Rare event simulation for diffusions

Jonathan Weare

University of Chicago

December 8, 2015

Consider the discrete time dynamical system

$$X_k^h = X_{k-1}^h + h\, b(X_{k-1}^h) + \sqrt{h}\, \sigma(X_{k-1}^h)\, \xi_k, \quad X_0^h = x.$$

where the ξ_k are independent and $\mathbf{E}[\xi_k] = 0$ and $\mathbf{E}\left[\xi_k \xi_k^{\mathsf{T}}\right] = I$, e.g. $\xi_k \sim \mathit{Gaussian}(0, I)$.

Equations of this type are used in the modeling of many real-world processes, e.g.

- A system of atoms interacting with each other and with a heat bath.
- The value of certain financial instruments.
- and many physical evolutions governed by ODE or PDE for which the model is not precisely known.

Notice that if $\sigma \equiv 0$ then we expect that as $h \to 0$,

$$X^h_{\lfloor t/h \rfloor} \to y_t$$

where y_t is the solution to the ODE

$$y_t = x + \int_0^t b(y_s) ds$$

or, equivalently, for a smooth function f, the function $u_h(t,x) = f(X^h_{\lfloor t/h \rfloor})$ converges as $h \to 0$ to the solution $u(t,x) = f(y_t)$ of the PDE

$$\partial_t u = \mathcal{L} u, \quad u(0,x) = f(x).$$

where

$$\mathcal{L} u = b \cdot Du$$

When $\sigma \neq 0$ we expect that

$$u_h(t,x) = \mathbf{E}\left[f(X^h_{\lfloor t/h \rfloor})\right]$$

converges as $h \to 0$ to the solution u of the PDE

$$\partial_t u = \mathcal{L}u, \quad u(0,x) = f(x).$$

where now

$$\mathcal{L}u = b \cdot Du + \frac{1}{2}\text{trace}(\sigma\sigma^{\mathsf{T}} D^2 u).$$

(The second order term appears when you expand $f(X^h_1) - f(x)$ in powers of h. The terms of order $\sqrt{h}$ vanish when you take the expectation. The two terms in $\mathcal{L}$ are the order h terms in the expansion.)

In fact, the solution to the PDE $\partial_t u = \mathcal{L} u$ with $u(0, x) = f(x)$ is $u(t, x) = \mathbf{E}\left[f(X_t)\right]$ where X_t solves

$$X_t = x + \int_0^t b(X_s) ds + \int_0^t \sigma(X_s) dW_s.$$

The second integral is an Itô integral.

For appropriate processes b_s and σ_s, if

$$X_t = x + \int_0^t b_s \, ds + \int_0^t \sigma_s \, dWs$$

and $w(t,x)$ is smooth, then we have Itô's formula

$$w(t, X_t) = w(0,x) + \int_0^t (\partial_t + \mathcal{L}_s) \, w(s, X_s) ds + \int_0^t Dw(s, X_s) \sigma(X_s) dW_s.$$

where

$$\mathcal{L}_t w = b_t \cdot Dw + \frac{1}{2} \text{trace} \left(\sigma_t \sigma_t^{\mathsf{T}} D^2 w \right).$$

Since b and σ are generally very approximate anyway, we won't worry about the difference between $\mathbf{E}\left[f(X_t)\right]$ and $\mathbf{E}\left[f(X^h_{\lfloor t/h \rfloor})\right]$.

To approximate averages with respect to X_t we can use

$$\mathbf{E}\left[f(X_t)\right] \approx \frac{1}{n}\sum_{j=1}^{n} f(X_t^j)$$

where the X_t^j are independent copies of X_t.

As you can check

$$\mathbf{mse} = \mathbf{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n} f(X_t^j) - \mathbf{E}\left[f(X_t)\right]\right)^2\right]$$

$$= \frac{1}{n}\left(\mathbf{E}\left[f^2(X_t)\right] - \mathbf{E}\left[f(X_t)\right]^2\right)$$

So the approximation converges like $1/\sqrt{n}$.

But notice that if $f(x) = 1$ for $x \in A$ and 0 otherwise then

$$\mathbf{mse} = \frac{1}{n}\left(\mathbf{P}\left[X_t \in A\right] - \mathbf{P}\left[X_t \in A\right]^2\right)$$

so the typical error is proportional to $\sqrt{\mathbf{P}\left[X_t \in A\right]}$.

If the event $\{X_t \in A\}$ is rare (i.e. if $\mathbf{P}\left[X_t \in A\right]$ is very small) then the error will be much bigger than the answer itself.

So the "straightforward" Monte Carlo approach won't work for rare events.

And in many situations a rare event may be precisely the characteristic of the system we are most interested in, e.g.

- A chemical reaction that is 10^{15} times less frequent than the typical timescale of a bond vibration.
- The most extreme and costly weather and climate events (like massive hurricanes or long droughts).
- A stock option highly dependent on the probability that the stock value reaches some unlikely level.

The “small noise” setting is a useful device when thinking about rare event simulation. Suppose

$$X_t = x + \int_0^t b(X_s)ds + \sqrt{\epsilon}\int_0^t \sigma(X_s)dW_s$$

and $f(x) = e^{-g(x)/\epsilon}$.

$u(t,x) = \mathbf{E}\left[e^{-g(X_t)/\epsilon}\right]$ solves $\partial_t u = \mathcal{L}u$ with $u(0,x) = e^{-g(x)/\epsilon}$, where now

$$\mathcal{L}u = b \cdot Du + \frac{\epsilon}{2}\text{trace}(\sigma\sigma^T D^2 u)$$

If $\partial_t u = \mathcal{L} u$ then $G^\epsilon = -\epsilon \log u$ solves the Hamilton-Jacobi equation

$$\partial_t G^\epsilon = \mathcal{H}(DG^\epsilon) + \frac{\epsilon}{2}\text{trace}(\sigma\sigma^\mathsf{T} D^2 G^\epsilon), \quad G^\epsilon(0,x) = g(x)$$

where

$$\mathcal{H}^\epsilon(p) = b \cdot p - \frac{1}{2}\|\sigma p\|_2^2.$$

Taking $\epsilon \to 0$ informally (though our conclusions are broadly correct) we arrive at the equation

$$\partial_t G = \mathcal{H}(DG), \quad G(0,x) = g(x).$$

This is a first order HJ equation with a convex Hamiltonian. The solution can be written

$$G(t,x) = \inf_{\phi} \int_0^t \frac{1}{2} \|\phi_s\|_2^2 ds + g(y_t^\phi)$$

where y_t^ϕ is the solution to the ODE

$$y_t^\phi = x + \int_0^t \left(b(y_s^\phi) + \sigma(y_s^\phi)\phi_s \right) ds$$

So

$$\mathbf{E}\left[e^{-g(X_t)/\epsilon}\right] = e^{-\frac{\gamma_1+o(1)}{\epsilon}}$$

where $\gamma_1 = G(t,x) = \inf_\phi \int_0^t \frac{1}{2}\|\phi_s\|_2^2 ds + g(y_t^\phi)$.

And by exactly the same argument,

$$\mathbf{E}\left[e^{-2g(X_t)/\epsilon}\right] = e^{-\frac{\gamma_2+o(1)}{\epsilon}}$$

where $\gamma_2 = \inf_\phi \int_0^t \frac{1}{2}\|\phi_s\|_2^2 ds + 2g(y_t^\phi)$.

Observe that $2\gamma_1 \geq \gamma_2$ so

$$\frac{\sqrt{\mathbf{mse}}}{\mathbf{E}\left[e^{-g(X_t)/\epsilon}\right]} = \frac{1}{n}\sqrt{e^{-\frac{\gamma_2-2\gamma_1+o(1)}{\epsilon}} - 1}$$

can be really big and standard MC won't work when ϵ is small.

So

$$\mathbf{E}\left[e^{-g(X_t)/\epsilon}\right] = e^{-\frac{\gamma_1+o(1)}{\epsilon}}$$

where $\gamma_1 = G(t,x) = \inf_\phi \int_0^t \frac{1}{2}\|\phi_s\|_2^2 ds + g(y_t^\phi)$.

And by exactly the same argument,

$$\mathbf{E}\left[e^{-2g(X_t)/\epsilon}\right] = e^{-\frac{\gamma_2+o(1)}{\epsilon}}$$

where $\gamma_2 = \inf_\phi \int_0^t \frac{1}{2}\|\phi_s\|_2^2 ds + 2g(y_t^\phi)$.

Observe that $2\gamma_1 \geq \gamma_2$ so

$$\frac{\sqrt{\mathbf{mse}}}{\mathbf{E}\left[e^{-g(X_t)/\epsilon}\right]} = \frac{1}{n}\sqrt{e^{-\frac{\gamma_2-2\gamma_1+o(1)}{\epsilon}} - 1}$$

can be really big and standard MC won't work when ϵ is small.

So we've established that rare events are a difficult computational problem. What can we do about it?

We need to "bias" the dynamical system so that the rare event becomes common while maintaining the ability to compute statistics corresponding to the unbiased dynamics.

So we've established that rare events are a difficult computational problem. What can we do about it?

We need to "bias" the dynamical system so that the rare event becomes common while maintaining the ability to compute statistics corresponding to the unbiased dynamics.

There are several methods by which we can bias the dynamics and reduce the error. For example

1. *importance sampling* – we can add an additional term to the dynamics nudging the system "toward" the rare event and reweight the samples.
2. *splitting* – we can evolve a large family of samples simultaneously and occasionally duplicate those making progress toward the rare event and remove those that aren't.
3. *control variates* – we can add a random variable to $f(X_t)$ that has zero mean but cancels out some of the variance of $f(X_t)$.
4. *path sampling* – use a Markov chain Monte Carlo scheme to sample directly from the ensemble of trajectories conditioned on the rare event.
5. *stratification* – partition the space of paths and sample trajectory fragments in each domain independently.

For any variance reduction strategy to be successful you have to know (or guess) some features of the rare event and input it into the scheme.

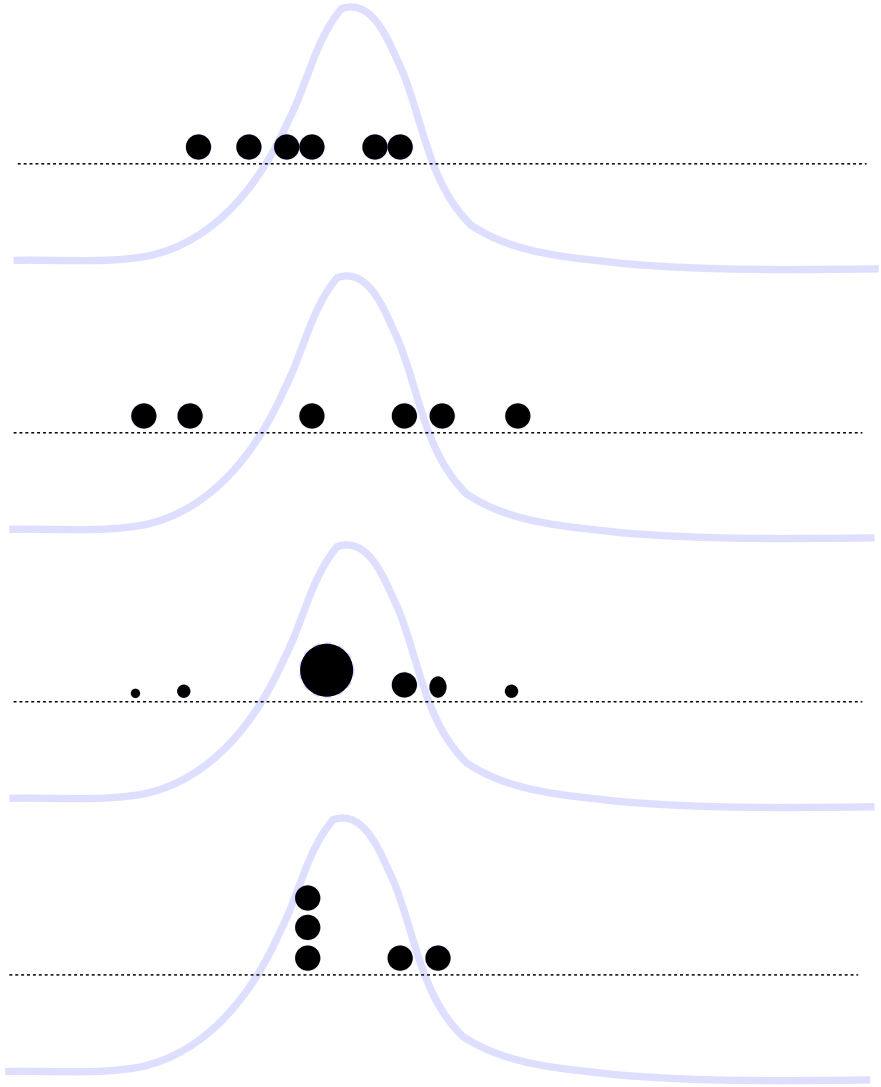
Diffusion Monte Carlo is a simple splitting scheme. Given a sequence of times $t_0 < t_1 < t_2 < \cdots$, it generates an ensemble of N_{t_k} points X_{t_k} so that

$$\mathbf{E}\left[\sum_{i=1}^{N_{t_k}} f(X_{t_k}^i)\right] = \mathbf{E}\left[f(X_{t_k}) e^{-\sum_{j=1}^{k} v(X_{t_{j-1}}, X_{t_j})}\right]$$

for any reasonable observable f.

The ensemble of points evolves in two steps:

1. Evolve each point according to the underlying dynamics from t_{k-1} to t_k.
2. To incorporate the additional "weight" factor $e^{-v(X_{t_{k-1}}, X_{t_k})}$ copy particles with large weight and kill those with low weight.



DMC proceeds from and ensemble $\left\{X^i_{t_{k-1}}\right\}$ of size $N_{t_{k-1}}$ at "time" t_{k-1} to an ensemble of size N_{t_k} at time t_k as follows:

1 : for $i = 1 : N_{t_{k-1}}$

2 : evolve the sample $X^i_{t_{k-1}}$ to time $\tilde{X}^i_{t_k}$

3 : generate a random integer $N^i \geq 0$ with $\mathbf{E}\left[N^i\right] = e^{-v(X^i_{t_{k-1}}, \tilde{X}^i_{t_k})}$

4 : add N^i copies of $\tilde{X}^i_{t_k}$ to the time t_k ensemble

5 : set $N_{t_k} = \sum_{i=1}^{N_{t_{k-1}}} N^i$

Suppose $\mathbf{E}\left[\sum_{i=1}^{N_{t_{k-1}}} f(X^i_{t_{k-1}})\right] = \mathbf{E}\left[f(X_{t_{k-1}})e^{-\sum_{j=1}^{k-1} v(X_{t_{j-1}},X_{t_j})}\right]$ for any f. Then

$$\begin{aligned}
\mathbf{E}\left[\sum_{i=1}^{N_{t_k}} f(X^i_{t_k})\right] &= \mathbf{E}\left[\sum_{i=1}^{N_{t_{k-1}}} N^i f(\tilde{X}^i_{t_{k-1}})\right] \\
&= \mathbf{E}\left[\sum_{i=1}^{N_{t_{k-1}}} \mathbf{E}\left[\mathbf{E}\left[N^i \mid \{\tilde{X}^j_{t_{k-1}}\}\right] f(\tilde{X}^i_{t_{k-1}}) \mid \{X^j_{t_{k-1}}\}\right]\right] \\
&= \mathbf{E}\left[\sum_{i=1}^{N_{t_{k-1}}} \mathbf{E}\left[e^{-v(X^i_{t_{k-1}},\tilde{X}^i_{t_k})} f(\tilde{X}^i_{t_{k-1}}) \mid X^i_{t_{k-1}}\right]\right] \\
&= \mathbf{E}\left[f(X_{t_k})e^{-\sum_{j=1}^{k} v(X_{t_{j-1}},X_{t_j})}\right]
\end{aligned}$$

What if we choose

$$v(x,y) = G(y) - G(x)?$$

Then DMC would compute

$$\mathbf{E}\left[\sum_{i=1}^{N_{t_k}} f(X_{t_k}^i)\right] = e^{G(X_0)}\mathbf{E}\left[f(X_{t_k})e^{-G(X_{t_k})}\right]$$

or (by redefining f)

$$e^{-G(X_0)}\mathbf{E}\left[\sum_{i=1}^{N_{t_k}} f(X_{t_k}^i)e^{G(X_{t_k}^i)}\right] = \mathbf{E}\left[f(X_{t_k})\right].$$

Recall the branching rule:

$$
\begin{aligned}
3: &\text{ generate a random integer } N^i \geq 0 \\
&\quad \text{with } \mathbf{E}[N_i] = e^{-\left(G(\tilde{X}^i_{t_k}) - G(X^i_{t_{k-1}})\right)} \\
4: &\text{ add } N^i \text{ copies of } \tilde{X}^i_{t_k} \text{ to the} \\
&\quad \text{time } t_k \text{ ensemble}
\end{aligned}
$$

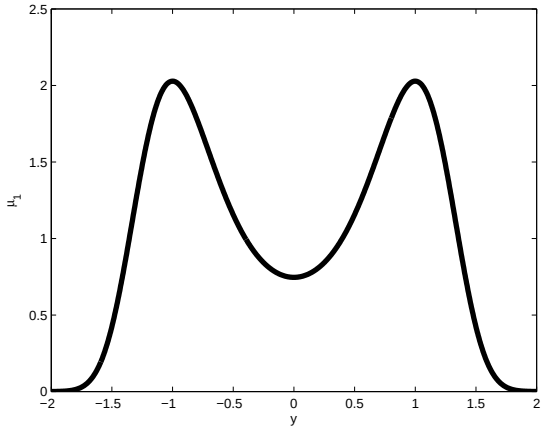
So if G decreases in a step more copies will be created.

By choosing G to be relatively small in a region we can sample that region more thoroughly

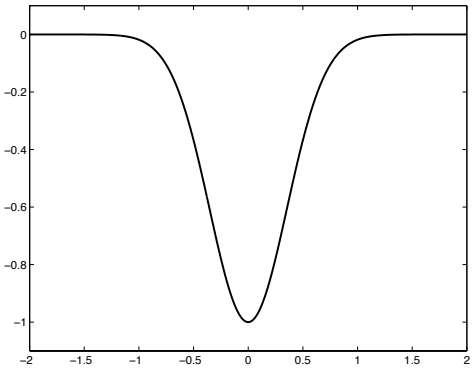
So, for example suppose

$$X_t = x - \int_0^t \nabla V(X_s) ds + \sqrt{2\mu} W_t$$

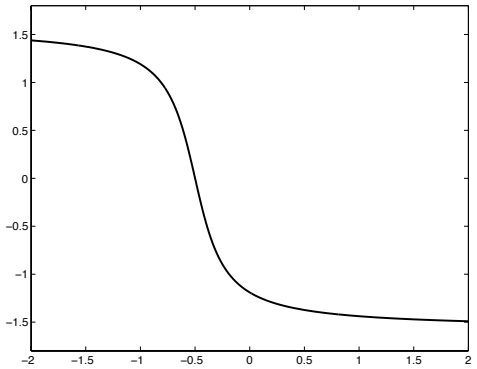
where the invariant measure $e^{-V/\mu}$ looks like



Then we might choose



or



if (**left**) we want to compute an average near the low probability saddle

or (**right**) we want to force the system from the left well to the right well.

A simple rare event example:

$$X_t = x - \int_0^t \nabla V(X_s) ds + \sqrt{2\mu} W_t$$

Starting from the lower well and running for 1 unit of time.

We use a DMC scheme with

$$v(x,y) = G(y) - G(x), \qquad G(x) = -\lambda \|x - x_A\|$$

where x_A is the minimum in the lower basin. We want to compute

$$\mathbf{P}_{x_A}\left(X_1 \in B\right), \qquad B = \{x : \|x - x_B\| < 0.25\}.$$

λ is chosen so that the expected number of particles ending in B is close to 1.

μ	λ	estimate	variance $\times$workload	brute force variance
16	5	0.5133	0.3357	0.2499
8	15	0.2839×10^{-1}	0.5519×10^{-2}	0.2758×10^{-1}
4	25.5	0.4813×10^{-5}	0.1521×10^{-8}	0.4813×10^{-5}
2	33	0.1262×10^{-13}	0.2133×10^{-23}	0.1262×10^{-13}

By the way...

$$\mathbf{E}\left[f(W_t)e^{-\int_0^t u(W_s)ds}\right]$$
$$\approx \mathbf{E}\left[f(W_t)e^{-\sum_{j=1}^{\lfloor t/dt\rfloor}\frac{1}{2}\left(u(W_{jh})+u(W_{(j-1)h})\right)ds}\right]$$

and if ψ is the solution to the imaginary time Schrödinger equation

$$\partial_t\psi = \frac{1}{2}\Delta\psi - u\psi$$

then

$$\int f(x)\psi(t,x)\,dx = \mathbf{E}\left[f(W(t))e^{-\int_0^t u(W(s))ds}\right]$$

where W is a Brownian motion.

DMC is frequently used to solve the imaginary time Schrödinger equation in high dimensions.

We haven't talked about a mathematically justifiable "optimal" choice of the function G within DMC.

We'll consider the optimal design question in more detail for importance sampling.

In importance sampling, instead of sampling the original process X you sample a biased process $\hat{X}$ solving

$$\hat{X}_t = x + \int_0^t \left(b(\hat{X}_s) + \sigma(\hat{X}_s) v(s, \hat{X}_s) \right) ds + \int_0^t \sigma(\hat{X}_s) dW_s$$

which satisfies

$$\mathbf{E}\left[f(\hat{X}_t) Z_t \right] = \mathbf{E}\left[f(X_t) \right]$$

where

$$Z_t = e^{-\int_0^t v^{\mathsf{T}}(s,\hat{X}_s) dW_s - \frac{1}{2}\int_0^t \|v(s,\hat{X}_s)\|_2^2 ds}.$$

To see this we just need to know that $u(t,x) = \mathbf{E}\left[f(\hat{X}_t)Z_t\right]$ solves $\partial_t u = \mathcal{L}u$ with $u(0,x) = f(x)$.

Note that Itô's formula for the process $Y_t = -\log Z_t$ reads

$$f(Y_t) = f(0) + \frac{1}{2\epsilon}\int_0^t \left(f'(Y_s) + f''(Y_s)\right)\|v(s,\hat{X}_s)\|_2^2 ds + \int_0^t f'(Y_s)v^{\mathsf{T}}(s,\hat{X}_s)dW_s.$$

Applied to $Z_t = e^{Y_t}$ we find

$$Z_t = 1 - \int_0^t Z_s v^{\mathsf{T}}(s,\hat{X}_s)dW_s$$

Fixing t and applying Itô's formula to $u(t-s, \hat{X}_s)$,

$$u(t-s,\hat{X}_s) = u(t,x) + \int_0^t (-\partial_t + \mathcal{L})\, u(t-s,\hat{X}_s) ds + \int_0^t \sigma(\hat{X}_s) v(s,\hat{X}_s) \cdot Du(t-s,\hat{X}_s) ds + \int_0^t Du(t-s,\hat{X}_s)\sigma(\hat{X}_s) dW_s$$

The second term on the right hand side vanishes because $\partial_t u = \mathcal{L} u$.

Finally, we can define the process $Y_s = \left(u(t-s, \hat{X}_s), Z_s\right)$ and apply Itô's formula to a function $h(Y_s)$,

$$h(Y_s) = h(Y_0) + \int_0^t \left(\bar{b}_s \cdot Dh(Y_s) + \frac{1}{2}\text{trace}\left(\bar{\sigma}_s \bar{\sigma}_s^{\mathsf{T}} D^2 h(Y_s)\right)\right) ds + \int_0^t Dh(Y_s)\bar{\sigma}_s dW_s$$

where

$$\bar{b}_s = \begin{bmatrix} \sigma(\hat{X}_s)v(s, \hat{X}_s) \cdot Du(t-s, \hat{X}_s) \\ 0 \end{bmatrix}, \quad \bar{\sigma}_s = \begin{bmatrix} Du(t-s, \hat{X}_s)\sigma(\hat{X}_s) \\ -Z_s v^{\mathsf{T}}(s, \hat{X}_s) \end{bmatrix}.$$

This formula with $h(u, z) = uz$ tells us that

$$u(t-s, \hat{X}_s)Z_s = u(t, x) - \int_0^t u(t-s, \hat{X}_s)Z_s v^{\mathsf{T}}(s, \hat{X}_s) dW_s + \int_0^t Z_s Du(t-s, \hat{X}_s)\sigma(\hat{X}_s) dW_s$$

So we find that $\mathbf{E}\left[u(t-s,\hat{X}_s)Z_s\right] = u(t,x) = \mathbf{E}\left[f(X_t)\right]$ for all $s \leq t$. In particular

$$\mathbf{E}\left[f(\hat{X}_t)Z_t\right] = \mathbf{E}\left[u(0,\hat{X}_t)\right] = \mathbf{E}\left[f(X_t)\right].$$

The **mse** corresponding to the new estimator

$$\frac{1}{n}\sum_{j=1}^{d} f(\hat{X}_t^i)Z_t^i \approx \mathbf{E}\left[f(X_t)\right]$$

is

$$\mathbf{mse} = \frac{1}{n}\left(\mathbf{E}\left[f^2(\hat{X}_t)Z_t^2\right] - \mathbf{E}\left[f(X_t)\right]^2\right).$$

We want to choose v so that $\mathbf{E}\left[f^2(\hat{X}_t)Z_t^2\right]$ is as close as possible to $\mathbf{E}\left[f(X_t)\right]^2$.

Using Itô's formula a few times just as before we can show that the function

$$w(s,x) = \mathbf{E}_{s,x}\left[f^2(\hat{X}_t)Z_t^2\right]$$

for $s \leq t$ solves the PDE

$$\partial_s w + \mathcal{L}w + \|v\|_2^2 w - \sigma v \cdot Dw = 0, \quad w(t,x) = f^2(x).$$

Notice that if, for $s \leq t$, we choose

$$v(s,x) = \sigma^{\mathsf{T}}(x) D\left(\log u(t-s,x)\right)$$

then $w(s,x) = u(t-s,x)^2$ solves the PDE.

So, in that case

$$\mathbf{E}\left[f^2(\hat{X}_t)Z_t^2\right] = \mathbf{E}\left[f(X_t)\right]^2$$

and the Monte Carlo estimate has zero variance.

So there's a choice of v that leads to a perfect estimator... but it requires knowing the answer before you start.

Any good choice of v is going to be some kind of approximation of this perfect choice even if it's an ad-hoc choice based only on physical intuition.

Let's try to find some provably good, but still practically useful choices of v.

One way to do that is to use our small noise device.

Assume again that $f(x) = e^{-g(x)/\epsilon}$, multiply the diffusion coefficient σ by $\sqrt{\epsilon}$, and multiply v by $1/\sqrt{\epsilon}$, so that

$$\hat{X}_t = x + \int_0^t \left(b(\hat{X}_s) + \sigma(\hat{X}_s) v(s, \hat{X}_s) \right) ds + \sqrt{\epsilon} \int_0^t \sigma(\hat{X}_s) dW_s$$

and

$$Z_t = e^{-\frac{1}{\sqrt{\epsilon}} \int_0^t v^{\mathsf{T}}(s,\hat{X}_s) dW_s - \frac{1}{2\epsilon} \int_0^t \|v(s,\hat{X}_s)\|_2^2 ds}.$$

we find that now for $s \leq t$,

$$w(s,x) = \mathbf{E}_{s,x} \left[e^{-2g(\hat{X}_t)/\epsilon} Z_t^2 \right]$$

solves the PDE

$$\partial_s w + b \cdot Dw + \frac{\epsilon}{2} \text{trace} \left(\sigma \sigma^{\mathsf{T}} D^2 w \right) + \frac{1}{\epsilon} \|v\|_2^2 w - \frac{1}{2} \sigma v \cdot Dw = 0, \quad w(t,x) = e^{-2g(x)/\epsilon}.$$

We can again take the log transformation, $V^\epsilon = -\epsilon \log w$ and find that

$$\partial_s V^\epsilon + b \cdot DV^\epsilon + \frac{\epsilon}{2}\text{trace}\left(\sigma\sigma^\mathsf{T} D^2 V^\epsilon\right) - \frac{1}{2}\|v + \sigma DV^\epsilon\|_2^2 - \frac{1}{2}\|v\|_2^2 = 0, \quad V^\epsilon(t,x) = 2g(x).$$

When $v(s,x) = -\sigma^\mathsf{T}(x) DG^\epsilon(t-s,x)$, the equation is solved by $V^\epsilon(s,x) = 2G^\epsilon(t-s,x)$.

Taking the $\epsilon \to 0$ limit we find that $V^\epsilon \to V$ solving

$$\partial_s V + b \cdot DV - \frac{1}{2}\|v + \sigma DV\|_2^2 - \frac{1}{2}\|v\|_2^2 = 0, \quad V(t,x) = 2g(x).$$

$$2G(t,x) - V(0,x) = -\lim_{\epsilon \to 0} \epsilon \log \mathbf{E}\left[f(X_t)\right]^2 + \lim_{\epsilon \to 0} \epsilon \log \mathbf{E}\left[f^2(\hat{X}_t) Z_t^2\right]$$

is always positive and is the rate of exponential growth of the **mse** with ϵ^{-1}.

As a first step we should try to choose v so that this difference is 0.

Not surprisingly $v(s,x) = -\sigma^{\mathsf{T}}(x) DG(t-s,x)$ works.

Subsolutions work too. Suppose for some pair F, v,

$$\partial_s F + b \cdot DF - \frac{1}{2}\|v + \sigma DF\|_2^2 - \frac{1}{2}\|v\|_2^2 \geq 0, \quad F(t,x) \leq 2g(x).$$

A comparison principle tells us then that $F \leq V$. So if we also have $F(0,x) \geq 2G(t,x)$ then it must be that $V(0,x) = 2G(t,x)$.

Ideally we'd have more than just $V(0,x) = 2G(t,x)$ (subexponential in ϵ^{-1} growth of **mse**). For example, we'd like the **mse** to remain bounded as $\epsilon \to 0$.

It turns out that if we choose $v(s,x) = -\sigma^{\mathsf{T}}(x)DG(t-s,x)$ then, for most initial conditions, **mse** will remain bounded or even go to zero.

For $v(s,x) = -\sigma^{\mathsf{T}}(x) DG(t-s,x)$,

$$\hat{X}_t = x + \int_0^t \left(b(\hat{X}_s) + \sigma(\hat{X}_s)\hat{u}_s(s,\hat{X}_s) \right) ds + \sqrt{\epsilon} \int_0^t \sigma(\hat{X}_s) dW_s$$

where

$$\hat{u}(s,x) = \arg\min_u \left\{ \int_s^t \frac{1}{2} \|u_r\|_2^2 dr + g(\phi_t^u) \right\}$$

with

$$\dot{\phi}_r^u = b(\phi_r^u) + \sigma(\phi_r^u) u_r, \quad \phi_r^u(s) = x.$$