The design and analysis of a generalized DPR algorithm for rare event simulation

Thomas Dean* and Paul Dupuis†

July, 2008

Abstract

We consider a general class of branching methods with killing for the estimation of rare events. The class includes a number of existing schemes, including RESTART and DPR. A method for the design and analysis is developed when the quantity of interest can be embedded in a sequence whose limit is determined by a large deviation principle. A notion of subsolution for the related calculus of variations problem is introduced, and two main results are proved. One is that the number of particles and the total work scales subexponetially in the large deviation parameter when the branching process is constructed according to a subsolution. The second is that the asymptotic performance of the schemes as measured by the variance of the estimate can be characterized in terms of the subsolution. Examples are given, and methods that use killing are compared with the analogous schemes without killing.

1 Introduction

The need to calculate the probabilities of rare events arises in many fields, for example operations research, engineering, physics and chemistry to name just a few. By a rare event, we mean one whose probability obeys a large deviations rule (see [10, 15]), and in particular that the probability is one of a sequence of probabilities p_n such that for some $\gamma > 0$

$$\lim_{n\to\infty} -\frac{1}{n}\log p_n = \gamma.$$

*Nomura Centre for Mathematical Finance, Mathematical Institute, University of Oxford, Oxford, OX1 3LB, UK. Research of this author supported in part by the National Science Foundation (NSF-DMS-0404806 and NSF-DMS-0706003).

†Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, RI, 02912, USA. Research of this author supported in part by the National Science Foundation (NSF-DMS-0404806 and NSF-DMS-0706003), the Army Research Office (W911NF-05-1-0289), and the Air Force Office of Scientific Research (FA9550-07-1-0544).

Often it is necessary to estimate these probabilities numerically and there is much interest in developing accurate and efficient Monte Carlo algorithms to do this.

In Monte Carlo one simulates K independent and identically distributed (iid) copies of a random variable $\hat{s}^n$ whose mean is p_n, and then estimates p_n by the sample mean. Since the variance of the estimate is K^{-1} times the variance of a single sample, performance can be measured in terms of the variance of $\hat{s}^n$, and since $E[\hat{s}^n] = p_n$ it follows that minimizing the variance is equivalent to minimizing the second moment. A standard measure of performance for rare event simulation is the asymptotic relative error $\lim_{n\to\infty} \left[\log E\left[(\hat{s}^n)^2\right] / \log p_n\right]$. However this measure ignores the computational cost of the algorithm, a cost that can become significant in some circumstances, e.g., for a poorly designed splitting algorithm. For our purposes a more suitable measure of performance is the asymptotic work-normalised relative error

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{E\left[(\hat{s}^n)^2\right] E\left[w^n\right]}{p_n^2},$$

where w^n denotes the computational cost of generating a single sample of the n^{th} estimator. This measure is essentially the same as the asymptotic work-normalised error proposed by Glasserman et. al. in [19].

For typical problems of rare event estimation (e.g., hitting probabilities), the work required by naive Monte Carlo grows subexponentially, and a straightforward calculation shows that the asymptotic work-normalised relative error is equal to the large deviations rate γ. Hence the ratio of the work-normalised variance to the square of the probability being estimated grows exponentially in n. It follows that an enormous number of samples is required for the standard deviation of the resulting estimator to be sufficiently small, and so alternative Monte Carlo methods are sought. By Jensen's inequality the asymptotic work-normalised relative error is always greater than or equal to zero. The aim is thus to find an alternative Monte Carlo method such that this value is as close to zero as possible. A Monte Carlo method that achieves the value zero is said to be asymptotically optimal.

The two most common alternative Monte Carlo methods in the context of rare event estimation are importance sampling and multi-level splitting. For an overview of the field of rare event simulation see [1]. The present paper is concerned with the design and analysis of a particular multi-level splitting method called the Direct Probability Redistribution (DPR) algorithm [20, 21], which is a generalization of the RESTART (REpetitive Simulation Trials After Reaching Thresholds) algorithm [2].

Multi-level splitting algorithms are often used simulate probabilities of the form $P\{X \in \mathcal{A}\}$, where $X = \{X_i, i = 0, \ldots\}$ is a discrete time stochastic process and $\mathcal{A}$ is some subset of the path space $\mathcal{D}$. (Splitting for continuous time processes will not be considered in this paper.) The multi-level splitting philosophy is to simulate particles that evolve according to the law of $\{X_i\}$, and at certain times split those particles considered more likely to lead to a trajectory that belongs to the set $\mathcal{A}$. For example, $\mathcal{A}$ might be the trajectories which reach some unlikely set B before hitting a likely set A, after starting in neither A nor B. In this case, the splitting will favor migration towards B. Particles which are split are given an appropriate weighting to ensure that the algorithm remains unbiased. Broadly speaking there are two types of multi-level splitting algorithm, those with and without killing, where *stopping* is distinguished from killing. In the example just mentioned particles are stopped upon entry into either A or B. Killing involves abandoning a particle prior to entry into either A or B, presumably because continuation of the trajectory is not worth the computational effort. Care must be taken that any killing will not introduce bias.

To the authors' knowledge there is only one type of multi-level splitting algorithm without killing - the splitting algorithm (see [19] for further references). The standard implementation of this algorithm requires a sequence of sets $C_1 \supset \cdots \supset C_J$, called *splitting thresholds*, and a sequence of positive integers $R_1, \ldots, R_J$, called *splitting rates*. A single particle is started at the initial position X_0 and evolves according to the law of $\{X_i\}$. When a particle enters a set C_j for the first time it produces $R_j - 1$ offspring. After splitting has occurred all particles evolve independently of each other. Each particle is stopped according to the stopping rule associated with $\{X_i\}$ and the algorithm terminates when all the particles generated have been stopped. The probability of interest is approximated by $N/\prod_{i=1}^{J} R_i$, where N is the number of particles simulated whose trajectories belong to $\mathcal{A}$. A more general version of this algorithm lets the splitting rates R_i take non-negative real values, in which case the number of offspring is randomized.

Although the splitting algorithm can be very effective there is one clear source of inefficiency when dealing with rare events. The vast majority of the particles generated will not have trajectories that belong to the set $\mathcal{A}$, and so much of the computational effort is devoted to generating trajectories that do not make any direct contribution. Multi-level splitting algorithms with killing were introduced as a way to mitigate this problem. The first such algorithm was the RESTART algorithm, introduced in [2] and [3]. Its implementation is identical to the standard splitting algorithm except that particles are split every time they enter a splitting threshold and particles are killed when they exit the splitting threshold in which they were born. The initial particle is assumed to be born in the set C_0, which by convention is equal to the state space of the process $\{X_i\}$, and so this particle is never killed.

The standard version of the RESTART algorithm requires that the splitting rates be integer valued and that the process not upcross more than one splitting threshold in each time step. The DPR algorithm, introduced in [20] and [21], generalises this. Although in motivation it differs, in implementation it is identical to the RESTART algorithm except that the above restrictions are lifted. Because the DPR algorithm is a generalisation of the RESTART algorithm the discussion in the rest of the paper focuses on the DPR algorithm.

At present RESTART and DPR algorithms have been applied to the estimation of hitting and stationary probabilities of stochastic processes [3, 5, 21]. In [4] a formula is obtained for the variance of the RESTART algorithm, but to date no analogous expressions have been obtained for the DPR algorithm. Further there does not exist any systematic and rigorous framework for the design and analysis (bounds on the asymptotic work-normalised relative error) for RESTART or DPR algorithms.

The purpose of this paper is to extend the framework for designing and analysing splitting algorithms for rare event simulation as presented in [9] to the case of DPR. This framework uses subsolutions to a Hamilton-Jacobi-Bellman (HJB) partial differential equation (PDE) naturally associated with the probability via a large deviations analysis. A practical advantage of the approach is that for an interesting and growing class of problems one can find appropriate subsolutions to the HJB equation that generate asymptotically optimal splitting algorithms. The work in [9] is itself an extension of the results presented in [13], which focus on algorithms based on importance sampling for rare event estimation. There is, however, an important distinction in the types of subsolution required for splitting type schemes and importance sampling. For importance sampling, one needs functions that are either *classical* sense subsolutions, or more generally the minimum of a finite collection of functions which satisfy the equation in a classical sense at all points where the gradient is defined. In contrast, splitting type schemes require only a subsolution in the viscosity sense.

The main results in this paper are as follows. In Section 3 a generalised DPR (GDPR) algorithm is defined for estimating expected values of the form

$$E\left[\sum_{i=0}^{\tau} f(X_i)\right], \tag{1.1}$$

rather than just hitting and stationary probabilities. Standing assumptions which will be in force throughout this paper are that τ is the (a.s. finite) time of first entry into some closed set M, and $f(x)$ is a non-negative measurable function. In Section 4 a specific and natural implementation of the GDPR algorithm is proposed and formulae for the computational cost and second moment of the GDPR algorithm are derived.

Sections 5 and 6 consider the asymptotic problem. In Section 5 a method for designing GDPR algorithms based on the subsolution framework in [9] is developed. Expressions for the asymptotic work-normalised relative error of such algorithms are derived using the formulae developed in the previous section, and subsolutions which lead to asymptotically optimal performance are identified. Section 6 describes a PDE based method for characterizing subsolutions which is simpler to use in practice. The last three sections are devoted to numerical results and a comparison with ordinary splitting, some short conclusions and an appendix containing proofs that were deferred from the body of the paper.

2 Notation and Terminology

Some conventions are as follows. The letter i exclusively denotes a discrete time variable. Random processes are denoted using capitals, e.g., $\{X_i, i = 0, \ldots\}$, with X_i the state of the process at time i. For simplicity it is assumed that $X_0 = x_0$ is deterministic, though all results can easily be generalized to the case when this is no longer true. The letter D is used to denote the state space of a random process, $\mathcal{D}$ to denote the corresponding path space, and δ_x to denote the standard Dirac probability measure at x. It is assumed that $D \subset \mathbb{R}^d$ for some d. Although we will later consider processes $\{X_i, i = 0, \ldots\}$ as elements of a sequence that satisfies a large deviation property, for notational simplicity the large deviation index is initially suppressed.

The stationary measure of $\{X_i\}$ [assuming one exists] is denoted by π, and the induced measure on the state space at time i is denoted $\pi_i(dx)$. Thus for $A \subset D$

$$\pi_i(A) = E_{x_0}\left[1_A\left(X_i\right)\right].$$

The following non-standard notation is also used. Branching processes are denoted with overbars, e.g., $\left\{\bar{X}_i, i = 0, \ldots\right\}$. Each branching process has a $\mathbb{Z}_+$-valued process $N_i, i = 0, \ldots$ associated with it, where N_i equals the number of particles present in the branching process at time i. For each $i = 0, \ldots$ and $j = 1, \ldots, N_i$, $\bar{X}_{i,j}$ denotes the state of the j^{th}-particle at time i, and $(\bar{X}_{0,j}, \ldots, \bar{X}_{i,j})$ denotes the path history of the j^{th}-particle at time i. By convention a particle created by splitting inherits the parent particle's path history. We also define a pair of measures relevant to branching processes by

$$\bar{\pi}_i(A) \doteq E_{x_0}\left[\sum_{j=1}^{N_i} 1_A\left(\bar{X}_{i,j}\right)\right] \quad \text{and} \quad \bar{\delta}_{\bar{X}_i} \doteq \sum_{j=1}^{N_i} \delta_{\bar{X}_{i,j}}.$$

Note that these are typically not probability measures. The first is referred to as un-normalized induced measure and the second as an un-normalized empirical measure.

Recall that multi-level splitting algorithms are defined via sequences of nested sets $C_1 \supset \cdots \supset C_J$ and splitting rates $R_1, \ldots, R_J$. By convention C_0 is equal to D, $C_{J+1} = \emptyset$, and $R_0 = 1$. Defining algorithms through levels and rates quickly becomes notationally cumbersome. In addition, it is not well suited to the analysis of a sequence of problems indexed by a large deviation parameter. In this paper "importance functions" will be used to identify the algorithm data. An importance function is a non-negative step function $V(x)$ such that there is a sequence of sets $\tilde{C}_1 \supset \cdots \supset \tilde{C}_{\tilde{J}}$, with the property that $V(x)$ is constant on each $\tilde{C}_j/\tilde{C}_{j+1}$, $V(x) > V(y)$ for all $x \in \tilde{C}_j$, $y \in \tilde{C}_j^c$ and $V(x) = 0$ for all $x \in \tilde{C}_1^c$. For a given importance function it is convenient to use the notation $\tilde{C}_0 = \tilde{C}_1^c$ and to let $V_j, j = 1, \ldots, \tilde{J}$ denote the value taken by the importance function on the set $\tilde{C}_j/\tilde{C}_{j+1}$. As we will see later on, it is possible to obtain a collection of importance functions corresponding to a collection of values of the large deviation index from a single "generating" function in a convenient manner.

Each importance function $V(x)$ naturally defines a sequence of splitting thresholds and splitting rates by $J = \tilde{J}$ and $C_j = \tilde{C}_j$, $R_j = \exp(V_j - V_{j-1})$ for $j = 1, \ldots, J$. By convention $\exp(V_0 - V_{-1}) = 1$. Conversely, given a sequence of splitting thresholds and splitting rates one can define the related importance function by

$$V(x) \doteq \max_{j: x \in C_j} \left\{ \log \left(\prod_{k=0}^{j} R_k \right) \right\}.$$

Hence there is a one-to-one correspondence between importance functions and sequences of splitting thresholds and splitting rates, and so there is no loss of generality in using importance functions to define multi-level splitting schemes. Given $V(x)$ and $x \in D$ let $\rho(x)$ be the unique integer j such that $x \in C_j/C_{j+1}$.

The term Monte Carlo (DPR, GDPR, etc.) *algorithm* will be used to refer to a single Monte Carlo algorithm and the term Monte Carlo (DPR, GDPR, etc.) *scheme* will be used to refer a sequence of Monte Carlo (DPR, GDPR, etc.) algorithms used to estimate a sequence of probabilities or expected values.

Finally some terminology is needed to concisely describe how offspring are generated when a particle splits. For the RESTART algorithm this is simple. Every time a particle enters a splitting threshold C_j, a deterministic number of offspring $R_j - 1$ are generated. Since particles are destroyed when they exit the splitting threshold in which they are born, each particle has an integer attached to it to record this splitting threshold. These are referred to as the support thresholds of the particles. For DPR and GDPR the number of particles can be random, and since particles can jump more than one level in a single step, the construction of an unbiased algorithm requires that the support threshold also be randomized.

Let $\mathbb{S}$ be the elements $q \in \mathbb{Z}_+^\infty$ such that $q_j = 0$ for all sufficiently large j. Vectors $q \in \mathbb{S}$ will be referred to as splitting vectors, and the term splitting distribution will mean a probability measure $\mathcal{Q}$ on $\mathbb{S}$. The splitting process can then be described by randomly assigning to each particle that splits a vector $q \in \mathbb{S}$. The number of *new* particles will be equal to $\sum_{j=0}^{\infty} q_j$, and precisely q_j of the new particles will be given support threshold j. The distribution of this random vector will depend on the parent particle's current and immediately prior thresholds.

Although this notation is more complicated than necessary in the context of RESTART, it proves to be very useful for describing the DPR and GDPR algorithms.

Consider the DPR algorithm. If a particle moves from C_j/C_{j+1} to C_k/C_{k+1}, $k > j$, then splitting occurs. All offspring and the parent are located in C_k/C_{k+1}, and the support threshold of each new particle is an element of $\{j+1,\ldots,k\}$. Numbers of offspring and their support thresholds are independent of all past data except through the values of j and k. It follows that each DPR algorithm will induce a splitting distribution on the set $\mathbb{S}$ for each pair $1 \leq j < k \leq J$. Conversely, any collection of splitting distributions $\mathcal{Q}_{j,k}, 1 \leq j < k \leq J$ will define how particles are split for a particular DPR algorithm. Such collections provide a concise way of describing the splitting mechanisms of a given DPR algorithm, and will used in the next section to describe how particles in the GDPR algorithm should be split.

3 The GDPR Algorithm

For the rest of this paper the following condition will hold.

Condition 3.1 *$f \geq 0$ and τ is almost surely finite.*

The DPR algorithm can be motivated as follows. Suppose a process $\{X_i\}$ with state space D has stationary measure π, and that for some set A, $\pi(A)$ is so small that estimating $\pi(A)$ using standard Monte Carlo techniques is impractical. Suppose further that for an importance function $V(x)$ one could simulate a process $\hat{X} = \{\hat{X}_i, i = 0,\ldots\}$ taking values in the same state space with stationary measure

$$\hat{\pi}(B) = \int_B e^{V(x)}\pi(dx) \Big/ \int_D e^{V(x)}\pi(dx).$$

One could then approximate $\pi(A)$ by approximating

$$\int_A e^{-V(x)}\hat{\pi}(dx) \Big/ \int_D e^{-V(x)}\hat{\pi}(dx)$$

via simulation.

Following the standard logic of acceleration methods generally, the hope is that with a well chosen importance function $V(x)$ the variance of the estimator using $\hat{X}$ is made lower than that of the original estimator by building in information regarding the underlying process and the event of interest. The DPR method essentially follows this approach, except that a branching process is used rather than an ordinary Markov process, and the amplification factor $\exp(V(x))$ which multiplies $\pi(dx)$ is produced by the branching (see [21, Figure 4]). The essential points are:

- A birth-death branching process $\bar{X} = \left\{\bar{X}_i, i = 0,\ldots,T\right\}$ is simulated by splitting the original process.

- Given an importance function $V(x)$ the process $\bar{X}$ has un-normalized induced measure $\bar{\pi}_i(dx) = \exp(V(x))\pi_i(dx)$.

- $\pi(A)$ is then approximated by

$$\frac{1}{T}\sum_{i=0}^{T}\int_D 1_A\left(x\right) e^{-V(x)}\bar{\delta}_{\bar{X}_i}(dx). \tag{3.1}$$

To obtain recursive formulas for the GDPR algorithm we will need to allow for varying initial conditions. If we examine a generic particle at some time after the algorithm has started, then it will be in a set of the form C_j/C_{j+1} and have a killing threshold in $\{0,\ldots,j-1\}$. Treating this as an initial condition requires that the GDPR algorithm be defined in such a way that it actually simulates a branching process, defined via splitting of X, that has the un-normalized induced measure

$$\bar{\pi}_i(dx) = e^{V(x)-V(x_0)}\pi_i(dx).$$

For an unbiased algorithm, the killing threshold must take a prescribed form that will be identified below. For the purposes of defining the algorithm with general initial condition we temporarily denote the distribution of the support threshold of the initial particle by $\mathcal{I}$ and call it the *initialising distribution*.

In order to accommodate stopping times, general cost functions as in (1.1) and general initial conditions, one should replace (3.1) by

$$e^{V(x_0)}\sum_{i=0}^{\infty}\int_D f(x)e^{-V(x)}\bar{\delta}_{\bar{X}_i^{\tau}}(dx) = e^{V(x_0)}\sum_{i=0}^{\infty}\int_D \bar{f}(x)\bar{\delta}_{\bar{X}_i^{\tau}}(dx),$$

[where $\bar{f}(x) \doteq f(x)\exp(-V(x))$] for a suitably defined version of the branching process, where contributions to the integral are terminated after each branched trajectory first enters $M \subset D$. This observation motivates the definition of a generalized DPR algorithm (GDPR) that follows. The killing of particles upon entry into M should be kept logically distinct from the killing introduced to enhance algorithmic efficiency. The superscript τ in $\bar{X}^{\tau}$ is used to indicate that trajectories are killed after entry into M. Also, the definition $\bar{f} \doteq fe^{-V}$ will be used extensively in the sequel.

The splitting thresholds, splitting rates and splitting processes of the algorithm will be defined using importance functions V and splitting and initialization distributions $\mathcal{Q}_{j,k}$ and $\mathcal{I}$ as described previously. The generalized DPR algorithm, with the dependence on these quantities suppressed in the notation, can be written in pseudo code as follows.

Generalized DPR Algorithm (GDPR)

```
Variables:
  $i$ current time
  $N^\tau_i$ number of particles at time $i$
  $\bar{X}^\tau_{i,j}$ position of $j^{th}$ particle at time $i$
  $C^\tau_{i,j}$ current threshold of $j^{th}$ particle at time $i$
  $L^\tau_{i,j}$ support threshold of $j^{th}$ particle at time $i$
  $\hat{s}(\bar{f})$ (at termination) an estimator of $E_x[\sum_{i=1}^{\tau} f(X_i)]$
  $j, k, l$ counting variables
  $Y_{i,j}$ free variables
Initialization Step:
  $N^\tau_0 = 1$, $\bar{X}^\tau_{0,1} = x_0$, $C^\tau_{0,1} = \rho(x_0)$, $\hat{s}(\bar{f}) = \bar{f}(\bar{X}_{0,1})$, $i = 0$
  generate a random variable $L$ with distribution $\mathcal{I}$
  $L^\tau_{0,1} = L$
Main Algorithm:
  while $N^\tau_i \neq 0$
    $N^\tau_{i+1} = 0$
    for $j = 1, \dots, N^\tau_i$
      Test to see if the particle is not killed due to
      stopping:
      if $\bar{X}^\tau_{i,j} \notin M$
        generate a random variable $Y_{i,j}$ with distribution
        $P\left(Y_{i,j} \in dy\right) = P\left(X_{i+1} \in dy \middle| X_i = \bar{X}^\tau_{i,j}\right)$

        Test to see if the particle is not killed
        due to threshold:
        if $\rho(Y_{i,j}) \geq L^\tau_{i,j}$
          $\hat{s}(\bar{f}) = \hat{s}(\bar{f}) + \bar{f}(Y_{i,j})$
          $N^\tau_{i+1} = N^\tau_{i+1} + 1$
          $\bar{X}^\tau_{i+1,N^\tau_{i+1}} = Y_{i,j}$
          $C^\tau_{i+1,N_{i+1}} = \rho(Y_{i,j})$
          $L^\tau_{i+1,N_{i+1}} = L^\tau_{i,j}$
        end

        Test to see if the particle should be branched:
        if $\rho(Y_{i,j}) > C^\tau_{i,j}$
          let $Q^{C^\tau_{i,j},\rho(Y_{i,j})}$ be an independent sample from
          the law $\mathcal{Q}_{C^\tau_{i,j},\rho(Y_{i,j})}$
          for $k = 1, \dots, J$
            for $l = 1, \dots, Q_k^{C^\tau_{i,j},\rho(Y_{i,j})}$
              $\hat{s}(\bar{f}) = \hat{s}(\bar{f}) + \bar{f}(Y_{i,j})$
            $N^\tau_{i+1} = N^\tau_{i+1} + 1$
            $\bar X^\tau_{i+1,N^\tau_{i+1}} = Y_{i,j}$
            $C^\tau_{i+1,N^\tau_{i+1}} = \rho(Y_{i,j})$
            $L^\tau_{i+1,N^\tau_{i+1}} = k$
          end
        end
      end
    end
  end
  $i = i + 1$
end
$\hat s(\bar f) = \exp(V(x_0))\hat s(\bar f)$
```

Note that $\hat{s}(\bar{f}) = e^{V(x_0)}\sum_{i=0}^{\infty}\int_D \bar{f}(y)\bar{\delta}_{\bar{X}_i^\tau}(dy)$ as claimed. An algorithm resulting from an importance function V, a collection of splitting distributions $\mathcal{Q}_{j,k}$ and an initializing distribution $\mathcal{I}$ will be said to be unbiased if

$$E\left[\hat{s}(\bar{f})\right] = E_{x_0}\left[\sum_{i=0}^{\tau} f(X_i)\right]$$

for all suitable f and τ. Recall that the splitting rates R_k are defined in terms of the importance function by $\exp(V_k - V_{k-1})$. Given an importance function V define distributions $\mathcal{L}_k$ on $\{0,\ldots,k\}$ and $\mathcal{L}_{j,k}, j \leq k$ on $\{j+1,\ldots,k\}$ by

$$\mathcal{L}_{j,k}(l) = (e^{V_l} - e^{V_{l-1}})/(e^{V_k} - e^{V_j}) \text{ and } \mathcal{L}_k(l) = (e^{V_l} - e^{V_{l-1}})/e^{V_k}. \tag{3.2}$$

Theorem 3.2 *Assume Condition 3.1. Suppose that an importance function V and stochastic process $\{X_i\}$ with initial condition $X_0 = x_0$ are given. If $\mathcal{I} = \mathcal{L}_{\rho(x_0)}$ and*

$$E_{\mathcal{Q}_{j,k}}\left[Q_l^{j,k}\right] = (R_l - 1)\prod_{r=j+1}^{l-1} R_r = \left(e^{V_l} - e^{V_{l-1}}\right)/e^{V_j} = \mathcal{L}_{j,k}(l)\left(e^{V_k} - e^{V_j}\right)/e^{V_j}, \tag{3.3}$$

then the resulting GDPR algorithm is unbiased.

When in the sequel we refer to *unbiased initial and splitting distributions*, it is assumed that they satisfy the conditions of Theorem 3.2, and indeed for the rest of this paper only unbiased distributions will be considered. The proof of Theorem 3.2 relies on the following lemma.

Lemma 3.3 *Assume Condition 3.1 and let $\bar{f}(x) = f(x)e^{V(x)}$. Let $i \in \mathbb{Z}_+$ be given. For a GDPR scheme with unbiased initialising and splitting distributions,*

$$e^{V(x_0)}E_{x_0}\left[\sum_{m=1}^{N_i^\tau} \bar{f}(\bar{X}_{i,m}^\tau)1_{\{L_{i,m}^\tau = l\}}\right] = E_{x_0}\left[\bar{f}(X_i)(e^{V_l} - e^{V_{l-1}})1_{\{\rho(X_i)\geq l\}}1_{\{\tau \geq i\}}\right]$$

and

$$e^{V(x_0)}E_{x_0}\left[\int_D \bar{f}(y)\bar{\delta}_{\bar{X}_i^\tau}(dy)\right] = E_{x_0}\left[\bar{f}(X_i)e^{V(X_i)}1_{\{\tau\geq i\}}\right] = E_{x_0}\left[f(X_i)1_{\{\tau \geq i\}}\right].$$

The proof of this Lemma is given in the appendix. The proof of Theorem 3.2 is given below.

Proof of Theorem 3.2. Suppose that $\{X_i\}$, V, f, τ are given and that the initialising and splitting distributions are unbiased. Let $\hat{s}(\bar{f}, \tau)$ and $\hat{s}(\bar{f}, \tau \wedge n), n = 1, \ldots$ denote the output of the GDPR algorithms for estimating $E_{x_0}\left[\sum_{i=0}^{\tau} f(X_i)\right]$ and $E_{x_0}\left[\sum_{i=0}^{\tau\wedge n} f(X_i)\right], n = 1, \ldots$ respectively. Note that the GDPR algorithm for estimating $E_{x_0}\left[\sum_{i=0}^{\tau\wedge n} f(X_i)\right]$ is equivalent to running the GDPR algorithm for simulating $E_{x_0}\left[\sum_{i=0}^{\tau} f(X_i)\right]$ and terminating it at time n. Thus $\hat{s}(\bar{f}, \tau\wedge n) \to \hat{s}(\bar{f}, \tau)$ a.s. Using the Monotone Convergence Theorem (MCT)

$$E_{x_0}\left[\sum_{i=0}^{\tau} f(X_i)\right] = \lim_{n\to\infty} E_{x_0}\left[\sum_{i=0}^{\tau\wedge n} f(X_i)\right].$$

If follows from Tonelli's theorem and Lemma 3.3 that

$$\begin{aligned}
E_{x_0}\left[\sum_{i=0}^{\tau\wedge n} f(X_i)\right] &= \sum_{i=0}^{n} E_{x_0}\left[f(X_i)1_{\{\tau\geq i\}}\right] \\
&= \sum_{i=0}^{n} e^{V(x_0)}E_{x_0}\left[\int_D \bar{f}(x)\bar{\delta}_{\bar{X}_i^\tau}(dx)\right] \\
&= e^{V(x_0)}E_{x_0}\left[\sum_{i=0}^{n}\int_D \bar{f}(x)\bar{\delta}_{\bar{X}_i^\tau}(dx)\right] \\
&= E_{x_0}\left[\hat{s}(\bar{f}, \tau\wedge n)\right].
\end{aligned}$$

Thus it follows from the MCT that

$$\begin{aligned}
E_{x_0}\left[\hat{s}(\bar{f}, \tau)\right] &= \lim_{n\to\infty} E_{x_0}\left[\hat{s}(\bar{f}, \tau\wedge n)\right] \\
&= \lim_{n\to\infty} E_{x_0}\left[\sum_{i=0}^{\tau\wedge n} f(X_i)\right] \\
&= E_{x_0}\left[\sum_{i=0}^{\tau} f(X_i)\right].
\end{aligned}$$

■

In order for the GDPR algorithm to be useful in practice it is necessary to know that it will terminate in finite time a.s. This is guaranteed by the following lemma.

Lemma 3.4 *Assume Condition 3.1. Then the GDPR algorithm almost surely terminates in a finite time.*

Proof. Choose an arbitrary $i \geq 0$. Define $\bar{f}(x)$ by $e^{V(x_0)}\bar{f}(x) = 1$ for all x, and note that this implicitly requires $f(x) = e^{V(x)-V(x_0)}$. Then

$$E_{x_0}\left[N_i^\tau\right] = E_{x_0}\left[\sum_{j=1}^{N_i^\tau} 1\right] = E_{x_0}\left[\int_D e^{V(x_0)}\bar{f}(x)\bar{\delta}_{\bar{X}_i^\tau}(dx)\right].$$

It follows from Lemma 3.3 that

$$E_{x_0}\left[N_i^\tau\right] = E_{x_0}\left[e^{V(X_i)-V(X_0)}1_{\{\tau\geq i\}}\right]. \tag{3.4}$$

Therefore by Markov's inequality and the fact that $V_J = \max_{j=1,\ldots,J}\{V_j\}$,

$$\begin{aligned} P_{x_0}\left(N_i^\tau > 0\right) &\leq E_{x_0}\left[e^{V(X_i)-V(X_0)}1_{\{\tau\geq i\}}\right] \\ &\leq P_{x_0}\left(\tau \geq i\right)e^{V_J} \\ &\rightarrow 0 \end{aligned}$$

as $i \rightarrow \infty$. Note that $N_i^\tau = 0$ if and only if the simulation has terminated by time i, since $N_i^\tau = 0$ obviously implies $N_k^\tau = 0$ for all $k \geq i$. Since $\tau < \infty$ a.s., the result now follows from the fact that

$$P_x(\text{GDPR algorithm does not terminate}) \leq P(N_i^\tau > 0)$$

for all i. ■

4 Performance Measures

Missing from the discussion so far is any specific form for the splitting distribution $\mathcal{Q}_{j,k}$. We now give a very natural example which will be used throughout the rest of the paper, which is in fact the one proposed in [20, 21] for the original DPR algorithm. We make no claim of optimality for this particular choice. However, the resulting GDPR algorithm is simple and efficient to implement and also lends itself to analysis. To simplify the presentation, we describe random variables with the desired distribution.

Recall that the multinomial distribution $M(N, p_1, \ldots, p_d)$ on d-tuples $x_1, \ldots, x_d \in \mathbb{Z}_+^d : x_1 + \cdots + x_d = N$ is defined by

$$P((x_1,\ldots,x_d)=(n_1,\ldots,n_d)) = \binom{N}{n_1 \cdots n_d} p_1^{n_1} \cdots p_d^{n_d}.$$

Given a scalar a let $\{a\} = a - \lfloor a \rfloor$ denote its fractional part. For $j < k$ let A^1, A^2 and b be independent random vectors and random variables such that b equals 1 with probability $\{(e^{V_k} - e^{V_j})/e^{V_j}\}$ and 2 otherwise, and A^1 and A^2 are distributed according to

$$M\left(\left\lceil \frac{e^{V_k} - e^{V_j}}{e^{V_j}} \right\rceil, \frac{e^{V_{j+1}} - e^{V_j}}{e^{V_k} - e^{V_j}}, \ldots, \frac{e^{V_k} - e^{V_{k-1}}}{e^{V_k} - e^{V_j}}\right)$$

and

$$M\left(\left\lfloor \frac{e^{V_k} - e^{V_j}}{e^{V_j}} \right\rfloor, \frac{e^{V_{j+1}} - e^{V_j}}{e^{V_k} - e^{V_j}}, \ldots, \frac{e^{V_k} - e^{V_{k-1}}}{e^{V_k} - e^{V_j}}\right),$$

respectively. The splitting distributions $\mathcal{Q}_{j,k}$ we consider are those which are equal to the distributions of the random variables $(Q_0^{j,k}, Q_1^{j,k}, \ldots)$, where

$$Q_l^{j,k} = 0 \text{ if } l \notin \{j+1,\ldots,k\}$$

and

$$\left(Q_{j+1}^{j,k},\ldots,Q_k^{j,k}\right) = \begin{cases} A_{l-j}^1 & \text{if } b=1 \\ A_{l-j}^2 & \text{if } b=2 \end{cases}.$$

It is easy to check that for any $j < l \leq k$

$$E[Q_l^{j,k}] = (e^{V_l} - e^{V_{l-1}})/e^{V_j}$$

and $Q_l^{j,k} = 0$ otherwise, and so the resulting algorithm is unbiased [see Theorem 3.2].

Note that we have already assumed that the initializing distribution is $\mathcal{I} = \mathcal{L}_{\rho(x_0)}$. In actual numerical implementation it is always the case that $x_0 \in C_0$, which implies that all mass is on $l = 0$. The following condition will be used for the rest of the paper.

Condition 4.1 *Given an importance function V, the GDPR algorithm is implemented using the splitting distributions described in this section.*

The performance of the GDPR algorithm depends on two factors: the second moment (and hence variance) of the estimator and the computational cost of each simulation. To avoid discussion of the specific implementation of the algorithm the computational cost is defined to be

$$w = \sum_{i=0}^{\infty} N_i^{\tau},$$

i.e., the sum of the lifetimes of all the particles simulated. In this section formulae for both are derived in terms of only the importance function and the underlying process. Throughout it is assumed that $\{X_i\}, \tau, V$ and f are given. The first result is an expression for the cost of the GDPR algorithm which follows directly from (3.4) and the definition of w.

Theorem 4.2 *Assume Condition 3.1. Then*

$$E_{x_0}[w] = e^{-V(x_0)} E_{x_0}\left[\sum_{i=0}^{\tau} e^{V(X_i)}\right].$$

Note that the expression does not depend on the choice of (unbiased) splitting mechanism.

Next we give bounds for the second moment of the estimator. These bounds will be used later to extend the framework developed in [9] for splitting schemes to the design and analysis of GDPR schemes. The proofs are given in the appendix.

Theorem 4.3 *Assume Conditions 3.1 and 4.1. Then*

$$E_{x_0}[(\hat{s}(\bar{f}))^2] \le e^{V(x_0)} E_{x_0}\left[\sum_{i=1}^{\tau} e^{-V(X_{i-1})}\left(f(X_{i-1}) + E_{X_i}\left[\sum_{k=0}^{\tau} f(X_k)\right]\right)^2\right]. \quad (4.1)$$

Theorem 4.4 *Assume Conditions 3.1 and 4.1. Then*

$$E_{x_0}[(\hat{s}(\bar{f}))^2] \ge e^{V(x_0)} E_{x_0}\left[\sum_{i=0}^{\tau} e^{-V(X_i)} f(X_i)^2\right].$$

5 Design and Asymptotic Analysis of GDPR Schemes

Thus far we have only addressed the problem of estimating a single expected value of the form (1.1). Now we shall turn to the problem of estimating a sequence of such expected values

$$E_{x_n}\left[\sum_{i=0}^{\tau^n} f^n(X_i^n)\right], \, n = 1, 2, \ldots \quad (5.1)$$

for which a large deviations rule holds. We assume $x_n \to x$ as $n \to \infty$ with each $x_n \notin M$. The asymptotic performance of GDPR schemes will be evaluated using the following measure of work-normalised error:

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{E_{x_n}\left[(\hat{s}^n)^2\right] E_{x_n}\left[w^n\right]}{E_{x_n}\left[\sum_{i=0}^{\tau^n} f^n(X_i^n)\right]^2}.$$

Suppose that $-(1/n)\log E_{x_n}\left[\sum_{i=0}^{\tau^n} f^n(X_i^n)\right] \to \gamma$ as $n \to \infty$. A use of Jensen's inequality as in the Introduction shows that the best possible rate of decay for the work-normalised error is zero, and this occurs only when the work grows subexponentially and the second moment $E_{x_n}\left[(\hat{s}^n)^2\right]$ decays at rate 2γ. Bounds on the asymptotic behavior of the work-normalised error will be obtained via Theorems 4.2, 4.3 and 4.4.

Within the general framework of (5.1) there are two cases of particular interest. The first is that of hitting probabilities, such as the probability that X^n hits some rare set B before hitting a typical set A, after starting at $x_n \notin A \cup B$. In this case $M = A \cup B$ and $f^n(x) = 1_B(x)$. The second case occurs as one of two approaches discussed in Section 7.2 for approximating stationary measures, and uses $\tau^n = \lfloor Tn \rfloor + 1$ for some fixed $T \in (0, \infty)$ and a general cost f^n. The theory presented in this section will require some fairly standard assumptions on the stability and large deviations behavior of $\{X_i^n\}$, and also some regularity properties on M and f^n. For example, in the case of hitting probabilities we will want to know that τ^n/n can essentially be taken as bounded, in the sense that there is some $T < \infty$ such that the event $\tau^n/n > T$ is unimportant as far as the large deviation asymptotics are concerned. This is an important qualitative assumption, and is related to stability properties of the law of large number limit processes obtained when $n \to \infty$.

We next state some basic assumptions, including a large deviation property for the continuous time processes defined by $X^n(t) = X_i^n$ for $t \in [i/n, i/n + 1/n)$.

Condition 5.1 1. *There is a fixed closed set $M \subset D$ such that for all n, $\tau^n = \inf\{i : X_i^n \in M\}$.*

2. *There is a lower semicontinuous and bounded from below function $F(y)$ such that for all $y \in D$ and n*

$$f^n(y) = \exp\left(-nF(y)\right).$$

3. *For every $T \in (0, \infty)$ the sequence $\{X^n, n = 1, 2, \ldots\}$ satisfies a large deviation principle (LDP) on $\mathcal{D}([0, T] : D)$ with a rate function of the form*

$$\int_0^T L(\phi(s), \dot{\phi}(s))ds$$

if $\phi \in \mathcal{D}([0, T] : D)$ is absolutely continuous and ∞ otherwise. This LDP is uniform with respect to initial conditions in compact sets [10, 15].

As remarked above, the conditions we use beyond the LDP can be partitioned into "stability" and "regularity" type conditions. We next give two conditions which will be sufficient (but not necessary) for what follows. In particular, the condition we refer to as "controllability" can be weakened, but without this assumption it may not be obvious if a large deviation limit can be assumed for the sequence of expected values or probabilities of interest. Moreover, the sufficient conditions we give now will by themselves cover many interesting problems.

Condition 5.2 *For any compact $\kappa \subset D$*

$$\limsup_{T\to\infty}\limsup_{n\to\infty}\sup_{x\in\kappa} -\frac{1}{n}\log P_x\left\{\tau^n/n \ge T\right\} = -\infty$$

and for any $l \in \mathbb{Z}^+$

$$\limsup_{n\to\infty}\sup_{x\in\kappa}\frac{1}{n}\log E_x(\tau^n)^l < \infty.$$

Given the large deviations principle, conditions such as this will follow when all zero cost trajectories with initial conditions in a compact set K are forced to enter M° by some fixed finite time (which can depend on K). See the discussion in [15, Lemma 2.2, Chapter 4]. For verification in the context of stable stochastic networks, where M often includes the origin though M° does not and hence the constructions of [15, Lemma 2.2, Chapter 4] do not directly apply, see [12, Appendix A] and [14, Lemma A.4]. An example where Condition 5.2 would not hold is when there are two attractors for the zero cost trajectories, M contains one of the attractors but not the other, and the process starts in the domain of attraction of the stable point that is not in M.

Condition 5.3 *1. Let $\varepsilon > 0$. Given any compact set $K \subset D$, there is $\delta > 0$ such that if $x, y \in K \cap (M^\circ)^c$ satisfy $\|x - y\| \le \delta$, then there is $\sigma \le \varepsilon$ and a trajectory ϕ connecting x to y such that $\phi(r) \notin M$ for all $r \in (0, \sigma)$ and $\int_0^\sigma L(\phi(s), \dot\phi(s))ds \le \varepsilon$.*

2. Let $T < \infty$ and a bounded and continuous function H be given. Consider any sequence of times $i_n \le \tau^n \wedge \lfloor nT \rfloor$ such that $i_n/n \to t \le T$ and $x_n \notin M$ such that $x_n \to x$. Then

$$\limsup_{n\to\infty}\frac{1}{n}\log E_{x_n} e^{-nH(X^n_{i_n})} \le -\inf\left[\int_0^t L(\phi(r), \dot\phi(r))dr + H(\phi(t))\right], \qquad (5.2)$$

where the infimum is over all ϕ that satisfy $\phi(r) \notin M$ for $r \in (0, t)$.

One can consider part 1 as a "controllability" condition. A simple sufficient condition is that $L(x, \beta)$ be continuous, bounded on each compact subset of $\mathbb{R}^d \times \mathbb{R}^d$, and regularity of the boundary of M. The formulation of (5.2) is nonstandard, in that it assumes an upper bound in terms of the infimum over a set that is not necessarily closed. Under conditions such as continuity and boundedness of the local rate function (on compact sets), one can show that the infimum over the indicated trajectories and their closure is the same, and so (5.2) follows from the usual large deviation upper bound. The motivation for the formulation as given is specifically to cover stochastic networks. With stochastic networks one often has $0 \in M$, though not $0 \in M^\circ$. Owing to discontinuities of the local rate at the origin, the corresponding infimum over the closed set of trajectories, while valid, is not at all tight. Prelimit process trajectories which stay near but do not touch the origin are quite rare in the sense that (5.2) holds as stated. However, the deterministic trajectory which remains at the origin has zero rate, and hence the closure operation in the infimum produces a bound that is not tight and not useful. Again, we refer to [12] and [14] for further discussion on this point.

The controllability condition is not necessary. However, as remarked previously, without such a condition it is harder to establish when large deviations limits exist [as opposed to separate upper and lower bounds], and hence such limits are then often verified on a case-by-case basis.

We make the definition

$$\mathcal{J}(y,z) = \inf_{\phi:\phi(0)=y;\phi(T)=z;\phi(s)\notin M, s\in(0,T);T<\infty} \int_0^T L(\phi(s),\dot{\phi}(s))ds. \tag{5.3}$$

for $y, z \notin M^\circ$. Under Condition 5.3 $\mathcal{J}(y,z)$ is uniformly continuous on compacts. The definition of $\mathcal{J}(y,z)$ is extended to $D \times D$ by letting $\mathcal{J}(y,z) = \infty$ if y or $z \in M^\circ$ with $y \neq z$, and $\mathcal{J}(y,z) = 0$ if $y = z \in M^\circ$.

It is clear that defining a GDPR scheme for a sequence of expected values of the form (5.1) is equivalent to defining a sequence $\{V^n\}$ of importance functions. We propose that this be done using what we call *GDPR scheme generating functions*, which will always be abbreviated to *generating functions*. A generating function U is a continuous function on D that is bounded from below. Once a generating function has been chosen it can be used to define a sequence of importance functions $\{V^n\}$ in the following manner. First choose a fixed positive real number Δ. Then for each n define an importance function V^n by

$$V^n(y) = 0 \vee \tilde{V}^n(y).$$

where

$$\tilde{V}^n(y) = \Delta \left\lfloor \frac{nU(x_n) - nU(y)}{\Delta} \right\rfloor .$$

Consider the problem of hitting probabilities. In this case one expects the importance function to increase as y approaches the rare set B. A corresponding generating function will thus typically decrease as y approaches B. The construction guarantees that $V^n(y)$ achieves its minimum of zero at $y = x_n$, and since U is bounded from below V^n/n is bounded from above. If $x_n \to x$, then $V^n(y)/n \to (U(x) - U(y)) \vee 0$ as $n \to \infty$ uniformly in $y \in D$.

For the generating function approach to be practical one must be able to identify those functions which lead to schemes with good asymptotic performance. It will turn out that good generating functions are characterized by their relationship to the cost function $\mathcal{J}(y,z)$.

Definition 5.4 *A continuous function $G(y)$ is a subsolution to (5.3) if $G(y) - G(z) \leq \mathcal{J}(y,z)$ for all $y, z \in D$.*

The definition of subsolution used here is phrased purely in terms of the calculus of variations problem. This is slightly different from previous definitions, such as the PDE formulation of subsolutions used in [9]. The definition via calculus of variations is somewhat more to the point of what is required and used in the proofs. The well-known relations between the two are discussed in the next section. In the sequel we will show that if F is bounded and continuous, if U is a subsolution to (5.3) and if $\{V^n\}$ is the sequence of importance functions defined as above, then

$$\lim_{n\to\infty} -\frac{1}{n}\log E_{x_n}\left[(\hat{s}^n)^2(\bar{f}^n)\right] = \inf_{y\in D}\{\mathcal{J}(x,y) + (U(x) - U(y)) \vee 0 + 2F(y)\}$$

and

$$\lim_{n\to\infty}\frac{1}{n}\log E_{x_n}[w^n] = 0.$$

Thus the work associated with such a scheme grows subexponentially, and the performance is determined by the value $U(x)$. The best possible rate of decay is $2W(x)$, where $W(x)$ is defined by

$$\gamma = W(x) = \inf_{y\in D}\left[\mathcal{J}(x,y) + F(y)\right]. \tag{5.4}$$

If y is a minimizing point in (5.4), then achieving this best rate will require $U(x) - U(y) = \mathcal{J}(x,y)$, i.e., that $U(x)$ takes the maximum possible value at x. The question of finding suitable subsolutions will be addressed in more detail in the next section.

We will further show that if U is not a subsolution then there exists some $y \in D$ such that if $x_n \to y$ then

$$\liminf_{n\to\infty}\frac{1}{n}\log E_{x_n}[w^n] > 0.$$

It follows that generating functions which are not subsolutions should not be used to design GDPR schemes, as it is possible that the computational costs of such schemes will grow exponentially.

We will also prove an extension that allows one to relax the boundedness and continuity assumptions on F. The relaxation will be needed when considering the problem of estimating stationary measures and hitting probabilities

For convenience we recall the results of Theorems 4.3 and 4.4, rewritten using the large deviation scaling and incorporating the fact that $V^n(x_n) = 0$:

$$E_{x_n}\left[\sum_{i=1}^{\tau^n} e^{-V^n(X^n_{i-1})}\left(e^{-nF(X^n_{i-1})} + E_{X^n_i}\left[\sum_{j=0}^{\tau^n} e^{-nF(X^n_j)}\right]\right)^2\right] \geq E_{x_n}\left[\left(\hat{s}^n(\bar{f}^n)\right)^2\right] \tag{5.5}$$

and

$$E_{x_n}\left[\left(\hat{s}^n(\bar{f}^n)\right)^2\right] \geq E_{x_n}\left[\sum_{i=0}^{\tau^n} e^{-V^n(X^n_i)} e^{-n2F(X^n_i)}\right]. \tag{5.6}$$

In the following theorem we distinguish between cases when $x \notin M$ and $x \in \partial M$. The reason is because lower bounds on probabilities involving the hitting time τ^n can depend on the detailed properties of the underlying process (and not just on its large deviation rate function) when $x \in \partial M$, and so we only state the upper bound in that case. However, if large deviation limits are available for the expected values or probabilities that are being estimated via GDPR, then the same argument used establish coincidence of the upper and lower large deviation bounds could be used here as well to show that the limit (and not just limit inferior) holds when $x \in \partial M$. Analogous comments also apply for Theorem 5.6, where we relax the continuity and boundedness of F.

Theorem 5.5 *Assume Conditions 3.1, 4.1, 5.1, 5.2 and 5.3. Suppose that $x_n \to x \notin M$, that F is bounded and continuous, and that U is a subsolution. Then*

$$\lim_{n\to\infty} -\frac{1}{n}\log E_{x_n}\left[\left(\hat{s}^n(\bar{f}^n)\right)^2\right] = \inf_{y\in D}\left[\mathcal{J}(x,y) + (U(x)-U(y))\vee 0 + 2F(y)\right]. \tag{5.7}$$

If $x_n \to x \in M$ with each $x_n \notin M$, then the corresponding limit inferior holds.

Proof. Let $R \doteq \inf_{y\in D}\left[\mathcal{J}(x,y) + (U(x)-U(y))\vee 0 + 2F(y)\right]$, and choose y that is within $\varepsilon > 0$ of the infimum in the definition of R. Let ϕ and T be within ε of the infimum in the definition of $\mathcal{J}(x,y)$. Given $\eta > 0$, since $x \notin M$ there is $\delta > 0$ such that $\phi(s)$ is at least distance δ from M for $s \in [0, T-\eta]$. Hence if a trajectory of X^n stays in the open ball of radius δ about ϕ then $\tau^n \geq \lfloor n(T-\eta)\rfloor$. Since $V^n(y)/n \to (U(x)-U(y))\vee 0$ uniformly, by (5.6)

$$\begin{aligned}
&\liminf_{n\to\infty} \frac{1}{n}\log E_{x_n}\left[\left(\hat{s}^n(\bar{f}^n)\right)^2\right] \\
&\quad\geq \liminf_{n\to\infty} \frac{1}{n}\log E_{x_n}\left[\sum_{i=0}^{\tau^n} e^{-V^n(X^n_i)} e^{-n2F(X^n_i)}\right] \qquad (5.8)\\
&\quad\geq \liminf_{n\to\infty} \frac{1}{n}\log E_{x_n}\left[e^{-V^n(X^n_{\lfloor n(T-\eta)\rfloor})} e^{-n2F(X^n_{\lfloor n(T-\eta)\rfloor})}\right] \\
&\quad\geq -\int_0^{T-\eta} L(\phi(s),\dot{\phi}(s))ds - (U(x)-U(\phi(T-\eta)))\vee 0 - 2F(\phi(T-\eta)).
\end{aligned}$$

Letting $\eta \downarrow 0$ gives

$$\liminf_{n\to\infty} \frac{1}{n} \log E_{x_n} \left[\left(\hat{s}^n(\bar{f}^n) \right)^2 \right] \geq -R - 2\varepsilon,$$

and since $\varepsilon > 0$ is arbitrary this proves the lower bound.

We now turn to the upper bound, which is based on (5.5). Using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, the left hand side of (5.5) is less than or equal to

$$\begin{aligned} &2E_{x_n} \left[\sum_{i=1}^{\tau^n} e^{-V^n(X_{i-1}^n)} e^{-2nF(X_{i-1}^n)} \right] \\ &+2E_{x_n} \left[\sum_{i=1}^{\tau^n} e^{-V^n(X_{i-1}^n)} \left[\sum_{j=i}^{\tau^{1,i,n}} e^{-nF(X_j^{1,i,n})} \right] \left[\sum_{j=i}^{\tau^{2,i,n}} e^{-nF(X_j^{2,i,n})} \right] \right], \end{aligned} \tag{5.9}$$

where $X_j^{k,i,n}$ are (conditionally) independent copies of X_j^n that start at X_i^n at $j = i$. It follows from Conditions 5.2 and 5.3 and the fact that $V^n(y)/n \to (U(x) - U(y)) \vee 0$ that the first term in (5.9) obeys the necessary large deviations upper bound. Later in the proof we will show that the large deviation asymptotics of the second quantity in (5.9) are the same as those of

$$E_{x_n} \left[\sum_{i=0}^{\tau^n \wedge \lfloor nT \rfloor} e^{-V^n(X_i^n)} \left[\sum_{j=i}^{\tau^{1,i,n} \wedge \lfloor nT \rfloor} e^{-nF(X_j^{1,i,n})} \right] \left[\sum_{j=i}^{\tau^{2,i,n} \wedge \lfloor nT \rfloor} e^{-nF(X_j^{2,i,n})} \right] \right] \tag{5.10}$$

for some fixed and finite T. Assuming this claim, observe that there are no more than order n^3 terms in the expected value, and it suffices to obtain the desired upper bound on each of these terms. Accordingly, consider $i_n \leq \tau^n \wedge \lfloor nT \rfloor$, $j_n^k \leq \tau^{k,i,n} \wedge \lfloor nT \rfloor$ and assume that $i_n/n \to t$, $j_n^k/n \to s^k, k = 1, 2$, with $t \leq s^k \leq T$. Owing to Condition 5.3, the convergence of V^n, and the fact that the infimum over the two trajectories on the intervals $[t, s^k]$ will obviously be the same,

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n} \log E_{x_n} \left[e^{-V^n(X_{i_n}^n)} e^{-nF(X_{j_n^1}^{1,i,n})} e^{-nF(X_{j_n^2}^{2,i,n})} \right] \leq - \inf \Bigg[\int_0^t L(\phi(r), \dot{\phi}(r)) dr \\ +(U(x) - U(\phi(t))) \vee 0 + 2 \int_t^s L(\phi(r), \dot{\phi}(r)) dr + F(\phi(s)) \Bigg], \end{aligned}$$

where the infimum is over ϕ with the property that $\phi(r) \notin M^\circ$ for $r \in (0, t) \cup (t, s)$. By the definition of $\mathcal{J}(x, y)$ the infimum is bounded below by

$$\inf_{y\in D, z\in D} \left[\mathcal{J}(x, y) + (U(x) - U(y)) \vee 0 + 2\mathcal{J}(y, z) + 2F(z) \right].$$

Using the subsolution property of $U(y)$ and that for any $x, y, z \in D$, $\mathcal{J}(x, y) \leq \mathcal{J}(x, z) + \mathcal{J}(z, y)$, the last infimum is bounded below by

$$\inf_{y\in D} \left[\mathcal{J}(x, z) + (U(x) - U(z)) \vee 0 + 2F(z) \right],$$

and the upper bound follows.

It remains to show that the second term in (5.9) has the same large deviation asymptotics as (5.10). We first prove that

$$\limsup_{T\to\infty}\limsup_{n\to\infty}\frac{1}{n}\log E_{x_n}\left[1_{\{\tau^n/n\geq T\}}\sum_{i=1}^{\tau^n} e^{-V^n(X^n_{i-1})}\left[\sum_{j=i}^{\tau^{1,i,n}} e^{-nF(X_j^{1,i,n})}\right]\left[\sum_{j=i}^{\tau^{2,i,n}} e^{-nF(X_j^{2,i,n})}\right]\right]\leq -\infty. \tag{5.11}$$

For fixed $T<\infty$ Hölder's inequality implies that for any $p>1$ and $q>1$ with $1/p+1/q=1$, the expected value is bounded above by

$$\begin{aligned} & e^{n2\|F\|_\infty}E_{x_n}\left[1_{\{\tau^n/n\geq T\}}\sum_{i=1}^{\tau^n}\tau^{1,i,n}\tau^{2,i,n}\right] \\ \leq\ & e^{n2\|F\|_\infty}E_{x_n}\left[1_{\{\tau^n/n\geq T\}}\left(\tau^n\right)^3\right] \\ \leq\ & e^{n2\|F\|_\infty}\left(E_{x_n}\left[1_{\{\tau^n/n\geq T\}}\right]\right)^{1/p}\left(E_{x_n}\left[\left(\tau^n\right)^{3q}\right]\right)^{1/q}. \end{aligned}$$

It follows from Condition 5.2 that in order to prove (5.11) it suffices to show that for $q>1$

$$\limsup_{n\to\infty}\frac{1}{n}\log E_{x_n}\left[\left(\tau^n\right)^{3q}\right]<\infty.$$

but this also follows from Condition 5.2.

To justify bounding the other random times by $\lfloor nT\rfloor$ we show

$$\limsup_{T\to\infty}\limsup_{n\to\infty}\frac{1}{n}\log E_{x_n}\left[\sum_{i=1}^{\tau^n\wedge\lfloor nT\rfloor} e^{-V^n(X^n_{i-1})}\left(1_{\{\tau^{1,i,n}/n\geq T\}}+1_{\{\tau^{2,i,n}/n\geq T\}}\right)\left[\sum_{j=i}^{\tau^{1,i,n}} e^{-nF(X_j^{1,i,n})}\right]\left[\sum_{j=i}^{\tau^{2,i,n}} e^{-nF(X_j^{2,i,n})}\right]\right]\leq -\infty. \tag{5.12}$$

In this case the expected value is bounded above by

$$2e^{n2\|F\|_\infty}\sum_{i=1}^{\lfloor nT\rfloor}E_{x_n}\left[1_{\{\tau^n/n\geq i\}}1_{\{\tau^{1,i,n}/n\geq T\}}\tau^{1,i,n}\tau^{2,i,n}\right].$$

A similar use of Hölder's inequality shows that for any $p>1$ and $q>1$ with $1/p+1/q=1$ the expected value is bounded above by

$$2e^{n2\|F\|_\infty}\sum_{i=1}^{\lfloor nT\rfloor}\left(E_{x_n}\left[1_{\{\tau^n/n\geq T\}}\right]\right)^{1/p}\left(E_{x_n}\left[\left(\tau^n\right)^{2q}\right]\right)^{1/q}.$$

Thus (5.12) follows from Jensen's inequality and Condition 5.2.

We complete the proof that the quantities (5.9) and (5.10) have the same large deviation asymptotics by showing that one can replace $V^n(X_i^n)$ in (5.10) by $V^n(X_{i-1}^n)$. However, it follows from the fact that rate functions have compact level sets that given any $M < \infty$ and $a > 0$, there exists a compact set K and $N < \infty$ such that if $n \geq N$, then the event $X_i^n \notin K$ or $\left\| X_{i-1}^n - X_{i-1}^n \right\| \geq a$ for any $i \leq \lfloor nT \rfloor$ has probability at most e^{-nM}. This justifies the replacement, and completes the proof. ■

The following theorem provides a partial relaxation of the continuity and boundedness conditions imposed on F.

Theorem 5.6 *Assume Conditions 3.1, 4.1, 5.1, 5.2 and 5.3. Suppose that $x_n \to x$ with each $x_n \notin M$, that $F = -\log 1_{\{x \in G\}}$ for some $G \subset D$, and that U is a subsolution. Then*

1. *If G is a closed subset of D*

$$\liminf_{n\to\infty} -\frac{1}{n} \log E_{x_n}\left[\left(\hat{s}^n(\bar{f}^n)\right)^2\right] \geq \inf_{y \in D}\left\{\mathcal{J}(x,y) + (U(x) - U(y)) \vee 0 + 2F(y)\right\}.$$

2. *If G is a open subset of D and $x \notin M$*

$$\limsup_{n\to\infty} -\frac{1}{n} \log E_{x_n}\left[\left(\hat{s}^n(\bar{f}^n)\right)^2\right] \leq \inf_{y \in D}\left\{\mathcal{J}(x,y) + (U(x) - U(y)) \vee 0 + 2F(y)\right\}.$$

3. *If $\overline{G} = \overline{G^\circ}$ and $x \notin M$*

$$\lim_{n\to\infty} -\frac{1}{n} \log E_{x_n}\left[\left(\hat{s}^n(\bar{f}^n)\right)^2\right] = \inf_{y \in D}\left\{\mathcal{J}(x,y) + (U(x) - U(y)) \vee 0 + 2F(y)\right\}.$$

Proof. Parts 1 and 2 can be shown to follow from Theorem 5.5 using exactly the same method as in the proof of Theorem 1.2.3 in [10] (note that compactness of level sets is part of the definition of an LDP). Part 3 then follows using Condition 5.3. ■

Theorem 5.7 *Assume Conditions 3.1, 4.1, 5.1, 5.2 and 5.3. If the generating function $U(x)$ is a subsolution then*

$$\lim_{n\to\infty} \frac{1}{n} \log E_{x_n}[w^n] = 0.$$

Proof. We know from Theorem 4.2 and the fact that $V^n(x_n) = 0$ for all n that

$$E_{x_n}[w^n] = E_{x_n}\left[\sum_{i=0}^{\tau^n} e^{V^n(X_i^n)}\right].$$

Since V^n/n is bounded, it follows as in the proof of the upper bound in Theorem 5.5 that the large deviation asymptotics of $E_{x_n}[w^n]$ are the same as those of

$$E_{x_n}\left[\sum_{i=0}^{\tau^n \wedge \lfloor nT \rfloor} e^{V^n(X_i^n)}\right]$$

for some sufficiently large but finite T. The convergence $V^n(y)/n \to (U(x)-U(y)) \vee 0$ and the same line of argument as in Theorem 5.5 shows

$$\limsup_{n\to\infty} \frac{1}{n} \log E_{x_n}\left[\sum_{i=0}^{\tau^n \wedge \lfloor nT \rfloor} e^{V^n(X_i^n)}\right] \leq - \inf_{y \in D} \left[\mathcal{J}(x,y) - (U(x)-U(y)) \vee 0\right].$$

By the subsolution property $U(x)-U(y) \leq \mathcal{J}(x,y)$, and so the upper bound follows. Since $E_{x_n}[w^n] \geq 1$ for all n the lower bound is automatic, which completes the proof. ■

Theorem 5.8 *Assume Conditions 3.1, 4.1 and 5.3. If the generating function $U(x)$ is not a subsolution then there exists some initial condition y such that if $x_n \to y$ then*

$$\liminf_{n\to\infty} \frac{1}{n} \log E_{x_n}[w^n] > 0.$$

In particular note that if U is a subsolution then the work associated with the scheme grows subexponentially while generating functions which are not subsolutions should not be used to design GDPR schemes, as it is possible that the computational costs of such schemes will grow exponentially.

Proof of Theorem 5.8. Suppose that there exist $y, z \notin M$ such that

$$U(y) - U(z) > \mathcal{J}(y,z).$$

It follows from the definition of $\mathcal{J}(y,z)$ that $y \notin M^\circ$ and $z \notin M^\circ$. Since U is continuous and we assume that $\mathcal{J}(y,z)$ is continuous we can restrict our analysis to the domain $M^c \times M^c$. It follows that there exist $\delta, \epsilon > 0$ such that $U(y) - U(\tilde{z}) > \mathcal{J}(y,z) + \epsilon$ for all $\tilde{z}$ such that $|\tilde{z} - z| < \delta$. Choose a sequence of initial conditions y_n such that $y_n \to y$. Then for all n

$$E_{y_n}\left[\sum_{i=0}^{\tau^n} e^{V^n(X_i^n)}\right] \geq e^{n(\mathcal{J}(y,z)+\epsilon)-2\Delta} P_{y_n}(|X_i^n - z| < \delta \text{ for some } 0 \leq i \leq \tau^n)$$

¿From Condition 5.3 we know that $\lim \frac{1}{n} \log P_{y_n}(|X_i^n - z| < \delta$ for some $0 \leq i \leq \tau^n) \geq -\mathcal{J}(y,z)$ and so the result follows. ■

6 Constructing Generating Functions

As discussed in the last section, one can construct good GDPR schemes by finding suitable generating functions $U(y)$. Recall that the asymptotic performance of a scheme based on a subsolution is given by

$$\inf_{y\in D}\left\{\mathcal{J}(x,y)+(U(x)-U(y))\vee 0+2F(y)\right\},$$

for which we have the a priori bound $2W(x) = 2\inf_{y\in D}\{\mathcal{J}(x,y)+F(y)\}$. Observe that if $y^*\in D$ minimizes in this definition then we can restrict attention to subsolutions with the property $U(x)\geq U(y^*)$, since otherwise the identically zero subsolution would do better. Observe also that U is subsolution in the sense of Definition 5.4 if and only if the same is true for $U+a$ for any constant a. Since both F and U are bounded and continuous we can normalize by requiring $U(y)\leq F(y)$ for all y, in which case the asymptotic performance is bounded below

$$\inf_{y\in D}\{\mathcal{J}(x,y)+U(x)+F(y)\}=W(x)+U(x),$$

with the best possible performance given if $U(x)$ achieves the maximum possible value of $W(x)$. This interpretation is closer to that used in a previous paper [9], where only lower bounds of this form were given on performance, and conditions such as $U(y)\leq F(y)$ became, in the hitting probabilities context used in [9], boundary conditions of the form $U(y)\leq 0, y\in\partial B$.

An obvious candidate for $U(y)$ is of course the solution $W(y)$ to the calculus of variations problem (5.4). Clearly $W(y)$ is bounded from below and continuous. Moreover (5.4) implies that $W(y)-W(z)\leq\mathcal{J}(y,z)$ for all $y,z\in D$. Thus $W(y)$ is a generating function which is a subsolution in the sense of Definition 5.4. Since it obviously achieves the maximum value, a GDPR scheme derived from $W(y)$ is asymptotically optimal.

While this property makes $W(y)$ a potentially good choice for generating function, in practice it is often unavailable. However, in many cases it is possible to obtain simple alternatives which are also asymptotically optimal. One approach is to use the theory of viscosity solutions for nonlinear PDE. The general form of the PDE depends on the particular structure of the problem. For example, for hitting probabilities one has a nonlinear PDE together with Dirichlet boundary conditions, while for the general continuous exponential $\exp -nF$ the nonlinear PDE is replaced by a quasivariational inequality. The main point is that, under broad conditions, subsolutions in the sense of Definition 5.4 (suitably normalized as discussed above) define subsolutions in the viscosity sense [7] for the appropriate PDE, and conversely, a subsolution in the viscosity sense is also a subsolution in the sense of Definition 5.4.

It turns out that the PDE characterization is more convenient for the explicit construction of subsolutions than the one based on the calculus of variations problem (see the many examples given in [13]), and examples will be given in the following section on numerical examples. Consider, for example, the construction of subsolutions for the problem of hitting probabilities. Under regularity conditions, the corresponding PDE is

$$\mathbb{H}(y, D\bar{W}(y)) = 0, y \in D\backslash M, \quad \bar{W}(y) = 0, y \in \partial B, \quad \bar{W}(y) = \infty, y \in \partial A, \qquad (6.1)$$

where for $y \in D$, $q \in \mathbb{R}^d$

$$\mathbb{H}(y, q) = \inf_{\beta \in \mathbb{R}^d} \left[\langle q, \beta \rangle + L(y, \beta) \right],$$

and L is the Lagrangian function in (5.3). It is sometimes possible to find simple functions which satisfy the subsolution property in $D\backslash M$. The concavity of $q \to \mathbb{H}(y, q)$ for each y implies that the pointwise minimum of a finite collection of subsolutions also satisfies the subsolution property in $D\backslash M$. Hence, one can try to build a function which also satisfies the boundary condition in M as such a minimum. Also, the pointwise maximum of such functions satisfies the subsolution property in $D\backslash M$. It should be noted that the constructions in [13] ultimately produce piecewise smooth *classical* subsolutions, which correspond to a stronger notion of subsolution than of the viscosity subsolutions required by the GDPR algorithm.

An alternative approach is possible for the problem of simulating hitting probabilities when the Freidlin-Wentzell quasipotential [15] exists. Consider an initial condition x, and suppose that all zero cost trajectories for the large deviation rate function are attracted to x. Then the quasipotential $U^{\mathrm{QP}}(y)$ is defined by

$$U^{\mathrm{QP}}(y) = \inf_{\phi:\phi(0)=x;\phi(T)=y;T<\infty} \int_0^T L(\phi(s), \dot{\phi}(s))ds.$$

which is the same as $\mathcal{J}(x, y)$ except trajectories need not avoid M. If the quasipotential $U^{\mathrm{QP}}(y)$ is well defined then the function

$$U^*(y) \doteq -U^{\mathrm{QP}}(y)$$

is a subsolution in the sense of Definition 5.4 and fulfills all the requirements of an asymptotically optimal generating function for the initial condition x, except that it is not necessarily bounded below. Observe that the infimum in $\inf_{y \in B} U^{\mathrm{QP}}(y)$ must correspond to a trajectory that leaves A immediately and touches B for the first time at T, and hence is also a candidate trajectory in the definition of $\mathcal{J}(x, y)$. Using $(U^*(x) - U^*(y)) \vee 0 = U^{\mathrm{QP}}(y)$, we have $\inf_{y \in D} \left\{ \mathcal{J}(x, y) + 2F(z) + U^{\mathrm{QP}}(y) \right\} = 2 \inf_{y \in B} \mathcal{J}(x, y)$, and so at least formally $U^*(y)$ yields an asymptotically optimal scheme.

Two issues that prohibit the direct application of this result are that $U^*(y)$ as defined is not bounded from below, and also that for the stability conditions to imply bounds such as Condition 5.2 one usually requires something like $x \in M^\circ$. The latter issue is not actually a problem in many applications to queueing systems and stochastic networks. In these cases x is the origin, and conditions such as Condition 5.2 hold without x being interior to M [13]. Alternatively one may consider an attracting point $x^* \in M^\circ$ and initial conditions x that are close to x^*, in which case the quasipotential defines a nearly asymptotically optimal scheme.

With regard to the first issue, U^* cannot be used directly since the boundedness from below is required in the proofs of Theorems 5.5 and 5.7 to ensure that the number of splitting thresholds is bounded for each n. There is, however, a remedy, and one can construct an asymptotically optimal generating function simply by replacing $U^{\mathrm{QP}}(y)$ by $U^{\mathrm{QP}}(y) \wedge \alpha$ for sufficiently large α. Consider for example the problem of hitting probabilities. Given the initial condition x and a corresponding minimizing y^* in the definition of $W(x)$, one can then choose any α larger than $U^{\mathrm{QP}}(y^*)$. If α is chosen smaller than this value then the resulting schemes will be asymptotically suboptimal. If chosen larger then there is (relatively) little harm. The scheme will still be asymptotically optimal (and perhaps might even have a smaller variance for a given finite value of n), but the expected computational cost per sample run will be somewhat larger (see Theorem 4.2). A concrete example of how the quasipotential can be used to construct GDPR schemes in practice is given in the next section.

7 Numerical Examples

In this section we present numerical results. We study three problems: hitting probabilities for queueing networks, stationary measures for queueing networks and a problem concerning rare events of the sample mean of a sequence of iid random variables. For each case we present an estimate based on a stated number of runs, standard errors and (formal) confidence intervals for each estimate based on an empirical estimate of the variance and the total CPU time required to calculate each estimate.

It is instructive to compare the results for estimating hitting probabilities given in this section with those obtained for the splitting algorithm in [9]. In both cases the problems considered and the importance functions used are identical, further the simulations were all run on the same computer and so the total CPU time can be used as a fair measure of the computational cost of the algorithms. As can be seen the standard deviations of the GDPR and splitting algorithms are almost identical while in all cases the computational cost of the GDPR algorithm is much less than that of the splitting algorithm. For example, the first set of numerical experiments discussed below consider total population and individual buffer overflow problems for a tandem Jackson network, each for several values of n. While the standard errors are essentially the same, the computational time for the GDPR scheme ranges from

14% to 4% of the corresponding time required for ordinary splitting. Indeed the cost of the splitting algorithm relative to that of the GDPR algorithm can be seen to increase with the large deviations parameter n, and the numerical experiments suggest that the ratio of the cost of the splitting algorithm to that of the GDPR algorithm grows without bound as $n \to \infty$. Further the results in [8] show that the GDPR algorithm shows a similar advantage over the splitting algorithm when applied to the other queueing problems discussed in this section. Thus in practice the evidence indicates that the GDPR algorithm provides a large improvement in performance relative to that of the splitting algorithm in a wide range of contexts.

7.1 Hitting Probabilities

We study three problems: buffer overflow for a tandem Jackson network with one shared buffer, simultaneous buffer overflow for a tandem Jackson network with separate buffers for each queue, and buffer overflow for a simple Markov modulated queue.

We start with the problem of estimating the probability of a buffer overflow event for a simple tandem Jackson network. Let $Q_1(t), Q_2(t)$ denote the state at time t and assume the stability condition $\lambda < \min\{\mu_1, \mu_2\}$ (see [16] for a discussion of such processes). Suppose that the two queues share a single buffer and that we are interested in

$$p^n = P_{(1,0)}\left(Q_1(\tau^n) + Q_2(\tau^n) \geq n\right),$$

where $\tau^n = \inf\{t > 0 : Q_1(t) + Q_2(t) \in 0 \cup [n, \infty)\}$. It is well known that

$$\lim_{n\to\infty} -\frac{1}{n}\log p^n = \rho_1 \wedge \rho_2,$$

where $\rho_i = \log \frac{\mu_i}{\lambda}$. Without loss of generality one can assume that $\mu_2 \leq \mu_1$ (see [12]) and hence $\rho_1 \wedge \rho_2 = \rho_2$. Let

$$\left(X_1^n(i), X_2^n(i)\right) = \frac{1}{n}\left(Q_1(t_i), Q_2(t_i)\right), \tag{7.1}$$

where t_i is the time of the i^{th} change in state of $(Q_1(t), Q_2(t))$. Then the probabilities p^n can be written as $p^n = p^n((1,0))$, where

$$p^n((a,b)) = P_{(a,b)}\left\{X_1^n + X_2^n \text{ reaches 1 before returning to } (0,0)\right\}.$$

We will construct an asymptotically optimal scheme by in terms of a viscosity subsolution. Define the Hamiltonian $\mathbb{H}$ by

$$\mathbb{H}(p) = -2\log[\lambda e^{-p_1} + \mu_1 e^{(p_1-p_2)} + \mu_2 e^{p_2}]. \tag{7.2}$$

A smooth function $\bar{V}$ will be a subsolution to the relevant PDE if the following inequalities hold [7]:

$$\begin{array}{rl}
\mathbb{H}(D\bar{V}(y)) \geq 0, & y_1 + y_2 \leq 1, y_1 > 0, y_2 > 0 \\
\mathbb{H}(D\bar{V}(y) - a(0,1)) \vee \left\langle (0,1), -D\bar{V}(y) + a(0,1) \right\rangle \geq 0, & y_1 = 0, y_2 \in (0,1), a \geq 0 \\
\mathbb{H}(D\bar{V}(y) - a(1,0)) \vee \left\langle (1,-1), -D\bar{V}(y) + a(1,0) \right\rangle \geq 0, & y_1 \in (0,1), y_2 = 0, a \geq 0 \\
\bar{V}(y) \leq 0, & y_1 + y_2 = 1, y_1 \geq 0, y_2 \geq 0.
\end{array}$$

The second and third inequalities are the viscosity formulation of the appropriate Neumann boundary conditions [11], with for example $D\bar{V}(y) - a(0,1)$ describing the general form of all sub-differentials on the boundary $y_1 = 0$ and with respect to the domain $y_1 + y_2 \leq 1, y_1 \geq 0, y_2 \geq 0$. The last inequality is a Dirichlet boundary condition, and the Dirichlet boundary condition at the origin $[\bar{V}(0) \leq \infty]$ is omitted since it holds automatically. If a smooth function satisfies these properties, then by using the form of the large deviation rate function and a verification argument standard arguments as in the references show that it is a subsolution in the sense of Definition 5.4.

This problem is simple enough that one can find a subsolution by inspection. Indeed, with the choice $\bar{V}(y) = \rho_2(1-y_1-y_2)\vee 0$ we have $\bar{V}(0) = W(0) = \rho_2$, and direct calculation gives $D\bar{V}(y) = -\rho_2(1,1)$ and $\mathbb{H}(D\bar{V}(y)) = 0$. The Dirichlet boundary condition holds with equality. With regard to the Neumann boundary conditions the concavity of $\mathbb{H}$ means that $\mathbb{H}(D\bar{V}(y) - a(0,1)) \geq 0$ will not be true when $a > 0$, but since $\left\langle (0,1), -D\bar{V}(y) \right\rangle, \langle (0,1), (0,1) \rangle, \left\langle (1,-1), -D\bar{V}(y) \right\rangle$ and $\langle (1,-1), (1,0) \rangle$ are all non-negative the boundary conditions hold. Thus it follows from the discussion in Section 6 that a GDPR scheme derived from $\bar{V}$ above will be asymptotically optimal.

Table 1 below shows the results of estimating the probabilities p^n when $\lambda = 1$, $\mu_1 = \mu_2 = 4.5$ for various values of n using a GDPR scheme derived from the above generating function using $\Delta = \ln 4.5$. This value of Δ was chosen to coincide with the large deviation rate since then each level of the form $\{(q_1, q_2) : q_1 + q_2 = j\}$ corresponds to a splitting threshold. This choice gives good results although in performance is not too sensitive in this regard, and changing Δ by a factor of two has a qualitatively small effect. Each estimate is based on a run of 20,000 samples. The variances of the estimators are estimated in the standard manner, see [22]. The theoretical value was obtained by solving the matrix equations obtained by doing a one time step analysis.

Consider now the same tandem Jackson network but suppose that we are interested in estimating the probabilities

$$p^n((1,0)) = P_{(1,0)}\left(Q_1(\tau^n) \wedge Q_2(\tau^n) \geq n\right),$$

where $\tau^n \doteq \inf\{t : Q_1(t) \vee Q_2(t) = 0 \text{ or } Q_1(t) \wedge Q_2(t) \geq n\}$. For this example we will use the quasipotential to define the subsolution. It is shown in [18] that

n	30	40	50
Theoretical Value	2.63×10^{-18}	1.03×10^{-24}	3.80×10^{-31}
Estimate	2.63×10^{-18}	1.06×10^{-24}	3.83×10^{-31}
Std. Err.	0.08×10^{-18}	0.04×10^{-24}	0.15×10^{-31}
95% C.I.	$[2.47, 2.79] \times 10^{-18}$	$[0.99, 1.14] \times 10^{-24}$	$[3.54, 4.13] \times 10^{-31}$
Time Taken (s)	3	6	8

Table 1: Hitting Probabilities, Single Shared Buffer

$$\lim_{n\to\infty} -\frac{1}{n}\log p^n((1,0)) = \log\rho_1 + \log\rho_2 \doteq \gamma.$$

To investigate the large deviations properties of these probabilities we rescale the processes as in the last example.

It is well known (see [16]) that the queueing process $(Q_1(t), Q_2(t))$ has invariant measure $\mu(q_1,q_2) = \rho_1^{-q_1}\rho_2^{-q_2}$, and hence the rescaled process $(X_1^n(i), X_2^n(i))$ has invariant measure $\mu^n(x_1,x_2) = \rho_1^{-nx_1}\rho_2^{-nx_2}$ for points (x_1,x_2) of the form $(q_1,q_2)/n$. Using the relation between the quasipotential and the large deviation properties of the stationary distribution, it follows that the quasipotential of $(X_1^n(i), X_2^n(i))$ is

$$U^{\mathrm{QP}}(x_1,x_2) = x_1\log\rho_1 + x_2\log\rho_2.$$

Deriving a generating function from this quasipotential in the manner discussed in Section 6 gives the function

$$U(x_1,x_2) = ([1-x_1]\log\rho_1 + [1-x_2]\log\rho_2) \vee 0. \tag{7.3}$$

Since $U(0,0) = \gamma$ it follows that a GDPR scheme derived in this manner is asymptotically optimal.

Table 2 shows the results of estimating the probabilities p^n when $\lambda = 1, \mu_1 = 3$ and $\mu_2 = 2$ for various values of n using a GDPR scheme derived using (7.3) with $\Delta = \ln 6$. Each estimate is based on a run of 20,000 samples.

n	10	20	30
Theoretical Value	9.64×10^{-8}	1.60×10^{-15}	2.64×10^{-23}
Estimate	9.70×10^{-8}	1.57×10^{-15}	2.64×10^{-23}
Std. Err.	0.16×10^{-8}	0.03×10^{-15}	0.06×10^{-23}
95% C.I.	$[9.39, 10.0] \times 10^{-8}$	$[1.51, 1.63] \times 10^{-15}$	$[2.53, 2.75] \times 10^{-23}$
Time Taken (s)	3	12	26

Table 2: Hitting Probabilities, Separate Buffers

Finally, since many queueing models are non-Markovian we present an example involving a non-Markovian process. Consider a tandem network whose arrival rates are modulated by an underlying process $R(t)$ which takes values in the set $\{1, 2\}$, such that the times taken for the modulating process to switch states are independent exponential random variables with rate $\gamma(1)$ if R is in state 1 and $\gamma(2)$ otherwise. Let $\lambda(1), \mu_1(1), \mu_2(1)$ and $\lambda(2), \mu_1(2), \mu_2(2)$ be the arrival and service rates of the network in the first and second states respectively. The notion of viscosity subsolution is the same as that of the tandem Jackson network except that the Hamiltonian is given by

$$\mathbb{H}(p) = -\log C^*(p),$$

$C^*(p)$ is the largest eigenvalue of a matrix $A(p)$ defined by

$$A(p)_{j,k} = \begin{cases} \mathbb{H}_j(p)\frac{\lambda(j)+\mu_1(j)+\mu_2(j)}{\lambda(j)+\mu_1(j)+\mu_2(j)+\gamma(i)} & j = k \\ \frac{\gamma(j)}{\lambda(j)+\mu_1(j)+\mu_2(j)+\gamma(j)} & \text{otherwise} \end{cases},$$

and $\mathbb{H}_j(p)$ is equal to (7.2) with service rates corresponding to those of the Markov modulated network in the j^{th} state for $j \in \{1, 2\}$.

Consider again the single shared buffer problem. Let $\lambda(1) = 1, \mu_1(1) = 3.5, \mu_2(1) = 2.5, \gamma(1) = 0.2$ and $\lambda(2) = 1, \mu_1(2) = 4.5, \mu_2(2) = 4.5, \gamma(2) = 0.5$. Exactly as in the case of the Jackson network the Neumann boundary conditions will hold, and it can be shown by numerically evaluating the Hamiltonian that the function $\bar{V}(x_1, x_2) = 1.00029(1 - x_1 - x_2) \vee 0$ is a viscosity subsolution and that $\bar{V}(0, 0)$ is equal to the large deviations rate of the problem. As above it follows that a GDPR scheme derived from this generating function will be asymptotically optimal. Table 3 gives results for simulations using a GDPR scheme derived from $\bar{V}$ by choosing $\Delta = 2 \times 1.00029$. Each estimate was derived using 20,000 runs.

n	30	40	50
Theoretical Value	6.36×10^{-13}	2.88×10^{-17}	1.30×10^{-21}
Estimate	6.36×10^{-13}	2.87×10^{-17}	1.29×10^{-21}
Std. Err.	0.18×10^{-13}	0.09×10^{-17}	0.05×10^{-21}
95% C.I.	$[6.00, 6.72] \times 10^{-13}$	$[2.69, 3.05] \times 10^{-17}$	$[1.20, 1.38] \times 10^{-21}$
Time Taken (s)	1	2	2

Table 3: Hitting Probabilities, Non-Markovian Process

7.2 Estimating Stationary Measures

We now consider the problem of estimating stationary measures. Typically this is done in one of two ways. Suppose that we have a process $\{X\}$ with stationary measure π, such that for all Borel $C \subset D$ and all initial conditions x

$$\lim_{i\to\infty} E_x \left[I\left(X_i \in C\right)\right] = \pi(C). \tag{7.4}$$

It is well known that (7.4) holds under suitable mixing and stability conditions, see for example [23]. When (7.4) does hold $\pi(C)$ is often approximated by

$$E_x\left[\frac{1}{K_2}\sum_{i=K_1+1}^{K_1+K_2} I\left(X_i\in C\right)\right], \tag{7.5}$$

which is itself estimated using Monte Carlo simulation. It should be noted that estimating the quantity (7.5) leads to a biased estimator for the stationary measure $\pi(C)$, however the length of the "burn-in" period K_1 is chosen so that the effects of the transient parts of the behavior of the process $\{X\}$ can be ignored. This approach to estimating stationary measures will be referred to as the occupation measure method.

An alternative approach is the regenerative method. To simplify the exposition we assume that the state space of the process $\{X\}$ is discrete. It is possible to drop this restriction, see [6] and the references therein for more details. Suppose that a stationary measure exists as well as a positive recurrent state O. Then it is well known, see for example [25], that for any $C\subset D$

$$\pi(C)=\frac{E_O\left[s_{O,C}\right]}{E_O\left[s_O\right]}, \tag{7.6}$$

where s_O is the time of first return to the state O and $s_{O,C}$ is the amount of time spent in C prior to the first return to O. Thus $\pi(C)$ can be estimated by estimating these expected values and taking their ratio. This estimator has two advantages. Firstly it is nearly unbiased in that the estimator for the numerator is unbiased, while relatively much more accurate estimates of the denominator are easy to obtain. The second is that it does not require the calculation of any fundamental properties of the process, such as the decay rate of the transient. In this method the point O is often referred to as the regeneration point.

We now show how these two methods may be used in the context of rare event simulation and the GDPR algorithm. Assume that we have a sequence of processes $\{X^n\}$ and a subset $C\subset D$, and wish to estimate the stationary measures $\pi^n(C)$. Assume further that (7.4) and (7.6) hold for each n and that there exists $O\in D$, positive recurrent for each n, such that O is an attracting point of the limiting dynamics of $\{X^n\}$. Define $\mathcal{J}$ by (5.3) with the stopping set M equal to O, and further assume that for all $E\subset D$ such that $\overline{E}=\overline{E^\circ}$

$$\lim_{n\to\infty}-\frac{1}{n}\log\pi^n(E)=\inf_{y\in E}\mathcal{J}(O,y). \tag{7.7}$$

We first consider the problem of estimating $\pi^n(C)$ using the occupation measure method. To apply this method one must find suitable values K_1^n and K_2^n such that (7.5) is a sufficiently good approximation to $\pi^n(C)$ for each n. We start by assuming one can take $K_2^n = K$ for all n for some $K \gg 1$. This assumption is essentially the same as one made in the original implementations of the RESTART and DPR algorithms, [4] and [21]. In these papers it is implicitly assumed that K_1^n can be chosen to be equal to 0 for all n, however in many cases the transient behavior of the process may take a considerable amount of time to decay. Some rigorous methods for determining a suitable amount of burn-in are discussed in [17], however to produce the results presented in this paper we used the following heuristic.

Consider the calculus of variations problem (7.7) and suppose that there exists some trajectory $\phi^*(s)$ and terminal time T^* such that $\phi^*(0) = O$, $\phi^*(T^*) \in C$ and $\int_0^{T^*} L(\phi^*(s), \dot{\phi}^*(s))ds = \inf_{y \in C} \mathcal{J}(O, y)$. Intuitively this suggests that the process requires a time of roughly nT^* for sufficient mixing to have occurred that the occupation measure of the set C is close to $\pi^n(C)$ and so we should choose a burn in time of length nT for some $T \gg T^*$. Once the times K_1^n and K_2^n have been specified it remains to design a GDPR for estimating the quantities (7.5) using a suitable subsolution to (7.7). Note that for each n we can write

$$E_x\left[\frac{1}{K_2}\sum_{i=K_1^n+1}^{K_1^n+K_2^n} I\left(X_i \in C\right)\right] = E_x\left[\sum_{i=0}^{\tau^n} e^{-nF^n(X_i^n)}\right],$$

where

$$\tau^n = \inf\left\{i : (X_i^n, i/(K_1^n + K_2^n)) \in D \times \{1\}\right\}$$

and

$$F^n(y,i) = \infty I\left(y \notin C \text{ or } i/(K_1^n + K_2^n) < (K_1^n + 1)/(K_1^n + K_2^n)\right).$$

Since F^n depends on n the form of this problem is not consistent with (5.1). However by observing that $\infty I\left(y \notin C\right) \leq F^n(y,i) \leq \infty I\left(y \notin C \text{ or } i/(K_1^n + K_2^n) < 1\right)$, and since for these functions the corresponding limits coincide, it follows that the results proved in the previous sections can still be applied.

Finally consider estimating $\pi^n(C)$ using the regenerative method. We assume that the regeneration point is always chosen to be the point O described in (7.7). We further assume that the quantities $E_O\left[s_O^n\right]$ can be accurately estimated using standard Monte Carlo simulation so the problem reduces to that of finding a good GDPR scheme for estimating $E_O[s_{O,C}^n]$. It is clear that this problem is of the form (5.1) and it follows from the assumptions made above that the large deviations properties of the expected values $E_O[s_{O,C}^n]$ are identical to those of the stationary measures $\pi^n(C)$, and in particular that $\lim_{n\to\infty} -\frac{1}{n}\log[Es_{O,C}^n]$ equals (7.7). Thus in order to design a GDPR scheme it is again remains only to find a suitable subsolution to (7.7).

Note that the assumption (7.7) has a particularly useful practical consequence, which is that the large deviations decay rate for $\{\pi^n(C)\}$ coincides with the decay rate for the probability of hitting C before the first return to O, after starting at O.

In particular, the calculus of variations problems and corresponding HJB PDEs are the same in both cases and so it follows that the same generating functions can be used and that the asymptotic work-normalised relative errors will be identical. We will now illustrate these ideas by considering two queueing problems.

Consider first the same stable tandem Jackson network as in Section 7.1, and suppose we want to estimate

$$p^n = \pi\left(Q_1(t) + Q_2(t) \geq n\right), \tag{7.8}$$

where π denotes the stationary measure of $(Q_1(t), Q_2(t))$. It is easy to see that for all n

$$p^n = \pi^n\left(\left\{(a,b) \in \mathbb{R}^2 : x + y \geq 1\right\}\right),$$

where π^n is the stationary measure of the rescaled process (7.1). Further it follows from the discussion above that we can design GDPR schemes for both the occupation and regenerative methods using the same generating function as was used in the previous section for estimating the corresponding hitting probabilities, and that the resulting GDPR scheme will again be asymptotically optimal. Numerical results are presented in Tables 4 and 5 below. The regenerative estimator was made using $20,000$ runs and the regeneration point was chosen to be $(0,0)$. For the occupation measure method K was chosen to be $20,000$ and the above heuristic argument suggested a choice of $10n$ for K_1^n. Due to the effects of correlation the variance of the estimator based on the occupation measure method cannot be estimated in the standard way. Instead the variance was estimated using the method proposed in [24].

n	20	30	40
Theoretical Value	1.43×10^{-12}	6.16×10^{-19}	2.39×10^{-25}
Estimate	1.55×10^{-12}	6.39×10^{-19}	2.34×10^{-25}
Std. Err.	0.32×10^{-12}	1.54×10^{-19}	0.57×10^{-25}
95% C.I.	$[0.82, 2.28] \times 10^{-12}$	$[3.38, 9.41] \times 10^{-19}$	$[1.22, 3.46] \times 10^{-25}$
Time Taken (s)	0.4	0.6	1

Table 4: Stationary Measures, Ergodic Method

n	20	30	40
Theoretical Value	1.43×10^{-12}	6.16×10^{-19}	2.39×10^{-25}
Estimate	1.37×10^{-12}	6.47×10^{-19}	2.40×10^{-25}
Std. Err.	0.08×10^{-12}	0.46×10^{-19}	0.19×10^{-25}
95% C.I.	$[1.21, 1.54] \times 10^{-12}$	$[5.56, 7.38] \times 10^{-19}$	$[2.03, 2.76] \times 10^{-25}$
Time Taken (s)	0.7	1	2

Table 5: Stationary Measures, Regenerative Method

We also revisit the non-Markovian model and consider again estimating the stationary measures (7.8). Define the rescaled process

$$(X_1^n(i), X_2^n(i), M^n(i)) = \left(\frac{Q_1(t_i)}{n}, \frac{Q_2(t_i)}{n}, M(t_i) \right),$$

where t_i is defined in the same manner as before. It can be shown that for all n and $C \in D$

$$p^n = \frac{\alpha \pi^n (C \times \{1\}) + \beta \pi^n (C \times \{2\})}{\alpha \pi^n (D \times \{1\}) + \beta \pi^n (D \times \{2\})}$$

where

$$\alpha = (\lambda(1) + \mu_1(1) + \mu_2(1) + \gamma(1))^{-1} \text{ and } \beta = (\lambda(2) + \mu_1(2) + \mu_2(2) + \gamma(2))^{-1}.$$

Tables 6 and 7 below show numerical results for estimating these probabilities. The quantities $\pi^n (D \times \{1\})$ and $\pi^n (D \times \{2\})$ were estimated using standard Monte Carlo methods and the quantities

$$\pi^n (C \times \{1\}) \tag{7.9}$$

and

$$\pi^n (C \times \{2\}) \tag{7.10}$$

were estimated using both the occupation measure and regenerative methods via GDPR schemes derived using the same generating function and Δ as in Section 7.1. We first present data for simulating the stationary measure. In this case K was chosen to be 20,000 and K_1 was chosen to be equal to $50n$. The estimates obtained using the regenerative method were made using $20,000$ runs. The quantity (7.9) was estimated by choosing the regeneration point to be $((0,0),1)$ and (7.10) was estimated by choosing the regeneration point to be $((0,0),2)$.

n	20	30	40
Theoretical Value	4.91×10^{-09}	2.23×10^{-13}	1.01×10^{-17}
Estimate	5.10×10^{-09}	2.26×10^{-13}	0.82×10^{-17}
Std. Err.	1.63×10^{-09}	0.71×10^{-13}	0.27×10^{-17}
95% C.I.	$[1.89, 8.31] \times 10^{-09}$	$[0.87, 3.65] \times 10^{-13}$	$[0.29, 1.35] \times 10^{-17}$
Time Taken (s)	1	1	2

Table 6: Stationary Measures, Non-Markovian Process, Ergodic Method

7.3 Rare Events for the Sample Mean

We conclude this section by considering the use of subsolutions to create GDPR schemes for a finite time problem. Let $X_1, X_2, \ldots$ be a sequence of iid $\mathrm{N}(0, I^N)$

n	20	30	40
Theoretical Value	4.91×10^{-09}	2.23×10^{-13}	1.01×10^{-17}
Estimate	4.92×10^{-09}	2.13×10^{-13}	0.94×10^{-17}
Std. Err.	0.23×10^{-09}	0.13×10^{-13}	0.06×10^{-17}
95% C.I.	$[4.47, 5.36] \times 10^{-09}$	$[1.89, 2.38] \times 10^{-13}$	$[0.82, 1.06] \times 10^{-17}$
Time Taken (s)	1	2	2

Table 7: Stationary Measures, Non-Markovian Process, Regenerative Method

random variables where I^N is the N-dimensional identity matrix and let $S^n = \frac{1}{n}\sum_{i=1}^n X_i$. Suppose that we are interested in simulating the expected values

$$E\left[\sum_{m=1,\ldots,M} e^{n\,\langle \bar{a}^m, S^n\rangle}\right]$$

for some sequence of vectors $\bar{a}^1, \ldots, \bar{a}^M \in \mathbb{R}^d$. For $j \in \{1, \ldots, n\}$ let $S_n(j) = \frac{1}{n}\sum_{i=1}^j X_i$. Then given sequences x_n, j_n and $x \in \mathbb{R}^N$, $t \in [0,1]$ such that $\lim_{n\to\infty} x_n = x$ and $\lim_{n\to\infty} j_n/n = t$, it can be shown (see [10]) that

$$W(x,t) = \inf_{m=1,\ldots,M}\left\{-\langle \bar{a}^m, x\rangle - \frac{(1-t)}{2}\|\bar{a}^m\|^2\right\}.$$

Further the HJB PDE corresponding to the calculus of variations problem that describes the large deviations properties of this process is

$$W_t + \mathbb{H}(DW) = 0,\; t < 1 \tag{7.11}$$

and the terminal condition

$$W(x,1) = \min_{m=1,\ldots,M}\left\{-\langle \bar{a}^m, x\rangle\right\}, \tag{7.12}$$

where $L(\beta) = \|\beta\|^2/2$ and $\mathbb{H}(q) = \inf_{\beta\in\mathbb{R}^N}[\langle q,\beta\rangle + L(\beta)] = -\|q\|^2/2$. Note that this problem can be put into the general framework by considering the time variable as simply another state variable, rescaling time by $1/n$ and letting the stopping time τ^n equal 1 for all n. A smooth function U will be a subsolution for this equation if $U_t + \mathbb{H}(DU) \geq 0$ and if $U(x,1) \leq -\langle \bar{a}^m, x\rangle$ for $m = 1, \ldots, M$. Despite the fact that Condition 5.1 part 2 is no longer true it is easy to show that the conclusions of Theorem 5.5 still hold. It can be verified by inspection that any affine function of the form

$$U(x,t) = -\langle \alpha, x\rangle + \gamma - (1-t)\frac{\|\alpha\|^2}{2},$$

$\alpha, \gamma \in \mathbb{R}^N$ satisfies $U_t + \mathbb{H}(DU) \geq 0$. In particular, the functions $U_{\bar{a}^m}(x,t) = -\langle \bar{a}^m, x\rangle - \quad (1-T)\|\bar{a}^m\|^2, m = 1, \ldots, M$, are viscosity solutions to the HJB

PDE with terminal conditions $W(x,1) = -\langle \bar{a}^m, x\rangle$ respectively. Using the fact that the pointwise minimum of a finite collection of viscosity solutions is again a viscosity solution it follow that the function $U(x,t) = \inf_{m=1,\ldots,M} U_{\bar{a}^m}(x,t)$ is a viscosity subsolution to the HJB PDE with terminal condition (7.12) such that $U(0,0) = W(0,0)$.

Thus we can look for a generating function of the form $\bar{V}(x,t) = U(x,t) \wedge \gamma$ for some suitable γ. The choice of $\gamma = \inf_{m=1,\ldots,M}\{-\frac{3}{2}\left\|\bar{a}^m\right\|^2\}$ leads to an asymptotically optimal GDPR scheme and indeed this is the largest choice of γ for which this is true. Numerical results for a GDPR scheme derived using this choice of generating function and $\Delta = 0.5$ with $\bar{a}^1 = (1,0)$ and $\bar{a}^2 = (0,1)$ are shown in Table 8 below. Each estimate was derived using 20,000 runs.

n	40	60	80
Theoretical Value	9.70×10^8	2.14×10^{13}	4.71×10^{17}
Estimate	9.60×10^8	2.19×10^{13}	4.35×10^{17}
Std. Err.	0.47×10^8	0.13×10^{13}	0.32×10^{17}
95% C.I.	$[8.68, 10.5] \times 10^8$	$[1.94, 2.44] \times 10^{13}$	$[3.72, 4.97] \times 10^{17}$
Time Taken (s)	1.1	1.6	2.1

Table 8: Rare Events for the Sample Mean

8 Conclusions

We have shown that a generalized version of the GDPR algorithm can be defined which can be used to simulate a wide range of expected values. Further the generating function and subsolutions approach provides a rigorous and flexible framework for the design and analysis of such algorithms. One interesting feature of the numerical results presented is that although the RESTART and DPR algorithms have traditionally been implemented for estimating stationary measures using the occupation measure method the results presented in this paper do not indicate a significant superiority over the regenerative method. Given the relative simplicity in designing and analyzing various aspects of the regenerative method (e.g., sample variance), this method seems preferable.

There remain several topics for future research, in particular a theoretical comparison of the GDPR and standard splitting algorithms is required before the topic of multi-level splitting can be said to fully understood. Further as there is much interest in simulating rare events for diffusion processes a future goal is to formulate a version of the GDPR algorithm for continuous time processes. Finally there is much interest in processes whose limiting behavior has multiple attracting points. The question of finding good importance functions in this case for either importance sampling or multi-level splitting has yet to be addressed.

9 Appendix

This section contains the proofs of Lemma 3.3, Theorem 4.3 and Theorem 4.4.

Proof of Lemma 3.3. We use the notation N^τ_i, etc., as defined in the statement of the GDPR algorithm, and assume that $\tau > 0$, since otherwise the lemma is trivial. Recall that we assume that τ is the first entry time of some set M. The second result is obtained by summing the first one over l. We will prove the first display by induction on i. Clearly the result holds for $i = 0$. Suppose the result has been proved up to some i^*. We can write

$$\begin{aligned} & e^{V(x_0)}E_{x_0}\left[\sum_{m=1}^{N^\tau_{i^*+1}} \bar f(\bar X^\tau_{i^*+1,m})1_{\{L^\tau_{i^*+1,m}=l\}}\right] \\ &\quad = \sum_{k=0}^{\rho(x_0)} \mathcal{L}_{\rho(x_0)}(k)e^{V(x_0)}E_{x_0,k}\left[\sum_{m=1}^{N^\tau_{i^*+1}} \bar f(\bar X^\tau_{i^*+1,m})1_{\{L^\tau_{i^*+1,m}=l\}}\right], \end{aligned}$$

where $E_{x_0,k}$ denotes expected value given $\bar X^\tau_{0,1} = x_0$ and $L^\tau_{0,1} = k$. Note that

$$\begin{aligned} & E_{x_0,k}\left[\sum_{m=1}^{N^\tau_{i^*+1}} \bar f(\bar X^\tau_{i^*+1,m})1_{\{L^\tau_{i^*+1,m}=l\}}\right] \\ &\quad = E_{x_0,k}\left[\sum_{r=1}^{N^\tau_1} E_{\bar X^\tau_{1,r},L^\tau_{1,r}}\left[\sum_{m=1}^{N^\tau_{i^*+1}} \bar f(\bar X^\tau_{i^*+1,m})1_{\{L^\tau_{i^*+1,m}=l\}}\right]\right]. \end{aligned}$$

Thus the original quantity can be written as

$$\sum_{k=0}^{\rho(x_0)} \mathcal{L}_{\rho(x_0)}(k)e^{V(x_0)}E_{x_0,k}\left[\sum_{r=1}^{N^\tau_1} E_{\bar X^\tau_{1,r},L^\tau_{1,r}}\left[\sum_{m=1}^{N^\tau_{i^*}} \bar f(\bar X^\tau_{i^*,m})1_{\{L^\tau_{i^*+1,m}=l\}}\right]\right]. \tag{9.1}$$

Conditioning on the value of $y = Y_{0,1}$ as it appears in the pseudo code naturally partitions the problem into three cases. In the first case $\rho(y) = \rho(x_0)$. In this case we have $N^\tau_1 = 1$, $\bar X^\tau_{1,1} = y$, and $L^\tau_{1,1} = L^\tau_{0,1}$ has distribution $\mathcal{L}_{\rho(x_0)} = \mathcal{L}_{\rho(y)}$. The conditional version of (9.1) can be written as

$$\sum_{k=0}^{\rho(y)} \mathcal{L}_{\rho(y)}(k)e^{V(x_0)}E_{y,k}\left[\sum_{m=1}^{N^\tau_{i^*}} \bar f(\bar X^\tau_{i^*,m})1_{\{L^\tau_{i^*,m}=l\}}\right].$$

In the second case $\rho(y) < \rho(x_0)$. In this case the particle is killed if and only if $L^\tau_{0,1} > \rho(y)$. If $L^\tau_{0,1} \le \rho(y)$ then $N^\tau_1 = 1$ and $L^\tau_{1,1} = L^\tau_{0,1}$. The conditioned version of

(9.1) is then

$$\begin{aligned}
&\sum_{k=0}^{\rho(y)} \mathcal{L}_{\rho(x_0)}(k) e^{V(x_0)} E_{y,k}\left[\sum_{j=m}^{N_{i^*}^\tau} \bar{f}(\bar{X}_{i^*,m}^\tau) 1_{\left\{L_{i^*,m}^\tau = l\right\}}\right] \\
&\quad = \sum_{k=0}^{\rho(y)} \frac{\mathcal{L}_{\rho(x_0)}(k)}{\mathcal{L}_{\rho(y)}(k)} \mathcal{L}_{\rho(y)}(k) e^{V(x_0)} E_{y,k}\left[\sum_{m=1}^{N_{i^*}^\tau} \bar{f}(\bar{X}_{i^*,m}^\tau) 1_{\left\{L_{i^*,m}^\tau = l\right\}}\right] \\
&\quad = e^{V_{\rho(y)} - V_{\rho(x_0)}} \sum_{k=0}^{\rho(y)} \mathcal{L}_{\rho(y)}(k) e^{V(x_0)} E_{y,k}\left[\sum_{m=1}^{N_{i^*}^\tau} \bar{f}(\bar{X}_{i^*,m}^\tau) 1_{\left\{L_{i^*,m}^\tau = l\right\}}\right].
\end{aligned}$$

Finally there is the case when $\rho(y) > \rho(x_0)$. Here there is the possibility that new particles are created (i.e., $N_1^\tau > 1$), though in all cases we have $\bar{X}_{1,r}^\tau = y$. When new particles are created, the associated thresholds are determined according to the measure $\mathcal{Q}_{j,k}$, and so by the assumption of unbiasedness on these measures (9.1) takes the form

$$\begin{aligned}
&\left[\sum_{j=\rho(x_0)+1}^{\rho(y)} \frac{e^{V_j} - e^{V_{j-1}}}{e^{V_{\rho(x_0)}}} e^{V(x_0)} E_{y,j}\left[\sum_{m=1}^{N_{i^*}^\tau} \bar{f}(\bar{X}_{i^*,m}^\tau) 1_{\left\{L_{i^*,m}^\tau = l\right\}}\right]\right. \\
&\qquad \left. + \sum_{k=0}^{\rho(x_0)} \mathcal{L}_{\rho(x_0)}(k) e^{V(x_0)} E_{y,k}\left[\sum_{m=1}^{N_{i^*}^\tau} \bar{f}(\bar{X}_{i^*,m}^\tau) 1_{\left\{L_{i^*,m}^\tau = l\right\}}\right]\right] \\
&\quad = \sum_{j=0}^{\rho(y)} \frac{e^{V_j} - e^{V_{j-1}}}{e^{V_{\rho(x_0)}}} e^{V(x_0)} E_{y,j}\left[\sum_{m=1}^{N_{i^*}^\tau} \bar{f}(\bar{X}_{i^*,m}^\tau) 1_{\left\{L_{i^*,m}^\tau = l\right\}}\right] \\
&\quad = e^{V_{\rho(y)} - V_{\rho(x_0)}} \sum_{j=0}^{\rho(y)} \mathcal{L}_{\rho(y)}(j) e^{V(x_0)} E_{y,j}\left[\sum_{m=1}^{N_{i^*}^\tau} \bar{f}(\bar{X}_{i^*,m}^\tau) 1_{\left\{L_{i^*,m}^\tau = l\right\}}\right].
\end{aligned}$$

It follows that

$$\begin{aligned}
&e^{V(x_0)} E_{x_0}\left[\sum_{j=1}^{N_{i^*+1}^\tau} \bar{f}(\bar{X}_{i^*+1,j}^\tau) 1_{\left\{L_{i^*+1,m}^\tau = l\right\}}\right] \\
&\quad = e^{V(x_0)} E_{x_0}\left[e^{V_{\rho(\bar{X}_{1,1}^\tau)} - V_{\rho(\bar{X}_{0,1}^\tau)}} \sum_{j=0}^{\rho(\bar{X}_{1,1}^\tau)} \mathcal{L}_{\rho(\bar{X}_{1,1}^\tau)}(j) E_{\bar{X}_{1,1}^\tau, j}\left[\sum_{m=1}^{N_{i^*}^\tau} \bar{f}(\bar{X}_{i^*,m}^\tau) 1_{\left\{L_{i^*,m}^\tau = l\right\}}\right]\right].
\end{aligned}$$

Conditioning again on $\bar{X}_{1,1}^\tau$, using that X_1 has the same distribution as $\bar{X}_{1,1}^\tau$ given

$\tau \geq 1$ and applying the induction hypothesis gives

$$\begin{aligned}
& e^{V(x_0)} E_{x_0}\left[\sum_{j=1}^{N^\tau_{i^*+1}} \bar{f}(\bar{X}^\tau_{i^*+1,j}) 1_{\left\{L^\tau_{i^*+1,m}=l\right\}}\right] \\
= \;& e^{V(x_0)} E_{x_0}\left[e^{V_{\rho(\bar{X}^\tau_{1,1})}-V_{\rho(\bar{X}^\tau_{0,1})}} E_{\bar{X}^\tau_{1,1}}\left[\sum_{m=1}^{N^\tau_{i^*}} \bar{f}(\bar{X}^\tau_{i^*,m}) 1_{\left\{L^\tau_{i^*,m}=l\right\}}\right]\right] \\
= \;& e^{V(x_0)} E_{x_0}\left[e^{V_{\rho(X_1)}-V_{\rho(X_0)}} E_{X_1}\left[\sum_{m=1}^{N^\tau_{i^*}} \bar{f}(\bar{X}^\tau_{i^*,m}) 1_{\left\{L^\tau_{i^*,m}=l\right\}}\right]\right] \\
= \;& E_{x_0}\left[e^{V_{\rho(X_1)}} E_{X_1}\left[\sum_{m=1}^{N^\tau_{i^*}} \bar{f}(\bar{X}^\tau_{i^*,m}) 1_{\left\{L^\tau_{i^*,m}=l\right\}}\right]\right] \\
= \;& E_{x_0}\left[E_{X_1}\left[\bar{f}(X_{i^*})(e^{V_l}-e^{V_{l-1}}) 1_{\{\rho(X_{i^*})\geq l\}} 1_{\{\tau \geq i^*\}}\right]\right] \\
= \;& E_{x_0}\left[\bar{f}(X_{i^*+1})(e^{V_l}-e^{V_{l-1}}) 1_{\left\{\rho(X_{i^*+1})\geq l\right\}} 1_{\{\tau \geq i^*+1\}}\right].
\end{aligned}$$

■

Proof of Theorem 4.3. Recall that τ is the first entry time of some closed set $M \subset D$. First consider the case where $\bar{f}$ is bounded and there is a $T < \infty$ such that $\tau \leq T$ a.s. Let $W(x) = e^{-V(x)} E_x[(\hat{s}(\bar{f}))^2]$ and let $Z(x, j; k), k = 0, \ldots$ denote iid sequences of random variables with the same distribution as $\hat{s}(\bar{f})$, conditioned on $\bar{X}^\tau_{0,1} = x$ and $L^\tau_{0,1} = j$. Since $\bar{f}$ and the time interval are bounded these random variables are also bounded.

The proof is based on finding a recursive equation for W. If $x_0 \notin M$ then there are three contributions to $\hat{s}(\bar{f})$ depending on the killing and/or splitting that takes place over the next time step. The first is due simply to the current state of the particle and is always present. The second is due to future contributions if the particle stays above the support threshold, and the third occurs if new particles are generated. To account for thresholds of both the existing particles and those which might be generated, let $\bar{Q}^{j,k}_l$ be random variables equal in distribution to $Q^{j,k} + 1_{\{L \leq k\}} e_L$, where e_l denotes the l^{th} unit vector and L is a random variable independent of $Q^{j,k}$ and with distribution $\mathcal{L}_j$. Using the splitting distributions for

$Q_l^{j,k}$ given above,

$$\begin{aligned} W(x_0) = e^{-V(x_0)} E_{x_0} \Bigg[\Bigg(& e^{V(x_0)} \bar{f}(x_0) + 1_{\{L_{0,1}^\tau \le \rho(\bar{X}_{1,1}^\tau)\}} e^{V(\bar{X}_{0,1}^\tau) - V(\bar{X}_{1,1}^\tau)} Z(\bar{X}_{1,1}^\tau, L_{0,1}^\tau; 0) \\ & + 1_{\{\rho(\bar{X}_{0,1}^\tau) < \rho(\bar{X}_{1,1}^\tau)\}} \left(\sum_{j=\rho(\bar{X}_{0,1}^\tau)+1}^{\rho(\bar{X}_{1,1}^\tau)} \sum_{m=1}^{Q_j^{\rho(\bar{X}_{0,1}^\tau),\rho(\bar{X}_{1,1}^\tau)}} e^{V(\bar{X}_{0,1}^\tau) - V(\bar{X}_{1,1}^\tau)} Z(\bar{X}_{1,1}^\tau, j; m) \right) \Bigg)^2 \Bigg] \\ = \; & e^{V(x_0)} \bar{f}(x_0)^2 + 2\bar{f}(x_0) E_{x_0} \left[\sum_{j=0}^{J} \sum_{m=1}^{\bar{Q}_j^{\rho(\bar{X}_{0,1}^\tau),\rho(\bar{X}_{1,1}^\tau)}} e^{V(\bar{X}_{0,1}^\tau) - V(\bar{X}_{1,1}^\tau)} Z(\bar{X}_{1,1}^\tau, j; m) \right] \\ & + e^{-V(x_0)} E_{x_0} \left[\left(\sum_{j=0}^{J} \sum_{m=1}^{\bar{Q}_j^{\rho(\bar{X}_{0,1}^\tau),\rho(\bar{X}_{1,1}^\tau)}} e^{V(\bar{X}_{0,1}^\tau) - V(\bar{X}_{1,1}^\tau)} Z(\bar{X}_{1,1}^\tau, j; m) \right)^2 \right], \end{aligned}$$

If $x_0 \in M$ then $W(x_0) = e^{-V(x_0)} \left(e^{V(x_0)} \bar{f}(x_0) \right)^2 = e^{-V(x_0)} f(x_0)^2$.

We now use the following facts: $L_{0,1}$ has distribution $\mathcal{L}_{\rho(X_0)}$; $\bar{X}_{1,1}^\tau$ has the same distribution (conditioned on $\bar{X}_{0,1}^\tau = X_0 = x_0$) as X_1; by unbiasedness [see (3.2) and (3.3)] and the definition of $\bar{Q}_l^{j,k}$, for all j, k, l

$$E \bar{Q}_l^{j,k} e^{V_j - V_k} = \mathcal{L}_k(l); \tag{9.2}$$

and that the future evolution of the algorithm is independent of the $\bar{Q}_j^{k,l}$. Together with the last display, these give

$$\begin{aligned} W(x_0) = & \; e^{V(x_0)} \bar{f}(x_0)^2 + 2\bar{f}(x_0) E_{x_0} \left[\sum_{j=0}^{\rho(X_1)} \mathcal{L}_{\rho(X_1)}(j) E_{X_1,j}[\hat{s}(\bar{f})] \right] \\ & + e^{V(x_0)} E_{x_0} \left[\sum_{j,k=1}^{J} e^{-2V(X_1)} \bar{Q}_j^{\rho(X_0),\rho(X_1)} \bar{Q}_k^{\rho(X_0),\rho(X_1)} E_{X_1,j}[\hat{s}(\bar{f})] E_{X_1,k}[\hat{s}(\bar{f})] \right] \\ & + e^{V(x_0)} E_{x_0} \left[\sum_{j=1}^{J} e_j^{-2V(X_1)} \bar{Q}_j^{\rho(X_0),\rho(X_1)} \left(E_{X_1,j} \left[(\hat{s}(\bar{f}))^2 \right] - \left(E_{X_1,j}[\hat{s}(\bar{f})] \right)^2 \right) \right]. \end{aligned} \tag{9.3}$$

We examine the various terms separately. First note that given a generic starting point x the definition of the algorithm dictates that $L_{0,1}^\tau$ will have distribution $\mathcal{L}_{\rho(x)}$, and so by Theorem 3.2

$$2\bar{f}(x_0) E_{x_0} \left[\sum_{j=0}^{\rho(X_1)} \mathcal{L}_{\rho(X_1)}(j) E_{X_1,j}[\hat{s}(\bar{f})] \right] = E_{x_0} \left[2\bar{f}(x_0) E_{X_1} \left[\sum_{k=0}^{\tau} f(X_k) \right] \right]. \tag{9.4}$$

Again using (9.2),

$$\begin{aligned}
& e^{V(x_0)} E_{x_0}\left[\sum_{j=0}^{J} e^{-2V(X_1)} \bar{Q}_j^{\rho(X_0),\rho(X_1)} E_{X_1,j}\left[\left(\hat{s}(\bar{f})\right)^2\right]\right] \\
= \quad & E_{x_0}\left[e^{-V(X_1)} \sum_{j=0}^{J} e^{V(X_0)-V(X_1)} \bar{Q}_j^{\rho(X_0),\rho(X_1)} E_{X_1,j}\left[\left(\hat{s}(\bar{f})\right)^2\right]\right] \\
= \quad & E_{x_0}\left[e^{-V(X_1)} \sum_{j=0}^{J} \mathcal{L}_{\rho(X_1)}(j) E_{X_1,j}\left[\left(\hat{s}(\bar{f})\right)^2\right]\right] \\
= \quad & E_{x_0}\left[W\left(X_1\right)\right]. \qquad (9.5)
\end{aligned}$$

This leaves only the quantity

$$\begin{aligned}
& e^{-V(x_0)} E_{x_0}\left[\left(\sum_{j=0}^{J}\sum_{l=0}^{J} e^{2V(X_0)-2V(X_1)} \bar{Q}_j^{\rho(X_0),\rho(X_1)} \bar{Q}_l^{\rho(X_0),\rho(X_1)} E_{X_1,j}\left[\hat{s}(\bar{f})\right] E_{X_1,l}\left[\hat{s}(\bar{f})\right]\right)\right] \\
& -e^{-V(x_0)} E_{x_0}\left[\left(\sum_{j=0}^{J} e^{2V(X_0)-2V(X_1)} \bar{Q}_j^{\rho(X_0),\rho(X_1)} \left(E_{X_1,j}\left[\hat{s}(\bar{f})\right]\right)^2\right)\right].
\end{aligned}$$

The terms with both l and j below $\rho(X_0)$ contribute nothing to this expression, since $\bar{Q}_j^{\rho(X_0),\rho(X_1)}$ is then either 0 or 1. We can thus drop these terms, and decompose the double sum as

$$\sum_{j=\rho(X_0)+1}^{\rho(X_1)} \sum_{l=\rho(X_0)+1}^{\rho(X_1)} + 2 \sum_{j=1}^{\rho(X_0)} \sum_{l=\rho(X_0)+1}^{\rho(X_1)}.$$

If $(Y_1, \ldots, Y_m)$ has multinomial distribution $M(N, p_1, \ldots, p_m)$ then we have the moments

$$EY_i = Np_i, \quad EY_i^2 = (N^2 - N)p_i^2 + Np_i, \quad EY_iY_j = (N^2 - N)p_ip_j, i \neq j,$$

and so straightforward calculation together with the definitions (3.2) give

$$\begin{aligned}
& e^{-V(x_0)} E_{x_0}\left[1_{\{\rho(X_1)>\rho(X_0)\}} e^{2V(X_0)-2V(X_1)}\left[B^2 - B\right]\left(\sum_{j=\rho(X_0)+1}^{\rho(X_1)} \mathcal{L}_{\rho(X_0),\rho(X_1)}(j) E_{X_1,j}\left[\hat{s}(\bar{f})\right]\right)^2\right] \\
& +2e^{-V(x_0)} E_{x_0}\left[1_{\{\rho(X_1)>\rho(X_0)\}}\left(\sum_{j=1}^{\rho(X_0)} \mathcal{L}_{\rho(X_1)}(j) E_{X_1,j}\left[\hat{s}(\bar{f})\right]\right)\left(\sum_{l=\rho(X_0)+1}^{\rho(X_1)} \mathcal{L}_{\rho(X_1)}(l) E_{X_1,l}\left[\hat{s}(\bar{f})\right]\right)\right],
\end{aligned}$$

where B is a random variable equal to $\lceil (e^{V_{\rho(X_1)}} - e^{V_{\rho(X_0)}})/e^{V_{\rho(X_0)}} \rceil$ with conditional probability $\{(e^{V_{\rho(X_1)}} - e^{V_{\rho(X_0)}})/e^{V_{\rho(X_0)}}\}$ and $\lfloor (e^{V_{\rho(X_1)}} - e^{V_{\rho(X_0)}})/e^{V_{\rho(X_0)}} \rfloor$ otherwise.

We use that the conditional expected value of $B^2 - B$ is bounded above by $[(e^{V_{\rho(X_1)}} - e^{V_{\rho(X_0)}})/e^{V_{\rho(X_0)}}]^2$, again the definition (3.2), and the non-negativity of f to get the following upper bound on the last display:

$$e^{-V(x_0)}E_{x_0}\left[1_{\{\rho(X_1)>\rho(X_0)\}}\left(\sum_{j=1}^{\rho(X_1)}\mathcal{L}_{\rho(X_1)}(j)E_{X_1,j}\left[\hat{s}(\bar{f})\right]\right)^2\right]. \qquad (9.6)$$

We now combine (9.3), (9.4), (9.5) and (9.6) to get that for $x_0 \notin M$

$$\begin{aligned} W(x_0) \le\ & e^{-V(x_0)}f(x_0)^2 + 2e^{-V(x_0)}E_{x_0}\left[f(x_0)E_{X_1}\left[\sum_{k=0}^{\tau}f(X_k)\right]\right] \\ & + e^{-V(x_0)}E_{x_0}\left[1_{\{\rho(X_1)>\rho(X_0)\}}E_{X_1}\left[\sum_{k=0}^{\tau}f(X_k)\right]^2\right] + E_{x_0}\left[W(X_1)\right]. \end{aligned}$$

Since all functions involved are bounded it follows that the process

$$\Sigma_i \doteq W(X_{i\wedge\tau}) + \sum_{j=1}^{i\wedge\tau}\left\{e^{-V(X_{j-1})}\left(f(X_{j-1}) + E_{X_j}\left[\sum_{k=0}^{\tau}f(X_k)\right]\right)^2\right\}$$

defined for $i \in \{0,\ldots,T\}$ is a submartingale. Thus, using that $W(X_{T\wedge\tau}) = e^{-V(X_\tau)}f(X_\tau)^2$,

$$\begin{aligned} e^{-V(x_0)}E_{x_0}[(\hat{s}(\bar{f}))^2] &= W(x_0) \\ &= \Sigma_0 \\ &\le E_{x_0}\left[\Sigma_T\right] \\ &= E_{x_0}\left[\sum_{i=1}^{\tau}e^{-V(X_{i-1})}\left(f(X_{i-1}) + E_{X_i}\left[\sum_{k=0}^{\tau}f(X_k)\right]\right)^2\right]. \end{aligned}$$

We next remove the restrictions on the stopping time τ. We add time as a state variable [i.e., work with the process (X_i, i)], and consider the analogous estimation problem where the stopping set is $M \times \{T\}$ (i.e., we stop if either X_i enters M or $i = T$), and with $f_T(y,i) = f(y)$ and $\bar{f}_T(y,i) = f_T(y,i)e^{-V(y)}$. One can then use $\hat{s}(\bar{f}_T)$ to denote an unbiased estimator for $E_{(x_0,0)}\left[\sum_{i=0}^{\tau\wedge M} f_T(X_i,i)\right]$ and observe that the distribution of this estimator will be equal to the distribution of the estimator $s(\bar{f})$ in the case that the corresponding algorithm is forcibly terminated at time T. Thus $\hat{s}(\bar{f}_T) = e^{V(x_0)}\sum_{i=0}^{T}\int_D \bar{f}(y)\bar{\delta}_{\bar{X}_i^T}(dy)$ and $\hat{s}(\bar{f}_T) \uparrow \hat{s}(\bar{f})$ a.s. By the MCT

$$E_{(x_0,0)}\left[\hat{s}(\bar{f}_T)\right] \to E_{x_0}\left[\hat{s}(\bar{f})\right], \quad E_{(x_0,0)}\left[(\hat{s}(\bar{f}_T))^2\right] \to E_{x_0}\left[(\hat{s}(\bar{f}))^2\right]$$

and for any i

$$E_{(x_i,i)}\left[\hat{s}(\bar{f}_T)\right] \to E_{x_i}\left[\hat{s}(\bar{f})\right].$$

Applying the MCT for a final time in the representation for $E_{(x_0,0)}[(\hat{s}(\bar{f}_T))^2]$ gives the desired result. Lastly one must remove the assumption that f (or $\bar{f}$) is bounded, but this follows again by the MCT. Thus (4.1) follows. ■

Proof of Theorem 4.4. We start with (9.3), and use (9.5) and that the remaining terms are non-negative to get

$$W(x_0) \geq e^{V(x_0)} \bar{f}(x_0)^2 + E_{x_0}\left[W(X_1)\right].$$

The result follows by arguing in the same way as in the proof of Theorem 4.3 using the supermartingale

$$\Sigma_i \doteq W(X_{i\wedge\tau}) + \sum_{j=1}^{i\wedge\tau} \left\{ e^{-V(X_{j-1})} f(X_{j-1})^2 \right\}.$$

■

References

[1] Proceedings of the 6th International Workshop on Rare Event Simulation, Bamberg, Germany, October 2006.

[2] M. Villen-Altamirano and J. Villen-Altamirano. RESTART: A method for accelerating rare event simulations. *Proc. of the 13th International Teletraffic Congress, Queueing, Performance and Control in ATM*, Elsevier, Amsterdam, 71–76, 1991.

[3] M. Villen-Altamirano and J. Villen-Altamirano. RESTART: A straightforward method for fast simulation of rare events. *Proc. of the 1994 Winter Simulation Conference*, 282–289, 1994.

[4] M. Villen-Altamirano and J. Villen-Altamirano. Analysis of RESTART simulation: Theoretical basis and sensitivity study. *European Transactions on Telecommunications*, 13:373–385, 2002.

[5] J. Villen-Altamirano. Rare event RESTART simulation of two-stage networks *European Journal of Operations Research*, 179:148–159, 2007.

[6] S. Asmussen and P. Glynn. *Stochastic Simulation: Algorithms and Analysis.* Springer-Verlag, New York, 2007.

[7] M. Bardi and I. Capuzzo-Dolcetta. *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations.* Birkhäuser, Boston, 1997.

[8] T. Dean. *A Subsolutions Approach to the Analysis and Implementation of Splitting Algorithms in Rare Event Simulation.* PhD Thesis. Brown University, 2008.

[9] T. Dean and P. Dupuis. Splitting for rare event simulation: A large deviation approach to design and analysis. *Stoc. Proc. App.*, to appear, DOI 10.1016/j.spa.2008.02.017.

[10] P. Dupuis and R. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations.* John Wiley & Sons, New York, 1997.

[11] P. Dupuis, H. Ishii and H.M. Soner. A viscosity solution approach to the asymptotic analysis of queueing systems, *Ann. Probab.,* 18:226–255, 1990.

[12] P. Dupuis, A. Sezer, and H. Wang. Dynamic importance sampling for queueing networks, *Ann. Appl. Probab.*, 17:1306–1346, 2007.

[13] P. Dupuis and H. Wang. Subsolutions of an Isaacs equation and efficient schemes for importance sampling, *Math. of Op. Res.*, 32:1–35, 2007.

[14] P. Dupuis and H. Wang. Importance sampling for Jackson networks, *preprint*, 2008.

[15] M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems.* Springer-Verlag, New York, 1984.

[16] E. Gelenbe and G. Pujolle. *Introduction to Queueing Networks.* John Wiley & Sons, 1987.

[17] W.R. Gilks, S. Richardson and D.J. Spiegelhalter (Eds.). *Markov Chain Monte Carlo In Practice.* Chapman & Hall, 1996.

[18] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. A large deviations perspective on the efficiency of multilevel splitting. *IEEE Trans. Automat. Control*, 43:1666–1679, 1998.

[19] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. Multilevel splitting for estimating rare event probabilities. *Op. Res.*, 47:585–600, 1999.

[20] Z. Haraszti and J. K. Townsend. The theory of direct probability redistribution and its application to rare event simulation. *Proc. of the IEEE International Conference on Communications 1998*, 1443–1450.

[21] Z. Haraszti and J. K. Townsend. The theory of direct probability redistribution and its application to rare event simulation. *ACM Transactions on Modelling and Computer Simulation*, 9:105–140, 1999.

[22] N. Madras. *Lectures on Monte Carlo Methods.* American Mathematical Society, 2002

[23] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag 1994.

[24] P. Moran. The estimation of standard errors in Monte Carlo simulation experiments. *Biometrika*, 62:1–4, 1975.

[25] J.R. Norris. *Markov Chains.* Cambridge University Press, 1998.