The Annals of Applied Probability
2020, Vol. 30, No. 6, 2963–2998
https://doi.org/10.1214/20-AAP1578

SPLITTING ALGORITHMS FOR RARE EVENT SIMULATION OVER LONG TIME INTERVALS

BY ANNE BUIJSROGGE[1], PAUL DUPUIS[2,*] AND MICHAEL SNARSKI[2,†]

[1]*Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, a.buijsrogge@utwente.nl*
[2]*Division of Applied Mathematics, Brown University, *paul_dupuis@brown.edu; †michael_snarski@brown.edu*

In this paper we study the performance of splitting algorithms, and in particular the RESTART method, for the numerical approximation of the probability that a process leaves a neighborhood of a metastable point during some long time interval $[0, T]$. We show that, in contrast to alternatives such as importance sampling, the decay rate of the second moment does not degrade as $T \to \infty$. In the course of the analysis we develop some related large deviation estimates that apply when the time interval of interest depends on the large deviation parameter.

1. Introduction. In this paper we study *exit probabilities* for discrete time stochastic processes, where the process escapes from some neighborhood of an attractor prior to a given time. In contrast to existing work on exit probabilities, see, for example, [14, 17], we allow the time by which the process has to exit to grow polynomially with the large deviations scaling parameter. One motivation for considering this scaling is that when the time interval is large and the escape probability is small, the probability closely approximates the inverse of the mean escape time, and in particular the exponential decay rate for the probability and growth rate for the escape time coincide. Although one can easily conjecture an expression for the decay rate for such probabilities, it does not follow from standard sample path large deviation estimates, which apply to bounded time intervals. A first contribution of this paper is to apply Freidlin–Wentzell type arguments to rigorously determine the decay rate.

The main focus of this paper is on Monte Carlo estimation of the probability, and in particular on the use of *splitting* methods; see for instance, [15]. In rare event simulation the two most commonly used methods are *importance sampling* and those based on interacting particles, which include splitting as a particular case. For a discussion on some of the differences and similarities of importance sampling and splitting we refer to [4]. For both importance sampling and splitting it turns out that one can design and analyze efficient Monte Carlo schemes using *subsolutions* to a Hamilton–Jacobi–Bellman equation that is naturally associated to the process through the large deviation rate function; see [5, 6, 12]. This method has been successfully applied in a variety of settings, such as queueing theory [10], and will also be used in the current paper.

It was shown in [11] that importance sampling has some shortcomings when applied to the problem of estimating probabilities to escape from the neighborhood of an attractor. For example, when the time interval over which escape can occur is large, one is tempted to consider subsolutions to the corresponding time independent Hamilton–Jacobi–Bellman equation for the basis for algorithm design. It is sometimes the case (e.g., for reversible systems) that useful subsolutions can be found much more easily for the time independent version. However, as discussed in [11], since the attractor is inside the interior of the domain of interest, importance sampling schemes based on such time independent subsolutions generically degrade as

Received February 2019; revised September 2019.
MSC2020 subject classifications. 65C05, 60F10, 60G99.
Key words and phrases. Splitting algorithms, RESTART, Monte Carlo methods, large deviations, metastable points.

the time interval gets large. In this paper we will consider the design of splitting schemes, and show that no such degradation occurs.

Within the setting of splitting schemes there are those which allow killing to improve efficiency without introducing bias, with one of the most well known of such schemes being the RESTART method [18]. The improvement in efficiency of RESTART over ordinary splitting increases in proportion to the time interval, and so in this paper we focus mainly on this version of splitting, though the theoretical results can be proved for ordinary splitting as well, and in fact with simpler proofs.

We note that there are other particle based methods that produce unbiased estimates for these problems, such as the interacting particle methods developed in [3, 7]. While these methods have a number of demonstrated successes, at the present time there are no general results on their performance in the small probability limit even for fixed time intervals, and for this reason are not discussed in the paper.

The outline of the paper is as follows. In Section 2 we introduce the model, state conditions that will hold throughout the paper and state some preliminary results. Then in Section 3 we determine the decay rate for the probability to escape from a domain within some time $T(n)$, where n is the large deviation parameter and $T(n)$ is allowed to grow polynomially in n. In Section 4 we introduce RESTART and determine the asymptotic decay rate. We conclude in Section 5 with some numerical results.

2. Model and preliminary results. The problem of interest is to estimate exit probabilities of a discrete time process $\{X_i^n\}$ from a bounded open set $\mathcal{D} \subset \mathbb{R}^d$ over a time interval $[0, T(n)]$. The index n serves a dual purpose: we assume that X^n satisfies a large deviations principle (LDP) with rate n, and we also assume $T(n) \to \infty$ as $n \to \infty$.

Our results hold for a wide class of processes, and the techniques of the proof are agnostic to the specific choice of model. All that is required is a sample path LDP which is uniform with respect to the initial condition, as characterized in Definition 2.3 and required by Condition 2.4, and that the local rate function has enough regularity to guarantee that the infima over closed and open sets agree, as in Condition 2.2. However, it will be useful for the purposes of discussion to make some of the quantities involved more concrete. A large class of Markov processes satisfying the assumptions to be imposed can be obtained from recursive chains of the form

$$X_{i+1}^n = X_i^n + \frac{1}{n} v_i(X_i^n), \tag{2.1}$$

where $v_i(\cdot)$ are independent and identically distributed random vectors fields on $\mathbb{R}^d$ whose distribution is given by a stochastic kernel $\mu(\cdot|x), x \in \mathbb{R}^d$, so that $P(v_i(x) \in A) = \mu(A|x)$. To help the reader fix ideas, we provide an example which satisfies the conditions that we provide below.

EXAMPLE 2.1. Consider the small-noise diffusion

$$dX^n(t) = b\big(X^n(t)\big)\,dt + \sqrt{\frac{1}{n}}\sigma\big(X^n(t)\big)\,dB(t), \tag{2.2}$$

where $B(\cdot)$ is a standard d-dimensional Brownian motion, b and σ are Lipschitz continuous, and $\sigma(x)\sigma(x)^T \geq cI$ in the sense of symmetric matrices for some $c > 0$ and all $x \in \mathbb{R}^d$. If we consider the Euler–Maruyama discretization of this process with the convenient step size $1/n$ we obtain the recursion

$$X_{i+1}^n = X_i^n + \frac{1}{n} b(X_i^n) + \frac{1}{n}\sigma(X_i^n)\xi_{i+1},$$

with ξ_i being i.i.d. standard Gaussians, which is of the form (2.1) with $\mu(\cdot|x)$ Gaussian with mean $b(x)$ and covariance matrix $\sigma(x)\sigma(x)^T$.

If we assume also that and the drift generated by $b(\cdot)$ attracts the whole domain $\mathcal{D}$ to an equilibrium point $0 \in \mathcal{D}$, then under boundedness of $\mathcal{D}$ and regularity of its boundary all of the conditions assumed below will be satisfied. (Note that since $\mathcal{D}$ is assumed bounded, we can redefine b and σ in any way that is convenient on the complement of $\mathcal{D}$ when checking if the conditions hold).

A particularly relevant special case of (2.2) for the problem of escape over long time intervals is the Ornstein–Uhlenbeck process, when $b(x) = -Ax$ and $\sigma(x) = \Sigma$ for suitable matrices A and Σ. This is so because under mild conditions such processes provide the approximating diffusion model for (2.1) in a neighborhood of 0 when 0 is a stable equilibrium for the ordinary differential equation with drift $b(x) = \int_{\mathbb{R}^d} y\mu(dy|x)$, and because the difficulties associated with simulating over long times intervals are all due to the behavior in such neighborhoods.

As an example that is not of the form (2.1) one could suppose that the drift $b(x)$ and diffusion $\sigma(x)$ depend on an exogeneous finite-state Markov chain; this would be an example of a "Markov-modulated" process; see [2] for the large deviation theory of such processes. Although we use (2.1) as a canonical model for the discussion, it bears repeating that the specific form of the process is not used in the proofs. The key assumption is that an appropriate uniform LDP is available, as required by Condition 2.4, with appropriate regularity conditions.

Returning to the model in (2.1), the process X^n naturally induces a family of probability measures $\mathbb{P}_x$, where $\mathbb{P}_x(X^n(0) = x) = 1$. We let $\mathbb{E}_x[\cdot]$ denote the expected value with respect to $\mathbb{P}_x$. Define the cumulant generating function of the vector fields by

$$H(x, \alpha) \doteq \log \int_{\mathbb{R}^d} e^{\langle \alpha, u \rangle} \mu(du|x).$$

We assume that the following conditions are satisfied by H and the stochastic kernel μ. As noted in Example 2.1, when verifying the conditions we can at our convenience redefine $\mu(\cdot|x)$ off the closure of $\mathcal{D}$.

CONDITION 2.1.

1. For all $\alpha \in \mathbb{R}^d$, we have $\sup_{x \in \mathbb{R}^d} H(x, \alpha) < \infty$.
2. The map $x \to \mu(\cdot|x)$ is continuous in the topology of weak convergence.

The cumulant generating function $H(x, \alpha)$ is used to define the local rate associated to the system (2.1) via the Legendre–Fenchel transformation. For $(x, \beta) \in \mathbb{R}^d$, let

$$L(x, \beta) \doteq \sup_{\alpha \in \mathbb{R}^d} \{\langle \alpha, \beta \rangle - H(x, \alpha)\}. \tag{2.3}$$

We now provide conditions on the local rate L. For models of the form (2.1), since L is the Legendre–Fenchel transform of a cumulant generating function it is automatic that $L(x, \cdot)$ is convex and that there is a unique $b(x)$ such that $L(x, b(x)) = 0$ for each $x \in \mathbb{R}^d$. The remaining properties below can all be related to properties of the kernel $\mu(\cdot|x)$ (see [1, 8]).

CONDITION 2.2. We assume that $L : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty]$ satisfies the following properties:

1. for every $x \in \mathbb{R}^d$ there is a unique $b(x)$ such that $L(x, b(x)) = 0$, and $b : \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous;

2. for every $x \in \mathbb{R}^d$ $L(x, \cdot)$ is convex;
3. for every compact set K, there is $r \in (0, \infty)$ such that $L(x, \beta) \leq 1/r$ if $x \in K$, $|\beta| \leq r$.

The most significant of these is the last property, which is used to establish that the infimum of the rate function (defined below) over the interior and closure of sets of trajectories coincide. This is not strictly necessary, but without it both the statement of results and their proofs become much more complicated. The second property guarantees the existence and uniqueness of solutions to the initial value problem $\dot{x}(t) = b(x(t))$, $x(0) = x_0$ for any $x_0 \in \mathbb{R}^d$. We call the ordinary differential equation $\dot{x}(t) = b(x(t))$ the *noiseless dynamics*.

To state the LDP for X_i^n, it is convenient to interpolate it as a piecewise constant process. For fixed $T \in (0, \infty)$, define

$$X^n(t) \doteq X_i^n, \quad t \in \big[i/n, (i+1)/n\big), i = 0, \ldots, \lfloor nT \rfloor,$$

where $\lfloor a \rfloor$ is the integer part of a. We consider X^n as taking values in $D([0, T]) \doteq D([0, T] : \mathbb{R}^d)$, with the usual Skorokhod topology [13]. We could also have considered a piecewise linear interpolation that takes values in $C([0, T]) \doteq C([0, T] : \mathbb{R}^d)$ and obtain an LDP with the same rate function, but then we would have to justify the use of the Markov property at only the interpolation times and this seems to lead to more complicated proofs. We will make use of the fact that the Skorokhod topology on $D([0, T])$ relativized to $C([0, T])$ coincides with the uniform topology given by the metric

$$d(\varphi_1, \varphi_2) \doteq \|\varphi_1 - \varphi_2\|_\infty = \sup_{0 \leq t \leq T} \big|\varphi_1(t) - \varphi_2(t)\big|.$$

We omit the implicit dependence of the metric d on the time interval $[0, T]$ for notational simplicity, as it will be clear from the context what the interval is. For any set $A \subset C([0, T])$ and $\varphi \in C([0, T])$, we define $d(\varphi, A) \doteq \inf_{\phi \in A} d(\varphi, \phi)$. For objects taking values in $\mathbb{R}^d$ we use absolute values $|\cdot|$ to denote the standard Euclidean norm, and we use $\mathcal{E}_\mu$ to denote a ball of radius $\mu > 0$ around 0: $\mathcal{E}_\mu \doteq \{x \in \mathbb{R}^d : |x| < \mu\}$.

Under the conditions we will assume, the rate function associated with the process X^n over the interval $[0, T]$ is

$$I_T(\varphi) \doteq \int_0^T L\big(\varphi(s), \dot{\varphi}(s)\big)\, ds$$

if $\varphi(t)$ is absolutely continuous, and $I_T(\varphi) = \infty$ otherwise. Note that, owing to part 1 of Condition 2.2, if $I_T(\varphi) = 0$ then φ satisfies the noiseless dynamics $\dot{\varphi}(t) = b(\varphi(t))$.

We phrase the uniform large deviations principle in terms of level sets of I_T, which are defined by

$$\Phi_{x,T}(s) \doteq \big\{\varphi \in D\big([0, T]\big) : I_T(\varphi) \leq s, \varphi(0) = x\big\}.$$

The formulation of the uniform large deviations principle presented here is taken from [14], p. 74.

DEFINITION 2.3 (Uniform large deviations principle). The sequence X^n satisfies a uniform large deviations principle if:

1. the functional I_T is lower semicontinuous on $D([0, T])$ and for each $T \in [0, \infty)$, compact $K \subset \mathbb{R}^d$ and $s < \infty$, the set $\bigcup_{x \in K} \Phi_{x,T}(s)$ is compact;
2. for any $\delta > 0$, $\gamma > 0$, $s_0 < \infty$ and any compact $K \subset \mathbb{R}^d$, there is $N \in \mathbb{N}$ such that

$$\mathbb{P}_x\big(\|X^n - \phi\|_\infty < \delta\big) \geq \exp\big(-n\big(I_T(\phi) + \gamma\big)\big)$$

for all $n \geq N$, all $x \in K$ and all $\phi \in \Phi_{x,T}(s_0)$;

3. for any $\delta > 0$, $\gamma > 0$, $s_0 < \infty$ and any compact $K \subset \mathbb{R}^d$, there exists $N \in \mathbb{N}$ such that

$$\mathbb{P}_x\big(d\big(X^n, \Phi_{x,T}(s)\big) \geq \delta\big) \leq \exp\big(-n(s-\gamma)\big)$$

for all $n \geq N$, $s \leq s_0$ and $x \in K$.

This definition allows us to phrase the final assumption on the process X_i^n.

CONDITION 2.4. For each $T \in (0, \infty)$, the sequence $\{X^n, n \in \mathbb{N}\}$ satisfies the uniform LDP in Definition 2.3 with a rate function $I_T(\varphi)$ of the form

$$I_T(\varphi) = \int_0^T L\big(\varphi(s), \dot{\varphi}(s)\big)\, ds$$

for some L satisfying Condition 2.2.

Having posed all conditions on this process, we return to our original problem. We seek to estimate the probability that the process escapes from a domain $\mathcal{D}$ before some time $T(n)$. We will require the domain $\mathcal{D}$ to satisfy some mild regularity properties which are stated below. We will make use of the following notation throughout the paper. For any subset A of a topological space, we let $\overline{A}$ denote its closure, A^c its complement, and $\partial A = \overline{A} \cap \overline{A^c}$ its boundary. We denote the first exit time of the process X^n from the set $\mathcal{D}$ by

$$\rho^n \doteq \inf\big\{t \geq 0 : X^n(t) \notin \mathcal{D}\big\}.$$

We assume that the process exits the set $\mathcal{D}$ in finite time with probability 1, that is, for all $x \in \mathcal{D}$, $\mathbb{P}_x(\rho^n < \infty) = 1$.

CONDITION 2.5. We impose the following conditions on the set $\mathcal{D} \subset \mathbb{R}^d$:

1. $\mathcal{D}$ is a bounded open subset of $\mathbb{R}^d$;
2. $\mathcal{D}$ satisfies a regularity condition at all points of its boundary: for any $\delta > 0$ and $p \in \partial\mathcal{D}$, there is a point q in the interior of $\mathcal{D}^c$ with $|p - q| < \delta$;
3. the noiseless dynamics $\dot{x}(t) = b(x(t))$ possess a unique equilibrium point $O \in \mathcal{D}$ which is asymptotically stable on a neighborhood of $\overline{\mathcal{D}}$: there is open $\mathcal{D}' \supset \overline{\mathcal{D}}$ such that for any $x_0 \in \mathcal{D}'$ and $\mu > 0$, there is $T = T(x_0, \mu) \in [0, \infty)$ such that $\dot{x}(t) = b(x(t))$ for $t \geq 0$ and $x(0) = x_0$ together imply $|x(T) - O| < \mu$. Without loss of generality, we take O to be the origin $0 \in \mathbb{R}^d$.

The assumption that $\mathcal{D}$ is bounded is not necessary for the results on the decay rate of RESTART to hold, since one can always restrict to an appropriately chosen bounded subset of the domain, that is, take a compact set $K \subset \mathcal{D}$ large enough that escape outside of K is superexponentially small. Such a generalization is straightforward but cumbersome and is therefore omitted. We also note that the noiseless dynamics may exit the domain when started at certain initial conditions $x_0 \in \mathcal{D}$, though they must eventually re-enter.

With regard to $T(n)$ we assume the following.

CONDITION 2.6. $T(n)$ grows polynomially in n, by which we mean

$$\lim_{n\to\infty} T(n) = \infty \quad \text{and} \quad \lim_{n\to\infty} \big[\log T(n)\big]/n = 0.$$

Throughout the paper several interdependent parameters are used. For instance, a statement which holds for all n greater than some threshold value N that depends on ε. We highlight the important dependencies by explicitly writing the parameters at the outset, for example, $N(\varepsilon)$, and thereafter using only N. We highlight only the most relevant dependencies and ignore others. For instance, N might also depend on the domain $\mathcal{D}$, but since $\mathcal{D}$ is fixed throughout we do not include it in the list of dependencies.

2.1. *Preliminary results.* In this section we will state some results that will be used in Section 3 to estimate the rate of decay of the exit probability from $\mathcal{D}$ by some time $T(n)$. The proofs of the lemmas in this section follow more or less directly from the conditions that were presented in Section 2 and hence these proofs can be found in Appendix A.

The uniform bound on the cumulant generating function $H(x,\alpha)$ allows us to establish an asymptotic bound on the maximum jump size of the process, which is the content of Lemma 2.7. The proof follows from Chebyshev's inequality and since $n^{-1}\log(nT(n)) \to 0$, and is omitted.

LEMMA 2.7. *Assume Condition 2.1. Then for all $\delta > 0$,*

$$\limsup_{n\to\infty} \sup_{x\in\mathbb{R}^d} \frac{1}{n} \log \mathbb{P}_x\Big(\max_{0\le i\le nT(n)} \big| X_{i+1}^n - X_i^n \big| \ge \delta \Big) = -\infty.$$

Next, we define a function which will be used throughout the paper and will be commonly referred to as the "cost to exit," starting at a given point. Let

$$W(x) \doteq \inf\big\{ I_T(\varphi) : \varphi(0) = x, \varphi(T) \notin \mathcal{D}, T < \infty \big\} \tag{2.4}$$

for $x \in \mathcal{D}$ and $W(x) \doteq 0$ otherwise. The function $W(\cdot)$ has the following properties.

LEMMA 2.8. *Under Condition 2.2, the function $W(\cdot)$ satisfies*:

1. *$W(\cdot)$ is continuous on $\mathbb{R}^d$,*
2. *for every $x \in \mathcal{D}$, $W(x) \le W(0)$.*

The last lemma in this section gives an upper bound on the probability that the process does not enter the μ-neighborhood of the origin within some finite time T, starting at some point x that is not in the μ-neighborhood of the origin. The proof of this lemma uses the same argument as [14], Lemma 2.2, Chapter 4.

LEMMA 2.9. *For any $\mu > 0$ and any $M < \infty$, there are $T(\mu, M) < \infty$ and $N(\mu, M) < \infty$ such that for any initial point $x \in \overline{\mathcal{D}} \setminus \mathcal{E}_\mu$, where $\mathcal{E}_\mu = \{x : |x| < \mu\}$, we have*

$$\mathbb{P}_x\big(X^n(t) \in \overline{\mathcal{D}} \setminus \mathcal{E}_\mu, 0 \le t \le T \big) \le e^{-nM}$$

for all $n \ge N$.

3. Long time estimates. The main result of this section is the following estimate.

THEOREM 3.1. *Suppose Conditions 2.1, 2.2, 2.4, 2.5 and 2.6 are met. Then for any $x \in \mathcal{D}$ we have*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_x\big(\rho^n \le T(n) \big) = -W(x). \tag{3.1}$$

Since $T(n) \to \infty$, this is not covered by an LDP of the form in Definition 2.3, and in fact requires a Freidlin–Wentzell analysis appropriate for large time problems as in [14]. Theorem 3.1 follows directly from Lemmas 3.3 and 3.4 below, which give upper and lower bounds for the decay of $\mathbb{P}_x(\rho^n \le T(n))$. We remark that in fact Theorem 3.1 holds uniformly in x, but it is not stated as such since only the upper bound in Lemma 3.3 is proved in its uniform version as it is required for the proof of Theorem 4.7. The lower bound is easier and holds uniformly as well, but since it is not used in the analysis of RESTART we only establish the pointwise estimate in the interest of saving space.

We start with the upper bound when the process starts in a small neighborhood of the attracting point 0.

LEMMA 3.2. *Assume the conditions of Theorem 3.1 are met. For any $\varepsilon > 0$, there is a $\mu(\varepsilon) > 0$ and $N(\varepsilon) < \infty$ such that for all $n \geq N$ and all $y \in \mathcal{E}_\mu = \{x : |x| < \mu\}$,*

$$\frac{1}{n} \log \mathbb{P}_y\big(\rho^n \leq T(n)\big) \leq -W(y) + \varepsilon. \tag{3.2}$$

PROOF. Let $\varepsilon > 0$. The function $y \to W(y)$ is continuous on $\mathcal{D}$ by Lemma 2.8, hence uniformly continuous on any bounded neighborhood of 0. For the given $\varepsilon > 0$, there is $\mu > 0$ such that

$$\big|W(x) - W(y)\big| < \varepsilon/8 \quad \text{whenever } |x - y| < 2\mu. \tag{3.3}$$

We choose such a $\mu > 0$ and any $\eta \in (0, \mu/2)$. We also assume μ small enough that $\mathcal{E}_\mu \subset \mathcal{D}$. Both μ and η will remain fixed for the rest of the proof. For readability we have divided the proof into 5 steps.

Step 1. *A priori estimates*. Define the sets

$$\Gamma_\mu^\eta \doteq \{x : \mu - \eta \leq |x| \leq \mu\},$$
$$\gamma_\mu^\eta \doteq \left\{x : \frac{\mu}{2} - \eta \leq |x| \leq \frac{\mu}{2}\right\}.$$

We will choose a time T_1 such that the probability that the process does not enter the $\mu/2$-neighborhood of the attracting point 0 is superexponentially small. By Lemma 2.9, there are $T_1(\mu)$ and $N_1(\mu)$ such that $T \geq T_1$ and $n \geq N_1$ together imply, for any $x \in \mathcal{D} \setminus \mathcal{E}_{\mu/2}$, that

$$\mathbb{P}_x\big(X^n(t) \in \mathcal{D} \setminus \mathcal{E}_{\mu/2}, t \in [0, T_1]\big) \leq e^{-n(\bar{W}+1)}, \tag{3.4}$$

where $\bar{W} \doteq \inf_{x \in \Gamma_\mu^\eta} W(x)$. Without loss of generality, we may assume that $\bar{W} > \varepsilon$.

We claim there is $\delta(\varepsilon) > 0$ such that any path φ which satisfies $\varphi(0) \in \Gamma_\mu^\eta \subset \mathcal{D}$ and $\varphi(t) \notin \mathcal{D}$ for some $t \in [0, T_1]$ also satisfies

$$d\big(\varphi, \Phi_{\varphi(0),T_1}(\bar{W} - \varepsilon/4)\big) \geq \delta > 0.$$

If not, for any $\delta > 0$ we could find a path φ^δ with $\varphi^\delta(0) \in \Gamma_\mu^\eta$ and $\varphi^\delta(t_*) \notin \mathcal{D}$ for some $t_* \in [0, T_1]$, and $d(\varphi^\delta, \Phi_{\varphi^\delta(0),T_1}(\bar{W} - \varepsilon/4)) < \delta$. By Lemma A.1, on the compact set $K = \{x : \inf_{z \in \mathcal{D}} |x - z| \leq 1\}$ there is $c \in (0, \infty)$ such that for all $x, y \in K$, we can construct φ_{xy} satisfying $\varphi_{xy}(0) = x$, $\varphi_{xy}(\tau) = y$ with $\tau \in (0, \infty)$ and $I_\tau(\varphi_{xy}) \leq c|x - y|$.

Choose $\delta \in (0, \varepsilon/8c)$, and without loss of generality assume $\delta < 1$. We obtain a path φ^δ starting in Γ_μ^η for which $\varphi^\delta(t_*) \notin \mathcal{D}$ at some $t_* \in [0, T_1]$. Under the assumption that $d(\varphi^\delta, \Phi_{\varphi^\delta(0),T_1}(\bar{W} - \varepsilon/4)) < \delta$, we can find $\phi \in \Phi_{\varphi^\delta(0),T_1}(\bar{W} - \varepsilon/4)$ such that $\|\phi - \varphi^\delta\|_\infty < \delta$. Since $\phi(0) = \varphi^\delta(0) \in \Gamma_\mu^\eta$, the definition of $\bar{W}$ implies that $\phi(t_*) \in \mathcal{D} \subset K$, while $\delta < 1$ ensures $\varphi^\delta(t_*) \in K$. Applying Lemma A.1 to the points $x = \phi(t_*)$ and $y = \varphi^\delta(t_*)$, we obtain a path φ_{xy} connecting x and y in time τ with cost $I_\tau(\varphi_{xy}) \leq c\delta < \varepsilon/8$. Concatenating ϕ on $[0, t_*]$ with φ_{xy} on $(t_*, t_* + \tau]$, we obtain a path starting in Γ_μ^η and ending outside of $\mathcal{D}$ at time $t_* + \tau < \infty$, with cost less than $I_{t_*}(\phi) + I_\tau(\varphi_{xy}) \leq \bar{W} - \varepsilon/8$. This contradicts the definition of $\bar{W}$. Thus $\delta > 0$ must exist as claimed.

As a consequence of the above discussion, for any $y \in \Gamma_\mu^\eta$,

$$\big\{\rho^n \leq T_1, X^n(0) = y\big\} \subset \big\{d\big(X^n, \Phi_{y,T_1}(\bar{W} - \varepsilon/4)\big) \geq \delta\big\}.$$

By the uniform large deviations upper bound on the interval $[0, T_1]$, there is $N_2(\varepsilon, T_1) < \infty$ for which $n \geq N_2$ implies, for any $y \in \Gamma_\mu^\eta$,

$$\mathbb{P}_y\big(\rho^n \leq T_1\big) \leq \mathbb{P}_y\big(d\big(X^n, \Phi_{y,T_1}(\bar{W} - \varepsilon/4)\big) \geq \delta\big) \leq e^{-n(\bar{W} - \frac{\varepsilon}{2})}. \tag{3.5}$$

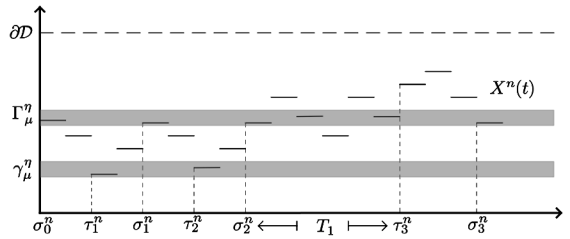

FIG. 1. *The piecewise constant process $X^n(t)$ moving between the sets Γ^η_μ and γ^η_μ. The process in the figure has not exited the domain $\mathcal{D}$ nor hit γ^η_μ at time τ^n_3, since it wandered around for time longer than T_1. Thus, σ^n_3 is triggered when the process returns to Γ^η_μ.*

With T_1 chosen, we introduce the stopping times $\sigma_0 \doteq 0$, and for $j \geq 1$,

$$\tau^n_j \doteq \inf\{t > \sigma^n_{j-1} : X^n(t) \in (\gamma^\eta_\mu \cup \mathcal{D}^c)\} \wedge (T_1 + \sigma^n_{j-1}),$$
$$\sigma^n_j \doteq \inf\{t > \tau^n_j : X^n(t) \in \Gamma^\eta_\mu\}.$$

The stopping times τ^n_j and σ^n_j also depend on η and μ, but we do not include this dependence to avoid an overload of notation. Figure 1 shows a sample path with its associated stopping times.

Step 2. *The probability of an excursion to $\partial\mathcal{D}$ within time $T(n)$ is approximately the probability of excursion over a single time interval $(\sigma^n_{j-1}, \tau^n_j)$ times the expected number of such intervals*. Define the random variables

$$B^n_j \doteq \begin{cases} 1 & \text{if } X^n(\tau^n_j) \notin \mathcal{D} \text{ or } \tau^n_j - \sigma^n_{j-1} \geq T_1, \\ 0 & \text{else.} \end{cases}$$

For any $y \in \Gamma^\eta_\mu$ it holds that

$$\begin{aligned} &\mathbb{P}_y(\rho^n \leq T(n)) \\ &\quad \leq \mathbb{P}_y(B^n_j = 1 \text{ for some } j \text{ with } \tau^n_j \leq T(n)) \\ &\quad = \mathbb{P}_y\left(B^n_j = 1 \text{ for some } j \text{ with } \left(\tau^n_1 - \sigma^n_0 + \sum_{i=2}^{j} \tau^n_i - \tau^n_{i-1}\right) \leq T(n)\right) \\ &\quad \leq \mathbb{P}_y\left(B^n_j = 1 \text{ for some } j \text{ with } \left(\sum_{i=1}^{j} \tau^n_i - \sigma^n_{i-1}\right) \leq T(n)\right) \\ &\quad \leq \mathbb{E}_y\left[\sum_{i=1}^{M^n} B^n_i\right], \end{aligned} \tag{3.6}$$

where $M^n \doteq \inf\{j \geq 1 : \sum_{i=1}^{j} \tau^n_i - \sigma^n_{i-1} > T(n)\}$. We can bound (3.6) by

$$\mathbb{E}_y\left[\sum_{i=1}^{M^n} B^n_i\right] \leq \mathbb{E}_y[M^n] \sup_{x \in \Gamma^\eta_\mu} \mathbb{E}_x[B^n_1] \tag{3.7}$$

for any $y \in \Gamma^\eta_\mu$. The proof of this statement can be found in Appendix B, Lemma B.1.

Step 3. *The expected number of intervals $(\sigma^n_{j-1}, \tau^n_j)$ in time $T(n)$ is approximately $T(n)$ divided by the expected duration of each interval. A lower bound on the expected duration*

gives an upper bound on the expected number of intervals. For any $y \in \Gamma_\mu^\eta$, we have

$$\begin{aligned}\mathbb{E}_y[M^n] \inf_{x\in\Gamma_\mu^\eta} \mathbb{E}_x[\tau_1^n] &\le \mathbb{E}_y\left[\sum_{i=1}^{M^n} \tau_i^n - \sigma_{i-1}^n\right] \\ &\le T(n) + \tau_{M^n}^n - \sigma_{M^n-1}^n \\ &\le T(n) + T_1.\end{aligned}$$

See Lemma B.1 for the proof of the first inequality.

It takes some positive time to travel from Γ_μ^η to either γ_μ^η or to escape $\mathcal{D}$, so $\mathbb{E}_x[\tau_1^n] \ge s > 0$ for some fixed s independent of $y \in \Gamma_\mu^\eta$ and all sufficiently large n. A precise proof can be found in Lemma B.2. Hence, for all $y \in \Gamma_\mu^\eta$,

$$\mathbb{E}_y[M^n] \le \frac{T(n) + T_1}{s}. \tag{3.8}$$

Step 4. *Combine estimates for* $y \in \Gamma_\mu^\eta$. From equations (3.6) and (3.7), we find for any $y \in \Gamma_\mu^\eta$

$$\mathbb{P}_y(\rho^n \le T(n)) \le \mathbb{E}_y[M^n]\Big(\sup_{x\in\Gamma_\mu^\eta} \mathbb{E}_x[B_1^n]\Big). \tag{3.9}$$

The first factor in (3.9) was bounded above in (3.8). For the second factor, note that for any $y \in \Gamma_\mu^\eta$,

$$\mathbb{E}_y[B_1^n] = \mathbb{P}_y(B_1^n = 1) = \mathbb{P}_y\big(\{X^n(\tau_1) \notin \mathcal{D}, \tau_1 \le T_1\} \cup \{\tau_1 \ge T_1\}\big). \tag{3.10}$$

We consider the two events separately and then bound the probability of the union by the sum of the probabilities. The probability of the first event is upper bounded in (3.5) for $n \ge N_2$. For the second event, note that

$$\{\tau_1 \ge T_1\} \subset \big\{X^n(t) \in \mathcal{D} \setminus \mathcal{E}_{\mu/2}, t \in [0, T_1]\big\} \cup \Big\{\max_{0\le i\le nT(n)} \big|X_{i+1}^n - X_i^n\big| \ge \eta\Big\}$$

for if the process did enter the $\mu/2$-neighborhood of the origin but did not trigger the event τ_1, then at some time index $i = 0, \dots, nT(n)$ it must have jumped over the set γ_μ^η. From Lemma 2.7, we can choose $N_3(\mu, \eta) < \infty$ for which $n \ge N_3$ implies

$$\mathbb{P}_y\Big(\max_{0\le i\le nT(n)} \big|X_{i+1}^n - X_i^n\big| \ge \eta\Big) \le e^{-n(\bar{W}+1)}. \tag{3.11}$$

The event $\{X^n(t) \in \mathcal{D} \setminus \mathcal{E}_{\mu/2}, t \in [0, T_1]\}$ is also superexponentially small, as the constant T_1 was originally chosen so that any $y \in \Gamma_\mu^\eta \subset (\mathcal{D} \setminus \mathcal{E}_{\mu/2})$ satisfies (3.4) for all $n \ge N_1$.

Set $N_4 = \max(N_1, N_2, N_3)$. Then, by (3.4), (3.5), (3.10) and (3.11), for any $n \ge N_4$ and any $y \in \Gamma_\mu^\eta$

$$\begin{aligned}\mathbb{E}_y[B_1^n] &\le \mathbb{P}_y(\rho^n \le T_1) + \mathbb{P}_y\big(X^n(t) \in \mathcal{D} \setminus \mathcal{E}_{\mu/2}, t \in [0, T_1]\big) \\ &\quad + \mathbb{P}_y\Big(\max_{0\le i\le nT(n)} \big|X_{i+1}^n - X_i^n\big| \ge \eta\Big) \\ &\le e^{-n(\bar{W}-\frac{\varepsilon}{2})} + 2e^{-n(\bar{W}+1)}.\end{aligned}$$

From (3.8) and (3.9) it follows that

$$\mathbb{P}_y(\rho^n \le T(n)) \le \frac{T(n) + T_1}{s}\big(e^{-n(\bar{W}-\frac{\varepsilon}{2})} + 2e^{-n(\bar{W}+1)}\big).$$

We can now take logarithms and scale by n. By using Condition 2.6 we can choose $N_5 \in (N_4, \infty)$, so that $n \geq N_5$ implies

$$\frac{1}{n}\log \mathbb{P}_y\big(\rho^n \leq T(n)\big) \leq -\bar{W} + \frac{\varepsilon}{2} + \frac{\varepsilon}{8}.$$

The additional error term of $\varepsilon/8$ appears to account for the polynomial factor and the additional exponential term which decays at rate $\bar{W} + 1$.

Since $|x - y| < 2\mu$ whenever $x, y \in \Gamma_\mu^\eta$, (3.3) ensures that $-\bar{W} \leq -W(y) + \varepsilon/8$. We conclude that whenever $n \geq N_5$ and $y \in \Gamma_\mu^\eta$, we have

$$\frac{1}{n}\log \mathbb{P}_y\big(\rho^n \leq T(n)\big) \leq -\bar{W} + \frac{\varepsilon}{2} + \frac{\varepsilon}{8} \leq -W(y) + \frac{3\varepsilon}{4}. \tag{3.12}$$

In particular, (3.2) holds for $y \in \Gamma_\mu^\eta$.

Step 5. *Extend estimate to* $y \in \mathcal{E}_\mu$. The estimate (3.12), which we have just shown holds for all $y \in \Gamma_\mu^\eta$, can be harnessed to obtain (3.2) for all $y \in \mathcal{E}_\mu$. Let $z \in \mathcal{E}_\mu \setminus \Gamma_\mu^\eta$, and define the stopping time

$$H_\Gamma^n \doteq \inf\big\{t \geq 0 : X^n(t) \in \Gamma_\mu^\eta\big\}.$$

In order to escape $\mathcal{D}$, a trajectory starting at z must pass through Γ_μ^η, or else jump over it. In the latter case we have

$$\mathbb{P}_z\big(\rho^n < H_\Gamma^n\big) \leq \mathbb{P}_z\Big(\max_{i=0,\dots,nT(n)} \big|X_{i+1}^n - X_i^n\big| \geq \eta\Big).$$

Owing to Lemma 2.7, the probability of jumping over Γ_μ^η is superexponentially small: we can choose $N_6(\mu, \eta) < \infty$ so large that

$$\mathbb{P}_z\Big(\max_{i=0,\dots,nT(n)} \big|X_{i+1}^n - X_i^n\big| \geq \eta\Big) \leq e^{-n(\bar{W}+1)} \tag{3.13}$$

for all $n \geq N_6$. By the law of total probability,

$$\begin{aligned}\mathbb{P}_z\big(\rho^n \leq T(n)\big) &= \mathbb{P}_z\big(\rho^n \leq T(n) \mid \rho^n < H_\Gamma^n\big)\mathbb{P}_z\big(\rho^n < H_\Gamma^n\big) \\ &\quad + \mathbb{P}_z\big(\rho^n \leq T(n) \mid \rho^n > H_\Gamma^n\big)\mathbb{P}_z\big(\rho^n > H_\Gamma^n\big) \\ &\leq e^{-n(\bar{W}+1)} + \mathbb{P}_{X^n(H_\Gamma^n)}\big(\rho^n \leq T(n) - H_\Gamma^n\big),\end{aligned} \tag{3.14}$$

where the last step follows from the strong Markov property applied at H_Γ^n. Since $X^n(H_\Gamma^n) \in \Gamma_\mu^\eta$ by definition of H_Γ^n, and since $H_\Gamma^n \geq 0$, it also holds that

$$\mathbb{P}_{X^n(H_\Gamma^n)}\big(\rho^n \leq T(n) - H_\Gamma^n\big) \leq \sup_{y \in \Gamma_\mu^\eta} \mathbb{P}_y\big(\rho^n \leq T(n)\big).$$

Recall that the choice of $\mu > 0$ in (3.3) ensures $W(z) \leq W(y) + \varepsilon/8$. Using (3.12) in (3.14), we obtain for the given $z \in \mathcal{E}_\mu \setminus \Gamma_\mu^\eta$ and $n \geq \max(N_5, N_6)$ that

$$\mathbb{P}_z\big(\rho^n \leq T(n)\big) \leq e^{-n(\bar{W}+1)} + e^{-n(W(y)-\frac{3\varepsilon}{4})} \leq 2e^{-n(W(z)-7\varepsilon/8)}.$$

By choosing n large enough and applying the logarithmic scaling, we can absorb the coefficient of 2 into the remaining $\varepsilon/8$ of room for error. This establishes the desired estimate for $z \in \mathcal{E}_\mu \setminus \Gamma_\mu^\eta$ and thus for all $y \in \mathcal{E}_\mu$. □

The next lemma extends the asymptotic upper bound to all points in $\mathcal{D}$. It is stated in its uniform version as required for the proof of Theorem 4.7.

LEMMA 3.3. *Under the conditions of Theorem 3.1,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}_z\big(\rho^n \le T(n)\big) \le -W(z) \tag{3.15}$$

uniformly in $z \in \mathcal{D}$.

PROOF. To establish the lemma, we claim it suffices to show that for any $z \in \mathcal{D}$ and for any $\varepsilon > 0$, there exist $\alpha(z,\varepsilon) > 0$ and $N(z,\varepsilon) < \infty$ such that

$$\frac{1}{n} \log \mathbb{P}_y\big(\rho^n \le T(n)\big) \le -W(y) + \varepsilon \tag{3.16}$$

for all $n \ge N(z,\varepsilon)$ and $y \in B_{\alpha(z,\varepsilon)}(z) \doteq \{y : |z-y| < \alpha(z,\varepsilon)\}$.

To justify the claim, first observe that for any choice of $\alpha(z,\varepsilon) > 0$, we have $\overline{\mathcal{D}} \subset \bigcup_{z\in\mathcal{D}} B_{\alpha(z,\varepsilon)}(z)$. Since $\overline{\mathcal{D}}$ is compact, there are $K < \infty$ and $\{z_i\}_{i=1}^K \subset \mathcal{D}$ such that $\overline{\mathcal{D}} \subset \bigcup_{i=1}^K B_{\alpha(z_i,\varepsilon)}(z_i)$. If $N \doteq \max_{i=1}^K N(z_i,\varepsilon) < \infty$, then (3.16) holds for any $n \ge N$ and any $y \in \mathcal{D}$. Since $\varepsilon > 0$ is arbitrary, (3.15) follows.

We now establish (3.16). The particular case of $z = 0$ is covered by Lemma 3.2, for which one can simply take $\alpha = \mu$, so it suffices to consider $z \in \mathcal{D} \setminus \{0\}$.

Next, we eliminate the case where $W(z) = 0$. Since W is continuous on $\mathbb{R}^d$, there is $\alpha(z,\varepsilon) > 0$ such that $0 < -W(y) + \varepsilon$ for all $y \in B_\alpha(z)$. Since $\mathbb{P}_y(\rho^n \le T(n)) \le 1$, it automatically follows that (3.16) holds for all $y \in B_\alpha(z)$, as required. Furthermore, we observe that $W(z) = 0$ for any $z \notin \mathcal{D}$, and so we may focus on points $z \in \mathcal{D} \setminus \{0\}$.

Let then $z \in \mathcal{D} \setminus \{0\}$ with $W(z) > 0$, and let $\varepsilon > 0$ be given. Without any loss of generality, we may assume that $0 < \varepsilon < W(z)$. By Lemma 3.2, there is $\mu_0(\varepsilon) > 0$ and $N_1(\varepsilon) < \infty$ such that $n \ge N_1$ implies, for all $y \in \mathcal{E}_{\mu_0} = \{x : |x| < \mu_0\}$, that

$$\frac{1}{n} \log \mathbb{P}_y\big(\rho^n \le T(n)\big) \le -W(y) + \frac{\varepsilon}{4}. \tag{3.17}$$

We claim that we can choose $\mu(z,\varepsilon) \in (0,\mu_0)$ and $\alpha(z,\varepsilon)$ such that: (1) $B_\alpha(z) \subset \mathcal{D} \setminus \mathcal{E}_\mu$ and (2) we have the inequality

$$W(y) \le \inf_{x\in\mathcal{E}_\mu} W(x) + \frac{\varepsilon}{2} \tag{3.18}$$

for all $y \in B_\alpha(z)$.

We first choose μ. By part 2 of Lemma 2.8 we have $W(z) \le W(0)$, and by the first part of the same lemma W is continuous at 0. Thus we can choose $\mu > 0$ such that $W(0) < W(x) + \varepsilon/4$ for all $x \in \mathcal{E}_\mu$, and without loss of generality $\mu < \min(|z|/2, \mu_0)$. In particular,

$$W(z) \le \inf_{x\in\mathcal{E}_\mu} W(x) + \frac{\varepsilon}{4}. \tag{3.19}$$

Next, since $\mu < |z|/2$ we can take $\alpha(z,\varepsilon,\mu) > 0$ small enough that $B_\alpha(z) \cap \mathcal{E}_\mu = \varnothing$ and moreover we may insist that $B_\alpha(z) \subset \mathcal{D}$ because $z \in \mathcal{D}$ and $\mathcal{D}$ is open. Since W is continuous at z, α can be taken even smaller to guarantee

$$\big|W(y_1) - W(y_2)\big| < \varepsilon/4 \tag{3.20}$$

for all $y_1, y_2 \in B_\alpha(z)$. Note that μ depends only on z and ε, so α inherently depends only on z and ε as well. Combined with (3.19), an application of the triangle inequality shows that this choice of α guarantees (3.18).

Define

$$\zeta_\mu^n \doteq \inf\big\{t > 0 : X^n(t) \notin \mathcal{D} \setminus \mathcal{E}_\mu\big\},$$

which is the first time the process enters the μ-neighborhood of the attracting point 0 or escapes the set $\mathcal{D}$. By Lemma 2.9, there are $T_1(\mu) < \infty$ and $N_2(\mu) < \infty$ so large that $n \geq N_2$ implies, for any $x \notin \mathcal{E}_\mu$,

$$\mathbb{P}_x\big(\zeta_\mu^n > T_1\big) \leq e^{-n(W(x)+1)}. \tag{3.21}$$

Then, for any $y \in B_\alpha(z)$,

$$\begin{aligned}\mathbb{P}_y\big(\rho^n \leq T(n)\big) &= \mathbb{P}_y\big(\rho^n \leq T_1\big) + \mathbb{P}_y\big(T_1 < \rho^n \leq T(n), \zeta_\mu^n \leq T_1\big) \\ &\quad + \mathbb{P}_y\big(T_1 < \rho^n \leq T(n), \zeta_\mu^n > T_1\big).\end{aligned} \tag{3.22}$$

We estimate each term separately. For the first term, we use the finite time uniform large deviations upper bound, part 3 of Definition 2.3. Let $\bar{W} \doteq \inf_{y \in B_\alpha(z)} W(y)$. Since the closure of $B_\alpha(z)$ is compact, an argument similar to the one in Step 1 of Lemma 3.2 shows that for the given $\varepsilon > 0$ there is $\delta(\varepsilon) > 0$ such that, for all $y \in B_\alpha(z)$,

$$\big\{\rho^n \leq T_1, X^n(0) = y\big\} \subset \big\{d\big(X^n, \Phi_{y,T_1}(\bar{W} - \varepsilon/8)\big) \geq \delta\big\}.$$

Applying the large deviations upper bound on the compact set $\overline{B_\alpha(z)}$ with $s_0 = \bar{W} - \varepsilon/8$, $\gamma = \varepsilon/8$ and $\delta > 0$ as above, we find $N_3(\varepsilon, T_1) < \infty$ for which $n \geq N_3$ implies, for all $y \in B_\alpha(z)$, $\mathbb{P}_y(\rho^n \leq T_1) \leq e^{-n(\bar{W}-\varepsilon/4)}$. By the choice of α in (3.20) we have $W(y) \leq \bar{W} + \varepsilon/4$, and so

$$\mathbb{P}_y\big(\rho^n \leq T_1\big) \leq e^{-n(\bar{W}-\varepsilon/4)} \leq e^{-n(W(y)-\varepsilon/2)} \tag{3.23}$$

for all $y \in B_\alpha(z)$.

For the second term, note that on the event $\{T_1 < \rho^n \leq T(n)\} \cap \{\zeta_\mu^n \leq T_1\}$, the process has entered the μ-neighborhood of 0, which is contained in the μ_0-neighborhood of 0 by the choice of μ. After entering $\mathcal{E}_\mu$ at time ζ_μ^n, it has $T(n) - \zeta_\mu^n$ time remaining to exit $\mathcal{D}$. By allowing $T(n)$ time we increase the probability of exiting. Thus, by (3.17) and the strong Markov property, $n \geq N_1$ implies

$$\begin{aligned}\mathbb{P}_y\big(T_1 < \rho^n \leq T(n), \zeta_\mu^n < T_1\big) &\leq \sup_{x \in \mathcal{E}_\mu} \mathbb{P}_x\big(\rho^n \leq T(n)\big) \\ &\leq e^{-n(\inf_{x \in \mathcal{E}_\mu} W(x) - \varepsilon/4)}.\end{aligned} \tag{3.24}$$

For the last term in (3.22), we use the estimate (3.21),

$$\mathbb{P}_y\big(T_1 < \rho^n \leq T(n), \zeta_\mu^n > T_1\big) \leq \mathbb{P}_y\big(\zeta_\mu^n > T_1\big) \leq e^{-n(W(y)+1)}, \tag{3.25}$$

which holds for all $y \in B_\alpha(z) \subset \mathcal{D} \setminus \mathcal{E}_\mu$ whenever $n \geq N_2$.

Set $N_4 \doteq \max(N_1, N_2, N_3)$. Using (3.23), (3.24) and (3.25) in (3.22), we find for all $n \geq N_4$ and $y \in B_\alpha(z)$

$$\mathbb{P}_y\big(\rho^n \leq T(n)\big) \leq e^{-n(W(y)-\varepsilon/2)} + e^{-n(\inf_{x \in \mathcal{E}_\mu} W(x) - \varepsilon/4)} + e^{-n(W(y)+1)}. \tag{3.26}$$

From (3.18), for any $y \in B_\alpha(z)$, $e^{-n(\inf_{x \in \mathcal{E}_\mu} W(x) - \varepsilon/4)} \leq e^{-n(W(y)-3\varepsilon/4)}$ for all $n \geq N_4$. It therefore follows from (3.26) that, for all $y \in B_\alpha(z)$,

$$\frac{1}{n} \log \mathbb{P}_y\big(\rho^n \leq T(n)\big) \leq -W(y) + \frac{3\varepsilon}{4} + \frac{3}{n},$$

and choosing $N_4 \leq N(z, \varepsilon) < \infty$ large enough that $3/n < \varepsilon/4$, we obtain (3.16) for all $y \in B_\alpha(z)$, as required. □

Next we prove the lower bound for the decay of $\mathbb{P}_x(\rho^n \leq T(n))$. We remind the reader that Lemma 3.4 also holds uniformly in x, but in the interest of space we only establish the pointwise version.

LEMMA 3.4. *Assume the conditions of Theorem 3.1. For any $x \in \mathcal{D}$, we have*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}_x\big(\rho^n \le T(n)\big) \ge -W(x).$$

PROOF. It suffices to show that for any $x \in \mathcal{D}$ and any $\varepsilon > 0$, there is $N(\varepsilon) < \infty$ such that $n \ge N$ guarantees

$$\frac{1}{n} \log \mathbb{P}_x\big(\rho^n \le T(n)\big) \ge -W(x) - \varepsilon.$$

To establish the estimate, we will take a trajectory which is within ε of the infimum in the definition of $W(x)$, and use part 3 of Condition 2.2 in conjunction with part 2 of Condition 2.5 to extend the trajectory a little past the boundary. This will allow us to apply the large deviations lower bound, which is stated in its uniform version as part 2 of Definition 2.3.

Fix $x \in \mathcal{D}$ and let $\varepsilon > 0$. By the variational definition of $W(x)$, there is a trajectory φ and $T < \infty$ such that $\varphi(0) = x$, $\varphi(T) \notin \mathcal{D}$ and $I_T(\varphi) \le W(x) + \varepsilon/4$. Define the compact set $K = \{y : \inf_{z\in\mathcal{D}} |y - z| \le 1\}$. By Lemma A.1, there is $c \in (0, \infty)$ such that for all $p, q \in K$, we can find φ_{pq} and τ which satisfy $\varphi_{pq}(0) = p$, $\varphi_{pq}(\tau) = q$, $\tau \in (0, \infty)$ and $I_\tau(\varphi_{pq}) \le c|p - q|$. Let $a \in (0, \varepsilon/4c)$ and assume without loss of generality that $a < 1$.

$I_T(\varphi) < \infty$ implies φ must be absolutely continuous, so there is a time $s \in (0, T]$ such that $\varphi(s) \in \partial\mathcal{D}$. Set $p \doteq \varphi(s)$ and note that $p \in K$. By the regularity property in part 2 of Condition 2.5, there is a point q in the interior of $\mathcal{D}^c$ which satisfies $|p - q| < a$, and since $a < 1$ we also have $q \in K$. From Lemma A.1 we obtain a trajectory φ_{pq} and $\tau > 0$ as described above with cost

$$I_\tau(\varphi_{pq}) \le ca < \varepsilon/4.$$

Let ϕ denote the concatenation of the paths φ on $[0, s]$ and φ_{pq} on $(s, s + \tau]$, that is,

$$\phi(t) = \begin{cases} \varphi(t), & t \in [0, s], \\ \varphi_{pq}(t - s), & t \in (s, s + \tau], \end{cases}$$

and note that ϕ is continuous at s. Moreover,

$$I_{s+\tau}(\phi) \le I_T(\varphi) + I_\tau(\varphi_{pq}) \le W(x) + \varepsilon/2.$$

Finally, since q lies in the interior of $\mathcal{D}^c$, there is $\delta > 0$ such that $|q - z| < \delta$ implies $z \in \mathcal{D}^c$. In particular, if $\|\phi' - \phi\|_\infty < \delta$, then $\phi'(s + \tau) \in \mathcal{D}^c$. If n is large enough to guarantee $T(n) \ge s + \tau$, then

$$\{\|X^n - \phi\|_\infty < \delta\} \subset \{\rho^n \le T(n)\}.$$

Applying part 2 of Definition 2.3 with $s_0 = W(x) + 1$, $\gamma = \varepsilon/2$, and the given $\delta > 0$, we find $N < \infty$ such that for all $n \ge N$,

$$\mathbb{P}_x\big(\|X^n - \phi\|_\infty < \delta\big) \ge \exp\big(-n\big(I_T(\phi) + \varepsilon/2\big)\big).$$

Taking logarithms and using $I_T(\phi) \le W(x) + \varepsilon/2$, we obtain the desired result. □

4. RESTART. In this section we describe how splitting can be used to estimate the probability that the process X^n exits before time $T(n)$ using simulation. In [11], importance sampling is used to estimate the probability that X^n exits within some fixed finite time T, and the decay rate of the second moment is analyzed in the limit as $n \to \infty$. In Section 4.3 we discuss why RESTART (and ordinary splitting) is preferred over importance sampling when the time grows polynomially with the large deviations scaling parameter n. Until Section 4.2, we will omit the dependence on the large deviations scaling parameter for simplicity.

Both ordinary splitting and RESTART are based on the following principle. If the evolution of the process seems "promising", we *split* the trajectory into multiple ones. After splitting the process, all copies are independent and moreover, they evolve according to the same dynamics as the original process. The splitting of trajectories happens at certain thresholds, which we explicitly define later, and the number of trajectories it splits into is called the splitting rate R. Note that we do not need to keep track of the full trajectories. Instead, we only need to keep track of the positions of the underlying branching process, so that at each time step i we have a labeled list of locations we refer to as *particles*.

With ordinary splitting schemes for exit problems on a finite time interval $[0, T]$, one simulates particles until one of the following two conditions are met: either the particle exits the domain of interest, or the maximum time T is reached (without exit from the domain). When T is large, in the latter case one spends significant computational effort to simulate particles which contribute little to the estimate relative to their simulation cost.

The RESTART algorithm overcomes this issue by "killing" particles of this type once they fall outside the threshold in which they were created. To compensate for these lost particles, splitting occurs whenever a particle enters a threshold, even if it has been previously visited by that particle. It turns out that this compensation is exactly what is needed to make the estimator unbiased. The proof of unbiasedness is not trivial and can be found in [6]. In this paper, we will focus on the RESTART algorithm, since for the problem of interest killing of particles is used more and more as the time gets larger. Thus, for the problem of interest RESTART has computational advantages over ordinary splitting and moreover, it is more complicated to analyze than ordinary splitting.

The main challenge in defining splitting schemes is the selection of the thresholds, that is, where splitting takes place. The splitting scheme we define will be phrased in terms of an importance function U, which is a continuous mapping $U : \mathbb{R}^d \to \mathbb{R}$ that is bounded from below. Only the relative values of $U(x)$ at different points matter, so we assume for simplicity that $U(x) \geq 0$ for all $x \in \mathbb{R}^d$.

The splitting thresholds are based on closed sets C_j, for $0 \leq j \leq J - 1 \doteq \lceil U(0)/\Delta \rceil - 1$, which are defined by

$$C_j \doteq \{x \in \mathcal{D} : U(x) \leq j\Delta\}, \tag{4.1}$$

and $C_J \doteq \mathcal{D}$, where $\Delta \doteq \log(R)$. We define the function $\sigma(x) \doteq j$ when $x \in C_j \setminus C_{j-1}$. Furthermore, we define a piecewise constant approximation to the importance function U by

$$\bar{U}(x) \doteq j\Delta \quad \text{for } x \in C_j \setminus C_{j-1}$$

for $0 \leq j \leq J$, where $C_{-1} \doteq \varnothing$ by convention.

Splitting occurs whenever a threshold with a lower index is entered. It may happen that several thresholds are crossed at one step, even though this may be unlikely. In that case, it is important to know which killing thresholds to assign to the newly created particles. As in [6], we let $q_\ell(j, k)$ denote the number of new particles that are given killing threshold ℓ, assuming the particle moves from $x \in C_j \setminus C_{j-1}$ to $x \in C_k \setminus C_{k-1}$ for $k < j$.

Suppose we have a splitting rate R, that is, the number of descendants from each particle is $R - 1$ when crossing one threshold. If the particle moves from $x \in C_j \setminus C_{j-1}$ to $x \in C_k \setminus C_{k-1}$ we define

$$q_\ell(j, k) \doteq \begin{cases} 0 & \text{if } j \leq k, \\ (R-1)R^{j-\ell-1} & \text{if } k \leq j-1 \text{ and } k \leq \ell \leq j-1. \end{cases}$$

This is the correct assignment of thresholds for unbiasedness, and is what one would get if the particle were moved past the $j - k$ thresholds sequentially (but in zero time), producing $R - 1$ descendents in each. If a particle escapes $\mathcal{D}$ at some time $i \leq T$ then it is killed on the

next step, so that there is only one contribution from that particle to the estimator (4.2). Note that this killing is distinct from the killing used to improve efficiency.

Before precisely defining the estimator and the RESTART algorithm, we need to introduce some more notation. Let N_i be the number of particles at time i, $Z_{i,j}$ the position of the jth particle at time i, $C_{i,j}$ the current threshold of the jth particle at time i and $K_{i,j}$ the killing threshold of the jth particle at time i. We define the unnormalized empirical measure $\bar{\delta}_{Z_i}$ by

$$\bar{\delta}_{Z_i} \doteq \sum_{j=1}^{N_i} \delta_{Z_{i,j}},$$

and the estimator γ then equals

$$\gamma \doteq \sum_{i=0}^{T} \int_{\mathbb{R}^d} 1_{\mathcal{D}^c}(x) e^{\bar{U}(x)-\bar{U}(0)} \bar{\delta}_{Z_i}(dx). \tag{4.2}$$

Note that the weight $\exp(\bar{U}(x)-\bar{U}(0))$ in (4.2) equals R^{j-J} when $x \in D_j = C_j \setminus C_{j-1}$. The state-dependence of $1_{\mathcal{D}^c}(x)e^{\bar{U}(x)-\bar{U}(0)}$ in (4.2) is to account for the fact that the process can exit $\mathcal{D}$ at different thresholds.

The RESTART Algorithm can be found in Appendix C, and can also be found in, for example, [1] and [6]. Note that the pseudocode represents a "parallel" version of the algorithm, since the particles for a given threshold are split and then simulated until it is terminated, which occurs when the particle reaches the next threshold, the target set, or is killed. One could also implement a "sequential" version in which a particle is simulated until it either reaches the target set or is killed, while recording the times and locations of splitting. Once the particle is terminated, the algorithm returns to the highest threshold for which particles remain to be simulated and starts anew.

4.1. *Subsolutions*. We recall that the proposed method for generating importance functions is through a function $\bar{V}$ which will satisfy a subsolution property associated to the underlying dynamics of the sequence of processes $\{X^n\}$.

To motivate the use of a subsolution, one can ask what kind of properties one seeks in a general importance function U. Recall that U specifies the splitting thresholds via (4.1). If U decreases too slowly then splitting occurs infrequently and the scheme is not very helpful, while if U decreases too rapidly then splitting occurs too often and there is an exponential growth in the number of particles, which makes the scheme impossible to simulate.

Let π_j denote the entrance distribution on the threshold C_j, and let p denote the probability of a particle distributed according to π_j reaching C_{j-1} before being killed. Out of R independent particles distributed according to π_j, we expect Rp particles to make it to C_{j-1}, so choosing $Rp \le 1$ avoids an explosion in the number of particles. When X^n satisfies a large deviations principle, for fixed $n \ge 1$ we have the heuristic approximation $p \approx e^{-nI_T(\varphi_j)}$, where φ_j is a trajectory satisfying $\varphi_j(0) \in \partial C_j$ and $\varphi_j(T) \in C_{j-1}$ with minimal cost. If we scale the splitting rate as $R = e^{n\Delta}$, then we want $e^{n\Delta - nI_T(\varphi_j)} \le 1$. The definition of threshold implies $U(\varphi_j(0)) - U(\varphi_j(T)) \approx \Delta$. Taking logarithms in $Rp \le 1$ and dividing by n, we obtain

$$U\big(\varphi_j(0)\big) - U\big(\varphi_j(T)\big) \le I_T(\varphi_j).$$

This rough estimate conveys why subsolutions as defined in Definition 4.1 may be appropriate importance functions. It also suggests the appropriate scaling of thresholds which will be used in the following section.

In general one could consider subsolutions that depend on time, in which case one would treat time as a state variable and split accordingly. However, when the time interval is large

one is tempted to use a simpler subsolution that does not depend on time. Although as mentioned previously such time-independent subsolutions are not effective for importance sampling [11], see also Section 4.3, we recall that a goal of the present paper is to show that they in fact work well for splitting, and in particular for the RESTART implementation.

The subsolution property is phrased in terms of a calculus of variations problem. The problem data are the local rate L and the escape set $\mathcal{D}$. For the phrasing in terms of a corresponding partial differential equation see [1, 5, 12], as well as Section 5.

DEFINITION 4.1. A continuous function $\bar{V} : \mathbb{R}^d \to \mathbb{R}$ is a subsolution if it is bounded from below and

$$\bar{V}(x) \le \inf_{\varphi(0)=x;\varphi(T)\notin\mathcal{D};\varphi(s)\in\mathcal{D},s\in(0,T),T<\infty} \left\{ \int_0^T L\big(\varphi(s),\dot{\varphi}(s)\big)\,ds + \bar{V}\big(\varphi(T)\big) \right\} \tag{4.3}$$

for all $x \in \mathcal{D}$, and $\bar{V}(z) \le 0$ for $z \notin \mathcal{D}$, where the infimum is taken over all absolutely continuous trajectories.

A subsolution $\bar{V}$ is never greater than the solution to the corresponding calculus of variations problem, which is given by $W(x)$ as defined in (2.4), with $W(x) = 0$ for $x \notin \mathcal{D}$. In particular, $\bar{V}(x) \le W(x)$ for all $x \in \mathcal{D}$.

It turns out that we can make a particularly convenient choice of subsolution which is appropriate for the problem of escape from $\mathcal{D}$ over a long time interval. To do so, we define the notion of a quasipotential.

DEFINITION 4.2. The quasipotential is the function $Q : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty]$ defined for $y, x \in \mathbb{R}^d$ by

$$Q(y,x) \doteq \inf_{\varphi(0)=y,\varphi(T)=x,T<\infty} \left\{ \int_0^T L\big(\varphi(s),\dot{\varphi}(s)\big)\,ds \right\},$$

where the infimum is over all absolutely continuous trajectories. The quasipotential with respect to the unique attracting point 0 is $Q(x) \doteq Q(0,x)$.

We note that $Q(x) < \infty$ for $x \in \mathcal{D}$, and that under Condition 2.2, $Q(x)$ is a continuous function—this is proved similarly to the way it is proved for W in Lemma 2.8. Finally, it follows from the definitions of Q and W that

$$\min_{x\in\partial\mathcal{D}} Q(x) = W(0). \tag{4.4}$$

The quasipotential is related to the subsolution property in the following way.

LEMMA 4.3. *For any constant $c \le \inf_{x\notin\mathcal{D}} Q(x)$, $\bar{V}(x) = -Q(x) + c$ is a subsolution in the sense of Definition 4.1.*

PROOF. First consider the case $x \notin \mathcal{D}$. Note that the infimum of Q on the boundary of $\mathcal{D}$ agrees with its infimum on the complement of $\mathcal{D}$, since the positivity of $L(x,\beta)$ ensures that any point in $\mathcal{D}^c$ cannot have lower cost than a point in $\mathcal{D}$. If $c \le \inf_{x\notin\mathcal{D}} Q(x)$, then for $x \notin \mathcal{D}$, $c - Q(x) \le 0$ and therefore $c - Q(x)$ satisfies the boundary condition.

For $x \in \mathcal{D}$, observe that Definition 4.1 for the choice $\bar{V}(x) = -Q(x) + c$ is equivalent to

$$-Q(x) \le \inf_{\varphi(0)=x;\varphi(T)\notin\mathcal{D};\varphi(s)\in\mathcal{D},s\in(0,T),T<\infty} \big\{ I_T(\varphi) - Q\big(\varphi(T)\big) \big\}.$$

If we temporarily assume the infimum is achieved and write $Q(\varphi(T)) \le Q(x) + I_T(\varphi)$, we can heuristically interpret this as saying that the cost to exit starting at 0 must be less than the cost to go from 0 to x and then exit starting from x.

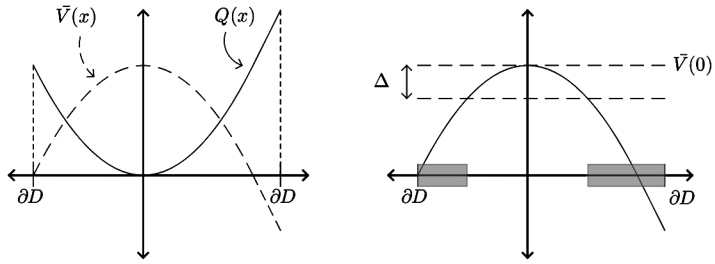

FIG. 2. *Left figure*: *The quasipotential* $Q(x)$ *for a simple one-dimensional domain, and* $\bar{V}(x) = -Q(x) + \min_{x\in\partial\mathcal{D}} Q(x)$. *Right figure*: *The dashed level lines indicate a spacing of* Δ, *starting from* $\bar{V}(0) = J\Delta$. *The shaded region indicates the set of* $x \in C_{J-1} = \{x : \bar{V}(0) - \bar{V}(x) \geq \Delta\}$.

To make this heuristic precise, we note that $Q(x, y) \leq I_T(\varphi)$ whenever $\varphi(0) = x$, $\varphi(T) = y$. Let φ be any function appearing in the characterization of the subsolution property, and let $y = \varphi(T) \notin \mathcal{D}$. It follows from the definition of Q and a standard dynamic programming argument that for any $x, y, z \in \mathbb{R}^d$

$$Q(z, y) \leq Q(z, x) + Q(x, y). \tag{4.5}$$

Rearranging gives

$$-Q(z, x) + c \leq -Q(z, y) + c + Q(x, y).$$

Letting $z = 0$ and using $Q(x, y) \leq I_T(\varphi)$ when $\varphi(0) = x, \varphi(T) = y$, we have

$$\bar{V}(x) \leq \bar{V}(y) + I_T(\varphi).$$

Since φ is any function appearing in the definition of a subsolution, (4.3) holds, that is, $\bar{V}$ is a subsolution. □

Figure 2 illustrates a subsolution given as the negative of the quasipotential plus a constant, which we interpret as follows. $Q(x)$ provides a lower bound on the cost of going from 0 to x, while a subsolution provides a lower bound on the cost to exit starting at x. If we were to choose the point x on the boundary which minimizes $Q(x)$, we would obtain a lower bound on the cost to escape starting at 0. The cost of every other point in $\mathcal{D}$ can be related to the cost at origin by the dynamic programming inequality (4.5).

We denote the minimum cost to exit via $c_Q \doteq \min_{x\notin\mathcal{D}} Q(x)$, and we remark that continuity of Q implies $c_Q = \min_{x\in\partial\mathcal{D}} Q(x)$. The constant c_Q will be referred to a few times in the following chapters, but we emphasize that it plays no role in the determination of the RESTART algorithm since the thresholds are determined by *differences* of the subsolution. Rather, we will see that the optimality of a subsolution $\bar{V}$ is determined by whether or not $\bar{V}(0) = c_Q$.

A final remark is that while in the current paper we only consider estimating probabilities, one could very well consider the estimation of so-called "risk-sensitive functionals" of the form $\mathbb{E}_x[e^{-F(X^n(\rho^n)))/\varepsilon} 1_{\{\rho^n \leq T(n)\}}]$, where $F : \mathbb{R}^d \to [0, \infty)$ is measurable. The same notion of subsolution would yield asymptotically optimal estimators, though one would have to appropriately modify the boundary conditions. For more information; see, for example, [5] and [12].

4.2. *Performance measures and the rare event setting*. A natural performance measure used to judge the quality of an estimator is its variance. If we restrict to unbiased estimators, then minimizing the variance is equivalent to minimizing the second moment. Let γ^n denote the RESTART estimator for the escape probability of X^n from $\mathcal{D}$, and let $\mathfrak{S}^n(\bar{V})$ denote the second moment $\mathbb{E}_0[(\gamma^n)^2]$ when the subsolution $\bar{V}$ is used to construct the splitting

scheme. Owing to the large deviations scaling of the escape probability, it is convenient to frame the minimization of the second moment as the maximization of the scaled quantity $-n^{-1}\log \mathfrak{S}^n(\bar{V})$.

Recall that Theorem 3.1 guarantees $-n^{-1}\log \mathbb{P}_0(\rho^n \le T(n)) \to W(0)$ as $n \to \infty$. It follows from Jensen's inequality and unbiasedness of γ^n that

$$\limsup_{n\to\infty} -\frac{1}{n}\log \mathfrak{S}^n(\bar{V}) \le \limsup_{n\to\infty} -\frac{2}{n}\log \mathbb{P}_0\big(\rho^n \le T(n)\big)^2 = 2W(0). \tag{4.6}$$

Thus there is a natural upper bound on the decay rate $-n^{-1}\log \mathfrak{S}^n$, which motivates the following definition.

DEFINITION 4.4. If the estimator γ^n satisfies

$$\liminf_{n\to\infty} -\frac{1}{n}\log \mathfrak{S}^n(\bar{V}) = 2W(0), \tag{4.7}$$

we say that γ^n is asymptotically optimal.

The expression for asymptotic optimality over a finite time interval, as in [11], is the same as in the above definition. However, the distinction between finite and increasing time intervals is important, and we shall see in Section 4.3 that the finite time result does not imply optimality in the sense of Definition 4.4.

In this section we introduce the objects associated with the splitting scheme with an appropriate n-dependent scaling. We state pre-asymptotic upper bounds for the second moment which will be used in Section 4.3 to establish asymptotic lower bounds on the decay rate in terms of the value of the subsolution at 0. Since the negative of the quasipotential plus a constant is a subsolution (see Lemma 4.3) and $c_Q = W(0)$ from (4.4), it will follow that $\bar{V}(x) = -Q(x) + c_Q$ will yield an estimator which is asymptotically optimal in the sense of Definition 4.4.

To begin, we specify the scaled spacing in terms of some fixed $\Delta > 0$, as we had before, and set $\Delta_n = \Delta/n$. The corresponding thresholds are

$$C_j^n = \big\{x : \bar{V}(0) - \bar{V}(x) \ge \big(J^n - j\big)\Delta_n\big\}$$

for $j = 0, \dots, J^n$, where J^n is the smallest positive integer such that $J^n\Delta_n \ge \bar{V}(0) - \bar{V}(x)$ for all $x \in \mathcal{D}$.

For any subsolution $\bar{V}$, $J^n\Delta_n$ approximates the maximum of $\bar{V}(0) - \bar{V}(x)$ on $\mathcal{D}$:

$$J^n\Delta_n \to M \doteq \max_{x\in\bar{\mathcal{D}}}\big\{\bar{V}(0) - \bar{V}(x)\big\}. \tag{4.8}$$

For simplicity we define the difference sets $D_j^n \doteq C_j^n \setminus C_{j-1}^n$, where $C_{-1}^n = \varnothing$ by convention.

Next we define the n-dependent piecewise continuous importance functions associated to a subsolution $\bar{V}$. We will denote these by $\bar{U}^n$, but we will speak of the splitting scheme associated to $\bar{V}$. For $x \in \mathbb{R}^d$ the piecewise constant function $\bar{U}^n$ is defined by

$$\bar{U}^n(x) \doteq \sum_{j=0}^{J^n} j\Delta_n 1_{D_j^n}(x). \tag{4.9}$$

For ease of notation we write $\bar{U}_j^n = \bar{U}^n(x)$ for $x \in D_j^n$. Observe that for $x \in D_j^n$, we have by (4.8) and (4.9)

$$\big|\bar{U}^n(x) - \bar{U}^n(0) - \big(\bar{V}(x) - \bar{V}(0)\big)\big| \le \Delta_n. \tag{4.10}$$

Thus, the differences $\bar{U}^n(x) - \bar{U}^n(y)$ converge uniformly to the differences $\bar{V}(x) - \bar{V}(y)$ as $n \to \infty$.

We make a remark on the computational cost of the algorithm. Let N_i^n denote the number of particles at time i. In our pre-asymptotic performance measure $-n^{-1}\log \mathfrak{S}^n(\bar{V})$, we neglect the cost of simulating all the branched trajectories, as counted by N_i^n, and it is perhaps more appropriate to consider the work-normalized error, see [16],

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{\mathfrak{S}^n(\bar{V})\mathbb{E}_0[w^n]}{\mathbb{P}_0(\rho^n \le T(n))},$$

where the computational work w^n is defined as

$$w^n \doteq \sum_{i=1}^{nT(n)} N_i^n. \tag{4.11}$$

It has been shown in [6], Theorem 5.7, that over finite time intervals the computational work of the RESTART scheme is sub-exponential when based on a subsolution, that is, when $T(n) \equiv T > 0$ for all $n \ge 1$, we have

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_0\big[w^n\big] = 0.$$

An argument similar to the one in [6] shows that the same result holds for work over the increasing time intervals $[0, T(n)]$. The work-normalized error achieves its optimal value of 0 if and only if the work is subexponential and the scheme is asymptotically optimal in the sense of Definition 4.4. Since the work is subexponential when the RESTART scheme is based on a subsolution, to demonstrate that the work-normalized error decays to its optimal value of zero, it suffices to establish asymptotic optimality in the sense of Definition 4.4. We will also record the work w^n in our numerical examples.

We have from Theorem 17.6 in [1] the following pre-asymptotic upper bound for the second moment of RESTART. The inequality first appeared in [6], Theorem 4.5, and is presented here in a slightly modified form, that is, in [6] the RESTART algorithm is not limited to estimating probabilities and hence their notation is different. The statement concerns the estimator without explicit dependence on a scaling parameter n, and so we denote the second moment of the RESTART estimator by $\mathfrak{S}(\bar{U})$.

THEOREM 4.5. *Suppose that A is open, B is closed with $A \cap B = \varnothing$, Y_i is a discrete time Markov chain, and that the stopping time $H \doteq \inf\{i \ge 1 : Y_i \in A \cup B\}$ is almost surely finite. Then if $\bar{U}$ is used to design the splitting scheme, for all $y_0 \in (A \cup B)^c$ we have*

$$\begin{aligned}\mathfrak{S}(\bar{U}) \le{}& e^{-\bar{U}(y_0)}\mathbb{E}_{y_0}\left[\sum_{i=1}^{H} e^{\bar{U}(Y_{i-1})}\big[\mathbb{P}_{Y_i}(Y_H \in B)\big]^2\right] \\ &+ e^{-\bar{U}(y_0)}\mathbb{E}_{y_0}\big[e^{\bar{U}(Y_H)}1_B(Y_H)\big],\end{aligned} \tag{4.12}$$

where $\mathfrak{S}(\bar{U})$ is the second moment of the RESTART estimator for $\mathbb{P}_{y_0}(Y_H \in B)$.

We will apply this theorem to deduce an analogous result for the present situation of exiting the domain $\mathcal{D}$ within some time $T(n)$. We do so by introducing time as a state variable and judiciously choosing A and B.

COROLLARY 4.6. *For an open domain $\mathcal{D}$, fix some $n \in \mathbb{N}$ and let $\rho^n \doteq \inf\{i \geq 0 : X_i^n \notin \mathcal{D}\}$. If $\bar{V}$ is used to determine the importance functions $n\bar{U}^n$ used to construct the splitting scheme, we have*

$$\mathfrak{S}^n(\bar{V}) \leq e^{-n\bar{U}^n(0)}\mathbb{E}_0\left[\sum_{i=1}^{nT(n)} e^{n\bar{U}^n(X_{i-1}^n)}\big[\mathbb{P}_{X_i^n}\big(\rho^n \leq T(n) - i/n\big)\big]^2\right]$$
$$+ e^{-n\bar{U}^n(0)}\mathbb{E}_0\big[e^{n\bar{U}^n(X_{\rho^n}^n)}1\{\rho^n \leq T(n)\}\big].$$

PROOF. Fix any $n \geq 1$ and set $A = \mathcal{D} \times (T(n), \infty)$ and $B = \mathcal{D}^c \times [0, T(n)]$. Note that A is open, B is closed and $A \cap B = \varnothing$. We will apply Theorem 4.5 to the Markov chain $Y_i = (X_i^n, i/n)$, with $\bar{U}(x,t) = n\bar{U}^n(x)$ being the importance function used to design the splitting scheme. The stopping time $H \doteq \inf\{i \geq 1 : Y_i \in (A \cup B)\}$ is finite almost surely by finiteness of $T(n)$ for each $n \geq 1$.

Given $y = (x, k/n) \in \mathcal{D} \times [0, T(n)]$, $\mathbb{P}_y(Y_H \in B)$ is the probability that there is some $i \in \{k+1, \ldots, nT(n)\}$ for which $X_i^n \notin \mathcal{D}$, conditioned on $X_k^n = x$. By the Markov property

$$\mathbb{P}_{Y_i}(Y_H \in B) = \mathbb{P}_{(X_i^n, i/n)}\big(\rho^n \leq T(n) - i/n\big),$$

where ρ^n on the right-hand side above is the first escape time from $\mathcal{D}$ for the process started at X_i^n. Applying this observation to the first term in (4.12), we obtain with $y_0 = (0,0)$

$$e^{-n\bar{U}(y_0)}\mathbb{E}_{y_0}\left[\sum_{i=1}^{H} e^{n\bar{U}(Y_{i-1})}\big[\mathbb{P}_{Y_i}(Y_H \in B)\big]^2\right]$$
$$\leq e^{-n\bar{U}^n(0)}\mathbb{E}_{(0,0)}\left[\sum_{i=1}^{nT(n)} e^{n\bar{U}^n(X_{i-1}^n)}\big[\mathbb{P}_{(X_i^n, i/n)}\big(\rho^n \leq T(n) - i/n\big)\big]^2\right].$$

The inequality follows because $H \leq nT(n)$, so one is taking a sum over a larger number of positive terms. The Markov chain X_i^n is time-homogeneous, so with an abuse of notation,

$$\mathbb{P}_{(X_i^n, i/n)}\big(\rho^n \leq T(n) - i/n\big) = \mathbb{P}_{X_i^n}\big(\rho^n \leq T(n) - i/n\big).$$

For the second term in (4.12), note that $1_B(Y_H) = 1\{\rho^n \leq T(n)\}$, and $Y_H = X_{\rho^n}^n$. By again using $y_0 = (0,0)$ and time-homogeneity of X_i^n, we conclude the proof. □

4.3. *Analysis of asymptotic performance.* The main result of this section is to establish a lower bound on the decay rate of the RESTART estimator for the problem of escape from a domain $\mathcal{D}$ over increasing time intervals $[0, T(n)]$. The lower bound is stated in terms of the value of the subsolution $\bar{V}$ at the origin. Recalling that $W(0)$ is the minimal cost to exit $\mathcal{D}$ starting at 0, we show that $\bar{V}(0) = W(0)$ ensures asymptotic optimality in the sense of Definition 4.4. In particular, when the subsolution is defined in terms of the quasipotential with the constant $c_Q = W(0)$ as in Lemma 4.3, the RESTART estimator is asymptotically optimal. An upper bound on the decay rate is automatic from Jensen's inequality, as discussed at the beginning of Section 4.2.

It would also be possible to establish upper and lower bounds on the decay rate in terms of a variational problem, and with such a formulation one could obtain a limit on the asymptotic decay rate (the upper and lower bounds would agree). Instead of formulating the analysis in terms of escape from $\mathcal{D}$, one would have to consider the problem of moving from any point $x \in \mathcal{D}$ to any other point $y \in \mathcal{D}$. While limits for the asymptotic decay rate (as opposed to just lower bounds) are useful for the comparison of estimators, the main goal of this paper is to demonstrate that the decay rate of splitting algorithms does not degrade over long time intervals, and this is accomplished with only lower bounds.

Before stating and proving the main result, we first point out that the increasing time intervals $[0, T(n)]$ can indeed ruin asymptotic optimality when using importance sampling instead of splitting. It has been shown in [11], where stochastic differential equation models are considered, that when importance sampling is based on a time-independent subsolution, the following lower bound on the second moment holds:

$$\mathfrak{S}^n(\bar{V}) \geq e^{C_1(T-K)} e^{-nC_2}.$$

Here T is the time interval for escape, C_1 and C_2 are positive constants, and $K < T$ is a fixed constant. If T were n dependent, for example, $T(n) = n^2$, then

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathfrak{S}^n(\bar{V}) = \infty.$$

Not only is the estimator not asymptotically efficient, the decay rate of this importance sampling scheme is worse than ordinary Monte Carlo simulation.

For one-dimensional stochastic differential equations, it is possible to construct time-dependent subsolutions which ensure that the importance sampling schemes do not degrade over long time intervals. These time-dependent subsolutions are difficult to construct and it is not known how to construct them in any generality for higher dimensions.

We show that these problems do not arise with splitting in Theorem 4.7. As discussed in [11], the lower decay rate of importance sampling is due to the exponential growth of the likelihood ratio when the trajectory stays near the origin, which it can do with large enough probability over a long time interval. In contrast, such trajectories do not affect the decay rate of splitting schemes since they are not multiplied by a likelihood ratio; all that matters is that on average, out of R particles born in a threshold, one of them reaches a threshold with a lower index.

The analysis is carried out for the RESTART scheme, but for ordinary splitting the result will be the same and the analysis will be easier. By using techniques similar to the ones in [5] it is straightforward to get estimates on the second moment.

The main result is stated in the following theorem. It places a lower bound on the asymptotic decay rate in terms of the value of the subsolution at the origin.

THEOREM 4.7. *Suppose Conditions 2.1, 2.2, 2.4, 2.5 and 2.6 are met. Let $\bar{V}$ be a subsolution in the sense of Definition 4.1, and let $\mathfrak{S}^n(\bar{V})$ denote the second moment of the estimator γ^n when $\bar{V}$ is used to determine the piecewise constant importance functions $\bar{U}^n$. We have the following lower bound*:

$$\liminf_{n\to\infty} -\frac{1}{n} \log \mathfrak{S}^n(\bar{V}) \geq \bar{V}(0) + W(0).$$

PROOF. We first consider a general subsolution $\bar{V}$ used to construct the importance function $\bar{U}^n(x)$ defined by (4.9). From Corollary 4.6,

$$\begin{aligned} \mathfrak{S}^n(\bar{V}) \leq{}& \mathbb{E}_0\left[\sum_{i=1}^{nT(n)} e^{n(\bar{U}^n(X^n_{i-1})-\bar{U}^n(0))}\left[\mathbb{P}_{X^n_i}\left(\rho^n \leq T(n) - i/n\right)\right]^2\right] \\ &+ \mathbb{E}_0\left[e^{n(\bar{U}^n(X^n(\rho^n))-\bar{U}^n(0))} 1\{\rho^n \leq T(n)\}\right]. \end{aligned} \tag{4.13}$$

There are two terms in (4.13), and both are positive and n-dependent. If two sequences of positive numbers $\{a^1_n\}$ and $\{a^2_n\}$ satisfy $\liminf_{n\to\infty} -n^{-1}\log(a^i_n) \geq a$, then

$$\liminf_{n\to\infty} -n^{-1}\log\left(a^1_n + a^2_n\right) \geq a. \tag{4.14}$$

It therefore suffices to check the limit on both terms individually.

The second term is easier to handle. If $x \notin \mathcal{D}$, then $\bar{V}(x) \le 0$ by the boundary condition for the subsolution property in Definition 4.1. Since $X^n(\rho^n) \notin \mathcal{D}$, it is straightforward to show using (4.10) and Theorem 3.1 that

$$\liminf_{n\to\infty} -\frac{1}{n}\log \mathbb{E}_0\big[e^{n(\bar{U}^n(X^n(\rho^n))-\bar{U}^n(0))}1\{\rho^n \le T(n) - i/n\}\big] \ge \bar{V}(0) + W(0).$$

We remark that the second term contributes to the overall rate of decay even though it might seem negligible compared to the first term in (4.13). This is because we expect the highest contributions to the variance to come from the particles which split closest to the escape set, which is what the term represents, since these particles are highly correlated.

Next we handle the first term. Again, let $\varepsilon > 0$. Owing to the polynomial growth of $T(n)$ and the logarithmic scaling in the statement of the theorem, it suffices to bound each term in the sum (similarly to (4.14)). Observe that for any $i \in \{0, \dots, nT(n)\}$,

$$\begin{aligned}&\mathbb{E}_0\big[e^{n(\bar{U}^n(X^n_{i-1})-\bar{U}^n(0))}\mathbb{P}_{X^n_i}\big(\rho^n \le T(n) - i/n\big)^2\big]\\ &\quad= \sum_{j=0}^{J^n} \mathbb{E}_0\big[1_{\{X^n_i \in D^n_j\}}e^{n(\bar{U}^n(X^n_{i-1})-\bar{U}^n(0))}\mathbb{P}_{X^n_i}\big(\rho^n \le T(n) - i/n\big)^2\big].\end{aligned}$$

Since J^n grows linearly with n, owing to the logarithmic scaling—similarly to (4.14)—we may estimate the expected value individually on each set D^n_j. To do this, we first resolve the time index disparity between the indicator $1_{\{X^n_i \in D^n_j\}}$ and the point X^n_{i-1} appearing in the argument of $\bar{U}^n$. We will show that the probability of X^n_i and X^n_{i-1} not lying in the same set D^n_j is negligible as $n \to \infty$. Define

$$\delta^n_i \doteq \big[\bar{U}^n(X^n_{i-1}) - \bar{U}^n(0)\big] - \big[\bar{U}^n(X^n_i) - \bar{U}^n(0)\big] = \bar{U}^n(X^n_{i-1}) - \bar{U}^n(X^n_i).$$

Since $\mathcal{D}$ is bounded and $\bar{U}^n$ is piecewise constant on $\mathcal{D}$,

$$B \doteq \sup_{x,y\in\mathcal{D}} \big|\bar{U}^n(x) - \bar{U}^n(y)\big| < \infty. \tag{4.15}$$

For any $\varepsilon_1 > 0$ and each $i \in \{0, \dots, nT(n)\}$, we have

$$\begin{aligned}&\mathbb{E}_0\big[1_{\{X^n_i \in D^n_j\}}e^{n(\bar{U}^n(X^n_{i-1})-\bar{U}^n(0))}\mathbb{P}_{X^n_i}\big(\rho^n \le T(n) - i/n\big)^2\big]\\ &\quad= \mathbb{E}_0\big[1_{\{X^n_i \in D^n_j\}}e^{n(\bar{U}^n(X^n_i)-\bar{U}^n(0))}e^{n\delta^n_i}\mathbb{P}_{X^n_i}\big(\rho^n \le T(n) - i/n\big)^2\big]\\ &\quad\le e^{n\varepsilon_1}\mathbb{E}_0\big[1_{\{X^n_i \in D^n_j\}}e^{n(\bar{U}^n(X^n_i)-\bar{U}^n(0))}\mathbb{P}_{X^n_i}\big(\rho^n \le T(n) - i/n\big)^2\big]\\ &\qquad+ \mathbb{E}_0\big[e^{n(B+\delta^n_i)}1\{|\delta^n_i| \ge \varepsilon_1\}\big].\end{aligned} \tag{4.16}$$

Inequality (4.10) implies that $\bar{U}^n(0) - \bar{U}^n(x) \to \bar{V}(0) - \bar{V}(x)$ uniformly in x, so for the given $\varepsilon_1 > 0$ there is $N_1(\varepsilon_1) < \infty$ such that $n \ge N_1$ and $|\delta^n_i| \ge \varepsilon_1$ implies

$$\big|\bar{V}(X^n_{i-1}) - \bar{V}(X^n_i)\big| \ge \varepsilon_1/2.$$

By the uniform continuity of $\bar{V}$ on $\overline{\mathcal{D}}$, there is $\eta(\varepsilon_1) > 0$ such that $|\bar{V}(z) - \bar{V}(y)| \ge \varepsilon_1/2$ implies $|z - y| \ge \eta$. Thus for the given $\varepsilon_1 > 0$ we have $\eta > 0$ such that

$$\big\{|\delta^n_i| \ge \varepsilon_1\big\} \subset \big\{|X^n_{i-1} - X^n_i| > \eta\big\}.$$

Lemma 2.7 implies that for any $\eta > 0$, we can find $N_2(\eta) < \infty$ large enough that, for all $i \in \{0, \dots, nT(n)\}$ and $n \ge N_2$,

$$\mathbb{P}_x\big(|X^n_i - X^n_{i-1}| \ge \eta\big) \le e^{-n(2B+\bar{V}(0)+W(0)+1)}.$$

From equation (4.15) we have $|\delta_i^n| \le B$, so the last term in (4.16) satisfies

$$\begin{aligned}\mathbb{E}_0\big[e^{n(\delta_i^n+B)}1\{|\delta_i^n| \ge \varepsilon_1\}\big] &\le e^{n2B}e^{-n(2B+\bar V(0)+W(0)+1)}\\ &= e^{-n(\bar V(0)+W(0)+1)},\end{aligned} \tag{4.17}$$

whenever $n \ge N_3 = \max(N_1, N_2)$. For large enough n, this exponential decay is strictly greater than the claimed decay of $\bar V(0) + W(0)$, hence it will become clear later that this term can be ignored when we establish the lower bound on the decay rate of the second moment of the estimator.

For fixed i and j, the third term in (4.16) satisfies

$$\begin{aligned}&e^{n\varepsilon_1}\mathbb{E}_0\big[1_{\{X_i^n\in D_j^n\}}e^{n(\bar U^n(X_i^n)-\bar U^n(0))}\mathbb{P}_{X_i^n}\big(\rho^n \le T(n) - i/n\big)^2\big]\\ &\quad\le e^{n\varepsilon_1}\sup_{z\in D_j^n}\mathbb{P}_z\big(\rho^n \le T(n)\big)\\ &\qquad\times \mathbb{E}_0\big[1_{\{X_i^n\in D_j^n\}}e^{n(\bar U_j^n-\bar U^n(0))}\mathbb{P}_{X_i^n}\big(\rho^n \le T(n) - i/n\big)\big].\end{aligned} \tag{4.18}$$

In the inequality above, we used the shorthand notation $\bar U_j^n = \bar U^n(x)$ for $x \in D_j$, and we enlarged the set $\{\rho^n \le T(n) - i/n\}$ to $\{\rho^n \le T(n)\}$. Since we are interested in taking limits as $n \to \infty$, it suffices to demonstrate the upper bound (or lower bound on the rate) for any sequence $\{j_n\}_{n\ge1}$ such that, as $n \to \infty$,

$$j_n\Delta_n \to \theta \in [0, M],$$

where we recall from (4.8) that $M = \max_{x\in\bar{\mathcal{D}}}\{\bar V(0) - \bar V(x)\}$ and $J^n\Delta_n \to M$.

The limits for j_n and J^n determine a limit for $U_{j_n}^n - U^n(0)$ as follows. Let

$$L(\theta) \doteq \big\{x : \bar V(0) - \bar V(x) = M - \theta\big\}.$$

If $x_n \in D_{j_n}^n = C_{j_n}^n \setminus C_{j_n-1}^n$ for all $n \ge 1$, then

$$x_n \in \big\{x : (J_n - j_n + 1)\Delta_n > \bar V(0) - \bar V(x) \ge (J_n - j_n)\Delta_n\big\}.$$

By passing to a convergent subsequence we may assume without loss of generality that $x_n \to \bar x$. Then $\bar x \in L(\theta)$ and $\bar V(x_n) - \bar V(0) \to \theta - M$. By (4.10) we have for the given $\varepsilon > 0$ that

$$\bar U_{j_n}^n - \bar U^n(0) \le \theta - M + \varepsilon/2$$

for all sufficiently large n. This will allow us to estimate the $\bar U_j^n - \bar U^n(0)$ term in (4.18) when j is replaced by j_n. Furthermore, the supremum over $z \in D_{j_n}^n$ in (4.18) will be replaced in the limit by one over $z \in L(\theta)$. Finally, the Markov property for $\{X_i^n\}$ implies that

$$\mathbb{E}_0\big[1_{\{X_i^n\in D_{j_n}^n\}}\mathbb{P}_{X_i^n}\big(\rho^n \le T(n) - i/n\big)\big] \le \mathbb{P}_0\big(\rho^n \le T(n)\big).$$

Thus an upper bound on (4.18) with $j = j_n$ is

$$e^{n(\theta-M+\varepsilon_1+\varepsilon/2)}\sup_{z\in D_{j_n}^n}\mathbb{P}_z\big(\rho^n \le T(n)\big)\mathbb{P}_0\big(\rho^n \le T(n)\big) \tag{4.19}$$

for all sufficiently large n.

To summarize, (4.19) provides an upper bound on the first term in (4.16) via (4.18), and the second term in (4.16) will turn out to be exponentially negligible compared to the first term owing to the estimate (4.17). It remains to determine the asymptotic behavior of (4.19).

We first show a lower bound on the decay of the second term of (4.19), that is,

$$\liminf_{n\to\infty} -\frac{1}{n}\log\Big(\sup_{z\in D^n_{j_n}} \mathbb{P}_z\big(\rho^n \le T(n)\big)\Big) \ge \inf_{z\in L(\theta)} W(z). \tag{4.20}$$

It suffices to establish the corresponding bound when n_k is any subsequence along which the left hand side of (4.20) converges. Let $\{n_k\}_{k\ge1}$ denote such a subsequence, and let $\varepsilon_2 > 0$ be given. By Lemma 3.3, there is $N_4(\varepsilon_2) < \infty$ such that, for all $z \in \overline{\mathcal{D}}$ and $n \ge N_4$,

$$-\frac{1}{n}\log \mathbb{P}_z\big(\rho^n \le T(n)\big) \ge W(z) - \varepsilon_2. \tag{4.21}$$

Let $K_1(N_4) < \infty$ be such that $n_k \ge N_4$ for all $k \ge K_1$. For each $k \ge 1$, choose $z_{n_k} \in D^{n_k}_{j_{n_k}}$ such that

$$\sup_{z\in D^{n_k}_{j_{n_k}}} \frac{1}{n_k}\log \mathbb{P}_z\big(\rho^{n_k} \le T(n_k)\big) < \frac{1}{n_k}\log \mathbb{P}_{z_{n_k}}\big(\rho^{n_k} \le T(n_k)\big) + \varepsilon_2. \tag{4.22}$$

By passing to a further subsequence if necessary, we may assume without any loss that $z_{n_k} \to \bar z \in L(\theta)$. If $z_{n_k} \in \overline{\mathcal{D}}$, (4.21) applies with $z = z_{n_k}$ and $n = n_k$ for all $k \ge K_1$, while for $z_{n_k} \notin \overline{\mathcal{D}}$ we have $W(z_{n_k}) = 0$ and the same estimate holds for any $k \ge 1$. Moreover, since W is continuous and $z_{n_k} \to \bar z$, there is $K_2(\varepsilon_2) < \infty$ such that $W(z_{n_k}) > W(\bar z) - \varepsilon_2$ for all $k \ge K_2$. Combining this with (4.22) and using monotonicity of the logarithm, we find, for all $k \ge \max(K_1, K_2)$,

$$\begin{aligned} -\frac{1}{n_k}\log\Big(\sup_{z\in D_{n_k}} \mathbb{P}_z\big(\rho^{n_k} \le T(n_k)\big)\Big) &> -\frac{1}{n_k}\log \mathbb{P}_{z_{n_k}}\big(\rho^{n_k} \le T(n_k)\big) - \varepsilon_2 \\ &\ge W(z_{n_k}) - 2\varepsilon_2 \\ &> W(\bar z) - 3\varepsilon_2. \end{aligned}$$

Since $\bar z \in L(\theta)$, we automatically have $W(\bar z) \ge \inf_{z\in L(\theta)} W(z)$. Since $\{n_k\}_{k\ge1}$ and $\varepsilon_2 > 0$ were arbitrary, we conclude that (4.20) holds.

We now return to (4.19). We first take logarithms, scale by $-n$, and take the limit inferior. Using superadditivity of the limit inferior, (4.20), and another application of Lemma 3.3 for the probability starting at 0, we obtain

$$\begin{aligned} &\liminf_{n\to\infty} -\frac{1}{n}\log\Big(e^{n(\theta-M+\varepsilon_1+\varepsilon/2)} \sup_{z\in D^n_{j_n}} \mathbb{P}_z\big(\rho^n \le T(n)\big)\mathbb{P}_0\big(\rho^n \le T(n)\big)\Big) \\ &\quad \ge M - \theta - \varepsilon_1 - \varepsilon/2 + \inf_{z\in L(\theta)} W(z) + W(0). \end{aligned}$$

To conclude the proof, we use the subsolution property of $\bar V$, which implies that $\bar V(x) \le W(x)$ for all x. For $z \in L(\theta)$, $\bar V(z) = \bar V(0) - M + \theta$, so $\inf_{z\in L(\theta)} W(z) \ge \bar V(0) - M + \theta$. Thus,

$$M - \theta - \varepsilon_1 - \varepsilon/2 + \inf_{z\in L(\theta)} W(z) + W(0) \ge \bar V(0) + W(0) - \varepsilon_1 - \varepsilon/2.$$

Since both ε and ε_1 were arbitrary, we send them to zero and arrive at the desired result for a general subsolution $\bar V$. □

COROLLARY 4.8. *If the subsolution is defined through* $\bar V(x) = -Q(x) + c_Q$, *then* $\bar V(0) = c_Q = W(0)$ *and the estimator is asymptotically optimal.*

PROOF. In the case where $\bar{V}(x) = -Q(x) + c_Q$, we recall that Lemma 4.3 shows $-Q(x) + c_Q$ is indeed a subsolution. From Theorem 4.7, we have the lower bound

$$\liminf_{n\to\infty} -\frac{1}{n}\log \mathfrak{S}^n(\bar{V}) \geq c_Q + W(0),$$

while Jensen's inequality implies the corresponding upper bound as in (4.6). Equation (4.4) implies $c_Q = W(0)$, which yields asymptotic optimality in the sense of Definition 4.4. □

5. Numerical results. In this section we present numerical results, for which we first introduce some general terminology. The simulation of a single γ^n defined in (4.2) requires the simulation of multiple trajectories started at points where a threshold is crossed. We refer to the simulation of a single γ^n as a *sample* of the estimator. For a given estimation problem, we generate multiple samples and average the result to obtain an estimate. We refer to the collection of all samples and the resulting estimate as a single *trial*, and we record the simulation time for a trial in seconds.

Suppose γ_k^n, $k = 1, \ldots, K$ are samples of the estimator, together comprising a single trial. We record the following quantities:

- the estimate $\bar{\gamma}^n = \frac{1}{K}\sum_{k=1}^K \gamma_k^n$,
- the formal confidence interval $[\bar{\gamma}^n - 1.96\sigma^n/\sqrt{K}, \bar{\gamma}^n + 1.96\sigma^n/\sqrt{K}]$, where $\sigma^n = \sqrt{\frac{1}{K-1}\sum_{k=1}^K (\gamma_k^n - \bar{\gamma}^n)^2}$,
- the relative error $\sigma^n/(\bar{\gamma}^n\sqrt{K})$,
- the empirical decay rate $(-\log(\frac{1}{K}\sum_{k=1}^K (\gamma_k^n)^2))/(-\log\bar{\gamma}^n)$,
- the average work $\frac{1}{K}\sum_{k=1}^K w_k^n$, where w_k^n is the work for the simulated sample γ_k^n, as in (4.11),
- the average maximum number of particles at any given time i in the simulation, $\bar{N}^n = \frac{1}{K}\sum_{k=1}^K \max_i N_{i,k}^n$, where $N_{i,k}^n$ denotes the number of particles at time i for the kth sample γ_k^n, and we also record the maximum over all trials, $\max_{i,k} N_{i,k}^n$,
- the simulation time, which is estimated directly on the computer.

We also provide the large deviations approximation $e^{-nW(0)}$.

All simulations were run in MATLAB on a 2013 Macbook Air with a 1.7 GHz i7 processor and 8 GB DDR3 RAM. The code is a near-verbatim transcript of the pseudo-code provided in Algorithm 1 and is therefore completely unoptimized. In particular, the simulation time is provided only as a baseline for comparison.

We next comment on the performance of the algorithms. Theorem 4.7 provides a lower bound on the decay rate achieved by a splitting scheme generated through a subsolution $\bar{V}$ in terms of $\bar{V}(0)$. In practice, the most difficult step is to verify that $\bar{V}$ is actually a subsolution. It turns out one can characterize subsolutions in terms of a partial differential equation (PDE) known as the Hamilton–Jacobi–Bellman equation (HJB). To streamline the presentation of examples, we briefly explain the relationships between the local rate $L(x, \beta)$, the quasipotential, and the yet-to-be defined HJB PDE; for a complete discussion, see [14], Theorem 4.3 in Chapter 5.

Recall that $H(x, \alpha)$ denotes the cumulant generating function and $L(x, \beta)$ its Legendre transform; see (2.3). Consider the problem of finding a function Z, continuous on $\overline{\mathcal{D}}$ and continuously differentiable on $\mathcal{D}$, such that $Z(0) = 0$, $Z(x) > 0$ for $x \in \mathcal{D}\setminus\{0\}$, and $\nabla Z(x) \neq 0$ for $x \in \mathcal{D}\setminus\{0\}$, and which satisfies the HJB PDE

$$H(x, \nabla Z) = 0. \tag{5.1}$$

It is a classical result of Freidlin and Wentzell's (Theorem 4.3 of Chapter 4 in [14]) that then $Z(x) = Q(x)$ for x on the sublevel set $\{x : Z(x) \le c_Q\}$, where we recall $c_Q = \min_{y\in\partial\mathcal{D}} Q(y)$ and $Q(x)$ is the quasipotential.

If we restrict to continuously differentiable subsolutions $\bar{V}$, one can show by using a verification argument that the subsolution property given by Definition 4.1 is equivalent to the differential inequality and boundary condition

$$\begin{cases} H(x, -\nabla\bar{V}) \le 0, & x \in \overline{\mathcal{D}}, \\ \bar{V}(x) \le 0, & x \in \partial\mathcal{D}. \end{cases} \tag{5.2}$$

Since $Q(x)$ satisfies (5.1), $\bar{V}(x) = -Q(x) + c_Q$ satisfies the above subsolution property.

The problem of finding subsolutions that result in an asymptotically optimal estimator can therefore be broken into two steps: first, find by any means possible a function $\bar{V}$ satisfying the differential inequality $H(x, -\nabla\bar{V}) \le 0$ on $\bar{\mathcal{D}}$, and second, verify asymptotic optimality by checking that the function $\bar{V}$ can be shifted by a constant so that $\bar{V}(0) = c_Q$ while satisfying the boundary condition $\bar{V}(x) \le 0$ on $\partial\mathcal{D}$. We remark that in practice we will have $H(x, -\nabla\bar{V}) = 0$ rather than an inequality and it will be automatic that $\bar{V}$ can be shifted so that $\bar{V}(0) = c_Q$.

EXAMPLE 5.1 (One-dimensional Ornstein–Uhlenbeck).

DYNAMICS AND DOMAIN. Consider the one-dimensional discrete time process given by

$$X^n_{i+1} = X^n_i - \frac{1}{n}X^n_i + \frac{1}{n}U_i, \qquad X_0 = 0,$$

where the U_i are standard normal random variables with mean 0 and variance 1. One can compare this discrete-time process with the continuous-time Ornstein–Uhlenbeck process

$$dX^n(t) = -X^n(t)\,dt + \frac{1}{\sqrt{n}}\,dB(t).$$

We take the domain to be $\mathcal{D} = (-1, 0.75)$.

SUBSOLUTION. For $v_i(x) = U_i - x$, the cumulant generating function is

$$H(x, \alpha) = \log \mathbb{E}\big[e^{\alpha v_i(x)}\big] = \frac{1}{2}\alpha^2 - \alpha x.$$

Though $\sup_{x\in\mathbb{R}} H(x, \alpha) = \infty$ for all $\alpha \ne 0$, the drift of the process can be modified outside of $\mathcal{D}$ to ensure Condition 2.1, without affecting the escape probabilities of the process.

We claim that the quasipotential starting at 0 is $Q(x) = x^2$. Indeed, Q satisfies the HJB PDE:

$$H\big(x, \nabla Q(x)\big) = \frac{1}{2}(2x)^2 - 2x^2 = 0.$$

It remains to identify c_Q. Since $\partial\mathcal{D} = \{-1, 0.75\}$, $c_Q = \min_{x\in\partial\mathcal{D}} Q(x) = 0.75^2$. Thus, a subsolution that results in an asymptotically optimal estimator is $\bar{V}(x) = -x^2 + 0.75^2$.

SIMULATION AND RESULTS. We consider $n = 10, 20, 30, 40$, with $T(n) = n^2/10$ and a splitting rate of $R = 4$. We run one trial for each n, and for each trial we generate $K = 2000$ samples (see Table 1).

The estimates have a reasonable confidence interval, given that each trial consists of $K = 2000$ samples. The decay rate tends to 2, as expected. The average work (and consequently the simulation time) increases for two reasons: first, the time interval is growing, and second, as n increases the discretization gets finer and more steps are required to exit $\mathcal{D}$. However, even though the average work seems to grow quickly, $n^{-1}\log w^n$ tends to zero.

TABLE 1
Results for the one-dimensional Ornstein–Uhlenbeck process

	(n, T)			
	(10, 10)	(20, 40)	(30, 90)	(40, 160)
LDP estimate	3.60×10^{-3}	1.30×10^{-5}	4.69×10^{-8}	1.61×10^{-10}
Estimate	2.47×10^{-2}	5.15×10^{-4}	5.46×10^{-6}	3.99×10^{-8}
Confidence interval	$[2.24, 2.69] \times 10^{-2}$	$[4.86, 5.44] \times 10^{-4}$	$[5.23, 5.69] \times 10^{-6}$	$[3.85, 4.12] \times 10^{-8}$
Relative error	0.45	0.28	0.02	0.02
Decay rate	1.56	1.87	1.95	1.97
$\max_{i,k} N^n_{i,k}$	1032	1295	4400	5166
Avg. max $\bar{N}^n$	27	113	247	414
Average work w^n	222	2426	9838	26,719
$n^{-1} \log w^n$	0.5403	0.3897	0.3065	0.2548
Simulation time (s)	13	136	550	1505

COMPARISON WITH IMPORTANCE SAMPLING. For this example we also present results using the importance sampling estimator in [12] with the same subsolution. As discussed at the beginning of Section 4.3, the decay rate of the importance sampling estimator drops when trajectories are allowed to linger around the rest point, which is seen in the results below.

To compare the computational cost of the estimators we use the simulation time. The simulation time is not meant to be an accurate measure of the computational cost, since the runtime is highly dependent on the implementation of the algorithm and architecture of the machine. We only use the simulation time for purposes of comparison. Moreover, we note that the MATLAB implementations we use highly favors the importance sampling estimator, in the sense that we expect the improvements to be much greater for RESTART than importance sampling when both are well written in a compiled language. The implementation of RESTART is written for transparency of code rather than efficiency and manages memory poorly. For instance, in the RESTART implementation, the arrays storing information about the particles can resize every time step. Despite this imbalance, we will find that splitting vastly outperforms importance sampling.

We first consider the case of $(n, T(n)) = (20, 40)$. To compute a ground "truth" value, we use $K = 1.5 \times 10^5$ samples of the RESTART estimator to find

$$\mathbb{P}\big(\rho^{20} \le 40\big) \approx 5.24 \times 10^{-4}.$$

In Table 2, we show estimates produced by the importance sampling estimator using $K = 2 \times 10^3, 4 \times 10^5$ and 10^6 samples for $(n, T(n)) = (20, 40)$ and using the same subsolution.

Observe that none of the estimates come close to the baseline value of 5.24×10^{-4}. By comparison, with $K = 2000$ the RESTART estimator gave an estimate of 5.15×10^{-4} in 136

TABLE 2
Importance sampling results for the one-dimensional Ornstein–Uhlenbeck process

	K		
	2×10^3	4×10^5	1×10^6
Estimate	8.29×10^{-5}	1.93×10^{-4}	2.19×10^{-4}
Relative error	0.12	0.11	0.13
Decay rate	1.64	1.00	0.84
Simulation time (s)	1.3	297	937

seconds. Note that the relative error remains roughly the same when increasing the number of samples, while this intuitively should give a decrease. Also note that the estimate of the decay rate decreases as the number of samples is increased. Both of these quantities do not behave as might be expected since there is an increase in variance due to "rogue" trajectories which linger around the rest point and contribute heavily to the second moment, though they are needed for unbiasedness of the importance sampling estimator and cannot be discarded. As the number of samples is increased, the importance sampling estimate approaches the baseline value.

The results for other pairs $(n, T(n))$ show similar results, with the importance sampling estimator not being even within an order of magnitude of the RESTART estimate when using the same number of samples. For example, when $(n, T(n)) = (30, 90)$ and using $K = 2.5 \times 10^6$ samples we obtain a relative error of 0.46, which is far worse than for RESTART based on 2000 samples. In comparison with $(n, T(n)) = (20, 40)$, we also see that the relative error is diverging, as predicted by the theoretical results of [11].

EXAMPLE 5.2 (Two-dimensional single-well potential).

DYNAMICS AND DOMAIN. Consider the two-dimensional single-well potential

$$P(x) = \left(x_1^2 + 2x_2^4\right)^2,$$

which is strictly convex and has its global minimum at 0. To obtain a Markov chain with stationary distribution

$$\pi^n(x)\,dx = \frac{1}{Z^n} e^{-nP(x)}\,dx, \quad Z_n = \int_{\mathbb{R}^2} e^{-nP(x)}\,dx,$$

we could proceed as in the previous example and use a discrete time approximation to the diffusion

$$dX^n(t) = -\nabla P(x)\,dt + \frac{1}{\sqrt{n}}\,dB(t).$$

One might be interested in other dynamics with the same stationary distribution. The Metropolis–Hastings algorithm is a well-known method for producing a Markov chain with a given stationary distribution. So long as the proposal distribution $q^n(\cdot|x)$ is supported on a compact set containing x, or has sub-exponential tails outside some compact set, the quasipotential will be a suitable importance function for the resulting Markov chain. One only needs to verify that the resulting process $\{X^n\}$ satisfies an LDP on $D([0, T])$, which can be done via standard weak convergence arguments.

To illustrate, we use a Metropolis–Hastings algorithm to simulate a Markov chain X^n with $X^n(0) = (0, 0)$ and stationary distribution π^n as above. The proposal distribution q^n used is that of a symmetric random walk,

$$q^n(x, \cdot) \sim x + \frac{1}{n}\xi v,$$

where v is a vector drawn uniformly from $\{e_1, -e_1, e_2, -e_2\}$, $e_1 = (1, 0)$ and $e_2 = (0, 1)$, and $\xi \sim \text{Unif}([-1, 1])$ is an independently drawn uniform random variable.

The domain to escape $\mathcal{D}$ is taken to be the unit ball around the initial point $(0, 0)$; $\mathcal{D} = \{x : x_1^2 + x_2^2 < 1\}$.

SUBSOLUTION. Under suitable assumptions, the stationary distribution π^n of the Markov chain identifies the quasipotential via

$$Q(x) = \lim_{n\to\infty} -\frac{1}{n} \log \pi^n(x).$$

TABLE 3
Results for the two-dimensional single-well potential

	(n, T)			
	(10, 20)	(20, 40)	(30, 60)	(40, 80)
LDP estimate	4.73×10^{-4}	2.24×10^{-7}	1.05×10^{-10}	5.01×10^{-14}
Estimate	7.91×10^{-4}	3.35×10^{-7}	1.42×10^{-10}	4.54×10^{-16}
Confidence interval	$[0.47, 1.11] \times 10^{-3}$	$[0.44, 6.26] \times 10^{-7}$	$[0, 3.77] \times 10^{-10}$	$[0, 1.35] \times 10^{-15}$
Relative error	0.29	0.44	0.84	1.00
Decay rate	1.37	1.59	1.67	1.78
$\max_{i,k} N^n_{i,k}$	4097	19,767	95,888	65,537
Avg max $\bar{N}^n$	29	117	319	380
Average work w^n	341	1498	3994	6243
$n^{-1} \log w^n$	0.58	0.36	0.27	0.22
Simulation time (s)	80	348	915	1631

See, for instance, Theorem 4.3, Chapter 4 in [14]. Thus $Q(x) = P(x)$ on $\mathcal{D}$. As in the previous example, we can verify that $-Q(x)$ satisfies the HJB PDE with a Hamiltonian associated to the jump vector $v_i(x) = -\nabla P(x) + u_i$, where u_i are i.i.d. standard normal random variables.

To find the value c_Q, we use Lagrange multipliers to find that the smallest value of $P(x)$ on $\partial\mathcal{D}$ is $(7/8)^2 = 0.7656$, at $(\sqrt{3}/2, 1/2)$. Thus, $c_Q = (7/8)^2$, and a subsolution $\bar{V}$ that results in an asymptotically optimal estimator is given by

$$\bar{V}(x) = -P(x) + 0.7656.$$

SIMULATION AND RESULTS. We consider values of $n = 10, 20, 30, 40$, with $T(n) = 2n$ and a splitting rate of $R = 4$. We run one trial for each n, and for each trial we generate $K = 2000$ samples (see Table 3).

We note that over many samples, the maximum number of particles can deviate significantly from the average maximum. For instance, with $n = 40$, the maximum over all samples is 65,537, while on average the maximum in a single sample is 380. This suggests that the subsolution-based scheme typically does not usually lead to an explosion in the number of particles, but it is inevitable that splitting can occur many times, especially with the RESTART algorithm.

Note also that the relative error is increasing as n increases, whereas in the previous example it was decreasing. This is in part due to the sizes of the estimated probabilities which are smaller in this example.

Finally, we remark that for $n = 30, 40$ the left side of the confidence interval is 0, as otherwise the reported value would be negative; this suggests more samples are needed. These are more time-consuming to generate in this example due to the fact that the dynamics are simulated via a Metropolis–Hastings algorithm.

EXAMPLE 5.3 ($M/M/\infty$ queue).

DYNAMICS AND DOMAIN. The $M/M/\infty$ queue is a Markov process with Poisson arrivals at a fixed rate λ and infinitely many servers available, each having exponential service time with rate μ. We are interested in the number of occupied servers $Q(t)$, the dynamics of which can be specified via the infinitesimal generator

$$(\mathcal{L}f)(x) = \lambda\big[f(x+1) - f(x)\big] + \mu x 1_{\{x \geq 1\}}\big[f(x-1) - f(x)\big].$$

TABLE 4
Results for the $M|M|\infty$ queue

	(n, T)		
	(10, 20)	(15, 30)	(20, 40)
LDP estimate	2.10×10^{-2}	4.41×10^{-4}	9.26×10^{-6}
Estimate	4.84×10^{-3}	1.80×10^{-4}	3.15×10^{-6}
Confidence interval	$[4.55, 5.11] \times 10^{-3}$	$[1.70, 1.91] \times 10^{-4}$	$[2.98, 3.31] \times 10^{-6}$
Relative error	2.93×10^{-2}	2.94×10^{-2}	2.66×10^{-2}
Decay rate	1.72	1.83	1.89
$\max_{i,k}^n N_{i,k}$	50	78	145
Avg max $\bar{N}^n$	10	16	26
Average work w^n	596	1498	3238
$n^{-1} \log w^n$	0.63	0.48	0.40
Simulation time (s)	130	270	577

We consider the scaled process $X^n(t) = \frac{1}{n} Q(nt)$, with $\lambda = \mu = 1$, and we take the domain to escape to be $\mathcal{D} = \{x : x \in [0, 2]\}$. In other words, we are looking for the probability of using more than $2n$ servers over $[0, T(n)]$.

SUBSOLUTION. It is well known that the stationary distribution of Q is given by $\pi(k) = e^{-\lambda/\mu}(\lambda/\mu)^k/k!$. Thus, the stationary distribution of X^n is, for $x = k/n$ and $k \geq 0$,

$$\pi^n(x) = e^{-n\lambda/\mu} \frac{(n\lambda/\mu)^{nx}}{(nx)!}.$$

For $\lambda = \mu = 1$, the quasipotential is given by

$$Q(x) = \lim_{n\to\infty} -\frac{1}{n} \log \pi^n(x) = x \log x - x + 1.$$

The most likely point of exit is at $x = 2$, so $c_Q = 2\log 2 - 1 \approx 0.3863$. A subsolution which yields an asymptotically optimal estimator is therefore given by $\bar{V}(x) = -Q(x) + c_Q$.

SIMULATION AND RESULTS. We consider values of $n = 20, 30, 40$ with $T(n) = n/2$ and a splitting rate of $R = 4$. We run one trial for each n, and for each n we generate $K = 4000$ samples (see Table 4).

In this example the simulation time is significantly lower and we can run more samples ($K = 4000$ as opposed to $K = 2000$ in the previous examples). This is in particular due to the fact that there are fewer thresholds, as determined by $\bar{V}(0) - \bar{V}(2) = c_Q$. One can also increase the splitting rate R to decrease the number of thresholds, but in such a case more particles are generated and the overall effect on the simulation time does not seem to change much.

EXAMPLE 5.4 (Multidimensional Ornstein–Uhlenbeck).

DYNAMICS AND DOMAIN. We consider once again the Ornstein–Uhlenbeck process, this time taking values in $\mathbb{R}^4$ with drift matrix A,

$$X_{i+1}^n = X_i^n - \frac{1}{n} A X_i^n + \frac{1}{n} u_i,$$

where $X_0 = (0, 0, 0, 0)$ and the u_i are normal random variables with mean vector $(0, 0, 0, 0)$ and covariance given by the identity matrix $I_{4\times 4}$. The drift matrix A is given by $A = PDP^T$,

where D is a diagonal matrix and P is orthogonal, $P^{-1} = P^T$. We take

$$P = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad D = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/8 \end{pmatrix},$$

$$A = \begin{pmatrix} 3/4 & 1/4 & 0 & 0 \\ 1/4 & 3/4 & 0 & 0 \\ 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/8 \end{pmatrix}.$$

Observe that the matrix A is symmetric and positive definite, with eigenvalues given by the diagonal of D. We take the domain to escape to be

$$\mathcal{D} = \{x \in \mathbb{R}^4 : |x_1| < 1/2, x_2 + x_3 < 1/8, x_3 < 1/4, x_4 < 1/8\}.$$

Note that $\mathcal{D}$ is not bounded since x_2, x_3, x_4 can take arbitrarily negative values, and escape can occur for arbitrarily negative x_2. As noted in the Introduction, this does not affect the use of the subsolution since the probability of escape at a location where x_2 is very negative is negligible compared to the probability of escape.

SUBSOLUTION. The associated Hamiltonian is

$$H(x, \alpha) = \frac{1}{2}\|\alpha\|^2 - \langle \alpha, Ax \rangle.$$

As in Example 5.1, the drift can be modified outside $\mathcal{D}$ to ensure $\sup_{x \in \mathbb{R}^d} H(x, \alpha) < \infty$. The quasipotential $Q(x)$ is $\langle x, Ax \rangle$.

A location of most likely escape is at $x = (1/2, -1/8, 1/4, 1/8)$, which determines the constant $c_Q = 0.1855$. The subsolution based on the quasipotential is given by $\bar{V}(x) = -Q(x) + c_Q$. Another subsolution derived from this one is given by $0.95\bar{V}(x)$. The factor of 0.95 ensures that this subsolution is *strict*, in the sense that the defining inequality of a subsolution is strict. One reason for using a strict subsolution is to better control the number of particles, but with an expected reasonable increase in variance; see, for instance, [5]. We run the simulation for both subsolutions and compare the results.

SIMULATION AND RESULTS. We consider values of $n = 40, 80, 120$ with $T(n) = n/10$ and a splitting rate of $R = 5$. We run one trial for each n and each subsolution, and for each n we generate $K = 1000$ samples. The choice of K is limited by the computational cost. Table 5 is based on the subsolution $\bar{V}(x)$, and Table 6 is based on $0.95\bar{V}(x)$.

We first make some remarks which apply to the results from both subsolutions. The simulation time is longer than in the previous example owing to the higher dimension of the process, and the relative errors are large owing to the small number of samples. In both cases the decay rate increases with n as the large deviation approximation becomes increasingly accurate.

Observe that the decay rate for the strict subsolution $0.95\bar{V}(x)$ is slightly lower for all three pairs $(n, T(n))$, though the disparity between the decay rates decreases as n increases. The relative errors are also larger for the strict subsolution. Nevertheless, the estimates seem comparable in both cases, and the confidence intervals also show significant overlap. Most importantly, we see significant decrease in the total work and therefore the simulation time, especially for the case of $n = 120$. We recall that the simulation time is not used as a measure of computational cost but rather as a basis of comparison. We refer to Tables 8 and 9 in [5] for a similar comparison.

TABLE 5
Results for the multidimensional Ornstein–Uhlenbeck process using the subsolution $\bar{V}(x)$

	(n, T)		
	(40, 4)	(80, 8)	(120, 12)
LDP estimate	5.98×10^{-4}	3.57×10^{-7}	2.13×10^{-10}
Estimate	2.38×10^{-5}	6.37×10^{-8}	8.89×10^{-11}
Confidence interval	$[1.04, 3.73] \times 10^{-5}$	$[3.89, 8.85] \times 10^{-8}$	$[0.59, 1.19] \times 10^{-10}$
Relative error	2.87×10^{-1}	1.98×10^{-1}	1.71×10^{-1}
Decay rate	1.58	1.77	1.85
$\max_{i,k}^{n} N_{i,k}$	293	1442	4997
Avg max $\bar{N}^n$	49	251	737
Average work w^n	1767	32,233	221,551
$n^{-1} \log w^n$	0.19	0.13	0.10
Simulation time (s)	100	1843	12,584

TABLE 6
Results for the multidimensional Ornstein–Uhlenbeck process using the subsolution $0.95\bar{V}(x)$

	(n, T)		
	(40, 4)	(80, 8)	(120, 12)
LDP estimate	5.98×10^{-4}	3.57×10^{-7}	2.13×10^{-10}
Estimate	5.49×10^{-5}	4.71×10^{-8}	8.10×10^{-11}
Confidence interval	$[1.90, 9.08] \times 10^{-5}$	$[2.15, 7.26] \times 10^{-8}$	$[0.45, 1.16] \times 10^{-10}$
Relative error	3.33×10^{-1}	2.77×10^{-1}	2.24×10^{-1}
Decay rate	1.52	1.74	1.83
$\max_{i,k}^{n} N_{i,k}$	272	797	1255
Avg max $\bar{N}^n$	38	136	270
Average work w^n	1301	16,533	72,050
$n^{-1} \log w^n$	0.18	0.12	0.09
Simulation time (s)	92	1142	5364

APPENDIX A: PROOFS FROM SECTION 2

The following is an easy consequence of Condition 2.2.

LEMMA A.1. *Under Condition 2.2, for any compact set* K *there is* $c \in (0, \infty)$ *such that for any* x *and* $y \in K$, *there exists* $\tau \in (0, \infty)$ *and a smooth function* φ, *with* $\varphi(0) = x$, $\varphi(\tau) = y$, $\tau = |x - y|/r$ *for which* $I_\tau(\varphi) < c|x - y|$.

PROOF OF LEMMA 2.8. First we prove that W is continuous at any point $x \in \mathcal{D}$. By Lemma A.1 applied to the compact set $\bar{\mathcal{D}}$, there is $c \in (0, \infty)$ such that for any $x, y \in \mathcal{D}$ satisfying $|x - y| < \varepsilon/2c$, the cost of going from x to y in some finite time $\tau \in (0, \infty)$ is less than $\varepsilon/2$. By concatenating this path from x to y with a near-minimizing one from y to some point not in $\mathcal{D}$, and using additivity of I_T on disjoint time segments, we obtain $|W(x) - W(y)| < \varepsilon$.

If $x \in \overline{\mathcal{D}} \setminus \mathcal{D}$ then $W(x) = 0$, and the above proof can be repeated on a neighborhood N of x, separating into the cases $y \in N \cap \mathcal{D}$ and $y \notin N \cap \mathcal{D}$. Finally, W is constant on the open set $\overline{\mathcal{D}}^c$ and hence continuous there as well.

Next we show that for any $\varepsilon > 0$ and any $x \in \overline{\mathcal{D}}$, $W(x) \le W(0) + \varepsilon$, which implies that $W(x) \le W(0)$ for all $x \in \mathcal{D}$. Note that we might have $W(x) = 0$, if the noiseless dynamics starting at x pass through the complement of $\mathcal{D}$ before reaching any sufficiently small neighborhood of the origin, in which case $W(x) \le W(0)$ automatically since $W(0) \ge 0$. Otherwise, since W is continuous at 0, there is $\delta > 0$ such that any $y \in \mathcal{E}_\delta = \{x : |x| < \delta\}$ satisfies $W(y) < W(0) + \varepsilon$. Since 0 is the unique attracting fixed point, for any $x \in \mathcal{D}$ there is $T > 0$ and a trajectory $\varphi \in C([0, T])$ with $\varphi(0) = x$ such that $I_T(\varphi) = 0$ and $\varphi(T) \in \mathcal{E}_\delta$. Thus,

$$W(x) \le I_T(\varphi) + W\big(\varphi(T)\big) < W(0) + \varepsilon,$$

as claimed. □

APPENDIX B: PROOFS FROM SECTION 3

LEMMA B.1. *For all* $y \in \Gamma_\mu^\eta$,

$$\mathbb{E}_y\left[\sum_{i=1}^{M^n} B_i^n\right] \le \mathbb{E}_y[M^n] \sup_{x \in \Gamma_\mu^\eta} \mathbb{E}_x[B_1^n] \tag{B.1}$$

and

$$\mathbb{E}_y\left[\sum_{i=1}^{M^n} \tau_i^n - \sigma_{i-1}^n\right] \ge \mathbb{E}_y[M^n] \inf_{x \in \Gamma_\mu^\eta} \mathbb{E}_x[\tau_1^n]. \tag{B.2}$$

PROOF OF LEMMA B.1. Consider

$$S_k^n = \sum_{i=1}^{k} \tau_i^n - \sigma_{i-1}^n.$$

Let $\mathcal{F}_i^n = \sigma(X_j^n, 0 \le j \le i)$, $\mathcal{F} = \sigma(\bigcup_{i=1}^\infty \mathcal{F}_i^n)$ and $\mathcal{G}_k^n = \mathcal{F}_{\sigma_k^n}^n$, where

$$\mathcal{F}_{\sigma_k^n}^n = \{A \in \mathcal{F} : A \cap \{\sigma_k^n \le m\} \in \mathcal{F}_m^n, m \ge 1\},$$

is the sigma-algebra of the stopping time σ_k^n. Since $\sigma_i^n \ge \tau_i^n$, S_k^n is $\mathcal{G}_k^n$-measurable. Let

$$M^n \doteq \inf\{k \ge 1 : S_k^n > T(n)\}.$$

Then M^n is a stopping time with respect to the filtration generated by S_k^n and hence $\mathcal{G}_k^n$. In particular, $\{M^n \ge k\} = \{M^n \le k-1\}^c$ is $\mathcal{G}_{k-1}^n$-measurable. We have

$$\begin{aligned}
\mathbb{E}_y\left[\sum_{i=1}^{M^n} \mathbb{E}_{X_{\sigma_{i-1}^n}^n} B_1^n\right] &= \mathbb{E}_y\left[\sum_{i=1}^{M^n} \mathbb{E}_y[B_i^n | X_{\sigma_{i-1}^n}^n]\right] = \mathbb{E}_y\left[\sum_{i=1}^{M^n} \mathbb{E}_y[B_i^n | \mathcal{G}_{i-1}^n]\right] \\
&= \sum_{i=1}^{\infty} \mathbb{E}_y[1_{\{M^n \ge i\}} \mathbb{E}_y[B_i^n | \mathcal{G}_{i-1}^n]] = \sum_{i=1}^{\infty} \mathbb{E}_y[1_{\{M^n \ge i\}} B_i^n] \\
&= \mathbb{E}_y\left[\sum_{i=1}^{M^n} B_i^n\right].
\end{aligned}$$

In the above display, the first two equalities follow from the strong Markov property conditioned on $X_{\sigma_{i-1}^n}^n$, and pulling out the infinite sum from the expectation sign is permitted by

Tonelli's theorem because all terms in the sum are positive. The result in (B.1) immediately follows. Similarly, we have

$$\mathbb{E}_y\left[\sum_{i=1}^{M^n}\tau_i^n-\sigma_{i-1}^n\right]=\mathbb{E}_y\left[\sum_{i=1}^{M^n}\mathbb{E}_{X^n_{\sigma_{i-1}^n}}\tau_1^n\right],$$

and thus (B.2) follows. □

LEMMA B.2. *There exist $s>0$ and $N<\infty$ such that for all $n\geq N$ and any $y\in\Gamma_\mu^\eta$, we have $\mathbb{P}_y(\tau_1^n\leq s)\leq\frac{1}{2}$.*

PROOF. We may assume that $s\leq 1$ without any loss, since the probability is decreased by decreasing s. The probability $\mathbb{P}_y(\tau_1^n\leq s)$ is the probability that X^n exits $\mathcal{D}$ or enters γ_μ^η before time s. If instead of the piecewise constant interpolation of X^n we had taken the piecewise continuous one, it would not be possible for X^n to jump over the set γ_μ^η and the probability $\mathbb{P}_y(\tau_1^n\leq s)$ would be increased. Consequently, it suffices to establish the desired inequality when X^n is the piecewise continuous interpolation. For the purposes of this proof it is convenient to do so, and by an abuse of notation we use X^n to denote this process, so that $X^n\in C([0,1])$ $\mathbb{P}_y$-almost surely for the remainder of the proof.

Let $\Delta=\min(\mu/2-\eta,\operatorname{dist}(\mathcal{E}_\mu,\partial\mathcal{D}))>0$ where $\mathcal{E}_\mu=\{x:|x|<\mu\}$. Any trajectory which starts in Γ_μ^η and exits $\mathcal{D}$ or enters γ_μ^η must travel a distance of at least $\Delta>0$ at some point in time. Let

$$A_s=\big\{\varphi\in C\big([0,s]\big):\big|\varphi(t)-\varphi(0)\big|\geq\Delta\text{ for some }t\in[0,s],\varphi(0)\in\Gamma_\mu^\eta\big\}.$$

Then $\{\tau_1^n\leq s\}\subset\{X^n\in A_s\}$. To get an upper bound on the probability of $\{X^n\in A_s\}$ we place a lower bound on the cost of any $\varphi\in A_s$.

By Condition 2.1, $\bar H(\alpha)=\sup_{x\in\mathbb{R}^d}H(x,\alpha)$ is finite for all $\alpha\in\mathbb{R}^d$. It is also convex, and its Legendre transform $\bar L(\beta)=\sup_{\alpha\in\mathbb{R}^d}\{\langle\alpha,\beta\rangle-\bar H(\alpha)\}$ satisfies $\bar L(\beta)\leq L(x,\beta)$ for all $x\in\mathbb{R}^d$. The finiteness of $\bar H$ implies that $\bar L$ is superlinear:

$$\lim_{c\to\infty}\inf_{\beta:|\beta|\geq c}\frac{1}{c}\bar L(\beta)=\infty.$$

Choose $c\in(0,\infty)$ such that $\inf_{\beta:|\beta|\geq c}\frac{1}{c}\bar L(\beta)\geq 1$, and then choose $s>0$ such that $\Delta/s=c$. Then for any $\varphi\in A_s$ we have either $I_s(\varphi)=\infty$ (when φ is not absolutely continuous) or

$$\int_0^s\bar L\big(\dot\varphi(r)\big)\,dr\geq\int_0^t\bar L\big(\dot\varphi(r)\big)\,dr\geq t\bar L\bigg(\frac{1}{t}\int_0^t\dot\varphi(r)\,dr\bigg)\geq t\frac{\Delta}{t}=\Delta,$$

where $t\in(0,s]$ is a point such that $|\varphi(t)-\varphi(0)|\geq\Delta$ and therefore $|\int_0^t\dot\varphi(r)\,dr/t|\geq\Delta/t$. Since $\bar L(\beta)\leq L(x,\beta)$ for all $x\in\mathbb{R}^d$, it follows that $I_s(\varphi)\geq\Delta$. According to Theorem 1.1. in [9] applied to the closed set A_s, for any $\varepsilon>0$ there is $N(\varepsilon)<\infty$ such for any $x\in\Gamma_\mu^\eta$ and any $n\geq N$

$$P_x\big(X^n\in A_s\big)\leq\exp\big(-n\big(I_s(\varphi)-\varepsilon\big)\big).$$

If we choose $\varepsilon=\Delta/2$ and use $I_s(\varphi)\geq\Delta$, then for n large enough we find $P_x\{X^n\in A_s\}\leq 1/2$. □

APPENDIX C: RESTART ALGORITHM

Initialization;
$N_0 = 1$, $Z_{0,1} = 0$, $C_{0,1} = J$, $K_{0,1} = J$, $\gamma = 0$, $i = 0$;
while $N_i \neq 0$ **do**
  $N_{i+1} = 0$;
  for $j = 1, \ldots, N_i$ **do**
    If the particle did not reach the boundary;
    if $Z_{i,j} \in \mathcal{D}$ **then**
      Generate dummy variable $Y_{i,j}$ according to the law of X_{i+1};

      If the particle is not killed;
      if $\sigma(Y_{i,j}) \leq K_{i,j}$ **then**
        $N_{i+1} = N_{i+1} + 1$;
        $Z_{i+1,N_{i+1}} = Y_{i,j}$;
        $C_{i+1,N_{i+1}} = \sigma(Y_{i,j})$;
        $K_{i+1,N_{i+1}} = K_{i,j}$;

        If the particle needs to be branched;
        if $\sigma(Y_{i,j}) < C_{i,j}$ **then**
          for $k = 1, \ldots, J$ **do**
            for $\ell = 1, \ldots q_k(C_{i,j}, \sigma(Y_{i,j}))$ **do**
              $N_{i+1} = N_{i+1} + 1$;
              $Z_{i+1,N_{i+1}} = Y_{i,j}$;
              $C_{i+1,N_{i+1}} = \sigma(Y_{i,j})$;
              $K_{i+1,N_{i+1}} = k$;
            end
          end
        end
      end
    end

    If the particle reached the boundary;
    if $Z_{i,j} \notin \mathcal{D}$ **then**
      $\gamma = \gamma + e^{\bar{U}(Z_{i,j})}$;
    end
  end
  i = i+1;
end
$\gamma = e^{-\bar{U}(0)}\gamma$;

Algorithm 1: RESTART Algorithm

Acknowledgments. A. Buijsrogge's research supported by the Netherlands Organization for Scientific Research (NWO), project number 613.001.105.

P. Dupuis' research supported in part by the National Science Foundation (DMS-1317199), then AFOSR (FA-9550-18-1-0214) and the Defense Advanced Research Projects Agency (W911NF-15-2-0122).

M. Snarski's research supported in part by NSERC (PGSD3-471576-2015), DOE (DE-SC0010539) and DARPA (W911NF-15-2-0122).

REFERENCES

[1] BUDHIRAJA, A. and DUPUIS, P. (2019). *Analysis and Approximation of Rare Events*: *Representations and Weak Convergence Methods*. *Probability Theory and Stochastic Modelling* **94**. Springer, New York. MR3967100 https://doi.org/10.1007/978-1-4939-9579-0

[2] BUDHIRAJA, A., DUPUIS, P. and GANGULY, A. (2018). Large deviations for small noise diffusions in a fast Markovian environment. *Electron. J. Probab.* **23** Art. ID 112. MR3878137 https://doi.org/10.1214/18-ejp228

[3] CÉROU, F., DEL MORAL, P. and GUYADER, A. (2011). A nonasymptotic theorem for unnormalized Feynman–Kac particle models. *Ann. Inst. Henri Poincaré Probab. Stat.* **47** 629–649. MR2841068 https://doi.org/10.1214/10-AIHP358

[4] DE BOER, P. T. (2006). Some observations on importance sampling and RESTART. In *Proc. of the* 6*th International Workshop on Rare Event Simulation*.

[5] DEAN, T. and DUPUIS, P. (2009). Splitting for rare event simulation: A large deviation approach to design and analysis. *Stochastic Process. Appl.* **119** 562–587. MR2494004 https://doi.org/10.1016/j.spa.2008.02.017

[6] DEAN, T. and DUPUIS, P. (2011). The design and analysis of a generalized RESTART/DPR algorithm for rare event simulation. *Ann. Oper. Res.* **189** 63–102. MR2833612 https://doi.org/10.1007/s10479-009-0664-7

[7] DEL MORAL, P. (2004). *Feynman–Kac Formulae*: *Genealogical and Interacting Particle Systems with Applications*. *Probability and Its Applications* (*New York*). Springer, New York. MR2044973 https://doi.org/10.1007/978-1-4684-9393-1

[8] DUPUIS, P. and ELLIS, R. S. (1997). *A Weak Convergence Approach to the Theory of Large Deviations*. *Wiley Series in Probability and Statistics*: *Probability and Statistics*. Wiley, New York. MR1431744 https://doi.org/10.1002/9781118165904

[9] DUPUIS, P., ELLIS, R. S. and WEISS, A. (1991). Large deviations for Markov processes with discontinuous statistics. I. General upper bounds. *Ann. Probab.* **19** 1280–1297. MR1112416

[10] DUPUIS, P., SEZER, A. D. and WANG, H. (2007). Dynamic importance sampling for queueing networks. *Ann. Appl. Probab.* **17** 1306–1346. MR2344308 https://doi.org/10.1214/105051607000000122

[11] DUPUIS, P., SPILIOPOULOS, K. and ZHOU, X. (2015). Escaping from an attractor: Importance sampling and rest points I. *Ann. Appl. Probab.* **25** 2909–2958. MR3375891 https://doi.org/10.1214/14-AAP1064

[12] DUPUIS, P. and WANG, H. (2007). Subsolutions of an Isaacs equation and efficient schemes for importance sampling. *Math. Oper. Res.* **32** 723–757. MR2348245 https://doi.org/10.1287/moor.1070.0266

[13] ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes*: *Characterization and Convergence*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, New York. MR0838085 https://doi.org/10.1002/9780470316658

[14] FREIDLIN, M. I. and WENTZELL, A. D. (1998). *Random Perturbations of Dynamical Systems*, 2nd ed. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **260**. Springer, New York. MR1652127 https://doi.org/10.1007/978-1-4612-0611-8

[15] GARVELS, M. J. J. (2000). The splitting method in rare event simulation. Thesis, Universiteit Twente, The Netherlands. MR2715371

[16] GLYNN, P. W. and WHITT, W. (1992). The asymptotic efficiency of simulation estimators. *Oper. Res.* **40** 505–520. MR1180030 https://doi.org/10.1287/opre.40.3.505

[17] SHWARTZ, A. and WEISS, A. (1995). *Large Deviations for Performance Analysis*: *Queues*, *Communications*, *and Computing*. *Stochastic Modeling Series*. CRC Press, London. MR1335456

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