Transformation Method: pushforward

Let $\mathcal{Y}, \mathcal{X}$ be two topological spaces equipped with their Borel $\sigma$-algebras.
Suppose that $f : \mathcal{Y} \to \mathcal{X}$ is Borel measurable;
Suppose that $q$ is a Borel probability measure on $\mathcal{Y}$ and let $Y \sim q$.
Write $\pi$ for the distribution of $X = f(Y)$, a Borel probability measure on $\mathcal{X}$.
Then $\pi$ is the push-forward of $q$ under $f$, written

$$\pi = f_* \mu.$$

It’s defined as

$$\pi(B) = (f_*)\mu(B) = q(f^{-1}(B)),$$
for all $B \in \mathcal{B}(\mathcal{X})$.

In terms of expectations

$$\int h \circ \varphi \, dq = \int h df_* \mu.$$
When $dq(x) = q(x)dx$, and $\varphi$ is a bijection, then $\pi$ also has a density given by the change of variables formula

$$\pi(x) = q \circ \varphi^{-1}(x) \left| \det(D\varphi^{-1})(x) \right|.$$
Transformation Method - Box-Muller Algorithm

**Gaussian distribution.** Let $U_1 \sim \mathcal{U}_{[0,1]}$ and $U_2 \sim \mathcal{U}_{[0,1]}$ be independent and set

$$ R = \sqrt{-2 \log(U_1)}, \quad \vartheta = 2\pi U_2. $$

Clearly $R, \vartheta$ independent and $R^2 \sim \text{Exp}(1/2), \ \vartheta \sim \mathcal{U}_{[0,2\pi]}$ with joint density

$$ q(r^2, \vartheta) = \frac{1}{2\pi} \frac{1}{2} \exp(-r^2/2). $$

Set $X = R \cos(\vartheta), \ Y = R \sin(\vartheta)$ a bijection.
By standard facts:

\[ f_{X,Y}(x, y) = f_{R^2, \theta}(r^2(x, y), \theta(x, y)) \left| \det \frac{\partial (r^2, \theta)}{\partial (x, y)} \right| \]

\[ = f_{R^2, \theta}(r^2(x, y), \theta(x, y)) \left| \det \frac{\partial (x, y)}{\partial (r^2, \theta)} \right|^{-1} \]

\[ = \frac{1}{2 \pi} \exp \left[ - \frac{x^2 + y^2}{2} \right] 2 = \frac{1}{2 \pi} \exp \left[ - \frac{x^2 + y^2}{2} \right], \]

since

\[ \left| \det \frac{\partial (x, y)}{\partial (r^2, \theta)} \right| = \left| \begin{array}{cc} \cos(\theta) & -r \sin \theta \\ \frac{2r}{\sin(\theta)} & r \cos \theta \end{array} \right| = \frac{1}{2}. \]

thus \((X, Y)\) are independent standard normal.
Transformation Method - Multivariate Normal

Let $Z = (Z_1, ..., Z_d)^{i.i.d.} \sim \mathcal{N}(0, 1)$.

Let $L$ be a real invertible $d \times d$ matrix satisfying $L L^T = \Sigma$, and $X = LZ + \mu$. Then $X \sim \mathcal{N}(\mu, \Sigma)$.

We have indeed $q(z) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}z^T z\right)$ and

$$\pi(x) = q(z) |\det \frac{\partial z}{\partial x}|$$

where $\frac{\partial z}{\partial x} = L^{-1}$ and $\det (L^{-1}) = \det (\Sigma)^{-1/2}$. Additionally,

$$z^T z = (x - \mu)^T (L^{-1})^T L^{-1} (x - \mu)$$

$$= (x - \mu)^T \Sigma^{-1} (x - \mu).$$

In practice, use a Cholesky factorization $\Sigma = L L^T$ where $L$ is a lower triangular matrix.
Sampling via Composition

Assume we have a joint pdf $\pi$ with marginal $\pi$; i.e.

$$
\pi(x) = \int \pi_{X,Y}(x,y) \, dy
$$

where $\pi(x,y)$ can always be decomposed as

$$
\pi_{X,Y}(x,y) = \pi_Y(y) \pi_{X|Y}(x|y).
$$

It might be easy to sample from $\pi(x,y)$ whereas it is difficult/impossible to compute $\pi(x)$.

In this case, it is sufficient to sample

$$
Y \sim \pi_Y \text{ then } X|Y \sim \pi_{X|Y}(\cdot|Y)
$$

so $(X,Y) \sim \pi_{X,Y}$ and hence $X \sim \pi$.

Latent variable models; HMMs;
Finite Mixture of Distributions

Assume one wants to sample from

\[ \pi(x) = \sum_{i=1}^{p} \alpha_i \cdot \pi_i(x) \]

where \( \alpha_i > 0 \), \( \sum_{i=1}^{p} \alpha_i = 1 \) and \( \pi_i(x) \geq 0 \), \( \int \pi_i(x) \, dx = 1 \).

We can introduce \( Y \in \{1,...,p\} \) and

\[ \pi_{X,Y}(x,y) = \alpha_y \times \pi_y(x). \]

To sample from \( \pi(x) \), first sample \( Y \) from a discrete distribution such that \( \mathbb{P}(Y = k) = \alpha_k \) then

\[ X| (Y = y) \sim \pi_y. \]
Rejection Sampling

**Basic idea:** Sample from instrumental proposal \( q \neq \pi \); correct through rejection step to obtain a sample from \( \pi \).

**Algorithm (Rejection Sampling).** Given two densities \( \pi, q \) with \( \pi(x) \leq M q(x) \) for all \( x \), we can generate a sample from \( \pi \) by

1. Draw \( X \sim q \), draw \( U \sim U_{[0,1]} \).
2. Accept \( X = x \) as a sample from \( \pi \) if

\[
U \leq \frac{\pi(x)}{M q(x)},
\]

otherwise go to step 1.

**Proposition**

*The distribution of the samples accepted by rejection sampling is \( \pi \).*
Rejection Sampling

Proof.

\[ P( X \in A \mid X \text{ accepted}) = \frac{P( X \in A, X \text{ accepted})}{P( X \text{ accepted})} \]

where

\[ P( X \in A, X \text{ accepted}) = \int_X \int_0^1 I_A(x) I\left(u \leq \frac{\pi(x)}{M q(x)}\right) q(x) \, du \, dx \]

\[ = \int_X I_A(x) \frac{\pi(x)}{M q(x)} q(x) \, dx \]

\[ = \int_X I_A(x) \frac{\pi(x)}{M} \, dx = \frac{\pi(A)}{M}. \]

\[ \square \]
Rejection Sampling

• Often we only know $\pi$ and $q$ up to some normalising constants; i.e.

$$
\pi = \tilde{\pi} / Z_\pi \quad \text{and} \quad q = \tilde{q} / Z_q
$$

where $\tilde{\pi}$, $\tilde{q}$ are known but $Z_\pi$, $Z_q$ are unknown.

*You still need to be able to sample from $q(\cdot)$.*

• If you can upper bound:

$$
\frac{\tilde{\pi}(x)}{\tilde{q}(x)} \leq \tilde{M},
$$

then using $\tilde{\pi}$, $\tilde{q}$ and $\tilde{M}$ in the algorithm is correct.

• Indeed we have

$$
\frac{\tilde{\pi}(x)}{\tilde{q}(x)} \leq \tilde{M} \Leftrightarrow \frac{\pi(x)}{q(x)} \leq \tilde{M} \frac{Z_q}{Z_\pi} = M.
$$
Rejection Sampling

Let $T$ denote the number of pairs $(X, U)$ that have to be generated until $X$ is accepted for the first time.

Lemma

\[
T \text{ is geometrically distributed with parameter } 1/M \text{ and in particular } \mathbb{E}(T) = M.
\]

In the unnormalised case, this yields

\[
\mathbb{P}(X \text{ accepted}) = \frac{1}{M} = \frac{Z_\pi}{\tilde{M} Z_q},
\]

\[
\mathbb{E}(T) = M = \frac{Z_q \tilde{M}}{Z_\pi},
\]

and it can be used to provide unbiased estimates of $Z_\pi/Z_q$ and $Z_q/Z_\pi$. 
Examples: Uniform from bounded subset of $\mathbb{R}^p$

- Let $B \subset \mathbb{R}^p$, a bounded subset of $\mathbb{R}^p$:

$$\pi (x) \propto \mathbb{1}_B (x).$$

Let $R$ be a rectangle containing $B \subset R$ and

$$q (x) \propto \mathbb{1}_R (x).$$

- Then we can use $\tilde{M} = 1$ and

$$\tilde{\pi} (x) / \left( \tilde{M}' \tilde{q} (x) \right) = \mathbb{1}_B (x).$$

- The probability of accepting a sample is then $Z_\pi / Z_q$. 

\[11 / 32\]
Example: Normal density

- Let $\tilde{\pi}(x) = \exp\left(-\frac{1}{2}x^2\right)$ and $\tilde{q}(x) = 1/(1 + x^2)$. We have
  \[
  \frac{\tilde{\pi}(x)}{\tilde{q}(x)} = (1 + x^2) \exp\left(-\frac{1}{2}x^2\right) \leq 2/\sqrt{e} = \tilde{M}
  \]
  which is attained at $\pm 1$.

- Let $X \sim \tilde{q}$. The acceptance probability is
  \[
  P\left(U \leq \frac{\tilde{\pi}(X)}{\tilde{M}\tilde{q}(X)} = \frac{Z_\pi}{\tilde{M}Z_q} = \frac{\sqrt{2\pi}}{2\sqrt{e}\pi} = \sqrt{\frac{e}{2\pi}} \approx 0.66,
  \right)
  \]
  and the mean number of trials to success is approximately $1/0.66 \approx 1.52$. 

Examples: Genetic Linkage model

• We observe

\[
(Y_1, Y_2, Y_3, Y_4) \sim \mathcal{M} \left( n; \frac{1}{2} + \frac{\theta}{4}, \frac{1}{4} (1 - \theta), \frac{1}{4} (1 - \theta), \frac{\theta}{4} \right)
\]

where \( \mathcal{M} \) is the multinomial distribution and \( \theta \in (0, 1) \).

• The likelihood of the observations is thus

\[
p(y_1, \ldots, y_4; \theta) = \frac{n!}{y_1! y_2! y_3! y_4!} \left( \frac{1}{2} + \frac{\theta}{4} \right)^{y_1} \left( \frac{1}{4} (1 - \theta) \right)^{y_2 + y_3} \left( \frac{\theta}{4} \right)^{y_4}
\]

\[\propto (2 + \theta)^{y_1} (1 - \theta)^{y_2 + y_3} \theta^{y_4}.
\]

• Bayesian approach where we select \( p(\theta) = \mathbb{1}_{[0,1]}(\theta) \) and are interested in

\[
p(\theta | y_1, \ldots, y_4) \propto (2 + \theta)^{y_1} (1 - \theta)^{y_2 + y_3} \theta^{y_4} \mathbb{1}_{[0,1]}(\theta).
\]
Examples: Genetic linkage model

- Rejection sampling using the prior as proposal $q(\theta) = \tilde{q}(\theta) = p(\theta)$ to sample from $p(\theta | y_1, ..., y_4)$.

- To use accept-reject, we need to upper bound

\[ \frac{\tilde{\pi}(\theta)}{\tilde{q}(\theta)} = \tilde{\pi}(\theta) = (2 + \theta)^{y_1} (1 - \theta)^{y_2 + y_3} \theta^{y_4} \]

- Maximum of $\tilde{\pi}$ can be found using standard optimization procedure to perform rejection sampling.

- For a realisation of $(Y_1, Y_2, Y_3, Y_4)$ equal to $(69, 9, 11, 11)$ obtained with $n = 100$ and $\theta^* = 0.6$, results shown in following figure.
Examples: Genetic linkage model

Figure: Histogram of 10,000 samples drawn from posterior obtained by rejection sampling (left); and histogram of waiting time distribution before acceptance (right).
Rejection Sampling Recap

Rejection sampling requires

- Samples from some distribution \( q \);

- evaluation of \( \pi(\cdot) \) point-wise, or unnormalized \( \tilde{\pi} \);

- an upper bound \( M \) on \( \pi(x)/q(x) \), or \( \tilde{\pi}/q \) and so on.

Sometimes the upper bound is not feasible.
Importance Sampling

- We want to compute

\[ I = \mathbb{E}_\pi(\varphi(X)) = \int_X \varphi(x) \pi(x) \, dx. \]

- We do not know how to sample from the target \( \pi \) but have access to a proposal distribution of density \( q \).

- We only require that \( \pi(x) > 0 \Rightarrow q(x) > 0 \);

  i.e. the support of \( q \) includes the support of \( \pi \).

- \( q \) is called the proposal, or importance distribution.
Importance Sampling

• We have the following identity

\[ I = \mathbb{E}_\pi (\varphi(X)) = \mathbb{E}_q (\varphi(X) w(X)), \]

where \( w : \mathbb{X} \to \mathbb{R}^+ \) is the importance weight function

\[ w(x) = \frac{\pi(x)}{q(x)}. \]

• Hence for \( X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} q, \)

\[ \hat{I}_n^{\text{IS}} = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i) w(X_i). \]
Importance Sampling Properties

Proposition

(a) **Unbiased:** \( \mathbb{E}_q[\hat{I}_n^{IS}] = I; \)

(b) **Strongly consistent:** If \( \mathbb{E}_q(\varphi(X)|w(X)) < \infty \) then

\[
\lim_{n \to \infty} \hat{I}_n^{IS} = I, \quad \text{a.s.}
\]

(c) **CLT:** \( \forall_q(\hat{I}_n^{IS}) = \sigma_{IS}^2/n \) where

\[
\sigma_{IS}^2 := \mathbb{V}_q (\varphi(X)w(X))
\]

If \( \sigma_{IS}^2 < \infty \) then

\[
\lim_{n \to \infty} \sqrt{n} \left( \hat{I}_n^{IS} - I \right) \overset{D}{\to} \mathcal{N} (0, \sigma_{IS}^2).
\]
Importance Sampling: Practical Advice

Consistency does not require $\sigma_{IS}^2 < \infty$ but highly recommended in practice (!).

**Sufficient condition**: If $\mathbb{E}_\pi (\phi^2(X)) < \infty$ and $w(x) \leq M$ for all $x$ for some $M < \infty$, then $\sigma_{IS}^2 < \infty$.

In practice ensure $w(x) \leq M$ although it is neither necessary nor sufficient, as seen in the following example.
Importance Sampling: Example

\[ \pi(x) = \mathcal{N}(x; 0, 1), \; q(x) = \mathcal{N}(x; 0, \sigma^2) \]

\[ w(x) = \frac{\pi(x)}{q(x)} \propto \exp \left[ -x^2 \left( 1 - \frac{1}{\sigma^2} \right) \right]. \]

For \( \sigma^2 \geq 1 \), \( w(x) \leq M \) for all \( x \),
and for \( \sigma^2 < 1 \), \( w(x) \to \infty \) as \( |x| \to \infty \).

For \( \varphi(x) = x^2 \), we have \( \sigma_{\text{IS}}^2 < \infty \) for all \( \sigma^2 > 1/2 \).

For \( \varphi(x) = \exp \left( \frac{\beta}{2} x^2 \right) \), we have \( I < \infty \) for \( \beta < 1 \)
but \( \sigma_{\text{IS}}^2 = \infty \) for \( \beta > 1 - \frac{1}{2\sigma^2} \).
Question

Is there a best proposal that minimizes the variance $\sigma_{IS}^2$?

Proposition

The optimal proposal minimising $\mathbb{V}_q \left( \hat{I}_n^{IS} \right)$ is given by

$$q_{opt}(x) = \frac{|\phi(x)| \pi(x)}{\int_{\mathcal{X}} |\phi(x)| \pi(x) \, dx}.$$
Optimal Importance Distribution II

Proof.
We have indeed
\[ \sigma_{IS}^2 = \mathbb{E}_q (\varphi(X)w(X)) = \mathbb{E}_q (\varphi^2(X)w^2(X)) - I^2. \]

We also have by Jensen’s inequality for any \( q \)
\[ \mathbb{E}_q (\varphi^2(X)w^2(X)) \geq \left( \int_X |\varphi(x)| \pi(x) \, dx \right)^2. \]

For \( q = q_{\text{opt}} \), we have
\[
\mathbb{E}_{q_{\text{opt}}} (\varphi^2(X)w^2(X)) = \int_X \frac{\varphi^2(x)\pi^2(x)}{|\varphi(x)| \pi(x)} \, dx \times \int_X |\varphi(x)| \pi(x) \, dx
= \left( \int_X |\varphi(x)| \pi(x) \, dx \right)^2.
\]
\[ \square \]
Optimal Importance Distribution

$q_{\text{opt}}(x)$ can never be used in practice!

For $\varphi(x) > 0$ we have $q_{\text{opt}}(x) = \varphi(x) \pi(x) / I$ and $\forall q_{\text{opt}} \left( \hat{I}_n^{\text{IS}} \right) = 0$ but this is because

$$\varphi(x) w(x) = \varphi(x) \frac{\pi(x)}{q_{\text{opt}}(x)} = I,$$

it requires knowing $I$!

This can be used as a guideline to select $q$; i.e. select $q(x)$ such that $q(x) \approx q_{\text{opt}}(x)$.

Particularly interesting in rare event simulation, not quite in statistics.
Normalised Importance Sampling

Standard IS has limited applications in statistics as it requires knowing $\pi(x)$ and $q(x)$ exactly.

Assume $\pi(x) = \tilde{\pi}(x)/Z_\pi$ and $q(x) = \tilde{q}(x)/Z_q$, $\pi(x) > 0 \implies q(x) > 0$ and define

$$\tilde{w}(x) = \frac{\tilde{\pi}(x)}{\tilde{q}(x)}.$$

An alternative identity is

$$I = \mathbb{E}_\pi(\varphi(X)) = \frac{\int_X \varphi(x) \tilde{w}(x) q(x) dx}{\int_X \tilde{w}(x) q(x) dx}.$$
Proposition (SLLN for NIS)

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} q$ and assume that $\mathbb{E}_q(|\varphi(X)| \cdot w(X)) < \infty$. Then

$$\hat{I}_{n}^{\text{NIS}} = \frac{\sum_{i=1}^{n} \varphi(X_i) \tilde{w}(X_i)}{\sum_{i=1}^{n} \tilde{w}(X_i)}$$

is strongly consistent.

Proof.

Divide numerator and denominator by $n$. Both converge almost surely by the strong law of large numbers.

But, for finite $n$ $\hat{I}_{n}^{\text{NIS}}$ is biased, see notes Chapter 3.
CLT for NIS

Proposition

If \( \forall_q (\varphi(X)w(X)) < \infty \) and \( \forall_q (w(X)) < \infty \) then

\[
\sqrt{n}(\hat{I}_n^{\text{NIS}} - I) \Rightarrow \mathcal{N}(0, \sigma_{\text{NIS}}^2),
\]

where

\[
\sigma_{\text{NIS}}^2 := \forall_q \left( [\varphi(X)w(X)] - lw(X) \right)
= \int \frac{\pi(x)^2 (\varphi(x) - I)^2}{q(x)} \, dx.
\]
Proof

Proof.

First notice that with $X_1, \ldots, X_n$ i.i.d. $\sim q$

$$\sqrt{n}(\hat{I}_n^{\text{NIS}} - I) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{w}(X_i) [\varphi(X_i) - I]$$

where since $\tilde{w}(x) = \tilde{\pi}/\tilde{q}$

$$\mathbb{E}_q \left[ \tilde{w}(X_n)(\varphi(X_i) - I) \right] = 0.$$ 

Since $\nabla_q (\varphi(X)w(X)) < \infty$ by standard CLT

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{w}(X_i) [\varphi(X_i) - I] \Rightarrow \mathcal{N} \left( 0, \nabla_q \left( \tilde{w}(X_1)[\varphi(X_1) - I] \right) \right).$$
Proof ctd...

Proof.

The strong law of large numbers applied to the denominator

\[
\frac{1}{n} \sum_{i=1}^{n} \tilde{\omega}(X_i) \to \mathbb{E}_q[\tilde{\omega}(X_1)] = \frac{Z_\pi}{Z_q}, \quad \text{a.s.}
\]

By Slutsky’s theorem, combining the two

\[
\sqrt{n}(\hat{I}_n^{\text{NIS}} - 1) \Rightarrow \mathcal{N}\left(0, \mathbb{V}_q(\tilde{\omega}(X_1)[\varphi(X_1) - 1]) \frac{Z^2_q}{Z^2_\pi}\right)
\]

\[
\sim \mathcal{N}\left(0, \sigma^2_{\text{NIS}}\right).
\]

Alternatively, use Delta method.
Toy Example: t-distribution

• We want to compute \( I = \mathbb{E}_\pi(|X|) \) where \( \pi(x) \propto (1 + x^2/3)^{-2} \) (t_3-distribution).

(a) Directly sample from \( \pi \).

(b) Use \( q_1(x) = g_{t_1}(x) \propto (1 + x^2)^{-1} \) (t_1-distribution).

(c) Use \( q_2(x) \propto \exp(-x^2/2) \) (normal).
Toy Example: t-distribution

Figure: Sample weights obtained for 1000 realisations of $X_i$, from the different proposal distributions.
Toy Example: t-distribution

Figure: Estimates $\hat{I}_n$ of $I$ obtained after 1 to 1500 samples. The grey shaded areas correspond to the range of 100 independent replications.