1 Introduction to Monte Carlo (Lectures 1 & 2)

The aims of this course are to introduce foundational theoretical concepts and methodological tools essential for the successful development, analysis, and application of advanced Machine Learning and Computational Statistical methods to address a wide range of challenging contemporary Engineering problems. Apart from other topics the bulk of the course will focus on Markov Chain Monte Carlo algorithms. To analyze them, we will cover the basics of some foundational mathematical concepts of Measure theory and Functional Analysis.

Markov Chain Monte Carlo agorithms are sampling methods, used to estimate expectations of the form $\mathbb{E}_{x\sim\pi}[f(x)]$, crucial for statistical inference. Before starting to study them it is important to understand in what contexts these algorithms appear in practice. This section will provide a framework, namely probabilistic Bayesian models, which is used in many applications.

1.1 Probabilistic Bayesian models

As a running example, consider the following fictional problem. Suppose that a mining company is interested in a particular area containing some mineral A that it wants to extract. In addition to mineral A, the mineral B is present in the area, making the extraction of mineral A more costly. The mining company wants to determine the amounts of mineral A and B in the area from a satellite image to forecast the earnings from mining there. Suppose that $\theta \in \Theta$ is some variable denoting the unknown material composition of the area of interest, i.e. the presence of A and B in each pixel. The company has calculated the projected earnings from the area for all possible θ, given by the function $e : \Theta \to \mathbb{R}$. Now they have acquired a satellite image $y \in Y$ and want to use it to predict θ and inform their decisions.

To be more specific, if y is an RGB image, we might write $Y = [0,1]^{w\times h\times 3}$ and $\Theta = [0,1]^{w\times h}$. Here, $w \times h$ is the image resolution, $y_{i,j}$ denotes the color intensity values of the pixel in (i,j)-th position, and $\theta_{i,j}$, $1-\theta_{i,j}$ denote the proportions of minerals A and B in the (i,j)-th pixel.

One way to forecast θ and the projected earnings is to use a probabilistic Bayesian model. Roughly speaking, a probabilistic Bayesian model is a joint probability distribution $p(y,\theta)$ on observed data $y \in Y$ and parameters that are to be estimated $\theta \in \Theta$. A Bayesian model is usually given in terms of the *likelihood* distribution $p(y \mid \theta)$, describing how the image y is formed if the area has a material composition described by θ, and the *prior* distribution $p(\theta)$, describing our prior beliefs of what θ is before observing any data.

In our example, if the mineral A is green and the mineral B is white, then the likelihood function could be given by the following forward model:

$$y = \theta \begin{bmatrix}0\\1\\0\end{bmatrix} + (1-\theta)\begin{bmatrix}1\\1\\1\end{bmatrix} + \eta$$

where η is some Gaussian noise[1]. I.e. we expect y to be roughly a linear combination of the white and green colors, where the coefficients are given by θ. As mentioned, prior to observing the picture the company did not have any information of what θ is and it is reasonable to assume that the prior on θ is a uniform distribution across Θ, i.e. $\theta_{i,j} \sim \mathcal{U}[0,1]$.

Once the data y has been observed, our belief of θ is the *posterior* distribution $p(\theta \mid y)$, given by the Bayes' theorem:

[1] It might be confusing that we have restricted our data to $[0,1]$ intervals, but the forward model assumes it can fall out of that range, as the gaussian noise η can attain any values. This is not a contradiction, the forward model is just a model and not the true mechanism of how images are formed. Maybe a better modelling decision could be to consider other noise distributions that wouldn't make the data fall out of that range.

$$\overbrace{p(\theta \mid y)}^{\text{posterior beliefs}} = \frac{\overbrace{p(\theta)}^{\text{prior beliefs}} \times \overbrace{p(y \mid \theta)}^{\text{data likelihood}}}{p(y)} \tag{1}$$

Note that the above equation reveals how fundamentally beliefs evolve over time in a Bayesian model — before seeing any data, our belief of θ is represented by the prior distribution, and upon observing evidence our beliefs change according to the Bayes' rule.

If such a Bayesian model is constructed, then one could estimate the expected earnings in the area by the mean of $e(\theta)$ under the posterior distribution

$$\mathbb{E}_{\theta \sim p(\cdot \mid y)}[e(\theta)] = \int_{\Theta} e(\theta) p(\theta \mid y) \mathrm{d}\theta \tag{2}$$

and quantify the risk by the variance:

$$\mathrm{Var}_{\theta \sim p(\cdot \mid y)}[e(\theta)] = \mathbb{E}_{\theta \sim p(\cdot \mid y)}[e(\theta)^2] - \mathbb{E}_{\theta \sim p(\cdot \mid y)}[e(\theta)]^2 \tag{3}$$

For a large part of this course, we will therefore be concerned with computing general integrals:

$$\mathbb{E}_{x \sim \pi}[f(x)] = \int_X f(x) \pi(x) \mathrm{d}x \tag{4}$$

Such expectations arise not only in the context mentioned above. Another example can be found in Deep Learning applications, where of big importance is computing the gradients $\nabla g(x)$ of some loss defined by function g. Often, for example in the context of Variational Autoencoders, gradients take the form of

$$\nabla g(x) = \mathbb{E}[h(x)]$$

and sampling algorithms need to be used to compute them.

1.2 Monte Carlo

One way to estimate the integral in (4) is to compute a simple Riemann discretization:

$$\mathbb{E}_{x \sim \pi}[f(x)] \approx \sum_{i=1}^{N} \Delta_i f(x_i) \pi(x_i) \tag{5}$$

However, there are a few issues: a) for fixed N, Riemann discretization will produce biased estimates b) the discretization error is difficult to characterize; and most importantly c) choosing the grid $x_1 \leq \cdots \leq x_N$ is challenging and essentially impossible, if $X = \mathbb{R}$, or if X is infinite-dimensional, which happens often as we will later see in the course.

The Monte Carlo approach is a whole suite of techniques used to compute integrals in (4) that elegantly overcomes all of the issues above. To illustrate this idea simply, first consider estimating the integral $I = \int_a^b f(x)\mathrm{d}x$ for some function f given in Figure 1. We know that the value of the integral is the area below

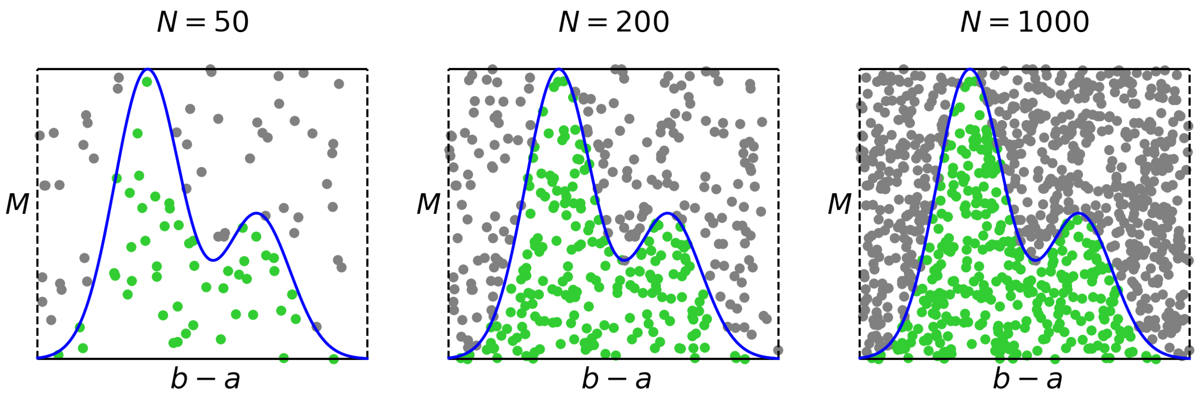


Figure 1

the curve and that the area of the bounding box of $f(x)$ is $A = (b-a) \times M$, where $M \geq \max f(x)$. So, if we randomly sample a point in the bounding box of $f(x)$ the probability that it lies beneath the curve is I/A and the probability that it lies above the curve is $1 - I/A$. And if we independently sample N points, then around NI/A of them should be below the curve. If N is large enough and K is the number of points lying beneath the curve, then we should see

$$\frac{NI}{A} \approx K \implies I \approx \frac{AK}{N} = \frac{(b-a)MK}{N}$$

So simply sampling N points and looking at how many of them are below the curve gives us an estimate $\hat{I} = \frac{AK}{N}$ of I.

The above method illustrates the core feature of Monte Carlo — taking random samples and averaging them to get the true value. However, it has one issue – we need to choose an upper bound M, which is not necessarily available to us. This is not difficult to fix. Indeed, notice that if we uniformly sample points only on the x-axis, then the average value of $f(x)$ will be $I/(b-a) = \frac{1}{b-a}\int_a^b f(x)\mathrm{d}x$. So, if $X_1, \dots, X_N$ are random samples from the uniform distribution $\mathcal{U}[a, b]$, we know that

$$I \approx \hat{I} = \frac{b-a}{N} \sum_{i=1}^{N} f(X_i)$$

For a more general integral $I = \mathbb{E}_{x\sim\pi}[f(x)] = \int_X f(x)\pi(x)\mathrm{d}x$ the Monte Carlo strategy translates to sampling N points $x_1, \dots, x_N \sim \pi$ and computing the average:

Definition 1. *A vanilla Monte Carlo estimate with N samples for the integral $I = \mathbb{E}_{x\sim\pi}[f(x)]$ is defined as*

$$\hat{I} = \frac{1}{N} \sum_{i=1}^{N} f(X_i) \tag{6}$$

where $X_1, \dots, X_N \sim \pi$.

In what ways is this estimate better than the Riemann discretization? Firstly, we do not need to come up with a discretization of the domain X. This is especially useful in cases when we are integrating over very high-dimensional (or even infinite-dimensional) variables x.

Secondly, it is very important to understand that unlike a Riemann discretization, the Monte Carlo estimate $\hat{I}$ is a random variable. Being one, it has properties such as expectation, variance, cumulative distribution function.

Definition 2. *Let $\hat{I}$ be an estimator of the integral $I = \mathbb{E}_{x\sim\pi}[f(x)] = \int_X f(x)\pi(x)\mathrm{d}x$. We say that $\hat{I}$ is unbiased if*

$$\mathbb{E}[\hat{I}] = I$$

It is trivial to verify that vanilla MC estimates are unbiased:

$$\mathbb{E}[\hat{I}] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{X_i\sim\pi}[f(X_i)] = \frac{1}{N}\sum_{i=1}^{N} I = I \tag{7}$$

So we know that on average we should get the correct answer. However, note that if the variance of $\hat{I}$ is large, the squared distance to I that we should expect is also big:

$$\mathrm{Var}(\hat{I}) = \mathbb{E}[(\hat{I} - I)^2] \tag{8}$$

We can write the variance of an MC esimate as

$$\mathrm{Var}(\hat{I}) = \frac{1}{N^2}\sum_{i=1}^{N} \mathrm{Var}(f(X_i)) = \sigma_f^2/N \tag{9}$$

where

$$\sigma_f^2 = \mathrm{Var}_{X\sim\pi}(f(X)) = \mathbb{E}_{X\sim\pi}[f(X)^2] - \mathbb{E}_{X\sim\pi}[f(X)]^2 = \mathbb{E}_{X\sim\pi}[f(X)^2] - I^2$$

is the single sample variance. So we see that the variance of $\hat{I}$ decreases linearly with the number of samples taken. However, note that it also depends on σ_f^2, so if it is exhorbitant, even large values of N will not help. We will see how this can be solved using variance reduction techniques below.

Finally, note that for vanilla MC estimators $\hat{I}$ we made the assumption that one can sample from the distribution π, which is almost never the case. In addition, because of the Bayes' rule (1), most often we only know the posterior distribution $p(\theta \mid y)$ up to a multiplicative constant — $p(y)$ is intractable.

1.3 Limit Theorems for Monte Carlo

We saw in equations (7) and (9) the expectation and variance of the MC estimator with fixed number of samples N. Limit theorems are a class of theorems analyzing the behaviour of $\hat{I}$ when N is taken to ∞.

Theorem 1 (Central Limit Theorem)**.** *Let $Y_1, Y_2, \ldots$ be a sequence of i.i.d. variables with mean μ and variance $\sigma^2 < \infty$. Let $S_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then*

$$\frac{S_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \text{ as } n \to \infty \tag{10}$$

Proof. Consider the moment generating functions M_{Z_n} and M_f of the random variables $Z_n = \frac{S_n - \mu}{\sigma/\sqrt{n}}$ and $f(Y_n)$:

$$\begin{aligned} M_f(t) &= \mathbb{E}[e^{-tf(Y)}] \\ M_{Z_n} &= \mathbb{E}[e^{-tZ_n}] \end{aligned}$$

Using the sum rule of mgfs and expanding using Taylor series, we get:

$$\begin{aligned} M_{Z_n}(t) &= M_f\left(\frac{t}{\sqrt{n\sigma^2}}\right)^n \\ &= \left(1 + \frac{t^2}{2n} + o(\frac{t^2}{n})\right)^n \\ &\to \exp\left(\frac{t^2}{2}\right) \end{aligned}$$

The RHS is the moment generating function of $\mathcal{N}(0,1)$ so we are done. □

Theorem 2 (Strong Law of Large Numbers)**.** *Let $Y_1, Y_2, \ldots$ be a sequence of i.i.d. variables with mean μ. Let $S_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then $S_n \to \mu$ with probability 1.*

Applying CLT and SLLN to the MC estimate we get that for large N

$$\hat{I} \approx \mathcal{N}(I, \sigma_f^2/N) \tag{11}$$

and $\hat{I}$ is guaranteed to converge to I with probability 1 as $N \to \infty$.

1.4 Variance reduction

As mentioned above, MC estimates can exhibit very high variance. For example, suppose that we want to compute the integral $I = \mathbb{E}_{x\sim\pi}[f(x)]$, where $\pi(x) = \mathbb{1}_{[0,1]}(x)$, i.e. the uniform distribution $\mathcal{U}[0,1]$, and $f(x) = x^2 \times \mathcal{N}(x; 0.5, 0.01)$, where $\mathcal{N}$ is the Gaussian pdf. The function $f(x)$ is showed in Figure 2 and Figure 3a shows the running vanilla MC estimate. Notice that, because $f(x)$ is highly concentrated around 0.5, most sampled points are outside the peak and have values incredibly close to 0. These points very slowly drag the average to 0, while the points in the peak contribute to sharp jumps in the estimate, i.e. they are informative. This behaviour results in a high variance MC estimate for I and is a problem, since we need to take a fairly large number of samples to get a reliable estimate. Variance reduction techniques aim to fix this issue. In this section we will cover two such techniques — Importance Sampling and Control Variates.

In the example above, points around the peak contribute most to the integral I and also have the highest variance in their values of the function $f(x)$. In contrast, the points outside the peak contribute very little to the overall value of the integral I and have almost a constant value of $f(x)$. We should spend more computational effort to estimate the part of integral in the peak rather than outside it. So, sampling points uniformly in the vanilla MC estimate is not a sensible choice. In the example above, out of the 1000 samples taken, only around a 100 of them fell within the peak. To fix this, we could do the following hack. Split the integral into 3 parts, one integral over the peak and two outside the peak. Then estimate the integral over

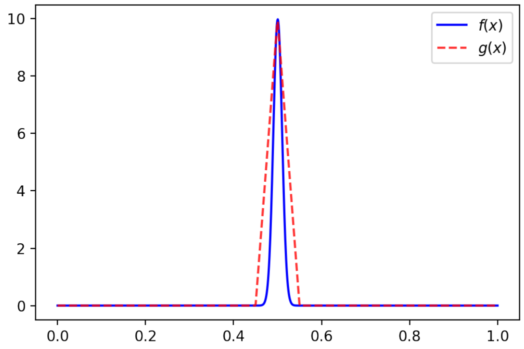
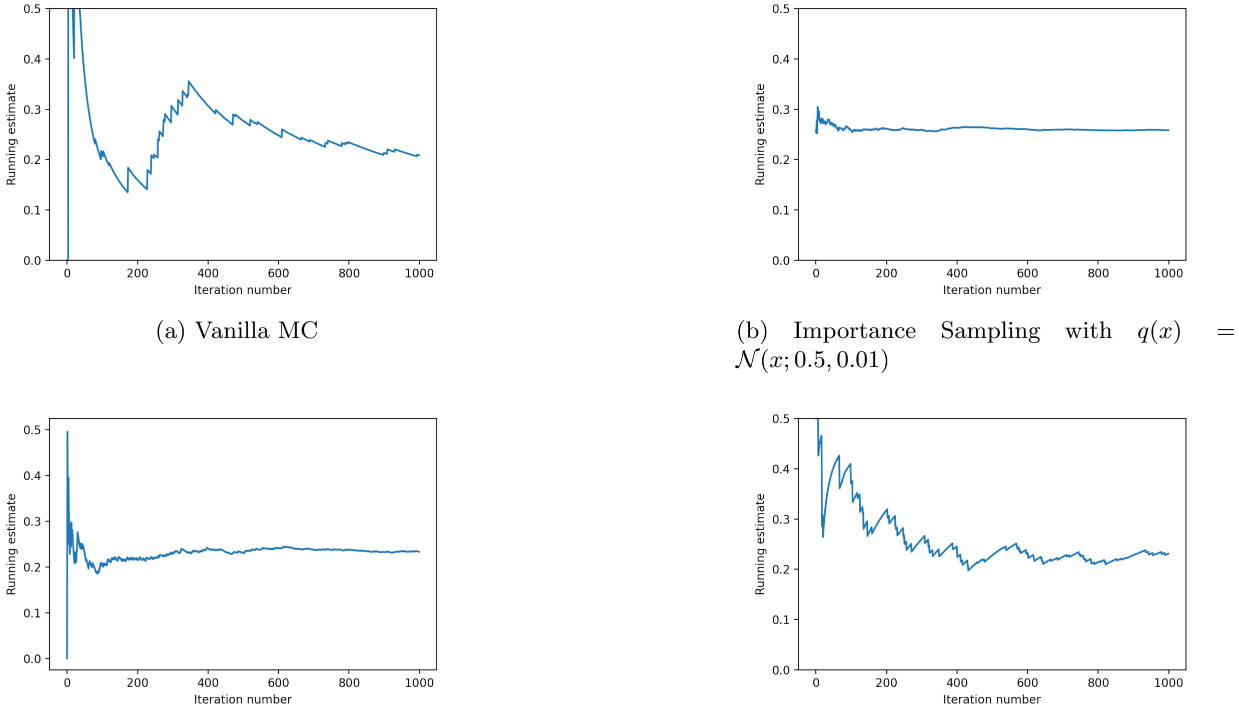


Figure 2



(a) Vanilla MC

(b) Importance Sampling with $q(x) = \mathcal{N}(x; 0.5, 0.01)$

(c) Sampling with splitting the $[0, 1]$ interval

(d) Control Variates with a triangular function

Figure 3

the peak with 900 samples and the ones outside the peak with 50 samples each, and sum them up. More precisely, we rewrite the expectation as

$$I = \mathbb{E}_{x\sim\mathcal{U}[0,1]}[f(x)] = 0.45 \times \mathbb{E}_{x\sim\mathcal{U}[0,0.45]}[f(x)] + 0.1 \times \mathbb{E}_{x\sim\mathcal{U}[0.45,0.55]}[f(x)] + 0.45 \times \mathbb{E}_{x\sim\mathcal{U}[0.55,1]}[f(x)] \quad (12)$$

and do vanilla MC on all three expectations. Figure 3c shows the running MC estimates as the number of samples are increased[2]. They are much more stable and it is clear that around 500 samples is sufficient for a decent estimate. Note that above we effectively just changed the sampling distribution to take more samples from the interval $[0.45, 0.55]$ and adjusted the function values $f(x)$ by the weights in (12). Importance Sampling (IS) is a generalization of this approach, where we change the sampling distribution.

To estimate the integral $I = \mathbb{E}_{x\sim\pi}[f(x)]$ a practitioner chooses a distribution $q(x)$, called the proposal, from which they want to sample. Then, an IS estimate is derived via a single line of maths:

$$\mathbb{E}_{x\sim\pi}[f(x)] = \int_X f(x)\pi(x)\mathrm{d}x = \int_X f(x)\frac{\pi(x)}{q(x)}q(x)\mathrm{d}x = \mathbb{E}_{x\sim q}\left[f(x)\frac{\pi(x)}{q(x)}\right] \quad (13)$$

Definition 3. *Given a proposal distribution $q(x)$ such that $q(x) > 0$ whenever $\pi(x) > 0$, the Importance Sampling estimate of $I = \mathbb{E}_{x\sim\pi}[f(x)]$ is*

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N} f(X_i)\frac{\pi(X_i)}{q(X_i)} \quad (14)$$

where $X_1, \ldots, X_N \sim q$.

Note the constraint on $q(x)$ in the definition. Of course, Importance Sampling works only if we samples from all regions of positive probability.

As we are hoping to engineer lower variance estimates via Importance Sampling, it would be good to understand how it behaves with different proposal distributions $q(x)$. Just like in the case of vanilla MC we have that the variance decreases linearly with the number of samples taken, but the constant σ_f^2 will now be different, and we will write it as $\sigma_{\bar{f}}^2$. As before it can be written as

$$\sigma_{\bar{f}}^2 = \mathrm{Var}_{X\sim q}\left(f(x)\frac{\pi(x)}{q(x)}\right) = \mathbb{E}_{X\sim q}\left[\left(f(X)\frac{\pi(X)}{q(X)}\right)^2\right] - I^2 \quad (15)$$

$\sigma_{\bar{f}}^2$ will be large, if $f(x)\pi(x)$ is high in regions where $q(x)$ is low. That is, for $\sigma_{\bar{f}}^2$ to be small $q(x)$ has to match $f(x)\pi(x)$. In fact, the optimal proposal distribution is given by $|f(x)|\pi(x)$ normalized to a probability density function:

$$q_{\mathrm{opt}}(x) = \frac{|f(x)|\pi(x)}{\int_X |f(x)|\pi(x)\mathrm{d}x} \quad (16)$$

To derive this in general, functional analysis tools are needed, but when $f(x)\pi(x)$ is positive, it's easy to verify that $\sigma_{\bar{f}}^2 = 0$[3]:

[2] Here I interleaved the sequences of 50, 900 and 50 samples, to make a fair comparison

[3] If $f(x)\pi(x)$ is negative at some points, the optimal distribution is the same, but the variance might be non-zero

$$\mathbb{E}_{X\sim q_{\text{opt}}}\left[\left(f(X)\frac{\pi(X)}{q(X)}\right)^2\right] = \int_X \frac{f(x)^2\pi(x)^2}{q(x)}\mathrm{d}x \tag{17}$$

$$= \left(\int_X f(x)\pi(x)\mathrm{d}x\right)^2 = I^2 \tag{18}$$

Although we introduced Importance Sampling as a variance reduction technique, it is often used in other contexts as well. Specifically, when it is not possible to produce samples from $\pi(x)$, but we can still evaluate it, expectations can be estimated via IS but not vanilla MC.

Another tool for variance reduction is Control Variates. In the example above, from Figure 2, we know that function $f(x)$ can be roughly approximated by a triangular function $g(x)$. As opposed to $\mathbb{E}_{x\sim\pi}[f(x)]$, the value of $\mathbb{E}_{x\sim\pi}[g(x)]$ is analytically available to us, which means that if we estimate $\mathbb{E}_{x\sim\pi}[f(x)-g(x)]$ we can recover $\mathbb{E}_{x\sim\pi}[f(x)] = \mathbb{E}_{x\sim\pi}[f(x)-g(x)]+\mathbb{E}_{x\sim\pi}[g(x)]$. Our hope is that, because $g(x)$ is a good approximation of $f(x)$, $f(x)-g(x)$ will not have such sharp peaks as $f(x)$ does and will be easier to estimate. Figure 3d shows a running estimate of I using this approach.

In general, the space X can be very high-dimensional when noticing such structures is impossible. In that case, some candidate control variates $g(x)$ might still be available to us, but we would not know their exact relationship with $f(x)$. However, it can be determined using some analysis. We formalize Control Variates as follows. In addition to the function $f(x)$ we have a function $g(x)$, for which the value of the integral $\mathbb{E}_{x\sim\pi}[g(x)]$ is analytically available to us. We also assume that $g(x)$ is informative of $f(x)$. Then the Control Variate estimate is defined as

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N}\Big(f(X_i) + \beta \times \big(g(X_i) - \mathbb{E}_{x\sim\pi}[g(x)]\big)\Big) \tag{19}$$

To use this estimate we need to find the coefficient β for which the variance will be reduced. The single sample variance of the Control Variate estimate is

$$\sigma_f^2 = \mathrm{Var}_{X\sim\pi}\big(f(X) + \beta \times (g(X) - \mathbb{E}_{x\sim\pi}[g(x)])\big) \tag{20}$$

$$= \mathrm{Var}_{X\sim\pi}\big(f(X)\big) + \beta^2\mathrm{Var}_{X\sim\pi}\big(g(X)\big) + 2\beta\mathrm{Cov}_{X\sim\pi}\big(f(X), g(X)\big) \tag{21}$$

which is minimized when

$$\beta = -\frac{\mathrm{Cov}_{X\sim\pi}\big(f(X), g(X)\big)}{\mathrm{Var}_{X\sim\pi}\big(g(X)\big)} \tag{22}$$

Several Control Variates can also be used at the same time. That is, if we have functions $g_1(x), \dots, g_k(x)$, we can consider the following estimate

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N}\Big(f(X_i) + \boldsymbol{\beta}^T\big(\boldsymbol{g}(X_i) - \mathbb{E}_{x\sim\pi}[\boldsymbol{g}(x)]\big)\Big) \tag{23}$$

where we write $\boldsymbol{g}(X) = (g_1(X), \dots, g_k(X))^T$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)^T$. Then the single sample variance is given by

$$\sigma_f^2 = \mathrm{Var}_{X\sim\pi}\big(f(X)\big) + \boldsymbol{\beta}^T\boldsymbol{C}\boldsymbol{\beta} + 2\boldsymbol{\beta}\mathrm{Cov}_{X\sim\pi}\big(f(X), \boldsymbol{g}(X)\big) \tag{24}$$

where $\boldsymbol{C}$ is the covariance matrix of $\boldsymbol{g}$. The above is minimized when

$$\boldsymbol{\beta} = -\boldsymbol{C}^{-1}\mathrm{Cov}_{X\sim\pi}\big(f(X), \boldsymbol{g}(X)\big) \tag{25}$$