Monte Carlo methods for the approximation and mitigation of rare events, I

Paul Dupuis

Division of Applied Mathematics
Brown University

Probability and Simulation (Recent Trends)

Centre Henri Lebesgue
Rennes

June 2018
Overview

Introduction
- Examples
- Model (finite time) problem
- Large deviation approximations
- Rare event issues, and goal of accelerated MC
- Methods

History and some difficulties
- Some early papers
- Problems with first approach to importance sampling–game interpretation
- Problems with splitting methods

Importance Functions as a common framework
- Generation of schemes for importance sampling and splitting
- Why the IF should be a subsolution to a HJB equation
- Statements of performance for IFs that are subsolutions
- Remarks on proofs
- Remarks on distinctions between methods
Overview

- Construction of subsolutions
  - Classes of problems and methods
  - An “on the fly” method for importance sampling
- Large time problems, a situation where importance sampling and splitting differ greatly
  - Game interpretation and the problems of importance sampling
  - Statements of performance
  - Example

Detailed exposition in Chapters 14-17 of

Examples: General framework

- Monte Carlo estimation of probabilities and expected values largely determined by rare events.
- Stochastic processes with light-tailed random variables.
- Typical examples: exit problems, risk-sensitive functionals, functionals of invariant distributions with simple structure.
- Exploit a law of large numbers (LLN) scaling, continuous-time limit.
- Methods also identify conditional most likely way event happens.
Example 1: Weighted serve-the-longer policy (wireless)

\[ p_n = P \{ Q_i \text{ exceeds } c_i \text{ in some } i = 1, 2 \text{ before } Q = (0, 0) | Q(0) = (1, 0) \} . \]

Standard large deviation scaling:

\[ X^n(t) = \frac{1}{n} Q(nt) \]

\[ p_n = P \{ X^n_i \text{ exceeds } c_i \text{ in some } i = 1, 2 \text{ before } X^n = (0, 0) | X^n(0) = (1/n, 0) \} . \]
Example 2: Chemical reaction network and metastability

Rates for molecule types to react:

\[ A \xrightarrow{r_1} B, \quad B \xrightarrow{r_2} A, \quad 2B + A \xrightarrow{r_3} 3B. \]

Suppose \( n \) molecules. If \( (C^n_A(t), C^n_B(t)) = \) (fraction type \( A \), fraction type \( B \)) at \( t \), then \( C^n_B(t) = 1 - C^n_A(t) \),

\[
\begin{align*}
C^n_A(t) & \to C^n_A(t) - \frac{1}{n} \text{ at rate } n \left[ r_1 C^n_A(t) + r_3 C^n_A(t) C^n_B(t)^2 \right] \\
C^n_A(t) & \to C^n_A(t) + \frac{1}{n} \text{ at rate } n \left[ r_2 C^n_B(t) \right].
\end{align*}
\]
Example 2: Chemical reaction network and metastability

LLN limit for $C_A^n$ as $n \to \infty$:

$$\dot{\phi} = -r_1 \phi + r_2 (1 - \phi) - r_3 \phi (1 - \phi)^2$$

with two stable equilibria when $r_3 > 3(r_1 + r_2)$:

$$p_n = P \{ C_A^n(T) \text{ near stable point } a | C_A^n(0) \text{ near stable point } b \}$$

and reverse rate.
Example 3: Not-so-rare but high cost per sample—SPDE

Spread of pollutant, with concentration $u^n(x, t), x \in D \subset \mathbb{R}^d, t \in [0, T]$,

$$u^n_t(x, t) = cu^n_{xx}(x, t) + \langle v(x), u^n_x(x, t) \rangle - \alpha u^n(x, t) + \frac{1}{n} \cdot N(dx, dt)$$

with boundary and initial conditions and $N(dx, dt)$ spatial-temporal Poisson noise. Quantity of interest

$$p_n = P \{ u^n(x_0, T) \geq u^* \}.$$ 

However, issue with sampling is high cost and $p_n$ small but not exceedingly so. Here a moderate deviation approximation may be useful. Similar issues with other systems with high sampling cost (e.g., mean field models).
Example 4: Helicopter rotor stall (UTRC/MIT project)

Quantity of interest

\[ p_\varepsilon = P \left\{ \sup_{0 \leq t \leq T} |\theta_\varepsilon(t)| \geq 15^\circ \right\}, \]

where \( \theta_\varepsilon(t) = \) aircraft pitch angle, and various diffusion (SDE model) dynamics for helicopter and wind dynamics.*

Other process models: non-Markovian (e.g., Markov modulated diffusion).

---

*Rare Event Simulation of a Rotorcraft System, Zhang, Marzouk, Min, Sahai, to appear in AIAA Journal.*
As a general discrete time Markov model consider iid random vector fields \( \{v_i(x), x \in \mathbb{R}^d\} \), with

\[
P\{v_i(x) \in A\} = \theta(A|x)
\]

and the process

\[
X_{i+1}^n = X_i^n + \frac{1}{n}v_i(X_i^n), \quad X_0^n = x.
\]

Continuous time interpolation:

\[
X^n(i/n) = X_i^n, \quad \text{piecewise linear interpolation for } t \neq i/n.
\]
Continuous time models:

- diffusion processes such as

\[ dX^\varepsilon = b(X^\varepsilon)dt + \sqrt{\varepsilon}\sigma(X^\varepsilon)dW \]

 corresponds to canonical model with

\[ \theta(\cdot|x) = N(b(x), \sigma(x)\sigma^T(x)) \]

(i.e., Euler approximation) with \( \varepsilon = 1/n \).

- continuous time pure jump (e.g., queueing model) do not need time discretization, have development entirely analogous to discrete time theory.
Hitting probability: assume LLN trajectories attracted to a point in $A$

and estimate

$$p_n(x) = P \{ X^n \text{ hits } B \text{ before } A | X^n(0) = x \}$$

for $x \in (A \cup B)^c$. Essentially a finite time problem.
Monte Carlo and rare event considerations

Define

\[ H(y, \alpha) = \log E \exp \langle \alpha, v_i(y) \rangle, \quad L(y, \beta) = \sup_{\alpha \in \mathbb{R}^d} [\langle \alpha, \beta \rangle - H(y, \alpha)], \]

assume \( H(y, \alpha) < \infty \) all \( \alpha \in \mathbb{R}^d \).

Under conditions \( \{X^n(\cdot)\} \) satisfies a Large Deviation Principle with rate function

\[ I_T(\phi) = \int_0^T L(\phi, \dot{\phi}) dt \]

if \( \phi \) is AC and \( \phi(0) = x \), and \( I_T(\phi) = \infty \) else. Heuristically, for \( T < \infty \), given \( \phi \), small \( \delta > 0 \) and large \( n \)

\[ P \left\{ \sup_{0 \leq t \leq T} \|X^n(t) - \phi(t)\| \leq \delta \right\} \approx e^{-nI_T(\phi)}. \]
Monte Carlo and rare event considerations

**Hitting probability:**

\[
-\frac{1}{n} \log p_n(x) \\
\to \inf \{ I_T(\phi) : \phi(0) = x, \phi \text{ enters } B \text{ prior to } A \text{ before } T, T < \infty \}.
\]

Let

\[ T_{B,A} = \{ \text{trajectories that hit } B \text{ prior to } A \} \]

\[ r(x) = \inf \{ I_T(\phi) : \phi(0) = x, \phi \text{ enters } B \text{ prior to } A \text{ before } T, T < \infty \} \]
Monte Carlo and rare event considerations

- For standard Monte Carlo we average iid copies of $1_{\{X^n \in T_{B,A}\}}$. One needs $K \approx e^{nr(x)}$ samples for bounded relative error $[\text{std dev}/p_n(x)]$.
- Alternative approach: construct iid random variables $s^n_1, \ldots, s^n_K$ with $E s^n_1 = p_n(x)$ and use the unbiased estimator
  \[
  \hat{q}_{n,K}(x) = \frac{s^n_1 + \cdots + s^n_K}{K}.
  \]

- Performance determined by variance of $s^n_1$, and since unbiased by $E (s^n_1)^2$.
- By Jensen’s inequality
  \[
  -\frac{1}{n} \log E (s^n_1)^2 \leq -\frac{2}{n} \log E s^n_1 = -\frac{2}{n} \log p_n(x) \to 2r(x).
  \]

- An estimator is called *asymptotically efficient* if
  \[
  \lim inf_{n \to \infty} -\frac{1}{n} \log E (s^n_1)^2 \geq 2r(x).
  \]
Two methods: Importance sampling and splitting

Two main methods (to date) for construction of random variables $s^n_i$ with $E s^n_i = p_n(x)$: importance sampling (IS) and splitting and related interacting particle methods.

- **Idea of importance sampling.** Simulate under a different distribution for which the event is not rare, correct using likelihood ratio to make unbiased.

- **Idea of splitting.** Many variations. Simplest is to trigger splits in such a way that rare event is encouraged. Divide unit mass of original particle among descendants to maintain unbiasedness.
Importance sampling

**Tilted distributions.** Recall

\[ X^n_{i+1} = X^n_i + \frac{1}{n} v_i(X^n_i), \quad X^n_0 = x \]

and \( P \{ v_i(y) \in dz \} = \theta(dz|y) \). Consider the exponential tilt with tilt parameter \( \alpha \) and

\[ \theta^\alpha(dz|y) = e^{\langle \alpha, y \rangle - H(y, \alpha)} \theta(dz|y). \]

Construct \( \bar{X}^n_i \) recursively by setting

\[ \bar{X}^n_{i+1} = \bar{X}^n_i + \frac{1}{n} \bar{Z}^n_i, \quad \bar{X}^n_0 = x, \]

where \( P \{ \bar{Z}^n_i \in dz \mid \text{data till time } i \} = \theta^{\bar{\alpha}^n_i}(dz|\bar{X}^n_i), \bar{\alpha}^n_i = F^n_i(\bar{X}^n_0, \ldots, \bar{X}^n_i) \).

Likelihood ratio up to time \( n \), new distribution with respect to old:

\[ \prod_{i=0}^{n-1} e^{\langle \bar{\alpha}^n_i, \bar{Z}^n_i \rangle - H(\bar{X}^n_i, \bar{\alpha}^n_i)} \]
Importance sampling

Let $\bar{N}^n = \min \{ i : \bar{X}_i^n \in A \cup B \}$ and set

$$s^n = 1\{\bar{X}^n \in T_{B,A}\} \prod_{i=0}^{\bar{N}^n - 1} e^{-\langle \bar{\alpha}_i^n, \bar{Z}_i^n \rangle + H(\bar{X}_i^n, \bar{\alpha}_i^n)}.$$

Recall that performance is measured by

$$E (s^n)^2 = E_x \left[ 1\{\bar{X}^n \in T_{B,A}\} \prod_{i=0}^{\bar{N}^n - 1} e^{-2\langle \bar{\alpha}_i^n, \bar{Z}_i^n \rangle + 2H(\bar{X}_i^n, \bar{\alpha}_i^n)} \right],$$

which in terms of original random variables with $\alpha_i^n = F^n_i(X_0^n, \ldots, X_{i-1}^n)$ is

$$E (s^n)^2 = E_x \left[ 1\{X^n \in T_{B,A}\} \prod_{i=0}^{N^n - 1} e^{-\langle \alpha_i^n, v_i(X_i^n) \rangle + H(X_i^n, \alpha_i^n)} \right].$$

*Potential problem:* the exponential in the last display is dangerous.
Splitting

Start a particle at initial condition of interest, then branch in a way that encourages outcome of rare event.

Key issues:

- When to branch?
- How much to branch?
- How to form an unbiased estimate?
A certain number [proportional to \( n \)] of splitting thresholds \( C_j^n \) are defined which enhance migration, e.g.,

A single particle is started at \( x \) that follows the same law as \( X^n \), but branches into a number of independent copies each time a new level is reached.
Splitting

For simplicity assume the number of new particles $M$ is deterministic, so at each branching a multiplicative weight $1/M$ assigned to each descendent. Evolution continues until every particle has reached either $A$ or $B$. Let

$$R^n_x = \text{total number of particles generated}$$

$$X^n_j(t) = \text{trajectory of } j\text{th particle},$$

$$W^n = \text{product of weights assigned to } j \text{ along path}$$

If $Kn$ thresholds, then

$$s^n = W^n \cdot (\# \text{ of particles reaching } B \text{ before } A) = \frac{1}{MK_n} \sum_{j=1}^{R^n_x} 1\{X^n_j \in T_{B,A}\}.$$ 

How to choose thresholds $C^n_r$, $M$, and weights?
Splitting

- **RESTART**\(^\dagger\) is a multi-level splitting scheme that adds killing of particles to increase efficiency.
- The vast majority of the particles generated will not have trajectories that reach \(B\).
- **RESTART** kills particles that *backtrack* too much.

**Implementation of RESTART**: Identical to the standard splitting algorithm except

- particles are split *every time they enter a splitting threshold*, and
- particles are killed when they exit the splitting threshold in which they were born (killing threshold associated with particle).

Recall that initial particle is in \(C_J = \mathbb{R}^d\), so this particle never killed (although eventually absorbed by \(A \cup B\)).

Remarks

- First use of IS in rare event context was Seigmund (1976), who considered a particular one dimensional problem.
- *Fixed rate* splitting schemes begin with Kahn and Harris (1951), further developed in Booth and Hendricks (1984). Schemes have different names in different communities (e.g., *forward flux* in chemistry).
- Related methods are *fixed effort splitting* and interacting particle methods which involve a *transportation step* and then a *resampling step* to control particle numbers. Analysis more difficult due to strong coupling of particles.
Problems with straightforward extension of IS

The design of IS from Seigmund (1976) was as follows: choose $\alpha_i^n$ depending only on time (not state) so that $\tilde{X}^n$ follows conditional most likely path from LD theory.

A *global* approach is required, and in particular state feedback is required.
Problems with straightforward extension of IS

First counter-example for traditional approach (Glasserman-Kuo). Tandem queue, total population overflow. Simulation results for $\lambda = 0.1, \mu_1 = 0.45, \mu_2 = 0.45$, and buffer size $n = 25$. True value $p_n = 4.04 \times 10^{-15}$ and sample size $K = 20,000$.

<table>
<thead>
<tr>
<th>No.</th>
<th>Estimate $\hat{p}_n$ ($\times 10^{-15}$)</th>
<th>Standard Error ($\times 10^{-15}$)</th>
<th>95% C.I. ($\times 10^{-15}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. 1</td>
<td>2.58</td>
<td>0.22</td>
<td>[2.14, 3.02]</td>
</tr>
<tr>
<td>No. 2</td>
<td>2.47</td>
<td>0.24</td>
<td>[1.99, 2.95]</td>
</tr>
<tr>
<td>No. 3</td>
<td>5.63</td>
<td>2.50</td>
<td>[0.63, 10.63]</td>
</tr>
<tr>
<td>No. 4</td>
<td>13.65</td>
<td>10.27</td>
<td>[-6.89, 34.19]</td>
</tr>
</tbody>
</table>
Problems with straightforward extension of IS

If we let $\alpha_i^n = u(\bar{X}_i^n)$ for some $u : \mathbb{R}^d \to \mathbb{R}^d$ then we have the second moment

$$E (s^n)^2 = E_x \left[ 1_{\{X^n \in \mathcal{T}_B, A\}} \prod_{i=0}^{N^n-1} e^{-\langle u(X_i^n), v_i(X_i^n) \rangle + H(X_i^n, u(X_i^n))} \right].$$

A LD analysis shows

$$-\frac{1}{n} \log E (s^n)^2 \to J(u, x)$$

$$= \inf_{\phi} \int_0^{\tau} \left[ L(\phi, \dot{\phi}) + \langle u(\phi), \dot{\phi} \rangle - H(\phi, u(\phi)) \right] dt,$$

where $\phi(0) = x$, $\tau = \inf\{t \geq 0 : \phi(t) \in B\}$ and $\phi$ hits $B$ before $A$. 

Problems with straightforward extension of IS

Formally optimal decay rate is \( W(x) = \sup_{u(.)} J(u, x) \), value of a differential game. One can characterize \( W \) as viscosity solution to

\[
\sup_u \inf_\beta \left[ L(x, \beta) + \langle u, \beta \rangle - H(x, u) + \langle DW(x), \beta \rangle \right] = 0,
\]

plus \( W(x) = 0, x \in B \) and \( W(x) = \infty, x \in A \). Note that min/max holds.

The “hope” behind extending Seigmund’s idea: the player \( u \) choosing the IS change of measure can ignore both the other player’s choice of \( \beta \) and the current state \( x \).
Problems with splitting in the rare event setting

- **Too small a probability that a single particle reaches** $B$. Same as issue ordinary Monte Carlo, occurs if thresholds are too far apart.

- **Exponential growth in the number of particles**. Occurs if thresholds too tightly spaced, branching process becomes supercritical. While estimator is bounded by one, exponential growth in computational effort for single sample as bad as exponential growth of variance possible with IS.
Problems with splitting in the rare event setting

- **Variation in the number of particles reaching** $B$. This occurs, especially in a multidimensional setting, when both of the spacing problems regarding thresholds just mentioned occur. Even if weights are the same the estimator will have large variance.
Importance Functions

Natural conceptual framework for design is through *importance functions*. Let \( V : \mathbb{R}^d \rightarrow \mathbb{R} \) be continuous and satisfy

\[
V(x) \leq 0 \text{ for } x \in B.
\]

Design of scheme:

- for importance sampling, and assuming \( V \) continuously differentiable, the change of measure if the simulated trajectory is at \( \tilde{X}_i^n \) is

\[
\theta^{-DV(\tilde{X}_i^n)}(dz|y) = e^{-DV(\tilde{X}_i^n),z}H(y,-DV(\tilde{X}_i^n))\theta(dz|y).
\]

- for splitting, assuming \( V \) is continuous, thresholds are defined by

\[
C_i^n = \{ y : V(y) \leq i (\log M) / n \},
\]

if \( M \) descendents at each time of splitting.
Recall

\[ r(x) = \lim_{n \to \infty} -\frac{1}{n} \log p_n(x) \]

\[ = \inf \left\{ \int_0^T L(\phi(s), \dot{\phi}(s)) ds : \phi(0) = x, \phi \text{ hits } B \text{ before } A, \ T < \infty \right\} . \]

Generalize to arbitrary starting position \( y \). Then \( r(y) \) satisfies a dynamic programming relation: for \( \Delta > 0 \)

\[ r(y) = \inf \left\{ \int_0^\Delta L(\phi(s), \dot{\phi}(s)) ds + r(y + \phi(\Delta)) : \phi(0) = y \right\} . \]
Heuristic analysis

This implies $r(y)$ is a weak sense (viscosity) solution to

$$
H(y, Dr(y)) = 0
$$

where $\mathbb{H}(y, \alpha) = -H(y, -\alpha)$. 
Heuristic analysis

The ideal situation with splitting (a sort of analogue of zero variance change of measure in IS):

After entering a threshold for the first time, the mean number of descendants to reach the next threshold is exactly one.

If we could use $r$ as the importance function, problems with splitting, importance sampling essentially solved!

**Splitting:** With thresholds defined by $r$,

$$P \{ \text{particle born at threshold } j \text{ reaches } j - 1 \} \approx e^{-n \inf \Delta(\phi)}$$
$$\approx e^{-n [r(C^n_j) - r(C^n_{j-1})]}$$
$$= e^{-n [\log M/n]}$$
$$= \frac{1}{M},$$

where inf is over paths connecting any point in $C^n_j$ to $C^n_{j-1}$. Critical (borderline) growth, and weights $M^{-K^n} = M^{-n[r(x)/(\log M)]} = e^{-nr(x)}$. 
**Heuristic analysis**

*Importance sampling:* With tilts defined by

\[
\theta^{-Dr(\bar{X}_i^n)}(dz|y) = e^{\langle -Dr(\bar{X}_i^n), z \rangle - H(y, -Dr(\bar{X}_i^n))} \theta(dz|y),
\]

likelihood ratio along *any trajectory* satisfies

\[
1\{\bar{X}_{\tilde{N}_n}^n \in B\} \prod_{i=0}^{\tilde{N}_n - 1} e^{\langle Dr(\bar{X}_i^n), \bar{Z}_i^n \rangle + H(\bar{X}_i^n, -Dr(\bar{X}_i^n))}
\]

\[
= 1\{\bar{X}_{\tilde{N}_n}^n \in B\} \prod_{i=0}^{\tilde{N}_n - 1} e^{n\langle Dr(\bar{X}_i^n), [\bar{X}_{i+1}^n - \bar{X}_i^n] \rangle - H(\bar{X}_i^n, Dr(\bar{X}_i^n))}
\]

\[
= 1\{\bar{X}_{\tilde{N}_n}^n \in B\} e^{n \sum_{i=1}^{\tilde{N}_n - 1} \langle Dr(\bar{X}_i^n), [\bar{X}_{i+1}^n - \bar{X}_i^n] \rangle}
\]

\[
\approx 1\{\bar{X}_{\tilde{N}_n}^n \in B\} e^{nr(\bar{X}_{\tilde{N}_n}^n) - nr(x)} = 1\{\bar{X}_{\tilde{N}_n}^n \in B\} e^{-nr(x)}.
\]
Heuristic analysis

While possible in special cases, $r$ is typically not available. However, it turns out that the critical properties of an importance function $V$ are

- it should be a subsolution to the PDE,
- its value at the starting point,
- it should be $C^1$ for important sampling and a viscosity subsolution for splitting.
Why Importance Functions should be subsolutions

Let

$$\mathbb{H}(y, \alpha) = -H(y, -\alpha).$$

A *classical sense* subsolution $V$ for the escape probability is smooth ($C^1$) and satisfies

A *viscosity sense* subsolution $V$ satisfies the boundary inequalities and $\mathbb{H}(y, p) \geq 0$ for any superdifferential $p$ of $V$ at $y \in (A \cup B)^c$ (need not be smooth). Analogous definitions for other problems (e.g., finite time probabilities, functionals).
Why Importance Functions should be subsolutions

A constraint on importance functions. A subsolution replaces equalities in the definition of the solution with correct inequalities to still allow rigorous performance analysis. If $V$ is used as an importance function and $V$ is not a subsolution, bad (exponentially bad) things can happen.

As Glasserman-Kuo example shows, performance no longer related to value of $V$ at starting point, and can be much worse than ordinary MC. Consider representation for second moment when $V$ used:

\[
E(s^n)^2 = E_x \left[ 1_{\{X^n \in T_B, A\}} \prod_{i=0}^{N^n-1} e^{\langle DV(X^n_i), v(X^n_i) \rangle + H(X^n_i, -DV(X^n_i))} \right].
\]

In regions where $H(x, -DV(x)) > 0$, exponential growth of likelihood ratio.
Why Importance Functions should be subsolutions

- For (ordinary) splitting in regions where $H(x, -DV(x)) > 0$ the mean number of particles reaching next threshold strictly larger than 1, exponentially many particles, and a “work-normalized” asymptotic efficiency impossible.

- For fixed effort type splitting particles can become strongly correlated in (multi-dimensional) regions where $H(x, -DV(x)) > 0$. 
Performance for schemes based on subsolutions

Let

$$\mathbb{H}(y, \alpha) = -H(y, -\alpha).$$

A *classical sense* subsolution $V$ for the escape probability is smooth ($C^1$) and satisfies

A *viscosity sense* subsolution $V$ satisfies the boundary inequalities and $\mathbb{H}(y, p) \geq 0$ for any superdifferential $p$ of $V$ at $y \in (A \cup B)^c$ (need not be smooth). Analogous definitions for other problems (e.g., finite time probabilities, functionals).
Theorem

Let $s^n$ be the splitting estimate for the escape probability $p_n(x)$ based on $V$ using either standard splitting or RESTART. Then the number of particles generated grows subexponentially in $n$ if and only if $V$ is a viscosity subsolution, in which case we also have

$$
\liminf_{n\to\infty} -\frac{1}{n} \log E(s^n)^2 \geq V(x) + r(x).
$$

Under additional regularity $\liminf$ and $\geq$ become $\lim$ and $\geq$.

With a strict subsolution, one can bound the mean number of particles uniformly in $n$. 
Performance for schemes based on subsolutions

Theorem
Let \( V \) be a classical subsolution and \( s^n \) be the importance sampling estimate for the escape probability \( p_n(x) \) based on \( V \). Then

\[
\lim_{n \to \infty} \inf \frac{1}{n} \log E (s^n)^2 \geq V(x) + r(x).
\]

Under additional regularity \( \lim \inf \) becomes \( \lim \) with a RHS bounded below by \( V(x) + r(x) \).

If \( V \) is not a subsolution, Glasserman example shows performance not characterized by value \( V(x) \), even if optimal.

Results somewhat improved over what is in papers, and proofs of these and results stated later found in Springer book.
Requirements for asymptotic optimality. Asymptotic optimality means $V(x) = r(x)$ (note $V(x) \leq r(x)$ automatic). Equation holds with equality along conditional most likely path starting at $x$:

Analogous results for finite time problems, expected values, other problem formulations.
Performance for schemes based on subsolutions

Owing to complexity, analogous analysis of interacting particle type schemes completed only for one dimension. Suggests following.

**Conjecture for an IPS**: Suppose that the splitting thresholds are based on importance function $V$ and consider any monotone nondecreasing function $g : \mathbb{R} \to \mathbb{R}$ with $g(0) = 0$ that $g \circ V$ is a subsolution for the associated HJB equation. There is a constant $C$ such that if $N$ particles are maintained,

$$\liminf_{n \to \infty} -\frac{1}{n} \log E(s^n)^2 \geq g(V(x)) - \frac{C}{N} + r(x).$$

Then sup on $g$ for tightest bound.

- In comparison to ordinary splitting, which requires that $V$ be a subsolution, here one extra degree of freedom.
- However, second moment rate for ordinary splitting $V(x) + r(x)$, so if $V$ is actually a subsolution with IPS we lose $C/N$ in rate of decay.
- Examples show IPS scheme has poor asymptotic performance if no monotone mapping $g$ which makes $g \circ V$ into a subsolution.
Importance sampling. Estimate

\[ 1\{\bar{X}_n^{\bar{N}} \in B\} \prod_{i=0}^{\bar{N}^n-1} e^{\langle DV(\bar{X}_i^n), \bar{Z}_i^n \rangle} + H(\bar{X}_i^n, -DV(\bar{X}_i^n)) . \]

- Express second moment for estimator as an exponential integral with respect to original distributions

\[ E_x \left[ 1\{X^n \in \mathcal{T}_B, A\} \prod_{i=0}^{N^n-1} e^{\langle DV(X_i^n), v_i(X_i^n) \rangle} + H(X_i^n, -DV(X_i^n)) \right] \]

- Use same method as that of LD analysis to write a stochastic control representation for log of second moment
- Use classical verification argument to bound representation (hence decay of second moment)
Remarks on proofs

Splitting. Estimate based on splitting rate $M$ and $Kn$ thresholds

$$s^n = \frac{\text{# particles reach } B \text{ before } A}{MK^n}, \quad E(s^n)^2 = E\left[ \sum_{j=1}^{R^n} 1\{X_j^n \in T_{B,A}\} M^{-Kn} \right]^2.$$ 

- Partition expectation into contributions from pairs of particles with last common ancestor at threshold $\kappa$.
- Bound each such using large deviation bounds (one particle gets to $C^n_\kappa$, two independent particles go from $C^n_\kappa$ to $B$), decay of $M^{-Kn}$, dynamic programming inequalities on $r$ and $V$ to get upper bound of form
  $$e^{-n[r(x)+V(x)]}.$$
Summary—distinctions between importance sampling and splitting

Subsolutions allow one to characterize sufficient (and also necessary) conditions for performance measures for important sampling and splitting, but there are important differences.

- Regularity required [classical for importance sampling, viscosity for splitting] and how it is used to define the scheme \([D \tilde{V} \text{ versus } \tilde{V}]\).
- Subsolutions are often naturally constructed as the pointwise min of smooth subsolutions. If \(V = \min_{k=1,\ldots,K} V_k\) satisfies the boundary condition \(V(x) \leq 0\) for \(x \in B\) and if for each \(k \ \mathbb{H}(y, DV_k(y)) \geq 0\), then for small \(\delta > 0\)

\[
V^\delta(x) = -\delta \log \left( \sum_{k=1}^K e^{-\frac{1}{\delta} V_k(x)} \right), \quad DV^\delta(x) = \frac{1}{\sum_{k=1}^K e^{-\frac{1}{\delta} V_k(x)}} \sum_{k=1}^K e^{-\frac{1}{\delta} V_k(x)} DV_k(x)
\]

is subsolution with only small change in value at starting point.
Summary—distinctions between importance sampling and splitting

- When one goes to the trouble to make importance sampling work, generally (but not always) performs somewhat better than splitting.

- For Markov modulated noise (not covered in talk) and other multiscale problems IS requires solution to an eigenvalue for each tilt parameter used. Although splitting does not explicitly, often needed to evaluate Hamiltonian. Possible advantage to splitting.

- Under the condition $\sup_x E e^{\sigma \|v_i(x)\|^2} < \infty$ for some $\sigma > 0$, one can prove non-asymptotic bounds for importance sampling. Analogue not known for splitting.
Construction of subsolutions

Various methods depending on structure of problem, time dependent/independent, expected value vs. probability, path dependent functionals, etc. (see the references):

For systems with $H(y, \alpha)$ piecewise constant structure (e.g., queueing networks)—a direct construction based on pointwise minimum of affine functions using critical roots of $\mathbb{H}(y, \alpha) = 0$. Serve-the-longer:

$$p_n = P \{X_i^n \text{ exceeds } c_i \text{ some } i = 1, 2 \text{ before } X^n = (0, 0) | X^n(0) = (1/n, 0) \}.$$
Construction of subsolutions

Two dimensional serve-the-longer (solved for arbitrary dimension):

Splitting uses \( \min\{ \langle x, r_1 \rangle - c_1, \langle x, r_2 \rangle - c_2 \} \), with \( c_i \) chosen to satisfy boundary inequality. IS uses mollification of \( \min\{ \langle x, r_1 \rangle - c_1, \langle x, r_2 \rangle - c_2, \langle x, r_0 \rangle - c_0 \} \).
Construction of subsolutions

Construction as pointwise minimum of solutions to boundary/terminal conditions admitting explicit solutions (e.g., occupancy models, random graphs, related combinatorial problems, mean field models for weakly interacting finite state Markov chains, etc.).

For combinatorial type problems one can identify in explicit form solutions to the corresponding PDE, which to takes forms such as

$$
\mathbb{H}(x, p, t) \doteq \inf_{\theta \in \mathcal{P}(\{1, \ldots, d\})} [\langle p, M\theta \rangle + R(\theta \| \gamma(x, t))],
$$

where $R$ is relative entropy and $\gamma : \mathbb{R}^d \times [0, T] \rightarrow \mathcal{P}(\{1, \ldots, d\})$ is given. Solutions written as constrained finite dimensional convex optimization problem.†

†Chapter 7 of book and references therein.
Construction of subsolutions

Construction in terms of solution to linear/quadratic/regulator (in particular useful for *moderate deviation* approximations).

- For problems with high cost-per-sample (e.g., SPDE) and interest in not-so-rare events one can use moderate deviation rate to suggest design.
- This leads to rate function/PDE corresponding to LQR.
- Subsolution based on solution to Riccati equation, which typically only solved once.
Construction of subsolutions

**Construction via Freidlin-Wentsell quasipotential.** Available in explicit form for reversible or (asymptotically reversible).

Suppose $0 \in A$ is a stable point for zero cost trajectories $I_T(\phi) = 0$. Let

$$Q(x) = \inf \left\{ \int_0^T L(\phi, \dot{\phi}) dt : \phi(0) = 0, \phi(T) = x, T < \infty \right\}.$$ 

Then $Q$ is available in explicit form and smooth for important classes of problems. By dynamic programming argument

$$Q(y) \approx \inf \{ Q(y - \Delta \beta) + \Delta L(y, \beta) \},$$

and expanding gives

$$\mathbb{H}(y, DQ(y)) = 0.$$ 

Thus

$$V(x) = -Q(x) + \inf_{y \in B} Q(y)$$

is a always a subsolution.
An “on the fly” method for IS

A method proposed by Vanden-Eijnden and Weare. Assume that the simulated trajectory for importance sampling currently at $x = \bar{X}_i^n$, and \textit{one knows} $\phi^*, T^*$ that optimize in

\[
\inf \{ I_T(\phi) : \phi(0) = x, \phi(T) \in B, T < \infty \}.
\]

Then if $r(\cdot)$ is smooth at $x$, from optimality of $\phi$ one has

\[
Dr(x) = -\frac{\partial}{\partial \beta} L(x, \phi(0))
\]

and can compute the next change of measure. Assuming $x' = \bar{X}_{i+1}^n$ is \textit{not far} from $\bar{X}_i^n$, use descent in path space to find candidate minimizer in

\[
\inf \{ I_T(\phi) : \phi(0) = x', \phi(T) \in B, T < \infty \}.
\]

On a regular basis, compute (hopefully) true global minimizer. Various potential problems.
A situation where IS and splitting differ greatly

Although often comparable, recent results show can be truly significant differences. In particular, when

- metastable point is in the domain of simulation.

\[ P_x \{ X^n \text{ hits } B \text{ before } A \text{ by } T \} \text{ vs } P_x \{ X^n \text{ hits } B \text{ by } T \} \]
A situation where IS and splitting differ greatly

Treatment in neighborhoods of metastable points difficult.

When $T$ is large (e.g., transition rate calculations) the Freidlin-Wentsell quasipotential can be used to define a subsolution with nearly optimal value at starting point (optimal in limit $T \to \infty$). Recall

$$Q(x) = \inf \left\{ \int_0^T L(\phi(s), \dot{\phi}(s))ds : \phi(0) = 0, \phi(T) = x, T < \infty \right\},$$

$$\mathbb{H}(y, DQ(y)) = 0,$$

and

$$V(x) = -Q(x) + \inf_{y \in B} Q(y)$$

is a subsolution with optimal value at 0.
Example. 1D Gauss-Markov with \( B = (-1, 1)^c \):

\[
dX^n(t) = -X^n(t)dt + n^{-1/2}dW(t).
\]

A non-asymptotic representation for the 2\(^{nd}\) moment. Suppose we use the IS based on \( V \). Given control process \( v \) let \( \hat{X}^n(0) = 0 \) and

\[
d\hat{X}^n(t) = -\hat{X}^n(t)dt + DV(\hat{X}^n(t))dt + v(t)dt + n^{-1/2}dW(t).
\]

Based on a representation theorem (Boué-D, Annals of Probability, 1998) we have the non-asymptotic bound

\[
-\frac{1}{n} \log E \left[ 1 \{ \bar{\tau}^n \leq T, \hat{X}^n(\bar{\tau}^n) \in B \} \right. \\
\left. \inf_v E \left[ \frac{1}{2} \int_0^{\hat{\tau}^n} |v(s)|^2 ds - \int_0^{\hat{\tau}^n} \left| DV(\hat{X}^n(s)) \right|^2 ds + \infty1 \{ \hat{\tau}^n > T \} \right. \\
\right].
\]
Game interpretation for IS

Consider

Using the control above one obtains

\[
E \left[ \frac{1}{2} \int_0^{\hat{T}^n} |v(s)|^2 \, ds - \int_0^{\hat{T}^n} |DV(X^n(s))|^2 \, ds + \infty 1 \{\hat{T}^n > T\} \right] \leq -\frac{1}{n} C_1[T-t^*] + C_2
\]

and thus

\[
E \left[ 1 \{\tilde{T}^n \leq T, \tilde{X}^n(\tilde{T}^n) \in B \} \right] R^n(\tilde{X}^n)^2 \geq e^{C_1[T-t^*]} e^{-nC_2}.
\]
Statements of performance

Now consider $T(n) \to \infty$ as $n \to \infty$ in sub-exponential fashion (e.g., $T(n) = n^2$). For the problem $P \{X^n \text{ hits } B \mid T(n)|X^n(0) = x\}$, obtain

$$\lim_{n \to \infty} -\frac{1}{n} \log E \left( s^n_{\text{splitting}} \right)^2 = 2 \ V(0)$$

versus

$$\lim_{n \to \infty} -\frac{1}{n} \log E \left( s^n_{\text{IS}} \right)^2 = -\infty.$$

In this context RESTART much more efficient than ordinary splitting. Proof for splitting very similar to proof for fixed time $[0, T]$, once one established point to point estimates for unbounded time intervals:

$$-\frac{1}{n} \log P \{X^n(T(n)) \in N_\delta(y) \mid X^n(0) = x\} \approx Q(y).$$

This is done using Freidlin-Wentsell machinery.
Relative error (standard deviation of the estimator divided by the probability of interest) for IS (left) and RESTART (right).

<table>
<thead>
<tr>
<th>$\varepsilon \setminus T$</th>
<th>1</th>
<th>1.5</th>
<th>2.5</th>
<th>10</th>
<th>14</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>.2</td>
<td>1.9</td>
<td>1.1</td>
<td>.85</td>
<td>12</td>
<td>64</td>
<td>180</td>
</tr>
<tr>
<td>.16</td>
<td>2.3</td>
<td>1.3</td>
<td>.88</td>
<td>11</td>
<td>63</td>
<td>810</td>
</tr>
<tr>
<td>.13</td>
<td>2.7</td>
<td>1.4</td>
<td>.92</td>
<td>11</td>
<td>38</td>
<td>250</td>
</tr>
<tr>
<td>.11</td>
<td>3.3</td>
<td>1.5</td>
<td>.91</td>
<td>9.6</td>
<td>59</td>
<td>130</td>
</tr>
<tr>
<td>.09</td>
<td>4.0</td>
<td>1.7</td>
<td>.98</td>
<td>10</td>
<td>58</td>
<td>100</td>
</tr>
<tr>
<td>.07</td>
<td>5.6</td>
<td>2.1</td>
<td>1.0</td>
<td>10</td>
<td>39</td>
<td>160</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\varepsilon \setminus T$</th>
<th>1</th>
<th>1.5</th>
<th>2.5</th>
<th>10</th>
<th>14</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>.2</td>
<td>4.4</td>
<td>2.3</td>
<td>2.0</td>
<td>.87</td>
<td>.75</td>
<td>.67</td>
</tr>
<tr>
<td>.16</td>
<td>5.6</td>
<td>3.5</td>
<td>2.3</td>
<td>.97</td>
<td>.81</td>
<td>.72</td>
</tr>
<tr>
<td>.13</td>
<td>7.1</td>
<td>4.0</td>
<td>2.6</td>
<td>1.1</td>
<td>.88</td>
<td>.78</td>
</tr>
<tr>
<td>.11</td>
<td>8.5</td>
<td>4.7</td>
<td>2.8</td>
<td>1.1</td>
<td>.95</td>
<td>.82</td>
</tr>
<tr>
<td>.09</td>
<td>11</td>
<td>5.6</td>
<td>3.1</td>
<td>1.3</td>
<td>1.0</td>
<td>.90</td>
</tr>
<tr>
<td>.07</td>
<td>18</td>
<td>7.2</td>
<td>3.7</td>
<td>1.4</td>
<td>1.1</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Summary

- We have described the role of subsolutions in the design of importance functions for various types of accelerated Monte Carlo.
- We considered the light tailed case, and assumed that some kind of large deviation asymptotic is valid.
- For some types of probabilities/expected values, the results obtained are in some sense similar for different schemes (IS versus splitting), though there are interesting differences.
- For other types of problems (in particular large time problems), significant differences appear, and interesting differences may continue to emerge.
- Some schemes still not well understood in this setup (e.g., highly interactive splitting methods).
Papers introducing methods


Papers developing IPS

References

Paper introducing RESTART


Papers showing state feedback essential for importance sampling


On-the-fly importance sampling

References

Sample papers developing subsolutions for importance sampling


Papers developing subsolutions for splitting

- Splitting for rare event simulation: A large deviations approach to design and analysis (T. Dean and D.), SPA, 119, (2009), 562–587.

Some large deviation analysis of IPS

Use of moderate deviations


Papers that analyse neighborhoods of metastable points

- Splitting algorithms for rare event simulation over long time intervals (A. Buijsrogge, D. and M. Snarski), preprint.

General reference


Forthcoming book

- Representations and Weak Convergence Methods for the Analysis and Approximation of Rare Events, A. Budhiraja and D, Springer-Verlag, 2018.