5 Variance reduction techniques

Small variance is vital with Monte Carlo methods, because m-fold reduction of variance means that we may use m-fold less samples to get an estimator with same variance. We saw above that importance sampling can be used to reduce variance of the Monte Carlo estimate. There are other useful techniques which we consider next.

5.1 Rao-Blackwellisation

Recall the *law of total variance.*

Proposition 5.1. *If* $\mathrm{Var}(Z) < \infty$*, then*

$$\mathrm{Var}(Z) = \mathbb{E}[\mathrm{Var}(Z \mid Y)] + \mathrm{Var}(\mathbb{E}[Z \mid Y]),$$

where $\mathrm{Var}(Z \mid Y) = \mathbb{E}[Z^2 \mid Y] - (\mathbb{E}[Z \mid Y])^2 \geq 0$.

Corollary 5.2. *If* $\mathrm{Var}(Z) < \infty$*, then*

$$\mathrm{Var}(\mathbb{E}[Z \mid Y]) = \mathrm{Var}(Z) - \mathbb{E}[\mathrm{Var}(Z \mid Y)] \leq \mathrm{Var}(Z).$$

That is, conditioning can only decrease variance.

Example 5.3 (Rao-Blackwellisation in $\mathbb{R}^2$). Suppose that p is a p.d.f. in $\mathbb{R}^2$, and we would like to compute

$$\mathbb{E}_p[f(X,Y)] = \iint f(x,y)p(x,y)\mathrm{d}x\mathrm{d}y.$$

Simple Monte Carlo would be to simulate $(X_k, Y_k) \overset{\text{i.i.d.}}{\sim} p$ and then compute the average $I_p^{(n)}(f) = n^{-1}\sum_{k=1}^n f(X_k, Y_k)$.

However, if the conditional law $p_{X|Y}(x \mid y)$ is available, and we can calculate the conditional expectation

$$h(y) := \mathbb{E}_p[f(X,y) \mid Y = y],$$

(that is, with $Z = f(X,Y)$, we have $\mathbb{E}[Z \mid Y] = h(Y)$), we may use instead

$$I_{p,\mathrm{RB}}^{(n)}(f) := \frac{1}{n}\sum_{k=1}^n h(Y_k), \tag{12}$$

which approximates the desired quantity $\mathbb{E}_p[f(X,Y)]$ and has smaller variance than $I_p^{(n)}(f)$ (and the improvement can be significant).

Remark 5.4. In Example 5.3, we need only the samples $(Y_k)_{k\geq 1}$ which are distributed according to the marginal disrtribution $p_Y(y) := \int p(x,y)\mathrm{d}x$. We have a choice to simulate either $(X_k, Y_k)_{k\geq 1} \overset{\text{i.i.d.}}{\sim} p$ and throwing away X_k, or simulating directly from the marginal distribution $(Y_k) \overset{\text{i.i.d.}}{\sim} p_Y$, whichever is easier.

Remark 5.5. Rao-Blackwellisation applies similarly also with importance sampling, and with other Monte Carlo methods, such as Markov chain Monte Carlo introduced later.

Remark 5.6 (*). The term *Rao-Blackwellisation* is used, because the method is often associated with sufficient statistics and the Rao-Blackwell theorem. *Marginalisation* or *conditioning* might be more appropriate, but Rao-Blackwellisation is widely used for historical reasons.

Remark 5.7 (*). Sometimes, it may be useful to employ some (biased) approximations $\hat{h}(y) \approx \mathbb{E}[f(X) \mid Y = y]$ in place of the true conditional expectation. Theoretical guarantees for such 'approximate Rao-Blackwellisation' are usually not available, but empirically this type of schemes may be useful.

5.2 Stratification

Example 5.8. Suppose we are interested to estimate $\mathbb{E}_p[f(X)]$ with

$$p(x) = \frac{1}{2}p_1(x) + \frac{1}{2}p_2(x),$$

where p_1 and p_2 are distributions on $\mathbb{X}$.

(a) We know how to sample $X_1, \dots, X_n \sim p$ using $Z_k^{(i)} \overset{\text{i.i.d.}}{\sim} p_i$ and $U_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0,1)$:

$$\begin{aligned} I_p^{(n)}(f) &= \frac{1}{n}\left[\sum_{k=1}^{n} \mathbf{1}\left(U_k \le \frac{1}{2}\right) f(Z_k^{(1)}) + \mathbf{1}\left(U_k > \frac{1}{2}\right) f(Z_k^{(2)})\right] \\ &\overset{d}{=} \frac{1}{n}\sum_{k=1}^{N_1} f(\tilde{Z}_k^{(1)}) + \frac{1}{n}\sum_{k=1}^{N_2} f(\tilde{Z}_k^{(2)}), \end{aligned}$$

where $(\tilde{Z}_k^{(i)}) \overset{\text{i.i.d.}}{\sim} p_i$, $N_1 = \sum_{k=1}^n \mathbf{1}\left(U_k \le \frac{1}{2}\right) \sim \text{Binom}(n, 1/2)$ and $N_2 = n - N_1 \sim \text{Binom}(n, 1/2)$ (note that N_1 and N_2 are not independent though!).

(b) Notice that $\mathbb{E}_p[f(X)] = (1/2)\mathbb{E}_{p_1}[f(X)] + (1/2)\mathbb{E}_{p_2}[f(X)]$, so we may use

$$\begin{aligned} I_p^{(n/2,n/2)}(f) &= \frac{1}{2} I_{p_1}^{(n/2)}(f) + \frac{1}{2} I_{p_2}^{(n/2)}(f) \\ &= \frac{1}{n}\sum_{k=1}^{n/2} f(Z_k^{(1)}) + \frac{1}{n}\sum_{k=1}^{n/2} f(Z_k^{(2)}). \end{aligned}$$

Which estimator should we use? The estimator $I_p^{(n/2,n/2)}(f)$, because it turns out that $\text{Var}\big(I_p^{(n/2,n/2)}(f)\big) \le \text{Var}(I_p^{(n)}(f))$. This is an example of *stratification* (with proportional allocation).

Theorem 5.9. *Suppose the distribution p is of the following mixture form:*

$$p(x) = \sum_{i=1}^{m} w_i p_i(x),$$

where $w_i > 0$ *and* $\sum_i w_i = 1$ *and* $p_1, \dots, p_m$ *are distributions.*

Let $\ell_1, \dots, \ell_m \in \mathbb{N}$ *with* $\sum_i \ell_i = n$, *and define the stratified estimator*

$$I_{\mathrm{p}}^{(\ell_1,\dots,\ell_m)}(f) := \sum_{i=1}^{m} w_i \bigg(\frac{1}{\ell_i}\sum_{j=1}^{\ell_i} f(X_j^{(i)})\bigg),$$

where $(X_j^{(i)})_{i,j}$ are all independent and $(X_j^{(i)}) \overset{i.i.d.}{\sim} p_i$. The estimator satisfies

(i) Unbiasedness: $\mathbb{E}[I_{\mathrm{p}}^{(\ell_1,\dots,\ell_m)}(f)] = \mathbb{E}_p[f(X)]$.

(ii) If $\boxed{\ell_i = w_i n}$ *(proportional allocation), then*

$$\mathrm{Var}\big(I_{\mathrm{p}}^{(\ell_1,\dots,\ell_m)}(f)\big) \le \mathrm{Var}\big(I_p^{(n)}(f)\big).$$

Proof. Unbiasedness (i) is direct, and

$$\begin{aligned}\mathrm{Var}\big(I_{\mathrm{p}}^{(\ell_1,\dots,\ell_m)}(f)\big) &= \sum_{i=1}^m w_i^2 \mathrm{Var}\bigg(\frac{1}{\ell_i}\sum_{j=1}^{\ell_i} f(X_j^{(i)})\bigg) \\ &= \sum_{i=1}^m \frac{w_i^2}{\ell_i}\mathrm{Var}_{p_i}\big(f(X)\big) \\ &= \frac{1}{n}\sum_{i=1}^m w_i \mathrm{Var}_{p_i}\big(f(X)\big),\end{aligned}$$

because $\ell_i = w_i n$. Consider then $X = \sum_{i=1}^m \mathbf{1}\,(s_{i-1} \le U < s_i)\, X^{(i)}$, where $U \sim \mathcal{U}(0,1)$, $s_0 = 0$, $s_i = \sum_{k=1}^i w_i$ and $X^{(i)} \sim p_i$, then $X \sim p$ (exercise!) and we notice that

$$\sum_{i=1}^m w_i \mathrm{Var}_{p_i}\big(f(X)\big) = \mathbb{E}[\mathrm{Var}(f(X) \mid U)] \le \mathrm{Var}_p\big(f(X)\big). \qquad \square$$

Example 5.10 (Stratification with inverse c.d.f.). Suppose F^{-1} is the (generalised) inverse c.d.f. corresponding to a distribution p, and we try to approximate $\mathbb{E}_p[f(X)]$. We may use the following stratified estimator

$$I_{p,\mathrm{strat}}^{(n)}(f) := \frac{1}{n}\sum_{k=1}^n f(X_k), \qquad X_k := F^{-1}(\tilde{U}_k), \qquad \tilde{U}_k := \frac{k-1+U_k}{n},$$

where $(U_k) \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0,1)$.

This is, in fact, proportionally allocated stratification, which follows by writing $\mathbb{E}_p[f(X)] = \mathbb{E}_u[f(F^{-1}(U)]$, where the uniform density can be written as

$$u(t) := \mathbf{1}\,(0 < t \le 1) = \sum_{k=1}^n w_k \tilde{u}_k(t),$$

where $w_k = 1/n$ and $\tilde{u}_k(t) = n\mathbf{1}\left(\frac{k-1}{n} < t \le \frac{k}{n}\right)$ are the densities of $\tilde{U}_k$.

Remark 5.11 (*). Stratification with proportional allocation is guaranteed to provide at least as good estimates as without stratification, but optimal allocation strategy would be $\ell_i \propto w_i\sqrt{\mathrm{Var}_{p_i}(f(X))}$. Because this depends on f and we may be interested in several f, and because $\mathrm{Var}_{p_i}(f(X))$ is usually not known, proportional allocation is often a safe choice.

5.3 Introducing dependence: antithetic variables

In some cases, it is possible to use the dependence of random variables to help decrease the variance. First such technique is so-called 'antithetic' variables.

Definition 5.12 (Antithetic variables)**.** Suppose $\hat{p}(x, y)$ is a joint distribution with p as its marginals[7]. In the discrete case, this means that for all $x, y \in \mathbb{X}$,

$$p(x) = \sum_{z \in \mathbb{X}} \hat{p}(x, z) \quad \text{and} \quad p(y) = \sum_{z \in \mathbb{X}} \hat{p}(z, y).$$

Suppose that $(X_n, Y_n)_{n \geq 1} \overset{\text{i.i.d.}}{\sim} \hat{p}$, then clearly $(X_n)_{n \geq 1} \overset{\text{i.i.d.}}{\sim} p$ and $(Y_n)_{n \geq 1} \overset{\text{i.i.d.}}{\sim} p$, but each X_k typically depends on the corresponding 'pair' Y_k. The antithetic variable estimator

$$I^{(n)}_{\hat{p},\text{anti}}(f) := \frac{1}{2n} \sum_{k=1}^{n} [f(X_k) + f(Y_k)]$$

is clearly unbiased and strongly consistent estimator of $\mathbb{E}_p[f(X)]$.

Proposition 5.13. *The variance of the antithetic variable estimator is*

$$\operatorname{Var}\big(I^{(n)}_{\hat{p},anti}(f)\big) = \frac{1}{2n}\big[\operatorname{Var}_p f(X) + \operatorname{Cov}_{\hat{p}}\big(f(X), f(Y)\big)\big].$$

Therefore, if $\operatorname{Cov}_{\hat{p}}\big(f(X), f(Y)\big) = \operatorname{Cov}\big(f(X_1), f(Y_1)\big) \leq 0$ *then* $\operatorname{Var}\big(I^{(n)}_{\hat{p},anti}(f)\big) \leq \operatorname{Var}\big(I^{(2n)}_p(f)\big)$*.*

Note that $I^{(2n)}_p(f)$ has the same total number of samples as $I^{(n)}_{\hat{p},\text{anti}}(f)$, so they have roughly equal computational complexity.

Useful antithetic variables can be found with the inverse c.d.f. method.

Proposition 5.14. *Suppose F^{-1} is a generalised inverse c.d.f. of p, and $f : \mathbb{R} \to \mathbb{R}$ is monotonic. Define $X_k = F^{-1}(U_k)$ and $Y_k = F^{-1}(1 - U_k)$ where $(U_k)_{k \geq 0} \overset{i.i.d.}{\sim} \mathcal{U}(0, 1)$. Then,* $\operatorname{Cov}(f(X_1), f(Y_1)) \leq 0$*.*

Proof. (*) Without loss of generality, we may assume f increasing. Then also $\bar{f}(x) = f(x) - \mathbb{E}_p[f(X)]$ and $g := \bar{f} \circ F^{-1}$ are increasing. If $\operatorname{Var}(g(U)) = 0$, the claim is trivial, so assume $\operatorname{Var}(g(U)) > 0$.

Because of symmetry

$$\operatorname{Cov}(f(X_1), f(Y_1)) = \int_0^1 g(u) g(1-u) \mathrm{d}u = 2 \int_0^{1/2} g(u) g(1-u) \mathrm{d}u.$$

Recall $\mathbb{E}[g(U)] = 0$, so there exists $u_0 \in (0, 1)$ such that $g(u) \leq 0$ for $u < u_0$ and

[7] Such $\hat{p}$ is also known as a *coupling* of p with itself.