Monte Carlo: Importance sampling

Art B. Owen
Stanford University

Adapted from “Monte Carlo theory, methods and examples”

`http://statweb.stanford.edu/~owen/mc/`

Importance sampling

Importance sampling is more complicated than other variance reduction methods.

Done well, it can turn a problem from intractable to easy.

It can also give infinite variance.

Sequential Monte Carlo

Importance sampling is a precursor.

See talks by N. Chopin

Outline

1) What IS is and why we need it
2) Self-normalized IS
3) Example
4) How to do it
5) Adaptive IS (briefly!)

Spiky integrands

Sometimes all the action is in a subset A of tiny probability.

$\mu = \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \approx \int_A f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ where $\mathbb{P}(\boldsymbol{X} \in A) \approx 0$

How it arises

1) Rare events $f(\boldsymbol{x}) = 1\{\boldsymbol{x} \in A\}, \quad \mathbb{P}(\boldsymbol{x} \in A) = \epsilon$
2) Singular integrands, e.g., $f(\boldsymbol{x}) \propto \|\boldsymbol{x} - \boldsymbol{x}_0\|^{-r}, \quad r < d = \dim(\boldsymbol{x})$

Examples

- Probability that an insurance company fails.
- Probability of electrical blackouts.
- Singular integrands in high energy physics.
- Graphics has both at once.

What to do

Get more samples $\boldsymbol{x}_i \in A$, the **important** region.

And then correct for that distortion.

Rare events

$$\mu = \int_A p(\boldsymbol{x})\,\mathrm{d}x = \epsilon \ll 1$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} 1\{\boldsymbol{x}_i \in A\}, \quad \boldsymbol{x}_i \overset{\text{iid}}{\sim} p$$

$$\mathbb{E}(\hat{\mu}) = \epsilon \quad \text{and} \quad \mathrm{Var}(\hat{\mu}) = \frac{\epsilon(1-\epsilon)}{n}$$

Coefficient of variation

$$\mathrm{cv} = \frac{\sqrt{\mathrm{Var}(\hat{\mu})}}{\mu} = \sqrt{\frac{1-\epsilon}{n\epsilon}} \approx \sqrt{\frac{1}{n\epsilon}}$$

To get cv $= 0.1$ we need $n \approx 100/\epsilon$.

Then $\epsilon = 10^{-9}$ takes $n \approx 10^{11}$.

Singularities

Sometimes not severe. For instance,

$$\int_{\mathbb{R}^d} \|\boldsymbol{x}-\boldsymbol{x}_0\|^{-r} p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \quad r>0$$

has finite variance if $r<d/2$ and p is bounded.

For fixed r the larger d gets, the less severe the singularity is.

The cv becomes manageable for large d/r.

Why?

It is a property of high dimensional geometry.

Sample from q

Choose a density q with $q(\boldsymbol{x}) > 0$ whenever $f(\boldsymbol{x})p(\boldsymbol{x}) \neq 0$. Then use

$$\hat{\mu}_q = \frac{1}{n}\sum_{i=1}^{n} \frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)}, \quad \boldsymbol{x}_i \overset{\text{iid}}{\sim} q.$$

Unbiased

Let $Q = \{\boldsymbol{x} \mid q(\boldsymbol{x}) > 0\}$. Then

$$\mathbb{E}(\hat{\mu}_q) = \int_Q f(\boldsymbol{x})\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_Q f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \mu$$

Safe harbour

We can pick q with $q(\boldsymbol{x}) > 0$ whenever $p(\boldsymbol{x}) > 0$.

That works for general f.

Choosing q

$$\mathrm{Var}(\hat{\mu}_q) = \frac{\sigma_q^2}{n}, \quad \text{where}$$

$$\sigma_q^2 = \int \Bigl(\frac{fp}{q} - \mu\Bigr)^2 q \,\mathrm{d}x = \int \frac{(fp - \mu q)^2}{q} \,\mathrm{d}x$$

From the numerator

We do well with $q \approx fp/\mu$. When $f \geqslant 0$, $q = fp/\mu$ is perfect, but unattainable: $f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)/q(\boldsymbol{x}_i) = \mu$.

Generally $q \propto |f|p$ is optimal.

From the denominator

Watch out for q close to zero. E.g., avoid light tailed q.

Todo list for IS

1) sample $\boldsymbol{x} \sim q$
2) compute fp/q given $\boldsymbol{x}$

Beyond variance

Chatterjee & Diaconis (2015) show that we need

$$n \approx \exp(\text{KL distance } p,\, q)$$

for generic f.

They use $\mathbb{E}_q(|\hat{\mu}_q - \mu|)$ and $\mathbb{P}_q(|\hat{\mu}_q - \mu| > \epsilon)$ instead of $\mathrm{Var}_q(\hat{\mu}_q)$.

95% confidence

Taking $\epsilon = .025$ in their Theorem 1.2 shows that we succeed with

$$n \geqslant 6.55 \times 10^{12} \times \exp(\mathrm{KL}).$$

Similarly, poor results are very likely for n much smaller than $\exp(\mathrm{KL})$.

The range for n is not precisely determined by these considerations (yet).

The weight function

Recall that

$$\sigma_q^2 = \int \frac{(fp)^2}{q}\,\mathrm{d}x - \mu^2$$

Let $w(\boldsymbol{x}) = p(\boldsymbol{x})/q(\boldsymbol{x})$.

That mean square can be written

$$\int \frac{(fp)^2}{q}\,\mathrm{d}\boldsymbol{x} = \mathbb{E}_q(w(\boldsymbol{x})^2 f(\boldsymbol{x})^2)$$

Bounded $w(\boldsymbol{x})$ is very helpful.

Unbounded w can give $\sigma_q = \infty$ even when $\sigma < \infty$.

Helpful identity

$$\mathbb{E}_q(w(\boldsymbol{x})^2 f(\boldsymbol{x})^2) = \mathbb{E}_p(w(\boldsymbol{x}) f(\boldsymbol{x})^2)$$

Effective sample size

Unequal weighting raises variance.

Kong (1992), Evans and Swartz (1995)

For IID Y_i with variance σ^2 and fixed* $w_i \geqslant 0$,

$$\mathrm{Var}\Bigl(\frac{\sum_i w_i Y_i}{\sum_i w_i}\Bigr) = \frac{\sum_i w_i^2 \sigma^2}{\bigl(\sum_i w_i\bigr)^2}$$

Write this as

$$\frac{\sigma^2}{n_e} \qquad \text{where} \quad n_e = \frac{(\sum_i w_i)^2}{\sum_i w_i^2}$$

If $\sum_i w_i > 0$ then

$$1 \leqslant n_e \leqslant n$$

If $\sum_i w_i = 0$ then

We don't need a diagnostic to tell us we have a problem.

*fixed?

Our $Y_i = f(\boldsymbol{x}_i)$ are actually linked to $w_i(\boldsymbol{x}_i)$.

Simple examples

$p = \mathcal{N}(0, I)$ and $q = \mathcal{N}(\theta, I)$

$$w(\boldsymbol{x}) = \exp(-\theta^{\mathsf{T}}\boldsymbol{x} + \theta^{\mathsf{T}}\theta/w)$$

$p = \mathcal{N}(0, 1)$ given $x > \tau > 1$ and $q = \tau + \mathrm{Exp}(1)/\tau$.

$$w(x) = \text{Exercise}$$

Tail weight

$p = \mathcal{N}(\mathrm{any})$ and $q = t_{(\nu)} \implies$ ok

$p = t_{(\nu)}$ and $q = \mathcal{N}(\mathrm{any}) \implies$ problematic.

Exercise:

$p = \mathcal{N}(\mu, I_d)$ and $q = \mathcal{N}(\mu, \sigma^2 I_d)$

Self-normalized I.S.

What if we cannot compute p/q? Suppose that

$$p(\boldsymbol{x}) = p_u(\boldsymbol{x})/c_p \quad \text{and} \quad q(\boldsymbol{x}) = q_u(\boldsymbol{x})/c_q$$

and we can compute p_u and q_u but not c_p or c_q. Then we use

$$\begin{aligned}
\tilde{\mu}_q &= \frac{1}{n}\sum_{i=1}^{n}\frac{p_u(\boldsymbol{x}_i)f(\boldsymbol{x}_i)}{q_u(\boldsymbol{x}_i)} \Big/ \frac{1}{n}\sum_{i=1}^{n}\frac{p_u(\boldsymbol{x}_i)}{q_u(\boldsymbol{x}_i)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{p(\boldsymbol{x}_i)f(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} \Big/ \frac{1}{n}\sum_{i=1}^{n}\frac{p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} \qquad \text{cancellation} \\
&\to \int p(\boldsymbol{x})f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \Big/ \int p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \qquad \text{law of large numbers} \\
&= \mu
\end{aligned}$$

About that denominator

Now we need $q(\boldsymbol{x}) > 0$ where $p(\boldsymbol{x}) > 0$,
even if $f(\boldsymbol{x}) = 0$.

Variance of SNIS

It is a ratio estimator:

$$\tilde{\mu}_q = \frac{1}{n}\sum_{i=1}^{n}\frac{p_u(\boldsymbol{x}_i)f(\boldsymbol{x}_i)}{q_u(\boldsymbol{x}_i)} \Bigg/ \frac{1}{n}\sum_{i=1}^{n}\frac{p_u(\boldsymbol{x}_i)}{q_u(\boldsymbol{x}_i)}$$

After Taylor expansions

$$\mathrm{Var}(\tilde{\mu}_q) \doteq \frac{1}{n}\sigma^2_{q,\mathrm{sn}} \qquad \sigma^2_{q,\mathrm{sn}} = \mathbb{E}_q\big(w^2(f-\mu)^2\big) \qquad w(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$$

versus $\quad \sigma^2_q = \mathbb{E}_q((fw-\mu)^2)$

Caveat

Taylor expansion gives $\mathrm{Var}($approximate $\tilde{\mu}_q)$

Optimal SNIS

$$q(\boldsymbol{x}) \propto p(\boldsymbol{x})|f(\boldsymbol{x}) - \mu| \quad \text{vs} \quad p|f| \text{ for ordinary IS} \quad \text{Hesterberg (1988)}$$

As a result

$$\sigma^2_{q,\mathrm{sn}} \geqslant \mathbb{E}_p(|f(\boldsymbol{X}) - \mu|)^2$$

For rare event A

Optimal SNIS has $\boldsymbol{x} \in A$ with probabilty $1/2$. (Exercise)

Variance cannot approach zero like with IS.

SNIS can still beat sampling from p.

Strongest case for SNIS

It is a replacement for acceptance-rejection when

1) p is unnormalized

2) p/q unbounded

IS vs acceptance rejection

- Acceptance-rejection requires bounded $w(\boldsymbol{x}) = p(\boldsymbol{x})/q(\boldsymbol{x})$
- We also have to know a bound.
- IS and SNIS require us to keep track of weights $w_i = w(\boldsymbol{x}_i)$
  Ok for one source of randomness; potentially awkward in a pipeline
- Plain IS requires normalized p/q
- Acceptance-rejection samples cost more (due to rejections)

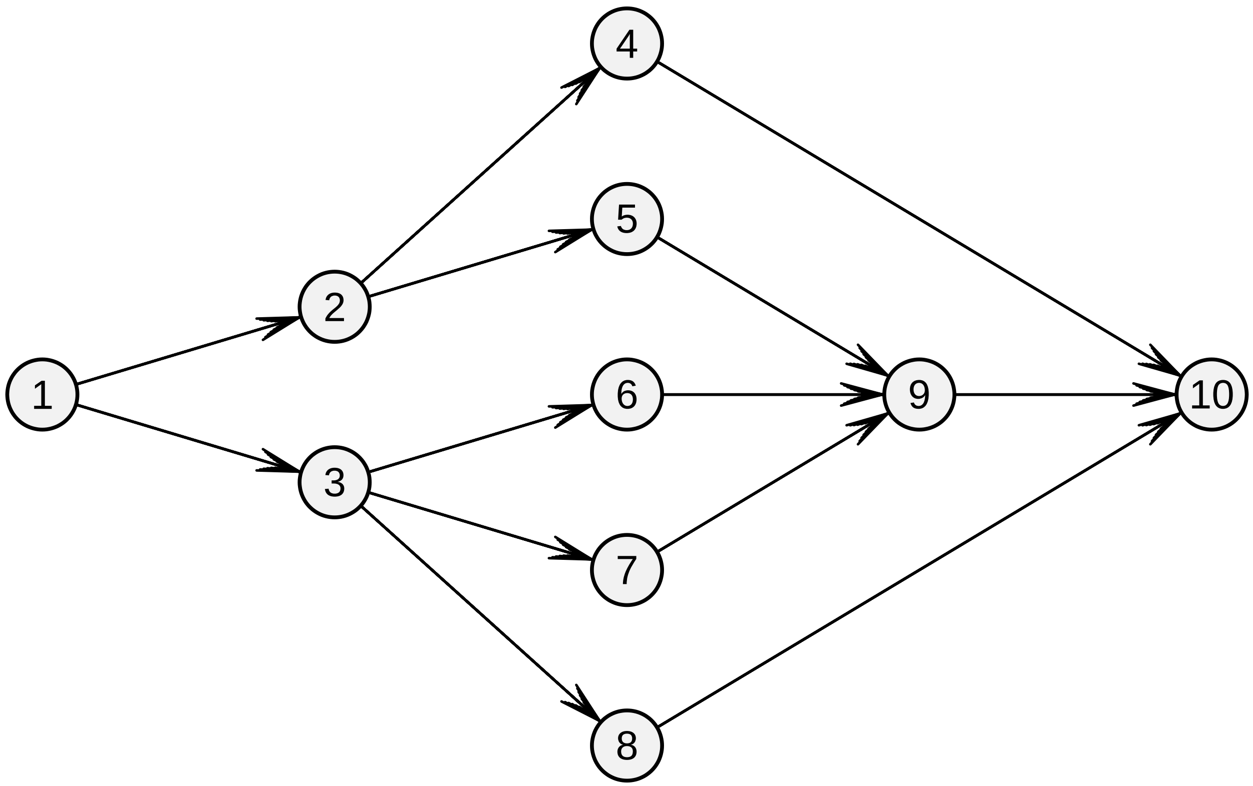
PERT example

A PERT problem from Chinneck. Time to write software.

j	Task	Predecessors	Days to complete
1	Planning	None	4
2	Database Design	1	4
3	Module Layout	1	2
4	Database Capture	2	5
5	Database Interface	2	2
6	Input Module	3	3
7	Output Module	3	2
8	GUI Structure	3	3
9	I/O Interface Implementation	5,6,7	2
10	Final Testing	4,8,9	2

Dependence

PERT graph (activities on nodes)



IS task

Replace all days by exponential random variables.

Find $\mathbb{P}(\text{takes} > 70 \text{ days})$.

PERT details

If everything goes as planned it takes exactly 15 days. (seems optimistic)

T_j is time spent on task j.

E_j is completion time of task j.

Project completes at E_{10}.

Exponential random times give $\mathbb{E}(E_{10}) \doteq 18$ with a long tail to the right.

What is $\mathbb{P}(E_{10} > 70)$? Only happened 2 of $10{,}000$ times.

Importance sampler

Change $T_j \sim \mathrm{Exp}(1) \times \theta_j \implies T_j \sim \mathrm{Exp}(1) \times \lambda_j, \quad j = 1, \dots, 10.$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{E_{i,10} > 70\} \prod_{j=1}^{10} \frac{\exp(-T_{ij}/\theta_j)/\theta_j}{\exp(-T_{ij}/\lambda_j)/\lambda_j}$$

Now choose $\lambda_j \geqslant \theta_j$ carefully.

PERT example

Full details in online notes.

First: 70 is about 4 times $\mathbb{E}(E_{10})$. So lets try $\lambda = 4\theta$. Use $n = 10{,}000$

Oops: we get $n_e \doteq 4.9$. One observation had 43% of the weight!

Second: Try searching for κ where $\lambda = \kappa\theta$ works well.

There really isn't one.

Critical path

In a deterministic setting: a task is on the critical path if delaying it by ϵ delays the total time by ϵ

Third: Just apply some κ to the 4 tasks (1,2,4,10) in the critical path.

$\kappa = 4$ works ok, so raise n to $200{,}000$.

PERT results

$\mathbb{P}(E_{10} > 70) \doteq 3.2 \times 10^{-5}$,
std. err. $\doteq 3.6 \times 10^{-7}$,
$n_e \doteq 7470$

IS reduced variance by about 1200

Lots of further tweaks possible   (e.g., integrate out T_{10})

Couldn't we just automate the process?

Not super rare

Maybe go for $\mathbb{P}(E_{10} > 365)$!

How to find q?

1) Pure inspiration

2) Exponential tilting

3) Hessians and Gaussians

4) Mixtures

Let's skip over item 1.

Changing a parameter

Nominal distribution	$p(\boldsymbol{x};\theta_0)$	$\theta_0 \in \Theta$
Sampling distribution	$p(\boldsymbol{x};\theta)$	$\theta \in \Theta$

Estimator

$$\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)\frac{p(\boldsymbol{x}_i;\theta_0)}{p(\boldsymbol{x}_i;\theta)}$$

The importance ratio often simplifies.

E.g., in exponential families.

Exponential tilting

Many important distributions can be written

$$\exp(\eta(\theta)^\mathsf{T} T(\boldsymbol{x}) - A(\boldsymbol{x}) - C(\theta)), \quad \theta \in \Theta$$

and often

$$\exp(\theta^\mathsf{T} \boldsymbol{x} - A(\boldsymbol{x}) - C(\theta)), \quad \theta \in \Theta$$

Nominal θ_0 sample with θ

Estimator

$$\hat{\mu}_\theta = \underbrace{e^{C(\theta)-C(\theta_0)}}_{\text{free of } \boldsymbol{x}_i} \times \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i) \underbrace{e^{(\theta_0-\theta)^\mathsf{T} \boldsymbol{x}_i}}_{\text{the tilt}}$$

Also called ‘exponential twisting’

Examples

Family	$p(\cdot\,;\theta)$	$w(\cdot)$	Θ
Normal	$\mathcal{N}(\theta, \Sigma)$	$\exp(\boldsymbol{x}^\mathsf{T}\Sigma^{-1}(\theta_0 - \theta) + \frac{1}{2}\theta^\mathsf{T}\Sigma^{-1}\theta - \frac{1}{2}\theta_0^\mathsf{T}\Sigma^{-1}\theta_0)$	$\mathbb{R}^d$
Poisson	$\mathrm{Poi}(\theta)$	$\exp(\theta - \theta_0)(\theta_0/\theta)^x$	$(0, \infty)$
Binomial	$\mathrm{Bin}(m, \theta)$	$(\theta_0/\theta)^x((1 - \theta_0)/(1 - \theta))^{m-x}$	$(0, 1)$
Gamma	$\mathrm{Gam}(\theta)$	$x^{\theta_0/\theta}\,\Gamma(\theta)/\Gamma(\theta_0)$	$(0, \infty)$

The normal family shown shares a non-singular Σ.

Exercise

Can we tilt when $\det(\Sigma) = 0$?

Hessian and Gaussian

Suppose that we find the mode $\boldsymbol{x}_*$ of $p(\boldsymbol{x})$ or better yet, of $h(\boldsymbol{x}) \equiv p(\boldsymbol{x})f(\boldsymbol{x})$.

Taylor approximation

$$\log(h(\boldsymbol{x})) \approx \log(h(\boldsymbol{x}_*)) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_*)^\mathsf{T} H_*(\boldsymbol{x} - \boldsymbol{x}_*)$$

$$h(\boldsymbol{x}) \approx h(\boldsymbol{x}_*)\exp\Bigl(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_*)^\mathsf{T} H_*(\boldsymbol{x} - \boldsymbol{x}_*)\Bigr), \quad \text{suggests}$$

$$q = \mathcal{N}(\boldsymbol{x}_*, H_*^{-1}).$$

The Hessian of $\log(h)$ at $\boldsymbol{x}_*$ is $-H_*$.

Requires positive definite H_*.

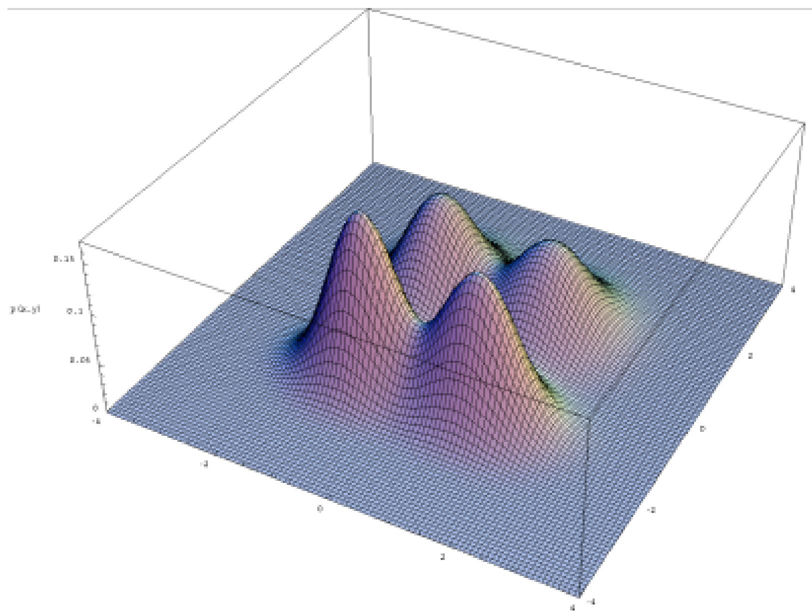
This is an IS version of the Laplace approximation.

For safety

Use a t distribution instead of $\mathcal{N}$.

Mixtures

What if there are multiple modes?



`https://en.wikipedia.org/wiki/Multimodal_distribution`

When sampling the highest mode, we might miss some others.

A mixture of unimodal densities can capture all the modes (that we know of).

Mixture distributions

Sample j randomly from $1, 2, \ldots, J$ with probabilities $\alpha_1, \ldots, \alpha_J$.

Here $\alpha_j \geqslant 0$ and $\sum_{j=1}^J \alpha_j = 1$.

Given j take $\boldsymbol{x} \sim q_j$.

For example

$$\sum_{j=1}^{J} \alpha_j \mathcal{N}(\theta_j, \sigma^2 I_d)$$

With large J, we get kernel density approximations.

These can approximate generic densities.

The approximation gets more difficult with large dimension.

West (1993), Oh & Berger (1993)

Mixtures continued

Multiple kinds of rare event

- J kinds of light path in graphics Veach, Guibas
- ~ 5000 ways the electrical grid can fail O, Maximov, Chertkov (2017)
- J ways a financial portfolio can be hurt
- J failure modes for a bridge

One component can oversample each failure mechanism **(that we know of)**.

Many integrands and/or many distributions

$$\mu_j = \int f_j(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \quad \text{or} \quad \mu_j = \int f(\boldsymbol{x})p_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \quad j = 1,\dots,J$$

Tune one component for each integrand of interest. Pool the values.

IS with mixtures

$$q_\alpha(\boldsymbol{x}) = \sum_{j=1}^{J} \alpha_j q_j(\boldsymbol{x})$$

$$\hat{\mu}_\alpha = \sum_{i=1}^{n} f(\boldsymbol{x}_i)\frac{p(\boldsymbol{x}_i)}{q_\alpha(\boldsymbol{x}_i)} = \sum_{i=1}^{n} \frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{\sum_{j=1}^{J} \alpha_j q_j(\boldsymbol{x}_i)}. \qquad (**)$$

$(**)$ Balance heuristic Veach & Guibas, also Horvitz-Thompson estimator.

Eric Veach got an Oscar for this!

Alternative

Suppose that $\boldsymbol{x}_i$ came from component $j(i)$. We could also use

$$\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)\frac{p(\boldsymbol{x}_i)}{q_{j(i)}(\boldsymbol{x}_i)}.$$

Exercise: This alternative has **higher** variance. (Hint: $1/x$ is convex)

But you don't have to compute every q_j for every $\boldsymbol{x}_i$.

Defensive mixtures

From Hesterberg (1995)

For our best guess $q(\cdot)$ take $q_1 \equiv p$ and let $q_2 = q$. Now,

$$w(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{q_\alpha(\boldsymbol{x})} = \frac{1}{\alpha_1 + \alpha_2 \times q_2(\boldsymbol{x})/p(\boldsymbol{x})} \leqslant \frac{1}{\alpha_1}, \quad \forall \boldsymbol{x}$$

Perhaps $\alpha_1 = 1/10$ or $1/2$.

q does not need to be heavy tailed any more because q_α is.

Variance bound

After some algebra,

$$\mathrm{Var}(\hat{\mu}_{q_\alpha}) \leqslant \frac{1}{n\alpha_1}(\sigma_p^2 + \alpha_2 \mu^2)$$

If however σ_q^2 was very small, defensive IS can lose out.

We don't have a bound vs σ_q^2.

Control variates and mixture IS

For normalized $q_j(\cdot)$ **we know** $\int q_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1,\, j = 1, \dots, J$

$$\mathbb{E}_{q_\alpha}\Bigl(\frac{q_j(\boldsymbol{x})}{q_\alpha(\boldsymbol{x})}\Bigr) = \int q_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1$$

Unbiased estimate

$$\hat{\mu}_{\alpha,\beta} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i) - \sum_{j=1}^{J}\beta_j q_j(\boldsymbol{x}_i)}{\sum_{j=1}^{J}\alpha_j q_j(\boldsymbol{x}_i)} + \sum_{j=1}^{J}\beta_j$$

Additional control variates can be added too.

Via regression

Regress $Y_i = f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)/q_\alpha(\boldsymbol{x}_i)$ on $Z_{ij} = q_j(\boldsymbol{x}_i)/q_\alpha(\boldsymbol{x}_i) - 1$.
Get $\hat{\mu} = \hat{\beta}_0$ (intercept) and se.

$\sum_j \alpha_j Z_{ij} = 1$ for all i, so drop one predictor.

Mixture IS results

O & Zhou (2000)

$$\mathrm{Var}(\hat{\mu}_{\alpha,\beta^{\mathrm{opt}}}) \leqslant \min_{1\leqslant j\leqslant J} \frac{\sigma_j^2}{n\alpha_j}$$

Properties

1) $\hat{\mu}_{\alpha,\beta}$ is unbiased if $q_\alpha > 0$ whenever $fp \neq 0$
2) If any q_j have $\sigma_{q_j}^2 = 0$ we get $\mathrm{Var}(\hat{\mu}_{\alpha,\beta^{\mathrm{opt}}}) = 0$
3) Even better to take exactly $n\alpha_j$ observations from q_j.

We could not expect better in general.

We might have $\sigma_j^2 = \infty$ for all but one j.

The bound is $\sigma_j^2/(n\alpha_j)$ as if we had just used the good one.
(Without knowing which one it was. Indeed it might be a different one for each of several different integrands.)

Summary of mixtures

Using mixtures, we can

- bound the importance ratio
- place a distribution near each singularity
- place a distribution near each failure mode
- tune a distribution to each f_j of interest
- tune a distribution to each p_j of interest
- use control variates to be almost as good as the optimal component

The mixture components can be based on intuition, tilting, Hessians.

Adaptive importance sampling

1) Data $\longrightarrow$ new q

2) $q \longrightarrow$ new data

- Active research area
- Survey of some highlights

What-if simulations

Reweight data from $p(\boldsymbol{x}; \theta_0)$ to estimate

$$\mu(\theta) \equiv \int f(\boldsymbol{x}) p(\boldsymbol{x}; \theta) \, \mathrm{d}\boldsymbol{x}, \quad \theta \neq \theta_0.$$

Now estimate what the IS variance 'would have been' from $p(\cdot; \theta)$.

Adaptation

$\theta_k \leftarrow \theta$ with low estimated variance from θ_{k-1} data.

What-if simulations

Family $p(\boldsymbol{x};\theta), \quad \theta \in \Theta$

We want $\mu(\theta) = \mathbb{E}(f(\boldsymbol{x});\theta) = \int f(\boldsymbol{x})p(\boldsymbol{x};\theta)\,\mathrm{d}\boldsymbol{x}, \quad \theta \in \Theta$

Sample from θ_0 and reweight for $\theta \neq \theta_0$

$$\hat{\mu}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i;\theta)}{p(\boldsymbol{x}_i;\theta_0)}, \quad \boldsymbol{x}_i \sim p(\cdot\,;\theta_0).$$

We can recycle our $f(\boldsymbol{x}_i)$ values

Common heavy-tailed q

$$\hat{\mu}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i;\theta)}{q(\boldsymbol{x}_i)}, \quad \boldsymbol{x}_i \sim q(\cdot\,)$$

Paraphrase (from memory) of Tukey and Trotter (1956)

“Any sample can come from any distribution.”

What-if continued

Estimate what the mean square would have been:

$$\mathrm{MS}_\theta = \int \frac{(f(\boldsymbol{x})p(\boldsymbol{x}))^2}{q(\boldsymbol{x};\theta)}\,\mathrm{d}\boldsymbol{x} = \int \frac{(f(\boldsymbol{x})p(\boldsymbol{x}))^2}{q(\boldsymbol{x};\theta)q(\boldsymbol{x};\theta_0)} q(\boldsymbol{x};\theta_0)\,\mathrm{d}\boldsymbol{x}$$

$$\widehat{\mathrm{MS}}_\theta = \frac{1}{n}\sum_{i=1}^n \frac{[f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)]^2}{q(\boldsymbol{x}_i;\theta)q(\boldsymbol{x}_i;\theta_0)}, \quad \boldsymbol{x}_i \sim q(\cdot;\theta_0)$$

Caution

Low n_e for large $\|\theta-\theta_0\|$

E.g., $p=\mathcal{N}(\theta_0,\Sigma)$ and $q=\mathcal{N}(\theta,\Sigma)$

$$n_e^* \geqslant \frac{n}{100} \iff (\theta-\theta_0)^\mathsf{T}\Sigma^{-1}(\theta-\theta_0) \leqslant \log(10) \doteq 2.30$$

O (2013), Chapter 9.14.

K rounds

Estimates: $\hat{\mu}_k$, $k = 1, \dots, K$ (unbiased)

$K = 2$ pilot and final

$K = n$ continual adaptation

Best linear unbiased combination

$$\sum_{k=1}^{K} \frac{\hat{\mu}_k}{\mathrm{Var}(\hat{\mu}_k)} \bigg/ \sum_{k=1}^{K} \frac{1}{\mathrm{Var}(\hat{\mu}_k)}$$

It is **not** safe to replace $\mathrm{Var}(\hat{\mu}_k)$ by $\widehat{\mathrm{Var}}(\hat{\mu}_k)$.
$\widehat{\mathrm{Var}}(\hat{\mu}_k)$ typically correlated with $\hat{\mu}_k$.

$\sqrt{k}$ weights

Use $\displaystyle \sum_{k=1}^{K} \sqrt{k}\hat{\mu}_k \Big/ \sum_{k=1}^{K} \sqrt{k}.$ O & Zhou (1999)

Near optimal

If $\mathrm{Var}(\hat{\mu}_k) \propto k^{-r_0}$ unknown $0 \leqslant r_0 \leqslant 1$

Then $$\sup_{1\leqslant K<\infty} \max_{0\leqslant r_0\leqslant 1} \frac{\mathrm{Var}(\text{using } r_1 = 1/2)}{\mathrm{Var}(\text{using } r_0)} = \frac{9}{8}.$$

AMIS

Adaptive Multiple Importance Sampling

Cornuet, Marin, Mira, Robert (2012)

1) Sample n_1 observations using θ_1
2) Estimate θ_2 from data and sample more
3) Keep estimating new θ_k and sampling more
4) Combine rounds by multiple importance sampling methods

Notes

Observation weights from one round depend on data from future rounds.

This breaks the Martingale property, so it is hard to get unbiased estimates.

SNIS is used throughout because the motivation is from Bayesian problems where p is usually not normalized.

APIS

Martino, Elvira, Luengo, Corander (2015)

Adaptive Population Importance Sampler

For an unnormalized $p(\cdot)$.

Choose n normalized distributions $q_i(\cdot\,;\theta_i,C_i)$, mean θ_i, covariance C_i.

Sample $\boldsymbol{x}_i \sim q_i$

Get SNIS weights $w_i \propto p(\boldsymbol{x}_i)/\sum_{i'} q_{i'}(\boldsymbol{x}_i;\theta_{i'},C_{i'})$.

Every m'th iteration, update the means θ_i but not the covariances C_i,
using previous $m-1$ iterations' data

Avoids "particle collapse".

Empirical assessment.

Cross-entropy

One of the most popular methods.

Rubinstein (1997), Rubinstein & Kroese (2004)

$\mu = \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ for $f \geqslant 0$ and $\mu > 0$.

We use $q(\cdot\,;\theta) = q_\theta$ for $\theta \in \Theta$. **Exponential family**

There is an optimal $q \propto fp$ but it is not usually in our family.

Variance based update

$$\theta^{(k+1)} = \arg\min_{\theta\in\Theta} \frac{1}{n_k}\sum_{i=1}^{n_k} \frac{(f(\boldsymbol{x}_i)p(\boldsymbol{x}_i))^2}{q_\theta(\boldsymbol{x}_i)}, \qquad \boldsymbol{x}_i = \boldsymbol{x}_i^{(k)} \sim q_{\theta^{(k)}}$$

The optimization may be too hard. Switch to a Kullback-Leibler distance

Update reduces to moment matching.

Skipping a ton of notation!

Cross-entropy

A common estimand is

$\mu = \mathbb{P}(g(\boldsymbol{x}) \geqslant \tau)$ for large τ, so

$f(\boldsymbol{x}) = 1\{g(\boldsymbol{x}) \geqslant \tau\}$

In the moment update:

$\theta \leftarrow$ a weighted average of $f(\boldsymbol{x}_i)$

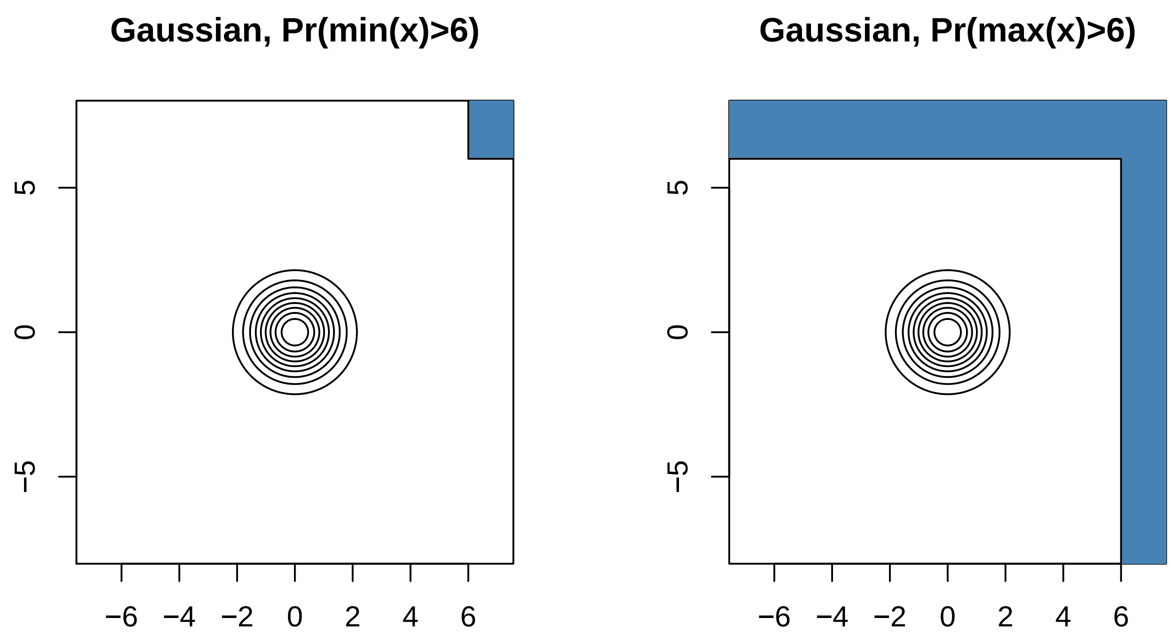
If $\tau \geqslant \max_i g(\boldsymbol{x}_i)$

Oops: $\theta \leftarrow 0/0$

Ingenious fix

They reduce τ to a high quantile of $g(\boldsymbol{x}_i)$ and go again.

Cross-ent examples

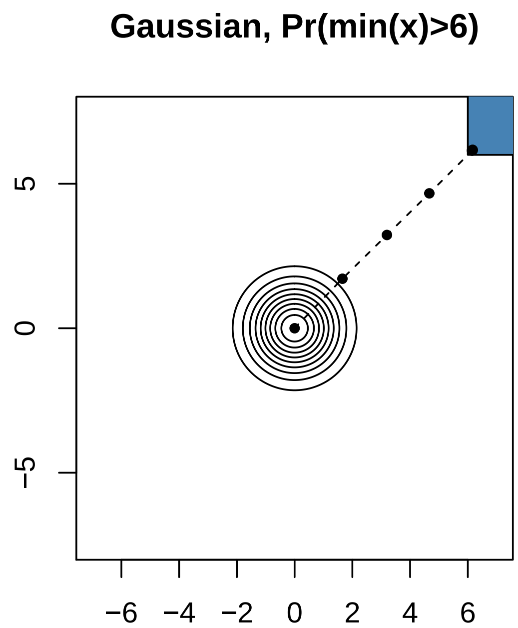


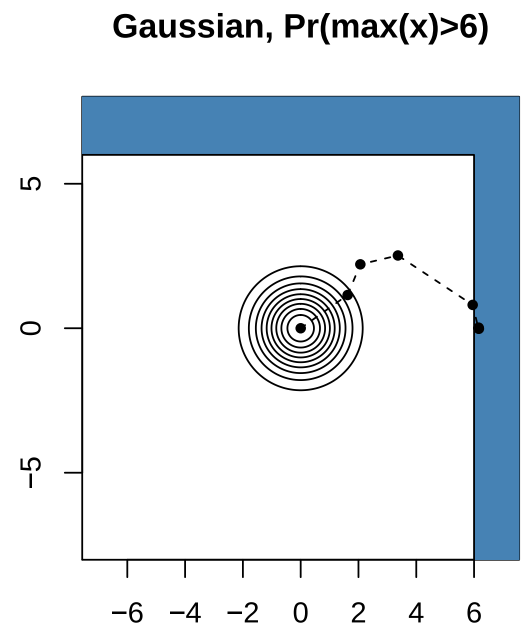
$\boldsymbol{x} \sim \mathcal{N}(\theta, I_2) \quad \theta = (0,0)^{\mathsf{T}}$.

Start with $\theta_1 = \theta = (0,0)^{\mathsf{T}}$.

Take $K = 10$ steps with $n = 1000$ each.

Cross-ent examples





$\theta_1 = (0, 0)^{\mathsf{T}}$.

For $\min(\boldsymbol{x})$, θ_k heads Northeast, and is ok.

For $\max(\boldsymbol{x})$, θ_k heads North (or East) and underestimates μ by about $1/2$.

More adaptive methods

There are enormously many of them. Still an active area.

Additional refs in online notes.

Also in slides for Los Alamos 2017 Winter School `http://statweb.stanford.edu/~owen/pubtalks/AdaptiveISweb.pdf`

More ideas

- Asymptotically exact IS in particle transport
  Booth (1985), Kollman (1993), Kong, Ambros, Spanier (2009)
- Nonparametric AIS and recursive partitioning
  Lepage (1978), Friedman & Wright (1979) Press & Farrar (1990)
- Stochastic convex programming
  Ryu & Boyd (2015)

Apologies to many left off the list.

Thanks

- Lecturers: Nicolas Chopin, Mark Huber, Jeffrey Rosenthal
- Guest speakers: Michael Giles, Gareth Roberts
- The London Mathematical Society: Elizabeth Fisher, Iain Stewart
- CRISM & The University of Warwick, Statistics
- Sponsors: Amazon, Google
- Partners: ISBA, MCQMC, BAYSM
- Poster: Talissa Gasser, Hidamari Design
- NSF: DMS-1407397 & DMS-1521145
- Planners: Murray Pollock, Christian Robert, Gareth Roberts
- Support: Paula Matthews, Murray Pollock, Shahin Tavakoli