# 1 Variance Reduction

We have seen that Monte Carlo integration typically has an error variance $\sigma^2/n$. We can reduce the error by sampling with a larger value of $n$, but the computing time grows with $n$. Sometimes we can **find a way to reduce** $\sigma$ instead. To do this, we construct a new Monte Carlo problem with the **same answer as our original one but with a lower** $\sigma$. These methods are known as variance reduction techniques. In this course, we will discuss four such methods.

Methods of variance reduction can sometimes bring enormous improvements compared to simple Monte Carlo. It is not uncommon for the value $\sigma^2$ to be **reduced many thousand fold**. It is also possible for a variance reduction technique to **bring a very modest improvement**, perhaps equivalent to reducing $\sigma^2$ by only 10%. What is worse, some methods **will raise** $\sigma^2$ in unfavorable circumstances. Hence, we should be careful in implementing variance reduction technique.

## 1.1 Measuring Efficiency

Variance reductions are used to **improve the efficiency** of Monte Carlo methods. Before looking at individual methods, we discuss how to measure efficiency.

Suppose for simplicity, that a **baseline method is unbiased** and estimates the desired quantity with variance $\sigma_0^2/n$, at a cost of $nc_0$, when $n$ function evaluations are used. To get an error variance of $\tau^2$, we need $n = \sigma_0^2/\tau^2$ and this will cost $c_0\sigma_0^2/\tau^2$. Here we are assuming that cost is measured in time and that overhead cost is small. If an **alternative unbiased method** has variance $\sigma_1^2/n$ and cost $nc_1$ under these conditions then it will cost us $c_1\sigma_1^2/\tau^2$ to achieve the same error variance $\tau^2$ that the baseline method achieved. The **efficiency of the new method**, **relative** to the standard method is defined by

$$E = \frac{c_0\sigma_0^2}{c_1\sigma_1^2}.$$

At any fixed level of accuracy, the old method takes $E$ **times as much work** as the new one.

There is no fixed rule for how large an efficiency improvement must be to make it worth using. In some settings, such as rendering computer graphics for animated motion pictures, where thousands of CPUs are kept busy for months, a 10% **improvement** ($E = 1.1$) **brings meaningful** savings. In other settings, such as a one-off computation, a 60-**fold gain** ($E = 60$) which turns a one minute wait into a one second wait, **may not justify the cost** of programming a more complicated method.

## 1.2 Antithetics

When we are using **Monte Carlo averages** of quantities $f(\boldsymbol{X})$ then the **randomness** in the algorithm leads to some **error cancellation**. In antithetic sampling we try to get even more cancellation. An antithetic sample is one that somehow gives the **opposite value** of $f(\boldsymbol{x})$, being low when $f(\boldsymbol{x})$ is high and vice versa. Ordinarily we get an opposite $f$ by sampling at a point that is somehow opposite to $\boldsymbol{x}$.

A set $\mathcal{D}$ is called **symmetric** about the point $\boldsymbol{c}$ if $\tilde{\boldsymbol{x}} = 2\boldsymbol{c} - \boldsymbol{x} \in \mathcal{D}$ for all $\boldsymbol{x} \in \mathcal{D}$. A density function $p$ on $\mathcal{D}$ is called symmetric if $p(\boldsymbol{x}) = p(\tilde{\boldsymbol{x}})$. Let $\mu = E(f(\boldsymbol{X}))$ for $\boldsymbol{X} \sim p$, where $p$ is a

symmetric density on the symmetric set $\mathcal{D}$. In this case, $\tilde{\boldsymbol{X}} \sim p$. For example, if $p$ is $N_q(0, \Sigma)$, then $\tilde{\boldsymbol{x}} = -\boldsymbol{x}$, with $\boldsymbol{c} = 0$. For $p$ is Uniform on $(0, 1)^d$, $\tilde{\boldsymbol{x}} = 1 - \boldsymbol{x}$ componentwise, with $\boldsymbol{c} = 0.5$.

The **antithetic sampling estimate** of $\mu$ is

$$\widehat{\mu}_{\text{anti}} = \frac{1}{n} \sum_{i=1}^{n/2} \left( f(\boldsymbol{X}_i) + f(\tilde{\boldsymbol{X}}_i) \right),$$

where $\boldsymbol{X}_i \overset{i.i.d.}{\sim} p$ and $n$ is an even number.

The **rationale** for antithetic sampling is that each value of $\boldsymbol{x}$ is **balanced by** its opposite $\tilde{\boldsymbol{x}}$ satisfying $(\boldsymbol{x} + \tilde{\boldsymbol{x}})/2 = \boldsymbol{c}$. Whether this balance is helpful depends on $f$ . Clearly if $f$ is **nearly linear** we could obtain a large improvement.

Clearly, $\widehat{\mu}_{\text{anti}}$ is **unbiased** for $\mu$ (*i.e.*, $E(\widehat{\mu}_{\text{anti}}) = \mu$). Suppose that $\sigma^2 = E((f(\boldsymbol{X}) - \mu)^2) < \infty$. Then the variance in antithetic estimate is

$$Var(\widehat{\mu}_{\text{anti}}) = \frac{n/2}{n^2} Var\left( f(\boldsymbol{X}) + f(\tilde{\boldsymbol{X}}) \right) = \frac{\sigma^2}{n}(1 + \rho),$$

where $\rho = Corr\left( f(\boldsymbol{X}), f(\tilde{\boldsymbol{X}}) \right)$. As $|\rho| \leq 1$, we have $0 \leq \frac{\sigma^2}{n}(1 + \rho) \leq \frac{2\sigma^2}{n}$. In the **best case**, antithetic sampling gives the **exact answer** from just **one pair** of function evaluations. In the worst case it doubles the variance. Both cases do arise. It is clear that a **negative correlation is favorable**.

Note that $f(\boldsymbol{x})$ can be written as

$$f(\boldsymbol{x}) = \frac{f(\boldsymbol{x}) + f(\tilde{\boldsymbol{x}})}{2} + \frac{f(\boldsymbol{x}) - f(\tilde{\boldsymbol{x}})}{2} \equiv f_E(\boldsymbol{x}) + f_O(\boldsymbol{x}).$$

The **even part** satisfies $f_E(\boldsymbol{x}) = f_E(\tilde{\boldsymbol{x}})$ and $\int f_E(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} = \mu$. The **odd part** satisfies $f_O(\boldsymbol{x}) = -f_O(\tilde{\boldsymbol{x}})$ and $\int f_O(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} = 0$. Also, the even and odd parts are **orthogonal** as

$$\begin{aligned}
\int f_E(\boldsymbol{x}) f_O(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} &= \int \left( \frac{f(\boldsymbol{x}) + f(\tilde{\boldsymbol{x}})}{2} \right) \left( \frac{f(\boldsymbol{x}) - f(\tilde{\boldsymbol{x}})}{2} \right) p(\boldsymbol{x}) d\boldsymbol{x} \\
&= \frac{1}{4} \int \left( f^2(\boldsymbol{x}) - f^2(\tilde{\boldsymbol{x}}) \right) p(\boldsymbol{x}) d\boldsymbol{x} \\
&= \frac{1}{4} \left( E(f^2(\boldsymbol{X})) - E(f^2(\tilde{\boldsymbol{X}})) \right) \\
&= 0, \text{ as } \boldsymbol{X} \text{ and } \tilde{\boldsymbol{X}} \text{ has same distribution.}
\end{aligned}$$

Hence, it follows that $\sigma^2 = \sigma_E^2 + \sigma_O^2$, where $\sigma_E^2 = \int \left( f_E(\boldsymbol{x}) - \mu \right)^2 p(\boldsymbol{x}) d\boldsymbol{x}$ and $\sigma_O^2 = \int f_O^2(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$. Note that **antithetic estimate** of $\mu$ can be written as $\widehat{\mu}_{\text{anti}} = \frac{2}{n} \sum_{i=1}^{n/2} f_E(\boldsymbol{X}_i)$ and therefore the **variance** of $\widehat{\mu}_{\text{anti}}$ is $2\sigma_E^2/n$. We see that antithetic sampling **eliminates the variance contribution** of $f_O$, but **doubles** the contribution from $f_E$. Antithetic sampling is **extremely beneficial** for integrands that are primarily **odd functions of their inputs**, having $\sigma_O^2 \gg \sigma_E^2$. The connection to correlation is via $\rho = (\sigma_E^2 - \sigma_O^2)/(\sigma_E^2 + \sigma_O^2)$. (Why?)

Variance **reduction is only a part** of the story because the cost of antithetic sampling using $n$ points could well be smaller than the cost of simple Monte Carlo with $n$ points. That will happen if it is **expensive to generate** $\boldsymbol{X}$, compared to the **cost of computing** $f$, but **inexpensive** to generate $\tilde{\boldsymbol{X}}$. For example, $\boldsymbol{X}$ might be a carefully constructed and expensive sample path from a Gaussian process while $\tilde{\boldsymbol{X}} = -\boldsymbol{X}$.

Because antithetic samples have **dependent values within pairs**, the usual variance estimate must be **modified**. Let $Y_i = f_E(\boldsymbol{X}_i)$, $i = 1, 2, \ldots, m$, where $m = n/2$. Then take

$$\widehat{\mu}_{\text{anti}} = \frac{1}{m}\sum_{i=1}^{m} Y_i, \qquad\qquad s^2_{\text{anti}} = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \widehat{\mu}_{\text{anti}})^2.$$

Then use $s^2_{\text{anti}}/m$ as the estimator of $Var(\widehat{\mu}_{\text{anti}})$. A CLT based 99% **confidence interval** for $\mu$ is

$$\widehat{\mu}_{\text{start}} \mp 2.58\sqrt{\widehat{Var}(\widehat{\mu}_{\text{anti}})} = \widehat{\mu}_{\text{start}} \mp 2.58\,\frac{s_{\text{anti}}}{\sqrt{m}}.$$