Optimal control of heavy haul train based on approximate dynamic programming

Xi Wang[1], Tao Tang[2] and Hui He[2]

Abstract
This research investigates the optimal control problem of heavy haul train for the minimization of longitudinal forces. As the heavy haul train is much heavier and longer than ordinary train, the in-train forces should be carefully manipulated to reduce the train's maintenance cost and, most importantly, to ensure operation safety. Specifically, the limitations of pneumatically controlled braking system increase the need for the optimal control strategy to accounting for future grades, speed restrictions and uncertain disturbances. In this article, the stochastic dynamic programming model is adopt to set up a rigorous mathematical formulation for heavy haul train control, and approximate dynamic programming algorithm with lookup table representation is introduced to find the optimal solution of the considered problem. By handling the existed uncertainties in a mathematical way, the post-decision state variable is utilized to represent the state of the heavy haul train after we have made a control decision but before any exogenous information has arrived. Finally, the computational results demonstrate the effectiveness and performance of the proposed model and algorithm.

Keywords
Heavy haul train, optimal control, air braking, approximate dynamic programming

Introduction

Background and literature review

As an efficient and reliable traffic mode, heavy haul railway transportation is currently one of the main methods that provide the services for transporting coals, petroleum, minerals and so on in many countries. In order to increase haulage capacity, trains have increased from a maximum of 10,000 to 20,000 ton on the Datong-Qinhuangdao and Shuohuang heavy haul railways in China.[1] Thus, the train control strategies will need to be developed and refined to heavier and longer trains accordingly.

In general, a desirable efficient operation strategy means the minimization of energy consumption, service quality in terms of punctuality and operation safety closely linked with in-train forces.[2–4] Over the last few decades, a number of optimizing algorithms and advanced control techniques, such as fuzzy control,[5] coasting control[6,7] and model predictive control (MPC),[8–10] have been proposed for the improvement of the three performance indicators. Nevertheless, as pointed out by Zhang and Zhuan,[8] it is impossible to make any improvement on energy performance without making the two other performance indicators worse

[1]National Engineering Research Center of Rail Transportation Operation and Control System, Beijing Jiaotong University, Beijing, China
[2]State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

Corresponding author:
Xi Wang, National Engineering Research Center of Rail Transportation Operation and Control System, Beijing Jiaotong University, Beijing 100044, China.
Email: xiwangbjtu@gmail.com

off, and we should choose the proper strategy regarding the desired performance in practice.

Since China's two main heavy haul dedicated lines are responsible for transporting coal from west to east, the routes are mainly downwards. When the train is generally travelling downwards, both dynamic braking and the air braking are applied to keep the train speed within the required range.[11] As only the locomotives in the heavy haul train can supply dynamic braking forces, the air braking force is on the leading position in regulating train's speed. Applying air braking will inevitably give rise to longitudinal forces within the train and lead to excessive tear-and-wear on braking unit. It is believed that the longitudinal coupler forces are the direct cause of coupler damages, which may lead to train operation disasters such as derailment.[12] Under this circumstance, the demand on a desirable train operation strategy to optimize the operation safety determined by in-train forces has escalated accordingly, combined with maintaining commercial punctuality. In this sense, finding the optimal control method for heavy haul trains to minimize the longitudinal forces is one of the core problems in both fields of theoretic research and practical operation.

Following this trend, substantial work has been done to optimize the operation safety determined by in-train forces,[13] designed an open-loop controller and took advantage of Lagrange multiplier approach to determine the power splitting between the neighbouring locomotives in order to minimize the longitudinal coupler forces. The authors of Nasr and Mohammadi[14] investigated the effects of the train brake delay time on longitudinal dynamic behaviour of the freight trains using simulation method and observed that the magnitude of the maximum tensile force relatively increased as brake delay time decreased. Zhang and Zhuan[12] introduced an MPC approach to schedule the heavy haul train to operate optimally during a long section, and the penalty for coupler damping was recommended to alleviate the cyclic vibration of couplers.

To the best of our knowledge although the majority of the literature had paid more attention to the optimal operations of heavy haul train under various scenarios, for example, cruising control,[15] station stopping,[16] delay recovery,[17] considering the real-world operation environments such as input constraints,[18] communication delays,[19] the uncertain of some performance parameters[20] and so on, the optimal air braking control strategy when travelling on steep descent is not as thoroughly studied as others, but its importance for transportation safety and efficiency should not be underestimated.

Unfortunately, the pneumatic braking system of heavy haul train has a number of limitations such as brake delays throughout the train and minimum brake application duration. Some pneumatic braking systems even do not have the graduated release function. These disadvantages of air braking impose stringent requirements on safety, reliability and service quality, thereby reducing the flexibility to optimize both longitudinal in-train forces and travelling time, simultaneously, increasing the complexity of the corresponding optimization models.[4] It is worth noting that due to the uncertain environment of real-world operations, such as wind gust, weather and gradient resistance, the detailed characteristics of train dynamic model cannot be captured accurately.[3,21] Recently, modelling the uncertainty has become a hot research issue in controller design, and adaptive robust control method is usually employed to achieve high-accuracy trajectory tracking performance.[22,23]

As far as I know, most solution methodologies in the literature, such as quadratic programming, linear quadratic regulator and MPC, have limitations in handling these real-life operational requirements under uncertainty. For this reason, some powerful algorithmic strategies are required to solve the constrained optimization problem in the presence of various forms of randomness.

Proposed approach

As an emerging technology for multistage stochastic, dynamic problems that arise in operations research, approximate dynamic programming (ADP) offers an extremely flexible modelling framework which makes it possible to combine the strengths of simulation with the intelligence of optimization.[24,25] To handle the existed uncertainties in a mathematical way, the post-decision state variable is introduced to represent the state of the system after we have made a decision but before any exogenous information has arrived. Exogenous information referring to the sources of uncertain factors can be viewed as information that becomes available over time under practical circumstance. This means that the value function is a deterministic function of the state and action, a feature that makes it suitable for a wide range of problems spanning real-time planning of locomotives,[26] dynamic routing and scheduling,[27,28] health resource deployment[29] and large-scale fleet management problems.[30]

With the consideration of practical constraints associated with the optimal operation of heavy haul train on steep descent, we are particularly interested in finding the corresponding approximate optimal control strategies to both enhance safety and guarantee service quality. This research aims to provide the following contributions to the optimal heavy haul train controller design:

1. In this study, the optimal control problem with respect to the cascade mass point model of heavy

haul train is transformed into a unified stochastic dynamic programming model. In the formulation, we can capture the impact of certain control action on the future, and then communicate this impact backward. By this way, decisions can be made more intelligently. Furthermore, operational regulations associated with the air braking are detailed and formulated as constraints on the possible solutions to ensure safety and punctuality.

2. A standard ADP algorithm with lookup table representation is formulated for finding the optimal operation strategy with the least in-train forces. Some critical algorithmic issues, such as such as post-decision state variable, exploration and exploitation, stepsize rules, are discussed to effectively solve the problem. By adopting different sets of parameters and simulation scenarios, the computational results show that the proposed approacher can efficiently generate optimal solutions for the considered problems within acceptable computational time.

The rest of this article is organized as follows. Section 'Problem description' gives a description for the non-linear dynamic model of the heavy haul train and then details the problem of running in a steep descent. In the next section, mathematical formulations of the optimal control problem are described. Section 'ADP approach' describes the algorithmic strategy, focusing primarily on the use of ADP to solve the problem of optimizing over time. Simulation results are discussed in section 'Simulation results and discussions' for demonstrating the validity and effectiveness of the proposed approaches. Some conclusions and further works are given in the final section.

Problem description

The dynamic model of heavy haul train

Essentially, heavy haul trains are distributed powered networked system constituted with many locomotives and wagons. Figure 1 is the sketch of the longitudinal motion. Assuming the train consists of n cars and q locomotives which locate at positions $j_1, j_2 \ldots, j_q$.

According to Newton's law, the longitudinal dynamics for each car can be established by the following equation

$$m_i \dot{v}_i = u_i + f_{i-1} - f_i - F_i^R, \quad i = 1, 2, \ldots, n \quad (1)$$

where m_i is the mass of the ith car, u_i is the traction or braking force added to the ith car, f_i denotes the in-train force between the ith and $(i + 1)$th car and F_i^R is the general running resistance.

In general, F_i^R consists of aerodynamic drag and rolling resistance M_i^R, track slope resistance L_i^R and curvature resistance C_i^R

$$F_i^R = M_i^R + L_i^R + C_i^R \quad (2)$$

The running resistance M_i^R depends on the physical properties of the high-speed train and its current speed combines both rolling resistance and air resistance. The former linearly increases as a function of the adhesion and the wheel rims. The latter quadratically increases as a function of the train velocity. According to the Davis[31] equation, constants c_0, c_1 and c_2 are introduced to approximate the running resistance as follows

$$M_i^R = m_i \left(c_0 + c_1 v_i + c_2 v_i^2 \right) \quad (3)$$

Line resistance L_i^R depends on the train mass and the slope angle θ

$$L_i^R = m_i g \sin \theta \quad (4)$$

where g is the gravity constant $g = 9.81$ N/kg and θ is measured in meters per thousand.

Concerning curve resistance, the value of C_i^R is approximated[32]

$$C_i^R = m_i g \frac{700}{R} \quad (5)$$

where R is the track curve radius.

As mentioned in Chou and Xia,[15] here we also assume that the coupler system is taken as a spring with damping. Thus, the in-train force can be established as follows
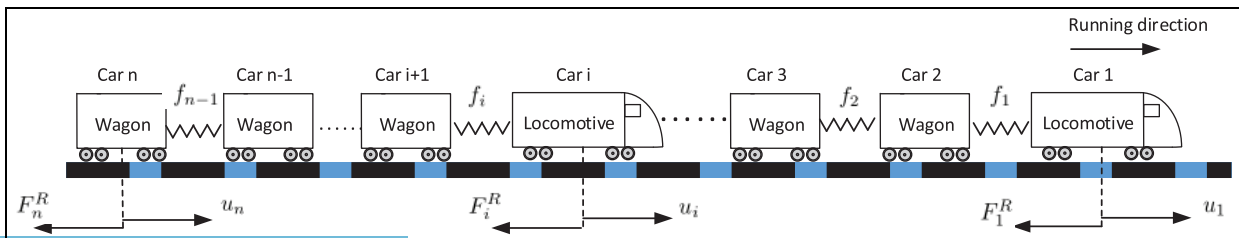

Figure 1. Longitudinal dynamics of heavy haul train.

$$f_i = k_i l_i + d_i(v_i - v_{i+1}), \quad i = 1, 2, \ldots, n-1 \quad (6)$$

where $k_i > 0$ is the stiffness coefficient and d_i is damping constants, which are determined by the characteristics of couplers in the heavy haul train. As l_i represents the absolute extension or compression length of the ith coupler corresponding to the original length without any elastic deformation, it could be positive or negative.

For a heavy haul train, the control inputs for locomotives can be either traction forces or braking forces, while the efforts of wagons are only braking forces. To

coasting, both traction force and braking force are switched off, which present opportunities for energy saving. Braking is to slow the train or to bring it to a stop. However, on track with steep downward gradients, the power-hold-coast-brake strategy may not be feasible and it will be necessary to replace the hold phase with one or more coast phases on the steep downhill sections.[7]

Definition 1. If the train speed increases on a grade when the maximum rheostatic brake is applied, then we say this grade is called a steep downhill. In this grade, we have

$$\begin{cases} m_i \dot{v}_i(t) = -b^r_{max} + f_{i-1} - f_i - F^R_i + w_i > 0, \quad i = j_1, j_2, \ldots, j_q, \\ m_i \dot{v}_i(t) = f_{i-1} - f_i - F^R_i + w_i > 0, \quad i = 1, 2, \ldots, n, \ i \neq j_1, j_2, \ldots, j_q \end{cases} \quad (8)$$

address the braking control issue, there are basically two types of braking units equipped in heavy haul train including rheostatic unit and pneumatic unit. The rheostatic brake is also called regenerative brake, which can be fed back to power other locomotives. For rheostatic braking system, the series excitation resistor can be adjusted to control the braking current so that continuous braking forces can be produced to slow down or stop the train. In pneumatic braking system, braking forces are applied by reducing the air pressure in the train air braking pipe.[14]

Additionally, the running train will inevitably suffer from the uncertain disturbance from real-world environments such as wind gust and weather condition, which may affect the transient longitudinal forces as well as service quality.[33] Thus, the parameter w can be expressed as the uncertain variable to characterize uncertain information. As air brakes are added as braking forces to all cars and rheostatic brake is equipped only to locomotives, the dynamic equation of n-cars heavy haul train (1) is equivalent to

Remark 1. Obviously, on a steep downhill, it may need particle air braking to prevent train from over-speeding. Considering the braking pipe needs to restore pressure completely to achieve effective release, the speed profile has a definite cyclic nature.[34] Thus, constraints should be used to guarantee the safe operations of heavy haul train. First, air-filled time in periodic train braking should be ensured. Second, as altering braking rates frequently gives rise to longitudinal forces within the train and leads to excessive tear-and-wear on braking unit, each braking notch should maintain certain time before transferred to another one. In addition, the switching of notches has to satisfy several guidelines. For example, coasting must be applied as an intermediate step if a driver wants to switch between motoring and braking. One switching should not jump too many notches.

In practice, there are more than one possible set of control instructions, which enable an inter-station run under the same runtime and speed requirements, but the resulting longitudinal forces may vary significantly.

$$\begin{cases} m_i \dot{v}_i(t) = u^a_i - b^r_i - b^p_i + f_{i-1} - f_i - F^R_i + w, \quad i = j_1, j_2, \ldots, j_q, \\ m_i \dot{v}_i(t) = -b^p_i + f_{i-1} - f_i - F^R_i + w, \quad i = 1, 2, \ldots, n, \ i \neq j_1, j_2, \ldots, j_q \end{cases} \quad (7)$$

where u^a_i is the traction efforts, b^r_i is the rheostatic brake forces and b^p_i is referred to air braking forces.

Control problem on steep decent

Generally, locomotive operation involves four possible operation modes: accelerating, cruising, coasting and braking. In accelerating phrase, traction effort is applied to accelerate the train and overcome the running resistance. Under most conditions, running resistance is positive so that partial traction effort is applied to maintain a constant speed in cruising state. During

A critical issue in above problem is the brake notches decision policy, that is, how to choose a brake notch, and when to perform or release the corresponding air brake. If these policies are not designed carefully, the longitudinal forces within the train may increase rapidly and the speed may exceed the limit, both of which will significantly affect the safety and efficiency of heavy haul train movement. Therefore, we focus on developing optimal control algorithms to achieve timely train speed adjustment with minimized in-train forces, which will be studied in the following section.

Mathematical formulations

This section will formulate the optimal control problem on steep downhill as a stochastic dynamic programming model with the minimized in-train forces criteria. In order to satisfy the requirements for punctuality, we should make sure that the train can cover the distance within the expected travel time T. Assuming the time interval between each continuous state is Δt, the control problem is divided into N stages, where N is determined by $T/\Delta t$. At each stage, we make a decision on a proper braking notch to select. The following discussion focuses on detailing each part of the model, such as states, actions, exogenous information, cost function, state transition function and objective function.

States

We regard each car in the heavy haul train as an intelligent agent. A policy defines the agent's way of behaving at each decision epoch. We measure the state S_t just before we make a decision. These decision epochs are modelled in discrete time, but the physical process occurs in continuous time. As shown in the following, the current position of the first car, relative coupler displacements and speeds are considered as state variables of the train

$$S_t = (t, p_t, L_t, V_t) \quad t = 1, 2, \ldots, N \quad (9)$$

where p_t is the position of the first car, the vector $L_t = [l_t^1, l_t^2, \ldots, l_t^{n-1}]^T$ gives the relative coupler displacements for all cars and $V_t \in \mathbb{R}^n$ is the speed vector of cars in the heavy haul train.

Remark 2. For simplicity, only the position of the first car is considered in the state variable. At equilibrium state, the distances between the $(i + 1)$th car and the ith car $L_{i-1,i}$ are some known values determined according to the length of cars and the natural length of coupler. As a result, the exact position of each car in heavy haul train can be obtained from $L_{i-1,i}$ and l_i.

Actions

In the train control process, actions may affect not only the immediate reward, but also the rewards of the following states. When heavy haul train is travelling on steep downward slope, the maximum rheostatic brake is used on each locomotive and air brake is applied on every car. Without the loss of generality, we suppose $\mathcal{B}$ be the space of all possible settings of air braking force. Thus, the following equations are obtained.

$$\begin{cases} u_i = b_{max}^r + b_i^p, & i = j_1, j_2, \ldots, j_q \\ u_i = b_i^p, & i = 1, \ldots, n, \; i \neq j_1, j_2, \ldots, j_q \end{cases} \quad (10)$$

where b_i^p is the air brake force for the ith car and $b_i^p \in \mathcal{B} \cup 0$.

As the maximum rheostatic brake can be treated as a constant for certain type of locomotive, the decision variables are defined as follows

$$\begin{aligned} x_t &= [x_t^1, \ldots, x_t^i, \ldots, x_t^n]^T \\ &= [b_1^p, \ldots, b_i^p, \ldots, b_n^p]^T \end{aligned}$$

According to Rao,[35] calculation equation of the braking force can be written as follows

$$b_i^p = \varphi_h \cdot \vartheta_h \cdot \beta_c \cdot 10^3 \quad (11)$$

where φ_h is the equivalent friction coefficient, ϑ_h denotes the equivalent emergency braking ratio and β_c represents the service braking coefficient.

In general, φ_h and ϑ_h have close relationship with the physical characteristics of locomotives and wagons, and β_c is selected according to the constant of train pipe pressure and the specific train pipe pressure reduction when the train implements air braking. Thus, the braking force is determined by the pressure reduction. According to the characteristics of pneumatic braking system, the possible pressure reduction amount is usually from 50 to 100 kPa with an interval of 10 kPa. In practical operation, the selected pressure reduction is not more than 100 kPa considering safety factors.

Heavy haul trains have a number of limitations that need to be carefully manipulated when determining the optimal action. In order to guarantee the requirements associated with safe operation and maintenance cost, constraints on decision variables x_t^i for each car should be contemplated considering both the operational factors and the capacities of air braking.

Braking force constraints. It is worth noting that although electronically controlled pneumatic (ECP) braking system has been developed to provide each car with different air braking efforts, this technology is not implemented in practical heavy haul train lines in China on a visible scale due to high operating cost. As a result, it is reasonable to assume that the braking force is identical for cars controlled by the same locomotive. The corresponding constraint is below

$$x_t^g = x_t^h, \; \forall j_k \leq g, \; h < j_{k+1}, \; k = 1, 2, \ldots, q - 1 \quad (12)$$

Air-filled time constraints. Considering the features of air braking, we should reserve enough time to fill the air tanks so that we can count on the air brakes. The following air-filled time constraints are formulated to capture this characteristic

$$\begin{cases} h - g \geq \sigma, & \forall 0 \leq g < r < h \leq N, \\ x_g^i \neq 0, \ x_r^i = 0, \ x_h^i \neq 0, & i = 1, \ldots, n \end{cases} \quad (13)$$

where σ denotes the minimum air-filled time.

Minimum hold on time constraints. In general, altering braking rates frequently not only gives rise to longitudinal forces within the train, but also leads to excessive tear-and-wear on braking unit. Therefore, switching forth and back between adjacent notches must be avoided. Once a braking notch is triggered, it should maintain certain time before transferred to another one.

If we use τ to denote the minimum hold on time, we have

$$\begin{cases} h - g \geq \tau, & \forall 1 \leq g < h \leq N, \\ x_{g-1}^i = 0, \ x_g^i \neq 0, \ x_h^i = 0, & i = 1, \ldots, n \end{cases} \quad (14)$$

Graduated release constraints. At present, most heavy haul train pneumatic braking systems do not equip the graduated releasing function in China. As a result, coasting must be applied as an intermediate step if we plan to switch to a brake notch with smaller braking force from a bigger braking force notch. Then, the corresponding constraints can be formulated as follows

$$x_h^i = 0, \ x_g^i \neq 0, \ i = 1, \ldots, n \quad (15)$$

where $1 \leq g < h \leq N$ and $h = \min\{h|x_g^i \neq x_h^i\}$.

Let $\mathcal{X}_t$ be the set of all x_t that satisfies constraints (10)–(15) at stage t. Furthermore, we define $X_t^\pi(S_t)$ as the decision function that determines decision x_t at stage t under policy π, given state S_t. Each element $\pi \in \Pi$ refers to a different policy and Π denotes the set of all implementable policies.

Exogenous information

In practice, the running heavy haul train will inevitably suffer from the wind gust, weather condition and other real-world environment factors. These unfavourable factors could be treated as uncertain external disturbances to heavy haul train. To handle the existed uncertainties in a mathematical way, we use exogenous information to describe the disturbances that arrive to the train exogenously, representing the sources of randomness. The exogenous information consists of the realization of the disruption statuses of all the cars. As the disruption statuses may change as we proceed to the next stage, the systems exogenous information can be written as

$$W_{t+1} = \widehat{D}_{t+1} \quad (16)$$

where $\widehat{D}_{t+1}$ denotes the disruption status realization that becomes known between stages t and $t+1$.

State transition function

The transition function details the system state transition. Given the current state $S_t = (t, p_t, L_t, V_t)$, if we choose an action $x_t = X_t^\pi(S_t)$ to control the train and then observe the new exogenous information $W_{t+1} = \widehat{D}_{t+1}$. The system transits to a new state S_{t+1} according to the following transition function

$$S_{t+1} = S^M(S_t, x_t, W_{t+1}) \quad (17)$$

As addressed above, we can derive from equations (6) and (8) that

$$\dot{\bar{z}}(t) = \bar{A}\bar{z}(t) + \bar{B}\bar{u}(t) + \Phi + W \quad (18)$$

where

$$\bar{z}(t) = [l_1, l_2, \ldots, l_{n-1}, v_1, v_2, \ldots, v_n]^T$$

$$\bar{u}(t) = [\underbrace{b_{j_1}^p, \ldots, b_{j_1}^p + b_{max}^r, b_{j_1+1}^p, \ldots, b_{j_q}^p + b_{max}^r, b_{j_q+1}^p, \ldots}_{n}]^T$$

$$\bar{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad A_{11} = 0_{(n-1)\times(n-1)}$$

$$A_{12} = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots \\ 0 & 1 & -1 & 0 & \cdots \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix}_{(n-1)\times n}$$

$$A_{21} = \begin{bmatrix} -\dfrac{k_1}{m_1} & 0 & 0 & \cdots & 0 \\ \dfrac{k_1}{m_2} & -\dfrac{k_2}{m_2} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \dfrac{k_{n-2}}{m_{n-1}} & -\dfrac{k_{n-1}}{m_{n-1}} \\ 0 & \cdots & 0 & 0 & \dfrac{k_{n-1}}{m_n} \end{bmatrix}_{n\times(n-1)}$$

$$A_{22} = \begin{bmatrix} -c_1 - c_2 v_1 - \dfrac{d_1}{m_1} & \dfrac{d_1}{m_1} & 0 \\ \dfrac{d_1}{m_2} & -c_1 - c_2 v_2 - \dfrac{d_1+d_2}{m_2} & \dfrac{d_2}{m_2} \\ 0 & \cdots & \dfrac{d_2}{m_{n-1}} \\ 0 & \cdots & 0 \\ & \cdots & 0 \\ & \cdots & 0 \\ -c_1 - c_2 v_{n-1} - \dfrac{d_{n-1}+d_{n-2}}{m_{n-1}} & \dfrac{d_{n-1}}{m_{n-1}} \\ \dfrac{d_{n-1}}{m_n} & -c_1 - c_2 v_n - \dfrac{d_{n-1}}{m_n} \end{bmatrix}_{n\times n}$$

$$\bar{B} = [0_{n \times (n-1)}, I_{n \times n}]^T, \quad W = [\underbrace{0, \ldots, 0}_{n-1} \quad \underbrace{w_1, \ldots, w_n}_{n}]^T$$

$$\Phi = [\underbrace{g \sin\theta - c_0, \ldots, g \sin\theta - c_0}_{n}]^T$$

In order to apply the stochastic dynamic programming framework, the continuous time-domain state-space equation (18) is discretized by the zero-order hold method with a sampling period Δt, and the detailed transition function can be described as follows

$$\begin{bmatrix} L_{t+1} \\ V_{t+1} \end{bmatrix} = A \begin{bmatrix} L_t \\ V_t \end{bmatrix} + B\bar{x}_t + \Phi + W \qquad (19)$$

where $A = e^{\bar{A}\Delta t}$, $B = \int_0^{\Delta t} e^{\bar{A}\tau} d\tau \bar{B}$ and $\bar{x}_t = [\underbrace{x_t^1, \ldots, x_t^{j_1} + b_{max}^r, x_t^{j_1+1}, \ldots, x_t^{j_q} + b_{max}^r, x_t^{j_q+1}, \ldots}_{n}]^T$.

Besides, according to Newton's law, p_{t+1} is presented as $p_{t+1} = p_t + v_t^1 \Delta t + acc\Delta t^2/2$, where $acc = (-b_{max}^r - x_t^1 + f_{i-1}(t) - f_i(t) - F_i^R(t) + w_t^1)/m_1$.

Cost function

The actual in-train forces or cost for period t would be given by $C_t(S_t, x_t) = \sum_{i=1}^{n-1} [(k_i l_{t+1}^i + d_i v_{t+1}^i - d_i v_{t+1}^{i+1})^2)]/(n-1)$, where l_{t+1}^i and v_{t+1}^i are calculated by the same logic as in the state transition function (19).

Objective function

The objective of the stochastic dynamic programming is to find the optimal policy $\pi \in \Pi$ to minimize the expected total cost (in-train forces) to ensure the safety of train operation, that is,

$$\min_{\pi \in \Pi} \mathbb{E} \sum_{t=0}^{N} C_t(S_t, x_t) = \min_{\pi \in \Pi} \mathbb{E} \sum_{t=0}^{N} C_t(S_t, X_t^\pi(S_t)) \quad (20)$$

where $x_t = X_t^\pi(S_t)$ is the decision made according to the decision function $X_t^\pi(S_t)$ under policy π, given the current state S_t.

ADP approach

If the state, decision and outcome spaces are finite discrete, the stochastic dynamic programming equation (20) can be solved using the classical backward dynamic programming algorithm by Bellman's equations

$$V_t(S_t) = \min_{\pi \in \mathcal{X}_t} (C_t(S_t, x_t) + \mathbb{E}[V_{t+1}(S_{t+1})]) \qquad (21)$$

where $V_t(S_t)$ is the value function of being in state S_t, in which $C_t(S_t, x_t)$ accounts for the immediate cost associated with the current state S_t and decision x_t, while the value function $V_{t+1}(S_{t+1}) = V_{t+1}(S^M(S_t, x_t, W_{t+1}))$ evaluates the future impact of the decision x_t under the realized exogenous information W_{t+1}.

Nevertheless, solving equation (21) encounters three curses of dimensionality (states, decisions and outcomes), and the dynamic programming approach for solving Bellman's equations becomes computationally intractable. As an alternative, the ADP approach is a powerful tool to overcome the curses of dimensionality, especially for complex and large-scale problems.[25]

In essence, ADP replaces the exact value function $V_t(S_t)$ in Bellman's equation with a statistical approximation $\overline{V}_t(S_t)$. Instead of calculating the exact state values backward in time, ADP steps forward in time by making decisions based on the approximate value function $\overline{V}_t(S_t)$.

In the following sections, we first introduce the post-decision state variable and then describe ADP algorithm with lookup table representation. Furthermore, we discuss some key algorithmic issues that we encountered in the design of ADP algorithm to effectively solve the problem, such as exploration and exploitation strategy, stepsize rules and so on.

Post-decision state variable

Our algorithmic strategy differs markedly from what is presented in merging math programming with the techniques of machine learning, particularly in use of the post-decision state variable.[25] Note that S_t is the state immediately before we make a decision, sometimes denoted as the pre-decision state. That is, before certain air braking command is applied, we can observe the train's state, S_t. The post-decision state S_t^x is the state immediately after action x_t, but before the arrival of new information W_{t+1}. Thus, we can express above transition function with two steps

$$S_t^x = S^{M,x}(S_t, x_t) \qquad (22)$$

$$S_{t+1} = S^{M,W}(S_t^x, W_{t+1}) \qquad (23)$$

In the first step (22), we consider the pure effect of decision-making, while in the second step (23), we pay attention to the effect of exogenous information. Given our action x_t, we have a deterministic transition from S_t to the so-called post-decision state $S^{M,x}(S_t, x_t)$. In this way, the expectation in equation (21) is eliminated and equation (21) is transformed to the following equivalent

$$V_t(S_t) = \min_{\pi \in \mathcal{X}_t} (C_t(S_t, x_t) + \overline{V}_{t+1}(S_{t+1}))$$

ADP algorithm

Rather than solving for the value of each state exactly, ADP steps forwards through time via simulation and proceeds by iteratively estimating and updating the

approximate value of being in a state. Algorithm 1 outlines the steps of ADP algorithm for solving the optimal heavy haul train control problem. This algorithm uses a single pass to simulate a sample trajectory using the current estimates of the value functions, and both the calculation and updates of the value function take place as the algorithm progresses forward time.

Algorithm 1. The ADP algorithm around post-decision states

1: Initialize $\overline{V}_t^0(S_t^x) = 0$, $t = 0, 1, \ldots, N$, and set the iteration counter $n = 1$.
2: Set initial train state $S_0^n = S_0$.
3: **for** $t = 0, 1, \ldots, N$ **do**
4:   **if** *exploitation* **then**
5:     solve $\widehat{v}_t^n = \min_{x_t \in /_t}(C_t(S_t^n, x_t) + \overline{V}_t^{n-1}(S_t^{x,n}))$.
6:     Let x_t^n be the action that solves the minimization problem:

$$x_t^n = \underset{x_t \in \mathcal{B}_t}{\arg\min}\,(C_t(S_t^n, x_t) + \overline{V}_t^{n-1}(S_t^{x,n})) \qquad (24)$$

7:   **end if**
8:   **if** *exploration* **then**
9:     randomly choose a solution x, and compute the observation $\widehat{v}_t^n$.
10:   **end if**
11:   Use the current observation $\widehat{v}_t^n$ to update $\overline{V}_{t-1}^n$ $(S_{t-1}^{x,n})$: $\overline{V}_{t-1}^n(S_{t-1}^{x,n}) = (1 - \alpha_{n-1})\overline{V}_{t-1}^{n-1}(S_{t-1}^{x,n}) + \alpha_{n-1}\widehat{v}_t^n$
12:   Obtain the post-decision state: $\widehat{v}_t^n$ to update $\overline{V}_{t-1}^n(S_{t-1}^{x,n})$:

$$S_t^{x,n} = S^{M,x}(S_t^n, x_t^n).$$

13:   Choose a sample disturbance vector ω^n with Monte Carlo simulation, and find the next pre-decision state

$$S_{t+1}^n = S^{M,W}(S_t^{x,n}, W_{t+1}(\omega^n))$$

14:   **if** train reaches the destination **then**
15:     Go to step 18;
16:   **end if**
17: **end for**
18: $n \leftarrow n + 1$.
19: **if** $n \leq \mathbb{N}$ **then**
20:   Go to step 2;
21: **end if**
22: Return the value function $(\overline{V}_t^{\mathbb{N}}(S_t^x))_{t=1}^N$.

With the approximate value function around post-decision state, we can solve for $\widehat{v}_t^n$ in step 5, which avoids the expectation within the min operator, but normally requires more effort in estimating $\overline{V}_t^{n-1}(S_t^{x,n})$.

Remark 3. The approximate value function $\overline{V}_{t+1}(\cdot)$ can take a variety of forms such as weighted sum of basis functions, piecewise linear functions, regression models, neural networks and the lookup table representation. As a generic model-free form, the lookup table is often used when the value function structure can hardly be clearly defined, which is just the case of heavy haul train control under study.

Exploration and exploitation strategy

If we constantly exploit the action with the minimum value, only the values of states with the minimum cost are updated, and the value of the rest states remains their initial values. This causes the approximate state values not improving as we do not explore other states. On the other hand, if we explore states and actions that may not look attractive, we could reduce the probability of being stuck in suboptimal solutions. In this sense, we should decide on the trade-off between exploration and exploitation when making a decision given a certain state.

One of the simple and intuitive ways of solving this problem is known as the ε-greedy policy, which guarantees that we will visit every (reachable) state infinitely often.[36] Under this policy, with probability ε we choose an action at random from the feasible region $\mathcal{X}$. With probability $1 - ε$, we choose the action according to equation (24), in which case we are exploiting our current knowledge of the value of each state.

Specifically, the ε-greedy policy is a fixed exploration rate strategy. It is worth noting that the value of being in a state is mostly dependent on the sample path and the initial solution in the early iterations. Therefore, it is reasonable to use the exploration strategy more frequently to improve the quality of the state values at the beginning, and the exploration rate decreases as the number of visits to the particular state increase. To do this, we introduce the exploration rate $\eta = \kappa/n'(S_t, x_t)$, where $n'(S_t, x_t)$ is the number of visits to a state-decision pair and κ is a constant. By this way, we can not only select a random probability of choosing the best action and choosing alternative actions, but also ensure that the probability of choosing the best action increases as we visit the state more.

Stepsize rules

After finding the next action with the exploration/exploitation strategy, we then confront the problem of approximating the current value of the visited state using proper stepsize rule, which plays an important role on the convergence performance. If we choose a too small stepsize, the rate of convergence will be slow. If the stepsize is too large, the performance will be unstable. Theoretically, there are two kinds of stepsize

rules including deterministic stepsize rules and stochastic stepsize. Deterministic stepsize do not change with practical data in the process of approximating state values, while stochastic stepsize rules adapt to collected data. For computational convenience, we implement with a deterministic stepsize rule, that is, the harmonic stepsize rule, as in equation (25)

$$\alpha_{n-1} = \frac{a}{a + n'(S_t, x_t) - 1} \quad (25)$$

where a is a constant. Note that the stepsize depends on the number of visits to the state-decision pair $n'(S_t, x_t)$, rather than the iteration counter n.

Simulation results and discussions

In this section, the detailed simulation results are provided to verify the effectiveness of the proposed control schemes. The routine of the optimal controller design can be categorized into two processes including the offline and online parts. For the offline part, the mathematical models described earlier are established according to practical conditions, and the proposed ADP algorithm is used to find out the optimal policy. Once the optimal policy is obtained, it is stored in a table format. Each entry in the table specifies the optimal action given the current train state. For the online part, the train looks up the policy table to find out the optimal action corresponding to its current state when running on railway lines, and then, it executes the action to get the next state. By this way, the optimal output is computed repeatedly until the train reaches the destination.

The simulation parameters given in Table 1 are based on the heavy haul train from Shuohuang heavy haul railways in China, where *SS*4*B* locomotive and C_{80} type wagon are applied in large scale.[35]

Table 1. Simulation parameters.

Parameters	Value	Unit
Locomotive mass	184×10^3	kg
Wagon mass	20×10^3	kg
Locomotive length	20	m
Wagon length	12	m
Coupler length $L_{i-1,i}$	1	m
Train pipe pressure	600	kPa
k_i	10×10^7	$N\,m^{-1}$
d_i	10×10^5	$N\,s\,m^{-1}$
c_0	7.6658×10^{-3}	$N\,kg^{-1}$
c_1	1.08×10^{-4}	$N\,s(m\,kg)^{-1}$
c_2	2.06×10^{-5}	$N\,s^2(m^2\,kg)^{-1}$
Max rheostatic brake b^r_{max}	323	kN
Number of cars	220	

Without loss of generality, it is assumed that there are four locomotives which are evenly spaced in the heavy haul train. As mentioned in Gao et al.,[19] the dynamics of wagons between locomotives are also neglected and regarded as rigid body for the sake of making the results explicit and reducing the unnecessary complexity. That is to say, the optimal control of four locomotive-wagon subgroups is considered. Noting that there are only a finite number control notches on-board to control the level of effort delivered by air braking, which are determined by the train pipe pressure reductions. According to experimental data collected from Shuohuang railways, two optional pressure reductions for heavy haul train are 50 and 70 kPa, and corresponding air braking forces could be calculated according to Rao.[35]

To further coincide with the real-world conditions, we take comprehensive measures to capture the parameters with regard to constraints on practical operation, including questionnaires with experienced drivers about how to drive in a safe and efficient way, observations in the Shuohuang railway by ourselves and so on. The related parameters are described as $\tau = 60\,s$ and $\sigma = 130\,s$.

After performing several preliminarily tests, we set the constant in the harmonic stepsize rule to $a = 5$ for a reasonable convergence and higher quality of solutions. We also consider the exploration rate for choosing random alternative decisions as $\kappa = 0.2$. To balance the experimental simplicity and control accuracy, Δt is set to be 2 s. In addition, we fix the iteration counter to $\mathbb{N} = 2 \times 10^5$ for the ADP algorithm to ensure the convergence. The algorithm is implemented in MATLAB on the Windows 7.0 platform and evaluated on a personal computer with a 3.3 GHz CPU and 4 GB memory.

Next, we present two different cases to test the algorithm. In Case 1, we consider the deterministic situation to test the performance of the ADP algorithm. In Case 2, associated with the real operation conditions, different speed limits and gradients are taken into consideration during the trip. Uncertain disturbance from real-world environment is also taken into account to verify the robustness of the proposed algorithm. Finally, to better illustrate the performance of the proposed method, we implement another set of experiments to compare the performances of the proposed algorithm with other approaches.

Case 1

For the convenience of describing trains movements, the length of relevant railway track section is given as 10 km, and the gradient is supposed to be $-12\permil$ along the whole track. As the heavy haul train is treated as a cascade mass point model, we assume that the
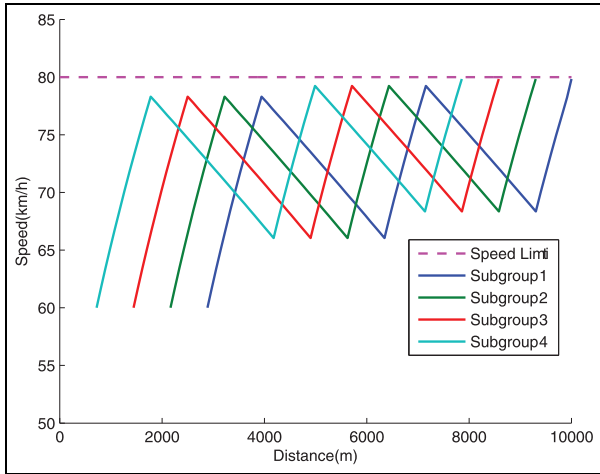
Figure 2. Travel trajectories of different subgroups under uniform track condition.

beginning location of the last car is at the initial site, and the initial coupler displacement between adjacent cars is 0. In other word, all cars are travelling on the same downward slope. Besides, the line limit speed and initial speed are set to be 80 km/h and $v_0 = 60$ km/h, respectively. For the purpose of maintaining the commercial punctuality, the train is supposed to complete the trip within 550 s.

Figure 2 gives the velocity profiles of each locomotive-wagon subgroup with respect to the distance under the optimal policy derived from the ADP

algorithm. As expected, heavy haul train traverses on its route without exceeding the speed limit. As the initial speed is far below the track permitted speed, a coasting operation is first obtained to allow the train accelerates due to the steep descent of the track. Next, braking force is applied to keep the train from overspeeding. Interestingly, a smaller pressure reduction (50 kPa) is selected to slow down the train instead of the bigger one to effectively control the in-train forces. It is obvious that the speed profiles have a cyclic feature to ensure the air-filled time, which are consistent with practice. In addition, it is worth mentioning that all the four locomotive-wagon subgroups adopt the same policy due to the similar operation conditions.

Figure 3 illustrates the learning procedure for the proposed scheme. Specifically, we plot the evaluated objective value of the ADP algorithm in every 2×10^3 iterations up to 2×10^5 iterations. At the beginning, the longitudinal in-train forces are far from the optimal one and gradually reduced from 1.432×10^5 to about 4.74×10^4 kN after 8×10^4 iterations. After about 1.6×10^5 iterations, the difference between adjacent objective values is zero, which means that the learned policy converges to the optimal one.

Case 2

To further test the effectiveness and robustness of the proposed approach, different speed limits and the
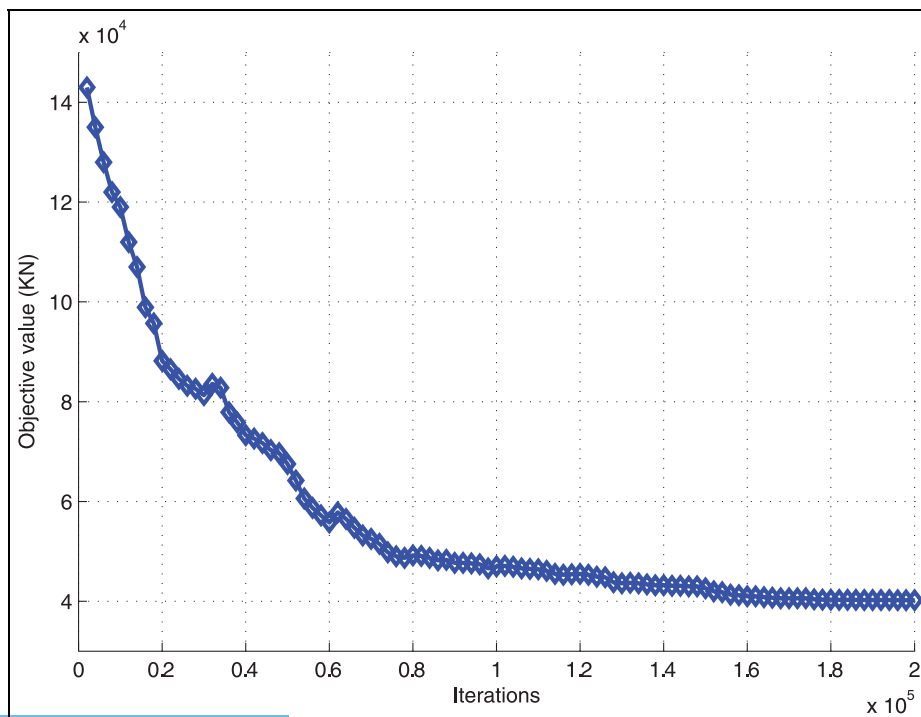
Figure 3. The convergence speed.
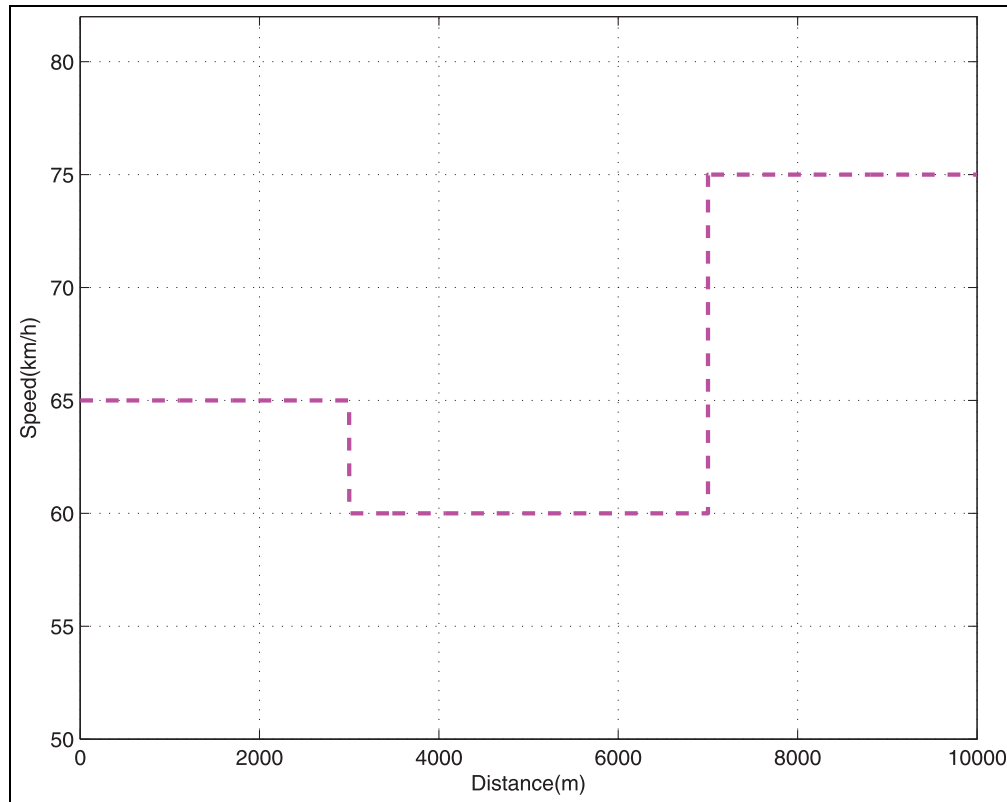
Figure 4. Speed limits.

gradients are considered as shown in Figures 4 and 5, respectively. For the sake of alleviating longitudinal damage to the coupler when releasing the air braking at low speed, the minimum allowable release speed is set to be 35 km/h. The planned trip time $T = 650$ s is longer than Case 1 due to the speed limits and initial speed is set to be $v_0 = 50$ km/h. Other parameters are the same as presented Case 1 unless otherwise noted. In addition, the uncertain disturbances to the longitudinal dynamics of the heavy haul train are assumed to be varying in the interval [0 m s^{-2}, 0.05 m s^{-2}], and the sample path is chosen every 1000 iterations.

Figure 6 shows the velocity/distance curves under our ADP policy. There is no doubt that the complexity of track conditions and speed limits make driving on steep descent more challenging. We can observe that our optimal policy can achieve safe driving as it never exceeds the speed limit, while maintaining the release speed bigger than minimum release speed. Thus, the train can run safely along such a continuous decent. Moreover, each locomotive-wagon subgroup follows a different optimal strategy compared to a unique one in case 1. This is because each intelligent agent needs to dynamically adapt the track condition to achieve the optimal performance. Given that the deceleration depends also on the gradient of the track, we observe that the on-board air braking notch with bigger

pressure reduction (70 kPa) is chosen at certain decision epoch to balance the deceleration of different subgroups in order to reduce the longitudinal coupler forces. The results show that ADP policy is able to learn information from the uncertain environment and find an optimal policy with real-time data.

Comparative analysis

As one of the most efficient algorithms from the reinforcement learning, Q-learning provides a nice mechanism using the value function.[36] The Q-factor $Q(s, a)$, which stores the value of a state-decision pair, is introduced to estimate the value of being in a state and taking a particular decision. To better illustrate the performance of the proposed method, we implement another set of experiments to compare the performance of the proposed ADP algorithm with Q-learning method. The harmonic stepsize rule and exploration rate $\eta = \kappa / n'(S_t, x_t)$ are used in the Q-learning method, and the control parameters remain the same as those used in Case 1 unless otherwise noted. The comparisons of convergence are shown in Figure 7.

We observe that Q-learning algorithm converges faster than ADP in early iterations, but ADP outperforms Q-learning after about 24,000 iterations. The optimal objective values obtained by the ADP and
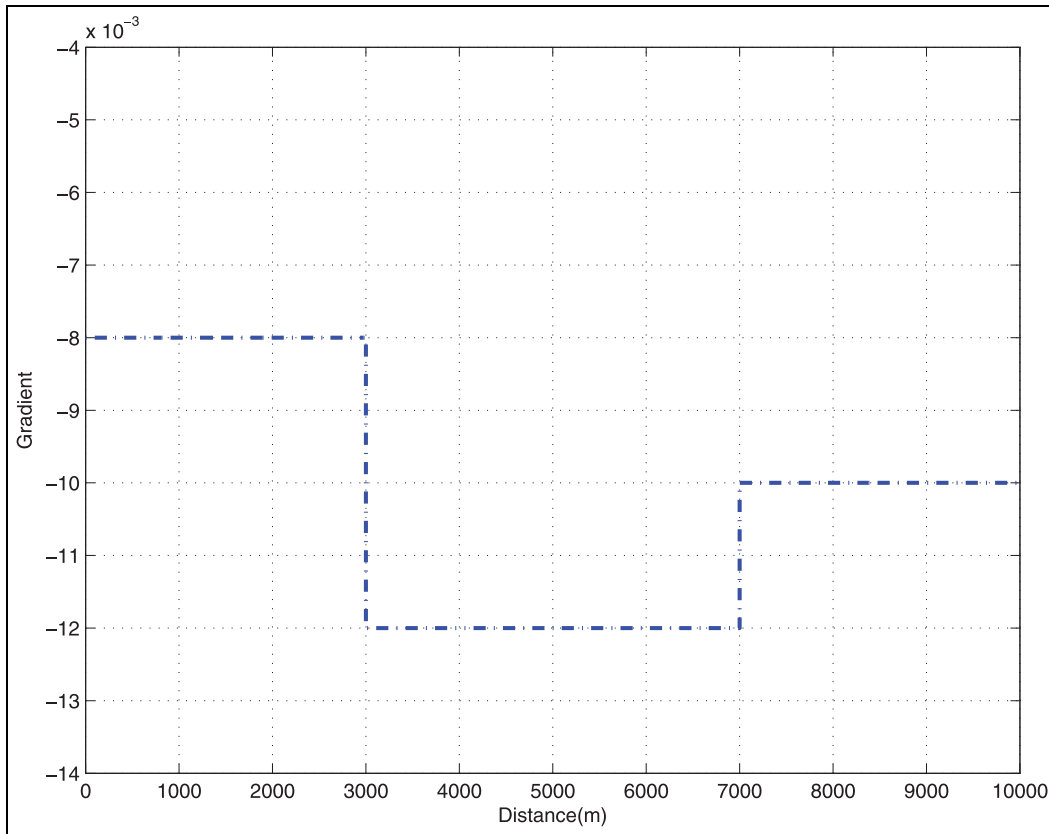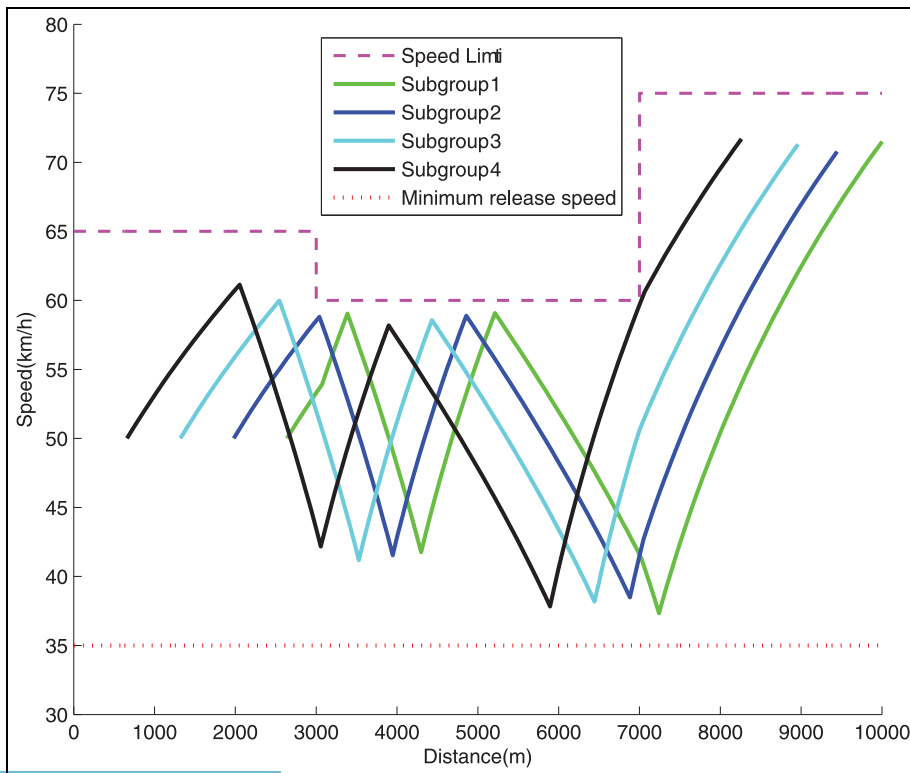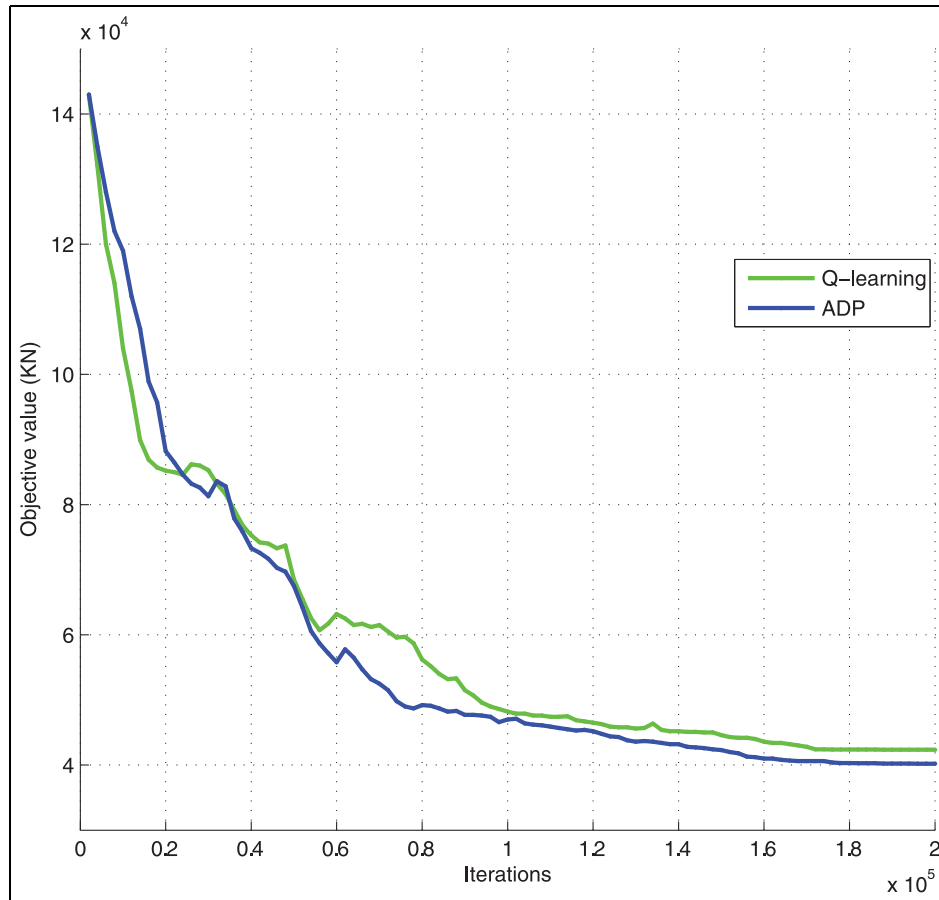
Figure 5. Gradients.

Figure 6. Travel trajectories of different subgroups under different track conditions.

Figure 7. The comparisons of convergence.

Table 2. Performance comparison.

Strategy	Objective value (kN)		
	Ave	Max	Min
LCS	4.73×10^4	4.87×10^4	4.65×10^4
ICS	5.24×10^4	5.33×10^4	5.12×10^4
The proposed strategy	4.02×10^4	5.81×10^4	5.56×10^4

LCS: leading controlled strategy; ICS: independently controlled strategy.

Q-learning algorithm are 4.02×10^4 and 4.14×10^4 kN, indicating that ADP algorithm has better performance in manipulating the in-train forces than Q-learning algorithm.

To the best of our knowledge, the optimal control action is either determined based on information of the leading locomotive (leading controlled strategy (LCS)) or calculated by each locomotive independently (independently controlled strategy (ICS)) in the majority of existing works. To show the advantages of the proposed method over LCS and ICS, we implement another set of experiments in the uncertain environment. We use the Monte Carlo simulation and run the simulation 100 times for all the mentioned approaches. The control

parameters remain the same as those used in Case 2 unless otherwise noted. The performance values of different methods are shown in Table 2.

As LCS and ICS mainly control the heavy haul train according to current information, future behaviours during the whole travel are usually not taken into consideration. Therefore, these two approaches are typical myopic ones, and they cannot accomplish an overall optimization for the heavy haul train movement due to the uncertainty during a long trip. It is shown from Table 2 that there are 17.8% and 30.3% gaps between the myopic objective values and the optimal one. On the contrary, our optimal policy, which captures the impact of decisions on the future and communicates

this impact backwards, is capable of giving the best performance compared with other two methods in all indices. Therefore, it is safe to conclude that considering the future impact when making control decisions is crucial in practical heavy haul train operation, and the proposed method is effective in enhancing safety and punctuality.

Conclusion and future work

This article deals with the optimal air braking control problem of heavy haul train on steep descent. In order to minimize the in-train coupler forces, while keeping to schedule and enhancing safety, a multistage stochastic dynamic programming model is designed with consideration of operational constraints and uncertain disturbances to the practical train movement. The model is capable of capturing the impact of decisions on the future, and then communicating this impact backwards so that decisions can be made more intelligently. To search for an optimal control strategy, ADP algorithm is introduced to step forward in time by making decisions based on the approximate value function. The performance of the proposed model and algorithm is validated by numerical experiments.

For future work, we plan to extend approximate value function to other forms such as weighted sum of basis functions, piecewise linear functions and regression models, which may outperform the lookup table ADP in terms of both solution accuracy and time.

References

1. Chang C, Wang C, Chen B, et al. A study of a numerical analysis method for the wheel-rail wear of a heavy-haul train. *Proc IMechE, Part F: J Rail and Rapid Transit* 2010; 224: 473–482.
2. Zhuan XT and Xia XH. Optimal scheduling and control of heavy haul trains equipped with electronically controlled pneumatic braking systems. *IEEE T Contr Syst T* 2007; 15: 1159–1166.
3. Yang LX, Li KP, Gao ZY, et al. Optimizing trains movement on a railway network. *Omega* 2012; 40: 619–633.
4. McClanachan M and Cole C. Current train control optimization methods with a view for application in heavy haul railways. *Proc IMechE, Part F: J Rail and Rapid Transit* 2011; 225: 1–12.
5. Yasunobu S and Miyamoto S. A fuzzy control for train automatic stop control. *Trans Soc Instrum Control Eng* 2002; 21: 1–9.
6. Howlett P. Optimal strategies for the control of a train. *Automatica* 1996; 32: 519–532.
7. Khmelnitsky E. On an optimal control problem of train operation. *IEEE T Automat Contr* 2000; 45: 1257–1266.
8. Zhang LJ and Zhuan XT. Optimal operation of heavy-haul trains equipped with electronically controlled pneumatic brake systems using model predictive control methodology. *IEEE T Contr Syst T* 2014; 22: 13–22.
9. Li SK, De Schutter B, Yang LX, et al. Robust model predictive control for train regulation in underground railway transportation. *IEEE T Contr Syst T* 2016; 24: 1075–1083.
10. Wang X, Zhao Y and Tang T. Fuzzy constrained predictive optimal control of high speed train with actuator dynamics. *Discrete Dyn Nat Soc* 2016; 2016: 5704743-1–5704743–14.
11. Sun Y, Cole C, Spiryagin M, et al. Longitudinal heavy haul train simulations and energy analysis for typical Australian track routes. *Proc IMechE, Part F: J Rail and Rapid Transit* 2014; 228: 355–366.
12. Zhang LJ and Zhuan XT. Development of an optimal operation approach in the MPC framework for heavy-haul trains. *IEEE T Contr Syst T* 2015; 16: 1391–1400.
13. Zhuan XT and Xia XH. Cruise control scheduling of heavy haul trains. *IEEE T Contr Syst T* 2006; 14: 757–766.
14. Nasr A and Mohammadi S. The effects of train brake delay time on in-train forces. *Proc IMechE, Part F: J Rail and Rapid Transit* 2010; 224: 523–534.
15. Chou M and Xia XH. Optimal cruise control of heavy-haul trains equipped with electronically controlled pneumatic brake systems. *Control Eng Pract* 2007; 15: 511–519.
16. Bai Y, Mao B, Ho T, et al. Station stopping of freight trains with pneumatic braking. *Math Probl Eng* 2014; 2014: 172549-1–172549-7.
17. Li X, Shou B and Ralescu D. Train rescheduling with stochastic recovery time: a new track-backup approach. *IEEE T Syst Man Cy A* 2014; 44: 1216–1233.
18. Li SK, Yang LX and Gao ZY. Adaptive coordinated control of multiple high-speed trains with input saturation. *Nonlinear Dynam* 2016; 83: 2157–2169.
19. Gao K, Huang ZW, Wang J, et al. Decentralized control of heavy-haul trains with input constraints and communication delays. *Control Eng Pract* 2013; 21: 420–427.
20. Li SK, Yang LX and Li KP. Robust output feedback cruise control for high-speed train movement with uncertain parameters. *Chinese Phys B* 2015; 24: 010503.

21. Sun WC, Li JF, Zhang LX, et al. Active suspension control with frequency band constraints and actuator input delay. *IEEE T Ind Electron* 2012; 59: 530–537.

22. Yao JY, Jiao ZX, Ma DW, et al. High-accuracy tracking control of hydraulic rotary actuators with modelling uncertainties. *IEEE/ASME T Mech* 2014; 19: 633–641.

23. Chen Z, Yao B and Wang Q. μ-synthesis based adaptive robust control of linear motor driven stages with high-frequency dynamics: a case study with comparative experiments. *IEEE/ASME T Mech* 2015; 20: 1482–1490.

24. Powell WB, Simao HP and Bouzaiene-Ayari B. Approximate dynamic programming in transportation and logistics: a unified framework. *EURO J Transp Logist* 2012; 1: 237–284.

25. Powell WB. *Approximate dynamic programming: solving the curses of dimensionality*. New York: John Wiley & Sons, 2007.

26. Powell WB, Bouzaiene-Ayari B, Lawrence C, et al. Locomotive planning at Norfolk Southern: an optimizing simulator using approximate dynamic programming. *Interfaces* 2014; 44: 567–578.

27. Sever D, Dellaert N, Van Woensel T, et al. Dynamic shortest path problems: hybrid routing policies considering network disruptions. *Comput Oper Res* 2013; 40: 2852–2863.

28. Stimpson D and Ganesan R. A reinforcement learning approach to convoy scheduling on a contested transportation network. *Optim Lett* 2015; 9: 1641–1657.

29. Schmid V. Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *Eur J Oper Res* 2012; 219: 611–621.

30. Simao HP, Day J, George AP, et al. An approximate dynamic programming algorithm for large-scale fleet management: a case application. *Transport Sci* 2009; 43: 178–197.

31. Davis WJ. *The tractive resistance of electric locomotives and cars*. Fairfield, CT: General Electric, 1926.

32. Chevrier R, Pellegrini P and Rodriguez J. Energy saving in railway timetabling: a bi-objective evolutionary approach for computing alternative running times. *Transport Res C: Emer* 2013; 37: 20–41.

33. Li SK, Yang LX, Li KP, et al. Robust sampled-data cruise control scheduling of high speed train. *Transport Res C: Emer* 2014; 46: 274–283.

34. Ma DW. The running safety and longitudinal force of heavy haul trains on downhill slope of daqin line. *Rolling Stock* 2005; 43: 1–5.

35. Rao Z. *Train traction calculation*. Beijing, China: China Railway, 2006.

36. Sutton RS and Barto AG. *Reinforcement learning: an introduction*. Cambridge, MA: MIT Press, 1998.