Unbiased Estimation without Exact Simulation

(Rhee & Glynn, 2013; Glynn & Rhee, 2014)

Gatsby

September 19, 2014

Introduction

- Quantity of interest: $\mathbb{E}_\pi f(X) = \int f(x) d\pi(x)$
  - Generate random objects of interest $X_1, \ldots, X_c \sim \pi$, and use the empirical mean: $\frac{1}{c} \sum_{i=1}^{c} f(X_i)$
  - Canonical $O(c^{-1/2})$ convergence rate / estimation error, for computational budget c

Introduction

- Quantity of interest: $\mathbb{E}_\pi f(X) = \int f(x) d\pi(x)$
  - Generate random objects of interest $X_1, \ldots, X_c \sim \pi$, and use the empirical mean: $\frac{1}{c}\sum_{i=1}^{c} f(X_i)$
  - Canonical $O(c^{-1/2})$ convergence rate / estimation error, for computational budget c
- How to generate the objects of interest? Often can obtain arbitrarily close approximations to π, but closer the approximation, more computation is required... $\rightarrow$ slower convergence rate

Introduction

- Quantity of interest: $\mathbb{E}_\pi f(X) = \int f(x) d\pi(x)$
  - Generate random objects of interest $X_1, \ldots, X_c \sim \pi$, and use the empirical mean: $\frac{1}{c}\sum_{i=1}^{c} f(X_i)$
  - Canonical $O(c^{-1/2})$ convergence rate / estimation error, for computational budget c
- How to generate the objects of interest? Often can obtain arbitrarily close approximations to π, but closer the approximation, more computation is required... $\rightarrow$ slower convergence rate
- Approximation error / **bias** much harder to quantify.

Introduction

- Quantity of interest: $\mathbb{E}_\pi f(X) = \int f(x) d\pi(x)$
  - Generate random objects of interest $X_1, \ldots, X_c \sim \pi$, and use the empirical mean: $\frac{1}{c}\sum_{i=1}^{c} f(X_i)$
  - Canonical $O(c^{-1/2})$ convergence rate / estimation error, for computational budget c
- How to generate the objects of interest? Often can obtain arbitrarily close approximations to π, but closer the approximation, more computation is required... $\rightarrow$ slower convergence rate
- Approximation error / **bias** much harder to quantify.
- **How to turn sequences of well behaved biased estimators into an unbiased estimator without sacrificing the convergence rate?**

Example 1: SDEs

- Quantity of interest: $\mathbb{E}_\pi f(X)$
- $X = (X(t) : t \geq 0)$ is the solution to

$$dX = \mu(X)dt + \sigma(X)dB,$$

- Cannot generate X exactly, but can use discrete-time approximation X_h, e.g., by Euler discretization with grid $0, h, 2h, \ldots$
- $f(X_h)$ is a biased estimator with bias that drops with h, and comes with the cost of $\Theta(1/h)$
- Select h and the number of replications carefully to balance bias and variance $\rightarrow$ slower convergence rate

Example 2: MCMC

- Markov chain $\{X_n\}_{n\geq 0}$ with equilibrium distribution π
- Has it equilibriated yet? Want to estimate $\mathbb{E}f(X_\infty)$, but only have finite time. How to quantify the bias?

Assumptions

- Let $Y = f(X)$ be a real-valued random variable with a finite second moment.
- Let $\{Y_t = f(X_t)\}_{t=1}^{\infty}$ be a sequence of real-valued random variables with finite second moments. Denote $Y_0 \equiv 0$.

Assumption (A1)

$\lim_{t\to\infty} \mathbb{E}\,|Y_t - Y|^2 = 0$, i.e., $Y_t \xrightarrow{L^2} Y$ (Y_t converges to Y in quadratic mean).

- stronger than convergence of first two moments, i.e., $\lim_{t\to\infty} \mathbb{E}Y_t = \mathbb{E}Y$, $\lim_{t\to\infty} \mathbb{E}Y_t^2 = \mathbb{E}Y^2$

Assumptions

- Let T be an integer-valued random variable independent of Y and $\{Y_t\}_{t=1}^{\infty}$ with $\mathbb{P}[T \geq t] > 0 \ \forall t \in \mathbb{N}$.

Assumption (A1+)

$\sum_{t=0}^{\infty} \frac{\mathbb{E}|Y_{t-1}-Y|^2}{\mathbb{P}[T \geq t]} < \infty$ (thus not only that $Y_t \xrightarrow{L^2} Y$ but convergence happens faster than the tail of T decreases).

Telescoping estimator

Theorem

Assuming ***(A1+)***,

$$Z = Z(T) = \sum_{t=1}^{T} \frac{Y_t - Y_{t-1}}{\mathbb{P}[T \geq t]}$$

is an unbiased estimator of $\mathbb{E}Y$ *with*

$$\mathbb{E}Z^2 = \sum_{t=1}^{\infty} \frac{\mathbb{E}|Y_{t-1} - Y|^2 - \mathbb{E}|Y_t - Y|^2}{\mathbb{P}[T \geq t]}.$$

- Under **(A1+)**, variance is finite and easily estimated by replication, i.e., using $\mathrm{Var}[Z_1, \ldots, Z_m]$, where $Z_j = Z(T_j)$ for i.i.d. $T_1, \ldots, T_m$, which gives confidence intervals for $\mathbb{E}Y$ → easy to construct algorithms with desired error tolerance.

Telescoping estimator

Theorem

Assuming ***(A1+)***,

$$Z = Z(T) = \sum_{t=1}^{T} \frac{Y_t - Y_{t-1}}{\mathbb{P}[T \geq t]} = \sum_{t=1}^{T} \frac{\Delta_t}{\mathbb{P}[T \geq t]}$$

is an unbiased estimator of $\mathbb{E}Y$ *with*

$$\mathbb{E}Z^2 = \sum_{t=1}^{\infty} \frac{\mathbb{E}|Y_{t-1} - Y|^2 - \mathbb{E}|Y_t - Y|^2}{\mathbb{P}[T \geq t]} = \sum_{t=1}^{\infty} \frac{\mathbb{E}\Delta_t^2 + 2\sum_{s=t+1}^{\infty} \mathbb{E}\Delta_t \Delta_s}{\mathbb{P}[T \geq t]}.$$

- Variance depends on the joint distribution of Y_t's only throught the L^2 norms of $Y_t - Y$: only the marginal distribution of Y_t affects the algorithm $\rightarrow$ we can often replace Y_{t-1} with $Y'_{t-1} \stackrel{D}{=} Y_{t-1}$, which will drive $\Delta_t = Y_t - Y'_{t-1}$ faster to 0.

Proof sketch

- $Z'_r = Z'_r(T) = \sum_{t=1}^{\min(T,r)} \frac{Y_t - Y_{t-1}}{\mathbb{P}[T \geq t]}$.

$$\begin{aligned} \mathbb{E}Z'_r &= \mathbb{E}\sum_{t=1}^{r} \frac{\mathbf{1}\left[T \geq t\right]}{\mathbb{P}\left[T \geq t\right]}\left(Y_t - Y_{t-1}\right) \\ &= \sum_{t=1}^{r} \mathbb{E}\left(Y_t - Y_{t-1}\right) \\ &= \mathbb{E}Y_r. \end{aligned}$$

- $Z'_r \stackrel{a.s.}{\to} Z$ as $r \to \infty$
- construct a subsequence of $\{Z'_r\}_{r=1}^{\infty}$ that must converge in L^2 using **(A1+)**. This L^2-limit then must be Z, so:

$$\mathbb{E}Y \leftarrow \mathbb{E}Y_r \quad = \quad \mathbb{E}Z'_r \to \mathbb{E}Z$$

Work-variance tradeoff

- t_n - time required to generate Y_n
- τ - time required to generate each Z:
  $\mathbb{E}\tau = \mathbb{E}\left[t_1 + \ldots + t_T\right] = \sum_{j=1}^{\infty} t_j \mathbb{P}\left[T \geq j\right]$
- $\text{Var}Z = \sum_{j=1}^{\infty} \frac{\mathbb{E}\Delta_j^2 + 2\sum_{s=j+1}^{\infty} \mathbb{E}\Delta_j\Delta_s}{\mathbb{P}[T \geq j]} = \sum_{j=1}^{\infty} b_j / \mathbb{P}\left[T \geq j\right]$
- For a given computational budget c, denote by $m(c)$ the number of replicates of Z we can generate in c time,
  $m(c) = \max\left\{m \geq 0 \,:\, \sum_{j=1}^{m} \tau_j \leq c\right\}$, and let $\bar{Z}_{(c)} = \frac{1}{m(c)} \sum_{j=1}^{m(c)} Z_j$
- With $m(c)$ replicates, workload is $m(c)\mathbb{E}\tau$, and the variance is $\frac{\text{Var}\, Z}{m(c)}$.
- From (Glynn and Whitt 1992) it follows that if $\mathbb{E}\tau < \infty$ and $\text{Var}Z < \infty$ then

$$\sqrt{c}\left(\bar{Z}_{(c)} - \mathbb{E}Y\right) \rightsquigarrow \mathcal{N}\left(0, \mathbb{E}\tau \text{Var}Z\right).$$

Work-variance tradeoff in SDEs

- Y_n corresponds to the discretization with increment $h = 2^{-n}$ (doubling the number of time steps).
- $t_n = \Theta(2^n)$, and typically $b_n = \mathbb{E}\,|Y_{n-1} - Y|^2 - \mathbb{E}\,|Y_n - Y|^2 = O(2^{-2np})$, where $p > 0$ is the strong order of the discretization scheme
- choose N so that $\mathbb{P}\,[N \geq n] = 2^{-nr}$, with $1 < r < 2p$, for $p > 1/2$.
- Then:
  - $\mathrm{Var} Z = \sum_{n=1}^{\infty} b_n / \mathbb{P}\,[N \geq n] = O\left(\sum_{n=1}^{\infty} 2^{-n(2p-r)}\right) < \infty$
  - $\mathbb{E}\tau = \sum_{n=1}^{\infty} t_n \mathbb{P}\,[N \geq n] = O\left(\sum_{n=1}^{\infty} 2^{-n(r-1)}\right) < \infty$

Work-variance tradeoff in SDEs

- Y_n corresponds to the discretization with increment $h = 2^{-n}$ (doubling the number of time steps).
- $t_n = \Theta(2^n)$, and typically $b_n = \mathbb{E}\left|Y_{n-1} - Y\right|^2 - \mathbb{E}\left|Y_n - Y\right|^2 = O(2^{-2np})$, where $p > 0$ is the strong order of the discretization scheme
- choose N so that $\mathbb{P}\left[N \geq n\right] = 2^{-nr}$, with $1 < r < 2p$, for $p > 1/2$.
- Then:
  - $\mathrm{Var} Z = \sum_{n=1}^{\infty} b_n / \mathbb{P}\left[N \geq n\right] = O\left(\sum_{n=1}^{\infty} 2^{-n(2p-r)}\right) < \infty$
  - $\mathbb{E}\tau = \sum_{n=1}^{\infty} t_n \mathbb{P}\left[N \geq n\right] = O\left(\sum_{n=1}^{\infty} 2^{-n(r-1)}\right) < \infty$
- Thus, $c^{-\frac{1}{2}}$ convergence. The fastest previous rate is $c^{-\frac{p}{2p+1}}$.

Unequilibriated MCMC

- Markov chain $\{X_n\}_{n\geq 0}$, with i.i.d. transitions φ_n, with $X_n = \varphi_n(X_{n-1})$
- Problem: if we just generate $\{X_n\}_{n\geq 0}$, and set $\Delta_n = f(X_n) - f(X_{n-1})$ for debiasing, then we would need $\Delta_n \to 0$ in L^2, which does not happen
- Idea: need to couple the values in Δ_n, i.e., replace X_{n-1} with $\tilde{X}_{n-1}$, s.t., $\tilde{X}_{n-1} \overset{\mathcal{D}}{=} X_{n-1}$, but $\tilde{X}_{n-1}$ is close to X_n

Coupling for contractive chains

- **(C1)** Chain is contractive on average: $\mathbb{E}\left\|\varphi_1(x) - \varphi_1(x')\right\|^2 \leq b\left\|x - x'\right\|^2$, for some $b < 1$.
- **(C2)** Function f is Lipschitz: $|f(x) - f(x')| \leq \kappa\left\|x - x'\right\|$, for some $\kappa < \infty$.
- Now, set

$$\begin{aligned} X_n &= (\varphi_n \circ \varphi_{n-1} \circ \cdots \circ \varphi_1)(x) \\ \tilde{X}_{n-1} &= (\varphi_n \circ \varphi_{n-1} \circ \cdots \circ \varphi_2)(x) \end{aligned}$$

  Note: $\left(X_n, \tilde{X}_{n-1}\right)$ can be recursively computed from $\left(X_{n-1}, \tilde{X}_{n-2}\right)$.
- Set $\Delta_n = f(X_n) - f(\tilde{X}_{n-1})$.

Coupling for contractive chains (2)

$$
\begin{aligned}
\mathbb{E}\Delta_n^2 &\leq \kappa^2 \mathbb{E}\left\|X_n - \tilde{X}_{n-1}\right\|^2 \\
&= \kappa^2 \mathbb{E}\left\|\varphi_n\left(X_{n-1}\right) - \varphi_n\left(\tilde{X}_{n-2}\right)\right\|^2 \\
&\leq \kappa^2 b \mathbb{E}\left\|X_{n-1} - \tilde{X}_{n-2}\right\|^2 \leq \cdots \\
&\leq \kappa^2 b^{n-1} \mathbb{E}\left\|X_1 - x\right\|^2 \to 0
\end{aligned}
$$

geometrically fast, so can match with appropriate distribution of T.

Alternative coupling

$$
\begin{aligned}
X_1^* &= \varphi_N(x) \\
X_2^* &= (\varphi_N \circ \varphi_{N-1})(x) \\
X_n^* &= (\varphi_N \circ \varphi_{N-1} \circ \cdots \circ \varphi_{N-n+1})(x)
\end{aligned}
$$

More complicated to implement as cannot recursively compute X_n^* from X_{n-1}^*. Computational effort quadratic in N. Also: different variance, since $\mathbb{E}\Delta_j\Delta_k \neq \mathbb{E}\Delta_j^*\Delta_k^*$.

Glivenko-Cantelli result

- Sometimes, not only interested in equilibrium expectation $\mathbb{E}f(X)$, but in the equilibrium distribution of $f(X)$
- How to estimate equilibrium cdf $F_\infty(y) = \mathbb{P}[f(X) \leq y] = \mathbb{E}\mathbf{1}[f(X) \leq y]$? Note that$\mathbf{1}[f(\cdot) \leq y]$ is not Lipschitz.
- In the context of exact simulation: generate $X_1, \ldots, X_c \sim \pi$, and set $Y_j = f(X_j)$. The empirical distribution function is given by $F_m(y) = \frac{1}{m}\sum_{j=1}^m \mathbf{1}[Y_j \leq y]$.
- Glivenko-Cantelli theorem shows uniform convergence of $F_m(y)$ to $F(y)$ if Y's are i.i.d.

$$\sup_y |F_m(y) - F_\infty(y)| \to 0, \qquad \text{a.s.}$$

Glivenko-Cantelli result

- The empirical (signed) measure induced by the debiasing scheme

$$\gamma_m(\cdot) \quad = \quad \frac{1}{m}\sum_{j=1}^{m}\left(\sum_{n=1}^{N_j}\frac{\delta_{Y_{n,j}}(\cdot)-\delta_{\tilde{Y}_{n-1,j}}(\cdot)}{\mathbb{P}[N\geq n]}\right),$$

and thereby empirical debiased distribution

$$F_m(y) \quad = \quad \frac{1}{m}\sum_{j=1}^{m}\left(\sum_{n=1}^{N_j}\frac{\mathbf{1}[Y_{n,j}\leq y]-\mathbf{1}\left[\tilde{Y}_{n-1,j}\leq y\right]}{\mathbb{P}[N\geq n]}\right).$$

- Glivenko-Cantelli theorem still holds: $\sup_y |F_m(y)-F_\infty(y)| \to 0$ a.s.

Summary

- **P. W. Glynn and C. Rhee, Exact Estimation for Markov Chain Equilibrium Expectations, 2014**
- **C. Rhee and P. W. Glynn, Unbiased Estimation with Square Root Convergence for SDE Models, 2013 arXiv:1207.2452**

- Exact estimation can be easier than exact simulation
- No bias and controllable variance - canonical convergence rate as a function of the computational budget
- Easy to handle work-variance tradeoff
- It can work under less restrictive conditions (π-irreducibility vs positive Harris recurrence in the Markov chain example)