# MS : MONTE CARLO INTEGRATION

In this chapter, we address two related problems :

(i) how to simulate observations according to some distribution

(ii) evaluate integrals $E\,\varphi(X) = \int \varphi(x)\,p(x)\,dx$.

## I – BASIC SAMPLING TECHNIQUES.

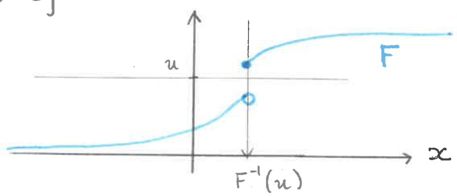
### I.1. Inverse function method.

Let $X$ be a RV with distribution function $F(x) = \int_{-\infty}^{x} p(u)\,du$

The generalised inverse of $F$ is

$F^{-1}(u) := \inf\{x \mid F(x) \geq u\}$

If $F^{-1}$ is known, then generating samples $\sim F$ is straightforward :




let $U \sim \mathcal{U}(0,1)$.

Then

$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x)$

$\to$ The RV $F^{-1}(U)$ has distribution $F$

$\hookrightarrow$ Once a generator of uniformly distributed RVs is available, we can draw samples $\sim F$, as long as $F^{-1}$ is explicitely known.

x Example : let $U \sim \mathcal{U}(0,1)$

$X \sim -\frac{1}{\lambda} \ln U$. Then $X \sim \mathrm{Exp}(\lambda)$ ■



Remarks : (i) Conversely, if $X \sim F$, and $F$ is continuous, then $F(X) \sim \mathcal{U}(0,1)$.

$F$ needs to be continuous.

Ex: $F = \mathcal{B}(1/2)$

Then $F(X) = \begin{cases} 1/2 & \text{w.p. } 1/2 \\ 1 & \text{w.p. } 1/2 \end{cases} \neq \mathcal{U}(0,1)$

(ii) $F^{-1}$ is not always explicitely known. A classic example is the normal distribution $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\,du$.

Alternatively, use rejection sampling.

### I.2. Rejection sampling

• Context : Generate samples $\sim p$, but $p$ is too complicated to do this directly (e.g. using the inverse function method). However, we have a simpler distribution $q$ such that

(a) we can draw samples $\sim q$

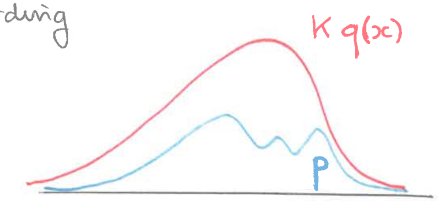
(b) there exists a constant $K$ such that $p(x) \leq Kq(x)$.

Note that necessarily, $K \geq 1$ since

$1 = \int p(x)\,dx \leq K \int q(x)\,dx = K$.

$q$ is known as a PROPOSAL DISTRIBUTION.

The idea is to sample according to $q$, and then to correct for using $q$ instead of $p$.



## Rejection Sampling

3

(i) Generate $U \sim \mathcal{U}(0,1)$
$X \sim q$ , independent

(ii) Accept $Y = X$ if $U \leq r(X) := \dfrac{p(X)}{K q(X)}$

(iii) Go back to (i) if rejection.

↳ Then $Y$ has distribution $p$.

Indeed, let $N := \inf \{ n \geq 1 \mid U_n \leq r(X_n) \}$,

first time we accept $Y$

number of times we go through the algorithm

where $(U_n, X_n) \overset{d}{=} (U, X)$.

Let $F$ be the distribution function associated with $p$.

We have

$$\mathbb{P}(Y \leq y, N = n)$$

$$= \mathbb{P}(U_1 > r(X_1), \dots, U_{n-1} > r(X_{n-1}), U_n \leq r(X_n), X_n \leq y)$$

Since $Y_n = X_n$ is accepted

$$= \left[ \mathbb{P}(U > r(X)) \right]^{n-1} \mathbb{P}(U_n \leq r(X_n), X_n \leq y)$$

$$\mathbb{P}(U > r(X)) = \mathbb{E}\, \mathbb{1}(U > r(X))$$

$$= \int_{\mathbb{R}} \left( \int_0^1 \mathbb{1}(u > r(x))\, du \right) q(x)\, dx$$

4

$$\mathbb{P}(U > r(X)) = \int_{\mathbb{R}} (1 - r(x))\, q(x)\, dx$$

$$= 1 - \int r(x)\, q(x)\, dx$$

$$= 1 - \int K^{-1} p(x)\, dx = 1 - K^{-1}.$$

$$\Rightarrow \mathbb{P}(Y \leq y, N = n) = (1 - K^{-1})^{n-1} \underbrace{\mathbb{P}(U_n \leq r(X_n), X_n \leq y)}$$

$$= \int_{\mathbb{R}} \left( \int_0^1 \mathbb{1}(u \leq r(x))\, du \right) \mathbb{1}(x \leq y)\, q(x)\, dx$$

$$= \int_{\mathbb{R}} \mathbb{1}(x \leq y)\, r(x)\, q(x)\, dx$$

$$= \int_{-\infty}^{y} K^{-1} p(x)\, dx = K^{-1} F(y)$$

$$\boxed{\mathbb{P}(Y \leq y, N = n) = (1 - K^{-1})^{n-1} \frac{F(y)}{K}}$$

↳ Marginals are:

- $\mathbb{P}(N = n) = \lim_{y \to \infty} \mathbb{P}(Y \leq y, N = n) = K^{-1}(1 - K^{-1})^{n-1}$

$\boxed{N \sim \text{Geom}(K^{-1})}$

- $\mathbb{P}(Y \leq y) = \sum_{n \geq 1} (1 - K^{-1})^{n-1} K^{-1} F(y)$

$$= K^{-1} F(y) \sum_{n \geq 1} (1 - K^{-1})^{n-1}$$

$$= K^{-1} F(y) \left( 1 - (1 - K^{-1}) \right)^{-1}$$

$$= F(y)$$

$\boxed{Y \sim F}$

■

The proof indicates that $N = \#$ iterations until we first accept a sample is $\sim \mathrm{Geom}(K^{-1}) \Rightarrow \mathbb{E} N = K$.

↳ We need on average $K$ attemps to generate a single observation $\sim p$.

$\Rightarrow$ Choose $(q, K)$ such that $K$ is as close to $1$ as possible. In other words, choose $q$ that looks like $p$ as much as possible.

× <u>Example</u>: $p \sim \mathcal{N}(0,1)$

$q \sim$ Cauchy $\qquad q(x) = \dfrac{1}{\pi} \dfrac{1}{1+x^2}$

can easily draw samples $\sim q$ using the inverse function method

Choice of $K$:

$$\frac{p(x)}{q(x)} = \sqrt{\frac{\pi}{2}} \underbrace{(1+x^2)\, e^{-x^2/2}}_{=:\, h(x)}$$

$h'(x) = x(1-x^2)\, e^{-x^2/2}$ vanishes for $x=0$ and $x=1$, with a global minimum at $x=\pm 1$; and $h(\pm 1) = 2e^{-1/2}$.

Thus $$\frac{p(x)}{q(x)} \leq \sqrt{\frac{2\pi}{e}}$$

Take $K = \sqrt{2\pi/e} \simeq 1.56$. The acceptance probability is $K^{-1} = 0.66$

• <u>Remarks</u>: (i) The proof that $Y$ has distribution $p$ can be easily adapted if instead of computing $\mathbb{P}(Y \leq y, N=n)$, one computes $\mathbb{P}(Y \in B, N=n)$, for $B \in \mathcal{B}(\mathbb{R})$.

$\Rightarrow$ Rejection methods remain valid in $\mathbb{R}^d$.



For example, let $B \subset [0,1] \times [0,1]$, $B \in \mathcal{B}(\mathbb{R}^2)$. We want to draw samples uniformly distributed over $B$. The idea is to generate points uniformly over $[0,1]^2$, until a point falls in $B$. Then indeed this point $\sim \mathcal{U}(B)$.

$$p(x,y) = |B|^{-1}\, \mathbb{1}_B(x,y), \quad |B| = \text{area of } B$$

$$q(x,y) = \mathbb{1}_{[0,1]^2}(x,y)$$

Take $K = |B|^{-1}$ and $r(x,y) = \mathbb{1}_B(x,y)$.

no need to know the constant $K$! which leads us to the following generalization:

(ii) Suppose that $p(x)$ is known up to a constant:

$$p(x) = \boxed{Z_p^{-1}}\, \boxed{p_0(x)}$$

unknown — can be computed.

(often the case in Bayesian statistics).

In addition, suppose that there exists a density $q_0(x)$ from which we can easily draw samples, and for which $p_0(x) \leq q_0(x) = q(x)$

It suffices to take $K = Z_p^{-1}$ (unknown) since

$$r(x) = \frac{p(x)}{K q(x)} = \frac{Z_p^{-1}\, p_0(x)}{Z_p^{-1}\, q_0(x)} = \frac{p_0(x)}{q_0(x)}$$

does not depend on $K$ & $Z_p$

# II. MONTE CARLO INTEGRATION

A typical application of the sampling techniques described in section I concerns the evaluation of integrals of the form

$$I = \mathbb{E}[\varphi(X)] = \int \varphi(x)\, p(x)\, dx,$$

where

- $\varphi : \mathbb{R}^d \to \mathbb{R}$ is known
- $X \sim p$, and we know how to generate samples from $p$.

A naïve MC estimator of $I$ is $\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i)$.

## II.1. Properties of $\hat{I}_n$.

→ $\hat{I}_n$ is unbiased: $\mathbb{E}\hat{I}_n = I$

→ SLLN: Provided $\mathbb{E}|\varphi(X)| < \infty$, $\hat{I}_n \to I$ a.s.

→ CLT: Provided $\mathbb{E}(\varphi(X))^2 < \infty$, $n^{1/2}(\hat{I}_n - I) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$

where $\sigma^2 =. \operatorname{Var} \varphi(X) = \mathbb{E}(\varphi(X))^2 - (\mathbb{E}\varphi(X))^2$

$$= \int \varphi^2(x)\, p(x)\, dx - I^2$$

In other words, "$\hat{I}_n$ converges to $I$ in $O(n^{-1/2})$".

↳ useful to construct confidence intervals.

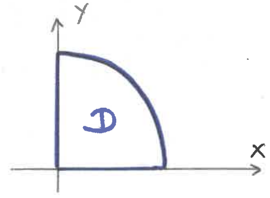
$$\hat{\sigma}_n^2 := \frac{1}{n} \sum_{i=1}^{n} \varphi^2(X_i) - \hat{I}_n^2 \xrightarrow{a.s.} \sigma^2$$

+ Slutsky theorem ⇒ $\dfrac{n^{1/2}(\hat{I}_n - I)}{\hat{\sigma}_n} \xrightarrow{d} \mathcal{N}(0,1)$.



× Example = Estimation of $\pi$

Let
- $(X, Y) \sim \mathcal{U}(C)$, for $C = [0,1] \times [0,1]$
- $\varphi(x,y) = \mathbb{1}(x^2 + y^2 \leq 1)$.
- $D = \{(x,y) \in \mathbb{R}_+^2 \mid x^2 + y^2 \leq 1\}$




Then

$$I = \iint_C \mathbb{1}_D(x,y)\, dx\, dy = \frac{\pi}{4}$$

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_D(X_i, Y_i) \xrightarrow{a.s.} \frac{\pi}{4}$$

$$\Leftrightarrow \quad 4\hat{I}_n \xrightarrow{a.s.} \pi$$

The variance of $\varphi(X,Y)$ is $\sigma^2 = I - I^2 = \frac{\pi}{4}\left(1 - \frac{\pi}{4}\right) \simeq 0.17$, which can be estimated using $\hat{\sigma}_n^2 = \hat{I}_n - \hat{I}_n^2$.

CLT: $\dfrac{n^{1/2}(\hat{I}_n - I)}{\sqrt{\hat{I}_n - \hat{I}_n^2}} \xrightarrow{d} \mathcal{N}(0,1)$

A $(1-\alpha)$ confidence interval for $I$ is then $\hat{I}_n \pm z_{1-\frac{\alpha}{2}} \hat{\sigma}_n n^{-1/2}$.

## II.2. Variance Reduction Techniques.

The MC estimator $\hat{I}_n$ of $I$ is $\approx \mathcal{N}(I, \sigma^2 n^{-1})$ for $n$ large. The approximation error is of order $\sigma^2/n$. To reduce the error, a common strategy is to reduce $\sigma^2$, so that for a given level of accuracy, the number of points to generate is reduced: a method decreasing $\sigma^2$ by a factor 2 allows half as many samples to draw to keep the same estimation error.

## II.2.a. Importance Sampling.

• Goal: to estimate $I = \mathbb{E}\,\varphi(X) = \int \varphi(x)\,p(x)\,dx$

If $\varphi$ is large where $p$ is small, the naïve MC estimator $\hat{I}_n$ is bad unless $n$ is very large.

× Examples: (i) $X \sim \mathcal{N}(0,1)$.

We wish to estimate $I = \mathbb{E}\,\varphi(X) = \mathbb{E}\,\mathbb{1}(X>6)$

$= \mathbb{P}(X>6)$.

$= \int 1(x>6)\,\phi(x)\,dx$

This integral is of order $10^{-9}$

⇒ Unless $n \approx 10^9$, it is very likely to obtain $\hat{I}_n = 0$.

(ii) $X \sim \mathcal{N}(m,1)$

$\varphi(x) = \exp\left(-mx + \frac{1}{2}m^2\right)$.

$\forall m, \quad I = \mathbb{E}\,\varphi(X) = \int \varphi(x)\,\phi(x)\,dx = \int \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx = 1.$

However, 95% of the $X_i$ lie between $m-2$ and $m+2$, while $\varphi$ decreases rapidly to 0 as $m$ increases.

⇒ For $m$ large, $\hat{I}_n \approx 0$, far from the theoretical value of 1.

• Strategy: do not sample from $p$ directly, but from some other density $q$, chosen in such a way that $q$ is large whenever $\varphi$ is large (and correct for this).




$$I = \mathbb{E}[\varphi(X)] = \int \varphi(x)\,p(x)\,dx$$

$$= \int \varphi(x)\,\frac{p(x)}{q(x)}\,q(x)\,dx$$

introduce $w(y) = \frac{p(y)}{q(y)}$

$$= \int \varphi(y)\,w(y)\,q(y)\,dy$$

$$= \mathbb{E}[\varphi(Y)\,w(Y)], \quad Y \sim q.$$

The IMPORTANCE SAMPLING estimator is

$\tilde{I}_n = \frac{1}{n}\sum_{i=1}^{n} w(Y_i)\,\varphi(Y_i)$, where $Y_i \sim q$ iid

→ SLLN: $\tilde{I}_n \xrightarrow{a.s.} I$

→ CLT: $n^{1/2}(\tilde{I}_n - I) \xrightarrow{d} \mathcal{N}(0, s^2)$, where

$s^2 = \mathrm{Var}[w(Y)\,\varphi(Y)]$

$= \int w^2(y)\,\varphi^2(y)\,q(y)\,dy - I^2$

which can be estimated using $\tilde{s}_n^2 = \frac{1}{n}\sum w^2(Y_i)\,\varphi^2(Y_i) - \tilde{I}_n^2$

Remarks (i) Importance sampling is useful as well in cases where we do not know how to simulate directly from p.

(ii) Compare $\sigma^2 = \mathrm{Var}[\varphi(X)]$ with $X \sim p$

$s^2 = \mathrm{Var}[w(Y)\,\varphi(Y)]$ $\quad Y \sim q$.

We want to find $q$ such that $s^2$ is as small as possible.

$s^2 = \mathrm{Var}[w(Y)\,\varphi(Y)] = \int w^2(y)\,\varphi^2(y)\,dy - I^2$

$Y \sim q$

(11)

$$s^2 = \int \frac{p^2(y)}{q^{\cancel{2}}(y)} \varphi^2(y) \cancel{q(y)}\, dy - I^2$$

$$= \mathbb{E}\left[\omega(X)\varphi^2(X)\right] - I^2 \quad , \quad X \sim p$$

$$\geqslant \left[\mathbb{E}|\varphi(X)|\right]^2 - I^2$$

Cauchy-Schwartz:

$U := \sqrt{\omega(X)}\, \varphi(X)$

$V := \dfrac{1}{\sqrt{\omega(X)}}$

$$\mathbb{E}|UV| \leqslant \sqrt{\mathbb{E}U^2\, \mathbb{E}V^2}$$

$$\Leftrightarrow \mathbb{E}|\varphi(X)| \leqslant \sqrt{\mathbb{E}\left(\omega(X)\varphi^2(X)\right) \times 1}$$

& equality is obtained for $q(x) = q^*(x) = \dfrac{|\varphi(x)|p(x)}{\int |\varphi(u)|p(u)\,du}$

The 'best' we can do. However, we cannot compute $q^*$ in practice. The result indicates that the density $q$ should be chosen in a way that it places mass where the product $|\varphi(x)|\,p(x)$ is large.

And if we could, one draw would be enough:

$$\tilde{I}_1 = \omega(Y_1)\varphi(Y_1) = \frac{p(Y_1)}{q^*(Y_1)}\varphi(Y_1) = \int |\varphi(u)|p(u)\,du = I.$$

$n=1$

x Remark: The distribution $p$ is usually known up to a normalizing constant: $p(x) = \boxed{Z_p^{-1}}\, p_o(x)$.

unknown

Put $q(x) := Z_q^{-1}\, q_o(x)$

(12)

Then $\mathbb{E}[\varphi(X)] = \int \varphi(x)\, p(x)\, dx$

$$= \frac{Z_q}{Z_p} \int \varphi(x) \boxed{\frac{p_o(x)}{q_o(x)}} q(x)\, dx$$

$$\simeq \boxed{\frac{Z_q}{Z_p}} \frac{1}{n} \sum_{i=1}^n \boxed{\omega_o(Y_i)}\, \varphi(Y_i) \quad ; \quad Y_i \sim q$$

The ratio $\dfrac{Z_q}{Z_p}$ can be evaluated using the same sample:

$$\frac{Z_p}{Z_q} = \frac{1}{Z_q} \int p_o(x)\, dx = \int \frac{p_o(x)}{q_o(x)} q(x)\, dx \simeq \frac{1}{n} \sum_{i=1}^n \omega_o(Y_i)$$

## IMPORTANCE SAMPLING

$$\mathbb{E}\,\varphi(X) \simeq \sum_{i=1}^n \tilde{\omega}_o(Y_i)\, \varphi(Y_i) \quad , \quad Y_i \sim q$$

where

$$\tilde{\omega}_o(Y_i) = \frac{\omega_o(Y_i)}{\sum_{j=1}^n \omega_o(Y_j)} = \frac{\dfrac{p_o(Y_i)}{q_o(Y_i)}}{\sum_{j=1}^n \dfrac{p_o(Y_j)}{q_o(Y_j)}}$$

Note that $\tilde{\omega}_o \geqslant 0$, and sum to 1.

x Example: Back to the estimation of $I = \mathbb{P}(X > 6)$, with $X \sim \mathcal{N}(0,1)$ – denote $p$ the standard normal density.

Consider $T \sim \mathrm{Exp}(1)$.

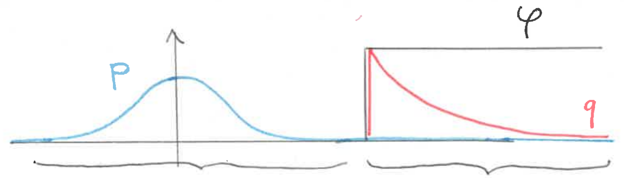
Then $Y = 6 + T$ has density $q(y) = e^{-(y-6)}\, \mathbb{1}(y > 6)$

Since $\mathbb{P}(Y \leqslant y) = \mathbb{P}(T \leqslant y - 6) = 1 - e^{-(y-6)}$.

Then $w(y)\,\varphi(y) = \dfrac{p(y)}{e^{-(y-6)}\mathbb{1}(y\geq 6)}\,\mathbb{1}(y\geq 6) = \dfrac{1}{\sqrt{2\pi}}\, e^{y-6-\frac{y^2}{2}}$

Consider samples $Y_i$, $i=1,..,n$, and the importance sampling estimator

$$\tilde{I}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(Y_i - 6 - \frac{Y_i^2}{2}\right)$$




The product $p(x)|\varphi(x)|$ is equal to zero here

this region is → where $p(x)|\varphi(x)|$ is maximum

↓ justifies our choice of q.

## II.2.b. Conditioning

- Goal: to estimate $I = \mathbb{E}\,\varphi(X) = \int \varphi(x)\,p(x)\,dx$, where $\mathbb{E}[\varphi^2(X)] < \infty$.
- Strategy: Conditioning leaves the mean unchanged, while reducing the variance.

Let $\psi(Y) = \mathbb{E}[\varphi(X) \mid Y]$

Then

→ $\mathbb{E}\,\psi(Y) = \mathbb{E}\,\mathbb{E}[\varphi(X)\mid Y] = \mathbb{E}\,\varphi(X)$

→ $\sigma^2 = \mathrm{Var}\,\varphi(X)$

$$= \mathrm{Var}\,\underbrace{\mathbb{E}(\varphi(X)\mid Y)}_{=\psi(Y)} + \mathbb{E}\,\mathrm{Var}(\varphi(X)\mid Y)$$

$$\geq \mathrm{Var}\,\psi(Y)$$

⇒ $\psi$ has mean $I$, and smaller variance than $\varphi$.



Use a variable $Y \sim q$ from which we can easily generate samples, and such that we can compute $\psi(y) = \mathbb{E}[\varphi(X) \mid Y=y]$.

Consider the estimator $\tilde{I}_n = \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i)$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\varphi(X) \mid Y_i].$$

{ * simulate $Y_i$ + * know the conditional exp.

→ SLLN: $\tilde{I}_n \xrightarrow{a.s} I$

→ CLT: provided $\mathbb{E}\,\psi^2(Y) < \infty$, then

$n^{1/2}(\tilde{I}_n - I) \xrightarrow{d} \mathcal{N}(0, s^2)$, with

$s^2 = \mathrm{Var}\,\psi(Y) = \mathrm{Var}\,\mathbb{E}[\varphi(X)\mid Y] = \mathbb{E}[\psi^2(Y)] - I^2$,

which can be estimated using

$s_n^2 = \frac{1}{n}\sum_{i=1}^{n} \psi^2(Y_i) - \tilde{I}_n^2$.

× Example = (continued from page 8). Estimation of $\pi$.

Since $X, Y \sim \mathcal{U}(0,1)$ are independent, we have

$$\mathbb{E}(\mathbb{1}_D(X,Y) \mid Y=y) = \mathbb{P}(X^2 + y^2 \leq 1)$$
$$= \mathbb{P}(X \leq \sqrt{1-y^2})$$
$$= \sqrt{1-y^2} = \psi(y).$$

⇒ $\tilde{I}_n = \frac{1}{n}\sum_{i=1}^{n}\sqrt{1-Y_i^2}$

The variance is $s^2 = \mathbb{E}\,\psi^2(Y) - I^2$

$$= \int_0^1 (1-y^2)\,dy - \left(\frac{\pi}{4}\right)^2$$
$$= \frac{2}{3} - \left(\frac{\pi}{4}\right)^2 \simeq 0.05$$

Compare with $\sigma^2 \simeq 0.17$ $\quad$ $s = 0.22$, $\sigma \simeq 0.41$ → $\tilde{I}_n$ is twice more accurate than $\hat{I}_n$. ∎

## II.2.c. Antithetic variables.

We present the approach in the case where the $X_i \sim p$ are simulated with the inverse function method; $X_i = F^{-1}(U_i)$, $U_i \sim \mathcal{U}(0,1)$, $F$ = distribution function of the $X_i$.

The standard MC estimator of $I = \mathbb{E}\,\varphi(X)$ is

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i) = \frac{1}{n}\sum_{i=1}^{n}\varphi(F^{-1}(U_i)).$$

Since $1-U_i \sim \mathcal{U}(0,1)$ as well, the estimator

$$\tilde{I}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\Big(\varphi(F^{-1}(U_i)) + \varphi(F^{-1}(1-U_i))\Big)$$

is also unbiased and convergent: $\tilde{I}_n \xrightarrow{a.s} I$. Moreover,

→ Variance of $\hat{I}_n$ is $\frac{1}{n}\mathrm{Var}[\varphi(X)]$

→ Variance of $\tilde{I}_n$ is

$$\frac{1}{4n^2}\sum_{i=1}^{n}\Big\{\mathrm{Var}\,\varphi(F^{-1}(U_i)) + \mathrm{Var}\,\varphi(F^{-1}(1-U_i)) + 2\,\mathrm{Cov}\big(\varphi(F^{-1}(U_i)), \varphi(F^{-1}(1-U_i))\big)\Big\}$$

$$= \frac{1}{2n}\Big\{\mathrm{Var}\,\varphi(F^{-1}(U)) + \mathrm{Cov}\big(\varphi(F^{-1}(U)), \varphi(F^{-1}(1-U))\big)\Big\}$$

$$= \frac{1}{2n}\Big\{\mathrm{Var}\,\varphi(X) + \underbrace{\mathrm{Cov}\big(\varphi(F^{-1}(U)), \varphi(F^{-1}(1-U))\big)}\Big\}$$

Cauchy-Schwartz

$$\leq \big(\mathrm{Var}\,\varphi(F^{-1}(U))\big)^{1/2}\big(\mathrm{Var}\,\varphi(F^{-1}(1-U))\big)^{1/2}$$

$$= \mathrm{Var}\,\varphi(X).$$

$$\Rightarrow \frac{\mathrm{Var}\,\tilde{I}_n}{\mathrm{Var}\,\hat{I}_n} \leq \frac{\frac{1}{2n}\{\mathrm{Var}\,\varphi(X) + \mathrm{Var}\,\varphi(X)\}}{\frac{1}{n}\mathrm{Var}\,\varphi(X)} = 1$$



↳ $\tilde{I}_n$ has smaller variance than $\hat{I}_n$, but it requires twice as many computations as $\hat{I}_n$ → no clear gain.

However, if $\varphi$ is monotonic, Chebyshev covariance inequality ensures that $\mathrm{Cov}\big(\varphi(F^{-1}(U)), \varphi(F^{-1}(1-U))\big) \leq 0$, so that

$$\frac{\mathrm{Var}\,\tilde{I}_n}{\mathrm{Var}\,\hat{I}_n} \leq \frac{1}{2}$$

Implementation cost is at least compensated by the variance reduction.

Indeed, let $X' \overset{d}{=} X$, $X, X'$ independent, $\varphi$ non-decreasing, $\psi$ non-increasing, such that $\varphi(X)$ and $\psi(X)$ are square integrable.

$$\mathrm{cov}\big(\varphi(X) - \varphi(X'), \psi(X) - \psi(X')\big)$$
$$= \mathbb{E}\big[(\varphi(X) - \varphi(X'))(\psi(X) - \psi(X'))\big]$$
$$= \int \underbrace{(\varphi(X(\omega)) - \varphi(X'(\omega))) \times (\psi(X(\omega)) - \psi(X'(\omega)))}\,\mathbb{P}(d\omega)$$

↗ $X(\omega) \leq X'(\omega)$ → product is $\leq 0$

↘ $X(\omega) \geq X'(\omega)$ → product is $\leq 0$

⇒ the covariance term is $\leq 0$. Expanding + using bilinearity of the covariance operator:

$$\mathrm{cov}\big(\varphi(X) - \varphi(X'), \psi(X) - \psi(X')\big)$$

cross product terms vanish

$$= \mathrm{cov}(\varphi(X), \psi(X)) + \mathrm{cov}(\varphi(X'), \psi(X'))$$
$$= 2\,\mathrm{cov}(\varphi(X), \psi(X)) \leq 0$$

& take $\underset{\text{non decr}}{\varphi = F^{-1}}$, $\underset{\text{non incr.}}{\psi = F^{-1}\circ h}$ $h(x) = 1-x$. ∎

(17)

× *Example*: Estimation of $\mathbb{E}(e^X)$, $X \sim \mathcal{N}(0,1)$ using antithetic variables.

We have $\varphi(x) = e^x$

With $h(x) = -x$, $X$ and $h(X)$ have the same $\mathcal{N}(0,1)$ distribution ⇒ compare

→ the naive MC estimator $\hat{I}_n := \frac{1}{n} \sum_{i=1}^n e^{X_i}$, with

→ $\tilde{I}_n := \frac{1}{n} \sum_{i=1}^n \frac{\varphi(X_i) + \varphi(h(X_i))}{2} = \frac{1}{n} \sum_{i=1}^n \frac{e^{X_i} + e^{-X_i}}{2}$.

Expect here as well that $\frac{\operatorname{var} \tilde{I}_n}{\operatorname{var} \hat{I}_n} \leq \frac{1}{2}$, since Chebychev covariance inequality holds true provided $h$ is non-increasing (see derivation on the previous page).

• $\sigma^2 := \operatorname{var}(e^X) = e(e-1)$ (after calculations) $\operatorname{var} \hat{I}_n = \frac{\sigma^2}{n}$

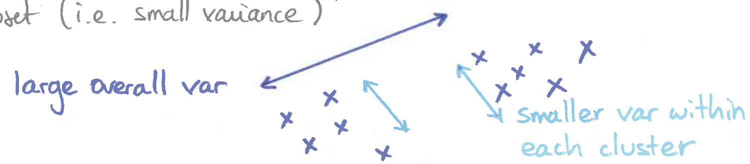
• $\operatorname{var} \tilde{I}_n = \frac{s^2}{n}$, with $s^2 = \frac{1}{2}(\operatorname{var} e^X + \operatorname{cov}(e^X, e^{-X}))$

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

We obtain $\frac{s^2}{\sigma^2} = \frac{e-1}{2e} \simeq 0.32$ [var reduced by a factor 3] ∎

## II.2.d. Stratification

The idea is to partition the set $\mathcal{X}$ of possible values of $X$ into subsets, in such a way that $X$ is relatively homogeneous on each subset (i.e. small variance).




(18)

Let $\mathcal{X} = \mathcal{X}_1 \cup \dots \cup \mathcal{X}_K$.

disjoint

Suppose $p_k = \mathbb{P}(X \in \mathcal{X}_k)$ known. In addition, assume that we know how to generate samples according to the conditional distribution $X \mid X \in \mathcal{X}_k$. Then:

$$I = \mathbb{E}\,\varphi(X) = \sum_{k=1}^K \mathbb{E}(\varphi(X) \mid X \in \mathcal{X}_k)\, \mathbb{P}(X \in \mathcal{X}_k)$$

$$= \sum_{k=1}^K p_k \underbrace{\mathbb{E}(\varphi(X) \mid X \in \mathcal{X}_k)}_{=: \mu_k}$$

this term can easily be estimated using $\frac{1}{n_k} \sum_{i=1}^{n_k} \varphi(X_{i,k})$, where $n_1 + \dots + n_K = n$, and $X_{1,k}, \dots, X_{n_k,k}$ are iid with distribution $X \mid X \in \mathcal{X}_k$.

Consider the estimator

$$\tilde{I}_n := \sum_{k=1}^K p_k \left( \frac{1}{n_k} \sum_{i=1}^{n_k} \varphi(X_{i,k}) \right)$$

→ SLLN: Provided $\mathbb{E}|\varphi(X)| < \infty$, $\tilde{I}_n \xrightarrow{a.s} I$ as $n \to \infty$.

→ If $\mathbb{E}[\varphi^2(X)] < \infty$, then $s_n^2 = \operatorname{Var} \tilde{I}_n$

$$= \sum_{k=1}^K \frac{p_k^2}{n_k} \underbrace{\operatorname{Var}(\varphi(X) \mid X \in \mathcal{X}_k)}_{=: \sigma_k^2}$$

$$= \sum_{k=1}^K \frac{p_k^2}{n_k} \sigma_k^2.$$

Using:

$\operatorname{Var} \varphi(X) = \operatorname{var} \mathbb{E}(\varphi(X) \mid Y) + \mathbb{E} \operatorname{var}(\varphi(X) \mid Y)$

Also, $\sigma^2 = \operatorname{Var} \varphi(X)$

$$= \sum_{k=1}^K p_k \sigma_k^2 + \sum_{k=1}^K p_k (\mu_k - I)^2$$

19

Thus, with $n_k := p_k n$, we get

$$\operatorname{var} \tilde{I}_n = s_n^2 = \frac{1}{n}\sum_{k=1}^{K} p_k \sigma_k^2 \leq \frac{\operatorname{var} \varphi(X)}{n} = \operatorname{Var} \hat{I}_n$$

$$\Rightarrow \boxed{\operatorname{var} \tilde{I}_n \leq \operatorname{var} \hat{I}_n}$$

variance reduction!

Remarks (i) In fact, we can go further and optimize the variance of $\tilde{I}_n$ with respect to $n_1, \dots, n_K$, subject to $n_1 + \dots + n_K = n$. The optimum solution $(n_1^*, \dots, n_K^*)$ is found to be $(n_1^*, \dots, n_K^*) = \left( \frac{p_1 \sigma_1}{\sum p_k \sigma_k} n, \dots, \frac{p_K \sigma_K}{\sum p_k \sigma_k} n \right)$

cannot be computed in practice, since the $\sigma_k$ are unknown, but we can proceed in two steps:

(a) estimate $\sigma_k$ using a first simulation

(b) perform a second simulation using the optimal allocation.

(ii) The methodology is similar to variance reduction techniques using conditioning. Compare:

- conditioning: simulate $Y$ & know the cond exp $\mathbb{E}(\varphi(X) \mid Y)$
- stratification: know the law of $Y$ & estimate $\mathbb{E}(\varphi(X) \mid Y)$.
  (i.e. the $p_k$)

variable $Y$ indicates in which stratum $X$ belongs to.

× Example: Estimation of $I = \mathbb{E}(\cos X) = \int_0^1 \cos x \, dx$, $X \sim \mathcal{U}(0,1)$. We have $\hat{I}_n = \frac{1}{n} \sum \cos X_i$, $X_i \sim \mathcal{U}(0,1)$ iid.

20

Next, consider the strata $\mathcal{X}_k = [x_{k-1}, x_k] = \left[\frac{k-1}{n}, \frac{k}{n}\right]$, $1 \leq k \leq n$.

$\mathbb{P}(X \in \mathcal{X}_k) = \frac{1}{n}$, $\forall k$

Moreover, $X \mid X \in \mathcal{X}_k \sim \mathcal{U}(x_{k-1}, x_k)$, so that the stratified estimator is $\tilde{I}_n = \frac{1}{n} \sum_{k=1}^{n} \cos U_k$, $U_k \sim \mathcal{U}(x_{k-1}, x_k)$.

→ We show next that $\tilde{I}_n$ converges to $I$ in $O(n^{-3/2})$ [much faster that the usual $O(n^{-1/2})$ rate]

- In fact, we prove the result in greater generality, for a differentiable function $\varphi$ [here $\varphi(x) = \cos x$], such that $M := \|\varphi'\|_\infty < \infty$.
- Recall the mean value theorem: for a continuous function $f$ on $[a,b]$, differentiable on $(a,b)$, there exists $c \in (a,b)$ such that $f(b) - f(a) = f'(c)(b-a)$.
- Thus, $\exists\, \theta_k \in (x_{k-1}, U_k)$ s.t. $\frac{\varphi(U_k) - \varphi(x_{k-1})}{U_k - x_{k-1}} = \varphi'(\theta_k)$.
- $$\operatorname{var}(\varphi(U_k)) = \operatorname{Var}\big(\varphi(x_{k-1}) + (U_k - x_{k-1})\varphi'(\theta_k)\big) = \operatorname{var}\big(\underbrace{(U_k - x_{k-1})}_{\leq 1/n} \underbrace{\varphi'(\theta_k)}_{\leq \|\varphi'\|_\infty = M}\big) \leq \frac{M^2}{n^2}$$
- $\operatorname{var} \tilde{I}_n = \operatorname{var}\left(\frac{1}{n} \sum \varphi(U_k)\right) \leq \frac{M^2}{n^3}$; so that

$\operatorname{var} \tilde{I}_n = O(n^{-3})$ indeed. ∎

# III. QUASI-MONTE-CARLO (QMC)



## III.1. Numerical Integration.

Unlike MC techniques, which uses random subdivisions of the support of integration, numerical methods use regular sub-divisions. We review briefly here the most common ones, and discuss their order of convergence.

× Goal: Approximation of $I = \int_a^b \varphi(x)\,dx$ , $a<b$

$\varphi$ = continuous on $[a,b]$

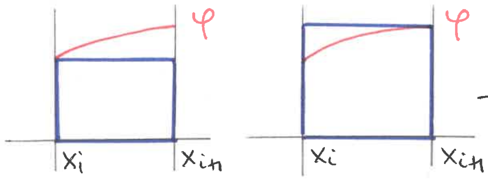
dimension $d=1$

integration with respect to the uniform density

× Notation: $x_i^{(n)} = a + i\dfrac{(b-a)}{n}$ . Note that $x_{i+1}^{(n)} - x_i^{(n)} = \dfrac{b-a}{n}$

When there is no confusion, we omit the superscript $n$, and write $x_i$ for $x_i^{(n)}$.

• Rectangle Method: the idea is rather simple: approximate $\varphi$ using piewise-constant functions, and replace/approximate the area under the curve with rectangular areas:




$\longrightarrow R_n^{(r)} = \dfrac{b-a}{n}\sum_{i=1}^{n}\varphi(x_i)$

$$R_n^{(\ell)} = \sum_{i=0}^{n-1}(x_{i+1}-x_i)\,\varphi(x_i) = \frac{b-a}{n}\sum_{i=0}^{n-1}\varphi(x_i)$$

convergence rate $O(n^{-1})$

↳ Result: if $\varphi$ is $\mathcal{C}^1$ on $[a,b]$ ; $M_1 = \sup_{[a,b]}|\varphi'|$ ,

then $\left| \int_a^b \varphi(x)\,dx - R_n^{(\ell/r)} \right| \leq \dfrac{M_1}{2n}(b-a)^2$



proof:

$$\left| \int_a^b \varphi(x)\,dx - \sum_{i=0}^{n-1}(x_{i+1}-x_i)\,\varphi(x_i) \right|$$

$$= \left| \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} (\varphi(x)-\varphi(x_i))\,dx \right|$$

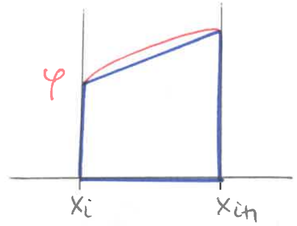
$$\leq \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} \underbrace{|\varphi(x)-\varphi(x_i)|}_{\leq M_1(x-x_i)}\,dx$$

(mean value theorem)

$$= \frac{M_1}{2}\sum_{i=0}^{n-1}(x_{i+1}-x_i)^2 = \frac{M_1}{2}\,n \times \frac{(b-a)^2}{n^2} \quad \blacksquare$$

• Trapezoidal Method: Average the left rectangle approximate $R_n^{(\ell)}$ with the right approximate $R_n^{(r)}$ : $R_n = \frac{1}{2}(R_n^{(\ell)} + R_n^{(r)})$ .

Geometrically, $R_n$ represents the area of a trapeze.




↳ Result: if $\varphi \in \mathcal{C}^2[a,b]$, $M_2 = \sup_{[a,b]}|\varphi''|$ ,

then

$$\left| \int_a^b \varphi(x)\,dx - R_n \right| \leq \frac{M_2}{12n^2}(b-a)^3$$

Faster $O(n^{-2})$ rate of convergence.

proof: for $n=1$, we need to bound $\int_a^b \varphi(x)\,dx - \dfrac{\varphi(a)+\varphi(b)}{2}(b-a)$ .

Consider $b$ as a variable, and study the function

$$f(x) = \int_a^x \varphi(u)\,du - \frac{\varphi(a)+\varphi(x)}{2}(x-a) , \quad x\in[a,b]$$

Note that $f(a) = 0$.

- $f'(x) = \varphi(x) - \frac{\varphi'(x)}{2}(x-a) - \frac{\varphi(a)+\varphi(x)}{2}$, $\quad f'(a) = 0$
- $f''(x) = \varphi'(x) - \frac{\varphi''(x)}{2}(x-a) - \frac{\varphi'(x)}{2} - \frac{\varphi'(x)}{2} = -\frac{\varphi''(x)}{2}(x-a)$.

$$\Rightarrow |f'(x)| = \int_a^x |f''(u)|\,du \le \int_a^x \frac{M_2}{2}(u-a)\,du = \frac{M_2}{4}(x-a)^2.$$

$$\Rightarrow |f(x)| = \int_a^x |f'(u)|\,du \le \int_a^x \frac{M_2}{4}(u-a)^2\,du = \frac{M_2}{12}(x-a)^3$$

Now, for $n \geq 2$, apply the same technique on $[x_i, x_{i+1}]$, and sum all the terms.

- Simpson Method: on each interval $[x_i, x_{i+1}]$, replace $\varphi$ with a second-order polynomial $P$, such that $\varphi$ and $P$ agree on $x_i$, $x_{i+1}$, and $\frac{1}{2}(x_i + x_{i+1})$.

↳ Result: if $\varphi \in \mathcal{C}^4[a,b]$, then we can achieve an error of order $O(n^{-4})$.

#Take Away
The more regular $\varphi$, the faster the numerical techniques are.

↑ Much better than the $O(n^{-1/2})$ rate of MC integration techniques, as soon as $\varphi$ is $\mathcal{C}^1$.

- But what happens in higher dimensions?

If $\varphi$ is $\mathcal{C}^s$ on $[0,1]^d$, then there exists methods with $O(n^{-s/d})$ rate of covergence. When $d$ gets large, the speed of convergence collapses → "curse of dimensionality".

⇒ In high dimension, MC techniques in $O(n^{-1/2})$ are preferable.



## III.2. QMC methods.

Let's consider the computation of $I = \int_0^1 \varphi(x)\,dx$ (uniform density).

So far, we know two techniques for approximating $I$:

↗ MC: random sequence $X_1, \dots, X_n$ iid $\mathcal{U}(0,1)$

$\hat{I}_n = \frac{1}{n}\sum \varphi(X_i)$

convergence in $O(n^{-1/2})$

↘ Numerical integration: deterministic sequence

Ex: rectangle method $x_1 = \frac{1}{n}, \dots, x_{n-1} = \frac{n-1}{n}, x_n = 1$

$R_n^{(e/r)}$; convergence in $O(n^{-1})$.

↑ Faster than MC, but going from $n$ to $(n+1)$ points is inefficient; as we need to compute $\varphi(\frac{i}{n+1})$; and we cannot make an explicit use of $R_n^{(e/r)}$ to compute $R_{n+1}^{(e/r)}$; unlike MC methods, which are recursive by nature, $\hat{I}_{n+1} = \frac{n}{n+1}\hat{I}_n + \frac{1}{n+1}\varphi(X_{n+1})$.

QMC techniques are a compromise between MC & numerical integration methods = they use deterministic sequences, acting "like" random sequences, and achieving faster rates of convergence than the traditional MC techniques.

Definition: Let $\{\xi_n\}$ be a sequence of $[0,1]^d$.
The discrepancy of $\{\xi_n\}$ is

$$D_n^*(\xi) = \sup_{B \in R^*} |\lambda_n(B) - \lambda(B)|$$
$$= \sup_{B \in R^*} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_B(\xi_i) - \lambda(B) \right|$$

$R^* = \{B \mid B = [0,u_1] \times \dots \times [0,u_d],\ 0 \le u_j \le 1\}$

Lebesgue measure

25

Ex: $d=1$

$$D_n^*(\xi) = \sup_{0\le u\le 1} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{[0,u]}(\xi_i) - u \right|$$

$\{\xi_i\}$ is a deterministic sequence acting as the uniform distribution on $[0,1]$. In probabilistic terms, think SLLN = as $n\to\infty$, expect $\frac{1}{n}\sum \mathbb{1}_{[0,u]}(\xi_i)$ to converge to $u$. The faster the convergence, the more "uniform" the sequence.

Ex: van der Corput sequence has $D_n^*(\xi) = O\left(\frac{\log n}{n}\right)$

. Halton sequence in dimension $d$ has $D_n^*(\xi) = O\left(\frac{(\log n)^d}{n}\right)$.

**Definition**: Hardy-Krause variation.

Let $\varphi: [0,1]^d \to \mathbb{R}$ of class $\mathcal{C}^d$. The Hardy-Krause variation of $\varphi$ is defined as

$$V(\varphi) = \sum_{j=1}^d \sum_{i_1<\dots<i_j} \int_{[0,1]^j} \left| \frac{\partial^j \varphi}{\partial x_{i_1}\dots\partial x_{i_j}} (x(i_1,\dots,i_j)) \, dx_{i_1}\dots dx_{i_j} \right|$$

All coordinates equal to 1 except those located at $i_1,\dots,i_j$; equal to $x_{i_1},\dots,x_{i_j}$.

Ex: • $d=1$, $V(\varphi) = \int_0^1 |\varphi'(x)|\,dx$

• $d=2$, $V(\varphi) = \int_0^1 \left| \frac{\partial \varphi}{\partial x_1}(x_1,1)\,dx_1 \right| + \int_0^1 \left| \frac{\partial\varphi}{\partial x_2}(1,x_2)\,dx_2 \right| + \iint \left| \frac{\partial^2\varphi}{\partial x_1 \partial x_2}(x_1,x_2)\,dx_1\,dx_2 \right|$

gets complicated quickly

26

**Theorem**: Koksma-Hlawka inequality:

$\forall \varphi: [0,1]^d \to \mathbb{R}$, $\forall$ sequence $\{\xi_n\}$ of $[0,1]^d$, we have

$$\left| \frac{1}{n}\sum_{i=1}^n \varphi(\xi_i) - \int_{[0,1]^d} \varphi(x)\,dx \right| \le V(\varphi) \times D_n^*(\xi)$$

the better the approximation,
- the less $\varphi$ varies
- the more uniform the sequence $\{\xi_n\}$.

It all makes sense!

the effect of $\varphi$ and $\{\xi_n\}$ are decoupled.

→ For $d=1$, we know that $D_n^*(\xi)$ cannot be smaller than $O\left(\frac{\log n}{n}\right)$

→ In dimension $d\ge 2$, we believe that the best we can do is a discrepancy of order $O\left(\frac{(\log n)^d}{n}\right)$.

Approximation methods based on such sequences are referred to as Quasi Monte-Carlo (QMC).

Ex: Halton, Faure, Sobol, Niederreiter, ...

The QMC estimator is then $\hat{I}_n = \frac{1}{n}\sum_{i=1}^n \varphi(\xi_i)$.

$\hat{I}_n$ converges to $I$ in $O\left(\frac{(\log n)^d}{n}\right)$.

OK if $d$ is not too large

Compare with the MC rate $O(1/\sqrt{n})$

& numerical techniques in $O(n^{-s/d})$ for $\varphi \in \mathcal{C}^s([a,b]^d)$.

OK if $d$ is large

OK if $d$ is small, & $\varphi$ quite regular