Advanced Simulation Methods

Chapter 3 - Importance Sampling and Variance Reduction Methods

1 Importance Sampling

In the rejection sampling algorithm, we simulate from a distribution π by sampling from a proposal distribution q and rejecting some of the proposed values. Importance sampling uses another correction scheme based on reweighting. In this context the proposal q is also known as an importance distribution.

1.1 Standard Importance Sampling

Let q, π be two pdfs on $\mathbb{X}$ such that $\pi(x) > 0 \Rightarrow q(x) > 0$. Then, for any[1] set A such that $\pi(A) > 0$

$$\begin{aligned}\pi(A) &:= \int_A \pi(x)\,dx \\ &= \int_A \underbrace{\frac{\pi(x)}{q(x)}}_{:=w(x)} q(x)\,dx \\ &= \int_A w(x)\,q(x)\,dx\end{aligned}$$

where $w : \mathbb{X} \to \mathbb{R}^+$ is the so-called importance weight function. This identity can be obviously generalised to the expectation of any function. Assume $\pi(x)\,\phi(x) > 0 \Rightarrow q(x) > 0$, then

$$\begin{aligned}I = \mathbb{E}_\pi(\phi(X)) &= \int_{\mathbb{X}} \phi(x)\,\pi(x)\,dx \\ &= \int_{\mathbb{X}} \phi(x)\,w(x)\,q(x)\,dx \\ &= \mathbb{E}_q(\phi(X)w(X)).\end{aligned}$$

Now let $X_1, ..., X_n$ be a sample of independent random variables distributed according to q then the estimator

$$\widehat{I}_n^{\mathrm{IS}} = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)w(X_i)$$

is consistent through the strong law of large numbers if $\mathbb{E}_q(|\phi(X)|\,w(X)) < \infty$. We also obtain easily the following result.

Proposition 1 ***(Bias and Variance of Standard Importance Sampling)***

(a) $\mathbb{E}_q\left(\widehat{I}_n^{\mathrm{IS}}\right) = I$,

(b) $\mathbb{V}_q\left(\widehat{I}_n^{\mathrm{IS}}\right) = \frac{1}{n}\mathbb{V}_q\left(\phi(X)w(X)\right)$ and if $\sigma^2_{\mathrm{IS}} := \mathbb{V}_q\left(\phi(X)w(X)\right) < \infty$

$$\sqrt{n}\left(\widehat{I}_n^{\mathrm{IS}} - I\right) \xrightarrow{\mathrm{D}} \mathcal{N}\left(0, \sigma^2_{\mathrm{IS}}\right)$$

Remark. A sufficient condition for $\mathbb{V}_q\left(\widehat{I}_n^{\mathrm{IS}}\right)$ to be finite is to have $\mathbb{V}_\pi(\phi(X))$ finite and $\pi(x)/Mq(x) \leq M < \infty$ for any $x \in \mathbb{X}$.

A natural question consists of choosing what is the best proposal distribution to minimize $\mathbb{V}_q\left(\widehat{I}_n^{\mathrm{IS}}\right)$.

[1] For $\mathbb{X} = \mathbb{R}^d$, we consider the Borel sigma algebra $\mathcal{F} = \mathcal{B}\left(\mathbb{R}^d\right)$, $A \in \mathcal{F}$ and the density is with respect to the Lebesgue measure dx.

Proposition 2 *The optimal proposal minimising* $\mathbb{V}_q\left(\widehat{I}_n^{IS}\right)$ *is given by*

$$q_{opt}(x) = \frac{|\phi(x)|\,\pi(x)}{\int_{\mathbb{X}}|\phi(x)|\,\pi(x)\,dx}.$$

Proof. We have indeed

$$\mathbb{V}_q\left(\phi(X)w(X)\right) = \mathbb{E}_q\left(\phi^2(X)w^2(X)\right) - I^2.$$

For $q = q_{\text{opt}}$, we have

$$\begin{aligned}\mathbb{E}_{q_{\text{opt}}}\left(\phi^2(X)w^2(X)\right) &= \int_{\mathbb{X}}\frac{\phi^2(x)\pi^2(x)}{|\phi(x)|\,\pi(x)}dx.\int_{\mathbb{X}}|\phi(x)|\,\pi(x)\,dx \\ &= \left(\int_{\mathbb{X}}|\phi(x)|\,\pi(x)\,dx\right)^2\end{aligned}$$

We also have by Jensen's inequality

$$\mathbb{E}_q\left(\phi^2(X)w^2(X)\right) \geq \mathbb{E}_q^2\left(|\phi(X)|\,w(X)\right) = \left(\int_{\mathbb{X}}|\phi(x)|\,\pi(x)\,dx\right)^2$$

so we can conclude. ■

This optimal variance estimator cannot typically be implemented; e.g for $\phi(x) > 0$ we have $q_{\text{opt}}(x) = \phi(x)\pi(x)/I$ and $\mathbb{V}_{q_{\text{opt}}}\left(\widehat{I}_n^{\text{IS}}\right) = 0$ but this cannot be implemented as this required knowing I! This can be however use as a guideline to select q; i.e. select $q(x)$ such that it approaches $q_{\text{opt}}(x)$ in some respect.

1.2 Normalised Importance Sampling

Practically standard importance sampling has limited applications as it requires knowing $\pi(x)$ exactly contrary to rejection sampling where $\pi(x)$ and $q(x)$ can be known only up to some normalising constants. However there is an alternative version of importance sampling known as normalised importance sampling which bypasses this problem. It relies on the following identity which holds whenever $\pi(x) > 0 \Rightarrow q(x) > 0$

$$\begin{aligned}I = \mathbb{E}_\pi(\phi(X)) &= \int_{\mathbb{X}}\phi(x)\,\pi(x)\,dx \\ &= \frac{\int_{\mathbb{X}}\phi(x)\,w(x)\,q(x)\,dx}{\int_{\mathbb{X}}w(x)\,q(x)\,dx} \\ &= \frac{\mathbb{E}_q(\phi(X)w(X))}{\mathbb{E}_q(w(X))}.\end{aligned}$$

Now let $X_1, ..., X_n$ be a sample of independent random variables distributed according to q then the estimator

$$\widehat{I}_n^{\text{NIS}} = \frac{\sum_{i=1}^n \phi(X_i)w(X_i)}{\sum_{i=1}^n w(X_i)}$$

is consistent through the strong law of large numbers as long as $\mathbb{E}_q(|\phi(X)|\,w(X)) < \infty$.

The normalised importance sampling estimator $\widehat{I}_n^{\text{NIS}}$ is a ratio of two estimators so we do not have simple expressions for its finite bias and variance but we can obtain their asymptotic (i.e. as $n \to \infty$) expression by relying on the delta method.

Proposition 3 ***(The multivariate Delta method).*** *Suppose* $Z_n = (Z_{n1}, ..., Z_{nk})$ *is a sequence of random vectors such that*

$$\sqrt{n}\left(Z_n - \mu\right) \xrightarrow{D} \mathcal{N}(0, \Sigma).$$

Let $g : \mathbb{R}^k \to \mathbb{R}$ *and let*

$$\nabla g = \left(\frac{\partial g}{\partial z_1} \cdots \frac{\partial g}{\partial z_k}\right)^T.$$

Let $\nabla g(\mu)$ *be* ∇g *evaluated* $z = \mu$ *and assume the elements of* $\nabla g(\mu)$ *are non-zero then*

$$\sqrt{n}\left(g(Z_n) - g(\mu)\right) \to \mathcal{N}\left(0, \nabla^T g(\mu)\ \Sigma\ \nabla g(\mu)\right).$$

Proposition 4 ***(CLT for Normalised Importance Sampling)***

Assume that $\mathbb{V}_q(\phi(X)w(X)) < \infty$ and $\mathbb{V}_q(w(X)) < \infty$ then

$$\sqrt{n}\left(\widehat{I}_n^{\mathrm{NIS}} - I\right) \xrightarrow{\mathrm{D}} \mathcal{N}\left(0, \sigma^2_{\mathrm{NIS}}\right)$$

We know that $\widehat{I}_n^{\mathrm{IS}}$ is unbiased whereas $\widehat{I}_n^{\mathrm{NIS}}$ is not. We give here an expression for the asymptotic bias.

Proposition 5 ***(Asymptotic Bias)***. *Assume that $\mathbb{V}_q(\phi(X)w(X)) < \infty$ and $\mathbb{V}_q(w(X)) < \infty$ then we have*

$$\begin{aligned}\lim_{n\to\infty} n\mathbb{E}_q\left(\widehat{I}_n^{NIS} - I\right) &= -cov_q\left(\phi(X)w(X), w(X)\right) + \mathbb{V}_q(w(X))I \\ &= -\int (\phi(x) - I)\frac{\pi^2(x)}{q(x)}dx.\end{aligned}$$

Remark. The bias being of order $1/n$, we can conclude that the mean square error of $\widehat{I}_n^{\mathrm{NIS}}$ is asymptotically governed by its variance term.

Example 1 *(Bayesian analysis of a Markov chain) Consider a two-state discrete time Markov chain (X_t) with transition matrix*

$$\begin{pmatrix} \alpha_1 & 1-\alpha_1 \\ 1-\alpha_2 & \alpha_2 \end{pmatrix}$$

that is $\mathbb{P}(X_{t+1} = 1 | X_t = 1) = 1 - \mathbb{P}(X_{t+1} = 2 | X_t = 1) = \alpha_1$ and $\mathbb{P}(X_{t+1} = 2 | X_t = 2) = 1 - \mathbb{P}(X_{t+1} = 1 | X_t = 2) = \alpha_2$. We assume that some physical constraints tell us that $\alpha_1 + \alpha_2 < 1$. Assume we observe $(X_1, ..., X_m) = (x_1, ..., x_m)$ and want to perform Bayesian inference about (α_1, α_2). We set the following prior

$$p(\alpha_1, \alpha_2) = 2\mathbb{I}_{\alpha_1+\alpha_2\leq 1}$$

then the posterior of interest is

$$p(\alpha_1, \alpha_2 | x_{1:m}) \propto \alpha_1^{m_{1,1}}(1-\alpha_1)^{m_{1,2}}(1-\alpha_2)^{m_{2,1}}\alpha_2^{m_{2,2}}\mathbb{I}_{\alpha_1+\alpha_2\leq 1}$$

where

$$m_{i,j} = \sum_{t=1}^{m-1} \mathbb{I}_{x_t=i}\mathbb{I}_{x_{t+1}=i}$$

The posterior does not admit a standard expression and its normalizing constant is unknown.

We are interested in estimating $\mathbb{E}[\varphi_i(\alpha_1, \alpha_2) | x_{1:m}]$ for $\varphi_1(\alpha_1, \alpha_2) = \alpha_1$, $\varphi_2(\alpha_1, \alpha_2) = \alpha_2$, $\varphi_3(\alpha_1, \alpha_2) = \alpha_1/(1-\alpha_1)$, $\varphi_4(\alpha_1, \alpha_2) = \alpha_2/(1-\alpha_2)$ and $\varphi_5(\alpha_1, \alpha_2) = \log\frac{\alpha_1(1-\alpha_2)}{\alpha_2(1-\alpha_1)}$.

We can sample from the posterior through rejection sampling using the prior as a proposal but this can be highly inefficient if m is large. We discuss various possible Importance Sampling proposals.

If there was no constraint on α_1, α_2 and $p(\alpha_1, \alpha_2)$ was uniform on $[0,1]\times[0,1]$, then the posterior would be

$$\begin{aligned}q_0(\alpha_1, \alpha_2) &= \mathcal{B}eta(\alpha_1; m_{1,1}+1, m_{1,2}+1) \\ &\quad \times \mathcal{B}eta(\alpha_2; m_{2,2}+1, m_{2,1}+1)\end{aligned}$$

so we could use this as importance distribution. Unfortunately, this is very inefficient, as for the given data $(m_{1,1}, m_{1,2}, m_{2,2}, m_{2,1})$ we have $q_0(\alpha_1 + \alpha_2 < 1) = 0.21$.

The form of the posterior also suggests using a Dirichlet distribution with density

$$q_1(\alpha_1, \alpha_2) \propto \alpha_1^{m_{1,1}}\alpha_2^{m_{2,2}}(1-\alpha_1-\alpha_2)^{m_{1,2}+m_{2,1}}$$

but $p(\alpha_1, \alpha_2 | x_{1:m})/q_1(\alpha_1, \alpha_2)$ is unbounded.

(Geweke, 1989) proposed using the normal approximation to the binomial distribution

$$q_2(\alpha_1,\alpha_2) \propto \exp\left(-(m_{1,1}+m_{1,2})(\alpha_1-\widehat{\alpha}_1)^2/(2\widehat{\alpha}_1(1-\widehat{\alpha}_1))\right)$$
$$\times \exp\left(-(m_{2,1}+m_{2,2})(\alpha_2-\widehat{\alpha}_2)^2/(2\widehat{\alpha}_2(1-\widehat{\alpha}_2))\right)\mathbb{I}_{\alpha_1+\alpha_2\leq 1}$$

where $\widehat{\alpha}_1 = m_{1,1}/(m_{1,1}+m_{1,2})$, $\widehat{\alpha}_1 = m_{2,2}/(m_{2,2}+m_{2,1})$. *To simulate from* $q_2(\alpha_1,\alpha_2)$, *we simulate first* $q_2(\alpha_1)$ *and then* $q_2(\alpha_2|\alpha_1)$ *which are univariate truncated Gaussian distributions. The ratio* $p(\alpha_1,\alpha_2|x_{1:m})/q_2(\alpha_1,\alpha_2)$ *is upper bounded.*

A final possible choice of q *consists of using*

$$q_3(\alpha_1,\alpha_2) = \mathcal{B}eta(\alpha_1; m_{1,1}+1, m_{1,2}+1)\, q_3(\alpha_2|\alpha_1)$$

where $p(\alpha_2|x_{1:m},\alpha_1) \propto (1-\alpha_2)^{m_{2,1}} p_2^{m_{2,2}} \mathbb{I}_{\alpha_2\leq 1-\alpha_1}$ *is (badly) approximated through* $q_3(\alpha_2|x_{1:m},\alpha_1) = \frac{2}{(1-\alpha_1)^2}\alpha_2 \mathbb{I}_{\alpha_2\leq 1-\alpha_1}$. *It is straightforward to check that* $p(\alpha_1,\alpha_2|x_{1:m})/q_3(\alpha_1,\alpha_2) \propto (1-\alpha_2)^{m_{2,1}} p_2^{m_{2,2}}/\frac{2}{(1-\alpha_1)^2}\alpha_2 < \infty$ *whenever* $m_{2,2} \geq 1$.

We present the empirical standard deviation of 4 of the sampling distributions for $N = 10,000$ *samples.*

Distribution	φ_1	φ_2	φ_3	φ_4	φ_5
q_1	*0.748*	*0.139*	*3.184*	*0.163*	*2.957*
q_2	*0.689*	*0.210*	*2.319*	*0.283*	*2.211*
q_3	*0.697*	*0.189*	*2.379*	*0.241*	*2.358*
π	*0.697*	*0.189*	*2.373*	*0.240*	*2.358*

Sampling from π *using rejection sampling works well but is computationally expensive.* q_3 *is computationally much cheaper whereas* q_1 *does extremely poorly as expected.*

2 Antithetic Variates

We are interested in computing

$$I = \int_0^1 \phi(x)dx = \mathbb{E}(\phi(U)),\ U \sim \mathcal{U}_{[0,1]}$$

Instead of

$$\widehat{I}_n = \frac{1}{n}\sum_{i=1}^n \phi(U_i),$$

we consider here

$$\overline{I}_n = \frac{1}{2n}\sum_{i=1}^n (\phi(U_i)+\phi(1-U_i)).$$

We obtain

$$\mathbb{V}\left(\overline{I}_n\right) = \frac{n}{4n^2}\mathbb{V}(\phi(U)+\phi(1-U))$$
$$= \frac{1}{2n}(\mathbb{V}(\phi(U)) + cov(\phi(U),\phi(1-U))).$$

If $\text{cov}(\phi(U),\phi(1-U)) < 0$, $\mathbb{V}\left(\overline{I}_n\right) \leq \mathbb{V}\left(\widehat{I}_n\right)$. The following lemma gives conditions for this to hold.

Lemma 1 *If the function* ϕ *is monotonic, then* $cov(\phi(U),\phi(1-U)) < 0$, *unless* ϕ *is constant on* $[0,1]$.

Proof. Let U_1, U_2 be independent and uniformly distributed on $[0,1]$. Then we have

$$cov(\phi(U),\phi(1-U)) = \frac{1}{2}\mathbb{E}[(\phi(U_1)-\phi(U_2))(\phi(1-U_1)-\phi(1-U_2))].$$

We assume that ϕ is monotonically increasing. If $U_1 < U_2$, then the first factor is negative and the second positive, and vice versa for $U_1 > U_2$. Thus, the integrand is always non-positive. To verify that the covariance is strictly negative, we investigate when the integrand is zero. One factor must be 0, that is almost surely either $\phi(U_1) = \phi(U_2)$ or $\phi(1-U_1) = \phi(1-U_2)$. Because ϕ is monotone, this is only possible if ϕ is constant.

3 Control Variates

Assume there exists a function φ such that $\int \varphi(x)\,\pi(x)\,dx$ is known and we want to compute $I = \int \phi(x)\pi(x)\,dx$. Without loss of generality, assume further that $\int \varphi(x)\,\pi(x)\,dx = 0$. Then for any λ

$$\widehat{I}_{n,c} = \frac{1}{n}\sum_{i=1}^{n}\left(\phi(X_i) - \lambda\varphi(X_i)\right)$$

is an unbiased estimator of I for $X_i \overset{\text{i.i.d}}{\sim} \pi$. Its variance is

$$\begin{aligned}\mathbb{V}\left(\widehat{I}_{n,c}\right) &= \frac{1}{n}\mathbb{V}\left(\phi(X_i) - \lambda\varphi(X_i)\right)\\ &= \frac{1}{n}\left\{\mathbb{V}\left(\phi(X_i)\right) + \lambda^2\mathbb{V}\left(\varphi(X_i)\right) - 2\lambda\text{cov}\left(\phi(X_i),\varphi(X_i)\right)\right\}.\end{aligned}$$

The optimal λ is

$$\lambda_{\text{opt}} = \frac{\text{cov}\left(\phi(X_i),\varphi(X_i)\right)}{\mathbb{V}\left(\varphi(X_i)\right)}$$

and the minimal variance is

$$\mathbb{V}_{\text{opt}}\left(\widehat{I}_{n,c}\right) = \frac{1}{n}\mathbb{V}\left(\phi(X)\right)\left\{1 - \text{corr}\left(\phi(X),\varphi(X)\right)^2\right\} \leq \frac{1}{n}\mathbb{V}\left(\phi(X)\right).$$

In general, λ_{opt} is unknown, but it can be estimated by

$$\widehat{\lambda}_{\text{opt}} = \frac{\sum_{i=1}^{n}\left(\phi(X_i) - \widehat{I}_n\right)\varphi(X_i)}{\sum_{i=1}^{n}\varphi(X_i)^2}.$$

This is consistent, and we obtain asymptotically the same variance as if λ_{opt} is known.