Proposition: Assume that $I = \mathbb{E}(f(X))$ exists. Then $\hat{I}_n$ defined above is an unbiased, strongly consistent estimator for I.

Proof: Using linearity of the expectation,

$$\mathbb{E}(\hat{I}_n) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(f(X_i))$$

$$= \frac{1}{n}\, n\,\mathbb{E}(f(X)) = \mathbb{E}(f(X)) = I.$$

So $\hat{I}_n$ is unbiased. Since $\mathbb{E}(f(X))$ exists we can apply the strong LLN to the r.v. $Z_i := f(X_i)$. This proves that $\hat{I}_n$ is strongly consistent. ∎

- This guarantees convergence of $\hat{I}_n$ as $n \to \infty$. In practice, we can only compute $\hat{I}_n$ for finite (but large) n.
  ⇒ need to have information on fluctuations of $\hat{I}_n$ around I for large n.

Proposition: Assume that $\mathbb{E}(f(X))$ and $\sigma^2 = \mathrm{Var}(f(X))$ exist. Then $\mathrm{Var}(\hat{I}_n) = \sigma^2/n$ and

$$\frac{\hat{I}_n - I}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as } n \to \infty.$$

In particular, $\hat{I}_n$ is asymptotically normal.

Proof: Since X_i, $i = 1, \dots, n$ are iid.,

$$\mathrm{Var}(\hat{I}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(f(X_i)) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Since $\mathbb{E}(\hat{I}_n) = I$, we can apply the CLT with $\mu = I$ to prove that $\hat{I}_n$ is asymptotically normal. ∎

We can use the fact that $\hat{I}_n$ is asymptotically normal to quantify how good an estimate $\hat{I}_n$ is of I.

In order to do that, we apply Chebychev's inequality to find:

$$\mathbb{P}\left(|\hat{I}_n - I| > a\frac{\sigma}{\sqrt{n}}\right) \leq \frac{1}{a^2}.$$

Chebychev's inequality:
X : $\mathbb{E}(X)=\mu$, $\mathrm{Var}(X)=\sigma^2$
$\mathbb{P}(|X-\mu| \geq k\sigma) \leq 1/k^2$

This bound is rigorous and holds uniformly for n. However, it is very coarse and we cannot use it in practice. Fortunately, since $\hat{I}_n$ is asymptotically normal, we know that, as $n \to \infty$,

$$\frac{\hat{I}_n - I}{\sigma/\sqrt{n}} \approx \mathcal{N}(0,1).$$ this implies that

$$\mathbb{P}\left(|\hat{I}_n - I| > a\frac{\sigma}{\sqrt{n}}\right) \approx 2(1-\Phi(a)),$$

where $\Phi(\cdot)$ is the CDF of a standard Gaussian distribution. Using this, we can construct confidence intervals for I: for a $(1-\alpha)100\%$ confidence interval, we choose $c = c_\alpha$ such that $2(1-\Phi(c_\alpha)) = \alpha$. Therefore, we obtain that

$$I \in \left(\hat{I}_n - c_\alpha\frac{\sigma}{\sqrt{n}},\ \hat{I}_n + c_\alpha\frac{\sigma}{\sqrt{n}}\right) \quad (*)$$

with $(1-\alpha)100\%$ confidence.

Note that to use $(*)$ we need to know σ, which we do not know in general. Instead, we can use an estimator $\hat{\sigma}_n(f)$:

$$\hat{\sigma}_n(f) = \frac{1}{n-1}\sum_{i=1}^{n}(f(x_i) - \hat{I}_n)^2$$

Using this, the confidence interval is

$$\left(\hat{I}_n - c_\alpha\frac{\hat{\sigma}_n}{\sqrt{n}},\ \hat{I}_n + c_\alpha\frac{\hat{\sigma}_n}{\sqrt{n}}\right).$$

From the previous inequality / confidence interval, we can also obtain

$$\mathbb{P}(|I_n - I| \leq \varepsilon) = \mathbb{P}\left(|I_n - I| \leq \frac{\varepsilon}{\sigma/\sqrt{n}} \frac{\sigma}{\sqrt{n}}\right)$$

$$\geq 1 - \frac{\hat{\sigma}_n^2}{n\varepsilon^2}.$$

So, for the confidence interval described above, we need

$$\boxed{n > \sigma_n^2 / \alpha \varepsilon^2}$$

this also means that the error scales like $1/\sqrt{n}$.!

$\Rightarrow$ the error is independent of dimension!

However, it also depends on the variance $\hat{\sigma}_n$ of the MC estimator, and this quantity can depend on the dimension, sometimes very badly.

this motivates the introduction of variance reduction techniques.

Extension of the central limit theorem

Roughly speaking, Bikelis' theorem provides an upper bound on the distance between the CDF of partial sums (when these are normalized by their standard deviation) and that of the asymptotic normal density in the central limit theorem.

Theorem 1 (Bikelis)**.** *Assume that $\{Z_i\}_{i\in\mathbb{N}}$ is a sequence of real-valued independent random variables (not necessarily identically distributed), such that $\mathbb{E}(Z_i) = 0$ for all i and that there exists $\gamma \in (0,1]$ such that $\mathbb{E}(|Z_i|^{2+\gamma}) < \infty$ for all i. Then there exists a universal constant $A \in [1/\sqrt{2\pi}, 1)$ such that:*

$$\forall x \in \mathbb{R}, \qquad |\Phi_n(x) - \Phi(x)| \leq \frac{A}{B_n^{1+\gamma/2}\,(1+|x|)^{2+\gamma}} \sum_{i=1}^{N} \mathbb{E}(|Z_i|^{2+\gamma}).$$

Here $\Phi(x)$ is the CDF of $\mathcal{N}(0,1)$ and

$$B_n = \sum_{i=1}^{N} \mathrm{var}(Z_i), \qquad \Phi_n(x) = \mathbb{P}\left(\frac{1}{\sqrt{B_n}} \sum_{i=1}^{N} Z_i \leq x\right).$$

Remark 1. If the random variables $\{Z_i\}_{i\in\mathbb{N}}$ are identically distributed with variance σ^2, the main statement in Bikelis' theorem takes a simpler form: under the assumptions,

$$\forall x \in \mathbb{R}, \qquad |\Phi_n(x) - \Phi(x)| \leq \frac{A\,n}{(\sqrt{n}\,\sigma)^{2+\gamma}\,(1+|x|)^{2+\gamma}}\,\mathbb{E}(|Z_1|^{2+\gamma}).$$

In addition, since the exact value of the universal constant A is unknown, in practice this inequality is used with $A = 1$. ⊘

Using Bikelis' theorem, it is possible to construct an accurate confidence interval for a Monte Carlo estimator without relying on an approximation. To this end, let us assume that $\mathbb{E}(|f(X_i) - I|^3) = \xi < \infty$. Applying Bikelis' theorem with $\gamma = 1$ and $Z_i = f(X_i) - I$, and denoting by Φ_n the CDF of $\frac{\sqrt{N}}{\sigma}(\hat{I}_n - I)$, we deduce

$$\mathbb{P}\left(\left|\frac{\hat{I}_n - I}{\sigma/\sqrt{n}}\right| \leq a\right) = \Phi_n(a) - \Phi_n(-a) \geq \Phi(a) - \Phi(-a) - \frac{2\,\xi}{\sqrt{n}\,\sigma^3\,(1+|a|)^3} =: \Psi_n(a).$$

Regardless of the value of n, the right-hand side of this equation is a continuous, strictly increasing function of a, taking a negative value at $a = 0$ and converging to 1 as $a \to \infty$. This implies, in particular, that for any $\alpha \in (0,1)$ there exists $c_\alpha^n = \Psi_n^{-1}(1-\alpha)$ and it holds with probability at least $(1-\alpha)$ that

$$I \in \left(\hat{I}_n - c_\alpha^n \frac{\sigma}{\sqrt{n}}, \hat{I}_n + c_\alpha^n \frac{\sigma}{\sqrt{n}}\right).$$

It is a simple exercise to check that $c_\alpha^n \to c_\alpha$ as $n \to \infty$, where c_α denotes the half-width, up to the factor $\sigma/\sqrt{n}$, of the confidence interval calculated via the central limit theorem.

2.4. Variance Reduction Techniques.

From the confidence interval,

$$I \in (\hat{I}_n - c_\alpha \sigma/\sqrt{n},\ \hat{I}_n + c_\alpha \sigma/\sqrt{n}),$$

it is clear that the number of samples n required to approximate I with a given tolerance ε depends strongly on σ : $\boxed{n > \sigma^2/\alpha\varepsilon^2}$.

More generally, we can measure the accuracy of the estimator $\hat{I}_n$ using the mean square error (MSE). Given an estimator $\hat{\theta}_n$ for θ, we compute

$$\boxed{MSE(\hat{\theta}_n) = \mathbb{E}((\hat{\theta}_n - \theta)^2)}.$$

We can decompose the MSE:

$$\begin{aligned}\mathbb{E}((\hat{\theta}_n - \theta)^2) &= \mathbb{E}((\hat{\theta}_n - \mathbb{E}\hat{\theta}_n + \mathbb{E}\hat{\theta}_n - \theta)^2) \\ &= (\mathbb{E}\hat{\theta}_n - \theta)^2 + \mathbb{E}(\hat{\theta}_n - \mathbb{E}\hat{\theta}_n)^2 \\ &= B_n^2 + V_n\end{aligned}$$

where $B_n = \mathbb{E}\hat{\theta}_n - \theta$ is the bias of the estimator and $V_n = Var(\hat{\theta}_n)$ is its variance.

Since the MC estimator $\hat{I}_n$ is unbiased, its MSE is

$$\boxed{MSE(\hat{I}_n) = Var(\hat{I}_n) = \sigma^2/n}$$

which is consistent with the error obtained using the CLT.

⇒ The performance of the MC estimator depends strongly on its variance and in some situations this can be huge ⇒ need a probibitely large number of samples. In order to avoid this, we introduce variance reduction techniques, which modify $\hat{I}_n$ to reduce its variance.

2.4.1. Antithetic Variables

The key idea of variance reduction techniques is to generate aditional variables which will improve the MC estimator $\hat{I}_n$.

With antithetic variables, we make use of the fact that if X_1 and X_2 are i.d. r.v.s with mean μ, then $\mathbb{E}\left(\frac{X_1+X_2}{2}\right) = \mu$, but

$$\mathrm{Var}\left(\frac{X_1+X_2}{2}\right) = \frac{1}{4}\left(\mathrm{Var}\, X_1 + \mathrm{Var}\, X_2 + 2\,\mathrm{Cov}(X_1, X_2)\right).$$

$\Rightarrow$ the variance is reduced as long as $\mathrm{Cov}(X_1, X_2) \leq 0$.

examples:

- If $X_1 \sim U(0,1)$, then $X_2 = 1 - X_1$ is an antithetic variable of X_1. Moreover, $\mathrm{Var}\left(\frac{X_1+X_2}{2}\right) = 0$!
- Similarly, if $X_1 \sim N(\mu, \sigma^2)$, then $X_2 = 2\mu - X_1$ is an antithetic variable of X_1.
- In general, when estimating $\int_0^1 g(x)\,dx$, one can use $X_1 = g(U)$, $X_2 = g(1-U)$ where $U \sim U(0,1)$. If $g(x)$ is monotone, we can guarantee that $\mathrm{Cov}(X_1, X_2) \leq 0$.

Note that:

- in general, cannot obtain perfectly negatively correlated antithetic variables.
- the variance reduction obtained with antithetic variables is not dramatic. (best case 50%)

Numerical example: Let $X = (X_1, X_2)$, $X \sim N(\mu, \Sigma)$, $\mu = (\mu_1, \mu_2)$, Σ covariance matrix.

We will estimate the expectation of

$$Y_j = e^{X_j}, \quad j = 1, 2.$$

$\rightarrow$ Example codes from Huynh et al.

2.4.2. Control Variates

We saw with antithetic variables that we can exploit negative correlation to reduce variance. We will see here how to use it for variables which are not antithetic. Suppose we wish to compute $I = \mathbb{E}(Z)$, with $Z = f(X)$ and suppose we can find a r.v. W (control) with known expectation $\mathbb{E}(W)$. Then, for some $\alpha \in \mathbb{R}$ define

$$Y = Z + \alpha(W - \mathbb{E}(W)).$$

Clearly, $\mathbb{E}(Y) = \mathbb{E}(Z + \alpha(W - \mathbb{E}(W))) = \mathbb{E}(Z)$.

However,

$$\boxed{\text{Var } Y = \text{Var } Z + \alpha^2 \text{Var } W + 2\alpha \text{Cov}(Z, W).}$$

We can choose α to minimise $\text{Var}(Y)$:

$$\alpha = -\frac{\text{Cov}(Z, W)}{\text{Var } W}$$

which gives

$$\boxed{\text{Var } Y = \text{Var } Z - \frac{\text{Cov}(Z, W)^2}{\text{Var } W} < \text{Var } Z!}$$

$\Rightarrow$ No matter how we choose W, we always reduce the variance!

In practice, the optimal constant α is not computable. We instead approximate it using the estimator

$$\hat{\alpha} = \frac{\hat{C}_{Z,W}}{\hat{C}_{W,W}}, \text{ where}$$

$$\hat{C}_{Z,W} = \frac{1}{N-1} \sum_{n=1}^{N} (Z_n - \hat{I}_{N,Z})(W_n - \hat{I}_{N,W})$$

$$\hat{C}_{W,W} = \frac{1}{N-1} \sum_{n=1}^{N} (W_n - \hat{I}_{N,W})^2$$

Z_n, W_n, $n = 1, \ldots, N$ samples of Z, W respectively,

$$\hat{I}_{N,Z} = \frac{1}{N} \sum_{n=1}^{N} Z_n \qquad \hat{I}_{N,W} = \frac{1}{N} \sum_{n=1}^{N} W_n.$$

Note that: Since $Z = f(X)$, we can also use $W = g(X)$ with g close to f.

Note that: if we define $\rho = \text{Corr}(Z,W) = \frac{\text{Cov}(Z,W)}{\sqrt{\text{Var}\,Z\,\text{Var}\,W}}$,

then $\text{Var}\,Y = \text{Var}\,Z\,\frac{1-\rho^2}{N}$. Therefore we want to choose a control variate W such that $|\rho|$ is as close to 1 as possible.

We can extend this to multiple control variates:

$$Y = Z + \alpha_1(W_1 - \mathbb{E}W_1) + \dots + \alpha_K(W_K - \mathbb{E}W_K).$$
$$= Z + \underline{\alpha}^T(\underline{W} - \underline{m}_W)$$

where $\underline{\alpha} = (\alpha_1,\dots,\alpha_K)^T$, $\underline{W} = (W_1,\dots,W_K)^T$, $\underline{m}_W = (\mathbb{E}W_1,\dots,\mathbb{E}W_K)^T$

We now have

$$\text{Var}\,Y = \text{Var}\,Z + 2\sum_{i=1}^{K}\alpha_i\,\text{Cov}(Z,W_i) + \sum_{i=1}^{K}\sum_{j=1}^{K}\alpha_i\alpha_j\,\text{Cov}(W_i,W_j)$$

and the optimal $\underline{\alpha}$ is given by the solution to

$M\underline{\alpha} = F$, where $M_{ij} = \text{Cov}(W_i,W_j)$ and $F_i = \text{Cov}(Z,W_i)$.

example 1: estimate $I = \int_0^1 e^x\,dx = e - 1 \approx 1.71828$

- Using simple MC: $\hat{I}_N = \frac{1}{N}\sum_{i=1}^{N} e^{U_i}$, $U_i \sim U(0,1)$

$$\text{Var}(e^U) = \underbrace{\mathbb{E}(e^{2U})}_{\int_0^1 e^{2x}dx} - \underbrace{(\mathbb{E}(e^U))^2}_{\left(\int_0^1 e^x dx\right)^2} \approx 0.242$$

$\Rightarrow \text{Var}(\hat{I}_N) = 0.242/N.$

- Use control variate:

Take $W \sim U(0,1) \Rightarrow \mathbb{E}W = 1/2$, $\mathrm{Var}\, W = 1/12$.

We can calculate

$$\mathrm{Cov}(e^U, W) = \int_0^1 x e^x dx - \int_0^1 x\, dx \int_0^1 e^x dx$$

$$= x e^x \Big|_{x=0}^{x=1} - \int_0^1 e^x dx - \frac{x^2}{2}\Big|_{x=0}^{x=1} e^x \Big|_{x=0}^{x=1}$$

$$= e - (e-1) - \frac{1}{2}(e-1) \approx 0.1411$$

$$\Rightarrow \mathrm{Var}\,(Z + \alpha(W - \mathbb{E}W)) = \mathrm{Var}\, Z - \frac{\mathrm{Cov}(Z,W)^2}{\mathrm{Var}\, W} \approx 0.0039$$

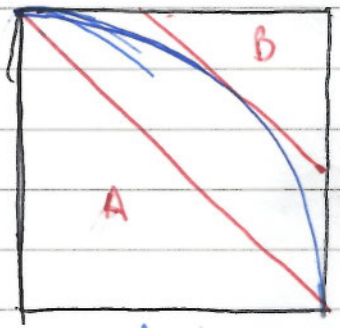
$\Rightarrow$ achieve variance reduction by a factor of 10!

<u>example 2</u>: Consider again the estimator for $\pi/4$ we computed last week: $Z = \frac{\mathbb{1}_C(X)}{\mathbb{1}_B(X)}$, where $X = (X_1, X_2)$, $X_i \sim U(-1,1)$ and $C = \{(x,y) : x^2 + y^2 \le 1\}$, $B = \{(x,y) \in (-1,1) \times (-1,1)\}$.

A similar estimator would be $Z' = 4\,\mathbb{1}\{U_1^2 + U_2^2 \le 1\}$.

We can check that

$W_1 = \mathbb{1}\{U_1 + U_2 \le 1\} = \mathbb{1}_A$ has mean $1/2$ and is positively correlated with Z.



We can check that $(1 - \rho^2) \approx 0.727$ which gives modest variance reduction.

Similarly, $W = \mathbb{1}\{U_1 + U_2 \le \sqrt{2}\} = \mathbb{1}_B$ has mean $(2-\sqrt{2})^2/2$ and is negatively correlated with Z. Here, $(1 - \rho^2) \approx 0.242$.

2.4.3. Importance Sampling

Suppose now that we want to compute, as usual,

$$I = \mathbb{E}(f(X)) = \int f(x)\,\pi(x)\,dx,$$

but where f now is nearly zero outside some region A such that $\mathbb{P}(A)$ is small. For example,

$$f(x,y) = 0.5\exp\left(-90(x-0.5)^2 - 45(y+0.1)^4\right) + \exp\left(-45(x+0.4)^2 - 60(y-0.5)^2\right)$$

for $(x,y) \in [-1,1] \times [-1,1]$ (see Liu, Monte Carlo Strategies in Scientific computing for figure).

→ in this case, with a regular grid, more than two thirds of computing time are wasted on evaluating grid points where $f(x) \approx 0$.

→ this is very common, especially in higher dimensions.

Note that: in cases like this, a simple Monte Carlo estimator using a uniform distribution also performs poorly!

The idea of importance sampling is to modify the sampling distribution π so that most of the sampling is done in the part of state space that contributes the most to $f(x)$ (= the "most important" part!).

This can improve performance drastically!
However, when performed poorly, it can yield estimators which have infinite variance

Note that: importance sampling is not just a variance reduction technique, it is a new sampling scheme!

How it works:

I will present the idea on $\mathbb{R}$, but note that it holds similarly for more general domains.

Again, we wish to compute $I = \int f(x)\,\pi(x)\,dx$.

Suppose that $\psi(x)$ is another probability density function on $\mathbb{R}$. We can rewrite

$$I = \int f(x)\,\pi(x)\,dx = \int f(x)\frac{\pi(x)}{\psi(x)}\,\psi(x)\,dx$$

$$= \mathbb{E}\left(f(Y)\,g(Y)\right),$$

where now $Y \sim \psi(\cdot)$ and $g(x) = \pi(x)/\psi(x)$ is a weight, which compensates for the fact that we are using a different sampling. $g(x)$ is known as likelihood ratio. ψ is the importance distribution, and π is the nominal distribution.

Definition: The importance sampling estimator for $I = \mathbb{E}f(X)$ is

$$\hat{I}_n^{is} = \frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)\,\pi(X_i)}{\psi(X_i)}, \quad X_i \sim \psi.$$

Proposition: If $\psi(x) > 0$ whenever $f(x)\,\pi(x) \neq 0$, then $\mathbb{E}(\hat{I}_n^{is}) = I$ (importance sampling estimator is unbiased) and $\mathrm{Var}(\hat{I}_n^{is}) = \sigma_\psi^2/n$, where

$$\sigma_\psi^2 = \int \frac{(f(x)\,\pi(x))^2}{\psi(x)}\,dx - I^2 = \int \frac{(f(x)\,\pi(x) - I\,\psi(x))^2}{\psi(x)}\,dx.$$

The proof is left as an exercise.

Note that for a poor choice of $\psi(x)$ we can end up with infinite variance!

We can also see that we reduce variance if the numerator, $f(x)\,\pi(x) - I\,\psi(x)$ is close to zero.

We can obtain an optimal ψ:

Proposition: The density ψ^* that minimises σ_ψ is given by

$$\psi^*(x) = \frac{|f(x)|\pi(x)}{\int |f(y)|\pi(y)\,dy}.$$

In particular, if $f \geq 0$ then $\sigma_\psi = 0$.

Proof: σ_ψ is minimised if and only if we minimise $\int \frac{f(x)^2 \pi(x)^2}{\psi(x)}\,dx$. We have

$$\int \frac{f(x)^2 \pi(x)^2}{\psi(x)}\,dx = \int \frac{f(x)^2 \pi(x)^2}{\psi(x)^2}\psi(x)\,dx$$

$$= \mathbb{E}_{Y\sim\psi}\left(\frac{f(Y)^2\pi(Y)^2}{\psi(Y)^2}\right)$$

$$\geq \left(\mathbb{E}_{Y\sim\psi}\left(\frac{|f(Y)|\pi(Y)}{\psi(Y)}\right)\right)^2$$

by Jensen's inequality.

> **Jensen's Inequality**
> If X is a r.v. and φ is convex, then
> $\varphi(\mathbb{E}(X)) \leq \mathbb{E}(\varphi(X))$

Now, since

$$\left(\mathbb{E}_{Y\sim\psi}\left(\frac{|f(Y)|\pi(Y)}{\psi(Y)}\right)\right)^2 = \left(\int |f(x)|\pi(x)\,dx\right)^2$$

we have, for any density ψ,

$$\mathrm{Var}\left(\hat{I}^{is}_{n,\psi}(\psi)\right) \geq \frac{1}{n}\left(\left(\int |f(x)|\pi(x)\,dx\right)^2 - I^2\right).$$

Plugging in the value of ψ^*, this inequality becomes identity $\Rightarrow$ ψ^* is the optimal density. If $f \geq 0$, then $|f| = f$. ∎

Note that in practice, we cannot compute ψ^*, since it requires computing I. However, this provides insight into what ψ should be: it is best to have mass (peaks) where $f\pi$ does. In general, choosing the right ψ requires experience.

example: Exponential Tilting

A common way of generating an importance distribution ψ is to use the moment generating function (MGF) of π:

$$M_\pi(t) = \mathbb{E}(e^{tX}) \qquad X \sim \pi.$$

We then consider the tilted density of π:

$$\psi(x) = \frac{\pi(x)\, e^{tx}}{M_\pi(t)}, \qquad -\infty < t < \infty$$

(assuming it exists)

If we want to sample more often from a region where X is large, we use a tilted density with $t > 0$.

If $X \sim N(\mu, \sigma^2)$, we can complete squares to obtain ψ:

$$\psi(x) \propto e^{-(x-\mu)^2/2\sigma^2} e^{tx} = e^{-(x-\mu-t\sigma^2)^2/2\sigma^2} e^{\mu t + t^2\sigma^2/2}$$

$$\Rightarrow \psi(x) = N(\mu + t\sigma^2, \sigma^2), \quad M_\pi = e^{\mu t + t^2\sigma^2/2}.$$

$\Rightarrow$ We can generate samples from ψ because it is Gaussian. In fact, this is true for any $\pi(x)$ within the exponential family.

We have $g(x) = \pi(x)/\psi(x) = e^{-tx} M_\pi(t) = e^{-t(x-\mu-t\sigma^2/2)}$

Algorithm: Exponentially tilted importance sampler for I:

1. Generate samples $y_i \sim N(\mu + t\sigma^2, \sigma^2)$ $\quad i = 1, \dots, n$
2. Compute $g_i = \exp(-t(y_i - \mu - t\sigma^2/2))$ $\quad i = 1, \dots, n$
3. Compute $\hat{I}_n = \frac{1}{n} \sum g_i \, \mathbb{1}(y_i > x_0)$

to compute $P(X > x_0)$, $X \sim N(\mu, \sigma^2)$.

To minimise the variance, choose t to minimise

$$\int_{x_0}^{\infty} \pi(x) \exp(-t(x - \mu - t\sigma^2/2))\, dx = M_\pi(t) \int_{x_0}^{\infty} \pi(x)\, e^{-tx}\, dx.$$

example: Sampling from bimodal distributions

In many applications, $\pi(x)$ is multimodal, possessing well-separated modes, or $f(x)\pi(x)$ is only nonzero in multiple distinct regions.

$\Rightarrow$ A natural choice for ψ is

$$\psi_\alpha = \sum_{j=1}^{J} \alpha_j \psi_j, \quad \text{where } \alpha_j \geq 0, \ \sum_{j=1}^{J} \alpha_j = 1$$

and ψ_j are distributions. One wishes to choose α_j, ψ_j to match the peaks of $f\pi$.

To sample from ψ_α, we generate a generalised Bernoulli r.v. (problem sheet) S taking values $i = 1, \dots, J$ with probability $\alpha_1, \dots, \alpha_J$. Then if $S = j$, return a sample from ψ_j. The estimator is

$$\hat{I}_n^{is} = \frac{1}{n} \sum_{i=1}^{n} \frac{f(y_i)\,\pi(y_i)}{\sum_{j=1}^{J} \alpha_j \psi_j(y_i)}.$$

A possible issue is that the estimator may not produce enough samples in some areas, leading to increase in variance.

example: Self-normalising importance sampling

In many applications, it is not possible to compute normalisation constants, i.e., we know π and/or ψ up to a constant: $\int \pi(x)\,dx = Z \neq 1$ or $\int \psi(x)\,dx = Z' \neq 1$.

In this case, we can use an alternative importance sampler: this is based on the following observation

$$\mathbb{E}_{x\sim\pi}(f(x)) = \frac{\mathbb{E}_{y\sim\psi}(f(y)\,g(y))}{\mathbb{E}_{y\sim\psi}(g(y))}.$$

This is because, if $g(x) = \pi(x)/\psi(x)$ (unnormalised),

$$\mathbb{E}_{x\sim\pi}(f(x)) = \frac{\int f(x)\,\pi(x)\,dx}{\int \pi(x)\,dx} = \frac{\int f(x)\,\frac{\pi(x)}{\psi(x)}\,\psi(x)\,dx}{\int \frac{\pi(x)}{\psi(x)}\,\psi(x)\,dx}$$

$$= \frac{\int f(x)\,g(x)\,\psi(x)\,dx}{\int g(x)\,\psi(x)\,dx} = \frac{\mathbb{E}_{y\sim\psi}(f(y)\,g(y))}{\mathbb{E}_{y\sim\psi}(g(y))}.$$

Algorithm: Self-Normalised importance sampler

1. Generate iid samples from ψ, y_i, $i = 1, \dots, n$
2. Compute $g_i = \pi(y_i)/\psi(y_i)$, $i = 1, \dots, n$
3. Generate the estimator

$$\hat{I}_n^{is} = \sum_{i=1}^{n} g_i f(y_i) \Big/ \sum_{i=1}^{n} g_i .$$

Note that $\sum g_i f(y_i)$ and $\sum g_i$ are unbiased and consistent estimators of $\mathbb{E}_{y\sim\psi}(f(y)\,g(y))$ and $\mathbb{E}_{y\sim\psi}(g(y))$. However, their ratio $\hat{I}_n^{is}$ will be biased for finite n.

However, it is possible to prove that if $\pi(x) > 0 \Leftrightarrow \psi(x) > 0$ then $\hat{I}_n^{is}$ is strongly consistent.