Statistical Models & Computing Methods

Lecture 3: Numerical Integration

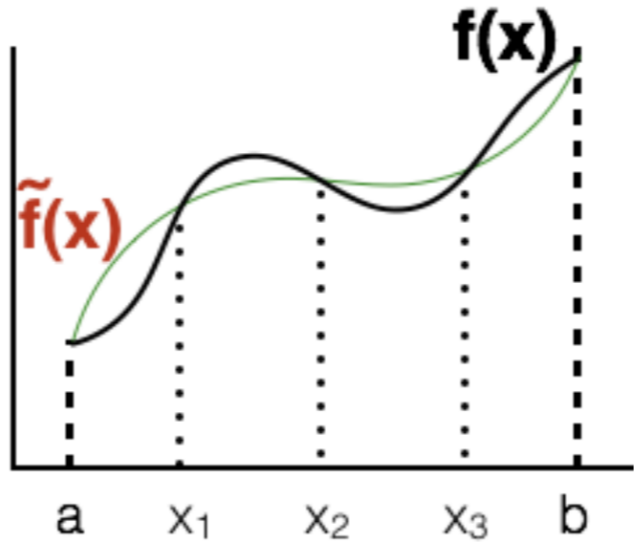
Cheng Zhang

School of Mathematical Sciences, Peking University

October 22, 2020

- Statistical inference often depends on intractable integrals $I(f) = \int_\Omega f(x)dx$
- This is especially true in Bayesian statistics, where a posterior distribution is usually non-trivial.
- In some situations, the likelihood itself may depend on intractable integrals so frequentist methods would also require numerical integration
- In this lecture, we start by discussing some simple numerical methods that can be easily used in low dimensional problems
- Next, we will discuss several Monte Carlo strategies that could be implemented even when the dimension is high

- Consider a one-dimensional integral of the form $I(f) = \int_a^b f(x)dx$
- A common strategy for approximating this integral is to use a tractable approximating function $\tilde{f}(x)$ that can be integrated easily
- We typically constrain the approximating function to agree with f on a grid of points: $x_1, x_2, \ldots, x_n$



- Newton-Côtes methods use equally-spaced grids
- The approximating function is a polynomial
- The integral then is approximated with a weighted sum as follows

$$\hat{I} = \sum_{i=1}^{n} w_i f(x_i)$$

- In its simplest case, we can use the Riemann rule by partitioning the interval $[a, b]$ into n subintervals of length $h = \frac{b-a}{n}$; then

$$\hat{I}_L = h \sum_{i=0}^{n-1} f(a + ih)$$

This is obtained using a piecewise constant function $\tilde{f}$ that matches f at the left points of each subinterval

- Alternatively, the approximating function could agree with the integrand at the right or middle point of each subinterval

$$\hat{I}_R = h\sum_{i=1}^{n} f(a+ih), \quad \hat{I}_M = h\sum_{i=0}^{n-1} f(a+(i+\frac{1}{2})h)$$

- In either case, the approximating function is a zero-order polynomial
- To improve the approximation, we can use the trapzoidal rule by using a piecewise linear function that agrees with $f(x)$ at both ends of subintervals

$$\hat{I} = \frac{h}{2}f(a) + h\sum_{i=1}^{n-1} f(x_i) + \frac{h}{2}f(b)$$

- We would further improve the approximation by using higher order polynomials
- Simpson's rule uses a quadratic approximation over each subinterval

$$\int_{x_i}^{x_{i+1}} f(x)dx \approx \frac{x_{i+1} - x_i}{6} \left(f(x_i) + 4f(\frac{x_i + x_{i+1}}{2}) + f(x_{i+1}) \right)$$

- In general, we can use any polynomial of degree k

- Newton-Côtes rules require equally spaced grids
- With a suitably flexible choice of $n+1$ nodes, $x_0, x_1, \ldots, x_n$, and corresponding weights, $A_0, A_1, \ldots, A_n$,

$$\sum_{i=0}^{n} A_i f(x_i)$$

  gives the exact integration for all polynomials with degree less than or equal to $2n+1$
- This is called Gaussian quadrature, which is especially useful for the following type of integrals $\int_a^b f(x)w(x)dx$ where $w(x)$ is a nonnegative function and $\int_a^b x^k w(x)dx < \infty$ for all $k \geq 0$

- In general, for squared integrable functions,

$$\int_a^b f(x)^2 w(x) dx \leq \infty$$

denoted as $f \in \mathcal{L}^2_{w,[a,b]}$, we define the inner product as

$$\langle f, g \rangle_{w,[a,b]} = \int_a^b f(x) g(x) w(x) dx$$

where $f, g \in \mathcal{L}^2_{w,[a,b]}$

- We said two functions to be *orthogonal* if $\langle f, g \rangle_{w,[a,b]} = 0$. If f and g are also scaled so that $\langle f, f \rangle_{w,[a,b]} = 1$, $\langle g, g \rangle_{w,[a,b]} = 1$, then f and g are orthonormal

- We can define a sequence of orthogonal polynomials by a recursive rule

$$T_{k+1}(x) = (\alpha_{k+1} + \beta_{k+1}x)T_k(x) - \gamma_{k+1}T_{k-1}(x)$$

- Example: Chebyshev polynomials (first kind).

$$T_0(x) = 1, \quad T_1(x) = x$$
$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

- $T_n(x)$ are orthogonal with respect to $w(x) = \frac{1}{\sqrt{1-x^2}}$ and $[-1, 1]$

$$\int_{-1}^{1} T_n(x)T_m(x)\frac{1}{\sqrt{1-x^2}}dx = 0, \quad \forall n \neq m$$

- In general orthogonal polynomials are note unique since $\langle f, g\rangle = 0$ implies $\langle cf, dg\rangle = 0$
- To make the orthogonal polynomial unique, we can use the following standarizations
  - make the polynomial orthonormal: $\langle f, f\rangle = 1$
  - set the leading coefficient of $T_j(x)$ to 1
- Orthogonal polynomials form a basis for $\mathcal{L}^2_{w,[a,b]}$ so any function in this space can be written as

$$f(x) = \sum_{n=0}^{\infty} a_n T_n(x)$$

where $a_n = \frac{\langle f, T_n\rangle}{\langle T_n, T_n\rangle}$

- Let $\{T_n(x)\}_{n=0}^{\infty}$ be a sequence of orthogonal polynomials with respect to w on $[a, b]$.
- Denote the $n+1$ roots of $T_{n+1}(x)$ by

$$a < x_0 < x_1 < \ldots < x_n < b.$$

- We can find weights $A_1, A_2, \ldots, A_{n+1}$ such that

$$\int_a^b P(x)w(x)dx = \sum_{i=0}^{n} A_i P(x_i), \quad \forall \deg(P) \leq 2n+1$$

- To do that, we first show: there exists weights $A_1, A_2, \ldots, A_{n+1}$ such that

$$\int_a^b P(x)w(x)dx = \sum_{i=0}^{n} A_i P(x_i), \quad \forall \deg(P) < n+1$$

- Sketch of proof. We only need to satisfy

$$\int_a^b x^k w(x)dx = \sum_{i=0}^{n} A_i x_i^k, \quad \forall\, k = 0, 1, \ldots, n$$

This leads to a system of linear equations

$$\begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_0 & x_1 & \ldots & x_n \\ \vdots & \vdots & \vdots & \vdots \\ x_0^n & x_1^n & \ldots & x_n^n \end{bmatrix} \begin{bmatrix} A_0 \\ A_1 \\ \vdots \\ A_n \end{bmatrix} = \begin{bmatrix} I_0 \\ I_1 \\ \vdots \\ I_n \end{bmatrix}$$

where $I_k = \int_a^b x^k w(x)dx$. The determinant of the coefficient matrix is a Vandermonde determinant, and is non-zero since $x_i \neq x_j, \forall i \neq j$

- Now we show that the above Gaussian Quadrature can be exact for polynomials of degree $\leq 2n+1$
- Let $P(x)$ be a polynomial with $\deg(P) \leq 2n+1$, there exist polynomials $g(x)$ and $r(x)$ such that

$$P(x) = g(x)T_{n+1}(x) + r(x)$$

with $\deg(g) \leq n, \deg(r) \leq n$, Therefore,

$$\begin{aligned}\int_a^b P(x)w(x)dx = \int_a^b r(x)w(x)dx &= \sum_{i=0}^{n} A_i r(x_i) \\ &= \sum_{i=0}^{n} A_i P(x_i)\end{aligned}$$

- We now discuss the Monte Carlo method mainly in the context of statistical inference
- As before, suppose we are interested in estimating $I(h) = \int_a^b h(x)dx$
- If we can draw iid samples, $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ uniformly from (a, b), we can approximate the integral as

$$\hat{I}_n = (b-a)\frac{1}{n}\sum_{i=1}^{n} h(x^{(i)})$$

- Note that we can think about the integral as

$$(b-a)\int_a^b h(x) \cdot \frac{1}{b-a} dx$$

where $\frac{1}{b-a}$ is the density of Uniform(a, b)

- In general, we are interested in integrals of the form $\int_{\mathcal{X}} h(x)f(x)dx$, where $f(x)$ is a probability density function
- Analogous to the above argument, we can approximate this integral (or expectation) by drawing iid samples $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ from the density $f(x)$ and then

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} h(x^{(i)})$$

- Based on the law of large numbers, we know that

$$\lim_{n\to\infty} \hat{I}_n \xrightarrow{p} I$$

- And based on the central limit theorem

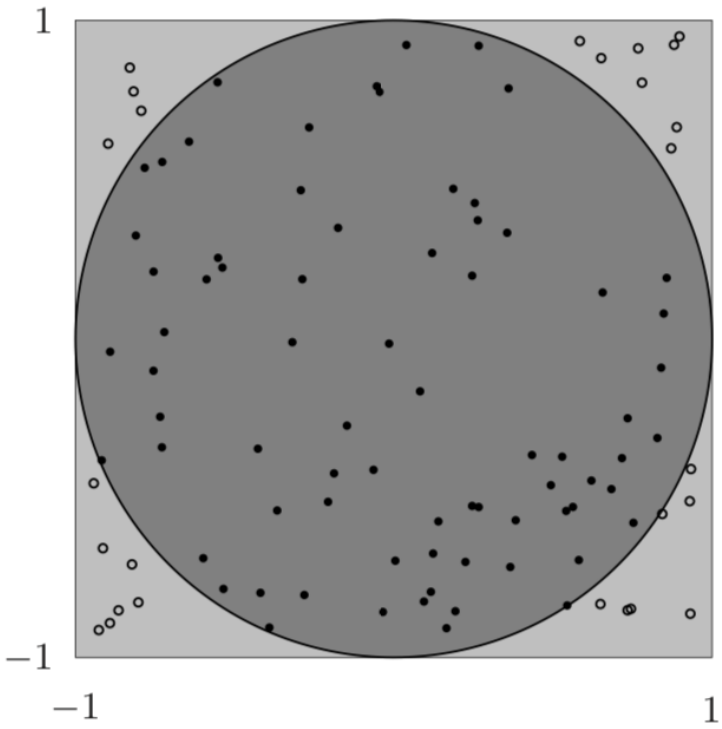
$$\sqrt{n}(\hat{I}_n - I) \to \mathcal{N}(0, \sigma^2), \quad \sigma^2 = \mathbb{V}\mathrm{ar}(h(X))$$

Example: estimating π

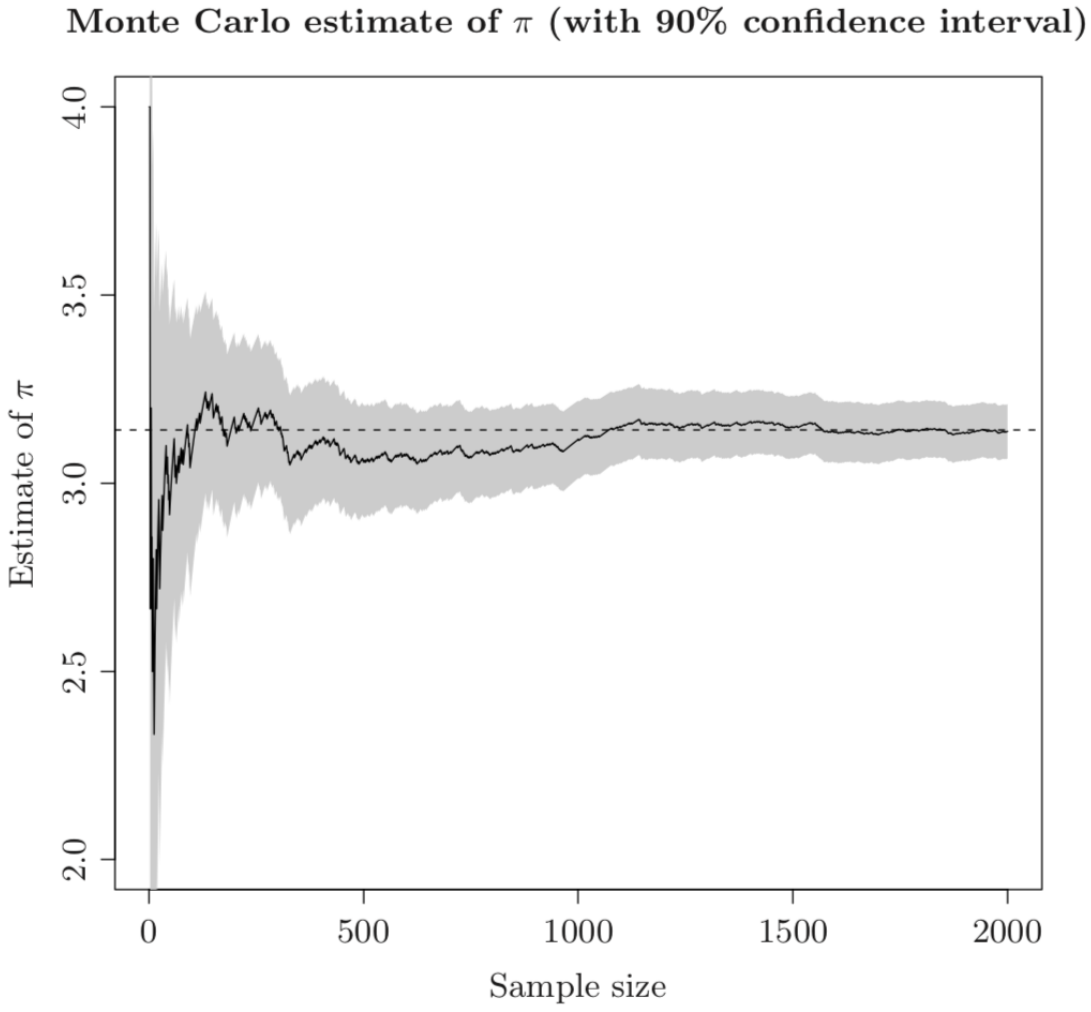
- Let $h(x) = \mathbf{1}_{B(0,1)}(x)$, then $\pi = 4 \int_{[-1,1]^2} h(x) \cdot \frac{1}{4}\, dx$
- Monte Carlo estimate of π

$$\hat{I}_n = \frac{4}{n} \sum_{i=1}^{n} \mathbf{1}_{B(0,1)}(x^{(i)})$$

$$x^{(i)} \sim \text{Uniform}([-1,1]^2)$$



Example: estimating π



- Convergence rate for Monte Carlo: $\mathcal{O}(n^{-1/2})$

$$p\left(|\hat{I}_n - I| \leq \frac{\sigma}{\sqrt{n\delta}}\right) \geq 1 - \delta, \quad \forall \delta$$

  often slower than quadrature methods ($\mathcal{O}(n^{-2})$ or better)
- However, the convergence rate of Monte Carlo does not depend on dimensionality
- On the other hand, quadrature methods are difficult to extend to multidimensional problems, because of the curse of dimensionality. The actual convergence rate becomes $\mathcal{O}(n^{-k/d})$, for any order k method in dimension d
- This makes Monte Carlo strategy very attractive for high dimensional problems

- Monte Carlo methods require sampling a set of points chosen randomly from a probability distribution
- For simple distribution $f(x)$ whose inverse cumulative distribution functions (CDF) exists, we can sampling x from f as follows

$$x = F^{-1}(u), \quad u \sim \text{Uniform}(0, 1)$$

where F^{-1} is the inverse CDF of f

- Proof.

$$p(a \leq x \leq b) = p(F(a) \leq u \leq F(b)) = F(b) - F(a)$$

- Exponential distribution: $f(x) = \theta \exp(-\theta x)$. The CDF is

$$F(a) = \int_0^a \theta \exp(-\theta x) = 1 - \exp(-\theta a)$$

therefore, $x = F^{-1}(u) = -\frac{1}{\theta}\log(1-u) \sim f(x)$. Since $1-u$ also follows the uniform distribution, we often use $x = -\frac{1}{\theta}\log(u)$ instead

- Normal distribution: $f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$. Box-Muller Transform

$$X = \sqrt{-2\log U_1}\cos 2\pi U_2$$
$$Y = \sqrt{-2\log U_1}\sin 2\pi U_2$$

where $U_1 \sim \text{Uniform}(0,1), \quad U_2 \sim \text{Uniform}(0,1)$

Intuition for Box-Muller Transform

- Assume $Z = (X, Y)$ follows the standard bivariate normal distribution. Consider the following transform

$$X = R\cos\Theta, \quad Y = R\sin\Theta$$

- From symmetry, clearly Θ follows the uniform distribution on the interval $(0, 2\pi)$ and is independent of R
- What distribution does R follow? Let's take a look at its CDF

$$\begin{aligned} p(R \le r) &= p(X^2 + Y^2 \le r^2) \\ &= \frac{1}{2\pi}\int_0^r t\exp(-\frac{t^2}{2})dt \int_0^{2\pi} d\theta = 1 - \exp(-\frac{r^2}{2}) \end{aligned}$$

Therefore, using the inverse CDF rule, $R = \sqrt{-2\log U_1}$

- If it is difficult or computationally intensive to sample directly from $f(x)$ (as described above), we need to use other strategies
- Although it is difficult to sample from $f(x)$, suppose that we can evaluate the density at any given point up to a constant $f(x) = f^*(x)/Z$, where Z could be unknown (remember that this make Bayesian inference convenient since we usually know the posterior distribution only up to a constant)
- Furthermore, assume that we can easily sample from another distribution with the density $g(x) = g^*(x)/Q$, where Q is also a constant

- Now we choose the constants c such that $cg^*(x)$ becomes the envelope (blanket) function for $f^*(x)$:

$$cg^*(x) \geq f^*(x), \quad \forall x$$

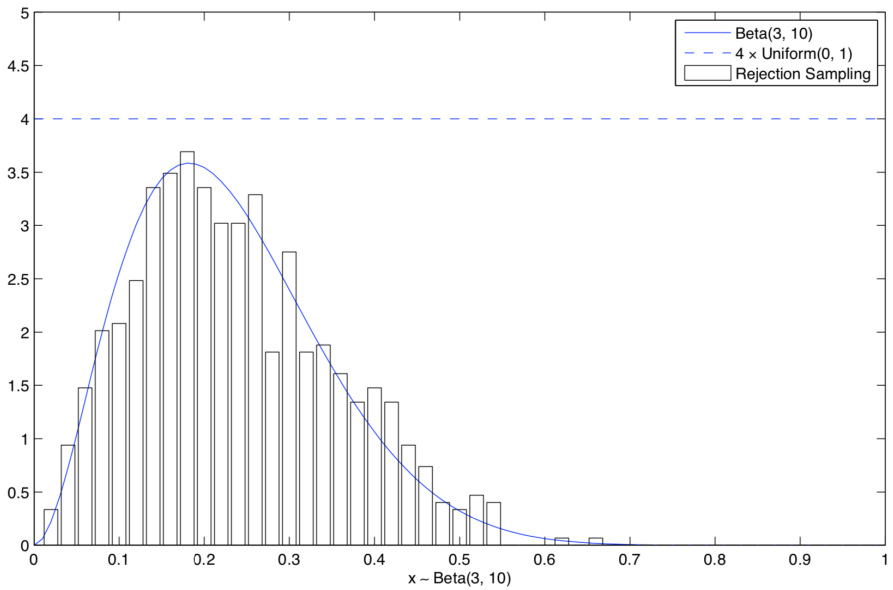
- Then, we can use a strategy known as *rejection sampling* in order to sample from $f(x)$ indirectly
- The rejection sampling method works as follows
  1. draw a sample x from $g(x)$
  2. generate $u \sim \text{Uniform}(0, 1)$
  3. if $u \leq \frac{f^*(x)}{cg^*(x)}$ we accept x as the new sample, otherwise, reject x (discard it)
  4. return to step 1

Rejection sampling generates samples from the target density, no approximation involved

$$
\begin{aligned}
p(X^R \le y) &= p(X^g \le y | U \le \frac{f^*(X^g)}{cg^*(X^g)}) \\
&= p(X^g \le y, U \le \frac{f^*(X^g)}{cg^*(X^g)}) / p(U \le \frac{f^*(X^g)}{cg^*(X^g)}) \\
&= \frac{\int_{-\infty}^{y} \int_0^{\frac{f^*(z)}{cg^*(z)}} dug(z)dz}{\int_{-\infty}^{\infty} \int_0^{\frac{f^*(z)}{cg^*(z)}} dug(z)dz} \\
&= \int_{-\infty}^{y} f(z)dz
\end{aligned}
$$

Example

- Assume that it is difficult to sample from the Beta(3, 10) distribution (this is not the case of course)
- We use the Uniform$(0, 1)$ distribution with $g(x) = 1,\ \forall x \in [0, 1]$, which has the envelop proporty: $4g(x) > f(x),\ \forall x \in [0, 1]$. The following graph shows the result after 3000 iterations



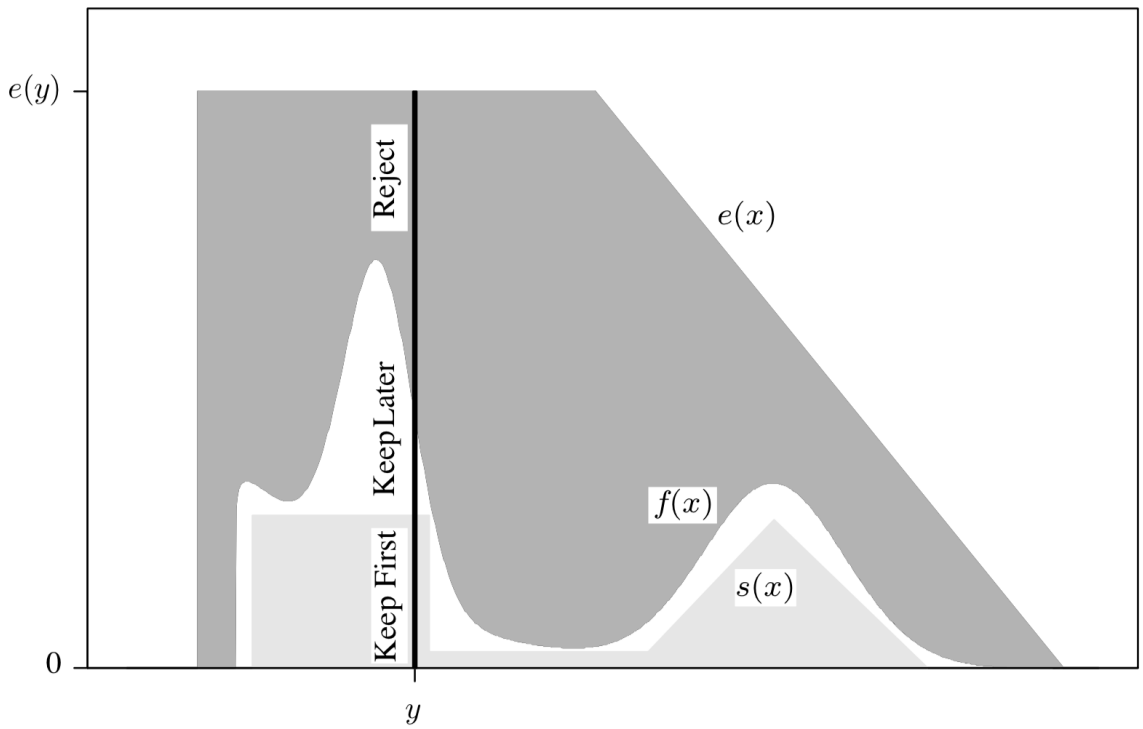
Rejection sampling becomes challenging as the dimension of x increases. A good rejection sampling algorithm must have three properties

- It should be easy to construct envelops that exceed the target everywhere
- The envelop distributions should be easy to sample
- It should have a low rejection rate

Squeezed Rejection Sampling

- When evaluating f^* is computationally expensive, we can improve the simulation speed of rejection sampling via *squeezed rejection sampling*
- Squeezed rejection sampling reduces the evaluation of f via a nonnegative squeezing function s that does not exceed f^* anywhere on the support of f: $s(x) \leq f^*(x), \forall x$
- The algorithm proceeds as follows:
  1. draw a sample x from $g(x)$
  2. generate $u \sim \text{Uniform}(0, 1)$
  3. if $u \leq \frac{s(x)}{cg^*(x)}$, we accept x as the new sample, return to step 1
  4. otherwise, determine whether $u \leq \frac{f^*(x)}{cg^*(x)}$. If this inequality holds, we accept x as the new sample, otherwise, we reject it.
  5. return to step 1

Squeezed Rejection Sampling



Remark: The proportion of iterations in which evaluation of f is avoided is $\int s(x)dx / \int e(x)dx$

- While Monte Carlo estimation is attractive for high dimension integration, it may suffer from lots of problems, such as rare events, and irregular integrands, etc.
- In what follows, we will discuss various methods to improve Monte Carlo approaches, with an emphasis on variance reduction techniques

- The simple Monte Carlo estimator of $\int_a^b h(x)f(x)dx$ is

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} h(x^{(i)})$$

  where $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ are randomly sampled from f
- A potential problem is the mismatch of the concentration of $h(x)f(x)$ and $f(x)$. More specifically, if there is a region A of relatively small probability under $f(x)$ that dominates the integral, we would not get enough data from the **important** region A by sampling from $f(x)$
- Main idea: Get more data from A, and then correct the bias

- **Importance sampling** (IS) uses importance distribution $q(x)$ to adapt to the true integrands $h(x)f(x)$, rather than the target distribution $f(x)$
- By correcting for this bias, importance sampling can greatly reduce the variance in Monte Carlo estimation
- Unlike the rejection sampling, we do not need the envelop property
- The only requirement is that $q(x) > 0$ whenever

$$h(x)f(x) \neq 0$$

- IS also applies when $f(x)$ is not a probability density function

- Now we can rewrite $I = \mathbb{E}_f(h(x)) = \int_{\mathcal{X}} h(x)f(x)\,dx$ as

$$\begin{aligned} I = \mathbb{E}_f(h(x)) &= \int_{\mathcal{X}} h(x)f(x)\,dx \\ &= \int_{\mathcal{X}} h(x)\frac{f(x)}{q(x)}q(x)dx \\ &= \int_{\mathcal{X}} (h(x)w(x))q(x) \\ &= \mathbb{E}_q(h(x)w(x)) \end{aligned}$$

where $w(x) = \dfrac{f(x)}{q(x)}$ is the importance weight function

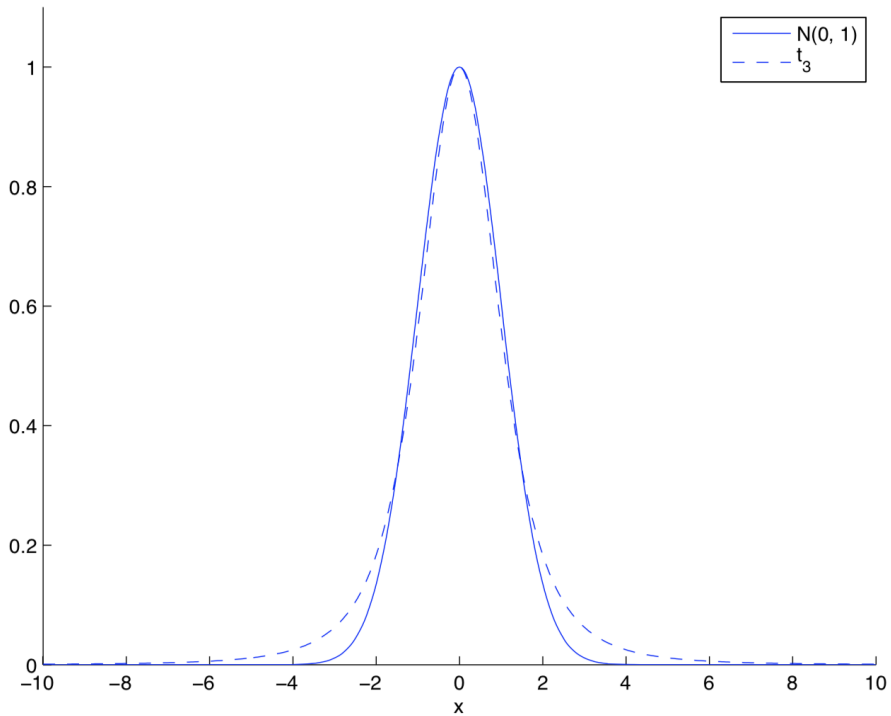
We can then approximate the original expectation as follows

- Draw samples $x^{(1)}, \ldots, x^{(n)}$ from $q(x)$
- Monte Carlo estimate

$$I_n^{\mathrm{IS}} = \frac{1}{n} \sum_{i=1}^{n} h(x^{(i)}) w(x^{(i)})$$

  where $w(x^{(i)}) = \frac{f(x^{(i)})}{q(x^{(i)})}$ are called importance ratios.
- Note that, now we only require sampling from q and do not require sampling from f

- We want to approximate a $\mathcal{N}(0,1)$ distribution with $t(3)$ distribution



- We generate 500 samples and estimated $I = \mathbb{E}(x^2)$ as 0.97, which is close to the true value 1.

Mean and Variance of IS

- Let $t(x) = h(x)w(x)$. Then $\mathbb{E}_q(t(X)) = I, X \sim q$

$$\mathbb{E}(I_n^{\text{IS}}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(t(x^{(i)}) = I$$

- Similarly, the variance is

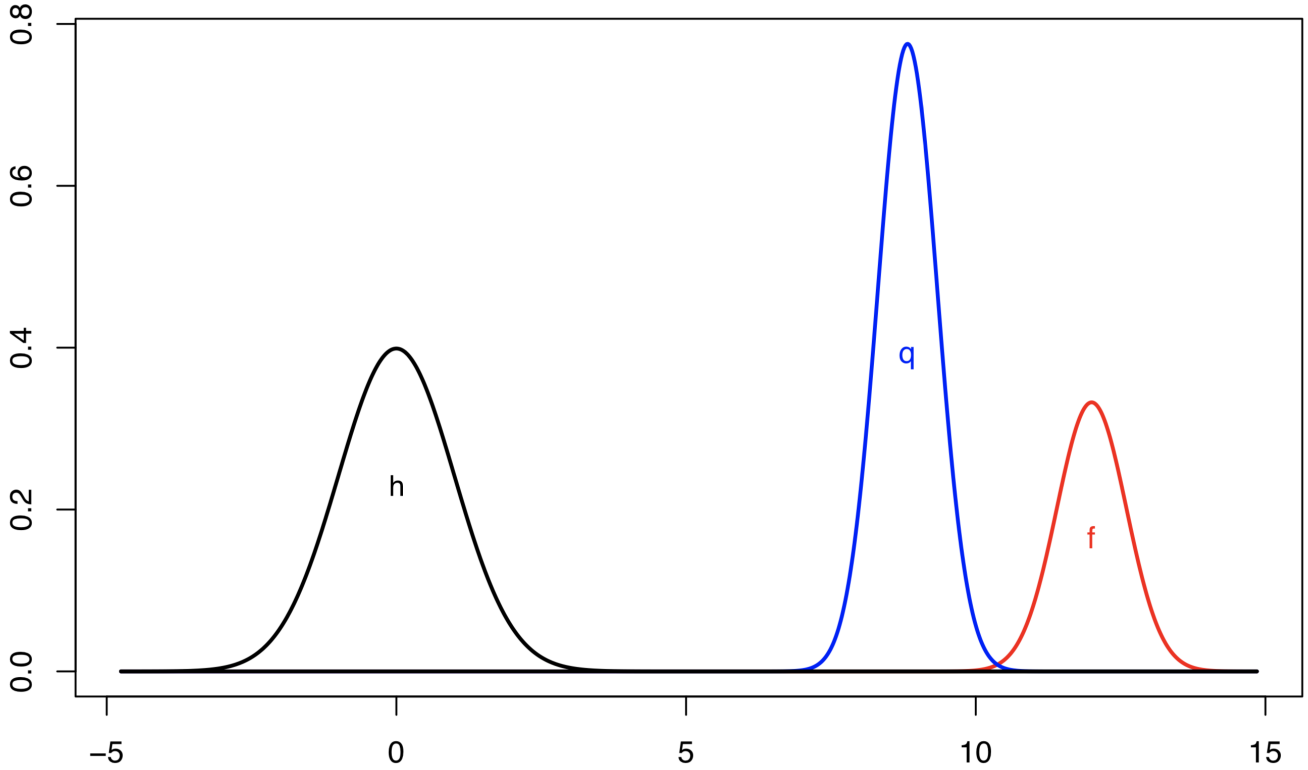
$$\begin{aligned}
\mathbb{V}\text{ar}_q(I_n^{\text{IS}}) &= \frac{1}{n}\mathbb{V}\text{ar}_q(t(X)) \\
&= \frac{1}{n}\int_{\mathcal{X}} \frac{(h(x)f(x))^2}{q(x)}\,dx - I^2 \qquad (1) \\
&= \frac{1}{n}\int_{\mathcal{X}} \frac{(h(x)f(x) - Iq(x))^2}{q(x)}\,dx \qquad (2)
\end{aligned}$$

- Recall the convergence rate for Monte Carlo is

$$p\left(|\hat{I}_n - I| \leq \frac{\sigma}{\sqrt{n\delta}}\right) \geq 1 - \delta, \quad \forall \delta$$

  For IS, $\sigma = \sqrt{\mathbb{V}\text{ar}_q(t(X))}$. A good importance distribution $q(x)$ would make $\mathbb{V}\text{ar}_q(t(X))$ small.

- What can we learn from equations (1) and (2)?
  - Optimal choice: $q(x) \propto h(x)f(x)$
  - $q(x)$ near 0 can be dangerous
  - Bounding $\frac{(h(x)f(x))^2}{q(x)}$ is useful theoretically



$\mathbb{V}\mathrm{ar}_q(t(X)) = 0$

Gaussian h and f $\Rightarrow$ Gaussian optimal q lies between.

- When f or/and q are unnormalized, we can esitmate the expectation as follows

$$I = \frac{\int_{\mathcal{X}} h(x) f(x)\, dx}{\int_{\mathcal{X}} f(x)\, dx} = \frac{\int_{\mathcal{X}} h(x) \frac{f(x)}{q(x)} q^*(x)\, dx}{\int_{\mathcal{X}} \frac{f(x)}{q(x)} q^*(x)\, dx}$$

  where $q^*(x) = q(x)/c_q$
- Monte Carlo estimate

$$I_n^{\text{SNIS}} = \frac{\sum_{i=1}^n h(x^{(i)}) w(x^{(i)})}{\sum_{i=1}^n w(x^{(i)})}, \quad x^{(i)} \sim q(x)$$

- Requires a stronger condition: $q(x) > 0$ whenever $f(x) > 0$

- Unfortunately, I_n^{SNIS} is biased. However, the bias is asymptotically negligible.

$$
\begin{aligned}
I_n^{\text{SNIS}} &= \frac{1}{n}\sum_{i=1}^{n} h(x^{(i)})f(x^{(i)})/q(x^{(i)}) \Big/ \frac{1}{n}\sum_{i=1}^{n} f(x^{(i)})/q(x^{(i)}) \\
&\xrightarrow{p} \int_{\mathcal{X}} h(x)f(x)/q(x)q^*(x)\,dx \Big/ \int_{\mathcal{X}} f(x)/q(x)q^*(x)\,dx \\
&= \int_{\mathcal{X}} h(x)f(x)\,dx \Big/ \int_{\mathcal{X}} f(x)\,dx \\
&= I
\end{aligned}
$$

- We use delta method for the variance of SNIS, which is a ratio estimate

$$\mathbb{V}\mathrm{ar}(I_n^{\mathrm{SNIS}}) \approx \frac{\sigma_{q,\mathrm{sn}}^2}{n} = \frac{\mathbb{E}_q(w(x)^2(h(x)-I)^2)}{n}$$

- We can rewrite the variance $\sigma_{q,\mathrm{sn}}^2$ as

$$\begin{aligned}\sigma_{q,\mathrm{sn}}^2 &= \int_{\mathcal{X}} \frac{f(x)^2}{q(x)}(h(x)-I)^2\,dx \\ &= \int_{\mathcal{X}} \frac{(h(x)f(x)-If(x))^2}{q(x)}\,dx\end{aligned}$$

- For comparison, $\sigma_{q,\mathrm{is}}^2 = \mathbb{V}\mathrm{ar}_q(t(X)) = \int_{\mathcal{X}} \frac{(h(x)f(x)-Iq(x))^2}{q(x)}\,dx$
- No q can make $\sigma_{q,\mathrm{sn}}^2 = 0$ (unless h is constant)

Optimial SNIS

- The optimal density for self-normalized importance sampling has the form (Hesterberg, 1988)

$$q(x) \propto |h(x) - I| f(x)$$

- Using this formula we find that

$$\sigma^2_{q,\text{sn}} \geq (\mathbb{E}_f(|h(x) - I|))^2$$

  which is zero only for constant $h(x)$
- Note that the simple Monte Carlo has variance $\sigma^2 = \mathbb{E}_f((h(x) - I)^2)$, this means SNIS can not reduce the variance by

$$\frac{\sigma^2}{\sigma^2_{q,\text{sn}}} \leq \frac{\mathbb{E}_f((h(x) - I)^2)}{(\mathbb{E}_f(|h(x) - I|))^2}$$

- The importance weights in IS may be problematic, we would like to have a diagnostic to tell us when it happens.
- Unequal weighting raises variance (Kong, 1992). For IID Y_i with variance σ^2 and fixed weight $w_i \geq 0$

$$\mathbb{V}\mathrm{ar}\left(\frac{\sum_i w_i Y_i}{\sum_i w_i}\right) = \frac{\sum_i w_i^2 \sigma^2}{(\sum_i w_i)^2}$$

- Write this as

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

- n_e is the **effective sample size** and $n_e \ll n$ if the weights are too imbalanced.

- Rejection Sampling requires bounded $w(x) = f(x)/q(x)$
- We also have to know a bound for the envelop distribution
- Therefore, importance sampling is generally easier to implement
- IS and SNIS require us to keep track of weights
- Plain IS requires normalized q
- Rejection sampling could be sample inefficient (due to rejections)

Exponential Tilting

- Consider that $f(x) = p(x; \theta_0)$ is from a family of distributions $p_\theta(x),\ \theta \in \Theta$
- A simple importance sampling distribution would be $q(x) = p(x; \theta)$ for some $\theta \in \Theta$.
- Suppose $f(x)$ belongs to an exponential family

$$f(x) = g(x) \exp(\eta(\theta_0)^T T(x) - A(\theta_0))$$

- Use $q(x) = g(x) \exp(\eta(\theta)^T T(x) - A(\theta))$, the IS estimate is

$$I_n^{\text{IS}} = \exp(A(\theta) - A(\theta_0)) \cdot \frac{1}{n} \sum_{i=1}^{n} h(x^{(i)}) \exp((\eta(\theta_0) - \eta(\theta))^T T(x^{(i)})$$

- Suppose that we find the mode x^* of $k(x) = h(x)f(x)$
- We can use Taylor approximation

$$\log(k(x)) \approx \log(k(x^*)) - \frac{1}{2}(x - x^*)^T H^*(x - x^*)$$

$$k(x) \approx k(x^*) \exp\left(-\frac{1}{2}(x - x^*)^T H^*(x - x^*)\right)$$

  which suggests $q(x) = \mathcal{N}(x^*, (H^*)^{-1})$
- This requires positive definite H^*
- Can be viewed as an IS version of the Laplace approximation

- Suppose we have K importance distributions $q_1, \ldots, q_K$, we can combine them into a mixture of distributions with probability $\alpha_1, \ldots, \alpha_K$, $\sum_i \alpha_i = 1$

$$q(x) = \sum_{i=1}^{K} \alpha_i q_i(x)$$

- IS estimate $I_n^{\text{IS}} = \frac{1}{n} \sum_{i=1}^{n} h(x^{(i)}) \frac{f(x^{(i)})}{\sum_{j=1}^{K} \alpha_j q_j(x^{(i)})}$
- An alternative. Suppose $x^{(i)}$ came from component $j(i)$, we could use

$$\frac{1}{n} \sum_{i=1}^{n} h(x^{(i)}) \frac{f(x^{(i)})}{q_{j(i)}(x^{(i)})}$$

Remark: This alternative is faster to compute, but has higher variance

- Designing importance distribution directly would be challenging. A better way would be to adapt some candidate distribution to our task through a learning process
- To do that, we first need to pick a family $\mathcal{Q}$ of proposal distributions
- We have to choose a termination criterion, e.g., maximum steps, total number of observations, etc.
- Most importantly, we need a way to choose $q_{k+1} \in \mathcal{Q}$ based on the observed information

- Suppose now we have a family of distributions (e.g., exponential family) $q_\theta(x) = q(x;\theta),\ \theta \in \Theta$
- Recall that the variance of IS estimate is

$$\frac{1}{n}\int_{\mathcal{X}} \frac{(h(x)f(x))^2}{q(x)}\,dx - I^2, \quad \text{therefore, we would like}$$

$$\theta = \arg\min_{\theta\in\Theta} \int_{\mathcal{X}} \frac{(h(x)f(x))^2}{q_\theta(x)}\,dx$$

- Variance based update

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

However, the optimization may be hard.

- Consider an exponential family

$$q_\theta(x) = g(x)\exp(\theta^T x - A(\theta))$$

- Now, replace variance by KL divergence

$$D_{KL}(k_*\|q_\theta) = \mathbb{E}_{k_*}\log\left(\frac{k_*(x)}{q_\theta(x)}\right)$$

- We seek θ to minimize

$$D_{KL}(k_*\|q_\theta) = \mathbb{E}_{k_*}(\log(k_*(x)) - \log(q(x;\theta)))$$

i.e., maximize

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

- Rewrite the negative cross entropy as

$$\begin{aligned}\mathbb{E}_{k_*}(\log(q(x;\theta))) &= \mathbb{E}_q\left(\frac{\log(q(x;\theta))k_*(x)}{q(x)}\right)\\ &= \frac{1}{I}\cdot\mathbb{E}_q\left(\frac{\log(q(x;\theta))h(x)f(x)}{q(x)}\right)\end{aligned}$$

- Update θ to maximize the above

$$\begin{aligned}\theta^{(k+1)} &= \arg\max_\theta \frac{1}{n_k}\sum_{i=1}^{n_k}\frac{h(x^{(i)})f(x^{(i)})}{q(x^{(i)};\theta^{(k)})}\log(q(x^{(i)};\theta))\\ &= \arg\max_\theta \frac{1}{n_k}\sum_{i=1}^{k} H_i\log(q(x^{(i)};\theta))\\ &= \arg\max_\theta \frac{1}{n_k}\sum_{i=1}^{k} H_i(\theta^T x^{(i)} - A(\theta))\end{aligned}$$

- The update often takes a simple moment matching form

$$\frac{\partial}{\partial \theta} A(\theta^{(k+1)}) = \frac{\sum_i H_i (x^{(i)})^T}{\sum_i H_i}$$

- Examples:
  - $q_\theta = \mathcal{N}(\theta, I)$

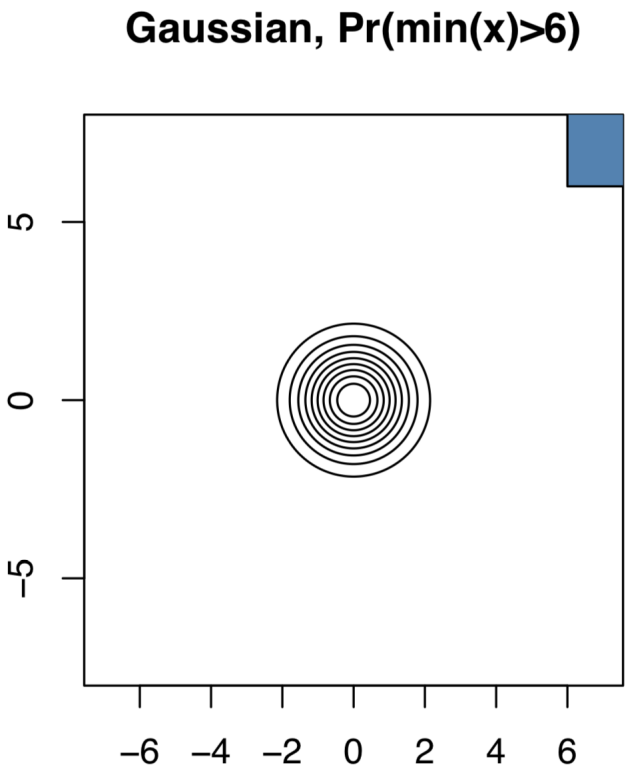
$$\theta^{(k+1)} = \frac{\sum_i H_i x^{(i)}}{\sum_i H_i}$$

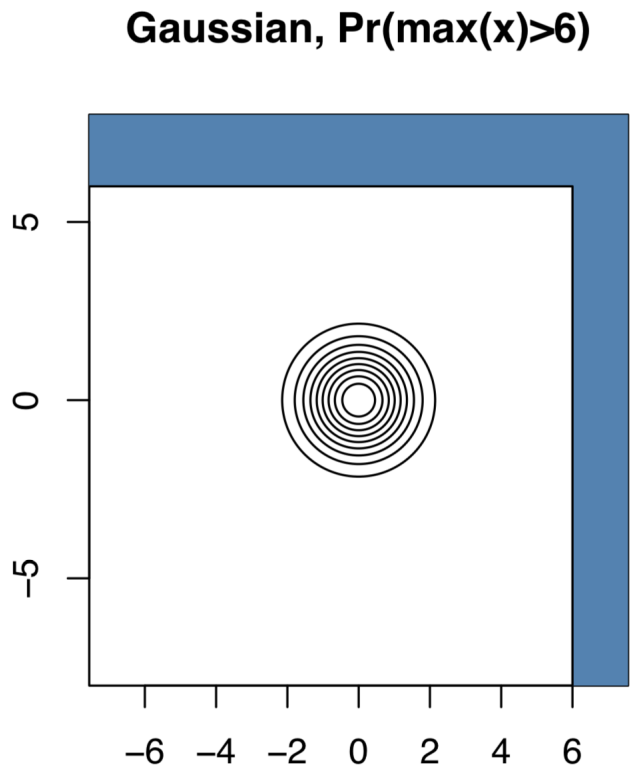
  - $q_\theta = \mathcal{N}(\theta, \Sigma)$

$$\theta^{(k+1)} = \Sigma^{-1} \frac{\sum_i H_i x^{(i)}}{\sum_i H_i}$$

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

Example

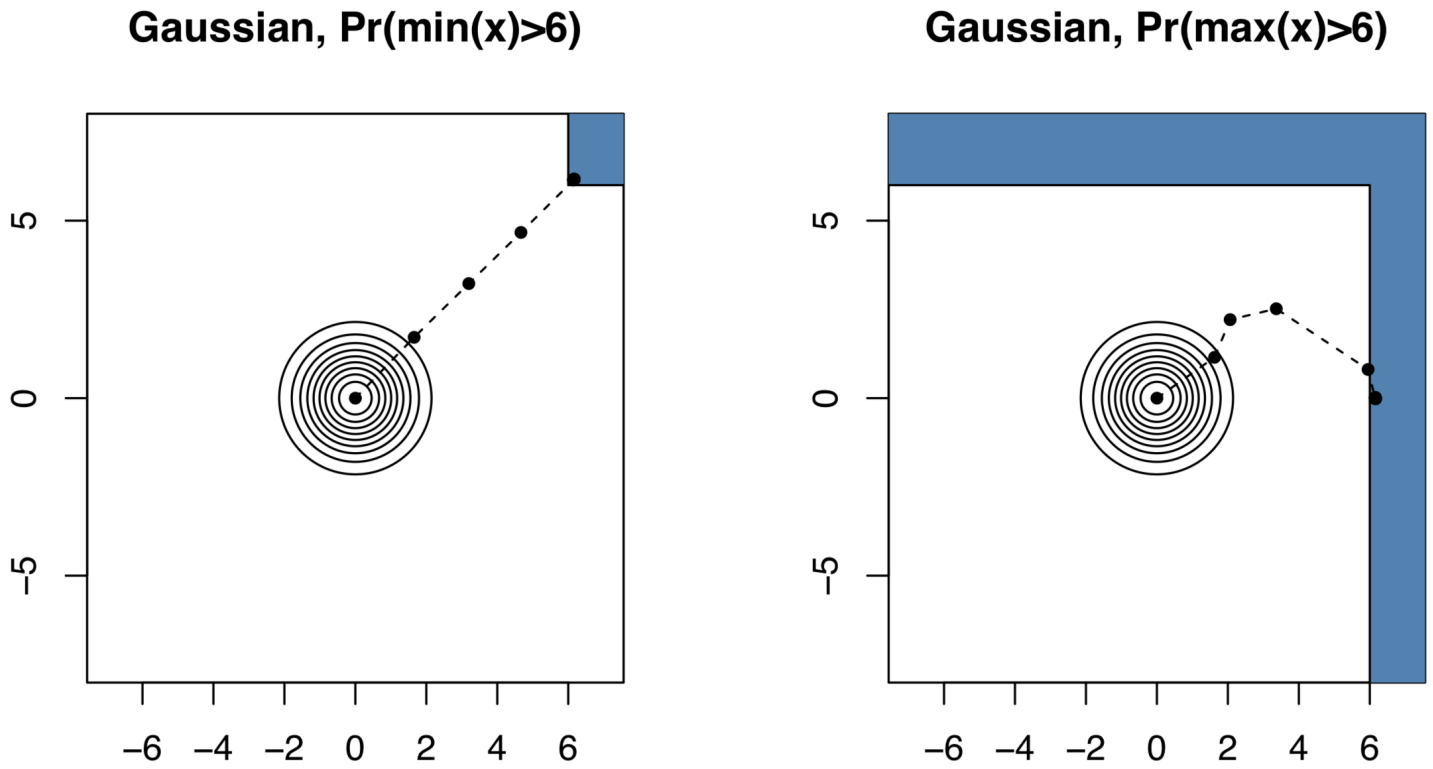




$\theta_1 = (0, 0)^T$

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

Example



For $\min(x), \theta^{(k)}$ heads Northeast, which is OK.
For $\max(x), \theta^{(k)}$ heads North or East, and miss the other part completely, leading to underestimates of I by about $1/2$

- The control variate strategy improves estimation of an unknown integral by relating the estimate to some correlated estimator with known integral
- A general class of unbiased estimators

$$I_{\mathrm{CV}} = I_{\mathrm{MC}} - \lambda(J_{\mathrm{MC}} - J)$$

where $\mathbb{E}(J_{\mathrm{MC}}) = J$. It is easy to show I_{CV} is unbiased, $\forall\lambda$

- We can choose λ to minimize the variance of I_{CV}

$$\hat{\lambda} = \frac{\mathbb{C}\mathrm{ov}(I_{\mathrm{MC}}, J_{\mathrm{MC}})}{\mathbb{V}\mathrm{ar}(J_{\mathrm{MC}})}$$

where the related moments can be estimated using samples from corresponding distributions

- Recall that IS estimator is

$$I_n^{\mathrm{IS}} = \frac{1}{n}\sum_{i=1}^{n} h(x^{(i)})w(x^{(i)})$$

- Note that $h(x)w(x)$ and $w(x)$ are correlated and $\mathbb{E}w(x) = 1$, we can use the control variate

$$\bar{w} = \frac{1}{n}\sum_{i=1}^{n} w(x^{(i)})$$

and the importance sampling control variate estimator is

$$I_n^{\mathrm{ISCV}} = I_n^{\mathrm{IS}} - \lambda(\bar{w} - 1)$$

λ can be estimated from a regression of $h(x)w(x)$ on $w(x)$ as described before

- Consider estimation of $I = \mathbb{E}(h(X, Y))$ using a random sample $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$ drawn from f
- Suppose the conditional expectation $\mathbb{E}(h(X, Y)|Y)$ can be computed. Using $\mathbb{E}(h(X, Y)) = \mathbb{E}(\mathbb{E}(h(X, Y)|Y))$, the *Rao-Blackwellized estimator* can be defined as

$$I_n^{\mathrm{RB}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(h(x^{(i)}, y^{(i)})|y^{(i)})$$

- Rao-Blackwellized estimator gives smaller variance than the ordinary Monte Carlo estimator

$$\begin{aligned} \mathbb{V}\mathrm{ar}(I_n^{\mathrm{MC}}) &= \frac{1}{n}\mathbb{V}\mathrm{ar}(\mathbb{E}(h(X, Y)|Y) + \frac{1}{n}\mathbb{E}(\mathbb{V}\mathrm{ar}(h(X, Y)|Y) \\ &\geq \mathbb{V}\mathrm{ar}(I_n^{\mathrm{RB}}) \end{aligned}$$

follows from the conditional variance formula

- Suppose rejection sampling stops at a random time M with acceptance of the nth draw, yielding $x^{(1)}, \ldots, x^{(n)}$ from all M proposals $y^{(1)}, \ldots, y^{(M)}$
- The ordinary Monte Carlo estimator can be expressed as

$$I_n^{\mathrm{MC}} = \frac{1}{n} \sum_{i=1}^{M} h(y^{(i)}) 1_{U_i \le w(y^{(i)})}$$

- Rao-Blackwellization estimator

$$I_n^{\mathrm{RB}} = \frac{1}{n} \sum_{i=1}^{M} h(y^{(i)}) t_i(Y)$$

where

$$t_i(Y) = \mathbb{E}(1_{U_i \le w(y^{(i)})} | M, y^{(1)}, \ldots, y^{(M)})$$

References

- P. J. Davis and P. Rabinowitz. Methods of Numerical Integration. Academic, New York, 1984.
- Hesterberg, T. C. (1988). Advances in importance sampling. PhD thesis, Stanford University.
- Kong, A. (1992). A note on importance sampling using standardized weights. Technical Report 348, University of Chicago.