Adaptive Importance Sampling

Art B. Owen

Stanford University

These are slides that I presented at Grid Science 2017 in Los Alamos NM, on Wednesday January 11.

I have amended them correcting typos and adding a few more references. I also add a few interstitial slides like this one to cover things that I either said or should have said during the presentation.

Outline

1) Basic Importance Sampling.

   often our only chance to solve a problem.

2) Mixture Distributions.

   the key to effective use of IS. Also an Oscar.

3) What-if Simulations.

   using data from one sim to learn how another would have worked.

4) Adaptive Importance Sampling.

   Still mostly intuitive and ad hoc.

What is missing

1) Some paywalled articles.

2) Many articles from application areas.

3) Some very advanced / hard to explain developments.

Google Scholar, Dec 2016

“adaptive importance sampling”

About 3,040 results (0.07 sec)

Since 2012: About 1,070 results (0.10 sec)

At the meeting, several people told me about the importance sampling work of Ian Dobson on estimating rare event probabilities for power system cascades. So I'm looking into those now.

The canonical problem

$$\text{Approximate} \quad \mu = \int_{\mathbb{R}^d} h(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

Use the first one that works of

1) Symbolic / closed form solutions
2) Numerical quadrature
3) Plain Monte Carlo
4) Importance Sampling
5) Adaptive Importance Sampling

Notes

- Some problems have $d = \infty$ and / or discrete $\boldsymbol{x}$
- We may need Markov chain Monte Carlo (MCMC) or sequential Monte Carlo (SMC).
- Sometimes we can use quasi-Monte Carlo (QMC).

There are problems where none of the five approaches on the previous slide will get the answer.

Quasi-Monte Carlo sampling can be used to great effect when the underlying integrands are just a little bit more regular than Monte Carlo uses. With randomized QMC, RQMC, one gets an answer that is seldom worse than plain MC or plain QMC and can be much better. This talk was just about MC.

MCMC and SMC can be brought to bear on problems that are less regular/favorable than plain MC uses. Those methods can handle even harder problems because they involve flexible explorations of the problem space. It can be hard to tell if/when they work.

Closed form

Sometimes we can get a closed form.

Symbolic computation (Mathematica, Maple, Sage, etc.) help.

Unfortunately

We can seldom do this for problems with real-world complexity,

so we resort to numerical methods.

Numerical quadrature

For instance tensor products of Simpson's rule.

Evaluate f at n points $\boldsymbol{x}_i \in \mathbb{R}^d$.

Can attain an error rate of $O(n^{-r/d})$ for $h \in C^{(r)}(\mathbb{R}^d)$.

Unfortunately

Low smoothness (small r) or high dimension (large d) make the methods ineffective.

Monte Carlo

Write $h(\boldsymbol{x}) = f(\boldsymbol{x})p(\boldsymbol{x})$ for a probability density function $p(\cdot)$. Then,

$$\mu = \int h(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \mathbb{E}(f(\boldsymbol{x})), \quad \boldsymbol{x} \sim p.$$

Estimate

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

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2), \quad \text{CLT} \quad \sigma^2 = \mathbb{E}((f(\boldsymbol{x}) - \mu)^2)$$

Upshot

Very competitive rate $O(n^{-1/2})$ when $r/d \ll 1/2$. We get confidence intervals too.

Unfortunately

$\dfrac{\sigma/\sqrt{n}}{\mu}$ might be large, or even infinite.

Places where MC has trouble

Rare events

- Bankruptcy / default for finance / insurance
- Energy grid failure Bent, Backhaus & Yamangil (2014),
  cascade simulations Chertkov (2012)
- Product failure, e.g., battery
- Monograph by Rubino & Tuffin (2009)

Spiky/singular integrands

- High energy physics (Feynman diagram),
  Aldins, Brodsky, Dufner, Kinoshita (1970)
  multiple singularities in $[0,1]^7$ that dominate the integral
- Graphical rendering
  Kollig & Keller (2006) Integrand includes factors like $\|\boldsymbol{x} - \boldsymbol{x}_0\|^{-1}$
- Particle transport, e.g., T. Booth

Rare events

$$f(\boldsymbol{x}) = 1_A(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{x} \in A \\ 0, & \text{else}, \end{cases}$$

$$\mu = \mathbb{P}(\boldsymbol{x} \in A) = \int_A p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \epsilon$$

Coefficient of variation of $\hat{\mu}$

$$\frac{\sigma/\sqrt{n}}{\mu} = \frac{\sqrt{\epsilon(1-\epsilon)}}{\sqrt{n\epsilon}} \approx \frac{1}{\sqrt{n\epsilon}}$$

To get $\mathrm{cv} = 0.1$ takes $n \geqslant 100/\epsilon$.

Singularities

$$f(\boldsymbol{x}) \approx \|\boldsymbol{x} - \boldsymbol{x}_0\|^{-r}, \quad \boldsymbol{x} \in A \subset \mathbb{R}^d$$
$$0 < a \leqslant p(\boldsymbol{x}) \leqslant b < \infty, \quad \text{for } \boldsymbol{x} \in A.$$

f is integrable over bounded A containing $\boldsymbol{x}_0$ iff $r < d$.

Then f^2 is integrable iff $2r < d$.

For $r \in [d/2, d)$

μ exists but $\sigma^2 = \infty$

Monte Carlo converges but slower than $O_p(1/\sqrt{n})$.

Other singularities

$$\|\boldsymbol{x} - \boldsymbol{x}_0\|_p{}^{-r}, \qquad \min(x_1, x_2, \ldots, x_d)^{-r},$$
$$\min(x_1, 1 - x_1, x_2, 1 - x_2, \ldots, x_d, 1 - x_d)^{-r}$$

Very often singularities are simply not a big problem for MC and the interesting thing is that high dimension can make singularities **less** harmful. This is a 'concentration of measure' result as described below.

Let f be singular on some subset of the input space. For simplicity, take subsets of the form $\Omega_{d,p}(\rho) = \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\|_p \leqslant \rho\}$ for integer $d \geqslant 1$ and $1 \leqslant p \leqslant \infty$ and $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_p^{-r}$.

Let $\boldsymbol{x}$ be uniformly distributed on $\Omega_{d,p} = \Omega_{d,p}(1)$ of volume $|\Omega_{d,p}|$. Then $\mu = \int_0^1 y^{d-1-r}\,\mathrm{d}y|\Omega_{d-1,p}|/|\Omega_{d,p}| = (d-r)^{-1}|\Omega_{d-1,p}|/|\Omega_{d,p}|$. The singularity is severe if, say, half of μ comes from a very tiny region around 0.

The radius ρ that captures half of the integral satisfies $\int_0^\rho y^{-r+d-1}\,\mathrm{d}y = (1/2)\int_0^1 y^{-r+d-1}\,\mathrm{d}y$, and it is $\rho = 2^{-1/(d-r)}$. The relative volume is $|\Omega_{d,p}(\rho)|/|\Omega_{d,p}(1)| = \rho^d = 2^{-d/(d-r)}$. If r is close to d then half of the integral comes from a very tiny fraction of the volume. If $r = 1$ and d grows then half of the integral comes from about half of the volume, hence, not much of a spike. If $r < d/2$, so the variance is finite, then this volume is larger than $2^{-d/(d-d/2)} = 1/4$. Once again, not a very prominent spike.

For singularities on a k dim manifold replace d by $d-k$.

Conclusion: rare events described by $0/1$ random variables can be much harder to handle than unbounded random variables.

Importance sampling

For rare events and (some) singularities there is a region A of relatively small $\mathbb{P}(\boldsymbol{x} \in A)$ that dominates the integral. Taking $\boldsymbol{x} \sim p$ does not get enough data from the **important** region A.

Main idea

Get more data from A, and then correct the bias.

Kahn (1950), Kahn & Marshall (1953), Trotter & Tukey (1956)

Good news Importance sampling can solve these sampling problems.

Bad news It can also fail spectacularly.

For a density q

$$f(\boldsymbol{x})p(\boldsymbol{x}) = f(\boldsymbol{x})\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}q(\boldsymbol{x})$$

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

For now, assume that $p(\boldsymbol{x})/q(\boldsymbol{x})$ is computable.

Unbiasedness

If $q(\boldsymbol{x}) > 0$ whenever $f(\boldsymbol{x})p(\boldsymbol{x}) \neq 0$ then for $\boldsymbol{x}_i \sim q$,

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

Sufficient condition

$q(\boldsymbol{x}) > 0$ whenever $p(\boldsymbol{x}) > 0$.

Yields unbiasedness for generic f.

Notation

$\mathbb{E}_q(\cdot)$, $\mathrm{Var}_q(\cdot)$ etc. are for $\boldsymbol{x}_i \sim q$.

$$w(\boldsymbol{x}) \equiv \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}, \quad \mathbb{E}_p(g(\boldsymbol{x})) = \mathbb{E}_q(g(\boldsymbol{x})w(\boldsymbol{x})).$$

Variance of IS

Based on O (2013), Ch 9.1

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

$$\sigma_q^2 = \mathbb{E}_q\left(\left(\frac{fp}{q}\right)^2\right) - \mu^2 = \int \frac{(fp)^2}{q}\,\mathrm{d}\boldsymbol{x} - \mu^2$$

After some algebra

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

Lessons

1) $q(\boldsymbol{x}) \propto f(\boldsymbol{x})p(\boldsymbol{x})$ is best, when $f \geqslant 0$.
2) q near 0 is dangerous.
3) Bounding $(fp)^2/q$ is useful theoretically.

Adaptive IS is about using sample data to select q.

Proportionality

For $f \geqslant 0$, taking $q = fp/\mu$ yields $\sigma_q^2 = 0$. Specifically

$$\hat{\mu}_q = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)\frac{p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)\frac{p(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)/\mu} = \mu$$

but we need to know μ to use it.

We look for q nearly proportional to fp.

More generally

$q \propto |f|p$ is always optimal, but not zero variance if f takes both positive and negative values.

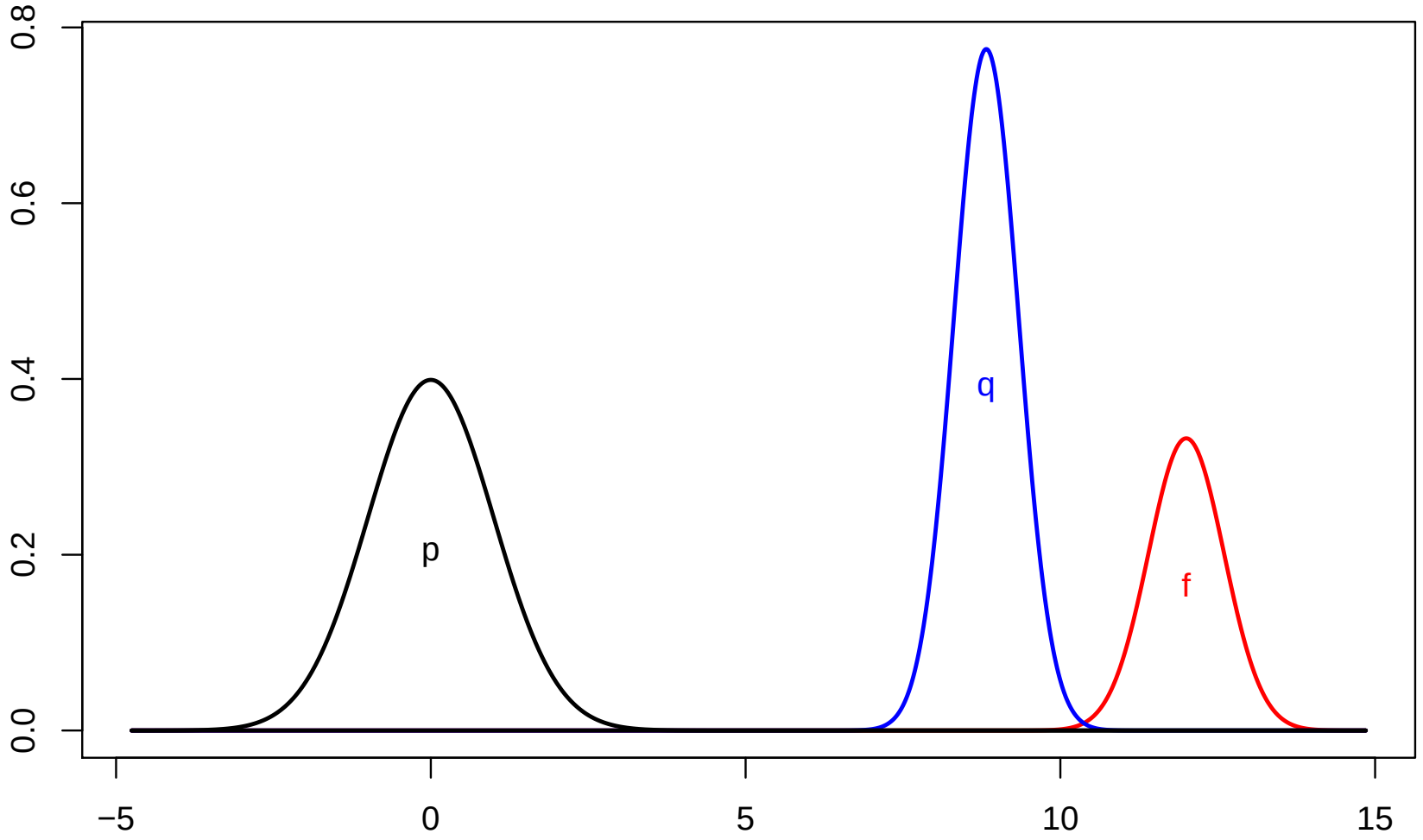
For $f \geqslant 0$

$q \propto f$ is not best (unless p is flat)

$q \propto p$ is not best (unless f is flat)

Importance involves **both** f and p via fp.

Example



$\mu \doteq 1.74 \times 10^{-24}, \quad \mathrm{Var}(\hat{\mu}_q) = 0$

Gaussian p and $f \implies$ Gaussian optimal q lies between

Here is a positivisation trick from O & Zhou (2000).

If f takes both positive and negative values then we can write

$$\mu = \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int f_+(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int f_-(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

for nonnegative functions $f_+(\boldsymbol{x}) = \max(0, f(\boldsymbol{x}))$ and $f_-(\boldsymbol{x}) = \max(0, -f(\boldsymbol{x}))$. Then we can use

$$\hat{\mu} = \frac{1}{n_+}\sum_{i=1}^{n_+}\frac{f_+(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{q_+(\boldsymbol{x}_i)} - \frac{1}{n_-}\sum_{i=1}^{n_-}\frac{f_-(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{q_-(\boldsymbol{x}_i)}$$

for $\boldsymbol{x}_i$ from $q_\pm$. There then exist zero variance sampling distributions $q_\pm$ that we might approximate. More generally we can use

$$\hat{\mu} = c + \frac{1}{n_+}\sum_{i=1}^{n_+}\frac{(f(\boldsymbol{x}_i)-c)_+p(\boldsymbol{x}_i)}{q_+(\boldsymbol{x}_i)} - \frac{1}{n_-}\sum_{i=1}^{n_-}\frac{(f(\boldsymbol{x}_i)-c)_-p(\boldsymbol{x}_i)}{q_-(\boldsymbol{x}_i)}$$

or

$$\hat{\mu} = \mathbb{E}_p(g(\boldsymbol{x})) + \frac{1}{n_+}\sum_{i=1}^{n_+}\frac{(f(\boldsymbol{x}_i)-g(\boldsymbol{x}_i))_+p(\boldsymbol{x}_i)}{q_+(\boldsymbol{x}_i)} - \frac{1}{n_-}\sum_{i=1}^{n_-}\frac{(f(\boldsymbol{x}_i)-g(\boldsymbol{x}_i))_-p(\boldsymbol{x}_i)}{q_-(\boldsymbol{x}_i)}$$

for some g with known expectation.

Tails of q

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

The term depending on q

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

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

$$= \int f^2 w p \, \mathrm{d}\boldsymbol{x} = \mathbb{E}_p(f^2 w)$$

Consequence

$$\mu^2 + \sigma_q^2 = \mathbb{E}_p(f^2 w) = \mathbb{E}_q(f^2 w^2)$$

$\therefore$ bounding $w(\boldsymbol{x}) = \dfrac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$ is advantageous for many f.

Gaussian p with t distributed q

Multivariate Gaussian, θ, Σ

$$p(\boldsymbol{x}) \propto \exp(-(\boldsymbol{x}-\theta)^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{x}-\theta)/2)$$

Multivariate t , θ, Σ, degrees of freedom $\nu > 0$

$$q(\boldsymbol{x}) \propto (1+(\boldsymbol{x}-\theta)^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{x}-\theta))^{-(\nu+d)/2}$$

The t distribution has heavier tails than the Gaussian

$\sup_{\boldsymbol{x}} p(\boldsymbol{x})/q(\boldsymbol{x})$ is bounded

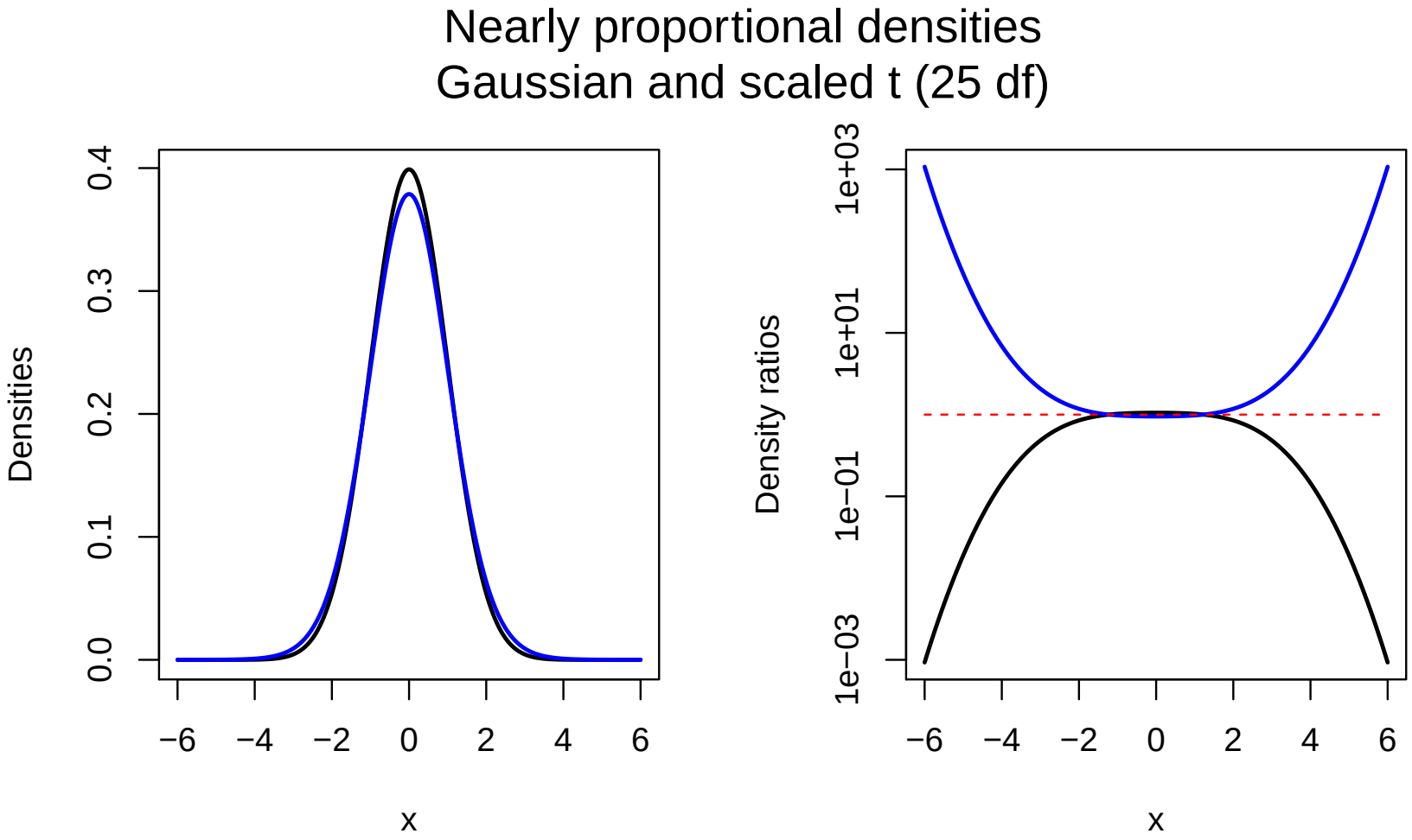
$\sup_{\boldsymbol{x}} p(\boldsymbol{x})\exp(\theta^{\mathsf{T}}\boldsymbol{x})/q(\boldsymbol{x})$ is bounded

Geweke (1989) Split t (at 0)

Reversing roles of t and Gaussian

Leads to unbounded w

q nearly proportional to p



Light tailed q can make p/q diverge.

Can have $\sigma_q^2 = \infty$ for $f(\boldsymbol{x}) = 1$.

Ironically it is the **unimportant** region that causes trouble.

Beyond variance

Chatterjee & Diaconis (2015) show that the sample size required is roughly $n \approx \exp(\text{KL distance } p, q)$ for generic f.

This analysis involves $\mathbb{E}_q(|\hat{\mu}_q - \mu|)$ and $\mathbb{P}_q(|\hat{\mu}_q - \mu| > \epsilon)$ instead of $\mathrm{Var}_q(\hat{\mu}_q)$. It avoids working with variances that can be harder to handle than means.

Taking $\epsilon = .025$ in Theorem 1.2 (for 95% confidence of a small error) 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).

Self-normalized importance sampling

Based on O (2013, ch 9.2)

$$p(\boldsymbol{x}) = c_p p_u(\boldsymbol{x}), \quad q(\boldsymbol{x}) = c_q q_u(\boldsymbol{x}), \quad 0 < c_p, c_q < \infty$$

$$\hat{\mu}_q = \frac{c_p}{c_q} \frac{1}{n} \sum_{i=1}^{n} \frac{f(\boldsymbol{x}_i) p_u(\boldsymbol{x}_i)}{q_u(\boldsymbol{x}_i)}.$$

If p_u, q_u computable, but c_p/c_q unknown

$$w_u(\boldsymbol{x}_i) = w_{u,q}(\boldsymbol{x}_i) = p_u(\boldsymbol{x}_i)/q_u(\boldsymbol{x}_i)$$

$$w_i = w_{i,q} = w_u(\boldsymbol{x}_i) \Big/ \sum_{i'=1}^{n} w_u(\boldsymbol{x}_{i'})$$

$$\tilde{\mu}_q = \sum_{i=1}^{n} w_i f(\boldsymbol{x}_i)$$

Self-normalized importance sampling

- $\tilde{\mu}_q$ has some bias, asymptotically negligible
- Self-normalized sampling competes with acceptance-rejection
- It does not attain zero variance for any q (and non-trivial f)

Adaptation

There are adaptive versions of both IS and SNIS.

SNIS is consistent

by the Law of Large Numbers

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

Recall the weights

$$
w_i = \frac{p_u(\boldsymbol{x}_i)}{q_u(\boldsymbol{x}_i)} \;\Big/\; \sum_{i'=1}^{n} \frac{p_u(\boldsymbol{x}_{i'})}{q_u(\boldsymbol{x}_{i'})}
$$

SNIS variance

Delta method variance approximation for a ratio estimate:

$$\widetilde{\mathrm{Var}}(\tilde{\mu}_q) = \frac{\sigma^2_{q,\mathrm{sn}}}{n}$$

$$\sigma^2_{q,\mathrm{sn}} = \mathbb{E}_q(w(\boldsymbol{x})^2(f(\boldsymbol{x}) - \mu)^2)$$

Rewrite and compare

$$\sigma^2_{q,\mathrm{sn}} = \int \frac{p^2}{q}(f-\mu)^2 \,\mathrm{d}\boldsymbol{x}$$

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

$$\text{vs} \quad \sigma^2_q = \int \frac{(fp - \mu q)^2}{q} \,\mathrm{d}\boldsymbol{x}$$

No q can make $\sigma^2_{q,\mathrm{sn}} = 0$ (unless f is constant).

Exactness

$$\hat{\mu}_q = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)\frac{p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)}$$

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

Constants $f(\boldsymbol{x}) = c$

$\tilde{\mu}_q = c$, while generally $\hat{\mu}_q \neq c$

We don't usually try to average constants, but getting that wrong means:

1) $\widehat{\mathbb{P}}(\boldsymbol{x} \in B) + \widehat{\mathbb{P}}(\boldsymbol{x} \notin B) \neq 1$, and
2) $\hat{\mathbb{E}}(f(\boldsymbol{x}) + c) \neq \hat{\mathbb{E}}(f(\boldsymbol{x})) + c$.

Optimal SNIS q^{opt}

$$q^{\mathrm{opt}} \propto |f(\boldsymbol{x}) - \mu| p(\boldsymbol{x})$$ Hesterberg (1988)

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

Best possible variance reduction from SNIS

$$\frac{\sigma^2}{\sigma^2_{q^{\mathrm{opt}},\mathrm{sn}}} = \frac{\mathbb{E}_p((f(\boldsymbol{x}) - \mu)^2)}{\mathbb{E}_p(|f(\boldsymbol{x}) - \mu|)^2}$$

Versus plain IS

For $f \geqslant 0$, best IS gets $\sigma^2_q = 0$.

Rare events

$$f(\boldsymbol{x}) = 1_{\boldsymbol{x}\in A}, \quad \mu = \mathbb{P}(\boldsymbol{x} \in A) = \epsilon$$

Optimal q

$$q^{\mathrm{opt}}(\boldsymbol{x}) \propto q_u^{\mathrm{opt}}(\boldsymbol{x}) = |f(\boldsymbol{x}) - \mu| p(\boldsymbol{x}) = \begin{cases} (1-\epsilon)p(\boldsymbol{x}) & \boldsymbol{x} \in A \\ \epsilon p(\boldsymbol{x}) & \boldsymbol{x} \notin A. \end{cases}$$

Normalize q

$$\int q_u^{\mathrm{opt}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 2\epsilon(1-\epsilon)$$

$$\int_A q^{\mathrm{opt}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \frac{\epsilon(1-\epsilon)}{2\epsilon(1-\epsilon)} = \frac{1}{2} \qquad (!!)$$

Versus plain IS

Best plain IS $q = pf/\mu$ puts all probability on A, not half.

Rare event efficiency

$$\sigma_{q^{\mathrm{opt}},\mathrm{sn}} = \mathbb{E}_p(|f(\boldsymbol{x}) - \mu|) = \epsilon(1-\epsilon) + (1-\epsilon)\epsilon = 2\epsilon(1-\epsilon) \approx 2\epsilon$$

Best possible coefficient of variation $\sigma_{q,\mathrm{sn}}/\mu$ is ≈ 2.

Versus plain MC as $\mu = \epsilon \to 0$

$$\frac{\sqrt{\widetilde{\mathrm{Var}}(\tilde{\mu}_{q^{\mathrm{opt}},\mathrm{sn}})}}{\mu} \approx \frac{2}{\sqrt{n}}$$ Optimal SNIS

$$\frac{\sqrt{\epsilon(1-\epsilon)/n}}{\mu} \approx \frac{1}{\sqrt{\epsilon}}\frac{1}{\sqrt{n}}$$ plain MC

Optimal SNIS would be **very good**. Gets cv 0.1 for $n = 400$ vs $100/\epsilon$

Plain IS can be even better (lower bound 0).

When to use SNIS

If we only have p_u or q_u then we cannot use IS and SNIS is available.

I.e., use SNIS when we can draw $\boldsymbol{x} \sim q$ and evaluate p_u/q_u but not p/q.

Suppose we could do both?

We can then use $\int p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \mathbb{E}_q(w(\boldsymbol{x})) = 1$ to define a **control variate**.
(Much more later.)

The resulting estimate is then exact for $q = fp/\mu$ and for $f(\boldsymbol{x}) = c$.

Asymptotically this combination beats **both** plain IS and SNIS.

That is, when we have normalized p and q, SNIS is not best.

Effects of dimension

$$p(\boldsymbol{x}) = \prod_{j=1}^{d} p_j(x_j \mid x_1, \ldots, x_{j-1})$$

$$q(\boldsymbol{x}) = \prod_{j=1}^{d} q_j(x_j \mid x_1, \ldots, x_{j-1})$$

$$w(\boldsymbol{x}) = \prod_{j=1}^{d} \frac{p_j(x_j \mid x_1, \ldots x_{j-1})}{q_j(x_j \mid x_1, \ldots x_{j-1})} \equiv \prod_{j=1}^{d} w_j(x_j \mid x_1, \ldots, x_{j-1}).$$

We can write p and q as above even if we generate them some other way.

Each $\mathbb{E}_q(w_j(\cdot \mid \cdot)) = 1$ and $\mathbb{E}_q(w_j(\cdot \mid \cdot)^2) \geqslant 1$.

As a result $\mathbb{E}_q(w(\boldsymbol{x})^2)$ can grow exponentially with d affecting both IS and SNIS.

If q_j gets closer to p_j as j increases we might be ok.

Or if f is specially related to p/q.

Gaussian p and q

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

$$w(\boldsymbol{x}) = \exp(-\theta^{\mathsf{T}}\boldsymbol{x} + \theta^{\mathsf{T}}\theta/2), \quad \mathbb{E}_q(w(\boldsymbol{x})^2) = \exp(\theta^{\mathsf{T}}\theta).$$

For $\theta = (1, 1, 1, \ldots, 1)^{\mathsf{T}}$ we get $\mathbb{E}_q(w(\boldsymbol{x})^2) = \exp(d)$.

Similar results from $\theta = (1, 0, 0, \ldots, 0)^{\mathsf{T}}$ and $\theta = (1/\sqrt{d}, 1/\sqrt{d}, \ldots, 1/\sqrt{d})^{\mathsf{T}}$.

Variance change

Taking $q = \mathcal{N}(0, \sigma^2 I)$

$$\mathbb{E}_q(w(\boldsymbol{x})^2) = \begin{cases} \left(\dfrac{\sigma^2}{\sqrt{2\sigma^2 - 1}}\right)^d, & \sigma^2 > 1/2 \\ \infty & \text{else.} \end{cases}$$

We need σ^2 closer to 1 as d grows, e.g., $1 + O(1/d)$.

Infinite dimension

$\boldsymbol{x} = (x_1, x_2, \dots)$.

The function $f(\boldsymbol{x})$ depends on only the first $M = M(\boldsymbol{x})$ components of $\boldsymbol{x}$.

Assume M is a stopping time

I.e., we can tell whether $M(\boldsymbol{x}) = m$ from $x_1, x_2, \dots, x_m$. For example:

$$M = \min\Big\{m \geqslant 1 \mid \sum_{i=1}^{m} x_i \geqslant 100\Big\}.$$

We **must** choose $q(\cdot)$ so that $\mathbb{P}_q(M < \infty) = 1$, so that we can compute f in finite time.

The estimate

$$\hat{\mu}_q = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i) \prod_{j=1}^{M(\boldsymbol{x})} \frac{p(x_j \mid x_1, \dots, x_{M(\boldsymbol{x})})}{q(x_j \mid x_1, \dots, x_{M(\boldsymbol{x})})}$$

Note: weight $w(\boldsymbol{x})$ only uses $\prod_j p/q$ for $j \leqslant M$.

We don't need the infinite product.

Infinite dimension ctd.

Technicality

$$\mathbb{E}_q(\hat{\mu}_q) = \mathbb{E}_p(f(\boldsymbol{x})1_{M(\boldsymbol{x})<\infty}).$$

If $\mathbb{P}_p(M(\boldsymbol{x}) < \infty) = 1$ then $\mathbb{E}_q(\hat{\mu}_q) = \mathbb{E}_p(f(\boldsymbol{x}))$.

The technicality is useful

Suppose we want $\mathbb{P}_p(M(\boldsymbol{x}) < \infty)$. Choose $f(\boldsymbol{x}) = 1$. Then

$$\mathbb{E}_q(\hat{\mu}) = \mathbb{P}_p(M(\boldsymbol{x}) < \infty).$$

Siegmund (1976)

Weights and effective sample size

Kong (1992)

$$\hat{\mu}_q = \frac{1}{n}\sum_{i=1}^{n} w_i f(\boldsymbol{x}_i), \quad w_i = \frac{p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)}, \quad W \equiv \sum_i w_i$$

If a small number of w_i dominate the others then $\hat{\mu}_q$ can be unstable.

An effectively smaller number of $f(\boldsymbol{x}_i)$ are being averaged.

Effective sample size

$$n_e = \frac{(\sum_i w_i)^2}{\sum_i w_i^2}$$

$n_e = n$ if all w_i are equal. $n_e = 1$ if only one w_i is positive

Slippery derivation

It follows from variance of a weighted sum of IID random variables.

In IS, the random variables and the weights are both functions of the same $\boldsymbol{x}$.

n_e is the same for IS and SNIS.

Effective sample size

$$n_e = \frac{(\sum_i w_i)^2}{\sum_i w_i^2} \approx \frac{n}{1+\mathrm{cv}(\boldsymbol{w})^2}$$

Population counterpart

$$n_e^* = \frac{n\mathbb{E}_q(w)^2}{\mathbb{E}_q(w^2)} = \frac{n}{\mathbb{E}_q(w^2)} = \frac{n}{\mathbb{E}_p(w)}$$

Larger w are like getting fewer data points.

Interpretation

Plain MC has $n_e = n$ and we don't want that. IS has $n_e < n$.

We just want to detect very badly imbalanced weights via small n_e.

Integrand specific version

Evans & Swartz (1995)

Use $w_i(f) = |f(\boldsymbol{x}_i)|p(\boldsymbol{x}_i)/q(\boldsymbol{x}_i)$, $W(f) = \sum_i w_i(f)$, $\tilde{w}_i(f) = w_i(f)/W(f)$.

$$n_e(f) = \frac{1}{\sum_{i=1}^n \tilde{w}_i(f)^2}$$

PERT example

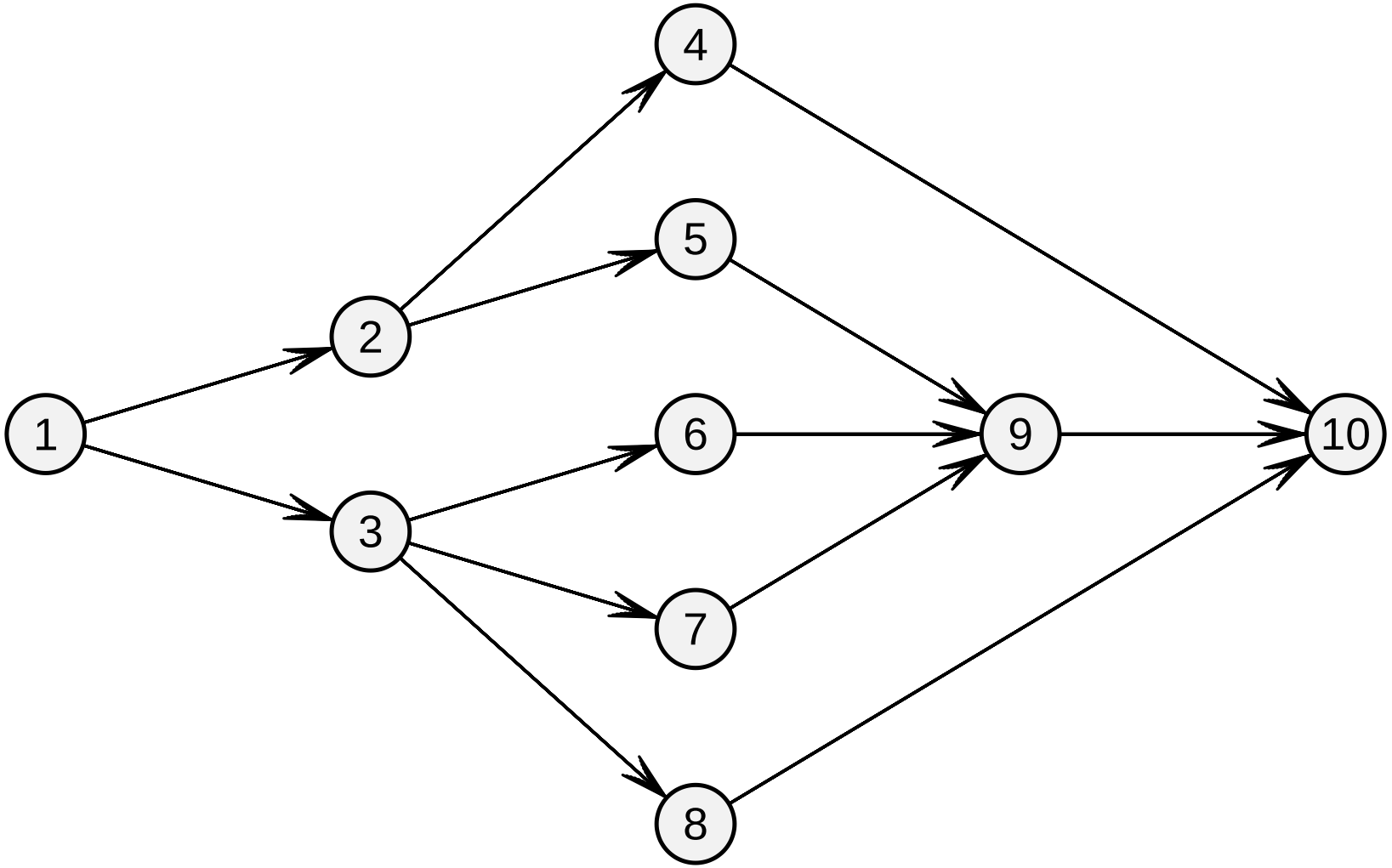
Ten steps in a sofware project. Some cannot start until previous are done.

j	Task	Predecessors	Mean Duration
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

Table 1: A PERT problem from Chinneck (2009, Chapter 11)

PERT ctd

PERT graph (activities on nodes)



PERT ctd

For given times, the project completes in $E = E_{10} = 15$ days.

(They seem optimistic.)

Start time	Time taken	End time
$S_j = \max_{k \to j} E_k$	T_j	$E_j = S_j + T_j$

E.g., from the graph: $S_9 = \max(E_5, E_6, E_7)$.

For independent exponential random variables T_j, we get $\mathbb{E}(E_{10}) \doteq 18$.

Suppose there's a large penalty when $E > 70$.

Plain MC

Got $E > 70$ twice in $n = 10{,}000$ tries.

Let's try importance sampling.

Change the exponential rates

Change T_i from exponential with mean θ_i to mean λ_i.

$$\hat{\mu}_\lambda = \frac{1}{n}\sum_{i=1}^{n} 1_{\{E_{i,10}>70\}} \prod_{j=1}^{10} \frac{\exp(-T_{ij}/\theta_i)/\theta_i}{\exp(-T_{ij}/\lambda_i)/\lambda_i}.$$

We can make $\mathbb{E}_\lambda(E_{10}) \approx 70$ by taking $\lambda = 4\theta$.

We get $\hat{\mu}_\lambda \doteq 2.01 \times 10^{-5}$ with a standard error of 4.7×10^{-6}.

Diagnostics

We also get $n_e \doteq 4.90$. This is quite low.

The largest weight was about 43% of the total.

The standard error did not warn us about this problem.

The standard error is more vulnerable to large weights than the mean is.

An effective sample size for the se is $n_{e,\sigma} \doteq 1.15$. O (2013), Chapter 9.3

Try different λ

Replace $\lambda = 4\theta$ by $\lambda = \kappa\theta$ and search for good κ. (None seemed good.)

Unequal scaling

Adding a day to some tasks does not always add a day to the project.

Details: $T_j = \theta_j + 1, T_k = \theta_k$ for $k \neq j$. Critical path: $j \in \{1, 2, 4, 10\}$.

Even doubling some tasks' time might not cause a delay.

Details: $T_j = 2\theta_j, T_k = \theta_k$ for $k \neq j$. Changes for: $j \notin \{3, 7, 8\}$.

Scale tasks $1, 2, 4, 10$ by κ

κ	3.0	3.5	4.0	4.5	5.0
$10^5 \times \hat{\mu}$	3.3	3.6	3.1	3.1	3.1
$10^6 \times \mathrm{se}(\hat{\mu})$	1.8	2.1	1.6	1.5	1.6
n_e	939	564	359	239	165
$n_{e,\sigma}$	165	88	52	33	23

Followup

Using $\kappa = 4$ on the critical path with $n = 200{,}000$ yields

$$\hat{\mu}_q = 3.18 \times 10^{-5}, \quad \mathrm{se} = 3.62 \times 10^{-7} \quad n_e \doteq 7470, \quad n_{e,\sigma} \doteq 992, \quad n_e(f) \doteq 7400$$

Vs plain MC

Plain MC has variance $\epsilon(1-\epsilon)/n$.

IS has variance about 1200-fold smaller here.

Vs SNIS

In this case SNIS had about double the estimated variance of plain IS.

Further improvements

Much more subtle tuning could be applied.

The 10'th random variable can be 'integrated out' in closed form, reducing variance.

Control variates could be applied.

Antithetics could be applied.

Summary of basic IS

- IS helps with rare events or spiky integrands
- It becomes unbiased by reweighting
- A good importance sampler q is nearly proportional to fp
- Take extreme care that q is not small
- Self-normalized IS is more widely applicable, but less effective
- Dimension brings challenges

PERT example

This was adaptive IS done by hand.

Devising q

We can go through a storehouse of distributions to find q.

For instance Devroye (1986).

General techniques

- Exponential tilting
- Laplace approximations
- Mixtures

Exponential tilting

Also called exponential twisting.

Exponential family

$$p(\boldsymbol{x};\theta) = \exp(\,\theta^{\mathsf{T}}\boldsymbol{x} - A(\boldsymbol{x}) - C(\theta)\,), \quad \boldsymbol{x} \in \mathcal{X}$$

Here $p(\boldsymbol{x}) = p(\boldsymbol{x};\theta_0)$ and $q(\boldsymbol{x}) = p(\boldsymbol{x};\theta)$.

Includes: $\mathcal{N}(\theta, I)$, Poisson(e^θ), Binomial, Gamma distributions.

More general case

$$p(\boldsymbol{x};\theta) = \exp(\,\eta(\theta)^{\mathsf{T}}T(\boldsymbol{x}) - A(\boldsymbol{x}) - C(\theta)\,), \quad \boldsymbol{x} \in \mathcal{X}$$

Here $\eta(\cdot)$ and $T(\cdot)$ are known functions.

IS with tilting

$$\begin{aligned}\hat{\mu}_\theta &= \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)\frac{p(\boldsymbol{x}_i;\theta_0)}{p(\boldsymbol{x}_i;\theta)}, \qquad \boldsymbol{x}_i \sim p(\cdot\,;\theta)\\ &= \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)\frac{\exp(\boldsymbol{x}_i^{\mathsf{T}}\theta_0 - A(\boldsymbol{x}_i) - C(\theta_0))}{\exp(\boldsymbol{x}_i^{\mathsf{T}}\theta - A(\boldsymbol{x}_i) - C(\theta))}\\ &= \frac{c(\theta)}{c(\theta_0)}\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)e^{\boldsymbol{x}_i^{\mathsf{T}}(\theta_0-\theta)}\end{aligned}$$

where $c(\theta) = \exp(C(\theta))$. This often simplifies.

For $\mathcal{N}(\theta, \Sigma)$

$$\hat{\mu}_\theta = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)\, e^{\boldsymbol{x}_i^{\mathsf{T}}\Sigma^{-1}(\theta_0-\theta)} \times \underbrace{e^{\frac{1}{2}\theta^{\mathsf{T}}\Sigma^{-1}\theta - \frac{1}{2}\theta_0^{\mathsf{T}}\Sigma^{-1}\theta_0}}_{c(\theta)/c(\theta_0)}$$

Further tilting

Define a family $q(\boldsymbol{x};\theta) \propto p(\boldsymbol{x})e^{h(\boldsymbol{x})^{\mathsf{T}}\theta}$.

The function $h(\cdot)$ defines 'features' of $\boldsymbol{x}$.

$\theta_j > 0$ makes higher values of $h_j(\boldsymbol{x})$ more probable.

$q(\boldsymbol{x};0) = p(\boldsymbol{x})$

If p is not in the exponential family, then we might not be able to compute $w(\boldsymbol{x})$.

If we can sample from $q(\boldsymbol{x};\theta)$ then we can use SNIS.

Hessian at the mode

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

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

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

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

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

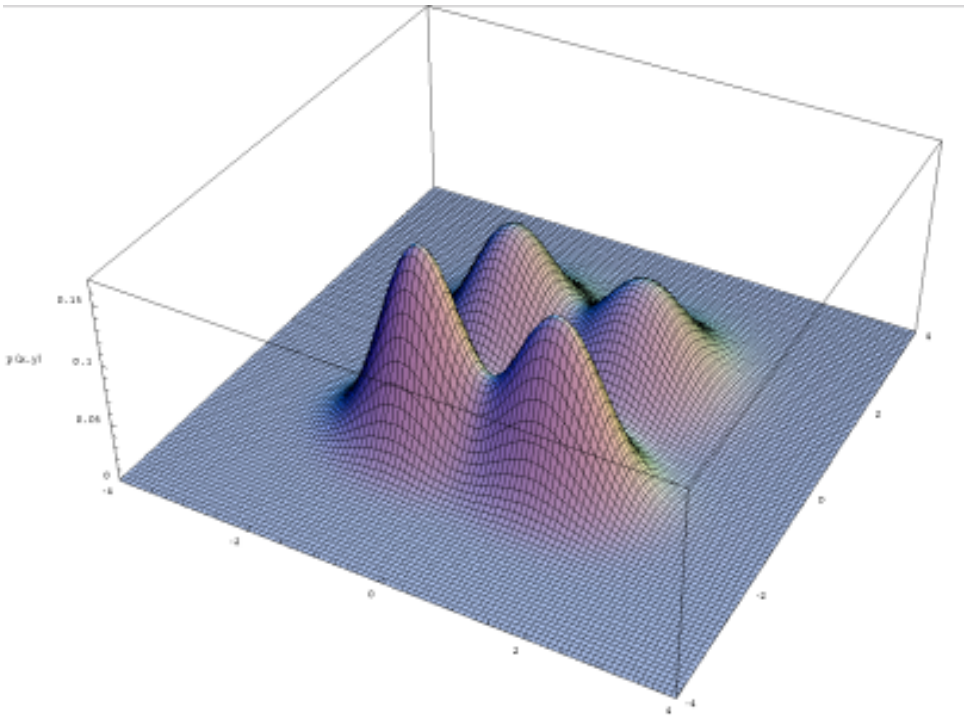
Requires positive definite H_*.

For safety

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

Mixtures

What if there are multiple modes?



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

Closely sampling the highest mode might lead us to undersample or even 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)

Multiple rare event sets

Suppose that $\boldsymbol{x}$ describes a bridge and its load.

There is a failure if $\boldsymbol{x} \in A_1$ or $\boldsymbol{x} \in A_2$.

Oversampling A_1 might lead us to miss A_2.

Finance example

Portfolio might fail to meet its goal under:

- A_1: oil sector underperforms
- A_2: dollar becomes too weak
- A_3: interest rates in China are too high
- A_4: pharmaceutical stocks do too well (portfolio shorted them)

Mixtures

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

Multiple integrands of interest

We want $\mu_j = \int f_j(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ for $j = 1, \ldots, J$.

Statistics: numerous posterior moments.

Graphics: red, green, blue for each pixel.

Grid science: J different designs.

Each $f_j(\cdot)p(\cdot)$ has its own peak.

Mixtures

One mixture component can be directed towards each integrand (that we have planned for).

We will have to take care that optimizing for f_1 does not spoil things for f_2. The same issue arises for multiple modes and failure regions.

Multiple nominal distributions

We want $\mu_j = \int f(\boldsymbol{x}) p_j(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$ for $j = 1, \ldots, J$.

The p_j may correspond to different scenarios (that we are aware of).

Mixtures

Multiple p_j and multiple f_k amount to multiple products $(pf)_{jk}$.

We can use a component for each.

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.

Alternative

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

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

It is also unbiased.

This alternative has **higher** variance.

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

Why higher variance?

Argument by convexity. Simple case $\alpha = (\frac{1}{2}, \frac{1}{2})$ for q_1 and q_2.

Recall: $\sigma_q^2 = \int (fp)^2/q \, \mathrm{d}\boldsymbol{x} - \mu^2$

Horvitz Thompson

$$\frac{1}{n}\left(\int \frac{(fp)^2}{(q_1+q_2)/2}\,\mathrm{d}\boldsymbol{x} - \mu^2\right)$$

Alternative

$$\frac{1}{2n}\left(\int \frac{(fp)^2}{q_1}\,\mathrm{d}\boldsymbol{x} - \mu^2\right) + \frac{1}{2n}\left(\int \frac{(fp)^2}{q_2}\,\mathrm{d}\boldsymbol{x} - \mu^2\right)$$

Horvitz-Thompson is better because $1/x$ is convex.

We gave the alternative a slight advantage by taking exactly $n/2$ observations from each q_j.

Defensive mixtures

From Hesterberg (1995)

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

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

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

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

Variance bounds

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

$$\mathrm{Var}(\hat{\mu}_{q_\alpha,\mathrm{sn}}) \leqslant \frac{1}{n\alpha_1}\sigma_{p,\mathrm{sn}}^2 = \frac{1}{n\alpha_1}\sigma_p^2$$

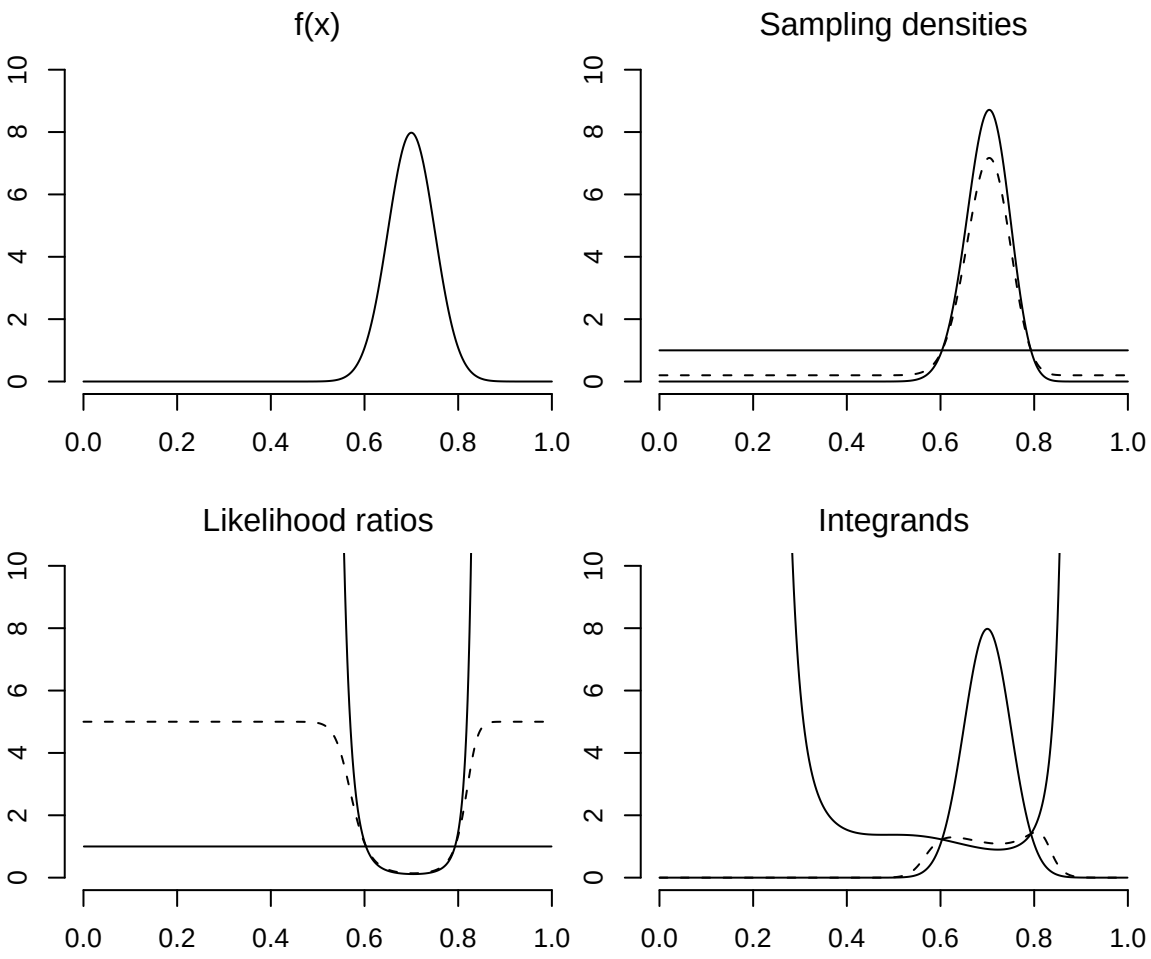
$$\mathrm{Var}(\hat{\mu}_{q_\alpha,\mathrm{sn}}) \leqslant \frac{1}{n\alpha_2}\sigma_{q,\mathrm{sn}}^2$$

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

We don't have a bound vs σ_q^2.

Loss of proportionality

Defensive importance sampling



- f Gaussian near 0.7
- p Uniform
- q is Beta $\approx f$
- q_α is dotted mixture
- $\alpha = (0.2, 0.8)$
- fp/q unbounded
- q_α not very proportional to fp
- **Solution = control variates.**

Control variates

Suppose that we know $\mathbb{E}(h(\boldsymbol{x})) = \theta$ in closed form.

If $h \approx f$ we can Monte Carlo the difference

$$\frac{1}{n}\sum_{i=1}^{n}\Big(f(\boldsymbol{x}_i) - h(\boldsymbol{x}_i)\Big) + \theta$$

for greatly reduced variance.

and h is similar to f then we can use it to estimate $\mathbb{E}(f(\boldsymbol{x}))$.

More generally

For known $\mathbb{E}(h(\boldsymbol{x})) = \theta \in \mathbb{R}^d$, let

$$\hat{\mu}_\beta = \frac{1}{n}\sum_{i=1}^{n}\Big(f(\boldsymbol{x}_i) - \beta^{\mathsf{T}} h(\boldsymbol{x}_i)\Big) + \beta^{\mathsf{T}}\theta$$

Optimal β can be estimated from the same data.

It is easy to see that $\mathbb{E}(\hat{\mu}_\beta) = \mu$ (unbiased) for any β.

The optimal β minimizes $\mathrm{Var}(\hat{\mu}_\beta)$. This β^{opt} is unknown and could well be harder to estimate than it is to estimate our ultimate goal μ.

A harder intermediate problem is usually to be avoided, but in this case we have a loop hole. Getting β wrong by ϵ will mean our variance is too large by a factor of $1 + O(\epsilon^2)$. We do not need a precise estimate of β. The typical ϵ is proportional to $1/\sqrt{n}$. We can ordinarily analyze control variates as if we were using the true unknown β.

[Aside: this asymptotic approximation is poor if the dimension of β is comparable to n.]

Control variates via regression

Define $Y_i = f(\boldsymbol{x}_i)$ and $Z_{ij} = h_j(\boldsymbol{x}_i) - \theta_j$.

$$\text{Fit} \quad Y_i \doteq \beta_0 + \sum_{j=1}^{J} \beta_j Z_{ij} \quad \text{by least squares.}$$

The intercept $\hat{\beta}_0$ is $\widehat{\mathbb{E}}(\mu)$.

Regression code also gives a standard error for the estimate.

Details in O (2013), Algorithm 8.3.

Centering Z_{ij} is essential.

Using estimated β

$\hat{\beta} - \beta^{\mathrm{opt}} = O_p(n^{-1/2})$, so we are about $n^{-1/2}$ from the optimum.

σ^2_β is quadratic in $\beta \implies \sigma^2_{\hat{\beta}} = \sigma^2_{\beta^{\mathrm{opt}}}(1 + O(1/n))$.

Control variates and mixture IS

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

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

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 $\sigma^2_{q_j} = 0$ 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.)

Multiple importance sampling

Veach & Guibas (1995)

Widely used in computer graphics.

Eric Veach got an Oscar for it in 2014.

$$\tilde{\mu} = \sum_{j=1}^{J} \frac{1}{n_j} \sum_{i=1}^{n_j} \omega_j(\boldsymbol{x}_{ij}) \frac{f(\boldsymbol{x}_{ij})p(\boldsymbol{x}_{ij})}{q_j(\boldsymbol{x}_{ij})}$$

$q_j =$ j'th light path sampler

Partition of unity

$\omega_j(\boldsymbol{x}) \geqslant 0$ and $\sum_{j=1}^{J} \omega_j(\boldsymbol{x}) = 1, \forall \boldsymbol{x}$

Can attain Horvitz-Thompson by 'balance heuristic' $\omega(\cdot) \propto n_j q_j(\cdot)$

The weight on $f(\boldsymbol{x}_{ij})p(\boldsymbol{x}_{ij})$ can depend on j in many ways.

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.

What-if simulations

We estimated $\hat{\mu} = \int f(\boldsymbol{x}) p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$ using $\boldsymbol{x}_i \sim q$

From the sample values $f(\boldsymbol{x}_i)$ we can estimate

1) $\int f(\boldsymbol{x}) p'(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$ for a different density $p' \neq p$
2) $\mathrm{Var}_q(\hat{\mu}_q)$ for $\boldsymbol{x}_i \sim q$
3) $\mathrm{Var}_{q'}(\hat{\mu}_{q'})$ for $\boldsymbol{x}_i \sim q'$ for a different density $q' \neq q$

Estimating what would have happened with a different q is the key to adaptive IS.

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 the others

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

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 simulations

The may be accurate for θ near θ_0 but otherwise may fail.

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

$n_e^* \geqslant \dfrac{n}{100}$ requires $(\theta - \theta_0)^\mathsf{T} \Sigma^{-1} (\theta - \theta_0) \leqslant \log(10) \doteq 2.30$ O (2013), Chapter 9.14.

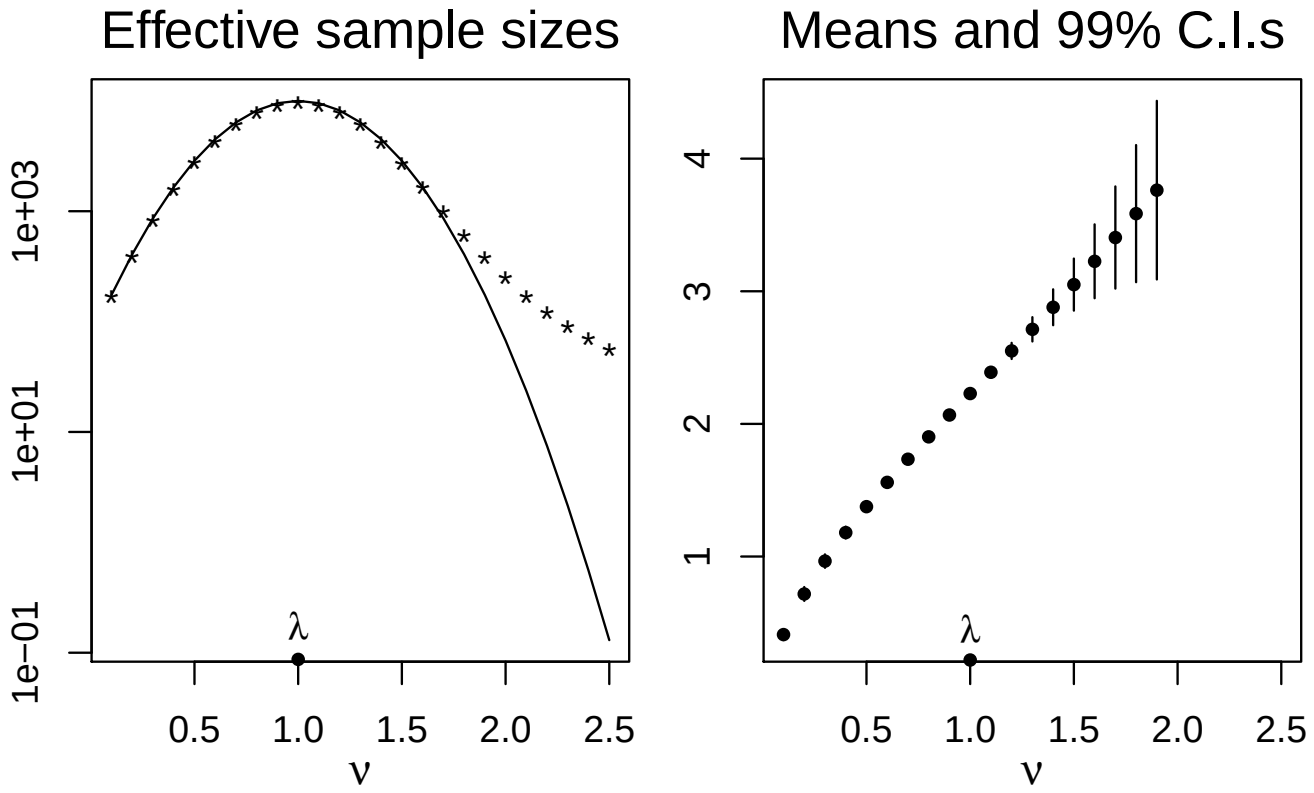
Example

$f(\boldsymbol{x}_i; \nu) = \max(x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$ where $x_{ij} \sim \text{Poi}(\nu)$.

Importance sample from λ instead of ν.

$$\begin{aligned}
\hat{\mu}_\nu &= \frac{1}{n} \sum_{i=1}^{n} \max(x_{i1}, \ldots, x_{i5}) \prod_{j=1}^{5} \frac{e^{-\nu} \nu^{x_{ij}}}{x_{ij}!} \Big/ \frac{e^{-\lambda} \lambda^{x_{ij}}}{x_{ij}!} \\
&= \frac{1}{n} \sum_{i=1}^{n} \max(x_{i1}, \ldots, x_{i5}) e^{5(\lambda - \mu)} \left(\frac{\nu}{\lambda} \right)^{\sum_j x_{ij}}
\end{aligned}$$

Poisson example



$x_{ij} \sim \text{Poi}(\lambda)$

Solid curve is pop'n effective sample size. Asterisks are sample estimates.

When $\nu > \lambda$ then q has lighter tails than p. Hence wide CIs and poor $\hat{n}_e$.

What-if for variance

For $q(\boldsymbol{x};\theta)$, $\sigma_\theta^2 = \int \frac{(f(\boldsymbol{x})p(\boldsymbol{x}))^2}{q(\boldsymbol{x};\theta)}\,\mathrm{d}\boldsymbol{x} - \mu^2 \equiv \mathrm{MS}_\theta - \mu^2$

Now

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

Sample estimate

For $\boldsymbol{x}_i \sim q(\boldsymbol{x};\theta_0)$,

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

Summary of what-if simulations

- We can learn about alternative p's and q's
- Effectiveness is over a limited range

The next slides are a survey of a selected set of adaptive importance sampling methods.

These are the ones that I thought were interesting and potentially useful in grid science.

Of course there are many others.

Adaptive importance sampling

Overview:

Use the $\boldsymbol{x}_i$ sampled so far to change q and sample some more.

Two stage

Get a pilot sample $\boldsymbol{x}_{i1} \sim q_1$.

Use them to choose q_2.

Take $\boldsymbol{x}_{i2} \sim q_2$ to estimate $\hat{\mu}$.

Multi-stage

Given q_1, for $k = 1, \ldots, K-1$, use $\boldsymbol{x}_{ik}$, $i = 1, \ldots, n_k$ to choose q_{k+1}.

Or use $\boldsymbol{x}_{ij}$, $1 \leqslant i \leqslant n_j$, $1 \leqslant j \leqslant k$ to choose q_{k+1}.

n-stage

Given q_1, take $\boldsymbol{x}_i$ from q_i computed from $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}$.

AIS details

We have to pick a family $\mathcal{Q}$ of proposal distributions.

We need an initial $q \in \mathcal{Q}$.

We have to choose a termination criterion:

e.g., K steps, total N observations, confidence interval width below ϵ.

We need a way to choose $q_{k+1} \in \mathcal{Q}$ from prior data.

We need a way to combine information from all K stages.

Combination

Suppose we get $\hat{\mu}_k$ for $k = 1, \ldots, K$.

Unbiased with variances σ_k^2.

Best linear unbiased combination

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

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

It is common for $\widehat{\mathrm{Var}}(\hat{\mu}_k)$ to be correlated with $\hat{\mu}_k$.

E.g., largest $\hat{\mu}_k$ may coincide with largest $\widehat{\mathrm{Var}}(\hat{\mu}_k)$.

Plug-in estimates could bias the estimate down.

Deterministic weighting is less risky.

Deterministic weights

Suppose true variance improves: $\mathrm{Var}(\hat{\mu}_k) \propto k^{-r_0}$ for some $0 \leqslant r_0 \leqslant 1$.

We should weight proportionally to k^{r_0} but we might not know r_0. So we use r_1.

If we take $r_1 = 1/2$ then we never raise variance by more than 12.5%.

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

Upshot

Only a small effiiciency loss from weighting stage k proportionally to $\sqrt{k}$.

$$\sum_{k=1}^{K} \sqrt{k} \hat{\mu}_k \Big/ \sum_{k=1}^{K} \sqrt{k}.$$

O & Zhou (1999)

Not efficient for exponential convergence (see below) but that is a rare circumstance.

Combination

Put weights ω_k on $\hat{\mu}_k$, such as $\omega_k \propto \sqrt{k}$.

Final estimate $\hat{\mu} = \sum_{k=1}^{K} \omega_k \mu_k$.

Assume

$$\mathbb{E}(\hat{\mu}_k \mid \text{data to steps } k-1) = \mu$$

$$\mathbb{E}(\widehat{\mathrm{Var}}(\hat{\mu}_k) \mid \text{data to steps } k-1) = \mathrm{Var}(\hat{\mu}_k \mid \text{data to steps } k-1)$$

Then by Martingale arguments

$$\mathbb{E}(\hat{\mu}) = \mu, \quad \text{and}$$

$$\widehat{\mathrm{Var}}(\hat{\mu}) \equiv \sum_k \omega_k^2 \widehat{\mathrm{Var}}(\hat{\mu}_k), \quad \text{satisfies}$$

$$\mathbb{E}(\widehat{\mathrm{Var}}(\hat{\mu})) = \mathrm{Var}(\hat{\mu})$$

Exponential convergence

Parametric set $\mathcal{Q} = \{q_\theta \mid \theta \in \Theta\}, \quad \Theta \subset \mathbb{R}^d$

Best case scenario

If $q_* = fp/\mu \in \mathcal{Q}$ then we can approach $\sigma_q^2 = 0$.

Let $q_* = q_{\theta_*}$. If we can estimate θ_* to within $O(n^{-1/2})$ from a first estimate we get much smaller variance at the second step.

The result can be a virtuous cycle of $\theta^{(k)} \to \theta_*$ and $\sigma_k^2 \to 0$.

Examples

1) Toy example with $p = \mathcal{N}(0, 1)$ and $f(x) = \exp(Ax)$.

2) Theory from Kollman (1993) and others for particle transport.

Toy example

$p = \mathcal{N}(0, 1)$, $f(x) = \exp(Ax)$ for $A > 0$ and $q_\theta = \mathcal{N}(\theta, 1)$.

We actually know $\mu = \mathbb{E}(f(\boldsymbol{x})) = \exp(A^2/2)$ so we don't really need MC.

Suppose we also know $\theta_* = A$ and $\mathbb{E}_p(f(\boldsymbol{x})) = \exp(A^2/2)$.
That is $\theta_* = \sqrt{2\log(\mathbb{E}(f(\boldsymbol{x})))}$.

Cycle through

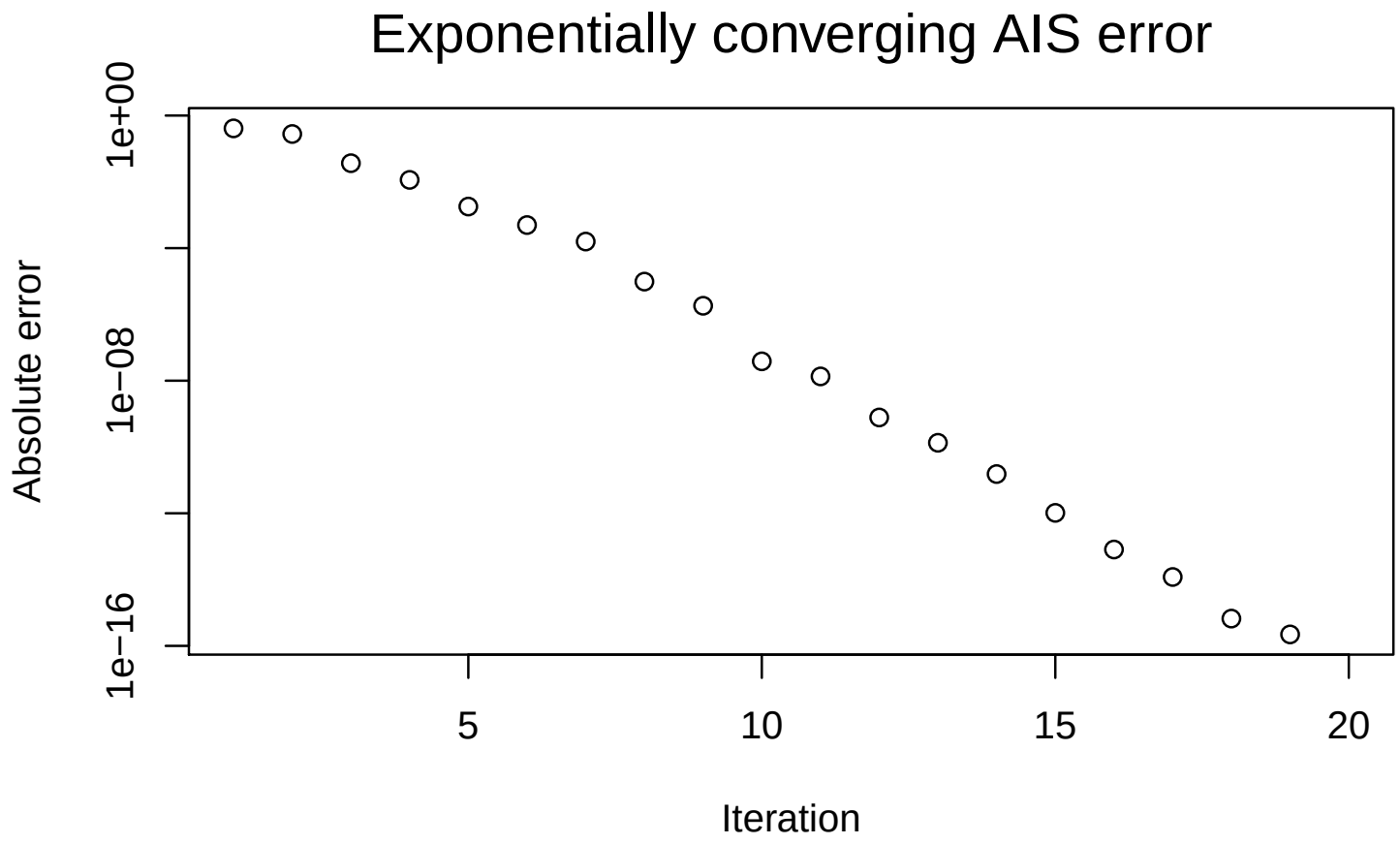
$$\boldsymbol{x}_i \sim q_{\theta^{(k)}}, \quad i = 1, \ldots, n$$

$$\hat{\mu}^{(k)} \leftarrow \frac{1}{n}\sum_{i=1}^{n} \frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{q_{\theta_k}(\boldsymbol{x}_i)}$$

$$\hat{\theta}^{(k+1)} \leftarrow \sqrt{2\log(\max(1, \hat{\mu}^{(k)}))}.$$

Without $\max(1, \cdot)$ the algorithm can fail.

Exponential convergence



Steps with $n = 20$.

Transport problems

Booth (1985) observed exponential convergence in some particle transport problems.

Kollman (1993) established exponential convergence under strong conditions:

start one simulation from each state of a Markov chain.

Extensions from Kollman, Baggerly, Cox, Picard (1999) and Baggerly, Cox, Picard (2000) and Kong, Ambrose & Spanier (2009)

Story

Markovian particles $\mathbb{P}(\boldsymbol{x}_{n+1} = \boldsymbol{x} \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n-1}) = \mathbb{P}(\boldsymbol{x}_{n+1} = \boldsymbol{x} \mid \boldsymbol{x}_{n-1})$

Terminal state Δ (absorption or left region of interest).

$\lim_{n\to\infty} \mathbb{P}(\boldsymbol{x}_n = \Delta) = 1.$

$$f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots) = \sum_{n=1}^{\infty} s(\boldsymbol{x}_{n-1} \to \boldsymbol{x}_n) = \sum_{n=1}^{\tau} s(\boldsymbol{x}_{n-1} \to \boldsymbol{x}_n)$$

hitting time τ, score function $s(\cdot \to \cdot)$ with $s(\Delta \to \Delta) = 0$.

Particle transport

The particle follows a Markov chain with transitions P.

We use transitions Q instead, with $Q_{ij} > 0$ whenever $P_{ij} > 0$.

Optimal Q

$$Q_{ij} = \frac{P_{ij}(s_{ij} + \mu_j)}{P_{i\Delta}s_{i\Delta} + \sum_{\ell=1}^{d} P_{i\ell}(s_{i\ell} + \mu_\ell)}$$

$$Q_{i\Delta} = \frac{P_{i\Delta}s_{i\Delta}}{P_{i\Delta}s_{i\Delta} + \sum_{\ell=1}^{d} P_{i\ell}(s_{i\ell} + \mu_\ell)}, \quad \text{where}$$

$$\mu_i = \mathbb{E}_Q\left(\sum_{n=1}^{\tau} s(\boldsymbol{x}_{n-1} \to \boldsymbol{x}_n) w_n(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \;\middle|\; \boldsymbol{x}_0 = i\right), \quad \text{and}$$

$$w_n = \prod_{j=1}^{n} \frac{P_{x_{j-1}x_j}}{Q_{x_{j-1}x_j}}.$$

Algorithm

Alternates between estimating μ and sampling from $Q(\mu)$.

Cross-entropy

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 $\theta \in \Theta$.

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

We would like

$$\theta = \arg\min_{\theta\in\Theta} \mathrm{MS}(\theta) \qquad \text{where} \qquad \mathrm{MS}(\theta) = \int \frac{(fp)^2}{q_\theta}\,\mathrm{d}\boldsymbol{x}$$

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 entropy.

Cross-entropy

Use an exponential family

$$q_\theta = \exp(\theta^\mathsf{T} \boldsymbol{x} - A(\boldsymbol{x}) - C(\theta))$$

Replace variance by KL distance

$$\mathcal{D}(q_* \| q_\theta) = \mathbb{E}_{q_*}\left(\log\left(\frac{q_*(\boldsymbol{x})}{q_\theta(\boldsymbol{x})}\right)\right)$$

This distance has $\mathcal{D}(q_* \| q_\theta) \geqslant 0$ and $\mathcal{D}(q_* \| q_*) = 0$.

Seek θ to minimize

$$\mathcal{D}(q_* \| q_\theta) = \mathbb{E}_{q^*}(\log(q_*(\boldsymbol{x})) - \log(q(\boldsymbol{x}; \theta)))$$

I.e., maximize

$$\mathbb{E}_{q^*}(\log(q(\boldsymbol{x}; \theta)))$$

Cross-entropy

For $q > 0$ whenever $q_* > 0$,

$$\mathbb{E}_{q_*}(\log(q(\boldsymbol{x};\theta))) = \mathbb{E}_q\left(\frac{\log(q(\boldsymbol{x};\theta))q_*(\boldsymbol{x})}{q(\boldsymbol{x})}\right) = \frac{1}{\mu}\mathbb{E}_q\left(\frac{\log(q(\boldsymbol{x};\theta))f(\boldsymbol{x})p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)$$

Choose $\theta^{(k+1)}$ to maximize

$$\begin{aligned} G^{(k)}(\theta) &= \frac{1}{n_k}\sum_{i=1}^{n_k}\frac{p(\boldsymbol{x}_i)f(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i;\theta^{(k)})}\log(q(\boldsymbol{x};\theta)) \\ &\equiv \frac{1}{n_k}\sum_{i=1}^{k} H_i \log(q(\boldsymbol{x}_i;\theta)) \\ &\equiv \frac{1}{n_k}\sum_{i=1}^{k} H_i\left(\theta^{\mathsf{T}}\boldsymbol{x}_i - A(\boldsymbol{x}_i) - C(\theta)\right) \end{aligned}$$

$C(\theta)$ is a known convex function. For $H_i \geqslant 0$. Assume (for now) that $\sum_i H_i > 0$.

The update often takes a simple moment matching form.

Cross-entropy update

$$\frac{\sum_i H_i \boldsymbol{x}_i^\mathsf{T}}{\sum_i H_i} = \frac{\partial}{\partial \theta} C(\theta^{(\theta+1)})$$

For $q_\theta = \mathcal{N}(\theta, I)$

$$\theta^{(k+1)} \leftarrow \frac{\sum_i H_i \boldsymbol{x}_i}{\sum_i H_i}$$

For $q_\theta = \mathcal{N}(\theta, \Sigma)$

$$\theta^{(k+1)} \leftarrow \Sigma^{-1} \frac{\sum_i H_i \boldsymbol{x}_i}{\sum_i H_i}$$

Other exponential family updates are typically closed form functions of sample moments.

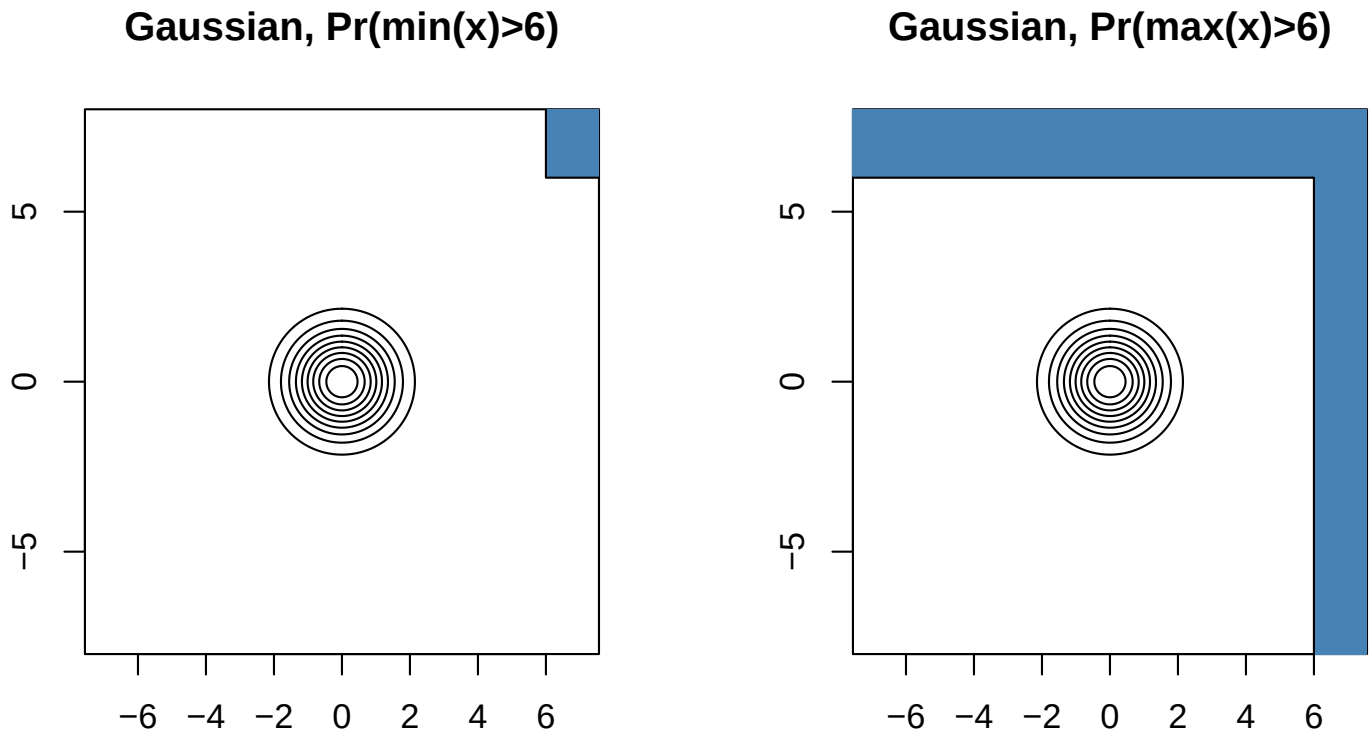
What if all $H_i = 0$?

Consider $f(\boldsymbol{x}) = 1_{g(\boldsymbol{x}) \geqslant \tau}$ for large τ.

Use some other $\tau^{(k)}$ instead. For example 99'th percentile of $g(\boldsymbol{x}_i)$.

Ideally $\tau^{(k)} \to \tau$.

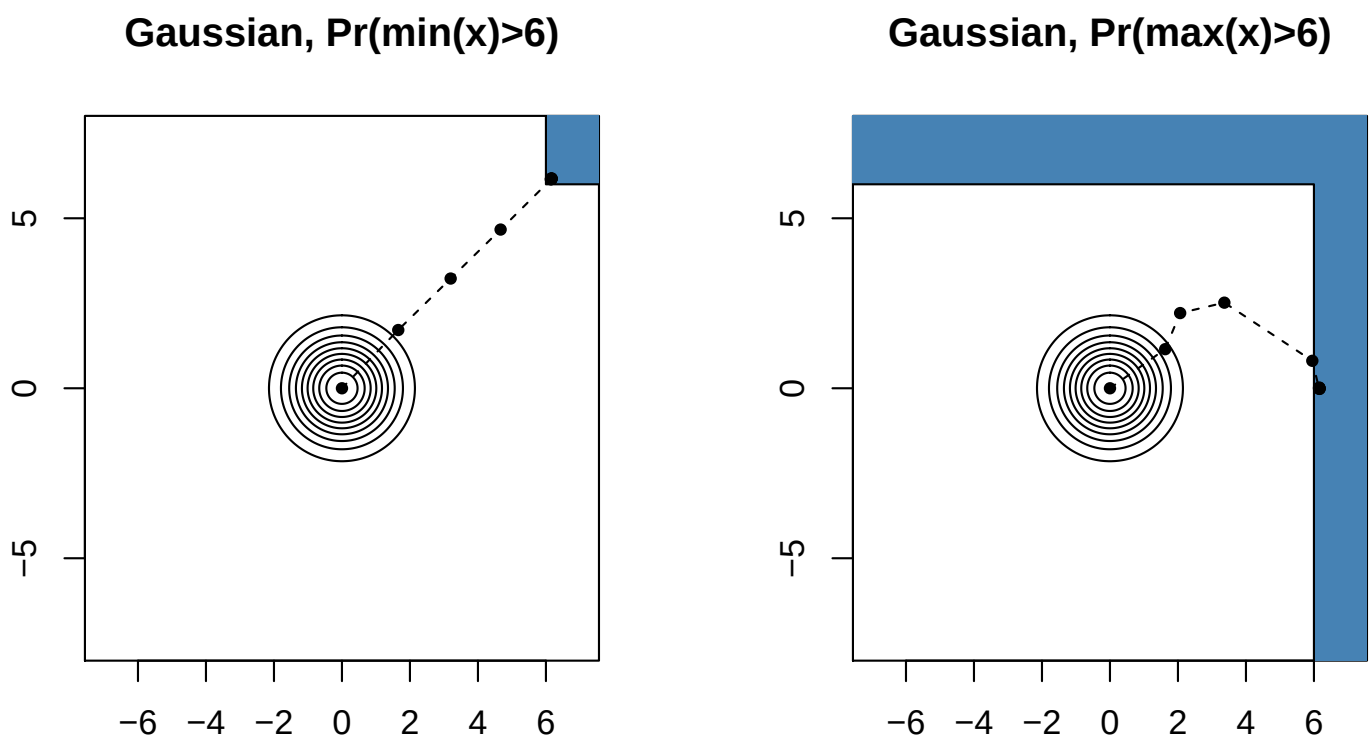
Cross-ent examples



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

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 $1/2$.

Nonparametric AIS

Flexible / infinite dimensional families $\mathcal{Q}$.

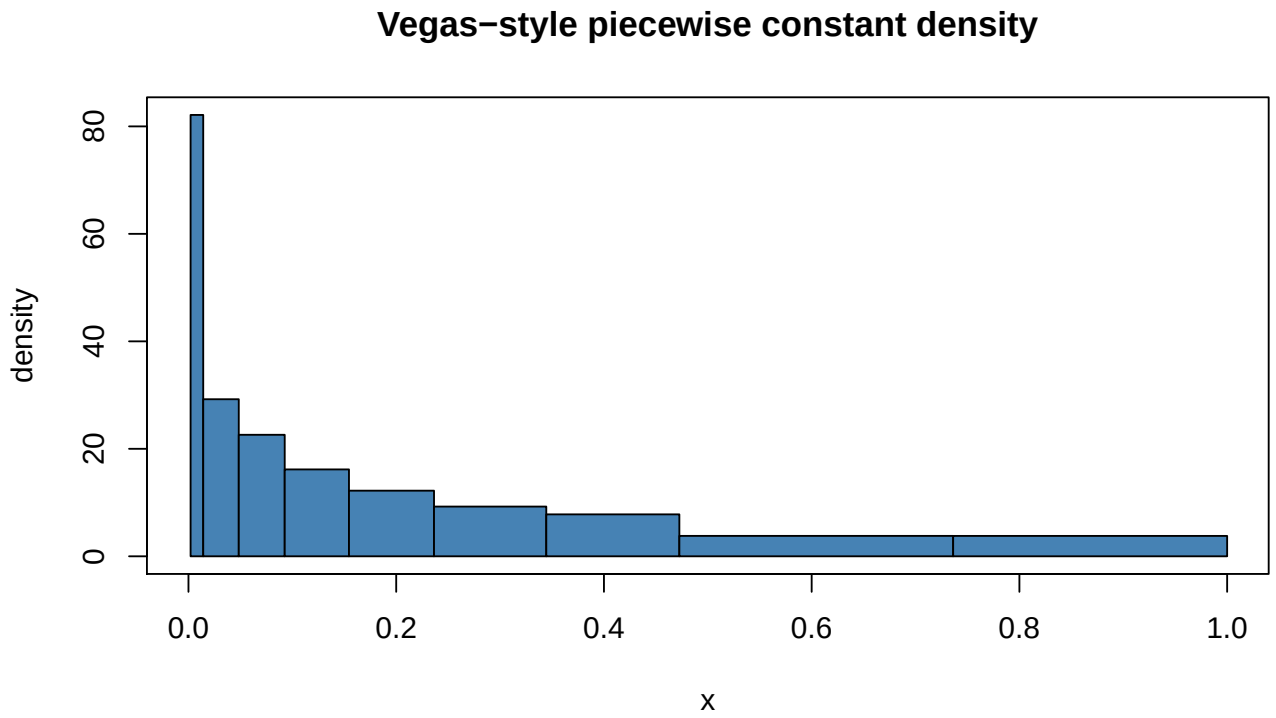
Often based on recursive rectangular partitioning of region like $[0, 1]^d$.

1) Vegas, Lepage (1978)
2) Divonne, Friedman & Wright (1979, 1981)
3) Miser Press & Farrar (1990)
4) Split Vegas Zhou (1998) $\approx$ Vegas within Miser

Vegas

$$q(\boldsymbol{x}) = \prod_{j=1}^{d} q_j(x_j)$$

Each q_j is piece-wise constant. Equal content, not equal width.



Very amenable to sampling spikes.

Comments on Vegas

The optimal q_j given other q_ℓ is

$$q_j(x_j) \propto \sqrt{\int_{(0,1)^{d-1}} \frac{f(\boldsymbol{x})^2}{\prod_{\ell\neq j} q_\ell(x_\ell)} \prod_{\ell\neq j} \mathrm{d}x_\ell}.$$

Update starts by iterating above. Some complex steps.

A problem with Vegas

Suppose $f(\boldsymbol{x})$ has two modes in $[0,1]^{10}$ one near $(0,0,\ldots,0)$ and one near $(1,1,\ldots,1)$

Each q_j needs two modes.

So q has 2^{10} modes, mostly irrelevant.

Divonne

Friedman & Wright (1979, 1981)

$p = \mathbf{U}[0,1]^d$ and f has two continuous derivatives.

Local optimization has less curse of dimension than integration.

Newton steps cost $O(d^3)$ per iteration.

Recursive partitioning

Hyper rectangles $R_\ell \subset [0,1]^d$ with $\cup_{\ell=1}^{L} R_\ell = [0,1]^d$ and $\mathbf{vol}(R_j \cap R_k) = 0$ for $j \neq k$.

$$\int_{[0,1]^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \sum_{\ell=1}^{L} \int_{R_\ell} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

Start with $L = 1$ and $R_1 = [0,1]^d$.

Divonne splits

Find min and max of f in each rectangle.

$$\overline{f}(R) = \max_{\boldsymbol{x} \in R} f(\boldsymbol{x})$$
$$\underline{f}(R) = \min_{\boldsymbol{x} \in R} f(\boldsymbol{x})$$
$$b(R) = \mathbf{vol}(R)|\overline{f}(R) - \underline{f}(R)| \qquad \text{(badness)}$$

Split the rectangle with largest 'badness'.

Non-binary splits

To do the split decide which of max, min is farthest from average f in the rectangle.

Define a subrectangle to isolate that mode.

Partition the original rectangle into numerous other subrectangles to complement the one around the mode.

An issue it is not designed for unbounded integrands and they will give it trouble.

Miser

Press & Farrar (1990) see also Numerical Recipes

Estimates $\int_{(0,1)^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ by recursive binary partitioning into rectangles.

Each rectangle can be split down the middle in d different directions.

The chosen direction is based on an estimate of which j will lead to the most accurate estimate.

They assume that Miser will be applied left and right, not Monte Carlo, and them employ an assumption about the Miser convergence rate to decide.

Having chosen the split, they allocate sample sizes left and right.

For a mode at $(1/2, 1/2, \ldots, 1/2)$ then split at $1/2 + \mathbf{U}(-\delta/2, \delta/2)$ to avoid it.

Split Vegas

Dissertation of Zhou (1998).

Within a rectangular region, fit Vegas.

If one of the marginal distributions is strongly bimodal then split the rectangle on that direction.

It can be subtle to indentify whether a given margin has two or more modes.

Mixture of products of beta

For $\int_{[0,1]^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$

Beta density $b(x;\alpha,\beta) \propto x^{\alpha-1}(1-x)^{\beta-1}$ $\alpha,\beta>0$

Can have a bump near $\alpha/(\alpha+\beta)$ or a spike near 0 and/or near 1.

d dimensional product of beta densities $\prod_{j=1}^d b(x_j;\alpha_j,\beta_j)$

Mixture of products of beta densities $q(\boldsymbol{x}) = \sum_{m=1}^M \gamma_m \prod_{j=1}^d b(x_j;\alpha_{jm},\beta_{jm})$

Adaptive part

Choose γ_m and α_{jm}, β_{jm} by Levenberg-Marquardt nonlinear least squares to minimize

$$\sum_{i=1}^n \Big(|f(\boldsymbol{x}_i)| - q(\boldsymbol{x}_i;\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})\Big)^2$$

O & Zhou (1999)

Many more details in thesis Zhou (1998).

Particle physics integrand

From Kinoshita (1970).

7 dimensional, positive and negative integrable singularities.

	MC	Vegas	MISER	ADBayes	MixBeta
Err/μ	0.154	0.092	0.073	0.210	0.035
$\widehat{\mathrm{Err}}/\mathrm{Err}$	1.163	34.0	0.750	19.5	0.394

$K = 20$ runs of $n = 10^5$ points each.

Except ADBAYES.

Known that $\mu \doteq 0.371005292$ as of 1991. (!!)

(Don't recall non-sampling costs.)

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

The weights on observations from one round can change later due to data from other 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.

Optimal mixture weights

$$\mathrm{MS}_{\alpha,\beta} \equiv \int \frac{(fp + \sum_j \beta_j (q_j - p))^2}{\sum_j \alpha_j q_j} \, \mathrm{d}\boldsymbol{x}$$

is **jointly** convex in α (in simplex) and β.

So is a sample estimate of it.

Therefore we can estimate optimal mixture weights.

Also we can constrain $\alpha_j \geqslant \epsilon > 0$ in a convex optimiation and variance does not increase much.

He & Owen (2014) sequentially optimize both α and β.

M-PMC

Mixture-Population Monte Carlo

Cappé, Douc, Guillin, Marin & Robert (2008)

$$q(\boldsymbol{x}) = \sum_{j=1}^{J} \alpha_j q_j(\boldsymbol{x}; \theta_j)$$

They target Kullback-Leibler distance to target dist'n.

Also use SNIS.

Aim to optimize over both α_j and θ_j (non-convex).

Use an EM algorithm for the θ updates.

Earlier Douc, Guillin, Marin & RObert (2007) target specific integrand.

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.

Stochastic convex programming

Ryu & Boyd (2015)

They take up the exponential family setting of cross-entropy.

They find in exponential families that the variance is a convex function.

So they don't have to use Kullback-Leibler.

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i;\theta_i)}$$

Potentially a new θ_i for each $\boldsymbol{x}_i$. (or small batches)

Stochastic gradient descent

$$\theta_{n+1} = \mathrm{Proj}(\theta_n - \alpha_n g_n)$$

Sequence $\alpha_n \to 0$ with $\sum \alpha_n = \infty$

Proj projects onto Θ

$g_n = g_n(\theta_n, \boldsymbol{x}_n)$ is n'th gradient estimate

$\hat{\mu}_n$ attains optimal variance$\times(1 + O(n^{-1/2}))$.

Kriging based estimators

Picheny (2009), Baleadent, Morio & Marzat (2013), Dalbey & Swiler (2014)

Rare event $\phi(\boldsymbol{x}) > \tau$. We get $\phi(\boldsymbol{x}_i)$, not just $f(\boldsymbol{x}) = 1_{\phi(\boldsymbol{x})>\tau}$.

Kriging

Given $\phi(\boldsymbol{x}_1), \ldots \phi(\boldsymbol{x}_n)$ it can predict $\phi(\boldsymbol{x})$ elsewhere by Gaussian process interpolation.

Swiler & West (2010) report that $\int p(\boldsymbol{x}) 1_{\hat{\phi}(\boldsymbol{x})>\tau} \, \mathrm{d}\boldsymbol{x}$ does not work well.

Better to use $\int p(\boldsymbol{x}) \Phi((\tau - \hat{\phi}(\boldsymbol{x})/s(\boldsymbol{x}))) \, \mathrm{d}\boldsymbol{x}$ where kriging gives $\hat{\phi}$ and a posterior standard deviation s. Chevalier, Ginsbourger, Picheny, Richet, Bect, Vazquez (2014)

Dalbey & Swiler (2014) use $\hat{\phi}$ to update q.

Thanks

- Yi Zhou, Hera He, co-authors
- Minyong Lee, discussion
- Toichuro Kinoshita (particle physics code and integral)
- U.S. National Science Foundation DMS-1521145, DMS-1407397
- Michael Chertkov
- Kacy Hopwood