4 Importance Sampling

Importance Sampling (IS) is the most basic and effective method for variance reduction of Monte Carlo with iid samples. The idea is to sample X from a different distribution than the original one, and to compensate for that by assigning *weights* to the samples. As we shall see, the IS sampling distribution $g(x)$ should ideally be proportional to $|H(x)|f(x)$.

4.1 The IS Estimator

Definition: Recall that we wish to estimate the expected value

$$\ell = E_f(H(X)) = \int H(x)f(x)dx \,,$$

(where $dx = dx_1 \dots dx_n$). Let g be a pdf that dominates f, in the sense that $g(x) = 0 \Rightarrow f(x) = 0$. Then

$$\ell = \int H(x)\frac{f(x)}{g(x)}g(x)dx = E_g(H(X)\frac{f(X)}{g(X)}) \,.$$

Consequently, if $X_1, \dots, X_N$ is an iid sample from g, then the following IS estimate

$$\hat{\ell} = \frac{1}{N}\sum_{i=1}^{N} H(X_i)\frac{f(X_i)}{g(X_i)}$$

is an *unbiased* estimator for ℓ.

The pdf g is called the IS distribution, or *trial distribution.* The ratio

$$W(x) = \frac{f(x)}{g(x)}$$

is the *likelihood ratio* of f and g (more formally, it is the Radon-Nikodym derivative of the respective measures). Denoting $w_i = \frac{f(X_i)}{g(X_i)}$, we can write the estimator as

$$\hat{\ell} = \frac{1}{N}\sum_{i=1}^{N} H(X_i)w_i \,, \quad w_i = \frac{f(X_i)}{g(X_i)}, \;\; X_i \sim g \,.$$

We refer the the w_i's as the IS *weights*, and to the sequence (X_i, w_i) as a *weighted sample* from g.

Monte Carlo Methods – Lecture Notes, N. Shimkin, Spring 2015

We note that the same IS estimator can be be used under the relaxed condition that g dominates Hf (rather than f alone), namely $g(x) = 0 \Rightarrow H(x)f(x) = 0$. In that case we formally set $0 \cdot \infty = 0$.

Bias and variance: The IS estimator is unbiased by construction, as

$$E_g(H(X)\frac{f(X)}{g(X)}) = \int H(X)\frac{f(X)}{g(X)}g(x)dx = \int H(X)f(X)dx = \ell\,.$$

The *sample variance* is given by

$$\begin{aligned} V_g &\triangleq \mathrm{Var}_g(H(X)W(X)) \\ &= E_g(H(X)^2W(X)^2) - \ell^2 \\ &= E_f(H(X)^2\frac{f(X)}{g(X)}) - \ell^2 \end{aligned}$$

Proposition 4.1 *V_g is minimized by choosing $g(x)$ proportional to $|H(x)|f(x)$, namely*

$$g^*(x) = \frac{|H(x)|f(x)}{\int |H(x)|f(x)dx}\,.$$

The minimal variance is

$$V_{g^*} = (E_f|H(X)|)^2 - \ell^2\,.$$

Proof: Apply Jensen's inequality to $E_g((HW)^2)$. □

We refer to g^* as the *optimal IS distribution.*

In particular, if $H(x) \geq 0$, we actually obtain $V_{g^*} = 0$. This means that ℓ can be precisely estimated using one sample!

Unfortunately, this observation is not useful. To see the problem, note that for $H > 0$, $g^*(x) = \frac{1}{\ell}H(x)f(x)$, which directly involves ℓ. The "estimate" here is obtained by sampling X_1 from g^*, and then outputting $H(X_i)W(X_1) = \ell$. Clearly, sampling plays no role here.

Our goal can therefore be stated as finding a trial distribution g which is easy to compute, and roughly approximates g^*.

Normalized IS: It is often the case that $f(x)$ is known only up to a multiplicative constant, namely $f(x) = Cf_0(x)$ with C unknown (recall the Boltzmann distribution example).

In that case we can use a normalized version of the IS estimator. Observe that $E_g(W(X)) = 1$, so that

$$\ell = E_g(H(X)W(X)) = \frac{E_g(H(X)W(X))}{E_g(W(X))} .$$

This suggests the following so-called *weighted sample estimator*:

$$\hat{\ell}_w = \frac{\sum_{i=1}^N H(X_i)w_i}{\sum_{i=1}^N w_i} , \quad w_i = \frac{f(X_i)}{g(X_i)}, \ \ X_i \sim g .$$

Since the weights appear both in the nominator and the denominator, is enough to know the w_i's (hence f and even g) up to a multiplicative constant.

Bias and Variance: It may be seen that the weighted sample estimator is no longer unbiased. However, the bias decreases rapidly with N.

The variance of the estimator is also increased by the randomness in the denominator. A rough estimate (which neglected dependence between the nominator and denominator) can be seen to be

$$\mathrm{Var}(\hat{\ell}) \approx \frac{1}{N} \frac{\mathrm{Var}_g(H(X)W(X))}{(E_g W(X))^2} (1 + \frac{\mathrm{Var}_g(W(X))}{(E_g W(X))^2} .$$

Note that each of these terms can be estimated using the weighted sample (X_i, w_i).

Example. Consider estimating $\ell = \mathbb{P}(X > \gamma)$, $X \sim \mathrm{Exp}(\mu)$, with $\mu\gamma >> 1$.

a. Compute κ^2, the squared coefficient of variation, for crude MC.

b. Compute g^*.

c. For $g(x) = \theta \mathrm{e}^{-\theta(x-\gamma)} 1_{\{x \geq \gamma\}}$, compute κ^2 as a function of α.

d. For $g(x) = \theta \mathrm{e}^{-\theta x} 1_{\{x \geq 0\}}$, find θ that minimizes the variance, and compute the corresponding κ^2.

4.2 Choosing g – The Variance Minimization Method

As choosing the trial distribution g equal to g^*, the optimal OS distribution, is infeasible, we often try to choose g as the "best" distribution out of a specific set $\mathcal{G}$ of probability distributions.

For example, a common choice (in the one-dimensional case) is the set of *exponentially titled* distributions,

$$\mathcal{G} = \{g(\cdot, \theta),\, \theta \in \Theta \subset \mathbb{R}\}, \quad g(x, \theta) = c(\theta)\mathrm{e}^{-\theta x} f_0(x)\,.$$

Here f_0 is the basic distribution, possibly taken as $f_0 = f$, and $c(\theta)$ is the normalization constant.

More generally, $\mathcal{G}$ is often taken as an *exponential family* of probability distributions, which has the following general form:

$$\begin{aligned} \mathcal{G} =& \{g(\cdot, v),\ v \in V \subset \mathbb{R}^{m_0}\}\,, \\ & g(x, v) = c_0(v)\mathrm{e}^{\theta(v)\cdot t(x)} h(x)\,. \end{aligned}$$

Here $\theta(v) = (\theta_1(v), \ldots, \theta_m(v))$ is a vector of functions of the parameters v, $t(x) = (t_1(x), \ldots, t_m(x))$, $h(x) \geq 0$, and $c_0(v)$ is the normalization constant.

By re-parameterization, any exponential family can be represented in the canonical form of a Natural Exponential Family (NEF):

$$\begin{aligned} \mathcal{G} =& \{g(\cdot, \theta),\ \theta \in \Theta \subset \mathbb{R}^{m}\}\,, \\ & g(x, \theta) = c(\theta)\mathrm{e}^{\theta\cdot t(x)} h(x)\,. \end{aligned}$$

Many commonly used distributions belong to an exponential family, including Bernoulli, binomial, Poisson, exponential, Pareto, Weibull, Laplace, chi-squared, normal, lognormal, gamma, beta, multivariate normal, Dirichlet, and multinomial. Some univariate examples:

1. Exponential Exp(λ): $\theta = -\lambda$, $t(x) = x$, $h(x) = 1$, $c(\theta) = -\lambda$
2. Poisson Poi(λ): $\theta = \ln(\lambda)$, $t(x) = x$, $h(x) = \frac{1}{x!}$, $c(\theta) = \exp(-\mathrm{e}^{\theta})$
3. Geometric G(p): $\theta = \ln(1-p)$, $t(x) = x - 1$, $h(x) = 1$, $c(\theta) = 1 - \mathrm{e}^{\theta}$
4. Binomial Bin(n, p): $\theta = \ln(\frac{p}{1-p})$, $t(x) = x$, $h(x) = \binom{n}{x}$, $c(\theta) = (1 + \mathrm{e}^{\theta})^{-n}$
5. Normal N(μ, σ^2): $\theta = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$, $t(x) = (x, x^2)$, $h(x) = 1$, $c(\theta) = \frac{\exp(\theta_1^2/4\theta_2)}{\sqrt{-\pi/\theta_2}}$

The choice of $\mathcal{G}$ should be such that some member of $\mathcal{G}$ can well approximate the shape of $g^* = c|H|f$.

Given the set $\mathcal{G} = \{g(\cdot,\theta)\}$, we wish to find a parameter θ that minimizes the estimator variance. This gives rise to the following (parametric) optimization problem:

$$\begin{aligned}
&\min_{\theta\in\Theta} \mathrm{Var}(\theta)\,, \\
&\qquad\qquad \mathrm{Var}(\theta) \triangleq \mathrm{Var}_{X\sim g(\cdot,\theta)}(H(X)W(X,\theta)),\;\; W(x,\theta) = \frac{f(x)}{g(x,\theta)}
\end{aligned}$$

Recall that

$$\begin{aligned}
\mathrm{Var}(\theta) &= \mathbb{E}_{X\sim g(\cdot,\theta)}(H(X)^2 W(X,\theta)^2) - \ell^2 \\
&= \mathbb{E}_{X\sim f}(H(X)^2 W(X,\theta)) - \ell^2\,.
\end{aligned}$$

Since the mean ℓ does not depend on g, we obtain the following equivalent optimization problem for θ:

$$\begin{aligned}
&\min_{\theta\in\Theta} V(\theta)\,, \\
&\qquad V(\theta) \triangleq \mathbb{E}_{X\sim f}(H(X)^2 W(X,\theta))\,.
\end{aligned}$$

Such an optimization problem, which involves an expected value in the cost function, is generally called a *stochastic* program. We refer to the specific problem here as the *Variance Minimization (VM) problem.*

An analytic solution to the VM problem is seldom feasible. However, in many cases of interest the function $V(\theta)$ is well behaved (e.g., convex and smooth), and can be minimized numerically. Assuming that the derivative and expectation can be interchanged (which holds under reasonable conditions), we obtain the gradient

$$\nabla V(\theta) = \mathbb{E}_f(H(X)^2 \nabla_\theta W(X,\theta)),$$

where

$$\nabla_\theta W(x,\theta) = \nabla_\theta \frac{f(x)}{g(x,\theta)} = -W(x,\theta)\nabla_\theta \ln g(x,\theta)\,.$$

In some cases of interest the gradient can be computed in closed form. The first order condition for optimality is $\nabla V(\theta) = 0$. This equation may then be solved numerically, e.g., using gradient descent.

If the expected value in the cost (or its gradient) is not tractable, an alternative is to use a *sampled approximation* of the VM problem. That is,

$$\begin{aligned}
&\min_{\theta\in\Theta} \hat{V}(\theta)\,, \\
&\qquad \hat{V}(\theta) = \frac{1}{K}\sum_{k=1}^{K} H(X_k)^2 W(X_k,\theta)\,,
\end{aligned}$$

where $(X_1, \ldots, X_K)$ is an iid sample from f. We refer to this problem as the *sampled VM program*. Note that, once the X_k's are available, we obtain a deterministic program. This problem is typically solved numerically, with the gradient computed similarly to the above.

A basic scheme that uses the sampled VM program proceeds as follows:

1. Obtain a test sample $X_1, \ldots X_K$ from f.
2. Choose θ by solving the sampled VM program.
3. Estimate ℓ using an IS estimator, with $g = g(\cdot, \theta)$.

Iterated Procedure: In some cases it might be ineffective to obtain the test sample $X_1, \ldots X_K$ from f, and we wish to take our test sample from some initial guess g_0 which may be closer to g^*. To that end, observe that

$$V(\theta) = \mathbb{E}_{X \sim g_0}(H(X)^2 W(X, \theta) W_0(X)), \quad W_0(x) \triangleq \frac{f(x)}{g_0(x)} .$$

This leads to the sampled cost

$$\hat{V}(\theta) = \frac{1}{K} \sum_{k=1}^{K} H(X_k)^2 W(X_k, \theta) W_0(X_k), \quad X_k \sim g_0 ,$$

from which we obtain the test distribution $g_1 = g(\cdot, \theta^*)$. This procedure of optimizing over θ may be repeated several times, each time sampling (X_k) from the test distribution g_{i-1} obtained in the previous round.

Such iterative refinement methods should be used with care, to avoid *degeneracy* of the distributions g_i.

4.3 Choosing g – The Cross Entropy Method

An alternative to minimizing the variance directly, is to choose g which is close to $g^*(x) = c|H(x)|f(x)$. A standard measure for the distance between two probability distributions is the Kullback-Leibler number (also known as the information divergence or relative entropy),

$$D_{KL}(f, g) = \mathbb{E}_g(\ln \frac{f(X)}{g(X)}) = \int f(x) \ln \frac{f(x)}{g(x)} dx$$

From Jensen's inequality,

$$D_{KL}(f,g) = -\mathbb{E}_f(\ln \frac{g(X)}{f(X)}) \geq -\ln \mathbb{E}_f(\frac{g(X)}{f(X)}) = 0\,,$$

with equality only if $f = g$. We note however that D_{KL} is *not* a metric, as it is not commutative, and does not satisfy the triangle inequality. Observe also that

$$\begin{aligned} D_{KL}(f,g) &= \int f(x)\ln(f(x))dx - \int f(x)\ln(g(x))dx \\ &= -H(f) + H(f,g)\,, \end{aligned}$$

where $H(f)$ is the *entropy* of f, and $H(f,g)$ the *Cross Entropy* (CE) between f and g.

Suppose that we wish to solve

$$\min_{\theta\in\Theta} D_{KL}(g^*, g(\cdot,\theta))\,.$$

As g^* is fixed, this is equivalent to

$$\min_{\theta\in\Theta} H(g^*, g(\cdot,\theta))\,,$$

which, in turn, is equivalent to

$$\max_{\theta\in\Theta} \;\; L(\theta)$$

$$L(\theta) \triangleq \int |H(x)| f(x) \ln g(x,\theta) dx = E_f(|H(X)| \ln g(X,\theta))$$

(note that the normalization constant c in g^* was dropped). The latter program is *CE optimization problem.*

The solution may be obtained as in the previous (VM) problem. Assuming that the derivative and expectation can be interchanged, we obtain the gradient

$$\nabla L(\theta) = \mathbb{E}_f(|H(X)|\nabla \ln g(X,\theta))\,.$$

The *sampled* CE optimization problem is given by

$$\max_{\theta\in\Theta} \hat{L}(\theta)\,,$$

$$\hat{L}(\theta) = \frac{1}{K}\sum_{k=1}^{K} |H(X_k)| \ln g(X_k,\theta)\,, \quad X_k \sim f\,.$$

We note that this program is similar to the MLE problem for estimating the parameter θ from samples (X_k), with the addition of "weights" $H(X_k)$.

An *iterative* scheme may be obtained, as before, by noting that

$$L(\theta) = E_{g_0}(|H(X)| W_0(X) \ln g(X, \theta)), \quad W_0(x) = \frac{f(x)}{g_0(x)} \,.$$

An advantage of the CE method relative to the VM method is that analytical solutions may be obtained in a wider set of problems. Numerical experiments show that the CE method may also be more stable for numerical optimization, and provides similar solutions (for θ) for moderate dimensions n of X, say $n \leq 50$. However, for higher dimensional problems, VM outperforms CE in terms of the resulting estimator variance.

Example: An analytic solution for exponential tilting. Consider the single-parameter exponential family that corresponds to exponential tilting:

$$g(x, \theta) = c(\theta) \mathrm{e}^{x\theta} g_0(x) = \mathrm{e}^{x\theta - \zeta(\theta)} g_0(x), \quad \theta \in \mathbb{R} \,,$$

where $\zeta(x) = -\ln c(x)$. We wish to maximize $L(\theta)$. Then the first-order condition $\nabla L(\theta) = 0$ implies

$$\zeta'(\theta) = \frac{\mathbb{E}_f(|H(X)| X)}{\mathbb{E}_f |H(X)|} \,.$$

Furthermore, if the parameter is chosen such that θ is the mean of $g(\cdot, \theta)$, namely $E_\theta(X) = \theta$, then $\zeta'(\theta) = \theta$, and consequently

$$\theta^* = \frac{\mathbb{E}_f(|H(X)| X)}{\mathbb{E}_f |H(X)|} \,.$$

4.4 Bayesian Inference

Consider the Bayesian point-estimation of an RV X based on measurement Y. Given are

$f_X(x)$	–	prior distribution of X (the 'state variable').
$f_{Y\|X}(y\|x)$	–	distribution of the measurement Y given state $X = x$ (the *likelihood function*).

We wish to compute the MMSE (Minimal Mean Square Error) estimate of X given Y:

$$\hat{X}(y) = E(X|Y = y)\,.$$

Example: A familiar problem in the engineering context is the linear model with additive noise:

$$Y = AX + V\,,$$

where A is a known matrix, and V the additive noise which is independent of X. More generally, we may consider the nonlinear model with additive noise,

$$Y = h(X) + V\,,$$

where h is a given function.

Recall that

$$E(X|Y = y) = \int x f_{X|Y}(x|y)dx\,,$$

where $f(x|y)$ can be calculated using Bayes formula

$$\begin{aligned} f_{X|Y}(x|y) =& \frac{1}{C(y)} f_X(x) f_{Y|X}(y|x)\,, \\ C(y) = f_Y(y) =& \int f_X(x) f_{Y|X}(y|x)dx\,. \end{aligned}$$

An analytical expression for $\hat{X}(y)$ is available only in special cases, and in general we require a numerical computation. Importance Sampling is one of the major tools used for this purpose.

For a given measured value y, let $g_y(x)$ be a trial distribution (in x) which dominates $f_X(x)f_{Y|X}(y|x)$. An IS estimate of $\hat{X}(y)$ is given by

$$\hat{\ell} = \frac{1}{N}\sum_{i=1}^{N} X_i W(X_i)\,,$$

where (X_i) is an iid sample from g, and

$$W(x) = \frac{f_{X|Y}(x|y)}{g_y(x)} = \frac{1}{C(y)} \frac{f_X(x) f_{Y|X}(y|x)}{g_y(x)} \,.$$

Since $C(y)$ is often hard to compute, we can use the *weighted* IS estimate:

$$\hat{\ell}_w = \frac{\sum_{i=1}^N X_i \tilde{W}(X_i)}{\sum_{i=1}^N \tilde{W}(X_i)} \,, \quad \tilde{W}(x) = \frac{f_X(x) f_{Y|X}(y|x)}{g(x)} \,.$$

The MSE of this estimator can be similarly estimated:

$$\text{MSE} = E(X - \hat{X}(y))^2 | Y = y) \approx \frac{\sum_{i=1}^N (X_i - \hat{\ell}_w)^2 \tilde{W}(X_i)}{\sum_{i=1}^N \tilde{W}(X_i)}$$

(possibly multiplied by $\frac{N}{N-1}$). Note that this is a different quantity than the variance $\text{Var}(\ell)$ of the MC estimator, that was discussed in Lecture 3.

Choosing g: The test distribution g may be simply chosen as the prior distribution f_X: $g(x) = f_X(x)$. This simplifies the calculation of the weights $\tilde{W}(X_i)$. Note however that the optimal (minimum variance) test distribution is proportional to $xf(x|y)$. Therefore, if $f(x|y)$ is significantly different from the prior $f(x)$, it may be a good idea to compute first a rough estimate of $f(x|y)$ (e.g., by a Gaussian approximation), and use it for g.

Empirical Distribution: In some cases it is required to generate an estimate for the entire posterior distribution, $f_{X|Y}(\cdot|y)$. This is used, for example, in state estimation of dynamic systems, using the so-called Particle Filter.

Using the weighted sample (X_i, w_i) from g, let

$$\hat{f}_N(x) = \frac{1}{N} \sum_{i=1}^N w_i \delta_{X_i}(x) \,.$$

Here δ_z is the delta function that puts unit mass at point z (in a continuous space this is the Dirac delta function, $\delta_z(x) = \delta(x - z)$, while for a discrete space this is the Kroeneker delta). It is easy to see that $\hat{f}$ is a probability distribution, and it provides an unbiased representation of $f_{X|Y}(\cdot|y)$ in the sense that, for any function $H(x)$,

$$E(\int H(x) \hat{f}_N(x) dx) = \int H(x) f_{X|Y}(x|y) dx$$

(verify that).