Stochastic Methods for Optimization and Sampling

M1, TSE

Sampling methods

Sampling

What is sampling?

If μ is a given (known) probability density over $\mathbb{R}$ or $\mathbb{R}^d$, how can we generate n iid copies $X_1 \dots X_n \sim \mu$?

We will see many techniques, among which

1. The inversion of the cdf
2. The rejection method
3. Monte-Carlo method.

Later on, we will see how to generate samples that are no longer iid, but that approximate μ well anyway.

1. Metropolis-Hastings,
2. Gibbs sampler.

Simulating Unif([0, 1])

In this course, we will assume that we *know* how to simulate n iid copies from the distribution $U([0,1])$.

This problem is actually more challenging than it seems.

Computers generally generate pseudo-random numbers, by defining sequences by induction.

Historically, the first pseudo-random number generator was

$$X_{n+1} = 16807 X_n \pmod{2^{31} - 1},$$

where X_0 (called "seed") is chosen by the practitioner.

One can however introduce true randomness: time, quantum phenomena etc.

Simulating Bernoulli random variables

Suppose we can simulate $U \sim \mathrm{Unif}([0,1])$.

How to simulate $\mathrm{Ber}(p)$ for $p \in [0,1]$?

Take $U \sim \mathrm{Unif}([0,1])$ and output $Y = \mathbb{1}_{U \leq p}$.

Indeed, $Y \in \{0,1\}$ and $\mathbb{P}(Y = 1) = \mathbb{P}(U \leq p) = p$.

Simulating Binomial random variables

Suppose we can simulate $U_1, \ldots, U_n \overset{iid}{\sim} \text{Unif}([0,1])$.

How to simulate $\text{Bin}(n,p)$ for $p \in [0,1]$?

We know that a binomial random variable with law $\text{Bin}(n,p)$ can be written as a sum of n iid $\text{Ber}(p)$ random variables.

It suffices to generate $U_1, \ldots, U_n \overset{iid}{\sim} \text{Unif}([0,1])$ and output $Y = \sum_{i=1}^{n} \mathbb{1}_{U_i \leq p}$.

Simulating a discrete random variable

Let $d \in \mathbb{N}$ and $p_1, \ldots, p_d \geq 0$ such that $p_1 + \cdots + p_d = 1$.

How to simulate a random variable X that takes value $j \in \{1, \ldots, d\}$ with probability p_j?

$$\mathbb{P}(X = j) = p_j, \quad \forall j \in \{1, \ldots, d\}.$$

Let $s_k = p_1 + \cdots + p_k$ for any $k \in \{1, \ldots, d\}$, and $s_0 = 0$.

Let $U \sim \text{Unif}([0,1])$, and output $Y = \sum_{j=1}^{d} j \mathbb{1}_{U \in [s_{j-1}, s_j)}$.

Indeed, $\mathbb{P}(Y = j) = \mathbb{P}(U \in [s_{j-1}, s_j)) = \int_{s_{j-1}}^{s_j} du = p_j$, since $s_j - s_{j-1} = p_j$.

A general technique

A general technique

If X has the same distribution as $f(U_1, \ldots, U_n)$ where $U_1, \ldots, U_n \overset{iid}{\sim} \text{Unif}([0,1])$, then it suffices to output $f(U_1, \ldots, U_n)$ to simulate X.

Simulating uniform random variables on an interval

Let $a \leq b$ be two real numbers. How to simulate $\mathrm{Unif}([a,b])$?

If $U \sim \mathrm{Unif}([0,1])$, then $a + (b-a)U$ has distribution $\mathrm{Unif}([a,b])$.

It suffices to output $a + (b-a)U$.

Indeed, the probability density of $\mathrm{Unif}([a,b])$ is

$$f_{\mathrm{Unif}([a,b])}(t) = \frac{1}{b-a} \mathbb{1}_{t \in [a,b]}.$$

Moreover, if $t \in [a,b]$, then

$$\mathbb{P}(a + (b-a)U \leq t) = \mathbb{P}\left(U \leq \frac{t-a}{b-a}\right) = \frac{t-a}{b-a} = \int_{-\infty}^{t} \frac{1}{b-a} \mathbb{1}_{x \in [a,b]} dx.$$

Example

How to simulate a random variable with density $f(x) = 2x\,\mathbb{1}_{x\in[0,1]}$?

Show that $\max(U_1, U_2)$ has density $2x\,1_{x\in[0,1]}$ by computing its cdf.

For any $t \in \mathbb{R}$:

$$\begin{aligned}\mathbb{P}\left(\max(U_1, U_2) \leq t\right) &= \mathbb{P}\left(U_1 \leq t \text{ and } U_1 \leq t\right)\\ &= \mathbb{P}\left(U_1 \leq t\right)\mathbb{P}\left(U_2 \leq t\right)\\ &= t^2\\ &= \int_0^t 2x\,\mathbb{1}_{x\in[0,1]}dx.\end{aligned}$$

This ensures that the density of $\max(U_1, U_2)$ is $2x\,\mathbb{1}_{x\in[0,1]}$.

Law of a random variable

A real-valued random variable X has density p with respect to the Lebesgue measure, if, and only if, for any continuous and bounded function $g : \mathbb{R} \longrightarrow \mathbb{R}$, we have

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x)p(x)dx.$$

Take-away: To identify the density p of X (if it exists), it suffices to show that, for any continuous function g, we have

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x)p(x)dx$$

Change of variable

To do so, the change of variable formula is often useful.

Change of variable formula

Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be a continuous function and $\phi : [a, b] \longrightarrow \mathbb{R}$ be C^1 and strictly monotonous. We have

$$\int_a^b f(\phi(t))\,\phi'(t)\,dt = \int_{\phi(a)}^{\phi(b)} f(u)du.$$

Informally, we say that we perform the change of variables $x = \phi(t)$, associated with $dx = \phi'(t)dt$.

Examples

What density do we sample from when we output U^2 for $U \sim \text{Unif}([0,1])$?

It suffices to compute the density of U^2.

$$\begin{aligned}
\mathbb{E}(g(U^2)) &= \int_0^1 g(u^2)du \\
&= \int_0^1 g(v)\frac{1}{2\sqrt{v}}dv \qquad \text{using } u = \sqrt{v}, du = \frac{1}{2\sqrt{v}}dv \\
&= \mathbb{E}(g(V)), \qquad \text{where } V \text{ has density } \frac{1}{2\sqrt{x}}\mathbb{1}_{x\in[0,1]}.
\end{aligned}$$

To sample from the probability density $\frac{1}{2\sqrt{x}}\mathbb{1}_{x\in[0,1]}$, it suffices to output U^2.

Examples

How to sample from the probability density $\alpha x^{\alpha-1}\mathbb{1}_{x\in[0,1]}$ for $\alpha \in (0,1)$?

Let's find a variable V such that for any continuous g, we have $\mathbb{E}g(V) = \int_{\mathbb{R}} g(x)\alpha x^{\alpha-1}\mathbb{1}_{x\in[0,1]}dx$.

We use the change of variable $y = x^\alpha$, $dy = \alpha x^{\alpha-1}dx$.

$$\begin{aligned}\int_{\mathbb{R}} g(x)\alpha x^{\alpha-1}\mathbb{1}_{x\in[0,1]}dx &= \int_{\mathbb{R}} g(y^{1/\alpha})\mathbb{1}_{y\in[0,1]}dy \\ &= \mathbb{E}g(U^{1/\alpha}) \quad \text{where } U \sim \text{Unif}([0,1]).\end{aligned}$$

It suffices to output $U^{1/\alpha}$.

Note that for $\alpha = 2$, we recover the density $2x\mathbb{1}_{x\in[0,1]}$ covered above, meaning that $\sqrt{U}$ has the same law as $\max(U_1, U_2)$!

Mixtures of distributions

How to simulate from density $\frac{1}{2}f_0 + \frac{1}{2}f_1$, where f_1, f_2 are densities over $\mathbb{R}$?

Let $i \sim \text{Ber}(1/2)$ and, conditionally on i, generate $X \sim f_i$ and output X.

More generally, how to simulate from $\alpha_1 f_1 + \cdots + \alpha_k f_k$, where $\alpha_1 + \ldots, +\alpha_k = 1$ and $\alpha_j \geq 0$ for any j?

Simulate a discrete random variable J over $\{1, \ldots, k\}$ such that $\mathbb{P}(J = j) = \alpha_j$ for any j.

Conditionally on J, generate $X \sim f_J$ and output X.

Inversion of the cdf

Method: Inversion of the cdf

Definition (Cumulative distribution function (cdf))

Let μ be a probability measure over $\mathbb{R}$. The cumulative distribution function (cdf) of μ, denoted by F_μ, is

$$F_\mu(x) = \mu\big((-\infty, x]\big) = \mathbb{P}(X \leq x) \quad \text{if } X \sim \mu.$$

Definition (Pseudo-inverse)

Let F_μ be the cdf of some probability measure μ. We define the pseudo-inverse G_μ of F_μ by letting, for any $t \in (0, 1)$,

$$G_\mu(t) = \inf\{x \in \mathbb{R} : F_\mu(x) \geq t\}.$$

Some properties

Proposition

1. For any $t \in [0,1]$ and $x \in \mathbb{R}$, we have $G_\mu(t) \leq x \Leftrightarrow F_\mu(x) \geq t$.
2. If the restriction of F_μ to (a,b) is bijective from (a,b) to $(0,1)$, of inverse F_μ^{-1}, then $G_\mu = F_\mu^{-1}$.

Theorem

Let $U \sim \text{Unif}([0,1])$, then $G_\mu(U)$ has distribution μ.

Proof: Since $U \in (0,1)$ a.s., $G(U)$ is well-defined. Let's compute the cdf of $G(U)$

$$\mathbb{P}(G(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$$

Since the cdf characterizes the law, $G(U)$ has distribution μ.

Methods of the inversion of cdf

Methods of the inversion of cdf

Let μ be a probability distribution over $\mathbb{R}$.

1. Compute its cdf F_μ
2. Compute its pseudo-inverse G_μ
3. output $G_\mu(U)$ where $U \sim \text{Unif}([0,1])$.

Advantage: Very general (true for any μ over $\mathbb{R}$).

Drawback: Computing F and G can be difficult.

Simulating exponential random variables

We recall that the exponential distribution with parameter $\lambda > 0$, denoted as $\mathrm{Exp}(\lambda)$, has density $f(x) = \mathbf{1}_{x \geq 0} \lambda e^{-\lambda x}$ over $\mathbb{R}$.

Proposition

Let $U \sim \mathrm{Unif}([0,1])$, then $-\frac{1}{\lambda}\log(U) \sim \mathrm{Exp}(\lambda)$

Rejection method

Rejection method

Goal: Simulate X with density f, but f is "too complicated".

Assumption: We know an auxiliary density g such that

1. It is easy to simulate Y with density g
2. There exists a constant $c > 0$ such that $f \leq cg$.
3. For any $x \in \mathbb{R}^d$, it is easy to evaluate $\dfrac{f(x)}{cg(x)}$.

Rejection method

Suppose $(Y_n)_n \overset{iid}{\sim} g$ and $(U_n)_n \overset{iid}{\sim} \text{Unif}([0,1])$ are independent.

Let $N = \min\left\{n \geq 0 : \; U_n \leq \dfrac{f(Y_n)}{cg(Y_n)}\right\}$.

Output Y_N.

Rejection method

Theorem

Let $(Y_n)_n \overset{iid}{\sim} g$ and $(U_n)_n \overset{iid}{\sim} \text{Unif}([0,1])$ be independent, and define

$$N = \min\left\{ n \geq 0 \;\middle|\; U_n \leq r(Y_n) \overset{\text{def}}{=} \frac{f(Y_n)}{cg(Y_n)} \right\}.$$

Then

1. The density of Y_N is f,
2. $N \sim \text{Geom}(1/c)$ and N is independent of Y_N.

Remark: In the theorem above, N is a random variable!

A random variable N has distribution $\text{Geom}(p)$ if for any $n \in \{1, 2, \dots\}$ we have $\mathbb{P}(N = n) = p(1-p)^n$, or equivalently

$$\forall n \in \{1, 2, \dots\} : P(N > n) = (1-p)^n.$$

Application of the rejection method

Let μ be a probability measure over $\mathbb{R}$, and A be a non-zero probability event for measure μ.

How to simulate from the probability measure $\mathbb{P}_\mu(\cdot|A)$ (probability measure conditionally on A)?

Keep on generating $X_1, \ldots, X_n \sim \mu$ iid until event A is satisfied.

Application: Conditional probability measure

More precisely: how to sample from $\mathbb{P}(\cdot|A)$?

Rejection method: Let g be the density of $\mathbb{P}$ and f that of $\mathbb{P}(\cdot|A)$.

Then $f(x) = \dfrac{g(x)}{\mathbb{P}_\mu(A)} \mathbb{1}_{x \in A} \leq c g(x)$ where $c = 1/\mathbb{P}_\mu(A)$.

Let $(X_n)_n \sim g$ iid and $(U_n)_n \sim \mathrm{Unif}([0,1])$ iid be independent.

- If $X_k \notin A$, then $r(X_k) = 0$, and $\mathbb{P}(U_k \leq r(X_n)) = 0$.
- If $X_k \in A$, then $r(X_k) = 1$, and $\mathbb{P}(U_k \leq r(X_n)) = 1$.

Hence, N is the first time that $X_n \in A$.

Keep on generating $X_1, \ldots, X_n \sim \mu$ iid until event A is satisfied.

Example: Simulating random variables in the unit disk

Sampling Unif r.v. in the unit disk (density $f(x) = \frac{1}{\pi}\mathbb{1}\{x^2+y^2 \leq 1\}$)?

Keep on sampling $X, Y \sim \text{Unif}([-1,1])$ iid until $X^2 + Y^2 \leq 1$. Why?

Let g be the density of $\text{Unif}[-1,1]^2$: $g(x,y) = \frac{1}{4}\mathbb{1}\{(x,y) \in [-1,1]^2\}$.

We have $f \leq \frac{4}{\pi}g$. Therefore, $c = 4/\pi$.

If $U \sim \text{Unif}([0,1])$, then for any $X_i, Y_i \sim \text{Unif}([-1,1])$:

- If $X_i^2 + Y_i^2 \leq 1$, then $\frac{f(X_i,Y_i)}{cg(X_i,Y_i)} = 1$ and $\mathbb{P}(U \leq r(X_i,Y_i)) = 1$,
- Otherwise, $r(X_i,Y_i) = \frac{f(X_i,Y_i)}{cg(X_i,Y_i)} = 0$ and $\mathbb{P}(U \leq r(X_i,Y_i)) = 0$.

Simulating normal random variables

Application: Simulating normal random variables

How to simulate $N(0,1)$, with density $f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$?

Take g as the density of a Laplace distribution with parameter $\lambda > 0$:

$$g(x) = \frac{\lambda}{2} \exp\left(-\lambda|x|\right).$$

- We can simulate g by simulating $\mathrm{Exp}(\lambda) \times \mathrm{Rad}(1/2)$.
- We have $\dfrac{f(x)}{g(x)} = \sqrt{\dfrac{2}{\pi}} \dfrac{1}{\lambda} \exp\left(\lambda|x| - \dfrac{x^2}{2}\right) \leq c \stackrel{\text{def}}{=} \sqrt{\dfrac{2}{\pi}} \dfrac{e^{\lambda^2/2}}{\lambda}$.

  (The maximum is attained for $x = \pm\lambda$)

  Taking $\lambda = 1$ gives the smallest value of $c \approx 1.31$.

We can use the rejection method with the density g of Laplace(1).

Box-Muller method

Proposition

If $U, V \sim \text{Unif}([0,1])$ are independent, then

$$\sqrt{-2\log(U)}\cos(2\pi V) \text{ and } \sqrt{-2\log(U)}\sin(2\pi V)$$

are $N(0,1)$ and independent.

A useful result: polar change of variables

To prove the proposition above, we will need the following result.

Proposition (Polar change of variables)

For any integrable function $f : \mathbb{R}^2 \to \mathbb{R}$, we have

$$\int_{\mathbb{R}^2} f(x,y)dxdy = \int_{\mathbb{R}_+ \times [0,2\pi)]} f\big(r\cos(\theta), r\sin(\theta)\big) r\, dr\, d\theta.$$

This change of variable is performed by replacing any occurrence of x by $r\cos(\theta)$ and any occurrence of y by $r\sin(\theta)$ and the differential $dxdy$ by $r\,dr\,d\theta$.

Proof: Box-Muller method

Proof The density of $(X, Y) \overset{iid}{\sim} N(0,1)$ is $f(x,y) = \frac{1}{2\pi} \exp\left(-\frac{x^2+y^2}{2}\right)$.

Polar change of variable:

$$(x, y) = (r\cos(\theta), r\sin(\theta)) \quad \text{and} \quad dxdy = r\,dr\,d\theta.$$

For any C^0, bounded function φ:

$$\begin{aligned}
\mathbb{E}(\varphi(X,Y)) &= \int \varphi(x,y) f(x,y) dxdy \\
&= \int_{\mathbb{R}_+ \times [0,2\pi)} \varphi(r\cos(\theta), r\sin(\theta)) \frac{1}{2\pi} \exp\left(-\frac{r^2}{2}\right) r\,dr\,d\theta \\
&= \int_{\mathbb{R}_+ \times [0,1)} \varphi(\sqrt{2s}\cos(2\pi\alpha), \sqrt{2s}\sin(2\pi\alpha)) \exp(-s)\, ds\, d\alpha
\end{aligned}$$

In the last line, we used the change of variable $s = r^2/2$ and $\alpha = \theta/2\pi$.

Proof (continued)

Therefore,

$$\mathbb{E}(\varphi(X,Y)) = \mathbb{E}\Big[\varphi(\sqrt{2E}\cos(2\pi V), \sqrt{2E}\sin(2\pi V))\Big]$$

where

$$E \sim \text{Exp}(1) \quad \text{and} \quad V \sim \text{Unif}([0,1]).$$

To simulate $E \sim \text{Exp}(1)$, just output $-\log(U)$ for $U \sim \text{Unif}([0,1])$.

Conclusion:

$$(\sqrt{-2\log(U)}\cos(2\pi V), \sqrt{-2\log(U)}\sin(2\pi V)) \overset{\text{law}}{=} (X,Y)$$

where $X, Y \overset{iid}{\sim} N(0,1)$.

Normal vectors

Definition

Gaussian vector A vector $X = (X_1, \ldots, X_d)$ is a Gaussian vector if any linear combination of its components is a Gaussian random variable.

In particular, each component X_i must be a Gaussian r.v. (why?)

Example (Gaussian marginals but not a Gaussian vector)

Let (X, bX) for $X \sim N(0,1)$ and $b \sim \text{Rad}(1/2)$ such that $b \perp\!\!\!\perp X$.

It is *not* a Gaussian vector since the sum of the two coordinates is $(1+b)X \sim \frac{1}{2}\delta_0 + \frac{1}{2}N(0,4)$ is not a Gaussian r.v.

Simulating Gaussian vectors

A Gaussian vector X is characterized by its mean $\mu = \mathbb{E}X$ and covariance matrix $\Sigma = \mathbb{E}XX^\top - \mathbb{E}X\mathbb{E}X^\top$. We write $X \sim N(\mu, \Sigma)$.

The matrix Σ is necessarily symmetric and non-negative. It is possible to find a matrix A such that $\Sigma = AA^\top$ (Cholesky decomposition).

To simulate $Y \sim N(\mu, \Sigma)$, one can generate $X_1, \ldots, X_d \overset{iid}{\sim} N(0, 1)$ and output

$$Y = \mu + AX$$

Monte-Carlo method

Monte-Carlo method

Standard method to approximate expectations/integrals.

Let $X \in \mathbb{R}^d$ be an integrable random variable (i.e. $\mathbb{E}\|X\| < \infty$).

How to *approximate* $\mathbb{E}X$?

Simulate $X_1, \dots, X_n$ iid with the same distribution as X, and output

$$\frac{1}{n}\sum_{i=1}^{n} X_i.$$

Indeed, by the strong law of large numbers, we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mathbb{E}[X] \quad a.s. \qquad \text{as } n \to \infty.$$

Applications

Let $\varphi : \mathbb{R}^d \to \mathbb{R}^p$ such that $\varphi(X)$ is integrable (i.e. $\mathbb{E}\|\varphi(X)\| < \infty$).

How to approximate $\mathbb{E}\varphi(X)$?

Generate $X_1, \ldots, X_n$ iid with the same law as X and output

$$\frac{1}{n}\sum_{i=1}^{n} \varphi(X_i).$$

Let A be an event. How to approximate $\mathbb{P}(X \in A)$?

Generate $X_1, \ldots, X_n$ iid with the same law as X and output

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \in A} = \frac{1}{n}\,\mathrm{Card}\{i \in \{1, \ldots, n\} : X_i \in A\}.$$

Intuition: $\mathbb{P}(X \in A)$ is the proportion of X_i's falling in A when drawing X_i's a large number of times.

Confidence interval

We have seen that $\frac{1}{n}\sum_{i=1}^{n} Z_i \longrightarrow \mathbb{E}[Z]$ *a.s.* as $n \to \infty$.

What about the *error* $\left|\sum_{i=1}^{n} Z_i - \mathbb{E}[Z]\right|$?

Theorem (Central Limit Theorem)

If $Z_1, \ldots, Z_n \in \mathbb{R}$ are iid and $\mathbb{E}Z_i^2 < \infty$, then the mean $\mu = \mathbb{E}Z$ and variance $\sigma^2 = \mathbb{V}Z$ exist, and

$$\sqrt{n}\left(\frac{\sum_{i=1}^{n} Z_i}{n} - \mu\right) \xrightarrow{Law} N(0, \sigma^2).$$

Confidence interval

Suppose we have

1. a family of probability distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$, for $\Theta \subseteq \mathbb{R}$,
2. a real number $\alpha \in (0, 1)$,
3. a sample $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}_\theta$ for some unknown $\theta \in \Theta$.

Definition (Confidence interval)

A confidence interval for θ with confidence level $1 - \alpha$ is a random interval $\big[a(X_1, \ldots, X_n),\, b(X_1, \ldots, X_n)\big]$ s.t.

$$\mathbb{P}_\theta\Big(\big[a(X_1, \ldots, X_n),\, b(X_1, \ldots, X_n)\big] \ni \theta\Big) = 1 - \alpha,$$

where $a(X_1, \ldots, X_n)$ and $b(X_1, \ldots, X_n)$ only depend on the observations $(X_1, \ldots, X_n)$.

Asymptotic confidence interval

Definition (Asymptotic confidence interval)

An asymptotic confidence interval for θ with asymptotic confidence level $1-\alpha$ is a random interval $\big[a(X_1,\dots,X_n),\, b(X_1,\dots,X_n)\big]$ s.t.

$$\mathbb{P}_\theta\Big(\big[a(X_1,\dots,X_n),\, b(X_1,\dots,X_n)\big] \ni \theta\Big) \underset{n\to\infty}{\longrightarrow} 1-\alpha,$$

where $a(X_1,\dots,X_n)$ and $b(X_1,\dots,X_n)$ only depend on the observations $(X_1,\dots,X_n)$.

Asymptotic confidence interval for Monte-Carlo

Suppose we would like to approximate $\mathbb{E}[\varphi(X)]$ and obtain a confidence interval for it.

We have $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(X_i) - \mathbb{E}\varphi(X)\right) \overset{\text{law}}{\to} N(0,\sigma^2)$ for $\sigma^2 = \mathbb{V}(\varphi(X))$.

Thus, a natural asymptotic confidence interval of level $1-\alpha$ is

$$\left[\frac{1}{n}\sum_{i=1}^{n}\varphi(X_i) - \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}\,,\ \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i) + \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right]$$

if $\sigma^2 = \mathbb{V}(\varphi(X))$ is known.

Asymptotic confidence interval – unknown variance

Suppose now that σ is unknown. We cannot directly obtain a confidence interval like above, but we can estimate σ:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\varphi(X_i)\right)^2 \longrightarrow \sigma^2 \quad a.s.$$

Conclusion: An asymptotic confidence interval for $\mathbb{E}\varphi(X)$ of asymptotic level $1-\alpha$ is

$$\left[\frac{1}{n}\sum_{i=1}^{n}\varphi(X_i) - \frac{\hat{\sigma}q_{1-\frac{\alpha}{2}}}{\sqrt{n}}\,,\ \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i) + \frac{\hat{\sigma}q_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right]$$

Importance sampling

Importance sampling

Let X be a random variable in $\mathbb{R}^d$ with density $f : \mathbb{R}^d \to \mathbb{R}$.

Goal: Approximate $\mathbb{E}[\varphi(X)] = \int_{\mathbb{R}^d} \varphi(x)\, f(x)\, dx$. Using Monte Carlo:

$$\frac{1}{n}\sum_{i=1}^{n} \varphi(X_i) \longrightarrow \mathbb{E}[\varphi(X)] \;\; a.s. \quad \text{if} \;\; X_1, \ldots, X_n \overset{iid}{\sim} f.$$

Problem: f can be hard to sample from: Obtaining $X_1, \ldots, X_n \overset{iid}{\sim} f$ is expensive. However, we assume we know a "nice" density g s.t.

- g is easy to sample from
- $\forall x \in \mathbb{R}^d$, the ratio $\frac{f(x)}{g(x)}$ exists and is easy to evaluate.

Importance sampling

Idea: Rewrite $\mathbb{E}\left[\varphi(X)\right]$ as an integral with respect to g

$$\mathbb{E}\left[\varphi(X)\right] = \int_{\mathbb{R}^d} \varphi(x)\, f(x)\, dx = \int_{\mathbb{R}^d} \varphi(x)\, \frac{f(x)}{g(x)}\, g(x)\, dx$$

$$= \mathbb{E}\left[\varphi(Y)\, \frac{f(Y)}{g(Y)}\right] \qquad \text{where } Y \sim g.$$

It now suffices to draw $Y_1, \ldots Y_n \overset{iid}{\sim} g$ (which is easy), and to output

$$\frac{1}{n} \sum_{i=1}^{n} \varphi(Y_i)\, \frac{f(Y_i)}{g(Y_i)} \longrightarrow \mathbb{E}[\varphi(X)] \quad a.s.$$

by the Law of Large Numbers.

Summary: Importance sampling method

Importance sampling (IS)

Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a function and $f, g : \mathbb{R}^d \to \mathbb{R}$ be probability densities such that

- g is easy to sample from
- $\forall x \in \mathbb{R}^d$, the ratio $\frac{f(x)}{g(x)}$ exists and is easy to evaluate.

Then the quantity $\mathbb{E}[\varphi(X)]$ can be approximated by

$$\frac{1}{n} \sum_{i=1}^{n} \varphi(Y_i) \, \frac{f(Y_i)}{g(Y_i)} \qquad \text{where } Y_1, \ldots, Y_n \overset{iid}{\sim} g.$$

Simulating rare events: MC vs IS

The MC method is limited if φ takes large values on rare events for f.

Example: Let $X \sim N(0,1)$ and suppose we want to approximate $\mathbb{P}(X > 6)$ using the Monte-Carlo method. We have

$$\mathbb{P}(X > 6) = \mathbb{E}\big[\mathbf{1}_{X>6}\big].$$

Letting $X_1, \ldots, X_n \overset{iid}{\sim} N(0,1)$, we should output $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \mathbf{1}_{X_i>6}$.

Problem: The summands $\mathbf{1}_{X_i>6}$ take the value 1 on a rare event:

$$\mathbb{P}(X > 6) \approx 10^{-9}.$$

Unless $n \gtrsim 10^9$, we most likely have only zeros in the sum above!

Simulating rare events: MC vs IS

To address this issue, let's use importance sampling.

Let g be the density of $N(7,1)$ and $Y_1, \ldots, Y_n \sim g$ iid and output

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i>6\}} \frac{f(Y_i)}{g(Y_i)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i>6\}} \exp\left(\frac{(Y_i-7)^2 - Y_i^2}{2}\right).$$

We have $\mathbb{P}(Y > 6) = \mathbb{P}(N(7,1) > 6) = \mathbb{P}(N(0,1) > -1) \approx 84\%$. Thus

- We have $Y_i > 6$ for around 84% of the Y_i's.
- We have $Y_i \leq 6$ for around 16% of the Y_i's.

Both events $(Y \leq 6)$ and $(Y > 6)$ are frequently observed under g.

This will work much better!

Variance reduction

Importance sampling also allows for variance reduction.

Using Monte-Carlo directly with $X_1, \dots, X_n \overset{iid}{\sim} f$, the CLT would yields

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(X_i) - \mathbb{E}\big[\varphi(X)\big]\right) \xrightarrow{\text{law}} N(0, \sigma^2),$$

where $\sigma^2 = \mathbb{V}(\varphi(X))$.

Using importance sampling with $Y_1, \dots, Y_n \sim g$ iid, the CLT gives

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\varphi(Y_i)f(Y_i)}{g(Y_i)} - \mathbb{E}\big[\varphi(X)\big]\right) \xrightarrow{\text{law}} N(0, s^2),$$

where $s^2 = \mathbb{V}\left[\dfrac{\varphi(Y)f(Y)}{g(Y)}\right]$. One can choose g to minimize s.

Variance reduction

Theorem

Consider

- Two probability densities f, g over $\mathbb{R}^d$,
- A function $\varphi : \mathbb{R}^d \to \mathbb{R}$,
- And define $w(x) = \frac{f(x)}{g(x)}$ for any $x \in \mathbb{R}^d$.

Let $X \sim f$ and $Y \sim g$, and assume that $\mathbb{E}\big[w^2(Y)\varphi^2(Y)\big] < \infty$. Then

$$s^2 = \mathbb{V}\big(w(Y)\,\varphi(Y)\big) \geq \mathbb{E}^2\big[\big|\varphi(X)\big|\big] - \mathbb{E}^2\big[\varphi(X)\big].$$

The function g^* minimizing s is

$$g^*(y) = \frac{\big|\varphi(y)\big|\, f(y)}{\displaystyle\int_{\mathbb{R}^d} \big|\varphi(x)\big|\, f(x)\, dx} = \frac{\big|\varphi(y)\big|\, f(y)}{\mathbb{E}|\varphi(X)|}.$$

Proof of the Theorem

We have

$$\begin{aligned} s^2 &= \mathbb{E}\big[w^2(Y)\varphi^2(Y)\big] - \mathbb{E}^2\big[\varphi(X)\big] \\ &= \mathbb{E}\big[w^2(Y)|\varphi|^2(Y)\big] - \mathbb{E}^2\big[\varphi(X)\big] \\ &\geq \mathbb{E}^2\big[w(Y)\big|\varphi(Y)\big|\big] - \mathbb{E}^2\big[\varphi(X)\big] \qquad \text{(Jensen)} \\ &= \mathbb{E}^2\big[|\varphi(X)|\big] - \mathbb{E}^2\big[\varphi(X)\big]. \end{aligned}$$

(Here we could have used Jensen, Cauchy-Schwarz, or $\mathbb{V}Z \geq 0$).

We have equality in $\mathbb{V}Z = 0$ iff Z is *a.s.* constant. Hence

$$\exists C > 0 : w(Y)\big|\varphi(Y)\big| = C \quad \Longleftrightarrow \quad g(Y) = \frac{f(Y)\big|\varphi(Y)\big|}{C},$$

which imposes $g(y) = \dfrac{f(y)\big|\varphi(y)\big|}{\displaystyle\int \big|\varphi(y)\big| f(y)dy}$ since $\int g$ must be 1.

Variance reduction: remarks

1. In the above theorem, s^2 is the asymptotic variance we have when using importance sampling.

2. The theorem above says that s can be reduced, but that the best one can hope for is $(s^*)^2 = \mathbb{E}^2\big[\big|\varphi(X)\big|\big] - \mathbb{E}^2\big[\varphi(X)\big]$.

3. Assuming further that $\varphi \geq 0$, we obtain $s^* = 0$: If we sample *one* observation from $g^*(y) = \dfrac{f(y)\varphi(y)}{\mathbb{E}[\varphi(X)]}$, the IS estimator is

$$\sum_{i=1}^{1} \frac{f(Y_1)\varphi(Y_1)}{g^*(Y_1)} = \mathbb{E}[\varphi(X)].$$

   We can estimate our target $\mathbb{E}[\varphi(X)]$ with one observation only!

Variance reduction: remarks

- Unfortunately, the optimal function $g^*(y) = \dfrac{|\varphi(y)|\, f(y)}{\mathbb{E}|\varphi(Y)|}$ involves the quantity $\mathbb{E}|\varphi(Y)|$ we are trying to estimate!

  It is therefore impossible to use g^* in practice.

- Moreover, even if g^* were known, nothing guarantees that g^* is easy to sample from.

- The advantage of this theorem is to show that the optimal g^* should take large values where $|\varphi| f$ is large.

Number of Iterations vs Runtime

For the same n, IS is more precise than MC if $\frac{s}{\sqrt{n}} < \frac{\sigma}{\sqrt{n}}$, i.e. $s < \sigma$.

Regarding *runtime*, computing $\varphi(Y_i)\frac{f(Y_i)}{g(Y_i)}$ can be slow for 2 reasons:

1. g may be difficult to sample from
2. The ratio $\frac{f(Y_i)}{g(Y_i)}$ can be slow to evaluate.

Previously, we ruled out such issues but they can arise in practice.

If evaluating $\varphi(X_i)$ takes 1s and $\varphi(Y_i)\frac{f(Y_i)}{g(Y_i)}$ takes τ seconds, we reach precision ϵ in

1. $t = n = O(\frac{\sigma^2}{\epsilon^2})$ seconds using MC,
2. $t = n\tau = O(\frac{\tau s^2}{\epsilon^2})$ seconds using IS.

IS is preferable over MC if $\tau s^2 < \sigma^2$, not just if $s^2 < \sigma^2$!