Rao-Blackwellization

Two points to this lecture:

1. How to improve the efficiency of a statistical estimator by adding more statistics.

2. When should you sample? And when use analytic techniques?

Recall efficiency, the no. of samples depends on variance of estimator.
How Many Independent Samples are Needed?

Want to estimate \( \mu = \mathbb{E} h(x) \mathbb{P}(X) \)

- e.g. The expected cost to build a pipeline in Sicily (Mackay rule)

Take \( m \) samples \( x^{(1)}, \ldots, x^{(m)} \) from \( \mathbb{P}(X) \)

\[
\text{Estimator } \bar{S}_m = \frac{1}{m} \sum_{i=1}^{m} h(x^{(i)}). \quad \text{This is a random variable which depends on the samples.}
\]

The expectation of the estimator is

\[
E_{\mathbb{P}[S_m]} = \sum_{x^{(1)}, \ldots, x^{(m)}} \frac{1}{m} \sum_{i=1}^{m} h(x^{(i)}) \mathbb{P}(X^{(i)}) \mathbb{P}(X^{(i)}) = \frac{1}{m} \sum_{i=1}^{m} h(x^{(i)}) \mathbb{P}(X^{(i)}) = \mathbb{E}_{\mathbb{P}}[h(x)]
\]

The variance of the estimator is:

\[
\text{Var}_{\mathbb{P}[S_m]} = E_{\mathbb{P}}[S_m^2] - (E_{\mathbb{P}}[S_m])^2
\]

\[
\text{Var}_{\mathbb{P}[S_m]} = \sum_{x^{(i)}, \ldots, x^{(m)}} \frac{1}{m} \sum_{i=1}^{m} h(x^{(i)}) h(x^{(i)}) - \mu^2
\]

So \( \text{Var}_{\mathbb{P}[S_m]} = \frac{1}{m} \bar{\sigma}^2 \)

where \( \bar{\sigma}^2 = E_{\mathbb{P}}[(h(x) - \bar{h})^2] \), with \( \bar{h} = E_{\mathbb{P}}[h(x)] = \sum_x \mathbb{P}(x) h(x) \).
One View of Rao-Blackwell.

An initial estimator $\hat{\theta}(x)$

Improve it by adding statistics $T(x)$ to get

$a_1(x) = \mathbb{E}(\hat{\theta}(x) | T(x))$

Claim: $a_1$ is always better than $\hat{\theta}$ (unless it is the same).

Technically, $\mathbb{E}(a_1(x) - \theta)^2 \leq \mathbb{E}(\hat{\theta}(x) - \theta)^2$

(Proof later in lecture.)

E.G. Data generated by Poisson process:

$P(M$ events in time $N) = e^{-\lambda N}(\lambda N)^M/M!$

Divide $N$ into $n$ time units

Let $X_i$ be the no. of counts in the $i^{th}$ time window.

Task: estimate $\lambda$

Data $x = (x_1, x_2, \ldots, x_n)$
Initial Estimator

\[ \lambda (x) = \begin{cases} 1, & \text{if } x_i = 0 \\ 0, & \text{otherwise.} \end{cases} \]  

\( \text{In period } 1, \ p(x_i \text{ counts}) = e^{-\lambda} x_i \)

\[ \sum_x \lambda (x) p(x_i \text{ counts}) = e^{-\lambda} (= p(x_i = 0)). \]

So \( \lambda (x) \) is an unbiased estimator of \( e^{-\lambda} \)

(\( \text{i.e. } E[\lambda(x)] = e^{-\lambda} \)), but the variance is high.

\[ \sum \frac{\lambda (x_i) - e^{-\lambda} x_i}{x_i} = e^{-\lambda} - e^{-2\lambda} \]

Better to use statistic: \( T(x) = x_1 + \ldots + x_n \).

\[ \lambda (x) = E[\lambda | x_1 + \ldots + x_n] \]

This is the probability that we have a total of \( x_1 + \ldots + x_n \text{ counts}, \) but \( x_i = 0 \) (i.e. all events are in t-e

This equals \( (1 - \frac{1}{n}) \) remaining \( n - 1 \) time periods.

If \( n \) is sufficient large

\[ x_1 + \ldots + x_n \approx n \lambda \text{ } \Rightarrow \text{expectation work. } \]

the Poisson Process

\[ (1 - \frac{1}{n}) ^ {x_1 + \ldots + x_n} \approx (1 - \frac{1}{n}) ^ {n \lambda} \approx e^{-\lambda} \]

So we estimate \( e^{-\lambda} \) well from one sample \( x \)

(or take \( \frac{1}{n} \{ \lambda (x_1) + \lambda (x_2) + \ldots + \lambda (x_n) \} \))
General Rao-Blackwellization

Page 51

Rule of Thumb: When sampling, do as much as possible analytically.

Goal: Estimate \( I = E_{\Xi} \sum h(\xi) = \int h(x) \pi(x) dx \)

Monte-Carlo says sample from \( \pi(x) \) to get

\( \text{i.i.d. } x^{(1)}, x^{(2)}, \ldots, x^{(m)} \)

\[ \frac{1}{m} \sum_{i=1}^{m} h(x^{(i)}) \]

\[ \lim_{m \to \infty} \frac{1}{m} (\hat{I}_m - I) \sim N(0, \sigma^2) \]

\[ \sigma = \text{Var} (h) \]

No. samples required \( \sim \sqrt{m \sigma^2} \).

(Ce.g. \( X = (X_1(Tx), T(x)) \))

Suppose we can decompose \( X \) into two parts

\( X = (X_1, X_2) \) / \( \pi_2(x_2) = \sum_{x_1} \pi(x_1, x_2) \),

and \( \mathbb{E} [h(x) | x_2] = \sum_{x_1} h(x_1, x_2) \pi(x_1 | x_2) \)

can be computed analytically.

New estimator

\[ \frac{1}{m} \sum_{i=1}^{m} \mathbb{E} [h(x) | x_2^{(i)}] \]

Where \( x_2^{(1)}, \ldots, x_2^{(m)} \) are i.i.d. samples from \( \pi_2(x_2) \).
We have two estimates \( \hat{I}_m \) and \( \tilde{I}_m \) for \( I \), which is better?

Both estimates are unbiased — the estimates are r.v.'s depending on the samples \( x^{(1)}, \ldots, x^{(m)} \) from distribution \( \pi(x) \).

\[
\bar{I}_m (x^{(1)}, \ldots, x^{(m)})
\]

Its expectation is

\[
\mathbb{E}[\bar{I}_m] = \sum_{x^{(1)}, \ldots, x^{(m)}} \bar{I}_m (x^{(1)}, \ldots, x^{(m)}) \pi(x^{(1)}) \ldots \pi(x^{(m)})
\]

\[
= \sum_{x^{(1)}, \ldots, x^{(m)}} \frac{1}{m} \sum_{i=1}^{m} h(x^{(i)}) \pi(x^{(1)}) \ldots \pi(x^{(m)})
\]

\[
= \sum_x h(x) \pi(x) = I. \quad \text{Hence unbiased}.
\]

Similarly, it can be checked that

\[
\mathbb{E}[\tilde{I}_m] = \sum_{x^{(1)}, \ldots, x^{(m)}} \tilde{I}_m (x^{(1)}, \ldots, x^{(m)}) \pi(x^{(1)}) \ldots \pi(x^{(m)})
\]

\[
= \sum_{x^{(1)}, x^{(2)}, \ldots} h(x^{(1)}, x^{(2)}) \pi(x^{(1)}) \pi(x^{(2)})
\]

\[
= I.
\]
Example

Consider $X_1, X_2$

Suppose $P(X_1 | X_2) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(X_1 - X_2)^2}{2\sigma^2}}$

$P(X_2) = \text{uniform on } [0, 1].$

$h(X_1, X_2) = (X_1^2 + X_2^2)$

Then

$$E \left[ h(X_1, X_2) | X_2 \right] = \int_{-\infty}^{\infty} \left( X_1^2 + X_2^2 \right) \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(X_1 - X_2)^2}{2\sigma^2}} \, dx_1$$

$$= (X_2^2 + \sigma^2) + X_2^2 = \sigma^2 + 2X_2^2.$$

Hence

$$E[h(X_1, X_2)] = \sigma^2 + 2 \sum_{i=1}^{m} \left( X_i \right)_i^2$$

Note: For this example we can compute $I_1 = \int_{0}^{1} (\sigma^2 + 2x^2) \, dx = \sigma^2 + 2\frac{1}{3}$ analytically.

Even better!

Define $I_2 = \int_{-\infty}^{\infty} X^2 e^{-\frac{(X - \mu)^2}{2\sigma^2}} \, dx = \mu^2 + \sigma^2.$

Because $\int_{-\infty}^{\infty} (X - \mu)^2 e^{-\frac{(X - \mu)^2}{2\sigma^2}} \, dx = \sigma^2$

$E\{ (X - \mu)^2 \} = \sigma^2 - \mu^2 = X^2 - 2X\mu + \mu^2$ — result follows.
How many samples are needed?

Recall that \( \sqrt{m} (T_m - T) \sim N(0, \sigma^2) \)

\[ \sigma^2 = \text{Var}_{T_m}(h(x)) \]

Claim: \( \text{Var}_\pi(h(x)) = \text{Var}_{\pi(x|x)}(E_\pi(h(x)|x)) \)

This implies that

\[ \text{Var}_{\pi(x|x)}(h(x)) \geq \text{Var}_{\pi(x|x)}(E_\pi(h(x)|x)) \]

with equality only if \( E_\pi(h(x)|x) = 0 \)

i.e., only if \( h(x) \) is a deterministic function of \( x \).

Proof of Claim:

\[
\text{Var}(h(x)) = \sum_{x_1} \left( \frac{h(x_1, x_2)}{\Pi(x_1, x_2)} - \left( \sum_{x_2} \frac{h(x_1, x_2)}{\Pi(x_1, x_2)} \right)^2 \right)^2 \Pi(x_1, x_2)
\]

\[
= \frac{1}{\mathcal{X}_2} \left( \sum_{x_1} \left( h(x_1, x_2) \frac{\Pi(x_1, x_2)}{\Pi(x_1, x_2)} - \left( \sum_{x_1} \frac{h(x_1, x_2)}{\Pi(x_1, x_2)} \Pi(x_1, x_2) \right)^2 \right)^2 \right)
\]

\[
= \mathbb{E}_{\pi(x|x)} \left( \text{Var}_{\pi(x|x)}(h(x)) \right) + \text{Var}_{\pi(x|x)}(E_\pi(h(x)|x))
\]