Exercise – Understanding the theory of importance sampling

1. Aim of the exercise

The aim of this exercise is to review the theory of importance sampling and weighted importance sampling.

2. Theory

The standard deviation of a Monte Carlo (MC) integration estimator decreases at the rate of $\mathcal{O}(n^{-1/2})$, as discussed in the theory of MC integration. The notation $\mathcal{O}(\cdot)$ describes how quickly the estimation error decreases as the sample size n increases. This error is often measured using the mean squared error (MSE). For MC integration, the estimator is unbiased, so the MSE equals the variance. Therefore, the convergence rate $\mathcal{O}(n^{-1/2})$ characterizes how both the variance and the MSE decrease with increasing n. While this rate is fundamental to standard MC estimators, it does not preclude the possibility of constructing more efficient estimators. Variance reduction techniques aim to improve the precision of estimates by lowering the variance, thereby achieving better accuracy without increasing the sample size.

Importance sampling (IS) is one of the most widely used variance reduction techniques. The introduction of this concept in statistics and econometrics is commonly credited to Kloek and van Dijk (1978). Importance sampling involves selecting a proposal distribution that favors important samples, specifically, those that lie in regions contributing significantly to the value of an integral. These regions typically correspond to areas with relatively high probability density. By concentrating sampling effort in these regions, importance sampling can lead to substantial reductions in variance (Rubinstein, 2016).

Importance sampling can be used to reduce the variance in MC integration. The goal is to select a proposal PDF, p_X, such that the variance of the resulting estimator, which emphasizes important samples, is smaller than that of the standard Monte Carlo estimator $\bar{g}_n$:

$$\frac{1}{n}\operatorname{Var}_{p_X}\left[g(X)\frac{f_X(X)}{p_X(X)}\right] < \frac{1}{n}\operatorname{Var}_{f_X}\left[g(X)\right] = \operatorname{Var}_{f_X}\left[\bar{g}_n\right].$$

The final equality holds because $\bar{g}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$ is the sample mean of n i.i.d. evaluations of $g(X)$, and the variance of a sample mean is equal to the variance of the individual terms divided by n. That is,

$$\operatorname{Var}_{f_X}(\bar{g}_n) = \operatorname{Var}_{f_X}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right) = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}_{f_X}(g(X_i)) = \frac{1}{n}\operatorname{Var}_{f_X}(g(X)).$$

This inequality illustrates how a well-chosen proposal distribution can concentrate sampling effort in regions that contribute most to the integral, thereby achieving lower variance and improved estimation efficiency without increasing the sample size.

The primary objective is to reduce the variance in approximating the integral estimated using MC integration:

$$\int_A \phi(x)\,dx = \int_A g(x)\cdot f_X(x)\,dx = \mathbb{E}_{f_X}[g(X)] =: \theta_{\mathrm{MC}}. \tag{1}$$

Let us call f_X the target PDF. The idea of importance sampling is as follows:

$$\begin{aligned}\int_A \phi(x)\,dx &= \int_{\mathbb{R}} g(x)\cdot f_X(x)\,dx \\ &= \int_{\mathbb{R}} g(x)\cdot \frac{f_X(x)}{p_X(x)}\cdot p_X(x)\,dx \\ &= \mathbb{E}_{p_X}\left[g(X)\cdot\frac{f_X(X)}{p_X(X)}\right] \\ &:= \mathbb{E}_{p_X}\left[g(X)\cdot W(X)\right] \\ &:= \theta_{\text{IS}} \\ &\overset{!}{=} \mathbb{E}_{f_X}\left[g(X)\right]\end{aligned} \tag{2}$$

where $W(X) := \frac{f_X(X)}{p_X(X)}$ is called the likelihood ratio. In stochastic calculus and measure theory, this random variable is better known as the Radon-Nikodym derivative. This shows that θ_{MC} and θ_{IS} are different representations of the integral

$$\int_A \phi(x)\,dx \tag{3}$$

where $A \subseteq \mathbb{R}$, and $\phi : \mathbb{R} \to \mathbb{R}$ is an integrable function.

That is, both estimators compute the same integral: $\theta_{\text{MC}} = \theta_{\text{IS}}$. Importance sampling rewrites the original integral as an expectation under a different probability distribution. The core idea is to shift the sampling burden from a difficult distribution f_X to a more convenient one p_X, while still recovering the correct expectation. Instead of sampling directly from f_X, we sample from p_X and compensate for the mismatch by weighting each sample appropriately. These weights, given by the ratio $\frac{f_X(x)}{p_X(x)}$, ensure that samples drawn from p_X contribute to the average as if they had been drawn from f_X. This reweighting corrects for the fact that we are drawing samples from p_X instead of f_X, allowing us to approximate expectations under f_X using samples from a different distribution.

Moreover, importance sampling can simplify the sampling process itself. If f_X is a difficult density to sample from directly, we can instead choose a proposal distribution p_X that is easier to sample from. Although the integrand becomes more complex, since it now includes the ratio $\frac{f_X(X)}{p_X(X)}$, this trade-off is often worthwhile because evaluating $f_X(x)$ is typically much easier than generating samples from it. Importance sampling thus avoids direct sampling from f_X by reweighting samples drawn from p_X using the likelihood ratio. Another powerful method that can address the problem of estimating expectations when direct sampling from f_X is difficult is the Markov Chain Monte Carlo. This method will be discussed in a separate exercise.

These show that importance sampling yields two advantages: a more efficient estimator of the integral in equation (3), and a procedure that enables us to perform Monte Carlo integration even when it is difficult to generate samples from f_X, provided we can evaluate $f_X(x)$ and sample from a suitable proposal distribution p_X.

Table 1 presents an overview of the conditions for choosing a proposal density p_X in importance sampling. Conditions 1 to 4 are the assumptions required for validity, and the remaining conditions are design guidelines that improve efficiency. We elaborate on these critical assumptions, beginning with support coverage and dominance.

Table 1: Assumptions for valid importance sampling and design guidelines for efficient importance sampling

IS.1	$\text{Supp}(g \cdot f_X) \subseteq \text{Supp}(p_X)$
IS.2	$p_X(x) = 0$ implies $g(x) \cdot f_X(x) = 0$
IS.3	$\frac{f_X(x)}{p_X(x)} < M \in \mathbb{R},\ \forall x \in \mathbb{R}$ and $\text{Var}_{f_X}[g(X)] < \infty$
IS.4	$\int_{\mathbb{R}} \left(\frac{f_X(x)}{p_X(x)}\right)^2 g(x)^2 p_X(x)\, dx < \infty$
IS.5	p_X is easy to sample from
IS.6	p_X closely matches the shape of $g(x) \cdot f_X(x)$
IS.7	p_X has heavier tails than f_X

Assumption IS.1 requires that the support of the proposal distribution covers the support of the integrand. This condition ensures that the proposal density p_X assigns positive probability to all regions where the integrand $g(x) \cdot f_X(x)$ is nonzero. If this condition is violated, parts of the integration domain may be ignored entirely, effectively truncating the integral and introducing bias into the estimate.

Assumption IS.2 means that if the proposal distribution p_X assigns zero probability to some value of x, then the target integrand $g(x)f_X(x)$ must also be zero at that point. In plain terms, we should never try to estimate something in a region where our proposal distribution never samples. This matters because in importance sampling we compute weights like $\frac{f_X(x)}{p_X(x)}$. If $p_X(x) = 0$ but $g(x)f_X(x) \neq 0$, then we would be dividing by zero, causing the weight to blow up and the estimator to be undefined. Assumption IS.2 prevents this by requiring that wherever p_X is zero, the integrand is also zero, ensuring that the ratio is always well-defined.

Assumption IS.3 requires that the likelihood ratio $\frac{f_X(x)}{p_X(x)}$ is bounded by some finite constant M for all values of x, and that the variance of $g(X)$ under the target distribution f_X is finite. This condition ensures that the importance weights remain controlled and cannot explode to infinity, preventing the estimator from being dominated by extreme values. If the ratio were unbounded, a few sampled points could carry disproportionately large weights, leading to instability and unreliable estimates. By bounding the ratio and requiring finite variance, IS.3 guarantees that the importance sampling estimator is well-defined, stable, and converges with manageable variability.

Assumption IS.4 requires that the second moment of the importance sampling estimator is finite. This condition ensures that the variance of the estimator exists and is well-defined. Intuitively, it means that when we reweight samples using the likelihood ratio $\frac{f_X(x)}{p_X(x)}$, the squared weights combined with the function $g(x)$ do not produce infinite contributions. If this integral were infinite, the estimator would have unbounded variance, making it unstable and unreliable. By requiring finiteness, IS.4 guarantees that importance sampling produces estimates that converge with controlled variability rather than being dominated by extreme values.

With these assumptions in place, we can now express how importance sampling allows us to compute expectations under a target distribution by reweighting samples drawn from a more convenient proposal distribution.

Proposition 1 (Importance sampling recovers target expectations)**.** *Let $X \sim p_X$ be a draw from a proposal distribution, and let f_X denote the density of a target distribution. Suppose that* *whenever $f_X(x) > 0$, it follows that $p_X(x) > 0$. Then, for any measurable function $h : \mathbb{R} \to \mathbb{R}$,*

$$\mathbb{E}_f[h(X)] = \mathbb{E}_p \left[h(X) \cdot \frac{f_X(X)}{p_X(X)} \right].$$

This identity expresses an equality between two expectations: one taken under the target distribution f_X, and the other under the proposal distribution p_X. The function h is arbitrary, as long as it is measurable and the expectations are well-defined. The key feature of the right-hand side is the presence of the weight $\frac{f_X(X)}{p_X(X)}$, which adjusts for the fact that the distribution of X is p_X, not f_X.

To understand why this identity holds, consider the change-of-measure principle. For any measurable function h, we can write:

$$\mathbb{E}_f[h(X)] = \int h(x) f_X(x)\, dx = \int h(x) \frac{f_X(x)}{p_X(x)} p_X(x)\, dx = \mathbb{E}_p \left[h(X) \cdot \frac{f_X(X)}{p_X(X)} \right],$$

where the second equality relies on the assumption that whenever $f_X(x) > 0$, we also have $p_X(x) > 0$. This ensures that the ratio $\frac{f_X(x)}{p_X(x)}$ is well-defined wherever the target density f_X is positive.

This identity is the foundation of importance sampling. It tells us that if we can evaluate the ratio $\frac{f_X(x)}{p_X(x)}$, then we can compute expectations under f_X by averaging weighted values of $h(X)$ over samples drawn from p_X. The weight $\frac{f_X(X)}{p_X(X)}$ amplifies or downscales each sample depending on how typical it is under the target distribution relative to the proposal.

In practice, this is useful when we wish to compute expectations under f_X but prefer to draw samples from p_X, for example because p_X is easier to sample from or better suited to the computational setting. The identity ensures that, with appropriate weighting, the resulting average still targets the correct expectation under f_X.

This identity forms the basis of the importance sampling estimator, which we now construct using Monte Carlo methods.

Proof. Let us apply the importance sampling identity to the function $h(x) = \mathbb{I}\{x \le \delta\}$. Then:

$$\mathbb{E}_f \left[\mathbb{I}\{X \le \delta\} \right] = \mathbb{E}_p \left[\mathbb{I}\{X \le \delta\} \cdot \frac{f_X(X)}{p_X(X)} \right].$$

We now compute both sides explicitly to verify this equality. First, consider the left-hand side:

$$\mathbb{E}_f \left[\mathbb{I}\{X \le \delta\} \right] = \int_{\mathbb{R}} \mathbb{I}\{x \le \delta\} \cdot f_X(x)\, dx.$$

This integral restricts the domain to $x \le \delta$, so we can write:

$$= \int_{-\infty}^{\delta} f_X(x)\, dx.$$

Next, consider the right-hand side:

$$\mathbb{E}_p \left[\mathbb{I}\{X \le \delta\} \cdot \frac{f_X(X)}{p_X(X)} \right] = \int_{\mathbb{R}} \mathbb{I}\{x \le \delta\} \cdot \frac{f_X(x)}{p_X(x)} \cdot p_X(x)\, dx.$$

Here, the $p_X(x)$ terms cancel:

$$= \int_{\mathbb{R}} \mathbb{I}\{x \leq \delta\} \cdot f_X(x)\, dx = \int_{-\infty}^{\delta} f_X(x)\, dx.$$

So both sides equal the same integral. Therefore:

$$\mathbb{E}_f \left[\mathbb{I}\{X \leq \delta\}\right] = \mathbb{E}_p \left[\mathbb{I}\{X \leq \delta\} \cdot \frac{f_X(X)}{p_X(X)}\right].$$

■

The importance sampling fundamental identity (Robert and Casella, 2010) gives an unbiased importance sampling, or likelihood ratio, estimator for any proposal PDF p_X that satisfies IS.1:

Theorem 1. *Let $X \sim p_X$ and let $X_1, ..., X_n \sim p_X$ be an i.i.d. random sample with realizations $x_1, ..., x_n$ such that $\mathbb{E}_{p_X}[|X_i|] < \infty$, $\forall i \in \{1, ..., n\}$. Then, an unbiased estimator of θ_{IS} is:*

$$\hat{\theta}_{IS} := \frac{1}{n}\sum_{i=1}^{n} g(x_i) \cdot \frac{f_X(x_i)}{p_X(x_i)} \xrightarrow{a.s.} \mathbb{E}_{p_X}\left[\frac{g(X) \cdot f_X(X)}{p_X(X)}\right] = \theta_{IS}. \tag{4}$$

Here, the almost surely convergence holds by the SLLN.

Proof. $\hat{\theta}_{IS}$ is an unbiased estimator of θ_{IS} because:

$$\begin{aligned} \mathbb{E}_{p_X}\left[\hat{\theta}_{IS}\right] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{p_X}\left[g(X_i) \cdot \frac{f_X(X_i)}{p_X(X_i)}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}} g(x_i) \cdot \frac{f_X(x_i)}{p_X(x_i)} \cdot p_X(x_i)\, dx \\ &= \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}} g(x_i) \cdot f_X(x_i)\, dx \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{f_X}\left[g(X_i)\right] \overset{i.d.}{=} \frac{1}{n} \cdot n \cdot \mathbb{E}_{f_X}\left[g(X)\right] = \mathbb{E}_{f_X}\left[g(X)\right] = \theta_{IS} \end{aligned} \tag{5}$$

where the first equality uses the linearity of expectation. ■

Observe that the MC integration estimator

$$\bar{g}_n := \frac{1}{n}\sum_{i=1}^{N} g(X_i) \tag{6}$$

is a special case of the importance sampling estimator in equation (4), if $p_X \overset{d}{=} f_X$ and $W(X) := \frac{f_X(X)}{p_X(X)} \equiv 1$.

To assess the efficiency of importance sampling, we analyze the variance of the estimator and explore strategies for minimizing it. Observe that the variance of the MC integration estimator can be rewritten in terms of θ_{IS}:

$$\begin{aligned} \mathrm{Var}[\bar{g}_n] = \frac{1}{n}\mathrm{Var}_{f_X}\left[g(X)\right] &= \frac{1}{n} \cdot \left[\mathbb{E}_{f_X}\left[g(X)^2\right] - \mathbb{E}_{f_X}\left[g(X)\right]^2\right] \\ &= \frac{1}{n} \cdot \left[\int_{\mathbb{R}} g(x)^2 \cdot f_X(x)\, dx - \theta_{IS}^2\right], \end{aligned} \tag{7}$$

where $X \sim f_X(x)$. The variance of the importance sampling estimator can be written in a similar fashion.

Proposition 2 (Variance of the importance sampling estimator)**.** *The variance of the importance sampling estimator is given by*

$$\mathrm{Var}_{p_X}\left[\hat{\theta}_{IS}\right] = \frac{1}{n}\mathrm{Var}_{p_X}\left[g(X)\cdot W(X)\right] = \frac{1}{n}\cdot\left[\int_{\mathbb{R}} g(x)^2 \cdot W(x)^2 \cdot p_X(x)\,dx - \theta_{IS}^2\right]. \tag{8}$$

This expression shows that the variance decreases at rate $1/n$ and depends on the second moment of the weighted integrand under the proposal distribution. To estimate this variance in practice, we use the unbiased sample variance of the weighted terms:

$$s_{IS}^2 := \widehat{\mathrm{Var}_{p_X}\left[\hat{\theta}_{IS}\right]} = \frac{1}{n-1}\sum_{i=1}^{n}\left[g(x_i)\cdot W(x_i) - \hat{\theta}_{IS}\right]^2. \tag{9}$$

Proof. Equation (8) follows by the definition of variance. Next, let $\sigma^2 := \mathrm{Var}_{p_X}\left[g(X_i)\cdot W(X_i)\right]$. The sample variance s_{IS}^2 is unbiased because:

$$\begin{aligned}
\mathbb{E}_{p_X}\left[s_{IS}^2\right] &= \frac{1}{n-1}\cdot\mathbb{E}_{p_X}\left[\sum_{i=1}^{n}\left(g(X_i)\cdot W(X_i) - \hat{\theta}_{IS}\right)^2\right] \\
&= \frac{1}{n-1}\cdot\mathbb{E}_{p_X}\left[\sum_{i=1}^{n} g^2(X_i)\cdot W^2(X_i) - 2\hat{\theta}_{IS}\cdot\sum_{i=1}^{n} g(X_i)\cdot W(X_i) + \sum_{i=1}^{n}\hat{\theta}_{IS}^2\right] \\
&= \frac{1}{n-1}\cdot\mathbb{E}_{p_X}\left[\sum_{i=1}^{n} g^2(X_i)\cdot W^2(X_i) - n\cdot\hat{\theta}_{IS}^2\right] \\
&= \frac{1}{n-1}\cdot\sum_{i=1}^{n}\mathbb{E}_{p_X}\left[g^2(X_i)\cdot W^2(X_i)\right] - \frac{n}{n-1}\cdot\mathbb{E}_{p_X}\left[\hat{\theta}_{IS}^2\right] \qquad (10) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\sigma^2 + \mathbb{E}_{p_X}\left[g(X_i)\cdot W(X_i)\right]^2\cdot\right) - \frac{n}{n-1}\cdot\left(\mathrm{Var}_{p_X}\left[\hat{\theta}_{IS}\right] + \mathbb{E}_{p_X}\left[\hat{\theta}_{IS}\right]^2\right) \\
&= \frac{1}{n-1}\cdot\left(n\cdot\sigma^2 + n\cdot\mathbb{E}_{p_X}\left[g(X_i)\cdot W(X_i)\right] - n\cdot\frac{n}{n^2}\cdot\sigma^2 - n\cdot\mathbb{E}_{p_X}\left[g(X_i)\cdot W(X_i)\right]\right) \\
&= \frac{1}{n-1}\cdot\left((n-1)\cdot\sigma^2\right) = \sigma^2 := \mathrm{Var}_{p_X}\left[\hat{\theta}_{IS}\right].
\end{aligned}$$

■

Observe that the mean squared error of $\hat{\theta}_{IS}$ is equal to its variance, because $\hat{\theta}_{IS}$ has no bias, i.e. it is an unbiased estimator, by Theorem 1.

To restate the goal of this section, we want to obtain a smaller variance for the Monte Carlo integration estimator. This is achieved when the variance of the importance sampling estimator is smaller than the variance of the plain Monte Carlo estimator. Equivalently, this means that their difference is positive. To see this explicitly, recall that the variance of $g(X)$ under the original sampling distribution f_X is defined as the second moment minus the squared mean,

$$\begin{aligned}
\mathrm{Var}_{f_X}[g(X)] &= \mathbb{E}_{f_X}[g(X)^2] - \left(\mathbb{E}_{f_X}[g(X)]\right)^2 \\
&= \int_{\mathbb{R}} g(x)^2 f_X(x)\,dx - \theta^2,
\end{aligned}$$

where

$$\theta = \int_{\mathbb{R}} g(x) f_X(x)\, dx.$$

In parallel, the variance of the weighted term $g(X)W(X)$ under the proposal distribution p_X is

$$\begin{aligned}\mathrm{Var}_{p_X}[g(X)W(X)] &= \mathbb{E}_{p_X}[(g(X)W(X))^2] - \left(\mathbb{E}_{p_X}[g(X)W(X)]\right)^2 \\ &= \int_{\mathbb{R}} g(x)^2 \left(\frac{f_X(x)}{p_X(x)}\right)^2 p_X(x)\, dx - \theta^2.\end{aligned}$$

Subtracting the second expression from the first cancels the common mean term θ^2, leaving

$$\mathrm{Var}_{f_X}[g(X)] - \mathrm{Var}_{p_X}[g(X)W(X)] = \int_{\mathbb{R}} g(x)^2 \left[f_X(x) - \frac{f_X(x)^2}{p_X(x)} \right] dx,$$

and the final equality is obtained by factoring the bracketed term:

$$\int_{\mathbb{R}} g^2(x) \left[1 - \frac{f_X(x)}{p_X(x)}\right] f_X(x)\, dx.$$

Since $g, f_X \geq 0$, variance reduction requires selecting a proposal PDF p_X such that

$$p_X(x) > f_X(x) \text{ in regions where } g^2(x) f_X(x) \text{ is large,}$$

and

$$p_X(x) < f_X(x) \text{ in regions where } g^2(x) f_X(x) \text{ is small.}$$

This condition follows directly from the variance difference expression, and it highlights that p_X should concentrate probability mass in the "important" regions of the domain where the integrand $\phi(x) = g^2(x) f_X(x)$ contributes most. This motivates the term *importance sampling.*

This motivates the search for an optimal proposal distribution that minimizes the variance of the estimator. In particular, the goal of this section can be stated more formally as minimizing the variance of $\hat{\theta}_{IS}$ with respect to p_X (Rubinstein, 2016):

$$\min_{p_X} \mathrm{Var}_{p_X}\left[g(X) \cdot W(X)\right]. \tag{11}$$

Fortunately, there exists a proposal PDF p_X that achieves the minimum variance, and in fact the minimum value of zero can be attained. This occurs when the weighted term $g(x)W(x)$ is constant almost surely under p_X, so that its variance vanishes. In other words, the optimal proposal distribution is proportional to the integrand itself:

$$p_X^*(x) \propto g(x) f_X(x).$$

With this choice, the importance sampling estimator reduces to the constant θ, and hence

$$\mathrm{Var}_{p_X^*}\left[g(X) \cdot W(X)\right] = 0.$$

This result shows that, in theory, importance sampling can eliminate variance entirely by sampling directly from the normalized integrand. In practice, however, p_X^* depends on the unknown quantity θ, so it cannot be implemented exactly. Nevertheless, it serves as a guiding principle: good proposal distributions should approximate the shape of $g(x) f_X(x)$ as closely as possible.

Theorem 2. *The PDF that minimizes the optimization problem (11) is*

$$p_X^*(x) = \frac{|g(x)| \cdot f_X(x)}{\int_{\mathbb{R}} |g(x)| \cdot f_X(x)\, dx}. \tag{12}$$

This minimizer p_X^* is often referred to as the optimal importance sampling PDF.

Proof. See Robert and Casella (2004). ■

Note that the characterization of the optimal proposal distribution reflects assumption IS.6 in Table 1, since the minimizer p_X^* is constructed to closely match the shape of $g(x) \cdot f_X(x)$.

Working with absolute values is often inconvenient, but the optimal importance sampling PDF in equation (12) simplifies when $g(x) \geq 0$ for all $x \in \mathbb{R}$.

Corollary 1. *If $g \geq 0$, then the optimal importance sampling PDF reduces to*

$$p_X^*(x) = \frac{g(x) \cdot f_X(x)}{\int_{\mathbb{R}} g(x) \cdot f_X(x)\, dx} = \frac{g(x) \cdot f_X(x)}{\theta_{IS}}. \tag{13}$$

For this simplified optimal importance sampling PDF, the variance of $\hat{\theta}_{IS}$ with respect to p_X^* equals zero.

Theorem 3.

$$\mathrm{Var}_{p_X^*}\left[\hat{\theta}_{IS}\right] \overset{(1)}{=} \mathrm{Var}_{p_X^*}\left[g(X) \cdot W^*(X)\right] \overset{(2)}{=} \mathrm{Var}_{p_X^*}\left[\theta_{IS}\right] \overset{(3)}{=} 0, \tag{14}$$

where the likelihood ratio is defined as $W^*(x) := \frac{f_X(x)}{p_X^*(x)}$.

Proof.

$$\begin{aligned}
\mathrm{Var}_{p_X^*}\left[\hat{\theta}_{IS}\right] &= \mathrm{Var}_{p_X^*}\left[\frac{1}{n}\sum_{i=1}^{n} g(X_i) \cdot W^*(X_i)\right] \\
&\overset{ind.}{=} \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}_{p_X^*}\left[g(X_i) \cdot W^*(X_i)\right] \\
&\overset{i.d.}{=} \frac{1}{n} \cdot n \cdot \mathrm{Var}_{p_X^*}\left[g(X) \cdot W^*(X)\right] \\
&\overset{(1)}{=} \mathrm{Var}_{p_X^*}\left[g(X) \cdot W^*(X)\right] \\
&= \mathrm{Var}_{p_X^*}\left[g(X) \cdot \frac{f_X(X)}{p_X^*(X)}\right] \\
&= \mathrm{Var}_{p_X^*}\left[g(X) \cdot f_X(X) \cdot \frac{\theta_{IS}}{g(X) \cdot f_X(X)}\right] \\
&\overset{(2)}{=} \mathrm{Var}_{p_X^*}\left[\theta_{IS}\right] \\
&= \mathrm{Var}_{p_X^*}\left[\mathbb{E}_{f_X}[g(X)]\right] \\
&\overset{(3)}{=} 0.
\end{aligned} \tag{15}$$

■

Despite its theoretical appeal, the optimal proposal PDF is often impractical. We therefore consider a more robust alternative: the auto-normalizing importance sampling estimator.

In general, the implementation of the optimal importance sampling PDF in Corollary 1 is problematic. The difficulty lies in the fact that we need to know θ_{IS} in order to determine p_X^*, but this is precisely the quantity we are trying to estimate. Brandimarte (2014) suggests that we should instead choose a proposal PDF p_X which approximates the product $g(x) \cdot f_X(x)$ as closely as possible. Thus, the real challenge is selecting a proposal PDF p_X that yields a significant reduction in the variance of the estimate. There is no guarantee that the variance will be reduced, because a poor choice of p_X could in fact increase variance. Hence, caution must be taken when designing a proposal distribution.

An importance sampling estimator with infinite variance should be avoided. We have

$$\mathrm{Var}_{p_X}\left[\hat{\theta}_{IS}\right] < \infty \quad \text{if and only if} \quad \mathrm{Var}_{p_X}\left[g(X) \cdot W(X)\right] = \mathbb{E}_{p_X}\left[g^2(X) \cdot W^2(X)\right] - \theta_{IS}^2 < \infty.$$

Since $\theta_{IS}^2 < \infty$ by the integrability of f (i.e. $\int_A |f(x)|\, dx < \infty$), it remains to show that

$$\mathbb{E}_{p_X}\left[g^2(X) \cdot W^2(X)\right] = \int_{\mathbb{R}} g^2(x) \cdot \frac{f_X(x)}{p_X(x)} \cdot p_X(x)\, dx < \infty. \tag{16}$$

Since $\int_A |f(x)|\, dx < \infty$, we know that $\int_{\mathbb{R}} g(x) \cdot f_X(x)\, dx$ is bounded. Therefore, we require that the likelihood ratio $\frac{f_X(x)}{p_X(x)}$ is bounded for all $x \in \mathbb{R}$. Suppose instead that the likelihood ratio is unbounded, i.e. there exists $x_0 \in \mathrm{int}\{\mathrm{Supp}(f_X) \cap \mathrm{Supp}(p_X)\}$ such that

$$\lim_{x \to x_0} \left|\frac{f_X(x)}{p_X(x)}\right| = \infty.$$

This occurs if $\lim_{x \to x_0} p_X(x) = 0$ while $\lim_{x \to x_0} f_X(x) \neq 0$, which implies that p_X has lighter tails than f_X. Hence, we want p_X to have heavier tails than f_X. Figure 1 illustrates this by showing that the standard Cauchy PDF has heavier tails than the standard normal PDF. Therefore, the standard normal is not a good proposal distribution if the target PDF is standard Cauchy. More generally, if the likelihood ratio is unbounded, the weights $\frac{f_X(x_i)}{p_X(x_i)}$ vary excessively, giving too much importance to a few values x_i.

Geweke (1989) proposes a sufficient condition for finite variance of $\hat{\theta}_{IS}$:

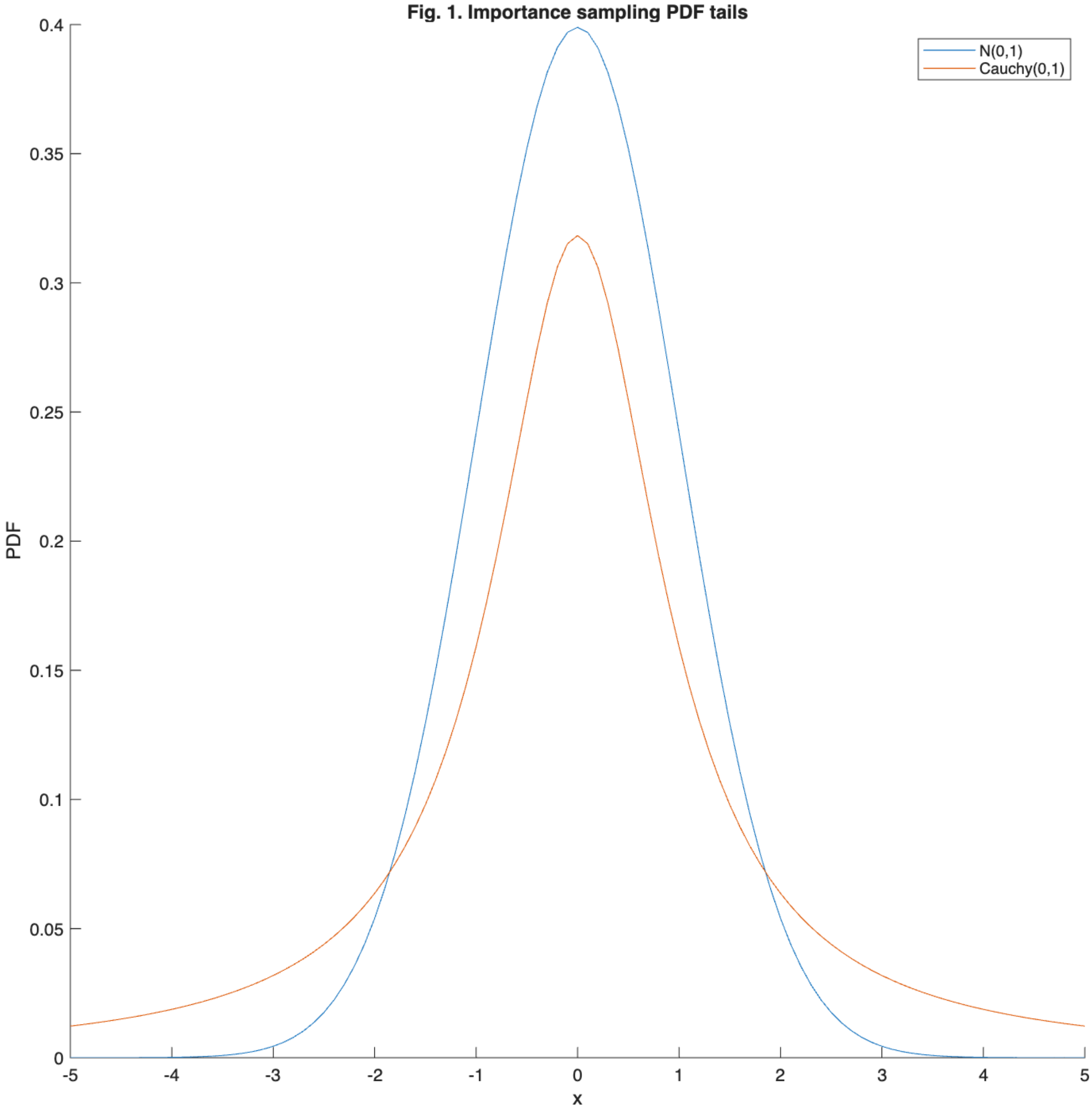
$$\frac{f_X(x)}{p_X(x)} < M \in \mathbb{R},\ \forall x \in \mathbb{R} \text{ and } \mathrm{Var}_{f_X}\left[g(X)\right] < \infty. \tag{17}$$

It is important to note that although the finite variance constraint in equation (16) is not necessary for the convergence of $\hat{\theta}_{IS}$, importance sampling performs poorly when (Robert and Casella, 2010)

$$\mathbb{E}_{f_X}\left[\frac{f_X(x)}{p_X(x)}\right] = \int_{\mathbb{R}} \frac{f_X(x)}{p_X(x)} \cdot f_X(x)\, dx = +\infty. \tag{18}$$

Poor performance refers both to the estimator's behaviour and to its comparison with the plain Monte Carlo estimator. Equation (18) thus provides a diagnostic test for possible poor performance. However, the sufficient condition in equation (17) prevents the scenario described in equation (18) from occurring.

Fig. 1. Importance sampling PDF tails



However, there are some difficulties. Condition IS.5 requires prior knowledge of the shape of f_X, and condition IS.4 is quite restrictive. Fortunately, there exists an alternative to the importance sampling estimator $\hat{\theta}_{IS}$ in Theorem 1. This alternative addresses the finite variance issue and generally results in a more stable estimator. It is called the auto-normalizing or weighted importance sampling estimator:

$$\hat{\theta}_{IS,w} := \frac{\sum_{i=1}^{n} g(x_i) \cdot w(x_i)}{\sum_{i=1}^{n} w(x_i)},$$

with

$$w(x_i) := \frac{\tilde{f}_X(x_i)}{\tilde{p}_X(x_i)} = \frac{\frac{1}{c_f} f_X(x_i)}{\frac{1}{c_p} p_X(x_i)}.$$

The new likelihood ratios $w(x_i)$ are interpreted as weights of the random sample $\{X_i\}_{i=1}^{n}$. In contrast to Monte Carlo integration, where all samples contribute equally, samples here are assigned different weights. While the raw sum of likelihood ratios $\sum_{i=1}^{n} w(x_i)$ is generally not equal to one, the normalized weights

$$\frac{w(x_i)}{\sum_{i=1}^{n} w(x_i)}$$

do sum to one, ensuring that the estimator is properly scaled.

A key advantage of this estimator is that it can be used even when f_X and p_X are only known up to normalizing constants $c_f, c_p \in \mathbb{R}$. This makes it particularly powerful when dealing with challenging integrands ϕ or in Bayesian applications, where target distributions are often specified only up to a constant.

Auto-normalizing importance sampling produces an estimator of θ_{IS} with finite variance but introduces a slight bias. However, this bias converges to zero as the sample size n increases, so the estimator remains consistent while offering improved stability in practice.

Theorem 4. *Let $X \sim p_X(x)$ and let $X_1, \ldots, X_n \sim p_X(x)$ be an i.i.d. random sample with $\mathbb{E}_{p_X}[|X_i|] < \infty$ for all $i \in \{1, \ldots, n\}$. Define the auto-normalizing importance sampling estimator as*

$$\hat{\theta}_{IS,w} := \frac{\sum_{i=1}^{n} g(x_i)\, w(x_i)}{\sum_{i=1}^{n} w(x_i)},$$

where x_i denotes the realization of X_i and $w(x) = \frac{f_X(x)}{p_X(x)}$ (or its normalized version if f_X and p_X are only known up to constants). Then, as $n \to \infty$,

$$\hat{\theta}_{IS,w} \xrightarrow{a.s.} \frac{\mathbb{E}_{p_X}[g(X)\, w(X)]}{\mathbb{E}_{p_X}[w(X)]}.$$

Moreover, since

$$\frac{\mathbb{E}_{p_X}[g(X)\, w(X)]}{\mathbb{E}_{p_X}[w(X)]} = \frac{\int_{\mathbb{R}} g(x)\, \frac{f_X(x)}{p_X(x)}\, p_X(x)\, dx}{\int_{\mathbb{R}} \frac{f_X(x)}{p_X(x)}\, p_X(x)\, dx} = \frac{\int_{\mathbb{R}} g(x)\, f_X(x)\, dx}{\int_{\mathbb{R}} f_X(x)\, dx} = \theta_{IS},$$

the estimator is consistent for θ_{IS}.

Proof. By the Strong Law of Large Numbers (SLLN), both the numerator and denominator of $\hat{\theta}_{IS,w}$ converge almost surely to their expectations:

$$\frac{1}{n} \sum_{i=1}^{n} g(x_i)\, w(x_i) \xrightarrow{a.s.} \mathbb{E}_{p_X}[g(X)\, w(X)]$$

and

$$\frac{1}{n}\sum_{i=1}^{n} w(x_i) \xrightarrow{a.s.} \mathbb{E}_{p_X}[w(X)].$$

Therefore,

$$\hat{\theta}_{IS,w} = \frac{\frac{1}{n}\sum_{i=1}^{n} g(x_i)\, w(x_i)}{\frac{1}{n}\sum_{i=1}^{n} w(x_i)} \xrightarrow{a.s.} \frac{\mathbb{E}_{p_X}[g(X)\, w(X)]}{\mathbb{E}_{p_X}[w(X)]}.$$

Expanding the expectations gives

$$\frac{\mathbb{E}_{p_X}[g(X)\, w(X)]}{\mathbb{E}_{p_X}[w(X)]} = \frac{\int_{\mathbb{R}} g(x)\, w(x)\, p_X(x)\, dx}{\int_{\mathbb{R}} w(x)\, p_X(x)\, dx}.$$

Now substitute $w(x) = \frac{c_f}{c_p}\frac{\tilde{f}_X(x)}{\tilde{p}_X(x)}$:

$$= \frac{\frac{c_f}{c_p}\int_{\mathbb{R}} g(x)\, \frac{\tilde{f}_X(x)}{\tilde{p}_X(x)}\, p_X(x)\, dx}{\frac{c_f}{c_p}\int_{\mathbb{R}} \frac{\tilde{f}_X(x)}{\tilde{p}_X(x)}\, p_X(x)\, dx}.$$

The constants $\frac{c_f}{c_p}$ cancel, leaving

$$= \frac{\int_{\mathbb{R}} g(x)\, \frac{f_X(x)}{p_X(x)}\, p_X(x)\, dx}{\int_{\mathbb{R}} f_X(x)\, dx}.$$

Finally, since $\int_{\mathbb{R}} f_X(x)\, dx = 1$, we obtain

$$= \int_{\mathbb{R}} g(x)\, f_X(x)\, dx = \theta_{IS}.$$

■

Theorem 5. *The auto-normalizing importance sampling estimator $\hat{\theta}_{IS,w}$ is a biased estimator of θ_{IS}, but the bias vanishes as $n \to \infty$.*

Proof. Liu (2001) provides a useful approximation of the expectation of $\hat{\theta}_{IS,w}$:

$$\mathbb{E}_{p_X}\left[\hat{\theta}_{IS,w}\right] \approx \theta_{IS} - \frac{1}{n}\mathrm{Cov}_{p_X}[w(X),\, g(X)\, w(X)] + \frac{1}{n}\theta_{IS}\mathrm{Var}_{p_X}[w(X)]\,. \tag{19}$$

The terms following θ_{IS} represent the approximate bias of the estimator. Since they are scaled by $\frac{1}{n}$, the bias converges to zero as $n \to \infty$. Hence, although $\hat{\theta}_{IS,w}$ is biased for finite n, it is asymptotically unbiased and consistent. ■

Agapiou et al. (2017) show that the bias and MSE of $\hat{\theta}_{IS,w}$ decrease at rate $\mathcal{O}(n^{-1})$, which is faster than the rate $\mathcal{O}(n^{-1/2})$ at which the standard deviation of the plain Monte Carlo estimator decreases.

In practice, a notion of sample variance is needed in order to assess the variability of $\hat{\theta}_{IS,w}$.

Theorem 6. *The sample variance of the auto-normalizing importance sampling estimator $\hat{\theta}_{IS,w}$ is*

$$\begin{aligned} s^2_{IS,w} := \widehat{\mathrm{Var}_{p_X}\left[\hat{\theta}_{IS,w}\right]} &= \frac{n}{n-1} \cdot \frac{\sum_{i=1}^{n} w^2(x_i) \left[g(x_i) - \hat{\theta}_{IS,w}\right]^2}{\left(\sum_{i=1}^{n} w(x_i)\right)^2} \\ &= \frac{n}{n-1} \cdot \sum_{i=1}^{n} w_i^2 \left[g(x_i) - \hat{\theta}_{IS,w}\right]^2, \end{aligned} \tag{20}$$

where $w_i := \frac{w(x_i)}{\sum_{j=1}^{n} w(x_j)}$ are the normalized weights.

Proof. Cochran (1977) provides a general technique for determining the sample variance of a ratio of weighted arithmetic means. Applying this directly to the auto-normalizing estimator yields the stated formula. ■

Note that the normalized weights w_i can be interpreted as probabilities, since they sum to one. This highlights the close analogy with the sample variance of the regular importance sampling estimator in equation (9), where the role of probabilities is played by the normalized weights.

While our focus remains on standard integration problems, it is worth noting that importance sampling extends far beyond this setting. The methods discussed here do not encompass the entirety of importance sampling. More advanced techniques include sequential importance sampling, exponential tilting, and sampling importance resampling. In addition, the concepts introduced in this section can be generalized to random vectors, in direct analogy with the extension of Monte Carlo integration to multi-dimensional problems. Nevertheless, in what follows we restrict attention to the application of importance sampling in standard integration problems. For further reading on these advanced techniques, see Agapiou et al. (2017) and Robert and Casella (2004).

Up to this point, our discussion of importance sampling has been theoretical, focusing on the properties of the estimators, their bias, and variance. For practical applications, however, it is useful to summarize the procedure in algorithmic form. The following pseudo-code outlines the steps required to implement the standard importance sampling estimator in practice.

The algorithm for calculating the importance sampling estimator is:

Algorithm: Pseudo-code for importance sampling

```
Input: Sample size $n$
for $i = 1, \ldots, n$
    Sample $X_i \sim p_X$
    Compute $g(X_i)$
    Compute likelihood ratio $W(X_i) = \frac{f_X(X_i)}{p_X(X_i)}$
end
Compute the estimator $\hat{\theta}_{IS} = \frac{1}{n} \sum_{i=1}^{n} g(X_i) \cdot W(X_i)$
```

In many practical situations, the target and proposal densities are only known up to normalizing constants. In such cases, the weighted (auto-normalizing) importance sampling estimator provides a more robust alternative. The following pseudo-code summarizes its implementation. The algorithm for weighted importance sampling is:

Algorithm: Pseudo-code for weighted importance sampling

```
Input: Sample size $n$
for $i = 1, \ldots, n$
    Sample $X_i \sim \tilde{p}$
    Compute $g(X_i)$
    Compute the weight $w(X_i) = \frac{\tilde{f}_X(X_i)}{\tilde{p}(X_i)}$
    Normalize: $w_i = \frac{w(X_i)}{\sum_{j=1}^{n} w(X_j)}$
end
Compute the estimator $\hat{\theta}_{IS,w} = \frac{1}{n} \sum_{i=1}^{n} w_i \cdot g(X_i)$
```

These algorithms highlight the practical steps required to implement importance sampling in code. The first algorithm corresponds to the standard estimator, while the second incorporates normalized weights to handle unnormalized densities and improve stability. Together, they provide a bridge between the theoretical properties discussed earlier and their application in simulation studies or Bayesian computation.

The theoretical results presented in this section are explored empirically in two exercises. These exercises are designed to illustrate how estimator performance varies with different choices of proposal distributions, and to demonstrate the practical implications of conditions such as IS.4 and IS.5. In particular, they highlight how tail behavior and bounded likelihood ratios affect the stability and efficiency of importance sampling estimators in practice.

3. Final notes

This file is prepared and copyrighted by Jelmer Wieringa and Tunga Kantarcı. This file and the accompanying MATLAB file are available on GitHub and can be accessed using this link. We have used the following references: 1. Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D. and Stuart, A.M., 2017. Importance sampling: intrinsic dimension and computational cost. Statistical Science, 32 (3), 405-431. 2. Brandimarte, P., 2014. Handbook in Monte Carlo simulation: applications in financial engineering, risk management, and economics. Hoboken, New Jersey: John Wiley & Sons. 3. Cochran, W.G., 1977. Sampling techniques. New York and London: Wiley. 4. Durrett, R. T. (Ed.), 2019. Probability: Theory and examples. Cambridge University Press: Cambridge, New York. 5. Geweke, J., 1989. Bayesian inference in econometric models using Monte Carlo integration. Econometrica, 57 (6), 1317-1339. 6. Kloek, T., van Dijk, H.K., 1978. Bayesian estimates of equation system parameters: An application of integration by Monte Carlo. Econometrica, 46 (1), 1-19. 7. Liu, J.S., 2001. Monte Carlo strategies in scientific computing. New York: Springer. 8. Metropolis, N. C., Ulam, S. M., 1949, The MC Method. Journal of the American Statistical Association , Vol. 44, No. 247, p. 335-341. 9. Robert, C. P., Casella, G. (Eds.), 2010. Introducing MC Methods with R. Springer: New York, London. 10. Robbins, H., 1948. On the asymptotic distribution of the sum of a random number of random variables, Proc. Natl. Acad. Sci. U.S.A. 34 (4) 162-163. 11. Rubinstein, R.Y., 2016. Simulation and the Monte Carlo method. New York: John Wiley & Sons. 12. Train, K.E., 2009. Discrete choice methods with simulation. Cambridge: Cambridge

University Press.