aversion continues to increase, the risk premium becomes too large, forcing the owner to reduce the sharing ratio and to increase the fixed fee in order to retain the contractor. Accordingly, in contracts (non-cooperation case) with a risk-neutral owner, with increasing level of contractor risk aversion, the proportion of outcome sharing to the contractor needs to reduce, and the fixed fee needs to increase.

The results indicate that the level of uncertainty in the outcome of the work affects the optimal form of the sharing arrangement. As outcome uncertainty increases,

it becomes increasingly expensive to transfer the risk associated with outcome underrun/overrun to the contractor. According to Eq. (57), the proportion of outcome sharing to the contractor needs to reduce for increasing cost uncertainty in order to manage the risk premium. For high cost uncertainty, the owner needs to bear a high proportion of the risk. This risk is not important to an assumed risk-neutral owner. For high cost uncertainty, the fixed fee needs to increase to motivate the contractor to accept the contract, as shown by Eq. (58); the opposite occurs when the cost uncertainty is low. Therefore, in contracts (non-cooperation case) with a risk-neutral owner, with increasing uncertainty in the outcome, the proportion of outcome sharing to the contractor needs to reduce, and the fixed fee needs to increase.

Based on Eqs. (57) and (58), the proportion of outcome sharing to the contractor needs to increase and the fixed fee needs to reduce as the effectiveness, k, of the contractor's effort towards the outcome increases; the opposite occurs when the contractor's effort effectiveness decreases. Accordingly, in contracts (non-cooperation case) with a risk-neutral owner, with increasing effectiveness of the contractor's effort towards the outcome, the proportion of outcome sharing to the contractor needs to increase, and the fixed fee needs to decrease. As an example, projects with early involvement of the contractor may represent projects where the contractor's effort effectiveness is increased.

### 5.3.3 Consortium of Contractors

For risk-neutral assumptions on the contractors within the consortium, Appendix E establishes the optimum parameters of Eq. (31) and these are given by,

$$n_1 = \frac{1}{1 + (k_2/k_1)^2} \tag{59}$$

$$n_2 = \frac{1}{1 + (k_1/k_2)^2} \tag{60}$$

$$F_1 = \text{MinFee}_1 - \frac{k_1^2}{2b} n_1^2 - \frac{k_2^2}{b} n_1 n_2 \tag{61}$$

$$F_2 = \text{MinFee}_2 - \frac{k_2^2}{2b} n_2^2 - \frac{k_1^2}{b} n_1 n_2 \tag{62}$$

where $k_i$ is a constant coefficient representing the effectiveness of the effort of contractor i towards the outcome; and $\text{MinFee}_i$ is the minimum fee required by contractor i to motivate it to agree to the contractual arrangement. The other symbols have the same meanings as previously.

For risk-averse assumptions on the contractors within the consortium, Appendix F establishes the optimum parameters of Eq. (31) and these are given by,

$$n_i = \frac{1}{1 + r_i \sigma^2 b / k_i^2} \qquad i = 1, 2 \tag{63}$$

$$F_1 = \text{MinFee}_1 - \frac{k_1^2}{2b} n_1^2 - \frac{k_2^2}{b} n_1 n_2 + \frac{1}{2} n_1^2 r_1 \sigma^2 \tag{64}$$

$$F_2 = \text{MinFee}_2 - \frac{k_2^2}{2b} n_2^2 - \frac{k_1^2}{b} n_1 n_2 + \frac{1}{2} n_2^2 r_2 \sigma^2 \tag{65}$$

where $r_i$ is the level of risk aversion of contractor i and the other symbols have the same meanings as previously.

***Outcome sharing among contractors***. Equations (59) and (60) demonstrate that, in a consortium of risk-neutral contractors, the outcome should be shared between the contractors based on their levels of effort effectiveness towards the outcome. Accordingly, the proportion of outcome sharing among risk-neutral contractors within a consortium needs to be high for contractors with high effort effectiveness towards the project outcome.

For a consortium of risk-averse contractors, Eq. (63) demonstrates that in the case where contractors have the same level of risk aversion, namely $r_1 = r_2$, but different levels of effort effectiveness towards the project outcome, that contractor with a higher level of effort effectiveness should receive/bear a higher proportion of monetary underrun/overrun associated with the project outcome. This conclusion is similar to that obtained for a consortium of risk-neutral contractors. Accordingly, the proportion of outcome sharing among risk-averse contractors (in a consortium), with the same level of risk aversion, needs to be high for contractors with high effort effectiveness towards the project outcome.

Equation (63) also shows that for the case where contractors have the same level of effort effectiveness towards the outcome, namely $k_1 = k_2$, but different levels of risk aversion, that contractor with a higher level of risk aversion should receive/bear a lower proportion of monetary underrun/overrun associated with the project outcome. This implies that the proportion of outcome sharing among contractors (in a consortium), with the same level of effort effectiveness towards the project outcome, needs to be lower for contractors with higher levels of risk aversion.

***Outcome sharing between owner and consortium.*** The consortium's share of the outcome is the sum of the contractors' outcome shares. Based on Eq. (63), where the contractors' levels of risk aversion increase or the level of uncertainty in the project outcome increases, the consortium's share of the outcome needs to decrease in order to contain the cost of the risk borne by the contractors (risk premium). Hence, the proportion of outcome sharing to a consortium of risk-averse contractors needs to reduce with increasing uncertainty level in the outcome, or with increasing risk aversion of the contractors.

Based on Eqs. (64) and (65), to motivate risk-averse contractors to accept the contract, the consortium's fixed fee needs to increase as the outcome uncertainty or risk aversion levels of the contractors increase. This implies that with a consortium of

risk-averse contractors, the consortium's fixed fee needs to increase with increasing outcome uncertainty, or with increasing contractor risk aversion.

Equations (59) and (60) together demonstrate that, at the optimum, a consortium of risk-neutral contractors should receive/bear all potential monetary underrun/overrun associated with the project outcome.

### 5.3.4 Multiple Outcomes

For defined risk assumptions on the contractor (risk-averse ranging to risk-neutral) and the owner (risk-neutral), Appendix G derives the optimum parameters of Eq. (39) and these are given by,

$$\mathbf{n} = \left(\mathbf{I} + r\left(\mathbf{K}\mathbf{B}^{-1}\mathbf{K}^{T}\right)^{-1}\mathbf{P}\right)^{-1}\mathbf{q} \tag{66}$$

$$F = \text{MinFee} - \mathbf{n}^{T}\mathbf{K}\mathbf{B}^{-1}\mathbf{K}^{T}\mathbf{n} + \frac{1}{2}\mathbf{n}^{T}\mathbf{K}\left(\mathbf{B}^{-1}\right)^{T}\mathbf{K}^{T}\mathbf{n} + \frac{1}{2}r\mathbf{n}^{T}\mathbf{P}\mathbf{n} \tag{67}$$

where r is the contractor's level of risk aversion; **I** is a $\mu$-dimensional identity matrix; **P** is a covariance matrix; the superscript $-1$ denotes the inverse of the matrix; the superscript T denotes the transpose of the matrix and **q** is an unitary column vector. The other symbols have the same meanings as previously.

Consider the application of this result to a contract with outcome sharing based on cost and time performance such as Eq. (37) (a two-outcome case). Assume that matrices **K** and **B** are diagonal; this can be generalized, allowing non diagonal matrices **K** and **B**. Using Eq. (66), the optimal sharing ratios for Eq. (37) can be obtained. These are,

$$n_c = \frac{1 + rb_t\sigma_t^2/k_t^2 - rb_c\rho\sigma_c\sigma_t/k_c^2}{\left(1 + r\sigma_c^2 b_c/k_c^2\right)\left(1 + r\sigma_t^2 b_t/k_t{}^2\right)(r\rho\sigma_c\sigma_t)^2\, b_c b_t/(k_c k_t)^2} \tag{68}$$

$$n_t = \frac{1 + rb_c\sigma_c^2/k_c^2 - rb_t\rho\sigma_c\sigma_t/k_t^2}{\left(1 + r\sigma_c^2 b_c/k_c^2\right)\left(1 + r\sigma_t^2 b_t/k_t^2\right) - (r\rho\sigma_c\sigma_t)^2\, b_c b_t/(k_c k_t)^2} \tag{69}$$

and using Eq. (67), the optimal fixed fee of Eq. (37) can be obtained from,

$$F = \text{Minfee} + \frac{1}{2}\left(r\sigma_c^2 - \frac{k_c^2}{b_c}\right)n_c^2 + \frac{1}{2}\left(r\sigma_t^2 - \frac{k_t^2}{b_t}\right)n_t^2 + r_c n_c n_t \rho\sigma_c\sigma_t \tag{70}$$

where $\sigma_c^2$ is variance of actual cost; $\sigma_t^2$ is variance of the monetary value of actual time; $k_c$ is a coefficient converting units of effort to monetary units, showing the contractor's effort effectiveness towards the actual cost; $k_t$ is a coefficient converting units of effort to monetary units, showing the contractor's effort effectiveness towards the actual time; $b_c$ is a constant coefficient reflecting the influence on the contractor's

cost of the contractor's effort towards any cost underrun; $b_t$ is a constant coefficient reflecting the influence on the contractor's cost of the contractor's effort towards the time underrun, and $\rho$ is the correlation coefficient between the actual cost and time uncertainties.

Equations (68) and (69) demonstrate that in the case where the contractor's effort effectiveness towards the actual cost is the same as that for the actual time, that is $k_c = k_t$, and the contractor's effort towards the cost underrun is the same as that for the time underrun, that is $b_c = b_t$, then the time sharing ratio should be higher than the cost sharing ratio, that is $n_t > n_c$, if the level of uncertainty in the monetary value of actual time is lower than the level of uncertainty in the actual cost, that is $\sigma_t < \sigma_c$. This implies that in a multiple outcome sharing arrangement, where contractor effort effectiveness towards outcomes is the same and the contractor's cost of this effort across outcomes is the same, the sharing ratio is higher for outcomes with lower uncertainty.

Equations (68) and (69) also show that in the case where $b_t = b_c$ and $\sigma_t = \sigma_c$, the time sharing ratio should be higher than that of the cost ratio, that is $n_t > n_c$, if the contractor's effort effectiveness towards the actual time is higher than that for the actual cost, that is $k_t > k_c$. Accordingly, in a multiple outcome sharing arrangement where the levels of uncertainty across outcomes are the same, and the contractor's cost of effort across outcomes is the same, the sharing ratio is higher for outcomes with higher contractor effort effectiveness.

Equations (68) and (69) further show that in the case where $k_t = k_c$ and $\sigma_t = \sigma_c$, the time sharing ratio should be less than that for the cost, that is $n_t < n_c$, if the contractor's cost of effort towards the time underrun is higher than that for the cost underrun, that is $b_t > b_c$. This implies that in a multiple outcome sharing arrangement where contractor effort effectiveness towards outcomes is the same and the levels of uncertainty across outcomes are the same, the sharing ratio is higher for outcomes with lower effort cost.

Based on Eqs. (68) – (70), the correlation between actual cost and time uncertainties affects the optimal sharing arrangement. It can be seen that by increasing the correlation between the actual cost and time uncertainty, that is by increasing $\rho$ from 0 to 1, the optimal sharing ratios need to decrease and the optimal fixed fee needs to increase. Accordingly, in contracts with multiple outcomes, with increasing outcome uncertainty correlation, the sharing ratios need to reduce, and the fixed fee needs to increase.

## 6 Empirical Support for the Theoretical Results

Strongly persuasive empirical support for the theoretical results presented here can be found in Hosseinian and Carmichael [43, 44] and Hosseinian [42].

***Risk-neutral behaviour.*** In construction projects there are generally considered to be five main motives driving contractors: profit; maintenance of owner-relationship; expansion to new fields or in size; resource utilization; and prestige work [8]. With the

last four (non-profit) motives, a risk-neutral attitude might be expected to apply [43]. With respect to the profit motive, Al-Subhi Al-Harbi [5] argues that the contractor is neutral towards losing a small amount of money since loss is part of any business; however, when the amount of the loss becomes high, possibly leading to bankruptcy, the contractor becomes risk-averse. Through measuring a sample of contractors' risk attitudes, Hosseinian and Carmichael [43] found 27 % of the participants to be risk-neutral. A survey of senior management involved in engineering projects by Lyons and Skitmore [56] reports that the majority of respondents consider themselves as risk-neutral. Uher and Toakley [71] report that the majority of the respondents in their survey identified themselves as either averse or neutral to risk.

## 7 Summary and Conclusions

This chapter demonstrated specific quantitative results for optimal sharing arrangements in target contracts. Different cases, based on the main variables that exist in contractual arrangements and project delivery, were discussed. Such variables include the risk attitudes of the parties (risk-neutral, risk-averse), single or multiple outcomes (cost, duration, quality), single or multiple agents (contractors, consultants), and cooperative or non-cooperative behaviour. The chapter opens up to the construction and related project industries a new way of thinking about construction contracts and contracting party interaction. The solutions given follow an ordered argument and are usable by practitioners. The chapter will be of interest to academics and practitioners (owners, contractors and consultants) involved in the design of target contracts. It provides an understanding of target arrangements, broader than existing treatments.

The following are the summary findings.

*(I) Cooperative contracting—owner and contractor*

- The outcome sharing arrangement is linear in the project outcome.
- The proportion of the optimal outcome sharing to the contractor decreases when the contractor's level of risk aversion increases and/or that of the owner decreases.
- Outcome uncertainty has no influence on the share to the contractor.

*(II) Cooperative contracting—owner, contractor and design consultant—traditional and managing contractor delivery*

- In both traditional and managing contractor delivery, the outcome sharing allocation to a design consultant increases by decreasing its level of risk aversion, or by increasing the owner's level or contractor's level of risk aversion.
- In traditional delivery, the outcome sharing allocation to a contractor increases by decreasing its level of risk aversion, or by increasing the owner's level or the design consultant's level of risk aversion.
- In managing contractor delivery, the outcome sharing allocation to a contractor increases by decreasing its level of risk aversion or the design consultant's level of risk aversion, or by increasing the owner's level of risk aversion.

- The outcome sharing allocation to a design consultant is the same in traditional or managing contractor delivery; however, the outcome sharing allocation to a contractor in managing contractor delivery is greater than that of traditional delivery.

*(III) Non-cooperative contracting—risk-neutral contractor*

- The contract should transfer to the contractor all potential monetary underruns/overruns (expressed relative to a target) associated with the contractor effort and events beyond the contractor's influence. That is, a risk-neutral contractor prefers to receive all potential monetary underruns associated with its effort and events beyond its influence, while it has to bear all potential monetary overruns due to events beyond its influence.
- For the time incentive contract, a risk-neutral contractor should be awarded all potential savings (expressed relative to a target) to the owner due to time underruns, while bearing all potential cost to the owner related to time overruns.

*(IV) Non-cooperative contracting—risk-neutral owner*

- The proportion of outcome sharing to the contractor should reduce, and the fixed fee should increase, for increasing contractor level of risk aversion.
- The proportion of outcome sharing to the contractor should reduce, and the fixed fee should increase, for increasing outcome uncertainty and reducing contractor effort effectiveness towards the outcome.

*(V) Non-cooperative contracting—consortium of contractors*

- The proportion of outcome sharing among contractors with the same risk-attitude should be higher for contractors with higher effort effectiveness.
- The proportion of outcome sharing among contractors with the same level of effort effectiveness should be lower for contractors with higher levels of risk aversion.
- A consortium of risk-neutral contractors should wholly receive or wholly bear any favourable or adverse outcome, respectively.
- The proportion of outcome sharing to a consortium of risk-averse contractors should reduce and the fixed component of the consortium fee should increase when the contractors within the consortium become more risk-averse or the level of the outcome uncertainty increases.

*(VI) Non-cooperative contracting—multiple outcomes*

- The ranking of the proportions of outcome sharing to the contractor, from low to high, should be based on the ranking of uncertainty levels across the outcomes, from high to low, when contractor effort effectiveness towards the outcomes are the same and the contractor's costs of effort across the outcomes are the same.
- The ranking of the proportions of outcome sharing to the contractor, from low to high, should be based on the ranking of the contractor's costs of effort, from high to low, for producing the outcomes, when contractor effort effectiveness towards the outcomes is the same and the levels of uncertainty across the outcomes are the same.

- The ranking of the proportions of outcome sharing to the contractor, from low to high, should be based on the ranking of the contractor's effort effectiveness towards outcomes, from low to high, when the levels of uncertainty across the outcomes are the same and the contractor's costs of effort across outcomes are the same.
- By increasing the correlation between outcomes, the fixed component of the contractor's fee should increase and the proportions of outcome sharing to the contractor should decrease.

The results provide guidance to those involved in designing contracts as to what is the best way to reward contractors and design consultants through the terms of a contract. Owners and contractors might have a general idea about risk sharing and contractor motivation; however, their decisions on contract form; at present; are not based on any rigorous model or theory. The findings presented here assist owners and contractors establish an optimal sharing arrangement. Contracting parties, at tender time, might negotiate any sharing, and the results should assist the contracting parties in this negotiation. Knowing the factors that influence the optimal form of outcome sharing facilitates the process of negotiation between the contracting parties to reach an agreement on the form of the optimal target arrangement. Any final sharing arrangement negotiated can be compared with the optimum result presented here. However, any sharing arrangement different to these optimum results may lead to translating unattractive risk to the contracting parties, resulting in a conflict of interest between the parties, and perhaps putting the project success at stake.

Although this chapter largely focuses on the owner-contractor relationship, the results provide guidance in writing target contracts between other contracting parties such as contractor-subcontractor, consultant-subconsultant, and owner-advisor.

## Appendix A. Cooperative Owner-Contractor

*Derivation of the solution to the maximization problem presented in expression* (8)

Consider where both the owner and contractor are risk-averse, though the level of aversion may range from very large to being risk-neutral. Risk aversion is characterized by a concave utility function. Exponential, power and linear-exponential are candidate functions [48]. Here, the exponential utility function, because it has been popularly adopted [40, 48], is used, and for the owner and the contractor, respectively, have the form,

$$U_o(x - Fee) = 1 - \exp\left[-r_o\,(x - Fee)\right] \tag{A1}$$

$$U_c(Fee) = 1 - \exp[-r_c Fee] \tag{A2}$$

where $r_o$ and $r_c$ are the owner's and the contractor's level of risk aversion, respectively. The shapes of the owner and contractor utility functions change with $r_o$ and $r_c$.

Substituting Eqs. (A1) and (A2) into expression (8), differentiating the result with respect to Fee, setting to zero, and simplifying,

$$-r_o \exp[-r_o (x - Fee)] + \lambda r_c \exp[-r_c Fee] = 0 \tag{A3}$$

from which an expression for $\lambda$ can be obtained.

Taking the derivative of Eq. (A3) with respect to x gives,

$$r_o^2 \left[1 - \frac{dFee}{dx}\right] \exp[-r_o (x - Fee)] - \lambda r_c^2 \frac{dFee}{dx} \exp[-r_c Fee] = 0 \tag{A4}$$

Substituting $\lambda$ from Eqs. (A3) into (A4),

$$\frac{dFee}{dx} = \frac{r_o}{r_o + r_c} \tag{A5}$$

Integrating Eq. (A5) with respect to x gives the optimal outcome sharing arrangement of Eqs. (44) and (45).

## Appendix B. Cooperative Owner-Contractor-Design Consultant (Traditional Delivery)

*Derivation of the solution to the maximization problem presented in expression* (12)

Consider a multiple-contract arrangement with traditional delivery, in which all parties are assumed to be risk-averse. Consistent with the above development, the parties' utilities are described by an exponential form and for the owner and agents, respectively, are,

$$U_o(x - \sum_{j=1}^{2} Fee_j) = 1 - \exp\left[-r_o \left(x - \sum_{j=1}^{2} Fee_j\right)\right] \tag{B1}$$

$$U_i(Fee_i) = 1 - \exp[-r_i Fee_i] \qquad i = 1, 2 \tag{B2}$$

where $r_o$, $r_1$ and $r_2$ are the levels of risk aversion of the owner, the contractor and the design consultant, respectively.

Substituting Eq. (B1) and (B2) into expression (12), differentiating the result with respect to $Fee_i$, i = 1, 2, setting to zero, and simplifying,

$$-r_o \exp\left[-r_o \left(x - \sum_{j=1}^{2} Fee_j\right)\right] + \lambda_i r_i \exp[-r_i Fee_i] = 0 \qquad i = 1, 2 \tag{B3}$$

from which expressions for $\lambda_i$, i = 1, 2 are obtained.

Differentiating Eq. (B3) with respect to x and substituting $\lambda_i$, i = 1, 2 from Eq. (B3),

$$r_o\left(1 - \frac{d\sum_{j=1}^{2} Fee_j}{dx}\right) - r_1\left(\frac{dFee_1}{dx}\right) = 0 \tag{B4}$$

$$r_o\left(1 - \frac{d\sum_{j=1}^{2} Fee_j}{dx}\right) - r_2\left(\frac{dFee_2}{dx}\right) = 0 \tag{B5}$$

The first terms in Eqs. (B4) and (B5) are the same. This then requires the second terms in Eqs. (B4) and (B5) to be the same. That is,

$$r_1\left(\frac{dFee_1}{dx}\right) = r_2\left(\frac{dFee_2}{dx}\right) \tag{B6}$$

Substituting $\left(\frac{dFee_2}{dx}\right)$ from Eqs. (B6) into (B4), and integrating with respect to x provides the contractor's optimal fee, given in Eq. (48), i = 1, and Eq. (49). Similarly substituting $\left(\frac{dFee_1}{dx}\right)$ from Eqs. (B6) into (B5), and integrating with respect to x provides the design consultant's optimal fee, given in Eqs. (48), i = 2, and (50).

## Appendix C. Cooperative Owner-Contractor-Design Consultant (Managing Contractor Delivery)

*Derivation of the solution to the maximization problem presented in expression* (18)

Using exponential utility functions, the utilities of the owner and contractor are described respectively by,

$$U_o(x - Fee_1) = 1 - \exp\left[-r_o\,(x - Fee_1)\right] \tag{C1}$$

$$U_1\,(Fee_1 - Fee_2) = 1 - \exp\left[-r_1\,(Fee_1 - Fee_2)\right] \tag{C2}$$

The design consultant's utility is obtained from Eq. (B2), with i = 2.

Substituting Eqs. (C2) and (B2), i = 2, into expression (18), differentiating the result with respect to $Fee_2$, setting to zero, and simplifying,

$$-\lambda_1 r_1 \exp\left[-r_1\,(Fee_1 - Fee_2)\right] + \lambda_2 r_2 \exp\left[-r_2 Fee_2\right] = 0 \tag{C3}$$

from which an expression for $\lambda_2$ can be obtained.

Taking the derivative of Eq. (C3) with respect to x gives,

$$\lambda_1 r_1^2 \left( \frac{dFee_1}{dx} - \frac{dFee_2}{dx} \right) \exp\left[-r_1 \left(Fee_1 - Fee_2\right)\right] - \lambda_2 r_2^2 \frac{dFee_2}{dx} \exp\left[-r_2 Fee_2\right] = 0 \tag{C4}$$

Substituting $\lambda_2$ from Eqs. (C3) into (C4) gives,

$$\frac{dFee_2}{dx} = \left( \frac{r_1}{r_1 + r_2} \right) \frac{dFee_1}{dx} \tag{C5}$$

Substituting Eqs. (C1), (C2) and (B2), i = 2, into expression (18), and differentiating the result with respect to $Fee_1$, setting to zero, and simplifying,

$$-r_o \exp\left[-r_o \left(x - Fee_1\right)\right] + \lambda_1 r_1 \exp\left[-r_1 \left(Fee_1 - Fee_2\right)\right] + \lambda_2 \frac{dFee_2}{dFee_1} r_2 \exp\left[-r_2 Fee_2\right] = 0 \tag{C6}$$

Taking the derivative of Eq. (C6) with respect to x gives,

$$r_o^2 \left(1 - \frac{dFee_1}{dx}\right) \exp\left[-r_o \left(x - Fee_1\right)\right] - \lambda_1 r_1^2 \left( \frac{dFee_1}{dx} - \frac{dFee_2}{dx} \right) \exp\left[-r_1 \left(Fee_1 - Fee_2\right)\right]$$
$$[8pt] - \lambda_2 \frac{dFee_2}{dFee_1} \frac{dFee_2}{dx} r_2^2 \exp\left[-r_2 Fee_2\right] = 0 \tag{C7}$$

Substituting $r_o \exp\left[-r_o \left(x - Fee_1\right)\right]$ from Eqs. (C6) into (C7), using Eq. (C5) and simplifying gives,

$$r_o \left(1 - \frac{dFee_1}{dx}\right) - \left( \frac{r_1}{r_1 + r_2} \right) \frac{dFee_1}{dx} r_2 = 0 \tag{C8}$$

Integrating Eq. (C8) with respect to x provides the contractor's optimal fee, given in Eq. (48), i = 1, and Eq. (51).

Substituting Eq. (48), i = 1, and Eqs. (51) into (C5), and integrating with respect to x, provides the design consultant's optimal fee, given in Eq. (48), i = 2 and Eq. (52).

## Appendix D. Non Cooperative Contracting, Single-Agent, Single-Outcome Case

*Derivation of the solution to the maximization problem presented in expressions* (27), (28) *and* (29)

Maximizing expression (29) with respect to e yields the optimal level of effort,

$$e = \frac{k}{b} n \tag{D1}$$

The optimal value of F is such that expression (28) holds as an equality, that is,

$$F = \text{MinFee} - nke + \frac{b}{2}e^2 + \frac{1}{2}n^2 r_c \sigma^2 \tag{D2}$$

Substituting Eqs. (D1) and (D2) into (27), the owner's problem can be restated as,

$$\underset{n}{\text{Max}} \left\{ \frac{k^2}{b}n - \text{MinFee} - \frac{k^2}{2b}n^2 - \frac{n^2 r \sigma^2}{2} - \frac{1}{2}(1-n)^2 r_o \sigma^2 \right\} \tag{D3}$$

Differentiating expression (D3) with respect to n and setting to zero, leads to the optimal sharing ratio,

$$n = \frac{1}{1 + r_c/(r_o + k^2/\sigma^2 b)} \tag{D4}$$

Substituting Eq. (D1) into (D2) leads to the optimal fixed fee,

$$\text{Fee} = \text{MinFee} + \frac{1}{2}\left(r_c \sigma^2 - \frac{k^2}{b}\right) n^2 \tag{D5}$$

***Special cases***

*Contracts with a risk-neutral contractor*

For the case where the contractor is risk-neutral, while the owner is either risk-neutral or risk-averse, the optimal sharing ratio and fixed fee are obtained by setting $r_c = 0$,

$$n = 1 \tag{D6}$$

$$\text{Fee} = \text{MinFee} - \frac{k^2}{2b} \tag{D7}$$

*Contracts with a risk-neutral owner*

For the case where the owner is risk-neutral while the contactor is risk-averse (ranging to risk-neutral) the optimal sharing ratio is obtained by setting $r_o = 0$,

$$n = \frac{1}{1 + r_c \sigma^2 b/k^2} \tag{D8}$$

And the optimal fixed fee is obtained from Eq. (D5).

## Appendix E. Contracts with a Consortium of Risk-Neutral Contractors

*Derivation of the solution to the maximization problem presented in expressions* (32), (33) *and* (34)

Differentiating expression (34) with respect to $e_i$ and setting to zero provides the optimal level of effort,

$$e_i = \frac{k_i}{b} n_i \qquad i = 1, 2 \tag{E1}$$

The optimal value of $F_i$ would be such that expression (33) holds as an equality, that is,

$$F_i = \text{MinFee}_i - n_i \sum_{j=1}^{2} k_j e_j + \frac{b}{2} e_i^2 \qquad i = 1, 2 \tag{E2}$$

Substituting Eqs. (E1) and (E2) into (32), the owner's problem can be restated as,

$$\underset{n_1, n_2}{\text{Max}} \left\{ E\left[U_O\right] = \underset{n_1, n_2}{\text{Max}} \left( \sum_{i=1}^{2} \frac{k_i^2}{b} n_i \right) - \sum_{i=1}^{2} \left( \text{MinFee}_i + \frac{k_i^2}{2b} n_i^2 \right) \right\} \tag{E3}$$

The sum of the contractors' sharing ratios, namely $n_1 + n_2$, is equal to the outcome sharing ratio to the consortium, that is the proportion the consortium receives in the consortium-owner relationship. The consortium proportion of outcome share takes values in the range 0 to 1. Thus the solution of the above maximization needs to satisfy,

$$\sum_{i=1}^{2} n_i \leq 1 \tag{E4}$$

Differentiating expression (E3) with respect to $n_i$, $i = 1, 2$, and setting to zero,

$$n_i = 1 \qquad i = 1, 2 \tag{E5}$$

This result does not satisfy Eq. (E4). Accordingly, the maximum of expression (E3) lies on the line $n_1 + n_2 = 1$ which is the boundary of the admissible region of the maximization problem.

Introducing a Lagrange multiplier $\lambda$, the maximization becomes,

$$\underset{n_1, n_2}{\text{Max}} \left\{ \left( \sum_{i=1}^{2} \frac{k_i^2}{b} n_i \right) - \sum_{i=1}^{2} \left( \text{MinFee}_i + \frac{k_i^2}{2b} n_i^2 \right) + \lambda \left( \sum_{i=1}^{2} n_i - 1 \right) \right\} \tag{E6}$$

Differentiating expression (E6) with respect to $n_i$, $i = 1, 2$, and $\lambda$, setting to zero, and simplifying leads to the optimal sharing ratios of Eqs. (59) and (60).

Substituting Eq. (E1) into (E2), leads to the optimal fixed fees of Eqs. (61) and (62).

## Appendix F. Contracts with a Consortium of Risk-Averse Contractors

*Derivation of the solution to the maximization problem presented in expressions* (32), (35) *and* (36)

Differentiating Eq. (36) with respect to $e_i$ and setting to zero provides the optimal level of effort, and this leads to Eq. (E1).

The optimal value of $F_i$ would be such that expression (35) holds as an equality, that is,

$$F_i = \text{MinFee}_i - n_i \sum_{j=1}^{2} k_j e_j + \frac{b}{2} e_i^2 + \frac{1}{2} n_i^2 r_i \sigma^2 \quad i = 1, 2 \tag{F1}$$

Substituting Eqs. (E1) and (F1) into expression (32), the owner's problem can be restated as,

$$\underset{n_1, n_2}{\text{Max}} \left\{ \left( \sum_{i=1}^{2} \frac{k_i^2}{b} n_i \right) - \sum_{i=1}^{2} \left( \text{MinFee}_i + \frac{k_i^2}{2b} n_i^2 + \frac{1}{2} n_i^2 r_i \sigma^2 \right) \right\} \tag{F2}$$

Differentiating expression (F2) with respect to $n_i$, $i = 1, 2$, and setting to zero, leads to the optimal sharing ratio of Eq. (63).

Substituting Eq. (E1) into (F1), leads to the optimal fixed components of Eqs. (64) and (65).

Where the contractors' levels of risk aversion approach zero and the contractors become risk-neutral, the solutions of expression (F2) lie on the line $n_1 + n_2 = 1$, and the optimal sharing ratios are obtained by Eqs. (59) and (60).

## Appendix G. Contracts with Multiple Outcomes

*Derivation of the solution to the maximization problem presented in expressions* (41), (42) *and* (43)

Maximizing expression (43) with respect to **e** yields the optimal effort levels,

$$\mathbf{e} = \mathbf{B}^{-1} \mathbf{K}^{\mathrm{T}} \mathbf{n} \tag{G1}$$

where the superscript $-1$ denotes the inverse of the matrix.

The optimal value of F is such that expression (42) holds as an equality, that is,

$$F = \text{MinFee} - \mathbf{n}^{\mathrm{T}} \mathbf{K} \mathbf{e} + \frac{1}{2} \mathbf{e}^{\mathrm{T}} \mathbf{B} \mathbf{e} + \frac{1}{2} r \mathbf{n}^{\mathrm{T}} \mathbf{P} \mathbf{n} \tag{G2}$$

Substituting Eqs. (G1) and (G2) into expression (41), the owner's problem can be restated as,

$$\underset{\mathbf{n}}{\text{Max}}\, \mathbf{q}^{\mathrm{T}} \mathbf{K} \mathbf{B}^{-1} \mathbf{K}^{\mathrm{T}} \mathbf{n} - \text{MinFee} - \frac{1}{2} \mathbf{n}^{\mathrm{T}} \mathbf{K} \mathbf{B}^{-1} \mathbf{K}^{\mathrm{T}} \mathbf{n} - \frac{1}{2} r \mathbf{n}^{\mathrm{T}} \mathbf{P} \mathbf{n} \tag{G3}$$

Differentiating expression (G3) with respect to **n** and setting to zero leads,

$$\mathbf{K}\mathbf{B}^{-1}\mathbf{K}^{\mathrm{T}}\mathbf{q} - \mathbf{K}\mathbf{B}^{-1}\mathbf{K}^{\mathrm{T}}\mathbf{n} - \mathrm{r}\mathbf{P}\mathbf{n} = 0 \tag{G4}$$

Multiplying by $(\mathbf{K}\mathbf{B}^{-1}\mathbf{K}^{\mathrm{T}})^{-1}$ leads to the optimal **n** of Eq. (66). Substituting Eq. (G1) into (G2) leads to the optimal fixed fee of Eq. (67).

## References


1. Abrahams A, Cullen C (1998) Project alliances in the construction industry. Aust Constr Law Newslett 62:31–36
2. Abudayyeh O (1994) Partnering: a team building approach to quality construction management. J Manag Eng ASCE 10(6):26–29
3. Abu Hijleh SF, Ibbs CW (1989) Schedule-based construction incentives. J Constr Eng Manag 115(3):430–443
4. Al-Bahar JF, Crandall KC (1990) Systematic risk management approach for construction projects. J Constr Eng Manag 116(3):533–546
5. Al-Subhi Al-Harbi KM (1998) Sharing fractions in cost-plus-incentive-fee contracts. Int J Proj Manag 16(2):73–80
6. ANAO (2000) Construction of the national museum of Australia and the Australian Institute of Aboriginal and Torres Strait Islander Studies. Audit Rep, Canberra, Australia, Australian National Audit Office
7. Ang AHS, Tang W (1975) Probability concepts in engineering planning and design, vol I. Wiley, New York
8. Antill JM (1970) Civil engineering management. Angus and Robertson, Sydney
9. Arditi D, Yasamis F (1998) Incentive/disincentive contracts: perceptions of owners and contractors. J Constr Eng Manag 124(5):361–373
10. Aulakh P, Kotabe M, Sahay A (1996) Trust and performance in cross-border marketing partnerships: a behavioral approach. J Int Bus Stud 27(5):1005–1032
11. Badenfelt U (2008) The selection of sharing ratios in target cost contracts. Eng Constr Architect Manag 15(1):54–65
12. Banker RD, Thevaranjan T (1997) Accounting earnings and effort allocation. Manag Finance 23(5):56–71
13. Barnes M (1983) How to allocate risks in construction contracts. Int J Proj Manag 1(1):24–28
14. Bartling B, Von Siemens FA (2010) The intensity of incentives in firms and markets: moral hazard with envious agents. Labour Econ 17(3):598–607
15. Basu AK, Kalyanaram G (1990) On the relative performance of linear versus nonlinear compensation plans. Int J Res Mark 7:171–179
16. Benjamin JR, Cornell CA (1970) Probability, statistics and decision for civil engineers. McGraw-Hill, New York
17. Berends TC (2000) Cost plus incentive fee contracting—experiences and structuring. Int J Proj Manag 18(3):165–171
18. Bolton P, Dewatripont M (2005) Contract theory. MIT Press, Cambridge, Mass, London
19. Bower D, Ashby G, Gerald K, Smyk W (2002) Incentive mechanisms for project success. J Manag Eng ASCE 18(1):37–43
20. Bresnen M, Marshall N (2000b) Motivation, commitment and the use of incentives in partnerships and alliances. Constr Manag Econ 18(5):587–598
21. Broome J, Perry J (2002) How practitioners set share fractions in target cost contracts. Int J Proj Manag 20(1):59–66

22. Carmichael DG (2000) Contracts and international project management. A. A. Balkema, Rotterdam
23. Carmichael DG (2002) Disputes and international projects. A A Balkema, Swets and Zeitlinger B V, Lisse
24. Carmichael DG (2004) Project management framework. A. A. Balkema, Rotterdam
25. Carmichael DG (2006) Project planning, and control. Taylor and Francis, Oxford
26. Chan APC, Chan DWM, Fan LCN, Lam PTI, Yeung JFY (2008) Achieving partnering success through an incentive agreement: lessons learned from an underground railway extension project in Hong Kong. J Manag Eng ASCE 24(3):128–137
27. Clemen RT, Reilly T (2001) Making hard decisions with decision tools. Duxbury/Thomson Learning, Pacific Grove
28. Cook EL, Hancher DE (1990) Partnering: contracting for the future. J Manag Eng ASCE 6(4):431–446
29. Coughlan AT, Sen SK (1989) Salesforce compensation: theory and managerial implications. Mark Sci 8(4):324–342
30. Das T, Teng BS (2001) Trust, control, and risk in strategic alliances: an integrated framework. Organ Stud 22(2):251–283
31. Department of Infrastructure and Transport (2011) National alliance contracting guidelines, guide to alliance contracting. www.infrastructure.gov.au. Accessed 6 March 2013
32. Department of Treasury and Finance (2006) Project alliancing practitioners' guide. http://www.dtf.vic.gov.au. Accessed 10 Feb 2013
33. Eisenhardt KM (1989) Agency theory: an assessment and review. Acad Manag Rev 14(1):57–74
34. El-Sayegh SM (2008) Risk assessment and allocation in the UAE construction industry. Int J Proj Manag 26(4):431–438
35. Eriksson P, Laan A (2007) Procurement effects on trust and control in client-contractor relationships. Eng Constr Architect Manag 14(4):387–399
36. Feltham GA, Xie J (1994) Performance measure congruity and diversity in multi-task principal/agent relations. Acc Rev 69(3):429–453
37. Harmon KMJ (2003) Conflicts between owner and contractors, proposed intervention process. J Manag Eng 19(3):121–125
38. Hauck AJ, Walker DHT, Hampson KD, Peters RJ (2004) Project alliancing at the national museum of Australia—collaborative process. J Constr Eng Manag 130(1):143–152
39. Holmstrom B (1979) Moral hazard and observability. Bell J Econ 10(1):74–91
40. Holmstrom B, Milgrom P (1987) Aggregation and linearity in the provision of intertemporal incentives. Econometrica 55(2):303–328
41. Holmstrom B, Milgrom P (1991) Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. J Law Econ Organ 7:24–52
42. Hosseinian SM (2013) Optimal outcome sharing arrangements in construction target contracts. Doctoral dissertation. The University of New South Wales, Sydney, Australia
43. Hosseinian SM, Carmichael DG (2013) An optimal incentive contract with a risk-neutral contractor. ASCE J Constr Eng Manag 139(8): 899–909
44. Hosseinian SM, Carmichael DG (2013) Optimal gainshare/painshare in alliance projects. J Oper Res Soc 64(8):1269–1278
45. Huang M, Chen G, Ching WK, Siu TK (2010) Principal-agent theory based risk allocation model for virtual enterprise. J Serv Sci Manag 3:241–250
46. Hughes D, Williams T, Ren Z (2012) Is incentivisation significant in ensuring successful partnered projects? Eng Constr Architect Manag 19(3):306–319
47. Kamann D, Snijdets C, Tazelaar F, Welling D (2006) The ties that bind: buyer-supplier relations in the construction industry. J Purch Supply Manag 12(1):28–38
48. Kirkwood CW (2004) Approximating risk aversion in decision analysis applications. Decis Anal 1(1):51–67
49. Kraus S (1996) An overview of incentive contracting. Artif Intell 83(2):297–346

50. Laffont JJ, Martimort D (2002) The theory of incentives: the principal-agent model. Princeton University Press, Princeton, N.J., Oxford
51. Lahdenpera P (2010) Conceptualizing a two-stage target-cost arrangement for competitive cooperation. Constr Manag Econ 28(7):783–796
52. Lambert R (2001) Contracting theory and accounting. J Acc Econ 32(1):3–87
53. Larson E (1997) Partnering on construction projects: a study of the relationship between partnering activities and project success. IEEE Trans Eng Manag 44(2):188–195
54. Love PED, Davis PR, Chevis R, Edwards DJ (2011) Risk/reward compensation model for civil engineering infrastructure alliance projects. J Constr Eng Manag 137(2):127–136
55. Love PED, Irani Z, Cheng EWL, Li H (2002) A model for supporting inter-organisational relations in the supply chain. Eng Constr Architect Manag 9(1):2–15
56. Lyons T, Skitmore M (2004) Project risk management in Queensland engineering construction industry: A survey. Int J Proj Manag 22(1): 51–61
57. Perry JG, Barnes M (2000) Target cost contracts: an analysis of the interplay between fee, target, share and price. Eng Constr Architect Manag 7(2):202–208
58. McGeorge D, Palmer A (2002) Construction management new directions, 2nd edn. Blackwell Science, Oxford
59. Petersen T (1993) The economics of organization: the principal-agent relationship. Acta Sociologica 36(3):277–293
60. Puddicombe MS (2009) Why contracts: evidence. J Constr Eng Manag 135(8):675–682
61. Rahman MM, Kumaraswamy MM (2002) Risk management trends in the construction industry: moving towards joint risk management. Eng Constr Architect Manag 9(2):131–151
62. Rahman MM, Kumaraswamy MM (2005) Assembling integrated project teams for joint risk management. Constr Manag Econ 23(4):365–375
63. Rahman MM, Kumaraswamy MM (2008) Relational contracting and teambuilding: assessing potential contractual and noncontractual incentives. J Manag Eng ASCE 24(1):48–63
64. Raju JS, Srinivasan V (1996) Quota-based compensation plans for multiterritory heterogeneous salesforces. Manag Sci 42(10):1454–1462
65. Ross J (2003) Introduction to project alliancing. http://www.pci-aus.com/files/resources/Alliancing_30Apr03_F.pdf. Accessed 15 Dec 2011
66. Sakal MW (2005) project alliancing: a relational contracting mechanism for dynamic projects. Lean Constr J 2(1):67–79
67. Sappington DEM (1991) Incentives in principal-agent relationships. J Econ Perspect 5(2):45–66
68. Sharma A (1997) Professional as agent: knowledge asymmetry in agency exchange. Acad Manag Rev 22(3):758–799
69. Shavell S (1979) Risk sharing and incentives in the principal and agent relationship. Bell J Econ 10(1):55–73
70. Stevens DE, Thevaranjan A (2010) A moral solution to the moral hazard problem. Acc Organ Soc 35(1):125–139
71. Turner JR, Simister SJ (2001) Project contract management and a theory of organization. Int J Proj Manag 19(8):457–464
72. Uher ET, Toakley RA (1999) Risk management in the conceptual phase of a project. Int J Proj Manag 17(3):161–169
73. Ward S, Chapman C, Curtis B (1991) On the allocation of risk inconstruction projects. Int J Proj Manag 9(3):140–147
74. Ward S, Chapman C (1994) Choosing contractor payment terms. Int J Proj Manag 12(4):216–221
75. Weitzman ML (1980) Efficient incentive contracts. Q J Econ 44(1):719–730
76. Wong PSP, Cheung SO (2005) Structural equation model of trust and partnering success. J Manag Eng ASCE 21(2):70–80
77. Zhao H (2005) Incentive-based compensation to advertising agencies: a principal-agent approach. Int J Res Mark 22(3):255–275