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 probability density functions on $\mathbb{X}$ such that $\pi(x) > 0 \Rightarrow q(x) > 0$. Then, for any[1] set A such that $\pi(A) > 0$

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

where $w : \mathbb{X} \to \mathbb{R}^+$ is the so-called importance weight function: $w : x \mapsto \pi(x)/q(x)$. 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 the following results.

Proposition 1.1. ***(Bias and Variance of Standard Importance Sampling)***

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

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

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

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 $\sigma^2_{\mathrm{IS}}(\varphi)$.

[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 1.2. *The optimal proposal minimising $\sigma^2_{IS}(\varphi)$ is given by*

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

Proof. We have indeed

$$\sigma^2_{\mathrm{IS}}(\varphi) = \mathbb{E}_q\left(\phi^2(X)w^2(X)\right) - I^2.$$

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

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

We also have by Jensen's inequality

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

so we can conclude. ■

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

Example 1.1. ***(Importance sampling for t-distributions)*** *Assume we are interested in computing*

$$I = \mathbb{E}_\pi(|X|) = \int_{\mathbb{R}} |x|\,\pi(x)dx$$

where π a t_3-distribution, that is, a t-distribution with 3 degrees of freedom. We propose 3 sampling schemes to compute I, using importance sampling.

1. *Use directly π as a sampling distribution;*
2. *use importance sampling with proposal density $q_1 : x \mapsto g_{t_1}(x)$, a t_1-distribution (also called a Cauchy distribution);*
3. *use importance sampling with proposal distribution q_2 being a standard normal distribution.*

Figure 1 illustrates the various choices of proposal distributions. The performance of the estimates are displayed in Figure 2 and the associated sample weights in Figure 3. We see that the normal distribution yields a poor estimate as the variance of the weights is infinite, whereas it can be shown that g_{t_1} yields a smaller variance estimate that π itself.

1.2 Normalised Importance Sampling

In practice, standard importance sampling has limited applications as it requires exact evaluations of $\pi(x)$, contrarily to rejection sampling where $\pi(x)$ and $q(x)$ only have to evaluated up to some normalising constants. However there is an alternative version of importance sampling known as normalised importance sampling which bypasses this problem. Assume that whenever $\pi(x) > 0 \Rightarrow q(x) > 0$, and that we can write $\pi(x) = \tilde{\pi}(x)/Z_\pi$ and $q(x) = \tilde{q}(x)/Z_q$, for some normalising constants Z_π and Z_q. We introduce the unnormalised weight function

$$\tilde{w} : x \mapsto \tilde{\pi}(x)/\tilde{q}(x) = w(x)\frac{Z_\pi}{Z_q}.$$

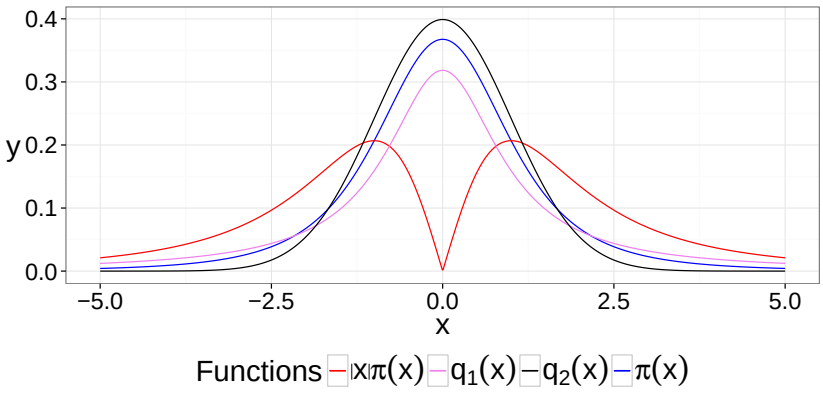

Figure 1: Different importance proposal distributions to estimate the area under the red curve.

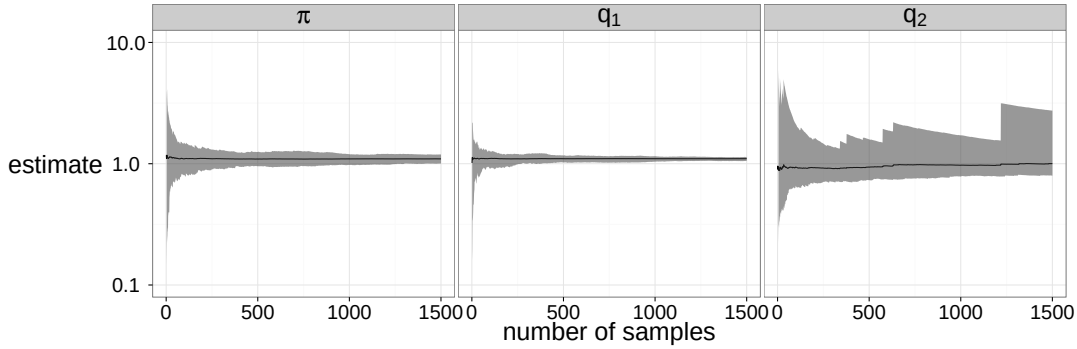

Figure 2: Estimates of I obtained after 1 to 1500 samples, using proposals π (left), q_1 (middle) or q_2 (right). The grey shaded areas correpond to the range of 100 independent replications, and the black line represents the mean.

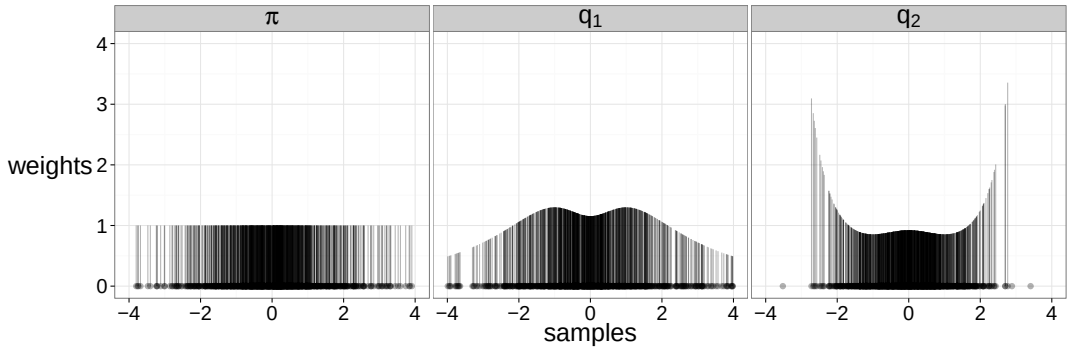

Figure 3: Sample weights obtained for 1000 realisations of X_i, using proposals π (left), q_1 (middle) or q_2 (right).

Then, we can write

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

where the importance weight function only involves the unnormalised probability density functions $\tilde{\pi}$ and $\tilde{q}$.

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

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

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, therefore we do not have simple expressions for its finite bias and variance. We can still obtain their asymptotic expression (i.e. as $n \to \infty$) using the delta method.

Proposition 1.3. ***(The multivariate delta method).*** *Suppose $Z_n = (Z_{n1}, ..., Z_{nk})$ is a sequence of random vectors such that*

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

where Σ is a $k \times k$ covariance matrix. 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 at μ and assume that the elements of $\nabla g\,(\mu)$ are non-zero, then

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

Proposition 1.4. ***(CLT for Normalised Importance Sampling)***

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

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

where

$$\begin{aligned}\sigma^2_{\text{NIS}} &= \mathbb{V}_q(\phi(X)w\,(X)) + I^2\mathbb{V}_q(w\,(X)) - 2I\mathbb{C}\text{ov}_q\,(\phi(X)w\,(X), w\,(X)) \\ &= \int (\phi\,(x) - I)^2\,\frac{\pi^2\,(x)}{q\,(x)}dx.\end{aligned}$$

Proof. We apply the delta method to $Z_n = (Z_{n1}, Z_{n2})$ where

$$Z_{n1} = \frac{1}{n}\sum_{i=1}^n \phi(X_i)w(X_i), \quad Z_{n2} = \frac{1}{n}\sum_{i=1}^n w(X_i)$$

and

$$\widehat{I}_n^{\text{NIS}} = \frac{Z_{n1}}{Z_{n2}} = g\,(Z_n)\,.$$

By the CLT, we have

$$\sqrt{n}\begin{pmatrix} Z_{n1} - \mathbb{E}_q(\phi(X)w\,(X)) \\ Z_{n2} - \mathbb{E}_q(w\,(X)) \end{pmatrix} \to \mathcal{N}\left(0, \begin{pmatrix} \mathbb{V}_q(\phi(X)w\,(X)) & \mathbb{C}\text{ov}_q\,(\phi(X)w\,(X), w\,(X)) \\ \mathbb{C}\text{ov}_q\,(\phi(X)w\,(X), w\,(X)) & \mathbb{V}_q(w\,(X)) \end{pmatrix}\right) \tag{1}$$

and

$$\nabla g = \begin{pmatrix} \frac{\partial g}{\partial z_1} \\ \frac{\partial g}{\partial z_2} \end{pmatrix} = \begin{pmatrix} 1/z_2 \\ -z_1/z_2^2 \end{pmatrix}$$

so

$$\nabla g\,(\mu) = \begin{pmatrix} 1/\mathbb{E}_q\,(w\,(X)) \\ -\mathbb{E}_q\,(\phi(X)w\,(X))\,/\mathbb{E}_q^2\,(w\,(X)) \end{pmatrix} = \begin{pmatrix} 1 \\ -\mathbb{E}_q\,(\phi(X)w\,(X)) \end{pmatrix}.$$

Hence we have

$$\begin{aligned}\nabla^{\mathrm{T}} g(\mu)\ \Sigma\ \nabla g(\mu) &= \mathbb{V}_q(\phi(X)w(X)) + \mathbb{E}_q^2(\phi(X)w(X))\,\mathbb{V}_q(w(X)) - 2\mathbb{E}_q(\phi(X)w(X))\,\mathbb{C}\mathrm{ov}_q(\phi(X)w(X), w(X)) \\ &= \mathbb{V}_q(\phi(X)w(X)) + I^2\mathbb{V}_q(w(X)) - 2I\mathbb{C}\mathrm{ov}_q(\phi(X)w(X), w(X))\end{aligned}$$

Rearranging the terms, we obtain the desired expression. ■

Remark: we can have either $\sigma^2_{\mathrm{IS}} < \sigma^2_{\mathrm{NIS}}$ or $\sigma^2_{\mathrm{IS}} > \sigma^2_{\mathrm{NIS}}$ as it is demonstrated here on a toy example. Indeed, we have

$$\begin{aligned}\sigma^2_{\mathrm{NIS}} - \sigma^2_{\mathrm{IS}} &= \int (\phi(x) - I)^2 \frac{\pi^2(x)}{q(x)} dx - \int \phi^2(x) \frac{\pi^2(x)}{q(x)} dx \\ &= I\left(I\int \frac{\pi^2(x)}{q(x)} dx - 2\int \phi(x) \frac{\pi^2(x)}{q(x)} dx\right).\end{aligned}$$

For $\pi(x) = \mathcal{N}(x; 0, 1)$, $q(x) = \mathcal{N}\left(x; 0, \sigma^2\right)$ we have

$$\begin{aligned}\frac{\pi^2(x)}{q(x)} &= \frac{1}{\sqrt{2\pi}}\sigma \exp\left(-\left(1 - \frac{1}{2\sigma^2}\right)x^2\right) \\ &= \sigma\sigma' \frac{1}{\sqrt{2\pi}\sigma'} \exp\left(-\frac{x^2}{2(\sigma')^2}\right)\end{aligned}$$

where $(\sigma')^2 = \sigma^2/\left(2\sigma^2 - 1\right)$. Hence, for $\phi(x) = x^2 + m$ and for $\sigma^2 > 1/2$

$$\begin{aligned}I\int \frac{\pi^2(x)}{q(x)} dx - 2\int \phi(x) \frac{\pi^2(x)}{q(x)} dx &= \sigma\sigma'\left\{(1+m) - 2\left((\sigma')^2 + m\right)\right\} \\ &= \sigma\sigma'\left\{\left[1 - 2(\sigma')^2\right] - m\right\}.\end{aligned}$$

For $\sigma^2 \in (1/2, \infty)$, we have $1 - 2(\sigma')^2 < 0$. Hence, by playing with m, the difference $\sigma^2_{\mathrm{NIS}} - \sigma^2_{\mathrm{IS}}$ can be made either positive or negative. As mentioned already, in many situations, only the normalised importance sampling estimator can be implemented anyway.

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

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

Sketch of proof. We are using the same notation as in the proof of the previous theorem. First note that

$$\frac{1}{Z_{n2}} = \frac{1}{1 - (1 - Z_{n2})} = 1 - (Z_{n2} - 1) + (Z_{n2} - 1)^2 + \ldots.$$

Then,

$$\begin{aligned}\widehat{I}_n^{\mathrm{NIS}} = \frac{Z_{n1}}{Z_{n2}} &= Z_{n1} - (Z_{n2} - 1)Z_{n1} + (Z_{n2} - 1)^2 \mathbb{E}_q(\phi(X)w(X)) \\ &\quad + (Z_{n2} - 1)^2 (Z_{n1} - \mathbb{E}_q(\phi(X)w(X))) + \ldots. \\ &= Z_{n1} - (Z_{n2} - 1)Z_{n1} + (Z_{n2} - 1)^2 \mathbb{E}_q(\phi(X)w(X)) + \ldots.\end{aligned}$$

Hence we have

$$n\widehat{I}_n^{\mathrm{NIS}} = nZ_{n1} - \sqrt{n}(Z_{n2} - 1)\sqrt{n}Z_{n1} + \left\{\sqrt{n}(Z_{n2} - 1)\right\}^2 \mathbb{E}_q(\phi(X)w(X)) + \ldots.$$

Now using Eq. (1) and by taking expectations, we obtain the result. ■

Remark. The bias being of order $1/n$, the square of the bias is in $1/n^2$, and we can conclude that the mean square error of $\widehat{I}_n^{\mathrm{NIS}}$ is asymptotically dominated by the variance term.

1.3 Case study: 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,

$$\begin{aligned} \mathbb{P}\left(X_{t+1}=1 \middle| X_t=1\right) &= 1-\mathbb{P}\left(X_{t+1}=2 \middle| X_t=1\right)=\alpha_1, \\ \mathbb{P}\left(X_{t+1}=2 \middle| X_t=2\right) &= 1-\mathbb{P}\left(X_{t+1}=1 \middle| X_t=2\right)=\alpha_2. \end{aligned}$$

We assume that some physical constraints tell us that $\alpha_1+\alpha_2<1$. Assume that we observe $(X_1,...,X_m)=(x_1,...,x_m)$; we shorten the notation by writing $x_{1:m}$ for $(x_1,\ldots,x_m)$. We want to perform Bayesian inference about (α_1,α_2). We set the following prior distribution:

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

The posterior of interest has density:

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

where

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

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

We are interested in estimating $\mathbb{E}\left[\varphi_i\left(\alpha_1,\alpha_2\right) \middle| x_{1:m}\right]$ for the following test functions:

$$\begin{aligned} \varphi_1\left(\alpha_1,\alpha_2\right) &= \alpha_1, \\ \varphi_2\left(\alpha_1,\alpha_2\right) &= \alpha_2, \\ \varphi_3\left(\alpha_1,\alpha_2\right) &= \alpha_1/\left(1-\alpha_1\right), \\ \varphi_4\left(\alpha_1,\alpha_2\right) &= \alpha_2/\left(1-\alpha_2\right), \\ \varphi_5\left(\alpha_1,\alpha_2\right) &= \log\frac{\alpha_1\left(1-\alpha_2\right)}{\alpha_2\left(1-\alpha_1\right)}. \end{aligned}$$

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; for a Markov chain of length 100, started at $X_1=1$, and simulated using $\alpha^\star=(0.2,0.4)$, we obtain an acceptance rate lower than 3%. Furthermore, to implement rejection sampling, we need an upper bound on the posterior density function, which can be found here using a numerical optimizer such as the one provided by the function `optim` in `R`. Therefore we discuss various possible importance proposal distributions.

We first consider the prior as a importance proposal distribution, and we denote it by q_0. The procedure is therefore similar to the rejection sampler described above, except that we do not need an upper bound on the posterior density function, and that each proposed sample contributes to the final estimate.

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

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

but the ratio $p\left(\alpha_1,\alpha_2 \middle| x_{1:m}\right)/q_1\left(\alpha_1,\alpha_2\right)$ is unbounded. Therefore, we do not expect this proposal distribution to yield good results.

Another possible choice of q consists of using

$$q_2\left(\alpha_1,\alpha_2\right)=\mathcal{B}eta\left(\alpha_1;m_{1,1}+1,m_{1,2}+1\right)q_2\left(\alpha_2 \middle| \alpha_1\right),$$

i.e. we match the correct marginal distribution for α_1. For α_2 given α_1, the posterior distribution is $p\left(\alpha_2 \middle| x_{1:m},\alpha_1\right) \propto \left(1-\alpha_2\right)^{m_{2,1}}\alpha_2^{m_{2,2}}\mathbb{I}_{\alpha_2\leq 1-\alpha_1}$, and we approximate it by $q_2\left(\alpha_2 \middle| x_{1:m},\alpha_1\right)=$

$\frac{2}{(1-\alpha_1)^2}\alpha_2 \mathbb{I}_{\alpha_2 \leq 1-\alpha_1}$. This conditional density corresponds to the law of Y defined as $Y = (1-\alpha_1)X$ where X follows a $\mathcal{B}eta\,(2,1)$ distribution. It is straightforward to check that $p\,(\alpha_1, \alpha_2|\, x_{1:m})\,/q_2\,(\alpha_1, \alpha_2) \propto (1-\alpha_2)^{m_{2,1}}\,\alpha_2^{m_{2,2}-1}\,(1-\alpha_1)^2\,/2 < \infty$ whenever $m_{2,2} \geq 1$.

We present the empirical root mean square errors corresponding to rejection sampling and importance sampling using the three choices of proposal distributions in the table. To compute the root mean square errors, we need the true value of each posterior expectations $\mathbb{E}\,[\varphi_i\,(\alpha_1, \alpha_2)|\, x_{1:m}]$. Since it is unavailable, we first approximate these using a rejection sampler based on one million proposed samples, and consider the result to be the ground truth. Then, for each proposed method, we use $10,000$ proposed samples, and 100 independent experiments are performed to compute the root mean square errors:

$$RMSE(\varphi_i) = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(\widehat{\varphi_i}^{(j)} - \mathbb{E}\,[\varphi_i\,(\alpha_1, \alpha_2)|\, x_{1:m}]\right)^2},$$

where $M = 100$ is the number of independent experiments, and each $\widehat{\varphi_i}^{(j)}$ is obtained using $10,000$ proposed samples. As we can see from the table, importance sampling using the prior distribution compares favourably to rejection sampling using the same prior distribution. The proposal distribution q_1 yields catastrophic results, as expected, while the custom distribution q_2 proves more efficient than the prior distribution q_0.

	Method	φ_1	φ_2	φ_3	φ_4	φ_5
1	rejection sampling	0.0036	0.0041	0.0055	0.0129	0.0313
2	IS using q_0	0.0014	0.0016	0.0021	0.0047	0.0125
3	IS using q_1	0.0587	0.0659	0.0834	0.1932	0.2937
4	IS using q_2	0.0011	0.0010	0.0016	0.0029	0.0090

Table 1: Root mean square errors associated to each method and each test function of interest.

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

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

We obtain

$$\begin{aligned}\mathbb{V}\left(\overline{I}_n\right) &= \frac{n}{4n^2}\mathbb{V}\left(\phi(U) + \phi(1-U)\right)\\ &= \frac{1}{2n}\left(\mathbb{V}\left(\phi(U)\right) + \mathbb{C}\mathrm{ov}\left(\phi(U), \phi(1-U)\right)\right).\end{aligned}$$

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

Lemma 2.1. *If the function ϕ is monotonic, then $\mathbb{C}ov(\phi(U), \phi(1-U)) < 0$, unless ϕ is constant on $[0,1]$.*

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

$$\mathbb{C}\mathrm{ov}\left(\phi(U), \phi(1-U)\right) = \frac{1}{2}\mathbb{E}\left[\left(\phi(U_1) - \phi(U_2)\right)\left(\phi(1-U_1) - \phi(1-U_2)\right)\right].$$

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

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

The optimal λ is

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

and the minimal variance is

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

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.