Importance Sampling
What is Importance Sampling?

- A simulation technique
- Used when we are interested in rare events
- Examples:
  - 
  - Bit Error Rate on a channel,
  - Failure probability of a reliable system
We saw some of it already

Q: We simulate $R = 10\,000$ samples and find no bit error. What can we say about the bit error rate?

A: with confidence 0.95, BER < $3.7 \times 10^{-4}$
THEOREM 2.2.4. [37, p. 110] Assume we observe \( z \) successes out of \( n \) independent experiments. A confidence interval at level \( \gamma \) for the success probability \( p \) is \([L(z); U(z)]\) with

\[
\begin{align*}
L(0) &= 0 \\
L(z) &= \phi_{N, z-1} \left( \frac{1+\gamma}{2} \right), \quad z = 1, \ldots, n \\
U(z) &= 1 - L(n - z)
\end{align*}
\] (2.26)

where \( \phi_{n, z}(\alpha) \) is defined for \( n = 2, 3, \ldots, z \in \{0, 1, \ldots, n\} \) and \( \alpha \in (0; 1) \) by

\[
\begin{align*}
\phi_{n, z}(\alpha) &= \frac{n_1 f}{n_2 + n_1 f} \\
n_1 &= 2(z + 1), \quad n_2 = 2(n - z), \quad 1 - \alpha = F_{n_1, n_2}(f)
\end{align*}
\] (2.27)

\( (F_{n_1, n_2}() \) is the CDF of the Fisher distribution with \( n_1, n_2 \) degrees of freedom. In particular, the confidence interval for \( p \) when we observe \( z = 0 \) successes is \([0; p_0(n)]\) with

\[
p_0(n) = 1 - \left( \frac{1 - \gamma}{2} \right)^{\frac{1}{n}} = \frac{1}{n} \log \left( \frac{2}{1 - \gamma} \right) + o \left( \frac{1}{n} \right) \text{ for large } n
\] (2.28)

Whenever \( z \geq 6 \) and \( n - z \geq 6 \), the normal approximation

\[
\begin{align*}
L(z) &\approx \frac{z}{n} - \frac{n}{n} \sqrt{\frac{z (1 - z)}{n}} \\
U(z) &\approx \frac{z}{n} + \frac{n}{n} \sqrt{z (1 - z)}
\end{align*}
\] (2.29)

can be used instead, with \( N_{0,1}(\eta) = \frac{1+\gamma}{2} \).
What is the Problem?

- Assume you can simulate a system
- You want to evaluate the probability of a rare event
- We want to say more than an answer like: \( p \in [0, 3.69 \times 10^{-4}] \)
i.e. we want a good relative accuracy on \( p \)

- Assume proba of rare event is \( 10^{-6} \): how many simulation runs do you need to obtain an estimate of \( p \) with 10% relative accuracy?
What is the Problem?

Assume proba of rare event is $10^{-6}$: how many simulation runs do you need to obtain an estimate of $p$ with 10% relative accuracy?
What is the Problem?

Assume proba of rare event is $10^{-6}$: how many simulation runs do you need to obtain an estimate of $p$ with 10\% relative accuracy?

- $R$ replications
- $N$ events
- $\hat{p} = \frac{N}{R}$

Confidence interval $\hat{p} \pm 1.96 \frac{\sigma}{\sqrt{R}}$

$\sigma^2 \approx \hat{p}(1 - \hat{p})$

Relative accuracy $= \frac{1.96 \sigma}{\sqrt{R \hat{p}}} = 1.96 \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{R}} = 1.96 \sqrt{\frac{1-\hat{p}}{R \hat{p}}}$

Relative accuracy $= 10\% \iff 1.96 \sqrt{\frac{1-\hat{p}}{R \hat{p}}} = 0.1 \iff R \approx \frac{1.96^2}{0.1^2 p} \approx \frac{400}{p}$
The Goal of Importance Sampling

- Obtain small probability $p$ with good accuracy
- ... while keeping $R$ small

- In the previous example, the direct approach requires $R = 4 \times 10^8$ runs to estimate $p \approx 10^{-6}$ with 10% accuracy

- We can do much better with Importance Sampling
The Idea of Importance Sampling

Formally, assume we simulate a random variable $X$ in $\mathbb{R}^d$, with PDF $f_X()$. Our goal is to estimate $p = \mathbb{E}(\phi(X))$, where $\phi$ is the metric of interest. Frequently, $\phi(x)$ is the indicator function, equal to 1 if the value $x$ corresponds to a failure of the system, and 0 otherwise.

We replace the original PDF $f_X()$ by another one, $f_{\tilde{X}}()$, called the PDF of the importance sampling distribution, on the same space $\mathbb{R}^d$. We assume that

$$\text{if } f_X(x) > 0 \text{ then } f_{\tilde{X}}(x) > 0$$

i.e. the support of the importance sampling distribution contains that of the original one. For $x$ in the support of $f_X()$, define the weighting function

$$w(x) = \frac{f_X(x)}{f_{\tilde{X}}(x)}$$

(7.15)
The Idea of Importance Sampling (cont’d)

- If we simulate $X$, how do we estimate $p$?

- If we simulate $\hat{X}$ instead of $X$, we cannot use $E(\phi(\hat{X}))$

- But: $E\left(\phi(\hat{X})w(\hat{X})\right) = p$
  Show this!
Importance Sampling Monte Carlo

which is the fundamental equation of importance sampling. A Monte Carlo estimate of $p$ is thus given by

$$
\hat{p} = \frac{1}{R} \sum_{r=1}^{R} \phi(\hat{X}_r)w(\hat{X}_r) \quad (7.17)
$$

where $\hat{X}_r$ are $R$ independent replicates of $\hat{X}$. 
Example: Bit Error Rate (BER)

**Example 7.16: Bit Error Rate and Exponential Twisting.** The Bit Error Rate on a communication channel with impulsive interferers can be expressed as [25]:

\[ p = \mathbb{P}(X_0 + X_1 + \ldots + X_d > a) \quad (7.18) \]

where \( X_0 \sim N_{0, \sigma^2} \) is thermal noise and \( X_j, j = 1, \ldots, d \) represents impulsive interferers. The distribution of \( X_j \) is discrete, with support in \( \{\pm x_{j,k}, k = 1, \ldots, n\} \cup \{0\} \)

\[
\begin{align*}
\mathbb{P}(X_j = \pm x_{j,k}) &= q \\
\mathbb{P}(X_j = 0) &= 1 - 2nq
\end{align*}
\]

where \( n = 40, q = \frac{1}{512} \) and the array \( \{\pm x_{j,k}, k = 1, \ldots, n\} \) are given numerically by channel estimation (Table 7.2, for \( d = 9 \)). The variables \( X_j, j = 0, \ldots, d \) are independent. For large values of \( d \), we could approximate \( p \) by a gaussian approximation, but it can easily be verified that for \( d \) of the order of 10 or less this does not hold [25].

<table>
<thead>
<tr>
<th>( k )</th>
<th>( j=1 )</th>
<th>( j=2 )</th>
<th>( j=3 )</th>
<th>( j=4 )</th>
<th>( j=5 )</th>
<th>( j=6 )</th>
<th>( j=7 )</th>
<th>( j=8 )</th>
<th>( j=9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.4706</td>
<td>0.0547</td>
<td>0.0806</td>
<td>0.0944</td>
<td>0.4884</td>
<td>0.3324</td>
<td>0.4822</td>
<td>0.3794</td>
<td>0.2047</td>
</tr>
<tr>
<td>2</td>
<td>0.8429</td>
<td>0.0683</td>
<td>0.2684</td>
<td>0.2608</td>
<td>0.0630</td>
<td>0.1022</td>
<td>0.1224</td>
<td>0.0100</td>
<td>0.0282</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
A direct Monte Carlo estimation (without importance sampling) gives the following results ($R$ is the number of Monte Carlo runs required to reach 10% accuracy with confidence 95%, as of Eq.(7.14)):

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>$a$</th>
<th>BER estimate</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>3</td>
<td>$(6.45 \pm 0.6) \times 10^{-6}$</td>
<td>$6.2 \times 10^4$</td>
</tr>
</tbody>
</table>
\[X = (X_0, X_1, \ldots, X_d)\]
\[X_0 \sim N(0, \sigma^2)\]
\(X_j\) discrete, on \(\{x_{j,1}, \ldots, x_{j,9}\}\)
\[P(X_{j,k} = x_{j,k}) = q_k\]
Estimate \(p = P(X_0 + \cdots + X_d > a)\)
\[\phi(X) = 1_{X_0 + \cdots + X_d > a}\]

\[\hat{X} = (\hat{X}_0, \hat{X}_1, \ldots, \hat{X}_d)\]
\(\hat{X}_0\) on \((-\infty, +\infty)\)
\(\hat{X}_j\) discrete, on \(\{x_{j,1}, \ldots, x_{j,9}\}\)
Estimate \(p = E\left(w(\hat{X})\phi(\hat{X})\right)\)
\[ X = (X_0, X_1, \ldots, X_d) \]
\[ X_0 \sim N(0, \sigma^2) \]
\( X_j \) discrete, on \( \{x_{j,1}, \ldots, x_{j,9}\} \)
\[ P(X_{j,k} = x_{j,k}) = q_k \]
Estimate \( p = P(X_0 + \cdots + X_d > a) \)
\[ \phi(X) = 1_{x_0 + \cdots + x_d > a} \]

\[ \hat{X} = (\hat{X}_0, \hat{X}_1, \ldots, \hat{X}_d) \]
\( \hat{X}_0 \) on \( (-\infty, +\infty) \)
\( \hat{X}_j \) discrete, on \( \{x_{j,1}, \ldots, x_{j,9}\} \)
Estimate \( p = E \left( w(\hat{X})\phi(\hat{X}) \right) \)

**Exponential Twist**

\[ P(\hat{X}_j = x_{j,k}) = e^{\theta x_{j,k}} P(X_j = x_{j,k}) \times ct \]
\[ = e^{\theta x_{j,k}} q_k \eta_j(\theta) \]

\[ \eta_j(\theta)^{-1} = \sum_k e^{\theta x_{j,k}} q_k \]
$X = (X_0, X_1, ..., X_d)$
$X_0 \sim N(0, \sigma^2)$

$f_{X_0}(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2}{2\sigma^2}}$

$\hat{X}_0$ on $(-\infty, +\infty)$

**Exponential Twist**
\[ X = (X_0, X_1, \ldots, X_d) \]
\[ X_0 \sim N(0, \sigma^2) \]
\[ f_{X_0}(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2}{2\sigma^2}} \]

\[ \hat{X}_0 \text{ on } (-\infty, +\infty) \]

**EXPERIMENTAL TWIST**

\[
\begin{align*}
    f_{\hat{X}_0}(x) &= \eta e^{\theta x} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2}{2\sigma^2}} \\
    &= e^{-\frac{x^2 + \theta x}{2\sigma^2}} \times \eta \frac{1}{\sqrt{2\pi\sigma}} \\
    &= e^{-\frac{x^2 - 2\sigma^2 \theta x}{2\sigma^2}} \times \eta \frac{1}{\sqrt{2\pi\sigma}} \\
    &= e^{-\frac{x^2 - 2\sigma^2 \theta x + \sigma^4 \theta^2}{2\sigma^2}} \times e^{\frac{\sigma^4 \theta^2}{2\sigma^2}} \times \eta \frac{1}{\sqrt{2\pi\sigma}} \\
    &= e^{-\frac{(x - \sigma^2 \theta)^2}{2\sigma^2}} \times ct \\
    \hat{X}_0 &\sim N(\sigma^2 \theta, \sigma^2) \\
    \eta &= e^{-\frac{\sigma^2 \theta^2}{2}}
\end{align*}
\]
\( X = (X_0, X_1, \ldots, X_d) \)
\( X_0 \sim N(0, \sigma^2) \)
\( X_j \) discrete, on \( \{x_{j,1}, \ldots, x_{j,9}\} \)
Estimate \( p = P(X_0 + \cdots + X_d > a) \)
\( \phi(X) = 1_{X_0 + \cdots + X_d > a} \)

\( \hat{X} = (\hat{X}_0, \hat{X}_1, \ldots, \hat{X}_d) \)
\( \hat{X}_0 \) on \( (-\infty, +\infty) \)
\( \hat{X}_j \) discrete, on \( \{x_{j,1}, \ldots, x_{j,9}\} \)
\( P(\hat{X}_j = x_{j,k}) = \eta_j(\theta) e^{\theta x_{j,k}} P(X_j = x_{j,k}) \)
\( f_{\hat{X}_0}(x) = \eta_0(\theta) e^{\theta x} f_{X_0}(x) \)

\( w(x_0, \ldots, x_d) = \frac{f_X(x)}{f_{\hat{X}}(x)} = \frac{e^{-\theta(x_0 + \cdots + x_d)}}{\eta_0(\theta) \cdots \eta_d(\theta)} \)
Estimate \( p = E \left( w(\hat{X}) \phi(\hat{X}) \right) \)
Importance Sampling Monte Carlo

We perform $R$ Monte Carlo simulations with $\hat{X}_j$ in lieu of $X_j$; the estimate of $p$ is

$$p_{est} = \frac{1}{R} \sum_{r=1}^{R} w\left(\hat{X}_0^r, ..., \hat{X}_d^r\right) 1\{\hat{X}_0^r + ... + \hat{X}_d^r > a\}$$

(7.20)

- We do this for several values of $\theta$ and find the same estimate $p \approx 6.45 \times 10^{-6}$

- What is different?
Importance Sampling Monte Carlo

We perform $R$ Monte Carlo simulations with $\hat{X}_j$ in lieu of $X_j$; the estimate of $p$ is

$$p_{est} = \frac{1}{R} \sum_{r=1}^{R} w \left( \hat{X}_0^r, \ldots, \hat{X}_d^r \right) 1_{\{\sum X_0^r + \ldots + X_d^r > a\}}$$

(7.20)

- We do this for several values of $\theta$ and find the same estimate $p \approx 6.45 \times 10^{-6}$
- What is different? Hopefully $R$, the number of runs

![Graph showing the relationship between $R$ and $\theta$.]
Choosing an Importance Sampling Distribution

What is a good importance sampling distribution? One that minimizes the number of runs.

This can be quantified with the variance of the importance sampling estimator.

More formally, we can evaluate the efficiency of an importance sampling estimator of \( p \) by its variance:

\[
\hat{v} = \text{var} \left( \phi(X)w(X) \right) = \mathbb{E} \left( \phi(X)^2 w(X)^2 \right) - p^2
\]

Assume that we want a \( 1 - \alpha \) confidence interval of relative accuracy \( \beta \). By a similar reasoning as in Eq.(7.14), the required number of Monte Carlo estimates is

\[
R = \hat{v} \frac{\eta^2}{\beta^2 p^2}
\]  

(7.21)

Thus, it is proportional to \( \hat{v} \). In the formula, \( \eta \) is defined by \( N_{0,1}(\eta) = 1 - \frac{\alpha}{2} \); for example, with \( \alpha = 0.05, \beta = 0.1 \), we need \( R \approx 400\hat{v}/p^2 \).
Example 7.17: Bit Error Rate, re-visited. We can apply Algorithm 2 directly. With the same notation as in Example 7.16, an estimate of $\hat{\sigma}$, the variance of the importance sampling estimator, is

$$\hat{\sigma}_{est} = \frac{1}{R} \sum_{r=1}^{R} w \left( \hat{X}_0^r, ..., \hat{X}_d^r \right)^2 1_{\{\hat{X}_0^r + ... + \hat{X}_d^r > a\}} - p_{est}^2$$

(7.22)

We computed $\hat{\sigma}_{est}$ for different values of $\theta$; Figure 7.12 shows the corresponding values of the required number of simulation runs $R$ (to reach 10% accuracy with confidence 95%), as given by Eq.(7.21)).
The smallest variance is for

\[ E(\phi(\hat{X})) \approx 0.5 \]
Choosing an Importance Sampling Distribution (1)

Rule of thumb:
- The events of interest, under the importance sampling distribution should be not rare
  not certain
Choosing an Importance Sampling Distribution (2)

- The optimal importance sampling distribution is the one that minimizes

\[ \mathbb{E}\left(\phi(\hat{X})^2 w(\hat{X})^2\right) \]

- Is this the same as minimizing the variance of the importance sampling estimator?
function MAIN
\[ \eta = 1.96; \beta = 0.1; \text{pCountMin} = 10; \]  \( \triangleright \beta \) is the relative accuracy of the final result  
\[ \text{GLOBAL } R_0 = 2 \frac{\eta^2}{\beta^2}; \]  \( \triangleright \) Typical number of iterations  
\[ R_{\text{max}} = 1E + 9; \]  \( \triangleright R_0 \) chosen by Eq.(7.14) with \( p = 0.5 \)  
\[ c = \frac{\beta^2}{\eta^2}; \]  \( \triangleright \) Maximum number of iterations

Find \( \theta_0 \in \Theta \) which minimizes varest(\( \theta \))

\[ \text{pCount0} = 0; \text{pCount} = 0; m_2 = 0; \]

for \( r = 1 : R_{\text{max}} \) do

\[
\text{draw a sample } x \text{ of } \hat{X} \text{ using parameter } \theta_0;
\]

\[ \text{pCount0} = \text{pCount0} + \phi(x); \]

\[ \text{pCount} = \text{pCount} + \phi(x)w(x); \]

\[ m_2 = m_2 + (\phi(x)w(x))^2; \]

if \( r \geq R_0 \) and \( \text{pCountMin} < \text{pCount} < r - \text{pCountMin} \) then

\[ p = \frac{\text{pCount}}{r}; \]

\[ v = \frac{m_2}{r} - p^2; \]

if \( v \leq cp^2r \) then break

end if

end if

end for

return \( p, r \)

end function
26:  function VAREST(θ)  ▷ Test if $\mathbb{E}\left(\phi(\hat{X})\right) \approx 0.5$ and if so estimate $\mathbb{E}\left(\phi(\hat{X})^2w(\hat{X})^2\right)$
27:      CONST $\hat{p}_{\text{min}} = 0.3$, $\hat{p}_{\text{max}} = 0.7$;
28:      GLOBAL $R_0$;
29:      $\hat{p} = 0$; $m_2 = 0$;
30:      for $r = 1 : R_0$ do
31:          draw a sample $x$ of $\hat{X}$ using parameter $\theta$;
32:          $\hat{p} = \hat{p} + \phi(x)$;
33:          $m_2 = m_2 + (\phi(x)w(x))^2$;
34:      end for
35:      $\hat{p} = \frac{\hat{p}}{R}$;
36:      $m_2 = \frac{m_2}{R}$;
37:      if $\hat{p}_{\text{min}} \leq \hat{p} \leq \hat{p}_{\text{max}}$ then
38:          return $m_2$;
39:      else
40:          return $\infty$;
41:      end if
42:  end function
A Generic Algorithm

- Ideas: empirically find importance sampling distribution such that
  - Average occurrence of event of interest is close to 0.5
  - Minimizes $\mathbb{E}(\phi(X)^2w(X)^2)$
  - Can be computed by Monte Carlo with small number of runs

- The algorithm does not say how to do one important thing: which one?
Conclusion

- If you have to simulate rare events, importance sampling is probably applicable to your case and will provide significant speedup.

- A generic algorithm can be used to find a good sampling distribution.