The sample size required in importance sampling

Sourav Chatterjee
Importance sampling

Let $\mu$ and $\nu$ be two probability measures on a set $\mathcal{X}$, with $\nu \ll \mu$.

Let $\rho = \frac{d\nu}{d\mu}$.

Let $X_1, X_2, \ldots \overset{i.i.d.}{\sim} \mu$.

Let $f : \mathcal{X} \to \mathbb{R}$ be a measurable function.

The goal is to estimate the integral

$$I(f) := \int_{\mathcal{X}} f(y) \, d\nu(y) = \int_{\mathcal{X}} f(x) \rho(x) \, d\mu(x).$$

The basic importance sampling estimate is

$$I_n(f) := \frac{1}{n} \sum_{i=1}^{n} f(X_i) \rho(X_i).$$

Question: How large should $n$ be, so that this estimate is accurate?
Problems

- In typical applications, \( \nu \) is nearly singular with respect to \( \mu \), which necessitates very large sample sizes.
- Usually, the estimated variance of \( I_n(f) \) is used as a diagnostic.
- Given \( \nu \), there is a big literature on choosing \( \mu \) so that the required sample size (as prescribed by the variance) is as small as possible, with the constraint that \( \mu \) is a measure that is “easy to generate from”.
- However, the sample size required for making the variance small may be much larger than the sample size required for guaranteeing that \( I_n(f) \) is close to \( I(f) \). In other words, it may be an overkill. We will see examples later.
- So, what is the right approach?
Recall: Base measure $\mu$, target measure $\nu$.

Let $Y \sim \nu$.

Let $L$ be the Kullback–Leibler divergence of $\mu$ from $\nu$. That is, $L = \mathbb{E}(\log \rho(Y))$.

The theorem says that if $s$ is the standard deviation of $\log \rho(Y)$, then a sample of size $\exp(L + O(s))$ is sufficient and a sample of size $\exp(L - O(s))$ is necessary for importance sampling to perform well.
The main result, precise statement

- Recall: \( \rho = \frac{d\nu}{d\mu}, \ Y \sim \nu, \ L = \mathbb{E}(\log \rho(Y)) = \text{KL}(\nu \| \mu). \)

Theorem (C. & Diaconis, 2015)

If \( n = \exp(L + t) \) for some \( t \geq 0 \), then

\[
\mathbb{E}|I_n(f) - I(f)| \leq \|f\|_{L^2(\nu)}(e^{-t/4} + 2\sqrt{\mathbb{P}(\log \rho(Y) > L + t/2)}).
\]

Conversely, let \( 1 \) denote the function that is identically equal to 1. If \( n = \exp(L - t) \) for some \( t \geq 0 \), then

\[
\mathbb{P}(I_n(1) \geq 1/2) \leq e^{-t/2} + 2\mathbb{P}(\log \rho(Y) \leq L - t/2).
\]
A simple example

- Let $\mu = \text{Binomial}(N, p)$ and $\nu = \text{Binomial}(N, r)$, where $r > p$.
- Then
  \[
  \log \rho(x) = x \log \frac{r}{p} + (N - x) \log \frac{1 - r}{1 - p}.
  \]
- Let $Y \sim \nu$. Then $L = \mathbb{E}(\log \rho(Y)) = NH(r, p)$, where
  \[
  H(r, p) = r \log \frac{r}{p} + (1 - r) \log \frac{1 - r}{1 - p}.
  \]
- Moreover, the standard deviation of $\log \rho(Y)$ is of order $\sqrt{N}$.
- Thus, the required sample size is $\exp(NH(r, p) + O(\sqrt{N}))$.
- On the other hand, if the variance is used to determine sample size, the required size would be $\exp(NV(r, p))$, where
  \[
  V(r, p) = \log \left( \frac{r^2}{p} + \frac{(1 - r)^2}{1 - p} \right).
  \]
- By Jensen’s inequality, $V(r, p) \geq H(r, p)$. 

Sourav Chatterjee

The sample size required in importance sampling
Figure: The dotted line represents $V(r, p)$ and the solid line represents $H(r, p)$. Here $p = 0.5$ and $r$ goes from 0.5 to 1 on the x-axis.
Often, the target density $\rho$ is known only up to an unknown constant.

That is, we are given that $\rho(x) = C\tau(x)$ where $\tau$ is known but $C$ is not.

In such cases, the self-normalized importance sampling estimate is used:

$$J_n(f) = \frac{\sum_{i=1}^{n} f(X_i)\tau(X_i)}{\sum_{i=1}^{n} \tau(X_i)}.$$

This is actually much more widely used than $I_n(f)$.

What is the required sample size for this one?

Answer: Same as before. Statement of theorem is slightly different.
Precise result for self-normalized estimate

- Recall: $\rho = \frac{d\nu}{d\mu}$, $Y \sim \nu$, $L = \mathbb{E}(\log \rho(Y)) = \text{KL}(\nu\|\mu)$.

Theorem (C. & Diaconis, 2015)

Let $n = \exp(L + t)$ for some $t \geq 0$. Let

$$
\epsilon := (e^{-t/4} + 2 \sqrt{\mathbb{P}(\log \rho(Y) > L + t/2)})^{1/2}.
$$

Then

$$
\mathbb{P}\left( |J_n(f) - I(f)| \geq \frac{2\|f\|_{L^2(\nu)}\epsilon}{1 - \epsilon} \right) \leq 2\epsilon.
$$

Conversely, suppose that $n = \exp(L - t)$ for some $t \geq 0$. Let $f(x)$ denote the function that is 1 when $\log \rho(x) \leq L - t/2$ and 0 otherwise. Then $I(f) = \mathbb{P}(\log \rho(Y) \leq L - t/2)$ and $\mathbb{P}(J_n(f) \neq 1) \leq e^{-t/2}$. 
Estimating probabilities of rare events

- Importance sampling is often used for estimating probabilities of rare events. Large literature, possibly beginning with the work of Siegmund (1976).
- Let $\mu$ and $\nu$ be two probability measures on the same space, with $\nu \ll \mu$. Let $\rho = \frac{d\nu}{d\mu}$.
- Suppose that $A$ is an event such that $\nu(A)$ is small but $\mu(A)$ is not.
- Let $X_1, X_2, \ldots \sim_{i.i.d.} \mu$. The importance sampling estimate of $\nu(A)$ is
  \[ I_n(1_A) = \frac{1}{n} \sum_{i=1}^{n} 1_A(X_i)\rho(X_i). \]
- Question: How large should $n$ be, so that $I_n(1_A)/\nu(A)$ is close to 1?
Sample size required for estimating small probabilities
(rough statement)

Let $Y \sim \nu$, and let $\nu_A$ be the law of $Y$ given $Y \in A$. Let $\rho_A = \frac{d\nu_A}{d\mu}$.

Let $L_A = \text{KL}(\nu_A\|\mu)$.

If $s_A$ is the standard deviation of $\log \rho_A(Y)$ conditional on the event $Y \in A$, then a sample of size $\exp(L_A + O(s_A))$ is sufficient and a sample of size $\exp(L_A - O(s_A))$ is necessary for $\ln(1_A)/\nu(A)$ to be close to 1.
Recall: \( Y \sim \nu \), \( \nu_A \) is law of \( Y \) given \( Y \in A \), \( \rho_A = \frac{d\nu_A}{d\mu} \),
\[
L_A = \mathbb{E}(\log \rho_A(Y) | Y \in A) = \text{KL}(\nu_A || \mu).
\]

**Theorem (C. & Diaconis, 2015)**

*If \( n = \exp(L_A + t) \) for some \( t \geq 0 \), then*

\[
\mathbb{E} \left| \frac{\ln(1_A)}{\nu(A)} - 1 \right| \leq e^{-t/4} + 2 \sqrt{\mathbb{P}(\log \rho_A(Y) > L_A + t/2 | Y \in A)}.
\]

*Conversely, suppose that \( n = \exp(L_A - t) \) for some \( t \geq 0 \). Then*

\[
\mathbb{P} \left( \frac{\ln(1_A)}{\nu(A)} \geq \frac{1}{2} \right) \leq e^{-t/2} + 2 \mathbb{P}(\log \rho_A(Y) \leq L_A - t/2 | Y \in A).
\]
Many more theorems and examples in the arXiv preprint “The sample size required in importance sampling” by Chatterjee and Diaconis.

Includes connections with statistical physics and phase transitions.

Contains a proposal for using the smallness of

$$\frac{\max_{1 \leq i \leq n} \rho(X_i)}{\sum_{i=1}^{n} \rho(X_i)}$$

as a diagnostic criterion for convergence of importance sampling, and proves that it works under certain circumstances.