Computational Statistics
M.Sc. Mathematical Engineering
Prof. A. Manzoni

A.A. 2021 – 2022
Politecnico di Milano

Hands-On 6
Rare Events & Importance Sampling

Why rare events yield inefficient MC simulations?

A remarkable context in which importance sampling is employed is the context of rare events. Suppose we want to estimate $\mu = P(A)$ for some rare event A. Using crude Monte Carlo, we shall simulate the event directly, and get

$$\hat{\mu}_{CMC} = \frac{1}{N}\sum_{i=1}^{N} X^{(i)}$$

where the $X^{(i)}$ are i.i.d. copies of the Bernoulli random variable $X = \mathbf{1}_A$. Note that, in this case,

$$\text{Var}\left[X^{(i)}\right] = \mu(1-\mu).$$

A computable $1-\alpha$ confidence interval should then be

$$\hat{I}_\alpha = \left[\hat{\mu}_{CMC} - z_{1-\alpha/2}\sqrt{\frac{\hat{\mu}_{CMC}(1-\hat{\mu}_{CMC})}{N}}, \hat{\mu}_{CMC} + z_{1-\alpha/2}\sqrt{\frac{\hat{\mu}_{CMC}(1-\hat{\mu}_{CMC})}{N}}\right]$$

for N large enough. The half-size of the Confidence Interval can be interpreted as (a probabilistic bound of) the absolute error of the estimation. If we divide it by the point-estimation, we can interpret the ratio as (a probabilistic bound of) the corresponding relative error

Hence, since $\mu \approx 0$, also $\hat{\mu}_{CMC} \approx 0$, so that the relative error becomes

$$err_{rel} = z_{1-\alpha/2}\sqrt{\frac{\hat{\mu}_{CMC}(1-\hat{\mu}_{CMC})}{N}} \cdot \frac{1}{\hat{\mu}_{CMC}} \approx \frac{1}{\sqrt{N\hat{\mu}_{CMC}}}$$

so that if we use the crude Monte Carlo approach, when the target μ is very small, so that its estimation $\hat{\mu}_{CMC}$ is also very small, the relative error is usually very large, unless N is huge.

To have an idea, if we want to compute a classic 95% confidence for the estimations and have a modest 10% of relative error in the answer, in the case of an event with probability $\mu = 10^{-9}$ would entail that

$$\frac{1}{\sqrt{N \cdot 10^{-9}}} \leq 10^{-1} \qquad \Rightarrow \qquad N \geq 10^{11}.$$

If each instance of the problem needs 1 sec of CPU time for its processing, we would need about 3168 years to estimate μ. Too bad. And if each instance needs just 1 msec (optimistic), we still need 3,168 years to get the answer. We must do something.

The general case

Assume we want to estimate

$$\ell = P_f(X \in A),$$

where X is an $\mathbb{R}$-valued random variable with density f, $A \subset \mathbb{R}$, and P_f is used to denote that the probability is calculated assuming $X \sim f$. Furthermore, suppose that ℓ is small (so the event $\{X \in A\}$ is rare). In this case, the relative error of the standard Monte Carlo estimator will be quite high because many samples are required in order to observe enough occurrences of $\{X \in A\}$.

The idea of importance sampling in this context is to reweight the probabilities of certain events happening so that $\{X \in A\}$ (which is the important event) happens much more often. Obviously, this needs to be done properly or the resulting estimator will be biased.

This is done by introducing a density g such that, when X has density g the event $\{X \in A\}$ is more likely and such that $g(x) = 0 \Rightarrow f(x)\mathbf{1}(X \in A) = 0$. Then, we can write

$$\begin{aligned} \ell = & \quad \mathbb{E}_f[\mathbf{1}(X \in A)] = \int_{\mathbb{R}} \mathbf{1}(X \in A) f(x) dx \\ = & \quad \int_{\mathbb{R}} \mathbf{1}(X \in A) \frac{f(x)}{g(x)} g(x) dx = \mathbb{E}_g \left[\frac{f(x)}{g(x)} \mathbf{1}(X \in A) \right] \end{aligned}$$

where $\mathbb{E}_f[\cdot]$ is calculated with $X \sim f$ and $\mathbb{E}_g[\cdot]$ is calculated with $X \sim g$. The term $f(X)/g(X)$, which reweights things so that everything is unbiased, is called the likelihood ratio.

Using this result, we obtain the estimator

$$\hat{\ell}_{IS} = \frac{1}{N} \sum_{i=1}^{N} \frac{f(X^{(i)})}{g(X^{(i)})} \mathbf{1}(X^{(i)} \in A)$$

where the $\{X^{(i)}\}$ are i.i.d. random variables with density g.

- **A first example**

Suppose we want to estimate

$$\ell = P(X > 4.5)$$

where $X \sim N(0, 1)$ using crude Monte Carlo or importance sampling with a Cauchy distribution. The following `Matlab` code can be used:

```
n=50000; % sample size
x_0 = 4.5; % extreme value

%% Using the Standard Monte Carlo
z_gauss = randn(n,1);
est_MC = sum(z_gauss>x_0)/n;

% Print the estimator
fprintf("\nTrue value, 1-normCDF(x_0): %.4e\n",1-normcdf(x_0))
fprintf("Using Monte Carlo Sampling: %.4e\n\n",est_MC)


%% Using Importance Sampling with a Cauchy distribution
z_cauchy = tan((rand(1,n) - 0.5)*pi);

cauchypdf = @(x) 1/pi./(1+x.^2);
w = normpdf(z_cauchy)./cauchypdf(z_cauchy);

est_IS=sum(w(z_cauchy>x_0))/n;

% Print the estimator
fprintf("\nTrue value, 1-normCDF(x_0): %.4e\n",1-normcdf(x_0))
fprintf("Using  Importance Sampling: %.4e\n\n",est_IS)
```

The difference is clearly visible:

```
True value, 1-normCDF(x_0): 3.3977e-06
Using Monte Carlo Sampling: 0.0000e+00

True value, 1-normCDF(x_0): 3.3977e-06
Using  Importance Sampling: 3.1896e-06
```

Problem 6.1 (Rare events probability / 1).
Suppose we want to estimate

$$\ell = P(X \in [5, \infty))$$

where $X \sim Exp(1)$. In this case, we could make the event of interest more likely by choosing g to be the density of a $Exp(1/5)$ random variable.
Implement in `Matlab` a MC method using importance sampling for the general problem of estimating $\ell = P(X > \gamma)$, using an exponential distribution with $\lambda = 1/\gamma$.

We have $A = [5, \infty)$, $f(x) = e^{-x}$, $g(x) = \frac{1}{5}e^{-\frac{1}{5}x}$, so that

$$\ell = \mathbb{E}_f[\mathbf{1}(X \in A)] = \int_5^\infty e^{-x} dx = e^{-5}$$

and $\mathbb{E}_g[X] = 5$. More in general, suppose that $g(x) = \lambda e^{-\lambda x}$, so that

$$\frac{f(x)}{g(x)} = \frac{1}{\lambda} \exp((\lambda - 1)x)$$

In `Matlab`:

```
N = 10^5 ;gamma = 5; lambda = 1/gamma;
results = zeros(N,1);   results_IS = zeros(N,1);

for i=1:N
    X = -log(rand);
    X_IS = -log(rand) / lambda;
    results(i) = (X > gamma);
results_IS(i) = exp((lambda-1)*X_IS) / lambda * (X_IS > gamma);
end

ell = exp(-5)
   ell = 0.0067

ell_hat = mean(results)
   ell_hat = 0.0070

re_hat = std(results) / (sqrt(N) * ell_hat).  % coefficient of variation
   re_hat = 0.0376

ell_hat_IS = mean(results_IS)
   ell_hat_IS =  0.0068

re_hat_IS = std(results_IS) / (sqrt(N) * ell_hat_IS)   % coefficient of variation
   re_hat_IS = 0.0081

re_hat / re_hat_IS  = 4.6590     % reduction in the coefficient of variation
```

How to systematically generate *good* proposal distributions?

We have seen that, regarding the choice of the distribution g, we can achieve variance reduction as soon as

$$\mathbb{E}_f\left[\frac{f(X)}{g(X)}\Psi(X)^2\right] \leq \mathbb{E}_f[\Psi(X)]^2$$

when we want to estimate $\mathbb{E}_f[\Psi(X)]$. In the special case of rare-event probability estimation, the requirement is that

$$\mathbb{E}_f\left[\frac{f(X)}{g(X)}\mathbf{1}(X \in A)\right] \leq \mathbb{E}_f[\mathbf{1}(X \in A)]$$

which will be satisfied is, for example, g gives a large weight to all x such that $x \in A$. We can event take this further and see that the above condition is equivalent to

$$\mathbb{E}_f\left[\frac{f(X)}{g(X)}\mathbf{1}(X \in A) \mid X \in A\right] P(X \in A) \leq P(X \in A) \quad \Rightarrow \mathbb{E}_f\left[\frac{f(X)}{g(X)}\mathbf{1}(X \in A) \mid X \in A\right] < 1;$$

that is, we require that the mean value of the likelihood ratio on the set A should be less than 1.

We have also seen that, in the general case, if we aim at esimating $\mu = \mathbb{E}[\Psi(X)]$, and $\Psi(X) \geq 0$, the optimal choice g^* is given by

$$g^*(x) = \frac{|\Psi(x)|f(x)}{\mathbb{E}_f[|\Psi(X)|]} = \frac{\Psi(x)f(x)}{\mu}$$

If $\Psi(x) = \mathbf{1}(x \in A)$, we have

$$g^*(x) = \frac{f(x)\mathbf{1}(x \in A)}{P(X \in A)} = f(x \mid X \in A),$$

that is, the original density conditioned on the event of interest happening.

Trying to find the best choice of g is essentially an infinite dimensional optimization problem. In practice, it is often easy to try to turn the problem into a finite dimensional one. This can be done, for instance, by restricting the choice of g to a parametric family of densities $\{g(x;\theta)\}_{\theta\in\Theta}$.

The most useful way to create such a family of densities is by so-called *exponential tilting*. Recall that the moment generating function[1] of a random variable $X \sim f$ is given by

$$M(\theta) = \mathbb{E}_f[\exp(\theta X)],$$

which is defined so long as the expectation on the right is finite. The *cumulant generating function* of a random variable is given by

$$\kappa(\theta) = \log \mathbb{E}_f[\exp(\theta X)] = \log M(\theta).$$

Using this, one can define a family of densities by

$$f(x;\theta) = \exp(\theta x - \kappa(\theta))f(x) = \frac{e^{\theta x}f(x)}{M(\theta)},$$

where $\theta \in \Theta = \{\theta : M(\theta) < \infty\}$, which is also called the density of the *exponential tilting of* X. (or *tilted density*). These are indeed densities, as the exponential function is non-negative and the density integrates to 1, derived from X by the so-called *exponential tilting*; such an "exponentially tilted measure" in many cases has the same parametric form as that of X.

For instance, if $f(x) = \lambda e^{-\lambda x}$, $x \in [0, \infty)$, then

$$f(x;\theta) = (\lambda - \theta)e^{-(\lambda-\theta)x}, \quad \theta < \lambda.$$

Using an exponential tilt, we obtain the estimator

$$\hat{\ell}_\theta = \frac{1}{N}\sum_{i=1}^{N}\frac{f(X^{(i)})}{f(X^{(i)};\theta)}\Psi(X^{(i)}) = \frac{1}{N}\sum_{i=1}^{N}\exp(\kappa(\theta) - \theta X^{(i)})\Psi(X^{(i)}).$$

The obvious question is how to choose θ. A possible option is to find the θ that minimizes the variance of the estimator[2], that is, we choose

$$\theta_{VM} = \arg\min_{\theta\in\Theta} Var_g[\exp(\kappa(\theta) - \theta X)\Psi(X))].$$

---

[1] The moment generating function is so named because it can be used to find the moments of the distribution. The series expansion of e^{tX}

$$e^{t\,X} = 1 + t\,X + \frac{t^2\,X^2}{2!} + \frac{t^3\,X^3}{3!} + \cdots + \frac{t^n\,X^n}{n!} + \cdots.$$

Hence,

$$M_X(t) = \mathbb{E}\left[e^{t\,X}\right] = 1 + t\mathbb{E}[X] + \frac{t^2\mathbb{E}[X^2)]}{2!} + \frac{t^3\mathbb{E}[X^3]}{3!} + \cdots + \frac{t^n\mathbb{E}[X^n)]}{n!} + \cdots = 1 + tm_1 + \frac{t^2m_2}{2!} + \frac{t^3m_3}{3!} + \cdots + \frac{t^nm_n}{n!} + \cdots,$$

where m_n is the n-th moment. Differentiating $M_X(t)$ i times with respect to t and setting $t = 0$, we obtain the i-th moment about the origin, m_i.

[2] Note that this is equivalent to minimizing the second moment of the estimator. Using the results from Section 5.2 (see Remark 5.5), the problem is thus to minimize

$$\mathbb{E}_f\left[\Psi^2(\mathbf{X})\frac{f(\mathbf{X})}{g(\mathbf{X})}\right] < \infty.$$

Note that this can be estimated by first using acceptance rejection to sample from $f(x \mid X \in A)$, then using numerical optimization to find the best choice of θ.

Problem 6.2 (Rare events probability / 2).
Suppose we wish to estimate $\ell = P(X > \gamma)$ where $X \sim N(0,1)$.

1. show that that $f(x)$ is a $N(\mu, \sigma^2)$ density, then the tilted density $f(x;\theta)$ is a $N(\mu+\sigma^2\theta, \sigma^2)$ density.
2. We can use the variance minimization approach to carry out importance sampling, as follows:
   - consider a small initial simulation (with $N_{init} = 10^4$ samples) and use numerical optimization to find the optimal parameter;
   - use this parameter to obtain a tilted distribution with can be used to perform importance sampling, then perform it using $N = 10^5$ samples.

   Apply this strategy by choosing $\gamma = 3$.

1. The moment generating function of a normal distribution with mean μ and variance σ^2 is given by

$$M(\theta) = exp\left(\mu\theta + \frac{\sigma^2\theta^2}{2}\right),$$

so that

$$\kappa(\theta) = \mu\theta + \frac{\sigma^2\theta^2}{2}.$$

This means that the tilted normal distribution is of the form

$$\begin{aligned} f(x;\theta) \propto & \exp(\theta x - \kappa(\theta)) \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ = & \exp\left(\theta x - \left(\mu\theta + \frac{\sigma^2\theta^2}{2}\right)\right) exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ = & \exp\left(-\frac{1}{2\sigma^2}(x - (\mu + \sigma^2\theta))^2\right), \end{aligned}$$

which is the PDF of a normal distribution with mean $\mu+\sigma^2\theta$ and variance σ^2. Thus, exponentially tilting a normal distribution is equivalent to shifting its mean by $\sigma^2\theta$.

2. We can use the following `Matlab` commands:

```
gamma = 3; mu = 0; sigma = 1;
N_init = 10^4;
N = 10^5;

X_init = mu + sigma * randn(N_init,1);
X_sample = X_init(find(X_init > gamma));
theta_vm = fminsearch(@(theta) second_moment(theta, mu, sigma, X_sample), 0);

X_is = theta_vm + randn(N,1);
results_IS = exp(mu*theta_vm + sigma^2*theta_vm^2/2 - theta_vm*X_is) .* (X_is > gamma);

X_CMC = randn(N + N_init,1);
results_CMC = (X_CMC > gamma);

prob_exact = 1 - normcdf((gamma-mu)/sigma)
   prob_exact = 0.0013

ell_hat_IS = mean(results_IS)
   ell_hat_IS = 0.0014
```

```
re_hat_IS = std(results_IS) / (sqrt(N)* ell_hat_IS)
   re_hat_IS = 0.0058

ell_hat_CMC = mean(results_CMC)
   ell_hat_CMC = 0.0012

re_hat_CMC = std(results_CMC) / (sqrt(N+N_init)* ell_hat_CMC)
   re_hat_CMC = 0.0857


function [ val ] = second_moment( theta, mu, sigma, X )

kappa_theta = mu * theta + sigma^2 * theta^2 / 2;
val = mean(exp(kappa_theta - theta*X));

end
```

An alternative way to choose θ is so that the mean of the tilted density is the point of interest (e.g., in the case where we wish to estimate $P(X > \gamma)$ we tilt the density so that $\mathbb{E}_\theta[X] = \gamma$.

Problem 6.3 (Rare events probability / 3).
We want to apply importance sampling by exploiting an exponential tilting, and selecting the tilted density to have mean γ by choosing

$$\theta_{ms} = \theta \qquad \text{such that} \frac{d}{d\theta}\kappa(\theta) = \gamma.$$

1. Show that if $X \sim f$ with cumulant generating function κ, then provided that the order of differentiation and expectation can be exchanged, prove that
$$\frac{d}{d\theta}\kappa(\theta) = \mathbb{E}_\theta[X].$$
2. Find the value of θ_{ms} in the case of a $N(0,1)$ distribution.
3. Implement in `Matlab` importance sampling to estimate $P(X > \gamma)$ according to the strategy found at the previous points.
4. Why did it work? A possible hint is provided by the following *Chernoff inequality*, which you can prove: if X has cumulant generating function κ, then
$$P(X > \gamma) \leq \inf_{\theta > 0} \exp\left(\kappa(\theta) - \theta\gamma\right).$$

1. Observe that, since
$$\kappa(\theta) = \log \mathbb{E}[\exp(\theta X)] \qquad \Rightarrow \qquad e^{\kappa(\theta)} = \mathbb{E}[\exp(\theta X)]$$
we obtain
$$\begin{aligned}\frac{d}{d\theta}\kappa(\theta) = & \frac{d}{d\theta}\log \mathbb{E}\left[e^{\theta X}\right] = \frac{1}{\mathbb{E}\left[e^{\theta X}\right]}\frac{d}{d\theta}\mathbb{E}\left[e^{\theta X}\right] \\ = & \frac{1}{e^{\kappa(\theta)}}\frac{d}{d\theta}\mathbb{E}\left[e^{\theta X}\right] = \frac{1}{e^{\kappa(\theta)}}\mathbb{E}\left[Xe^{\theta X}\right] \\ = & e^{-\kappa(\theta)}\int xe^{\theta x}f(x)dx = \int xe^{\theta x - \kappa(\theta)}f(x)dx = \mathbb{E}_\theta[X]\end{aligned}$$
having denoted by $\mathbb{E}_\theta[\cdot]$ the expectation with respect to the tilted distribution $f(x;\theta) = \exp(\theta x - \kappa(\theta))f(x)$.

2. In the case of a $N(0,1)$ distribution, the value of θ_{ms} such that

$$\frac{d}{d\theta}\kappa(\theta) = \gamma$$

is given by

$$\frac{d}{d\theta}\kappa(\theta) = \mu + \sigma^2\theta \quad \Rightarrow \quad \theta_{ms} = \frac{\gamma - \mu}{\sigma^2}.$$

3. Using this, we can then estimate $P(X > \gamma)$ via importance sampling as follows:

```
gamma = 4; mu = 0; sigma = 1;
N = 10^5;

theta = (gamma - mu ) / sigma^2;
X_is = mu + sigma^2*theta+ sigma*randn(N,1);
results_IS = exp(mu*theta + sigma^2*theta^2/2 - theta*X_is) .* (X_is > gamma);

X_CMC = mu + sigma * randn(N,1);
results_CMC = (X_CMC > gamma);

ell = 1 - normcdf((gamma-mu)/sigma)
   ell = 3.1671e-05

ell_hat_IS = mean(results_IS)
   ell_hat_IS = 3.1620e-05

re_hat_IS = std(results_IS) / (sqrt(N)* ell_hat_IS)
   re_hat_IS = 0.0067

ell_hat_CMC = mean(results_CMC)
   ell_hat_CMC = 3.0000e-05

re_hat_CMC = std(results_CMC) / (sqrt(N)* ell_hat_CMC)
   re_hat_CMC = 0.5773
```

4. We have that, for all $\theta > 0$,

$$\begin{aligned} P(X > \gamma) = \; & \mathbb{E}[\mathbf{1}(X > \gamma)] = \mathbb{E}\left[e^{\theta X - \theta X}\mathbf{1}(X > \gamma)\right] \\ \leq \; & e^{-\theta\gamma}\mathbb{E}\left[e^{\theta X}\mathbf{1}(X > \gamma)\right] \leq e^{-\theta\gamma}\mathbb{E}\left[e^{\theta X}\right] = e^{\kappa(\theta)-\theta\gamma}. \end{aligned}$$

The bound then follows by taking the infimum over all $\theta > 0$.

Note that, in general, this bound is minimized by choosing θ such that

$$\frac{d}{d\theta}(\kappa(\theta) - \theta\gamma) = 0 \quad \Rightarrow \frac{d}{d\theta}\kappa(\theta) = \gamma.$$