Importance Sampling

Prof. Nicholas Zabaras
Center for Informatics and Computational Science
https://cics.nd.edu/
University of Notre Dame
Notre Dame, Indiana, USA

Email: nzabaras@gmail.com
URL: https://www.zabaras.com/

October 8, 2018

Contents

- Review of MC Methods, sample representation of the MC estimator, Accept/Reject methods
- Importance sampling methods, Examples, Optimal importance sampling distribution, Normalized importance sampling, Asymptotic variance/Delta method, asymptotic bias, optimal normanized importance sampling, computing the ratio of normalized constants, Applications to Bayesian Inference, importance sampling in high dimensions, importance sampling vs rejection sampling
- Solving Ax=b with IS, a Bayes estimate for the Cauchy distribution with a normal prior, calculating integrals with singularity, sampling importance resampling, Monte Carlo and EM, summary of importance sampling methods

Following closely:

- C.P. Roberts and G. Casella, *Monte Carlo Statistical Methods*, Chapter 3 (google books, slides, video)
- J S Liu, Monte Carlo Strategies in Scientific Computing, Chapter 2
- A. Doucet, Statistical Computing and Monte Carlo Methods (2007)
- J-M Marin and C. P. Robert, Bayesian Core (Chapter 2)

Sample Representation of the MC Estimator

- Let $\pi(x)$ be a probability density on $\mathcal{X}$.

- Monte Carlo approximation is given by

$$\hat{\pi}_N(x) := \frac{1}{N}\sum_{i=1}^{N} \delta_{X^{(i)}}(x)\,,\ where\ X^{(i)} \overset{i.i.d.}{\sim} \pi$$

- For any function $\phi : \mathcal{X} \to \mathbb{R}$, we have then the following:

$$\mathbb{E}_{\hat{\pi}_N}(\phi(X)) = \frac{1}{N}\sum_{i=1}^{N} \phi(X^{(i)}) \simeq \mathbb{E}_{\pi}(\phi(X)$$

- The Monte Carlo estimator is unbiased :

$$\mathbb{E}_{\{X^{(i)}\}}\left[\mathbb{E}_{\hat{\pi}_N}(\phi(X)\right)] = \mathbb{E}_{\pi}(\phi(X)),$$

and has variance:

$$Var_{\{X^{(i)}\}}\left[\mathbb{E}_{\hat{\pi}_N}(\phi(X)\right)] = \frac{Var_{\pi}(\phi(X))}{N}$$

Direct Sampling Methods

We discussed already:

- Inverting the cdf
  - Practically impossible is high dimensions
  - In many cases, cdf is not known because the density is only known up to a normalizing constant
  - We discussed several problem-dependent transformation methods
- Rejection Sampling, which requires:
  - the target density (known up to a constant), $\pi(x)$
  - a proposal density, $q(x)$
  - a bounding constant, $C \geq \frac{\pi(x)}{q(x)} \; \forall x \in \mathcal{X}$
  - ✓ However, it is difficult to find good proposal densities $q(x)$, especially in high dimensions and
  - ✓ poor choices result in inefficient sampling i.e. low acceptance rate

Direct Sampling Methods

- Direct methods such as the inverse method, composition method, etc. are only appropriate for standard distributions.

- When $\pi \propto \pi^*$ does not admit a standard form, we use a proposal distribution q on $\mathcal{X}$ where we need q to dominate π, i.e.

$$M = \sup_{x \in \mathcal{X}} \frac{\pi^*(x)}{q^*(x)} < +\infty$$

Accept/Reject Algorithm

- We would like to sample from $\pi(x)$, but it is easier to sample from a *proposal distribution* $q(x)$
- $q(\boldsymbol{x})$ satisfies $\pi(\boldsymbol{x}) < M' q(\boldsymbol{x})$ for all $\boldsymbol{x}$ for some $M' \geq M$
- Procedure:
  - Sample Y from q *and* u *from* $\mathcal{U}[0,1]$
  - Accept (set $X = Y$) with probability $\pi^*(Y) \,/\, Mq^*(Y)$ (i.e. if $u \leq \pi^*(Y) \,/\, Mq^*(Y)$)
  - Reject otherwise and repeat.
- The accepted $X^{(i)}$ are sampled from $\pi(x)$!
- If M' is too large, we will rarely accept samples.

Accept/Reject Algorithm

- The implementation of the algorithm is simple once you have a bounding constant M.
- The performance of the algorithm degrades exponentially with the dimensionality of X.
- There is significant waste by rejecting samples that can provide some useful information.
- We are waiting a random time to obtain samples from π.

Is there a way to make use of all the samples drawn from the proposal density?

Importance Sampling

- Consider a target distribution $\pi(x)$ and a proposal distribution $q(x)$ with the constraint:

$$\pi(x) > 0 \Rightarrow q(x) > 0$$

- We are looking for a sampling method where ALL samples drawn from the proposal density $q(x)$ are usable.

$$\mathbb{E}_{\pi}\{f(\mathrm{x})\} \equiv \int f(\mathrm{x})\pi(\mathrm{x})d\mathrm{x} = \int_{x} \underbrace{\frac{f(\mathrm{x})\pi(\mathrm{x})}{q(\mathrm{x})}}_{f_q(\mathbf{x})} q(\mathrm{x})d\mathrm{x} = \int_{x} f(\mathrm{x}) \underbrace{\frac{\pi(\mathrm{x})}{q(\mathrm{x})}}_{w(x)} q(\mathrm{x})d\mathrm{x} = \mathbb{E}_q\left[w(X)f(X)\right]$$

- Note that with the above equation computing $E_{\pi}\{f(X)\}$ is transformed to computing

$$\mathbb{E}_q\left\{\underbrace{\frac{f(X)\pi(X)}{q(X)}}_{f_q(X)}\right\} = \mathbb{E}_q\left[w(X)f(X)\right] = \mathbb{E}_q\left[f_q(X)\right]$$

Importance Sampling

- Computing $\mathbb{E}_{\pi}\{f(X)\}$ is transformed to computing

$$\mathbb{E}_{q}\left\{\underbrace{\frac{f(X)\pi(X)}{q(X)}}_{f_q(X)}\right\}$$

- We can employ MC for this task:
  - Generate samples $X^{(i)} \sim q(x)\,(i.i.d.)$ and evaluate:

$$\mathbb{E}_{\hat{q}_N}(f_q(x)) = \frac{1}{N}\sum_{i=1}^{N} f_q(X^{(i)}) = \frac{1}{N}\sum_{i=1}^{N} \frac{\pi(X^{(i)})}{q(X^{(i)})} f(X^{(i)}) = \frac{1}{N}\sum_{i=1}^{N} w(X^{(i)}) f(X^{(i)})$$

where the importance weight is $w(X^{(i)}) \equiv \frac{\pi(X^{(i)})}{q(X^{(i)})}$.

Importance Sampling

- Monte Carlo approximation of $q(x)$ is:

$$\hat{q}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{X^{(i)}}(x) \;\; where \; X^{(i)} \overset{i.i.d.}{\sim} q$$

- It follows that an estimate of $\mathbb{E}_\pi[f(X)] = \mathbb{E}_q[w(X)f(X)]$ is

$$\mathbb{E}_{\hat{q}_N}\big(w(X)f(X)\big) = \frac{1}{N}\sum_{i=1}^{N} w(X^{(i)})f(X^{(i)})$$

- This corresponds to the following Monte Carlo approximation of $\pi(x)$:

$$\hat{\pi}_N(x) = \frac{1}{N}\sum_{i=1}^{N} w(X^{(i)})\delta_{X^{(i)}}(x) \;\; where \; X^{(i)} \overset{i.i.d.}{\sim} q$$

Importance Sampling

- The estimator is unbiased

$$\mathbb{E}_{\{X^{(i)}\}}\left[\mathbb{E}_{\hat{q}_N}\left(w(X)f(X)\right)\right] = \mathbb{E}_q\left(f_q(X)\right) = \mathbb{E}_q\left(\frac{\pi(X)}{q(X)}f(X)\right) = \mathbb{E}_\pi\left(f(X)\right)$$

- It has a variance

$$Var_{\{X^{(i)}\}}\left[\mathbb{E}_{\hat{q}_N}\left(w(X)f(X)\right)\right] = \frac{Var_q\left(w(X)f(X)\right)}{N} = \frac{\mathbb{E}_\pi\left(w(X)f^2(X)\right) - \mathbb{E}_\pi^2\left(f(X)\right)}{N}$$

- This is finite if $\mathbb{E}_\pi\left(w(X)f^2(X)\right) = \int \frac{\pi(x)^2}{q(x)^2} f^2(x)q(x)dx < \infty$.
- It is recommended that the selection of $q(x)$ should ensure

$$\mathbb{E}_\pi\left(w(X)\right) = \int \frac{\pi^2(x)}{q(x)}dx < +\infty$$

or even better:

$$\sup_{x\in\mathcal{X}} w(x) < +\infty$$

Importance Sampling

- We can estimate the expectation (integral) of interest by drawing samples from (practically) any distribution q
- Instead of assigning equal weight (1), the samples are weighted according to $w(x) = \frac{\pi(x)}{q(x)}$.
- The estimator is unbiased and converges to the correct value (Strong Law of Large Numbers)
- The rate of convergence (variance of the estimator) depends on the discrepancy between $f(\boldsymbol{x})\pi(\boldsymbol{x})$ and the importance sampling distribution $q(\boldsymbol{x})$

Computing the Normalizing Constant

- Suppose we want to compute the normalizing constant

$$Z = \int \pi(x)dx = \int \frac{\pi(x)}{q(x)} q(x)dx$$

- The important sampling estimator is

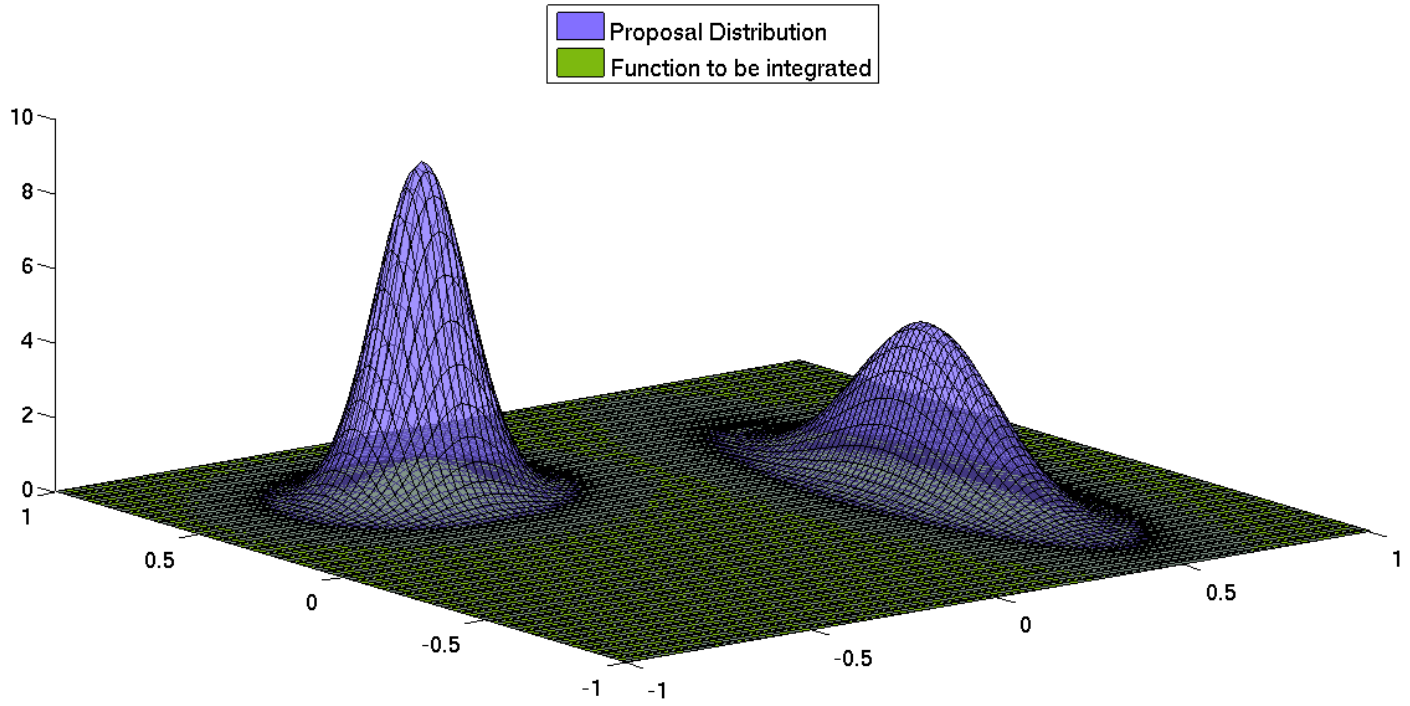
$$\hat{Z} = \frac{1}{N}\sum_{i=1}^{N} \frac{\pi(X^{(i)})}{q(X^{(i)})}, \quad X^{(i)} \sim q\ (i.i.d.)$$

Example of importance sampling

- Consider as an example the following integral (see here for a MatLab implementation)

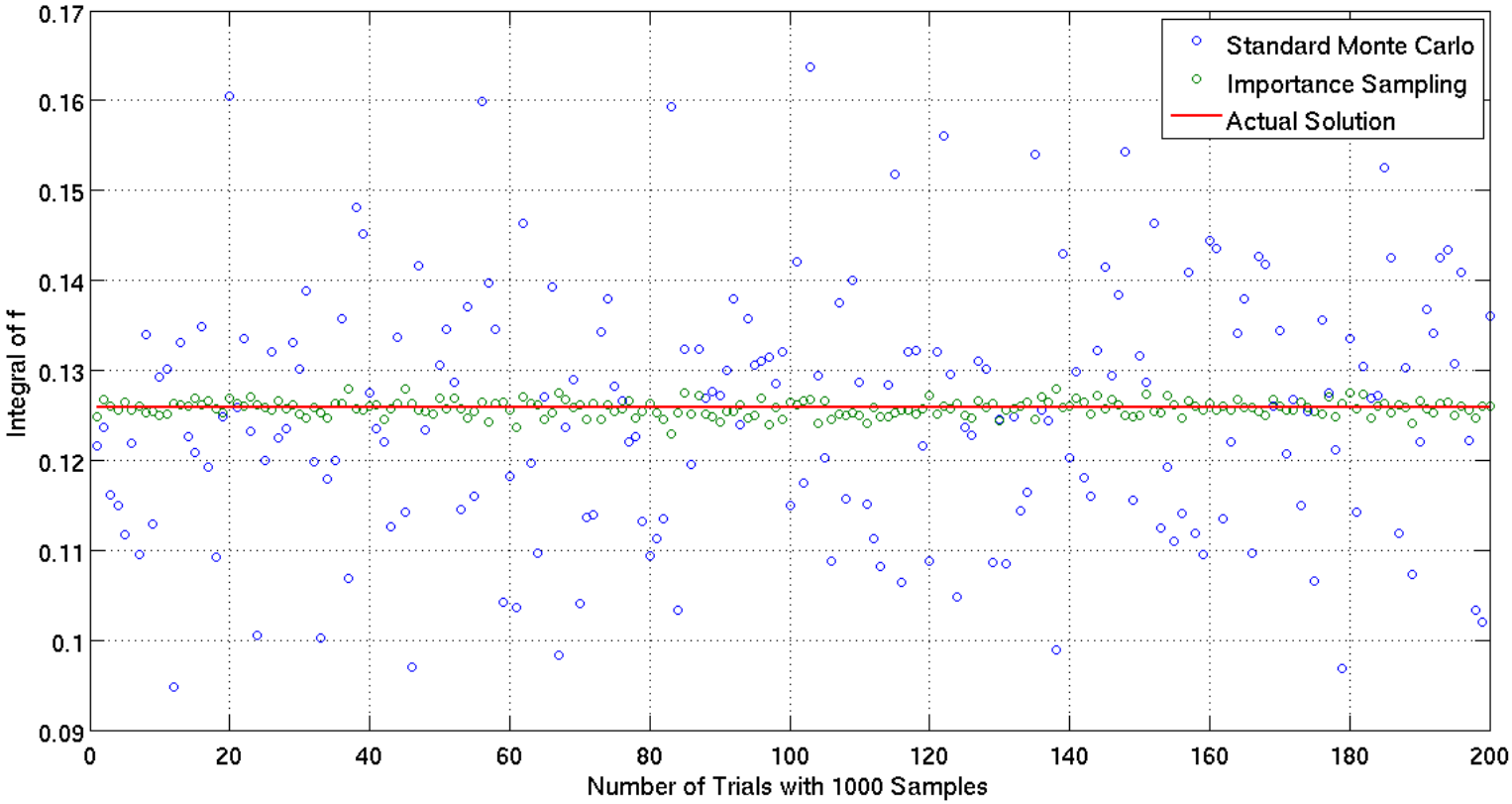
$$I = \iint_{[-1,1]\times[-1,1]} f(x,y)\,dxdy \qquad f(x,y) = 0.5e^{-90(x-0.5)^2-45(y+0.1)^4} + e^{-45(x+0.4)^2-60(y-0.5)^2}$$

$$Proposal\ Distribution: q(x,y) = 0.46\mathcal{N}\left(\begin{bmatrix}0.5\\-0.1\end{bmatrix},\begin{bmatrix}1/180 & 0\\0 & 1/20\end{bmatrix}\right) + 0.54\mathcal{N}\left(\begin{bmatrix}-0.4\\0.5\end{bmatrix},\begin{bmatrix}1/90 & 0\\0 & 1/120\end{bmatrix}\right)$$



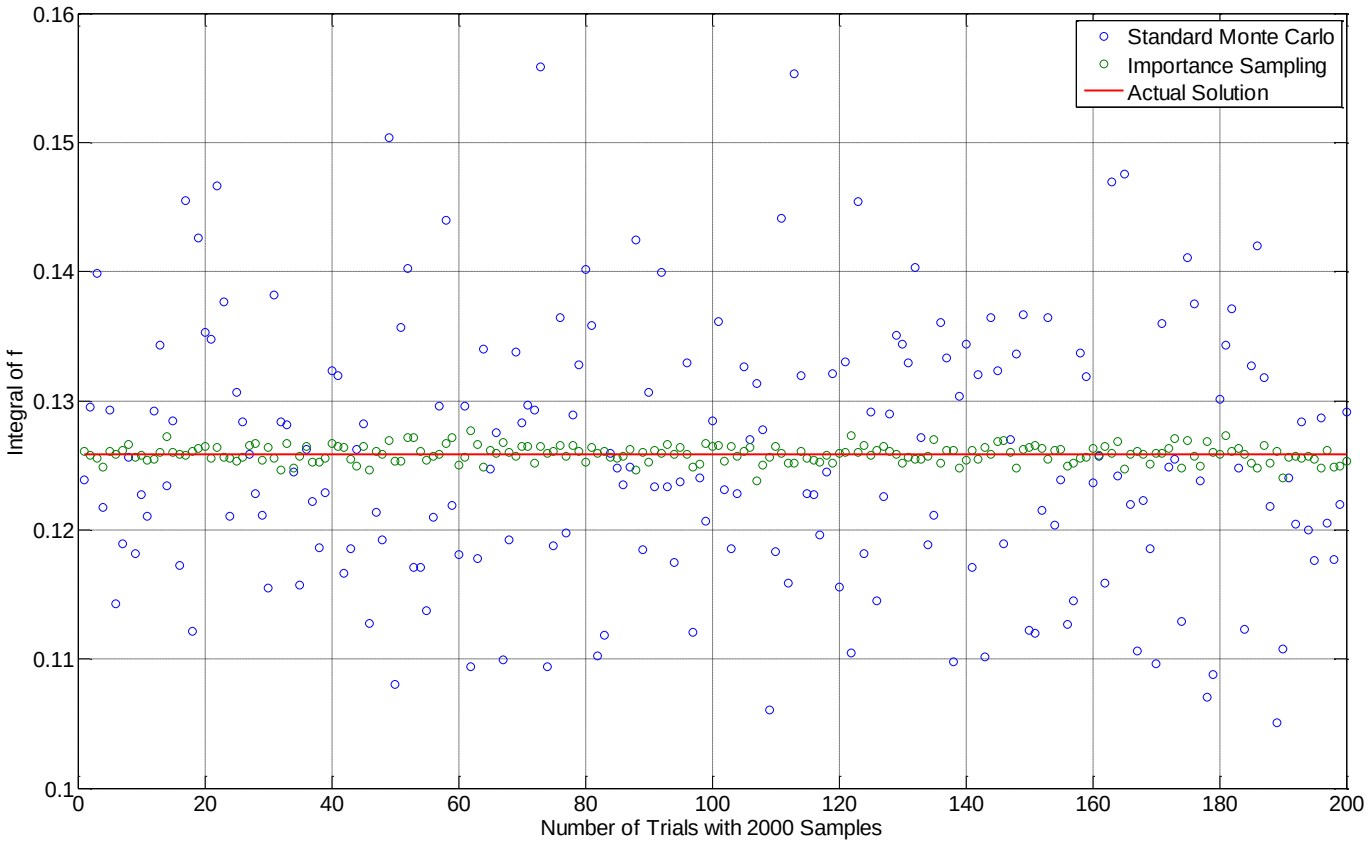
Example of Importance Sampling

- $N = 1000$, count $= 200$ (we take 1000 random sample points per run and run the simulation 200 times)
- The results of importance sampling are more accurate than the standard MC method.



Example of Importance Sampling

- $N = 2000$, count $= 200$ (we take 2000 random sample points per run and run the simulation 20 times)
- The results of importance sampling are more accurate than the standard MC method.



Sampling from a Gaussian Mixture

The basic idea for the algorithm used in the earlier example:

- introduce a variable z with probability $\Pr(z) = \begin{cases} p, & \text{when z=1} \\ 1-p, & \text{when z=2} \end{cases}$
- suppose the conditional distribution $q(x|z)$

$$q(x \mid z=1) = \mathcal{N}(\mu_1, \sigma_1)$$

$$q(x \mid z=2) = \mathcal{N}(\mu_2, \sigma_2)$$

- Thus, the marginal distribution of $q(x)$ is

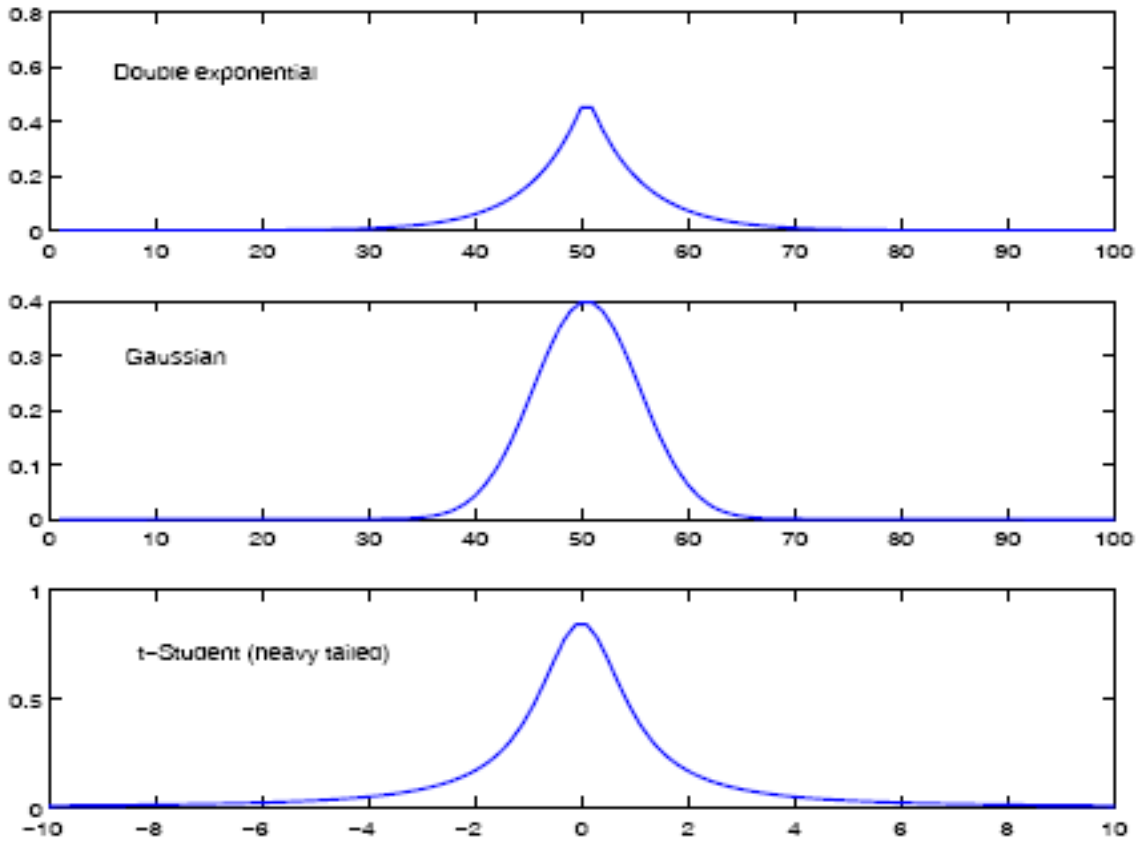
$$q(x) = \sum_z q(x \mid z)\Pr(z) = q(x \mid z=1)\Pr(z=1) + q(x \mid z=2)\Pr(z=2)$$

$$= p\mathcal{N}(\mu_1, \sigma_1) + (1-p)\mathcal{N}(\mu_2, \sigma_2)$$

- By introducing the "hidden" variable z, we can sample from $q(x)$ much easier by independently sampling from $\mathcal{N}(\mu_1, \sigma_1)$ *and* $\mathcal{N}(\mu_2, \sigma_2)$

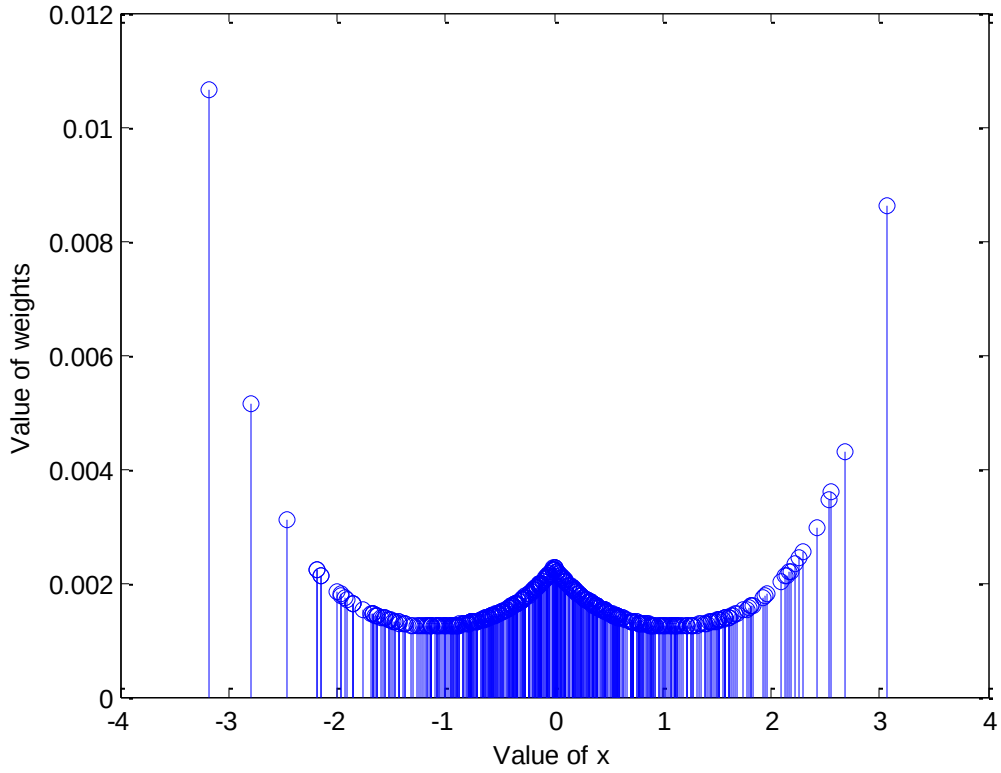
Example of Importance Sampling

- Target double exponential distribution. We use two importance sampling distributions – Gaussian and $\mathcal{T}$-Student's (see here for a MatLab implementation)



Example of Importance Sampling

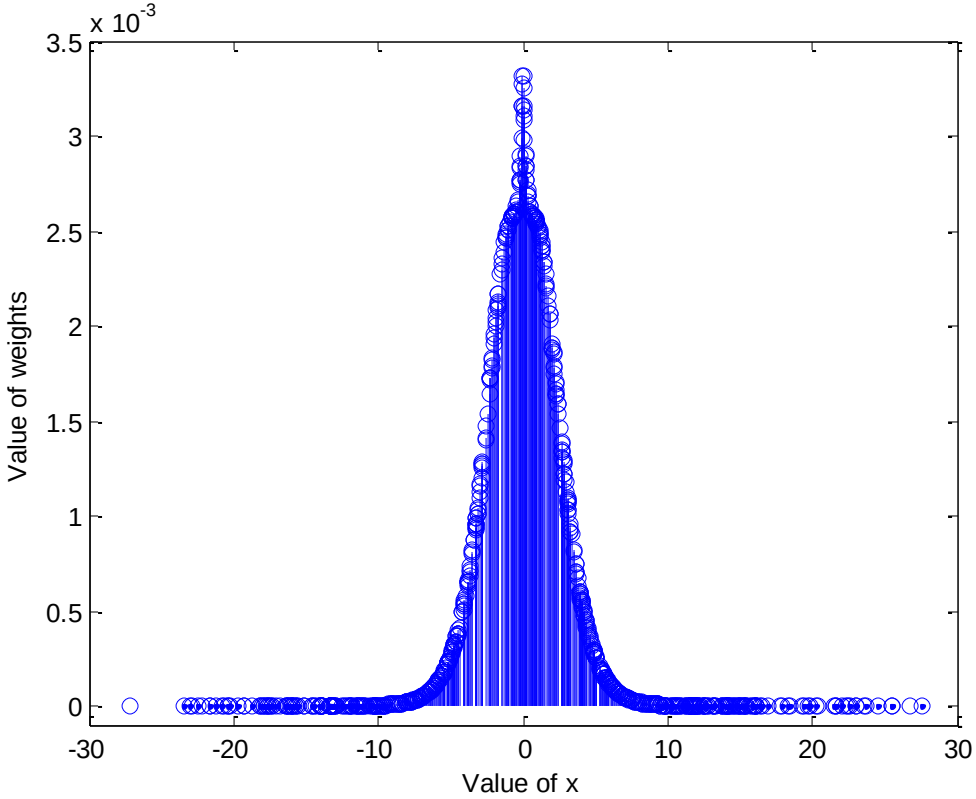
- Importance sampling approximation obtained using a Gaussian importance sampling distribution
- The approximation is poor as the Gaussian has thinner tails than the target.



See here for a MatLab implementation

Example of Importance Sampling

- Importance sampling approximation obtained using a Student's-$\mathcal{T}$ importance sampling distribution
- The approximation is now correct as the Student's-$\mathcal{T}$ has thicker tails than the target.



See here for a MatLab implementation

Optimal Importance Sampling Distribution

- For a given test function $f(x)$, one can minimize the IS variance using

$$q^{opt}(x) = \frac{|f(x)|\pi(x)}{\int_{\mathcal{X}} |f(x)|\pi(x)dx}$$

- Proof: $$Var_q\left(w(X)f(X)\right) = \int q(x)\underbrace{\frac{\pi^2(x)}{q^2(x)}}_{g^2(x)} f^2(x)dx - \left(\int \pi(x)f(x)dx\right)^2$$

- We use next that:

$$Var_q\left(g(X)\right) = \int q(x)g^2(x)dx - \left(\int q(x)g(x)dx\right)^2 \geq 0 \quad \forall\, g(x)$$

- From which we conclude:

$$\int q(x)\underbrace{\frac{\pi^2(x)}{q^2(x)}}_{g^2(x)} f^2(x)dx \geq \left(\int q(x)\underbrace{\frac{\pi(x)|f(x)|}{q(x)}}_{|g(x)|}dx\right)^2 = \left(\int \pi(x)|f(x)|dx\right)^2 \ (\textit{independent of } q)$$

- The optimum bound is obtained (check by substitution) as

$$\int q(x)\frac{\pi^2(x)}{q^2(x)} f^2(x)dx = \left(\int \pi(x)|f(x)|dx\right)^2 \Rightarrow q^{opt}(x) = \frac{|f(x)|\pi(x)}{\int_{\mathcal{X}} |f(x)|\pi(x)dx}$$

Optimal Importance Sampling Distribution

- For a given test function $f(x)$, one can minimize the IS variance using $$q^{opt}(x) = \frac{|f(x)|\pi(x)}{\int_{\mathcal{X}} |f(x)|\pi(x)dx}$$
- If $f(\boldsymbol{x})$ has a constant sign, then $Var_{q^{opt}} = 0$!
- Even though the optimal importance sampling density is analytically available, in practice:
  - We cannot readily draw samples from q^{opt}
  - We need to know $\int_{\mathcal{X}} |f(x)|\pi(x)dx$ in order to evaluate the weights $w(x) = \frac{\pi(x)}{q(x)}$.

Normalized Importance Sampling

- In most applications, standard importance sampling cannot be applied as the importance weights $w(x) = \pi(x)/q(x)$ cannot be evaluated in closed-form.

- In practice, we only know $\pi(x) \propto \pi^*(x)$ and $q(x) \propto q^*(x)$.

- Normalized importance sampling identity is based on:

$$\pi(x) = \frac{\pi^*(x)}{\int \pi^*(x)dx} = \frac{\frac{\pi^*(x)}{q^*(x)} q^*(x)}{\int \frac{\pi^*(x)}{q^*(x)} q^*(x)dx} = \frac{w^*(x)q^*(x)}{\int w^*(x)q^*(x)dx} \Rightarrow \pi(x) = \frac{w^*(x)q(x)}{\int w^*(x)q(x)dx}$$

where

$$w^*(x) = \frac{\pi^*(x)}{q^*(x)}$$

Normalized Importance Sampling

- For any test function $f(x)$, we can write:

$$\mathbb{E}_{\pi}(f(x)) = \int \pi(x) f(x) dx = \frac{1}{\int w^*(x) q(x) dx} \int w^*(x) q(x) f(x) dx =$$

$$= \frac{\mathbb{E}_q\left(w^*(X) f(X)\right)}{\mathbb{E}_q\left(w^*(X)\right)} = \frac{\mathbb{E}_q\left(w^*(X) f(X)\right)}{\mathbb{E}_q\left(w^*(X)\right)} = \frac{\mathbb{E}_q\left(w(X) f(X)\right)}{\mathbb{E}_q\left(w(X)\right)}$$

- Let us consider a Monte Carlo approximation of q:

$$\hat{q}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{X^{(i)}}(x) \;\; where \; X^{(i)} \overset{i.i.d.}{\sim} q$$

Normalized Importance Sampling

- Given a Monte Carlo approximation of q:

$$\hat{q}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{X^{(i)}}(x) \; where \; X^{(i)} \overset{i.i.d.}{\sim} q$$

- Then:

$$\hat{\pi}_N(x) = \sum_{i=1}^{N} W^{(i)} \delta_{X^{(i)}}(x), \; where \; W^{(i)} = \frac{w^*(X^{(i)})}{\sum_{j=1}^{N} w^*(X^{(j)})}$$

$$\mathbb{E}_{\hat{\pi}_N}(f(x)) = \sum_{i=1}^{N} W^{(i)} f(X^{(i)})$$

- The estimate is now the ratio of two MC estimates.

Normalized Importance Sampling

- The normalized importance sampling estimate is biased (the standard importance sampling is not)

- However, the normalized importance sampling is asymptotically unbiased (by the central limit it is asymptotically consistent).

- The asymptotic bias and variance can be derived using the delta method.

The Delta Method

- Assume $Z = g(A, B)$ with $\mathbb{E}(A) = \mu_A$ and $\mathbb{E}(B) = \mu_B$.
- A Taylor series expansion around $\mu = (\mu_A, \mu_B)$ gives:

$$Z \simeq g(\mu) + (A - \mu_A)\frac{\partial g}{\partial A}(\mu) + (B - \mu_B)\frac{\partial g}{\partial B}(\mu)$$

- It follows that

$$\mathbb{E}(Z) \simeq g(\mu)$$

$$Var(Z) \simeq \sigma_A^2 \left(\frac{\partial g}{\partial A}(\mu)\right)^2 + \sigma_B^2 \left(\frac{\partial g}{\partial B}(\mu)\right)^2 + 2\frac{\partial g}{\partial A}(\mu)\frac{\partial g}{\partial B}(\mu)\sigma_{A,B}$$

- In our case of the normalized importance sampling:

$$Z = \mathbb{E}_{\hat{\pi}_N}(f(X)) = \frac{\mathbb{E}_{\hat{q}_N}\left(w^*(X)f(X)\right)}{\mathbb{E}_{\hat{q}_N}\left(w^*(X)\right)} = \frac{A}{B}, A = \mathbb{E}_{\hat{q}_N}\left(w^*(X)f(X)\right),\ B = \mathbb{E}_{\hat{q}_N}\left(w^*(X)\right),$$

$$g(A, B) = \frac{A}{B},\ \mathbb{E}_q\left(\mathbb{E}_{\hat{\pi}_N}(f(X))\right) \simeq \frac{\mathbb{E}_q\left(\mathbb{E}_{\hat{q}_N}\left(w^*(X)f(X)\right)\right)}{\mathbb{E}_q\left(\mathbb{E}_{\hat{q}_N}\left(w^*(X)\right)\right)} = \frac{\mathbb{E}_q\left(w^*(X)f(X)\right)}{\mathbb{E}_q\left(w^*(X)\right)} = \mathbb{E}_\pi(f(X))$$

Asymptotic Variance of the Normalized IS

- Using $g(A,B)=\frac{A}{B}$, we can compute:

$$\frac{\partial g}{\partial A}(\mu)\frac{\partial g}{\partial B}(\mu)=\frac{1}{B}\left(-\frac{A}{B^2}\right)\Bigg|_{(\mu_A,\mu_b)}=-\frac{\mu_A}{\mu_B^3},\left(\frac{\partial g}{\partial A}(\mu)\right)^2=\frac{1}{\mu_B^2},\left(\frac{\partial g}{\partial B}(\mu)\right)^2=\frac{\mu_A^2}{\mu_B^4}$$

$$\mu_A=\mathbb{E}_q(w^*(X)f(X)),\ \mu_B=\mathbb{E}_q(w^*(X))$$

$$\sigma_A^2=\frac{Var_q(w^*(X)f(X))}{N},\ \sigma_B^2=\frac{Var_q(w^*(X))}{N}$$

$$\sigma_{A,B}=\frac{Cov_q\left(w^*(X)f(X),w^*(X)\right)}{N}=\frac{\mathbb{E}_q\left(w^*(X)^2f(X)\right)-\mathbb{E}_q\left(w^*(X)f(X)\right)\mathbb{E}_q\left(w^*(X)\right)}{N}$$

$$=\frac{\mathbb{E}_q\left(w^*(X)^2f(X)\right)-\mu_A\mu_B}{N}$$

Asymptotic Variance of the Normalized IS

- It follows that

$$Var\left(\mathbb{E}_{\hat{\pi}_N}(f(X))\right) \simeq \sigma_A^2\left(\frac{\partial g}{\partial A}(\mu)\right)^2 + \sigma_B^2\left(\frac{\partial g}{\partial B}(\mu)\right)^2 + 2\frac{\partial g}{\partial A}(\mu)\frac{\partial g}{\partial B}(\mu)\sigma_{A,B} =$$

$$= \frac{\sigma_A^2}{\mu_B^2} + \frac{\sigma_B^2\mu_A^2}{\mu_B^4} - 2\frac{\sigma_{A,B}\mu_A}{\mu_B^3}$$

- Straight forward substitution of the expressions for $\sigma_A^2, \sigma_B^2, \mu_A, \mu_B$, and using $\mathbb{E}_q\left(\mathbb{E}_{\hat{\pi}_N}(f(X))\right) \simeq \mathbb{E}_\pi(f(X))$ leads to the following CLT asymptotic result:

$$\sqrt{N}\left(\mathbb{E}_{\hat{\pi}_N}(f(X)) - \mathbb{E}_\pi(f(X))\right) \sim \mathcal{N}(0, \sigma_{IS}^2(f))$$

where

$$\sigma_{IS}^2(f) = \int \frac{\pi^2(x)}{q(x)}\left(f(x) - \mathbb{E}_\pi(f)\right)^2 dx = Var_q\left(w(X)\left(f(X) - \mathbb{E}_\pi(f)\right)\right)$$

Asymptotic Variance of the Normalized IS

- Even if it is not necessary, it is highly recommended to select $q(x)$ such that

$$\sup_{x \in \mathcal{X}} w(x) < \infty \; or \; equivalently \; \sup_{x \in \mathcal{X}} w^*(x) < \infty$$

- Normalized importance sampling often performs better than standard importance sampling.

Asymptotic Bias

- We can compute the bias using the δ-method with a 2nd order Taylor series expansion

$$Z \simeq g(\mu) + (A - \mu_A)\frac{\partial g}{\partial A}(\mu) + (B - \mu_B)\frac{\partial g}{\partial B}(\mu) +$$
$$+\frac{1}{2}(A - \mu_A)^2\frac{\partial^2 g}{\partial A^2}(\mu) + \frac{1}{2}(\mathrm{B} - \mu_B)^2\frac{\partial^2 g}{\partial B^2}(\mu) + (A - \mu_A)(B - \mu_B)\frac{\partial^2 g}{\partial A \partial B}(\mu)$$

- This gives

$$\mathbb{E}_{\mathrm{q}}\left(\mathbb{E}_{\hat{\pi}_{\mathrm{N}}}(\mathrm{f}(\mathrm{X}))\right) \simeq \mathrm{g}(\mu) + \frac{1}{2}\sigma_{\mathrm{A}}^2\frac{\partial^2 \mathrm{g}}{\partial \mathrm{A}^2}(\mu) + \frac{1}{2}\sigma_{\mathrm{B}}^2\frac{\partial^2 \mathrm{g}}{\partial \mathrm{B}^2}(\mu) + \sigma_{\mathrm{A,B}}\frac{\partial^2 \mathrm{g}}{\partial \mathrm{A}\partial \mathrm{B}}(\mu)$$

- Substitution of $\sigma_A^2, \sigma_B^2, \mu_A, \mu_B$ gives that asymptotically:

$$N\left(\mathbb{E}_q\left(\mathbb{E}_{\hat{\pi}_N}(f(X))\right) - \mathbb{E}_\pi\left(f(X)\right)\right) \sim -\int \frac{\pi^2(x)}{q(x)}\left(f(x) - \mathbb{E}_\pi(f)\right)dx$$

- *Thus we have Bias^2 of order $1/N^2$ & variance of order $1/N$.*

Optimal Normalized Importance Sampling

- For a given test function, one can minimize the normalized importance sampling variance using

$$q^{opt}(x) = \frac{\left| f(x) - \mathbb{E}_{\pi}(f) \right| \pi(x)}{\int_X \left| f(x) - \mathbb{E}_{\pi}(f) \right| \pi(x) dx}$$

- Proof: Noticing the form of the asymptotic variance, we write:

$$Var_q\left(w(X)\left(f(X) - \mathbb{E}_{\pi}(f) \right) \right) = \int q(x) \frac{\pi^2(x)}{q^2(x)} \left(f(x) - \mathbb{E}_{\pi}(f) \right)^2 dx - \left(\int \pi(x) \left(f(x) - \mathbb{E}_{\pi}(f) \right) dx \right)^2$$

where similarly to an earlier importance sampling proof

$$\int q(x) \frac{\pi^2(x)}{q^2(x)} \left(f(x) - \mathbb{E}_{\pi}(f) \right)^2 dx \geq \left(\int q(x) \frac{\pi(x) \left| f(x) - \mathbb{E}_{\pi}(f) \right|}{q(x)} dx \right)^2 = \left(\int \pi(x) \left| f(x) - \mathbb{E}_{\pi}(f) \right| dx \right)^2$$

- The optimum bound is obtained for $q^{opt}(x)$ as given above.

Choosing q(x) and Approximate Variance

- In practice, we are interested to choose $q(x)$ based on a specific $f(x)$.
- In general, we prefer $q(x)$ as close as possible to $\pi(x)$.
- For flat functions, one can approximate the variance by:

$$Var_q(\mathbb{E}_{\hat{\pi}_N}(f(X))) \simeq \Big(1 + Var_q(w(X))\Big)\frac{Var\Big(\mathbb{E}_\pi(f(X))\Big)}{N}$$

- This can be interpreted as N –weighted samples being approximately equivalent to M unweighted samples from π where

$$M = \frac{N}{1+Var_q(w(X))} \leq N$$

Computing the Ratio of Normalizing Constants

- One can calculate the ratio of normalizing constants.

$$\frac{\int \pi^*(x)dx}{\int q^*(x)dx} = \int \underbrace{\frac{\pi^*(x)}{q^*(x)}}_{w^*} \underbrace{\frac{q^*(x)}{\int q^*(x')dx'}}_{q(x)} dx = \int w^*(x)q(x)dx = \mathbb{E}_q(w^*(X))$$

- We can now approximate:

$$\mathbb{E}_{\hat{q}_N}[w^*(X)] = \frac{1}{N}\sum_{i=1}^{N} w^*(X^{(i)})$$

which is unbiased and has variance

$$Var\left[\mathbb{E}_{\hat{q}_N}[w^*(X)]\right] = \frac{Var_q(w^*(X))}{N}$$

- For the case $q(x) = \pi(x)$, then:

$$\mathbb{E}_{\hat{q}_N}[w^*(X)] = \frac{\int \pi^*(x)dx}{\int q^*(x)dx} \Rightarrow Var\left[\mathbb{E}_{\hat{q}_N}[w^*(X)]\right] = 0$$

Importance Sampling – Application to Bayesian Inference

- Let us return to our Bayesian model: prior $\pi(x)$ and likelihood $f(x|\theta)$

- The posterior distribution is given by:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta} \propto \pi^*(\theta|x), \text{ where } \pi^*(\theta|x) = f(x|\theta)\pi(\theta)$$

- We can use the prior distribution as a candidate distribution

$$q(\theta) = q^*(\theta) = \pi(\theta) \Rightarrow w^*(\theta) = f(x|\theta)$$

- We can also obtain an estimate of the marginal likelihood

$$\int_{\Theta} \pi(\theta) f(x|\theta) d\theta$$

Importance Sampling – Application to Bayesian Inference

- Assume that the likelihood is of a complicated form:

$$\int f(x, z \mid \theta) dz$$

- In this case you need to compute the importance weight

$$w\left(\theta^{(i)}\right) \propto \int f(x, z \mid \theta^{(i)}) dz$$

  which does not admit a closed-form expression.

- In such cases, an unbiased estimate of $w\left(\theta^{(i)}\right)$ is sufficient.

Importance Sampling in High-Dimensions

- Consider the following target distribution

$$\pi(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{\sum_{i=1}^{d} x_i^2}{2}}$$

- We take the following reasonable importance sampling distribution ($\sigma>1$)

$$q_\sigma(\boldsymbol{x}) = \mathcal{N}(0, \sigma^2 I) = \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{\sum_{i=1}^{d} x_i^2}{2\sigma^2}}$$

- Note that:

$$w_\sigma(x) = \frac{\pi(x)}{q_\sigma(x)} = \sigma^d e^{-\frac{1}{2}\sum_{i=1}^{d} x_i^2 (1-\frac{1}{\sigma^2})} \leq \sigma^d \ \ \forall x$$

Importance Sampling in High-Dimensions

$$q_\sigma(\boldsymbol{x}) = \mathcal{N}(0, \sigma^2 I) = \frac{1}{\left(2\pi\sigma^2\right)^{d/2}} e^{-\frac{\sum_{i=1}^{d} x_i^2}{2\sigma^2}} \qquad w_\sigma(x) = \frac{\pi(x)}{q_\sigma(x)} = \sigma^d e^{-\frac{1}{2}\sum_{i=1}^{d} x_i^2 \left(1-\frac{1}{\sigma^2}\right)} \leq \sigma^d \ \forall x$$

- We now note that:

$$Var_{q_\sigma}\left(\frac{\pi(x)}{q_\sigma(x)}\right) = \int q_\sigma \frac{\pi^2(x)}{q_\sigma^2(x)} dx - \left(\int \pi(x) dx\right)^2 =$$

$$= \int \frac{1}{\left(2\pi\right)^{d/2}} \sigma^d \exp\left\{-\frac{\sum_{i=1}^{d} x_i^2}{2}\left(2-\frac{1}{\sigma^2}\right)\right\} dx - 1 = \sigma^d \left(\frac{\sigma^2}{2\sigma^2-1}\right)^{d/2} - 1 = \underbrace{\left(\frac{\sigma^4}{2\sigma^2-1}\right)^{d/2}}_{>1} - 1$$

- It is easy to see that: $\sigma^4 > 2\sigma^2 - 1 \Leftrightarrow \left(\sigma^2 - 1\right)^2 > 0$. Therefore:

$$Var_{q_\sigma}\left(\frac{\pi(x)}{q_\sigma(x)}\right) \to \infty \ as \ d \to \infty$$

- The variance of the weights increases exponentially fast with dimensionality. This is despite the good choice of $q(x)$.

Summary: Importance Sampling

- While you can draw samples from practically any $q(x)$, wrong selection of $q(x)$ leads to poor estimates
- Difficult or impossible to select the proper $q(x)$ in high dimensions
- Importance sampling like all MC methods is embarrassingly parallel

Importance Sampling Vs. Rejection Sampling

- Given N samples from q, we estimate $\mathbb{E}_\pi(f(X))$ through importance sampling

$$\widehat{\mathbb{E}}_\pi^{IS}(f(X)) = \frac{\sum_{i=1}^N w^*(X^{(i)}) f(X^{(i)})}{\sum_{i=1}^N w^*(X^{(i)})}$$

or we only keep some of the samples through rejection and compute instead

$$\widehat{\mathbb{E}}_\pi^{RS}(f(X)) = \frac{1}{K}\sum_{k=1}^{K} f\left(X^{(i_k)}\right)$$

where K is a random variable (# of accepted samples).

- Which strategy performs the best?

Y. Chen, Another look at rejection sampling through importance sampling, Statistics & Probability Letters 72 (2005) 277–283

Rejection Sampling is a Special Case of Importance Sampling

- Define the (augmented) artificial target $\bar{\pi}(x, y)$ *on* $\mathcal{X} \times [0, 1]$

$$\bar{\pi}(x, y) = \begin{cases} \dfrac{Mq^*(x)}{\int \pi^*(t)dt} \text{ for } x \in \mathcal{X}, y \in \left[0, \dfrac{\pi^*(x)}{Mq^*(x)}\right] \\ 0, \text{otherwise} \end{cases}$$

then the marginal is:

$$\int \bar{\pi}(x, y)dy = \int_0^{\frac{\pi^*(x)}{Mq^*(x)}} \frac{Mq^*(x)}{\int \pi^*(t)dt} dy = \frac{Mq^*(x)}{\int \pi^*(t)dt} \frac{\pi^*(x)}{Mq^*(x)} = \pi(x)$$

- Now let us consider the proposal distribution

$$q(x, y) = q(x)\mathcal{U}_{[0.1]}(y) \text{ for } (x, y) \in \mathcal{X} \times [0,1]$$

Rejection Sampling is a Special Case of Importance Sampling

- Then rejection sampling is nothing but importance sampling on $\mathcal{X} \times [0, 1]$ where

$$w(x,y) = \frac{\bar{\pi}(x,y)}{q(x)\mathcal{U}_{[0,1]}(y)} = \frac{\frac{Mq^*(x)}{\int \pi^*(t)dt}}{\frac{q^*(x)}{\int q^*(t)dt}\mathcal{U}_{[0,1]}(y)} \Rightarrow$$

$$w(x,y) = \begin{cases} \frac{M\int q^*(x)dx}{\int \pi^*(x)dx} \; for \; Y^{(i)} \in \left[0, \frac{\pi^*(x)}{Mq^*(x)}\right] \\ 0, otherwise \end{cases}$$

- We have

$$\mathbb{E}^{RS}_{\hat{\pi}_N}(f(X)) = \frac{1}{K}\sum_{k=1}^{K} f\left(X^{(i_k)}\right) = \frac{\sum_{i=1}^{N} w(X^{(i)}, Y^{(i)}) f(X^{(i)})}{\sum_{i=1}^{N} w(X^{(i)}, Y^{(i)})}$$

- Thus rejection sampling performs importance sampling on an enlarged space.

Rejection Sampling is a Special Case of Importance Sampling

- The variance of the importance weights $w(X,Y)$ from rejection sampling is higher than $w(X)$ for standard importance sampling:

$$Var_q\left[w(X,Y)\right] \geq Var_q\left[w(X)\right]$$

- Indeed:

$$Var_{q(x)\mathcal{U}_{[0,1]}(y)}\left[w(X,Y)\right] = Var_{q(x)}\left[\underbrace{\mathbb{E}\left[w(X,Y)\,|\,X\right]}_{\int w(x,y)\underbrace{q(y|x)}_{\mathcal{U}_{[0,1]}(y)}dy}\right] + \mathbb{E}\left[Var\left[w(X,Y)\,|\,X\right]\right] =$$

$$= Var_{q(x)}\left[w(X)\right] + \mathbb{E}\left[Var\left[w(X,Y)\,|\,X\right]\right] \geq Var_{q(x)}\left[w(X)\right]$$

- For computing integrals, importance sampling should be the choice as rejection sampling is highly inefficient.

Importance Sampling for a DGM

- Ancestral Sampling with no Evidence: To sample from a distribution represented as a DGM, we can sample from $p(\mathbf{x})$ as follows:
  - first sample the root nodes, then sample their children, , etc.
  - This works because, in a DAG, we can always topologically order the nodes so that parents preceed children.
- Ancestal Sampling with Evidence: If all the variables are discrete, we can use the following simple procedure: perform ancestral sampling, but as soon as we sample a value that is inconsistent with an observed value, reject the whole sample and start again. This is very inefficient and not applicable to real-valued evidence.
- Likelihood Weighting: Use the following proposal: $q(\boldsymbol{x}) = \prod_{t \notin E} p\left(x_t | \boldsymbol{x}_{pa(t)}\right) \prod_{t \in E} \delta_{x_t^*}(x_t)$, E=set of observed nodes. Then the importance weights are:

$$\mathrm{w}(x) = \frac{p(x)}{q(x)} = \frac{\prod_{t \notin E} p\left(x_t | \boldsymbol{x}_{pa(t)}\right)}{\prod_{t \notin E} p\left(x_t | \boldsymbol{x}_{pa(t)}\right)} \prod_{t \in E} \frac{p\left(x_t | \boldsymbol{x}_{pa(t)}\right)}{1} = \prod_{t \in E} p\left(x_t | \boldsymbol{x}_{pa(t)}\right)$$

- Henrion, M. (1988). Propagation of uncertainty by logic sampling in Bayes' networks. In *UAI*, pp. 149– 164.
- Fung, R. and K. Chang (1989). Weighting and integrating evidence for stochastic simulation in Bayesian networks. In *UAI*.
- Shachter, R. D. and M. A. Peot (1989). Simulation approaches to general probabilistic inference on belief networks. In *UAI*, Volume 5.

Solving Ax=b with Importance Sampling

- Consider the system of equations $Ax = b,\ A \in \mathbb{R}^{n \times n}$
- Multiply this linear system with an invertable matrix G:

$$GAx = Gb,\ where: GA = I - B\ with \quad \rho(B) < 1$$

h

$spectral\ radius\ of\ B$

- Then the solution of the linear system is:

$$x = \sum_{k=0}^{\infty} B^k h$$

- Or in component form:

$$x_i = \sum_{k=0}^{\infty} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \dots \sum_{i_k=1}^{n} B_{ii_1} B_{i_1 i_2} \dots B_{i_{k-1} i_k} h_k$$

- G. E. Forsythe; Richard A. Leibler, Matrix Inversion by a Monte Carlo Method, *Mathematical Tables and Other Aids to Computation*, Vol. 4, No. 31. (Jul., 1950), pp. 127-129.
- J. H. Curtiss, Monte Carlo Methods for the Iteration of Linear Operators, *Journal of Mathematics and Physics*, Vol. 32 (1953) 209-232.
- John H. Halton, Sequential Monte Carlo techniques for the solution of linear systems, Journal of Scientific Computing, Vol. 9, Number 2 / June, (1994).

Solving Ax=b with Importance Sampling

$$x_i = \sum_{k=0}^{\infty} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \dots \sum_{i_k=1}^{n} B_{ii_1} B_{i_1 i_2} \dots B_{i_{k-1} i_k} h_k$$

- Introducing the following sequence of indices:

$$\gamma_k = (i_1, i_2, \dots, i_k), \; i_i \in \{1, 2, \dots, n\}$$

we can write the above equation as follows:

$$x_i = \sum_{\gamma_k} a_i(\gamma_k), \; where: a_i(\gamma_k) = \begin{cases} B_{ii_1} B_{i_1 i_2} \dots B_{i_{k-1} i_k} h_k \; if \; k > 0 \\ h_i \; if \; k = 0 \end{cases}$$

- We see that x_i is the average of a_i with respect to the uniform distribution of indices γ_k of any length k.

$$x_i \sim \mathbb{E}_{\pi}\left[a_i(\gamma_k)\right]$$

Not known normalization constant for π

Solving Ax=b with Importance Sampling

$$x_i \sim \mathbb{E}_\pi \left[a_i(\gamma_k) \right]$$

- We use an importance sampling approach:

$$x_i = \sum_{\gamma_k} a_i(\gamma_k) = \sum_{\gamma_k} \frac{a_i(\gamma_k)}{q(\gamma_k)} q(\gamma_k) = \mathbb{E}_q \left[\frac{a_i(\gamma_k)}{q(\gamma_k)} \right]$$

- We define the density q using "a random walk of k steps on indices":

$$q(\underset{(i_1, i_2, \ldots, i_k)}{\gamma_k}) = \underbrace{P_{ii_1} P_{i_1 i_2} \ldots P_{i_{k-1} i_k}}_{\text{transition probabilities}} \underbrace{P_{i_k}}_{\substack{\text{stopping probability} \\ \text{at index } i_k}}, \quad P_i = 1 - \sum_{j=1}^{n} P_{ij}$$

<1

- To obtain sequences of size k, we introduce a stopping probability at each state i.

Solving Ax=b with Importance Sampling

- Step 1: Draw *N* multi-indices from *q*

$$\gamma_k^{(j)} = \left(i_1^{(j)}, i_2^{(j)}, \ldots, i_k^{(j)} \right)$$

- Step 2: Compute

$$\hat{x}_i = \frac{1}{N} \sum_{j=1}^{N} \frac{a_i\left(\gamma_k^{(j)}\right)}{q\left(\gamma_k^{(j)}\right)}$$

where:

$$\frac{a_i(\gamma_k)}{q(\gamma_k)} = \begin{cases} \dfrac{B_{ii_1} B_{i_1 i_2} \ldots B_{i_{k-1} i_k} h_{i_k}}{P_{ii_1} P_{i_1 i_2} \ldots P_{i_{k-1} i_k} P_{i_k}} & if \;\; k > 0 \\ \dfrac{h_i}{p_i} \; if \;\; k = 0 \end{cases}$$

Solving Ax=b with Importance Sampling

- Consider the example:

$$\underbrace{\begin{bmatrix} 1.1 & -0.5 \\ -0.5 & 1.1 \end{bmatrix}}_{A} x = \underbrace{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}_{b}$$

- Let

$$A = I - B \Rightarrow B = \begin{bmatrix} -0.1 & 0.5 \\ 0.5 & -0.1 \end{bmatrix} \Rightarrow x = \sum_k B^k b$$

- The analytical solution is:

$$x = \begin{bmatrix} 1.67 \\ 1.67 \end{bmatrix}$$

Solving Ax=b with Importance Sampling

- Transition kernel

$$\begin{array}{c|cc|c} & 1 & 2 & \text{stop} \\ 1 & 1/3 & 1/3 & 1/3 \\ 2 & 1/3 & 1/3 & 1/3 \end{array}$$

- To estimate $x(1)$, we perform the algorithm in this way
  - step 1
    generate "a Markov Chain" from the transition kernel and starting from index 1, e.g. a chain such as

$$1 \xrightarrow[\Pr(1,i_1)]{B(1,i_1)} i_1 \xrightarrow[\Pr(i_1,i_2)]{B(i_1,i_2)} i_2 \xrightarrow[\Pr(i_2,i_3)]{B(i_2,i_3)} \cdots \xrightarrow[\Pr(i_{k-1},i_k)]{B(i_{k-1},i_k)} i_k \xrightarrow[\Pr(i_k,\text{stop})]{b(i_k)} \text{Stop}$$

    Then we get

$$x^{(n)} = \frac{B(1,i_1)B(i_1,i_2)\cdots B(i_{k-1},i_k)b(i_k)}{\Pr(1,i_1)\Pr(i_1,i_2)\cdots\Pr(i_{k-1},i_k)\Pr(i_k,\text{stop})}$$

  - step 2: repeat step 1 and average on $x^{(n)}$

- Note that in the implementation, we don't need to define explicitly the length *k* of the chains. Since we have specified a stopping probability for each state (i=1,…n), the chains generated will automatically be with different *k*.

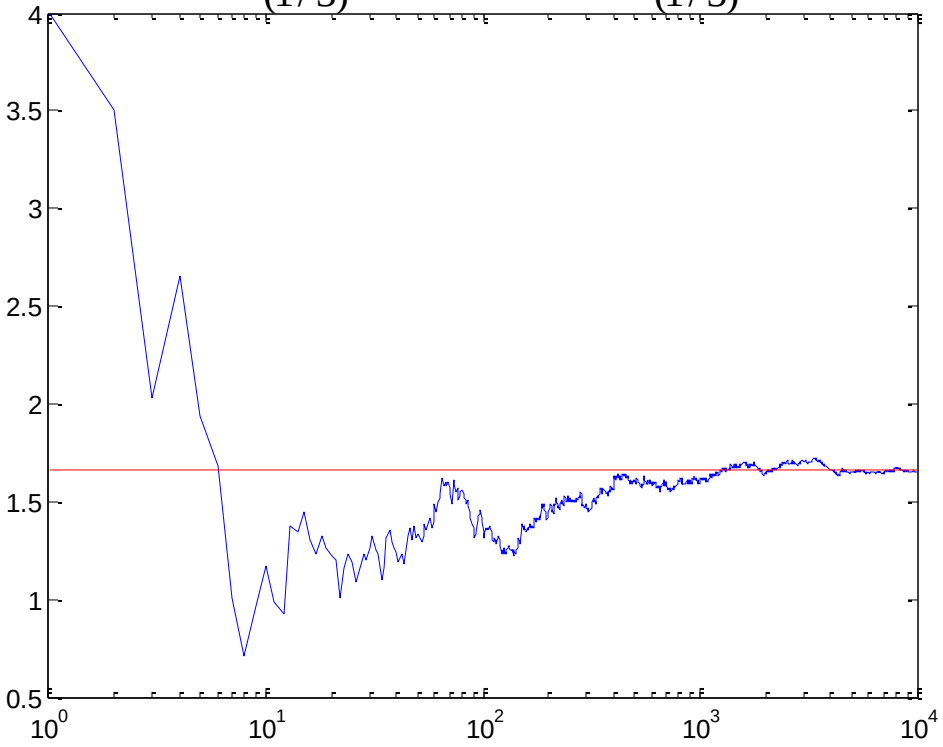
Solving Ax=b with Importance Sampling

- For example, for a Markov chain such that $1 \to 2 \to 2 \to 1 \to stop$

The estimated $x = \frac{B(1,2)B(2,2)B(2,1)b(1)}{(1/3)^4} = \frac{0.5 \times (-0.1) \times 0.5 \times 1}{(1/3)^4} = -2.025$

A MatLab implementation is provided here.

Another MatLab implementation for solving for both or a single variable is also available.



- The cost of this IS solver of linear equations is $\mathcal{O}(nsN)$, N=# of samples, s=average length of MCMC walks

Solving Ax=b using MC

- In this slightly different MatLab implementation, we choose A to be a symmetric positive definite matrix such that

$$eig(I - A) < 1$$

  - To do this you can set

$$A = I - \ Q' diag\ \ \varepsilon \times rand\ \ n,1\ \ \times Q$$

    Here Q is any orthogonal matrix, ϵ is a value less than 1 and n is the size of the system of interest.
  - It can be generated through a QR decomposition of any random matrix
  - N is chosen to be

$$Q\ \ R\ = qr\ \ rand\ \ n$$

Solving Ax=b using MC

- n is chosen to be 6
- ϵ is chosen to be 0.4

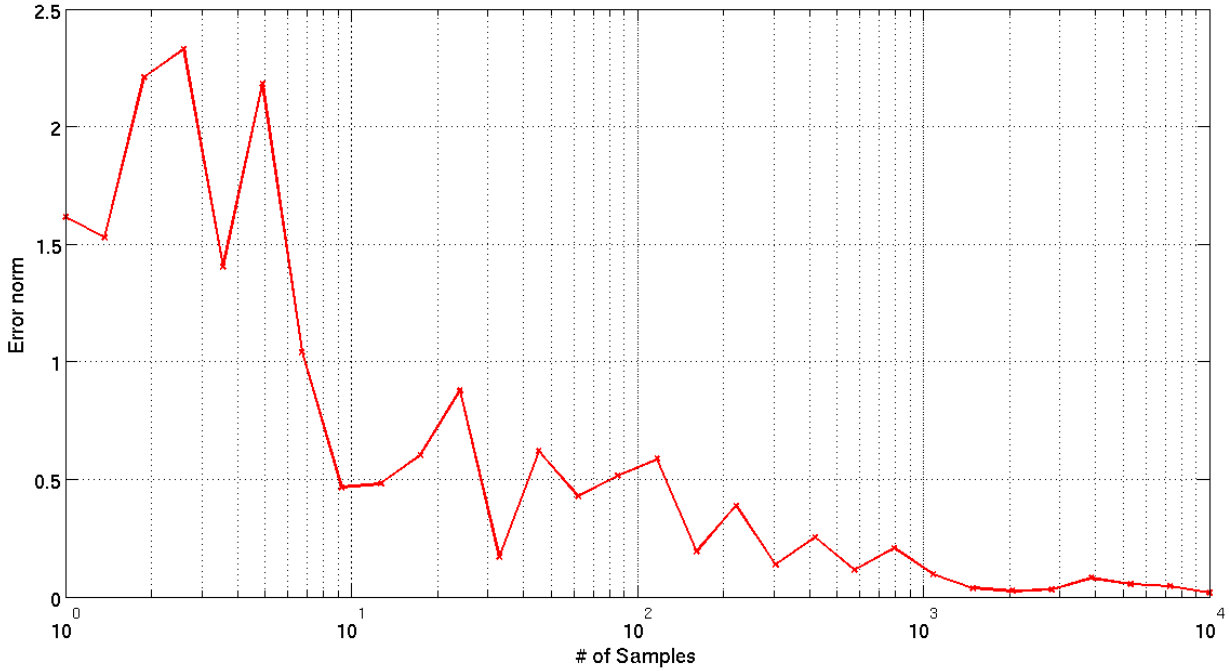
- One can solve for specific indices of the solution in x.
  - Indices chosen here as $index = [1,5]$
- Varying number of samples. We estimate the solution using the absorbing Markov Chain discussed earlier and in this paper.
- The transition probabilities and absorption probabilities are assumed to be identical and uniformly distributed to be $\frac{1}{n+1}$.

Solving Ax=b using MC

- The result is compared with the deterministic solution for the varying sample sizes

$$Error = \left|\left|x(index)^{MC} - x(index)^{Deterministic}\right|\right|_2$$

- A typical plot of this is shown below

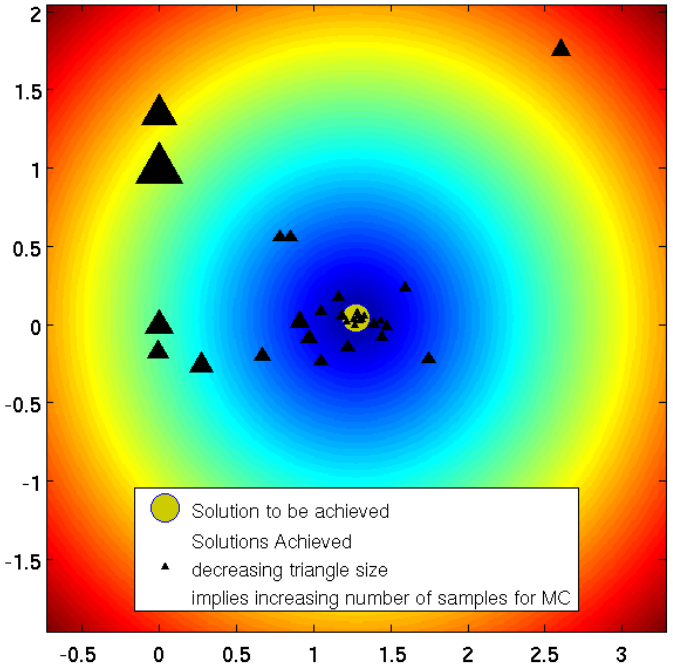


MatLab implementation

Solving Ax=b using MC

- In the plot below the progress of the solution towards the deterministic solution is shown as one increases the number of samples.
- On the horizontal axis is x(1) and vertical axis is x(5). These are the solutions we are interested in.
- This is plotted on a contour of error. The deterministic solution is circled in green.

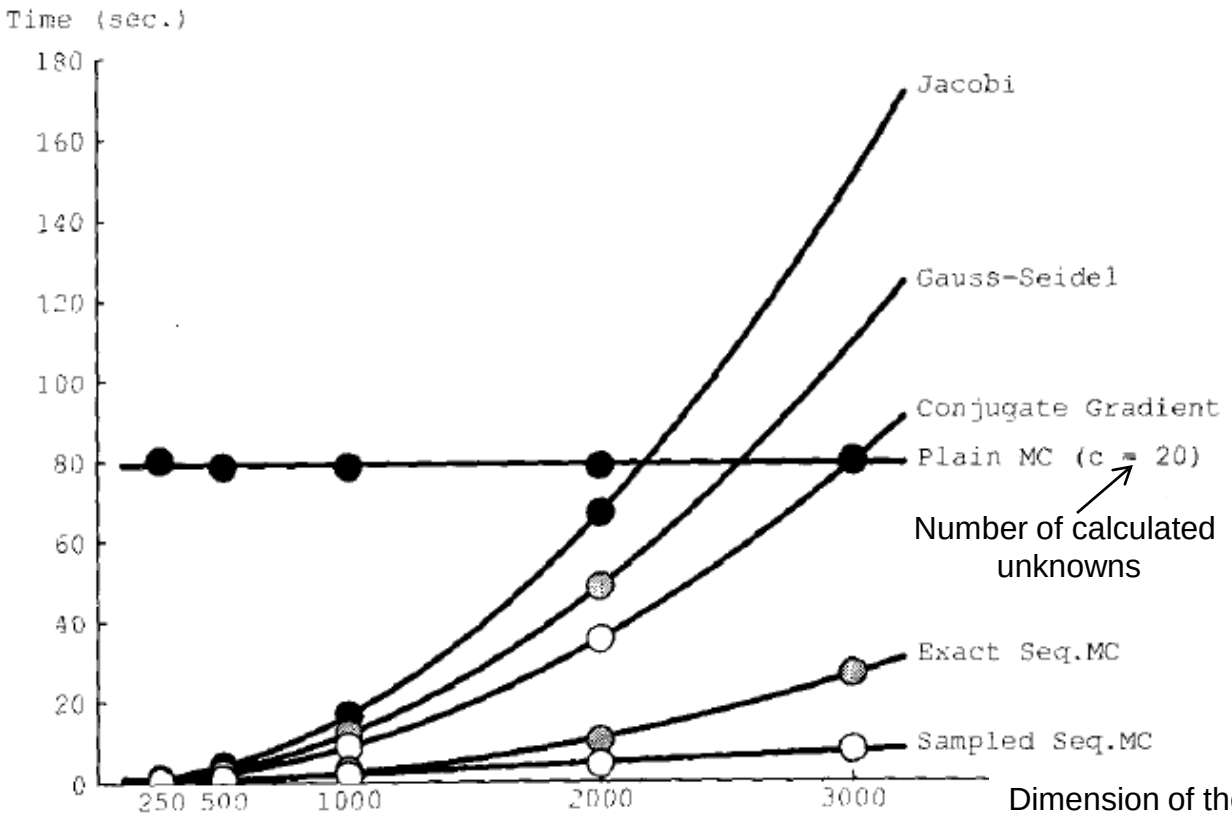
- The MC solutions are indicated as triangles. The larger the triangle the lesser the number of samples involved in the MC. As it becomes smaller more and more samples were taken.
- One can find a lot of small triangles closer to the green circle, indicating convergence as number of samples increases.
- The background is a contour of error. Blue is zero error and red is a large error.



MatLab implementation

Performance of MC in Solving Linear Systems

- A comparison is given below of the MC solver versus classical methods.



- **Direct Methods** (Gauss elimination, LU, Cholesky): $\mathcal{O}(n^3)$
- **Iterative Methods** (Jacobi, Gauss-Seidel) $\mathcal{O}(n^2 s)$, s=number of iterations
- **Monte Carlo Importance Sampling with n unknowns:** $\mathcal{O}$ (n s N) for uniform transition kernel or $\mathcal{O}$ (n^2 s N) otherwise. Here, s is the average length of walks and N the number of samples.
- **Monte Carlo Importance Sampling with m<<n unknowns:** If we only want to compute m out of the n components of the vector ***x***, then the number of operations needed drops down to $\mathcal{O}$ (m s N) for uniform transition kernel or $\mathcal{O}$ (m n s N) otherwise. MC then becomes highly favorable.

- John H. Halton, Sequential Monte Carlo techniques for the solution of linear systems, Journal of Scientific Computing, Vol. 9, Number 2 / June, (1994).

Sequential MC for Linear/Nonlinear Systems

- Nonlinear algebraic equations that involve iterative solution of linear systems can also be addressed using Monte Carlo Methods.
- Any fixed-point iteration problem of the form below (Picard iterations) is amenable to a similar approach, $\mathcal{L}$ a contraction mapping:

$$\boldsymbol{x}_{n+1} = \mathcal{L}\left(\boldsymbol{x}_n\right)$$

- Markov Chain Monte Carlo Solution of Integral Fredholm Equations is a typical example (see Doucet et al. 2010):

$$f\left(x\right) = g\left(x\right) + \int_D K(x,y)\, f\left(y\right) dy =$$

$$= g\left(x\right) + \sum_{k=1}^{\infty} \int_{D^k} \prod_{j=1}^{k} K(x_{j-1}, x_j)\, g(x_k)\, dx_{1:k}$$

- John H. Halton, Sequential Monte Carlo techniques for solving non-linear systems, Monte Carlo Methods and Applications, Volume 12, Number 2 / April (2006).
- A. Doucet, A. Johansen, V Tadic, On Solving Integral Equations using Markov Chain Monte Carlo, Applied Math and Computation, Vol. 216 (2010) 2869–2880.

Cauchy Example

- Consider i.i.d. sample $\mathcal{D}_n = (x_1, \ldots, x_n)$ from $\mathcal{C}(\theta, 1)$ and let the prior on θ be a flat prior. We use a normal importance function from the $\mathcal{N}(\mu, \sigma^2=1)$ to produce a sample $\theta_1, \ldots, \theta_M$ that approximates the Bayes estimator of θ by

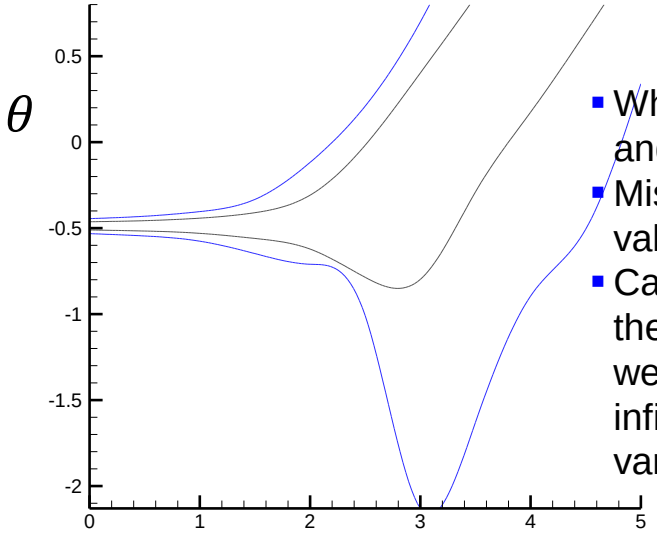
$$\hat{\delta}^{\pi}(\mathcal{D}_n) = \int \theta\pi(\theta|\mathcal{D}_n)d\theta = \frac{\sum_{t=1}^{M} \theta_t e^{(\theta_t-\mu)^2/2} \prod_{i=1}^{n}[1+(x_i-\theta_t)^2]^{-1}}{\sum_{t=1}^{M} e^{(\theta_t-\mu)^2/2} \prod_{i=1}^{n}[1+(x_i-\theta_t)^2]^{-1}}$$

- This is a very poor estimate and degrades considerably when μ increases:

Representation of the whole range and of the 90% range of variation of the importance sampling approximation to the Bayes estimate for

- *n = 10* observations from $\mathcal{C}(0, 1)$ distribution and
- *M = 1000* simulations of θ from $\mathcal{N}(\mu, 1)$ as a function of μ.

The range is computed using *1000* replications of the importance sampling estimates.



- Whole range and 90% range
- Miss the true valie for high μ
- Can show that the importance weights have infinite variance

For a MatLab implementation, see here

Computing an Unbounded Integral

- Let $x \sim \mathcal{T}(\upsilon, \theta, \sigma^2)$ be distributed according to the Student's $\mathcal{T}$ distribution

$$\pi(x) = \frac{\Gamma(\frac{\upsilon+1}{2})}{\sigma\sqrt{\upsilon\pi}\Gamma(\upsilon/2)}\left(1+\frac{(x-\theta)^2}{\upsilon\sigma^2}\right)^{-(\upsilon+1)/2}$$

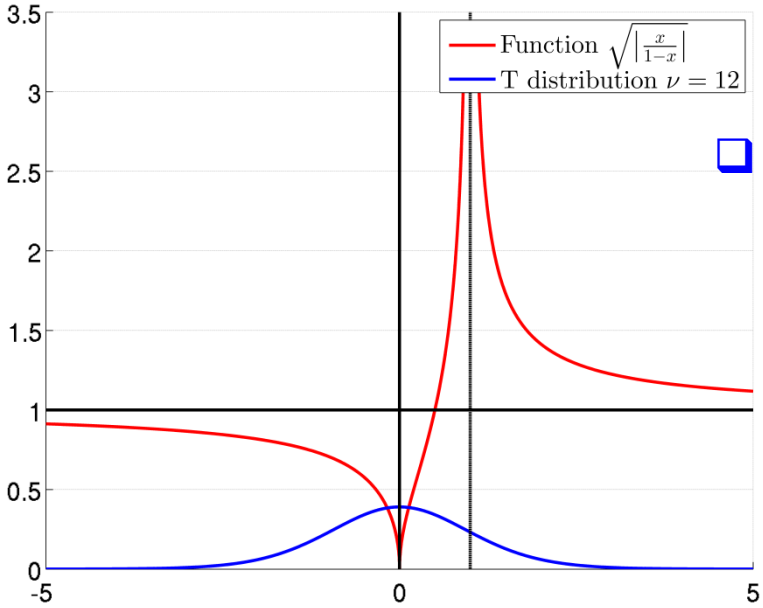
- Without loss of generality, take $\theta = 0, \sigma = 1$ to consider the standard distribution. We want to find the expectation of the function

$$f(x) = \sqrt{\left|\frac{x}{1-x}\right|}$$

with respect to the above distribution. The needed integral is over the entire real line. The function and the distribution are shown below.

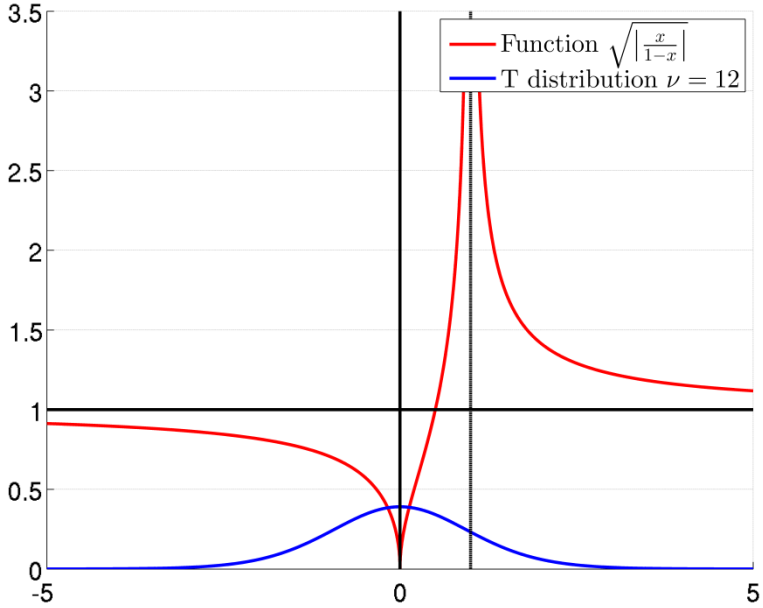
J-M Marin and C. P. Roberts, Bayesian Core (Chapter 2)

MatLab implementation can be found here



- You can sample from $\pi(x)$ by composition $\mathcal{N}(0,1)/\mathcal{Gamma}(\nu/2, \nu/2)$ using Monte Carlo.

Computing an Unbounded Integral



Matlab Code Snippet for Numerical Integration

```
nu=12;
g=@(y) f(y).*tpdf(y,nu);
value=integral(g,-inf,1,'Abstol',1e-
16)+integral(g,1,inf,'Abstol',1e-16);
```

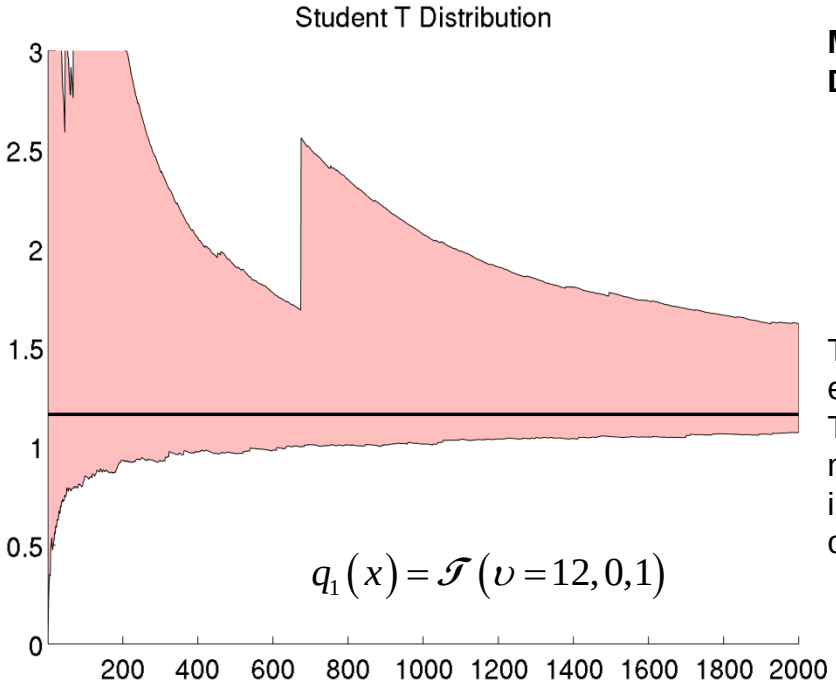
Result: 1.160058221169548

$$\mathcal{J} = \int f(x)\pi(x)dx = \int \sqrt{\left|\frac{x}{1-x}\right|}\pi(x)dx$$

$$\pi(x) = \mathcal{T}(12,0,1)$$

- The function has a singularity at $x = 1$. But the expectation does exist and one can use numerical integration to get the value.
- Matlab code snippet to do this is shown above and the estimate for the case of ν=12 is also given.
- Let us now do the same by Monte Carlo Methods.
  - Draw Samples from the Student's $\mathcal{T}$ distribution and evaluate the function.
  - The average value of the function obtained is the estimate of the integral.
  - Repeat the Monte Carlo Runs to get an indication of the variance of the result.
  - Plot the maximum estimate as a function of the number of samples drawn in the Monte Carlo run

Computing an Unbounded Integral



Monte Carlo Estimate using Samples from a Student's-$\mathcal{T}$ Distribution:

$$y_i = \frac{1}{N}\sum_{j=1}^{j=i} f\left(x_j\right)$$

$$x_j \sim \mathcal{T}\left(\upsilon = 12, 0, 1\right) \forall i = 1, \ldots, 2000$$

This process was repeated for 500 times thus getting 500 estimates at each *i*.
This graph shades the region between the maximum and the minimum estimate value of the integration for each *i*. This is indicative of the variance of the integrated value. The actual value of the integration is shown by the dark black line in the middle.

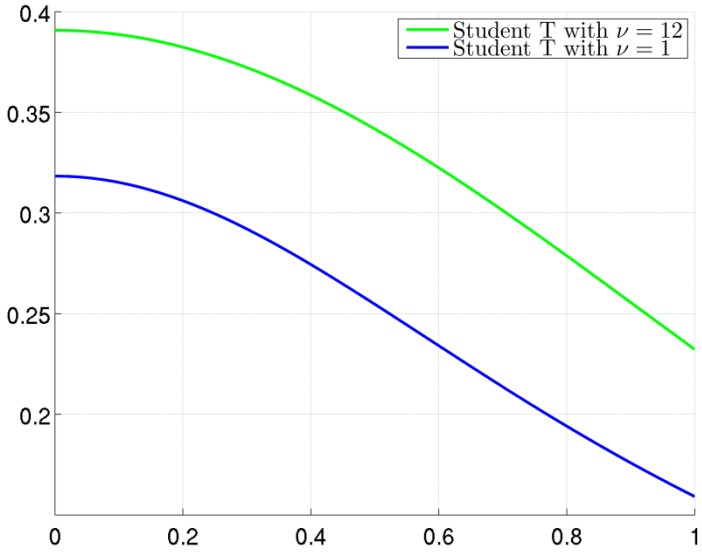
- The result of this process is shown above.
- Note the huge variance in the estimate:
  - Though the expectation of the function exists, its variance does not

$$\int_{-\infty}^{+\infty} f(x)^2 \mathcal{T}\left(x; \upsilon = 12\right) dx \to \infty$$

  - Recall that the variance of the Monte Carlo Estimate converges to

$$\frac{Var_{\mathcal{T}}(f)}{N}$$

Computing an Unbounded Integral



$$\int \left(\frac{x}{1-x}\right)\left(\frac{1}{1+x^2}\right)dx$$

$$= \frac{1}{4}\left(-2\tan^{-1}(x) - 2\log(1-x)\right.$$

$$\int_{-\infty}^{+\infty} f(x)^2 \mathcal{T}\left(x;\upsilon=12\right)dx \to \infty$$

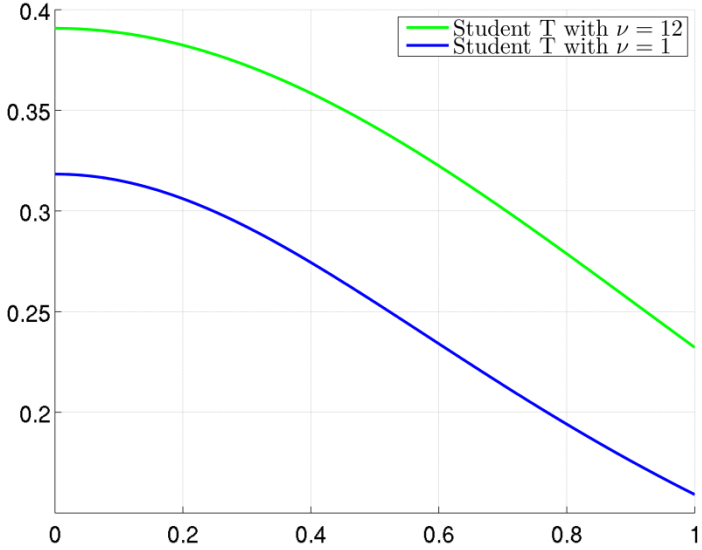
- To see this let us focus on the interval $(0,1)$
- The $\mathcal{T}$-distribution's pdf with $\nu = 12$ degrees of freedom is greater numerically than the $\mathcal{T}$-distribution's with $\nu = 1$ degree of freedom in the interval.
- Note: Student $\mathcal{T}$ distribution with 1 degree of freedom is a Cauchy distribution
- This is shown in the adjacent figure.
- So we can write

$$\frac{\Gamma(6.5)}{\sqrt{12\pi}\Gamma(6)}\int_0^1 \frac{x}{1-x}\left(1+\frac{x^2}{12}\right)^{-\frac{13}{2}}dx$$

$$> \frac{\Gamma(1)}{\sqrt{\pi}\Gamma(0.5)}\int_0^1 \frac{x}{1-x}(1+x^2)^{-1}dx$$

- Moreover the absolutes values can be safely removed in the interval.
- The analytical integration involving the Cauchy distribution is shown on the left.

Computing an Unbounded Integral



$$\int \left(\frac{x}{1-x}\right)\left(\frac{1}{1+x^2}\right) dx$$

$$= \frac{1}{4}\left(-2\tan^{-1}(x) - 2\log(1-x)\right)$$

$$\frac{\Gamma(6.5)}{\sqrt{12\pi}\Gamma(6)} \int_0^1 \frac{x}{1-x}\left(1+\frac{x^2}{12}\right)^{-\frac{13}{2}} dx$$

$$> \frac{\Gamma(1)}{\sqrt{\pi}\Gamma(0.5)} \int_0^1 \frac{x}{1-x}(1+x^2)^{-1} dx$$

- Substituting the limits of the integration i.e., 1 and 0, we can see that the upper limit shoots to ∞ and the lower limit to 0.
- So the integral

$$\frac{\Gamma(6.5)}{\sqrt{12\pi}\Gamma(6)} \int_0^1 \frac{x}{1-x}\left(1+\frac{x^2}{12}\right)^{-\frac{13}{2}} dx \to \infty$$

- This proves that the variance is ∞
- Note: This result is true for any $\nu \geq 1$ as can be shown by similar arguments
- This explains why standard Monte Carlo does not converge to the desired result.

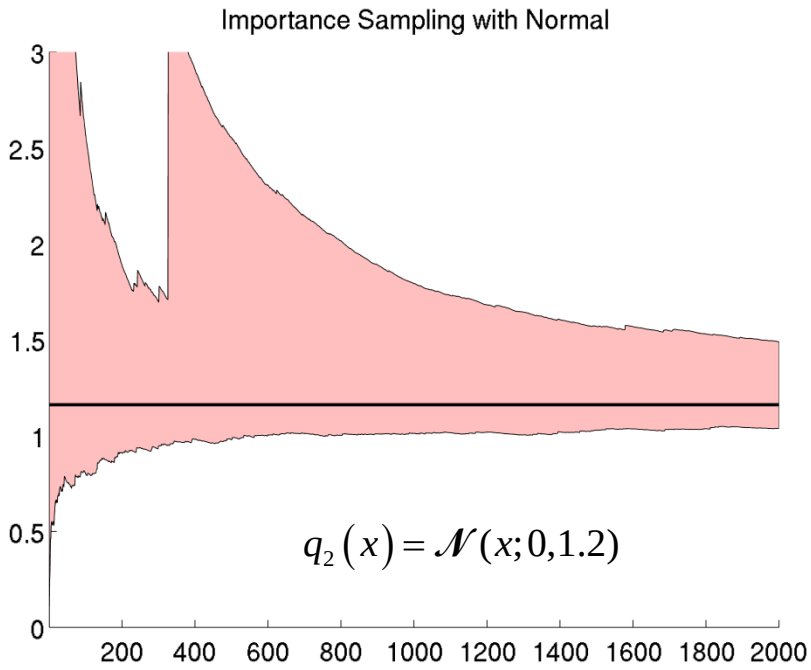
Computing an Unbounded Integral

- Let us try importance sampling with:
  - Proposal distribution a Gaussian (light tails)

$$\mathcal{N}(x;0,\frac{\upsilon}{\upsilon-2}),\ \upsilon>2$$

This coincides with the mean and variance of the Student's $\mathcal{T}$ distribution.

  - In this case $\upsilon = 12$. So our proposal is $\mathcal{N}(0,1.2)$.
- The result is shown below:



Monte Carlo Estimate using Importance Sampling taking samples from a Normal Distribution:

$$y_i = \frac{1}{N}\sum_{j=1}^{j=i}\frac{\mathcal{T}(x_j;\upsilon=12)}{\mathcal{N}(x_j;0,1.2)}f(x_j)$$

$$x_j \sim \mathcal{N}(0,1.2)\ \forall i = 1,\ldots,2000$$

In Importance sampling we have to take the weighted average of the function evaluated on samples drawn from the proposed distribution. In this case $\mathcal{N}(0,1.2)$. Weights are the ratio of the evaluation of the two probability distributions as given in the above formula.

This process was repeated for 500 times thus getting 500 estimates at each *i*.
This graph shades the region between the maximum and the minimum estimate value of the integration for each *i*. This is indicative of the variance of the integrated value. The actual value of the integration is shown by the dark black line in the middle.

Computing an Unbounded Integral

- We are seeing a similar behavior to what we observed in the first attempt.
  - The reason is the same, the variance is infinite
- The variance is given by

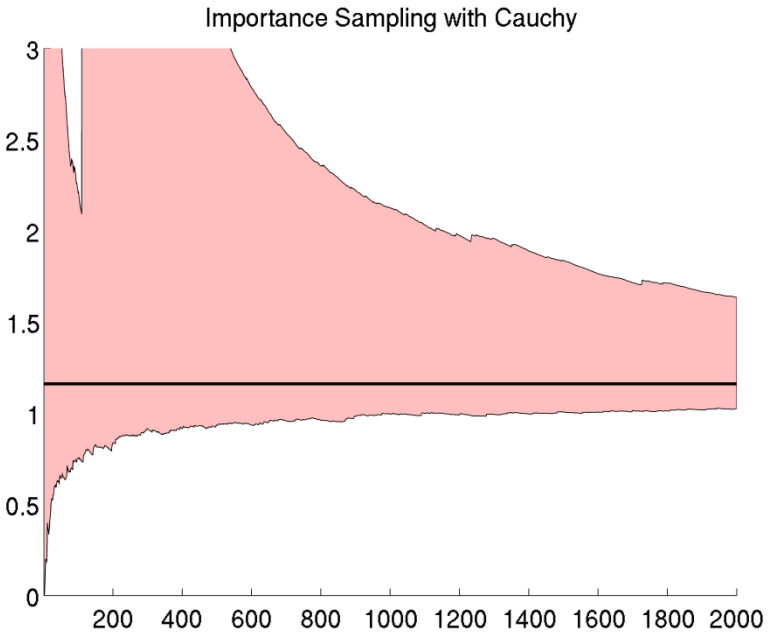
$$Var = \frac{\mathbb{E}_{\mathcal{I}}\left(\frac{\mathcal{I}(x)}{\mathcal{N}(x)} f^2(x)\right) - \mathbb{E}^2_{\mathcal{I}}\left(f(x)\right)}{N}$$

$$\mathbb{E}_{\mathcal{I}}\left(\frac{\mathcal{I}(x)}{\mathcal{N}(x)} f^2(x)\right) \propto \int_{-\infty}^{\infty} \frac{\left|\frac{x}{1-x}\right|\left(1+\frac{x^2}{12}\right)^{-6.5}}{\exp\left(-\frac{x^2}{2.4}\right)}\left(1+\frac{x^2}{12}\right)^{-6.5} dx = \int_{-\infty}^{\infty} \frac{\left|\frac{x}{1-x}\right|}{\left(1+\frac{x^2}{12}\right)^{13}} \exp\left(\frac{x^2}{2.4}\right) dx$$

- The integrand here has an exponential term that grows faster than the other terms and hence the integral is unbounded.
- This implies that the variance of the estimator also is unbounded.
- Hence, we observe the lack of convergence of the importance sampling w.r.t to the Normal distribution.

Computing an Unbounded Integral

- Similar behavior will be seen when one attempts to do it with a Cauchy distribution as shown next.



$$q_3(x) = \mathcal{C}(x;0,1)$$

$$= \frac{1}{\pi}\frac{1}{1+x^2} = \frac{\Gamma(1)}{\sqrt{\upsilon\pi}\underbrace{\Gamma(1/2)}_{\sqrt{\pi}}}\left(1+\frac{x^2}{\upsilon}\right)^{-(\upsilon+1)/2}\Bigg|_{\upsilon=1}^{1} = \mathcal{T}(1,0,1)$$

Monte Carlo Estimate using Importance Sampling taking samples from a Cauchy Distribution (chosen for its heavy tails):

$$y_i = \frac{1}{N}\sum_{j=1}^{j=i}\frac{\mathcal{T}(x_j;\upsilon=12)}{\mathcal{C}(x_j)}f(x_j)$$

$$x_j \sim \mathcal{C}(0,1)\ \forall i = 1,\ldots,2000$$

In importance sampling we have to take the weighted average of the function evaluated on samples drawn from the proposed distribution. In this case $\mathcal{C}$. Weights are the ratio of the evaluation of the two probability distributions as given in the above formula.

This process was repeated for 500 times thus getting 500 estimates at each *i*.

This graph shades the region between the maximum and the minimum estimate value of the integration for each *i*. This is indicative of the variance of the integrated value. The actual value of the integration is shown by the dark black line in the middle.

Computing an Unbounded Integral

- To obtain convergence, we need to find a distribution where the integrals don't blow up.
- A suggested distribution is the following Gamma:

$$|x-1| \propto \mathcal{G}(\alpha,1), \alpha = 0.5$$

$$q(x) = \frac{1}{2\Gamma(\alpha)}|x-1|^{\alpha-1}\exp\left(-|x-1|\right), x \in [-\infty,\infty]$$

  - It is symmetric about 1

- The variance for this distribution is given by

$$Var = \frac{\mathbb{E}_{\mathcal{I}}\left(\frac{\mathcal{I}(x)}{q(x)}f^2(x)\right) - \mathbb{E}^2_{\mathcal{I}}\left(f(x)\right)}{N}$$

$$\mathbb{E}_{\mathcal{I}}\left(\frac{\mathcal{I}(x)}{q(x)}f^2(x)\right) \propto \int_{-\infty}^{\infty} \frac{\left|\frac{x}{1-x}\right|\left(1+\frac{x^2}{12}\right)^{-6.5}}{|x-1|^{\alpha-1}\exp\left(-|x-1|\right)}\left(1+\frac{x^2}{12}\right)^{-6.5} dx = \int_{-\infty}^{\infty} \frac{|x||x-1|^{-\alpha}}{\left(1+\frac{x^2}{12}\right)^{13}}\exp\left(|x-1|\right)dx$$

- The underlying function is now integrable when $\alpha < 1$

Computing an Unbounded Integral

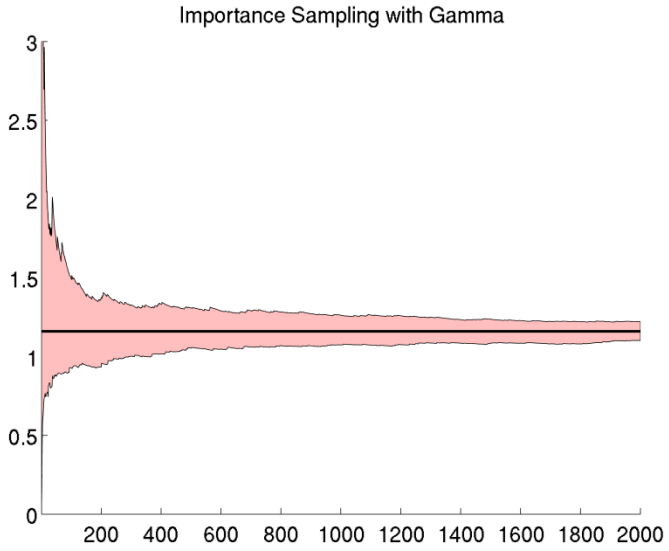
- To sample from $|x-1| \propto G(\alpha,1)$, use the following method:
  - Draw a sample z_i from the corresponding $\mathcal{G}(\alpha,1)$ distribution. z_i is positive.
  - Draw a uniform random number $u_i \in \mathcal{U}[0,1]$ If $u_i \leq 0.5$, then choose $x_i=1+z_i$ else choose $x_i=1-z_i$.
  - The proof is as follows:

$$q(x_i)=\begin{cases} P(u_i \leq 0.5)\,\mathcal{G}(x_i-1;\alpha,1) & x_i \geq 1 \\ P(u_i > 0.5)\,\mathcal{G}(1-x_i;\alpha,1) & x_i < 1 \end{cases}$$

$$q(x_i)=\begin{cases} \dfrac{1}{2\Gamma(\alpha)}(x_i-1)^{\alpha-1}\exp\left(-(x_i-1)\right) & x_i \geq 1 \\ \dfrac{1}{2\Gamma(\alpha)}(1-x_i)^{\alpha-1}\exp\left(-(1-x_i)\right) & x_i < 1 \end{cases}$$

$$q(x_i)=\frac{1}{2\Gamma(\alpha)}|x_i-1|^{\alpha-1}\exp\left(-|x_i-1|\right),\ x_i \in [-\infty,\infty]$$

Computing an Unbounded Integral



$$q(x) = \frac{1}{2\Gamma(\alpha)}|x-1|^{\alpha-1}\exp\left(-|x-1|\right)$$

Monte Carlo Estimate using Importance Sampling taking samples from a Gamma distribution symmetric around 1:

$$y_i = \frac{1}{N}\sum_{j=1}^{j=i}\frac{\mathcal{T}\left(x_j;\upsilon=12\right)}{q\left(x_j\right)}f\left(x_j\right)$$

$$x_j \sim q\ \forall i = 1,\ldots,2000$$

In Importance sampling we have to take the weighted average of the function evaluated on samples drawn from the proposed distribution. In this case p. Weights are the ratio of the evaluation of the two probability distributions as given in the above formula. This process was repeated for 500 times thus getting 500 estimates at each *i*.

This graph shades the region between the maximum and the minimum estimate value of the integration for each *i*. This is indicative of the variance of the integrated value. The actual value of the integration is shown by the dark black line in the middle.

- This results shows a clean convergence as predicted.
- Matlab Code for these simulations can be downloaded from this link

Generating Student's-$\mathcal{T}$ Samples

- We consider again the following student's-$\mathcal{T}$ distribution $\mathcal{T}(\upsilon, \theta, \sigma^2)$

$$\pi(x) = \frac{\Gamma(\frac{\upsilon+1}{2})}{\sigma\sqrt{\upsilon\pi}\Gamma(\upsilon/2)}\left(1+\frac{(x-\theta)^2}{\upsilon\sigma^2}\right)^{-(\upsilon+1)/2}$$

with the following parameters: $\theta = 0, \sigma = 1, \upsilon = 12.$

- We consider computing

$$\mu = \mathbb{E}_\pi\left(f(X)\right) = \int f(x)\pi(x)dx$$

with the following choices:

$$f(x) = \left(\frac{\sin x}{x}\right)^5 \mathbb{I}_{(2.1,\infty)}(x),\ f(x) = \sqrt{\left|\frac{x}{1-x}\right|},\ f(x) = \frac{x^5}{1+(x-3)^2}\mathbb{I}_{[0,\infty)}(x)$$

- We will study the performance of importance sampling with the following choices of importance distribution:

$$\mathcal{T}(\nu^*, 0, 1),\ \nu^* = 7 < \upsilon = 12,\ \mathcal{N}\left(0, \frac{\upsilon}{\upsilon-2}\right),\ \mathcal{C}(0,1)\ where\ \mathcal{C}(\alpha,\beta) = \frac{1}{\pi\beta\left(1+\left(\frac{x-\alpha}{\beta}\right)^2\right)}\mathbb{I}_{\mathbb{R}}(x), \alpha \in \mathbb{R}, \beta > 0$$

Generating Student-t Samples

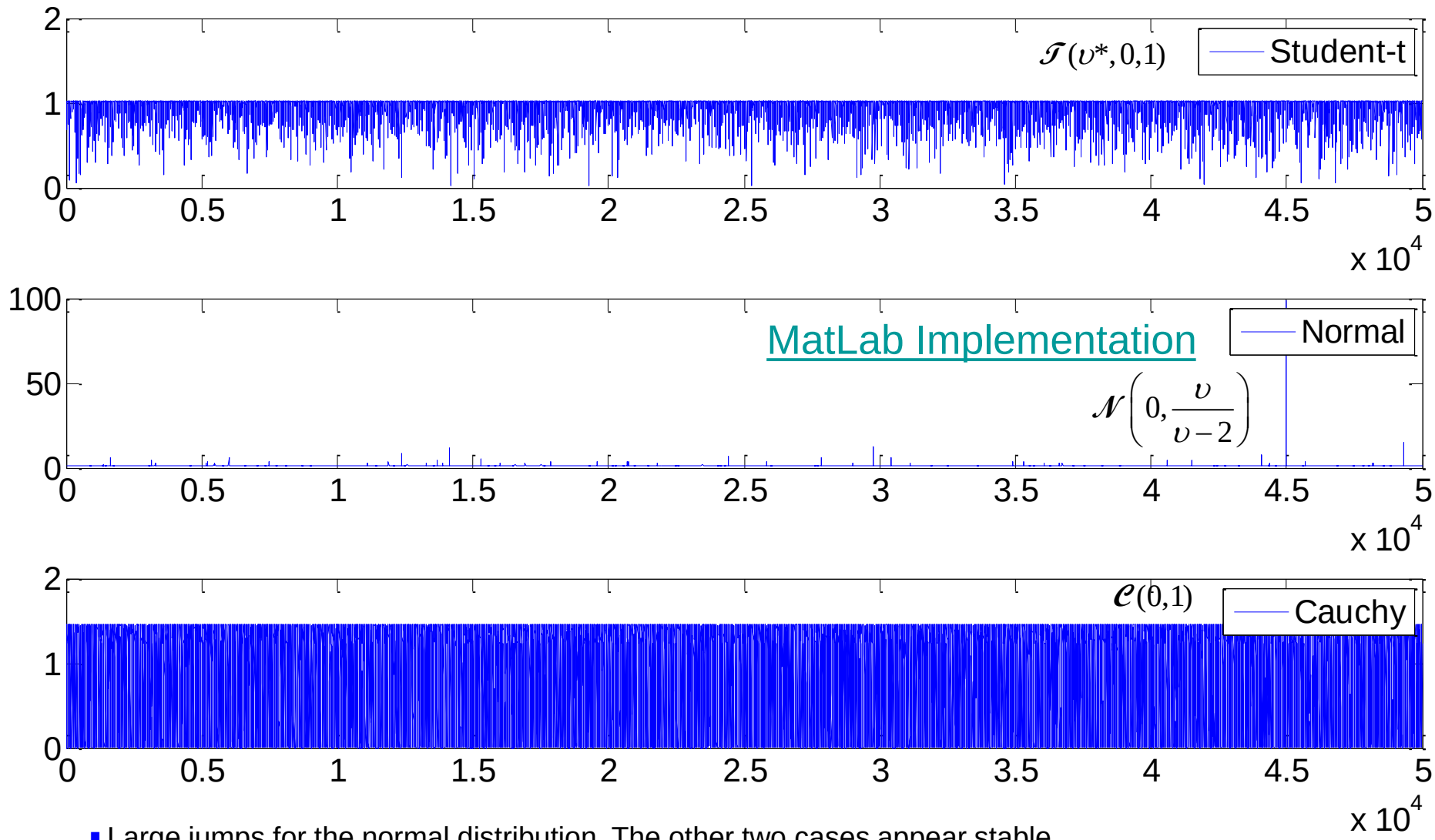
- The inverse CDF can be used to sample from the Cauchy distribution. Note that:

$$\mathcal{C}(\alpha,\beta)=\frac{1}{\pi\beta\left(1+\left(\frac{x-\alpha}{\beta}\right)^2\right)}\mathbb{I}_{\mathbb{R}}(x),\alpha\in\mathbb{R},\beta>0$$

$$\mathrm{F(x)}=\int_{-\infty}^{x}\frac{1}{\pi\beta\left(1+\left(\frac{u-\alpha}{\beta}\right)^2\right)}du=\frac{1}{2}+\frac{1}{\pi}\arctan\frac{x-\alpha}{\beta}\mathbb{I}_{\mathbb{R}}(x)$$

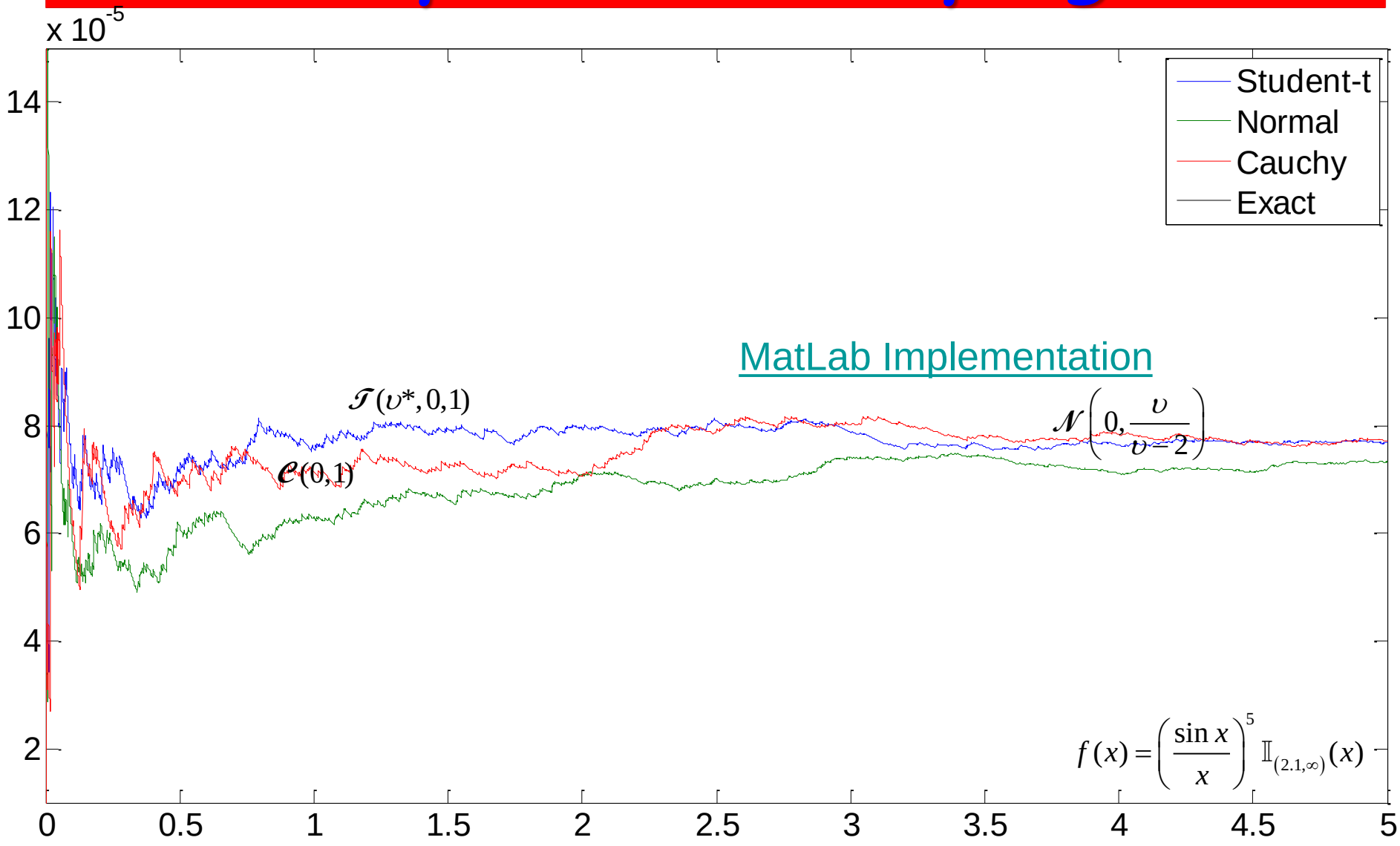
- Thus one can sample $U\sim\mathcal{U}[0,1]$, and then set $X=F^{-1}(U)$ to obtain $X\sim\mathcal{C}(\alpha,\beta)$

Importance Sampling Weights



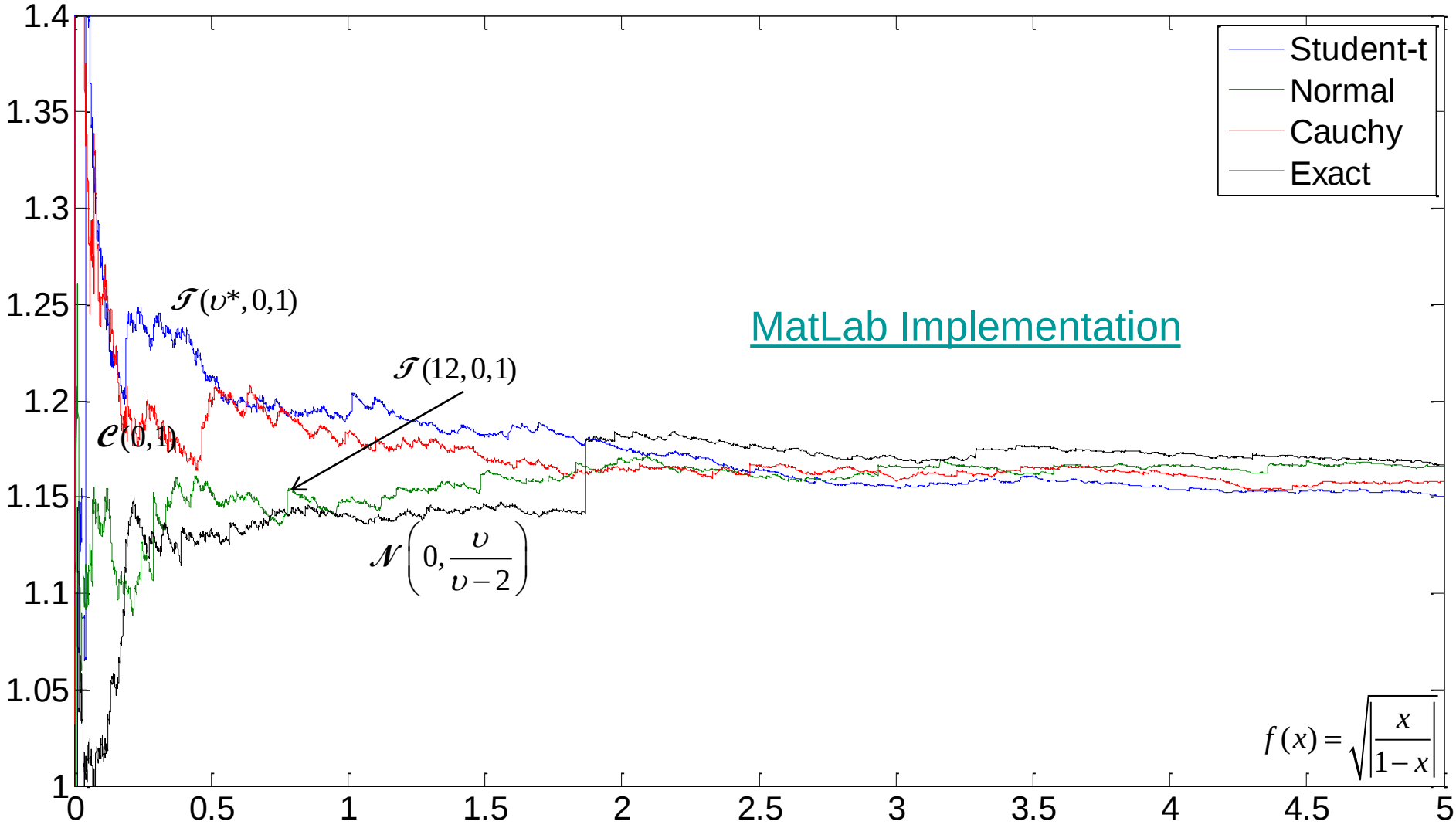
- Large jumps for the normal distribution. The other two cases appear stable.
- In the following Figures, the exact simulator refers to sampling from the $\mathcal{T}(12, 0, 1)$

Importance Sampling



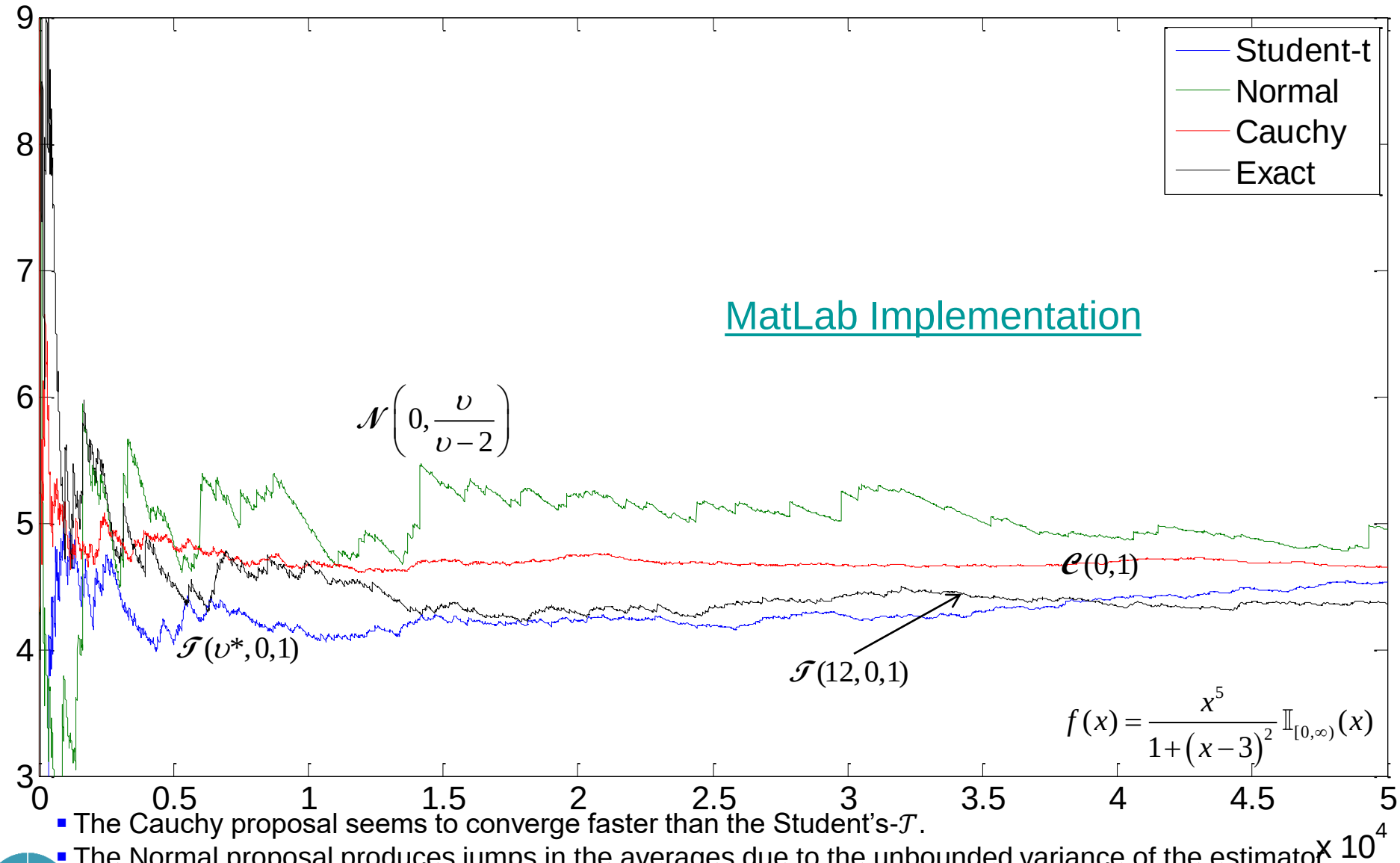
- The normal proposal appears to perform well for this function in comparison to Student's-$\mathcal{T}$ & Cauchy

Importance Sampling



- The Cauchy proposal seems to converge faster than the Student's-$\mathcal{T}$.
- The Normal proposal produces jumps in the averages due to the unbounded variance of the estimator.

Importance Sampling



- The Cauchy proposal seems to converge faster than the Student's-$\mathcal{T}$.
- The Normal proposal produces jumps in the averages due to the unbounded variance of the estimator.

Importance Sampling for Integration

- Let us compute

$$\int_{2.1}^{\infty} x^5 \pi(x) dx$$

where $\pi(x) = \frac{\Gamma((\upsilon+1/2))}{\sigma\sqrt{\upsilon\pi}\Gamma(\upsilon/2)}\left(1+\frac{x^2}{\upsilon}\right)^{-(\upsilon+1)/2}$ is a student's-$\mathcal{T}$ distribution with $\nu > 1$.

- We use the following proposal distributions

$$q_1(x) = \pi(x),\ \ q_2(x) = \frac{\Gamma(1)}{\sqrt{\pi}\Gamma(1/2)}\left(1+x^2\right)^{-1}\ (Cauchy\ distribution),\ q_3(x) = \mathcal{N}(x; 0, \frac{\upsilon}{\upsilon-2}),\ \upsilon > 1$$

- It is easy to see that $\frac{\pi(x)}{q_2(x)} < \infty,\ \int \frac{\pi(x)}{q_3(x)} dx = \infty$

$\frac{\pi(x)}{q_2(x)}$ is unbounded.

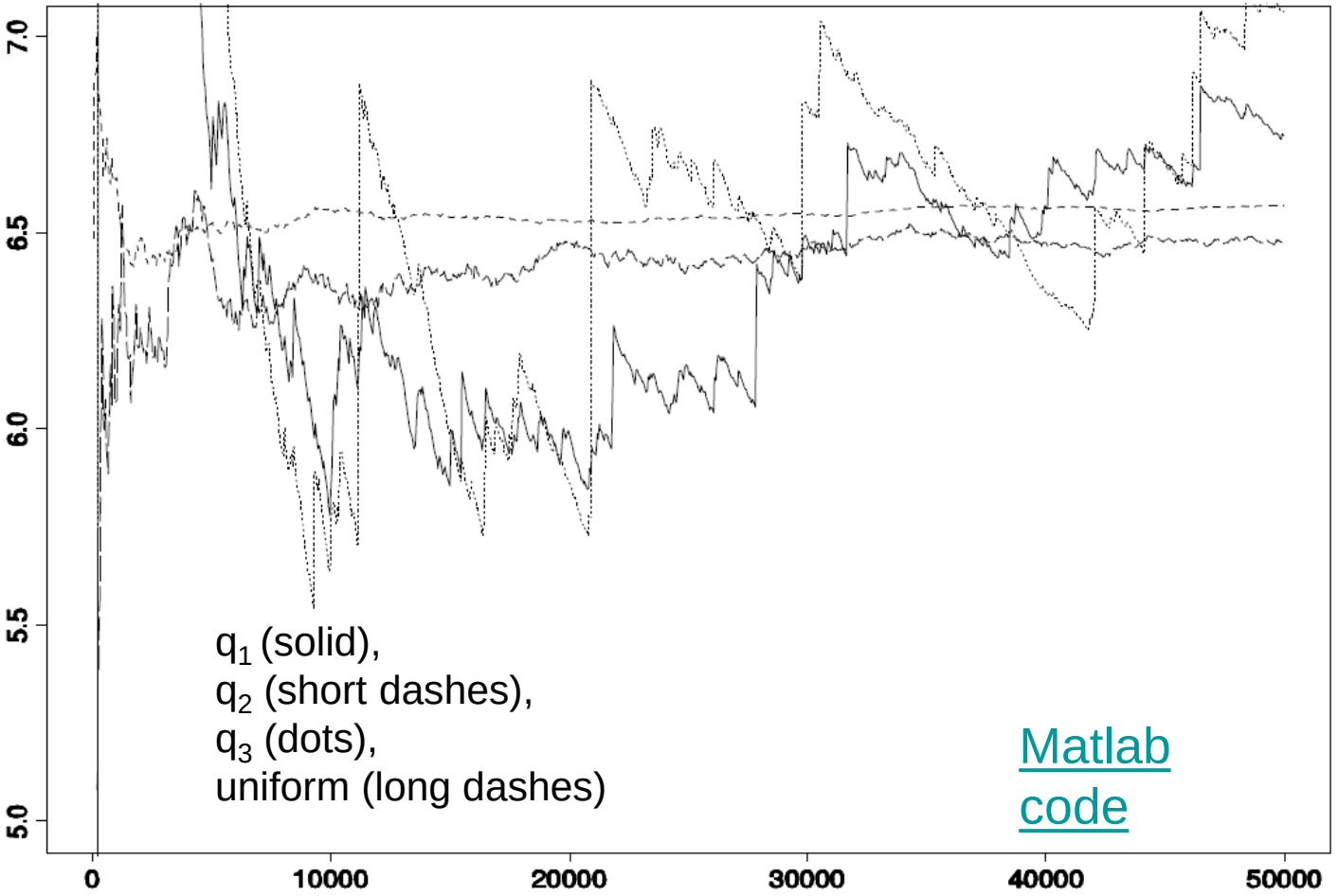
Importance Sampling for Integration

- Another solution is related to the fact that using a change of variable u = 1/x, we have

$$\int_{2.1}^{\infty} x^5 \pi(x) dx = \int_{0}^{1/2.1} u^{-7} \pi(1/u) du = \frac{1}{2.1} \int_{0}^{1/2.1} 2.1 u^{-7} \pi(1/u) du$$

which is the expectation of $u^{-7}\pi(1/u)$ with respect to $u \sim \mathcal{U}(0, 1/2.1)$

Importance Sampling for Integration



Performance for v = 12 with q_1 (Direct MC sampling), q_2 (Importance sampling with Cauchy distribution), q_3 (Importance sampling with Gaussian), and direct MC sampling from uniform distribution. Final values 6.271, 6.535, 5.513, 6.523 vs. true value 6.540

Bayesian Analysis of Markov Chain

- Application to Bayesian analysis of Markov chain. Consider a two state Markov chain with transition matrix F

$$\begin{bmatrix} p_1 & 1-p_1 \\ 1-p_2 & p_2 \end{bmatrix}$$

that is

$$\Pr(X_{t+1}=1 \mid X_t=1) = 1-\Pr(X_{t+1}=2 \mid X_t=1) = p_1$$
$$\Pr(X_{t+1}=2 \mid X_t=2) = 1-\Pr(X_{t+1}=1 \mid X_t=2) = p_2$$

with physical constraints $p_1+p_2 < 1$

- Assume we observe $x_1,\ldots,x_m$ and the prior is

$$\pi\ p_1, p_2\ \propto 2\mathbb{I}_{p_1+p_2<1}$$

Then the posterior is

$$\pi\ p_1, p_2 \mid x_{1:m}\ \propto p_1^{m_{1,1}}\ 1-p_1\ ^{m_{1,2}}\ 1-p_1\ ^{m_{1,2}}\ p_2^{m_{2,2}}\mathbb{I}_{p_1+p_2<1}$$

where

$$m_{i,j} = \sum_{t=1}^{m-1} \mathbb{I}_{X_t=i}\mathbb{I}_{X_{t+1}=j}$$

Bayesian Analysis of Markov Chain

- The posterior does not admit a standard expression and its normalization constant is unknown. We can sample from it using rejection sampling (works well but computationally expensive).

- We are interested in estimating $\mathbb{E}\left[\varphi_i \;\; p_1, p_2 \;\; | \, x_{1:m}\right]$ for

$$\varphi_1 \;\; p_1, p_2 \;\; = p_1, \varphi_2 \;\; p_1, p_2 \;\; = p_2, \varphi_3 \;\; p_1, p_2 \;\; = \frac{p_1}{1 - p_1},$$

$$\varphi_4 \;\; p_1, p_2 \;\; = \frac{p_2}{1 - p_2}, \varphi_5 \;\; p_1, p_2 \;\; = \log \frac{p_1 \;\; 1 - p_2}{p_2 \;\; 1 - p_1}$$

  using importance sampling.

- If there was no constraint on $p_1 + p_2 < 1$ and $\pi(p_1, p_2)$ was uniform on $[0,1] \times [0,1]$, then the posterior would be

$$\pi_0 \;\; p_1, p_2 \,|\, x_{1:m} \;\; \propto \mathcal{Be} \;\; p_1; \; m_{1,1} + 1 \;\; , \;\; m_{1,2} + 1 \;\; \mathcal{Be} \;\; p_2; \; m_{2,2} + 1 \;\; , \;\; m_{2,1} + 1$$

- For the given data $\left(m_{1,1}, m_{1,2}, m_{2,2}, m_{2,1}\right)$ this is not efficient as we have $\pi_0 \; (p_1 + p_2 < 1 | x_{1:m}) = 0.21$.

Bayesian Analysis of Markov Chain

- The form of the posterior suggests using a Dirichlet distribution with density

$$\pi_1\left(p_1, p_2 \mid x_{1:m}\right) \propto p_1^{m_{1,1}} p_2^{m_{2,2}} \left(1-p_1-p_2\right)^{m_{1,2}+m_{2,1}}$$

However, $\pi\left(p_1, p_2 \mid x_{1:m}\right) / \pi_1\left(p_1, p_2 \mid x_{1:m}\right)$ is unbounded.

Bayesian Analysis of Markov Chain

- (Geweke, 1989) proposed using two independent normal distributions to approximate the binomial distribution

$$\pi_2(p_1,p_2|x_{1:m}) \propto \exp\left(-\left(m_{1,1}+m_{1,2}\right)(p_1-\hat{p}_1)^2/\left(2\hat{p}_1(1-\hat{p}_1)\right)\right) \times \exp\left(-\left(m_{2,1}+m_{2,2}\right)(p_2-\hat{p}_2)^2/\left(2\hat{p}_2(1-\hat{p}_2)\right)\right)$$

where $\hat{p}_1 = \dfrac{m_{1,1}}{m_{1,1}+m_{1,2}}$, $\hat{p}_2 = \dfrac{m_{2,2}}{m_{2,1}+m_{2,2}}$. The ratio $\pi\left(p_1,p_2\,|\,x_{1:m}\right) \Big/ \pi_2\left(p_1,p_2\,|\,x_{1:m}\right)$ is upper bounded.

- A final choice consists of using

$$\pi_3\left(p_1,p_2\,|\,x_{1:m}\right) \propto \mathcal{B}e\left(p_1;\, m_{1,1}+1,\, m_{1,2}+1\right)\pi_3\left(p_2\,|\,x_{1:m},p_1\right)$$

where $\pi\left(p_2\,|\,x_{1:m},p_1\right) \propto \left(1-p_2\right)^{m_{2,1}} p_1^{m_{2,2}}\mathbb{I}_{p_2<1-p_1}$ is badly approximated through $\pi_3\left(p_2\,|\,x_{1:m},p_1\right) = \dfrac{2}{\left(1-p_1\right)^2}\,p_2\,\mathbb{I}_{p_2<1-p_1}$

It is straightforward to check that

$$\pi\left(p_1,p_2\,|\,x_{1:m}\right) \Big/ \pi_3\left(p_1,p_2\,|\,x_{1:m}\right) \propto \left(1-p_2\right)^{m_{2,1}} p_2^{m_{2,2}} \Big/ \frac{2}{\left(1-p_1\right)^2}\,p_2 < \infty$$

Bayesian Analysis of Markov Chain

- The conditional distribution $\pi_3(p_2 \mid x_{1:m}, p_1) = \frac{2}{(1-p_1)^2} p_2 \mathbb{I}_{p_2<1-p_1}$ can be sampled using the inverse CDF method. The CDF of this distribution is

$$F(p_2 \mid x_{1:m}, p_1) = \int_0^{p_2} \frac{2}{(1-p_1)^2} x \mathbb{I}_{p_2<1-p_1} dx = \frac{p_2^2}{(1-p_1)^2}, \text{ for } 0 < p_2 < 1-p_1$$

  - Generate a sample u~$\mathcal{U}(0,1)$
  - Obtain: $p_2 = F^{-1}(u) = (1-p_1)\sqrt{u}$

- Performance for *N = 10,000*

MatLab code

Distribution	φ1	φ2	φ3	φ4	φ5
π_1	0.681	0.079	2.145	0.086	3.231
π_2	0.707	0.190	2.435	0.239	2.359
π_3	0.697	0.189	2.380	0.239	2.357
π	0.697	0.189	2.374	0.240	2.359

Sampling Importance Resampling (SIR)

- We can draw unweighted samples from $p(x)$ by first using importance sampling (with proposal q) to generate a distribution of the form

$$p(\boldsymbol{x}) \approx \sum_s w_s \delta_{\boldsymbol{x}^s}(\boldsymbol{x}),\ \boldsymbol{x}^s \sim q(\boldsymbol{x})$$

- Here w_s are the normalized importance weights. We then sample with replacement from the Eq. above, where the probability that we pick $\boldsymbol{x^s}$ is w_s. Let this procedure induce a distribution denoted by $\hat{p}$.

- To see that this is valid, note that

$$\hat{p}(x \le x_0) = \sum_s \mathbb{I}(x^s \le x_0)\, w_s = \frac{\sum_s \mathbb{I}(x^s \le x_0) \frac{\tilde{p}(x^s)}{q(x^s)}}{\sum_s \frac{\tilde{p}(x^s)}{q(x^s)}} \to \frac{\int \mathbb{I}(x \le x_0) \frac{\tilde{p}(x)}{q(x)} q(x) dx}{\int \frac{\tilde{p}(x)}{q(x)} q(x) dx}$$

$$= \frac{\int \mathbb{I}(x \le x_0) \tilde{p}(x) dx}{\int \tilde{p}(x) dx} = \int \mathbb{I}(x \le x_0) p(x) dx = p(x \le x_0)$$

- This SIR result is an unweighted approximation $p(\boldsymbol{x}) \approx \frac{1}{S'} \sum_{s=1}^{S'} \delta_{\boldsymbol{x}^s}(\boldsymbol{x}), S' \ll S$.
  - Smith, A. F. M. and A. E. Gelfand (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician 46*(2), 84–88.

Sampling Importance Resampling (SIR)

- This algorithm can be used to perform Bayesian inference in low-dimensional settings.
- Suppose we want to draw (unweighted) samples from the posterior, $p(\boldsymbol{\theta}|D) = p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})/p(\mathcal{D})$. We can use importance sampling with $\tilde{p}(\boldsymbol{\theta}) = p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ as the unnormalized posterior, and $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$ as our proposal. The normalized weights have the form

$$w_s = \frac{\frac{\tilde{p}(\boldsymbol{\theta}_s)}{q(\boldsymbol{\theta}_s)}}{\sum_{s'} \frac{\tilde{p}(\boldsymbol{\theta}_{s'})}{q(\boldsymbol{\theta}_{s'})}} = \frac{p(\mathcal{D}|\boldsymbol{\theta}_s)}{\sum_{s'} p(\mathcal{D}|\boldsymbol{\theta}_{s'})}$$

- We can then use SIR to sample from $p(\boldsymbol{\theta}|\mathcal{D})$.
- If there is a big discrepancy between our proposal (the prior) and the target (the posterior), we will need a huge number of importance samples for this technique to work reliably, since otherwise the variance of the importance weights will be very large, implying that most samples carry no useful information.

Monte Carlo EM Algorithm

- Sampling methods can be used to approximate the E step of the EM algorithm. Consider a model with hidden variables **Z**, visible (observed) variables **X**, and parameters $\boldsymbol{\theta}$. The function that is optimized with respect to $\boldsymbol{\theta}$ in the M step is the expected complete-data log likelihood, given by

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \int p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \, lnp(\boldsymbol{Z}, \boldsymbol{X}|\boldsymbol{\theta}) \, d\boldsymbol{Z}$$

- We can use sampling methods to approximate this integral

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) \approx \frac{1}{L} \sum_{l=1}^{L} \ln p(\boldsymbol{Z}^{(l)}, \boldsymbol{X}|\boldsymbol{\theta}) \;, \; \boldsymbol{Z}^{(l)} \sim p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old})$$

- The Q function is then optimized in the usual way in the M step.

- Can extend to finding the MAP estimate by adding ln $p(\boldsymbol{\theta})$ to the function $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ before performing the M step.

Stochastic EM

- A particular instance of the Monte Carlo EM algorithm, called *stochastic EM*, arises if we consider a finite mixture model, and draw just one sample at each E step.

- Here the latent variable $\mathbf{Z}$ characterizes which of the K components of the mixture is responsible for generating each data point.

- In the E step, a sample of $\mathbf{Z}$ is taken from the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$ where $\mathbf{X}$ is the data set. This effectively makes a hard assignment of each data point to one of the components in the mixture.

- In the M step, this sampled approximation to the posterior distribution is used to update the model parameters in the usual way.

Data Augmentation Algorithm

- Consider a full Bayesian treatment in which we wish to sample from the posterior distribution over the parameter vector $\boldsymbol{\theta}$.
- In principle, we would like to draw samples from the joint posterior $p(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{X})$, but we shall suppose that this is computationally difficult.
- Suppose further that it is relatively straightforward to sample from the complete-data parameter posterior $p(\boldsymbol{\theta}|\mathbf{Z}, \mathbf{X})$.
- This inspires the *data augmentation* algorithm, which alternates between two steps known as the
  - I-step (imputation step, analogous to an E step) and
  - the P-step (posterior step, analogous to an M step).

Data Augmentation Algorithm

Objective: Sample from $p(\boldsymbol{\theta}|\boldsymbol{X})$. Assume that it is difficult to sample from $p(\boldsymbol{Z},\boldsymbol{\theta}|\boldsymbol{X})$ but easier to sample from $p(\boldsymbol{\theta}|\boldsymbol{Z},\boldsymbol{X})$.

1. Imputation Step: Since we cannot sample directly from $p(\boldsymbol{Z}|\boldsymbol{X})$, we draw $\boldsymbol{\theta}^{(l)}$, $\boldsymbol{Z}^{(l)}$ from $p(\boldsymbol{Z},\boldsymbol{\theta}|\boldsymbol{X})$ using

$$p(\boldsymbol{Z}|\boldsymbol{X}) = \int p(\boldsymbol{Z}|\boldsymbol{\theta},\boldsymbol{X})p(\boldsymbol{\theta}|\boldsymbol{X})d\boldsymbol{\theta}$$

   - Draw samples $\boldsymbol{\theta}^{(l)}$ from the current estimate for $p(\boldsymbol{\theta}|\boldsymbol{X})$, and
   - Use this to draw samples $\boldsymbol{Z}^{(l)}$ from $p(\boldsymbol{Z}|\boldsymbol{\theta}^{(l)},\boldsymbol{X})$.

2. Posterior Step:

$$p(\boldsymbol{\theta}|\boldsymbol{X}) = \int p(\boldsymbol{\theta}|\boldsymbol{Z},\boldsymbol{X})p(\boldsymbol{Z}|\boldsymbol{X})d\boldsymbol{Z}$$

   - Use the $\boldsymbol{Z}^{(l)}$ from the I-Step to compute a revised version of $p(\boldsymbol{\theta}|\boldsymbol{X})$,

$$p(\boldsymbol{\theta}|\boldsymbol{X}) \approx \frac{1}{L}\sum_{l=1}^{L} \mathrm{p}(\boldsymbol{\theta}|\boldsymbol{Z}^{(l)},\boldsymbol{X})$$

Summary: Importance Sampling

- Simulation from f (the true density) is not necessarily optimal
- Alternative to direct sampling from f is importance sampling, based on the alternative representation

$$\mathbb{E}_f\left[h(x)\right] = \int_{\chi}\left[h(x)\frac{f(x)}{q(x)}\right]q(x)dx = \mathbb{E}_q\left[h(x)\frac{f(x)}{q(x)}\right]$$

which allows us to use other distributions than f

Importance Sampling Algorithm

Evaluation of $\mathbb{E}_f\left[h(x)\right] = \int_{\chi} h(x)\,f(x)dx$

by

- Generating a sample $x_1, \ldots, x_m$ from a distribution q
- Using the approximation $\frac{1}{m}\sum_{j=1}^{m}\frac{f(x_j)}{q(x_j)}h(x_j)$

Summary: Justification

- Convergence of the estimator

$$\frac{1}{m}\sum_{j=1}^{m}\frac{f(x_j)}{q(x_j)}h(x_j)\rightarrow \mathbb{E}_f\left[h(x)\right]$$

- Converges for any choice of the distribution q as long as $\text{supp(q)} \supset \text{supp(f)}$
- Instrumental distribution q chosen from distributions easy to simulate
- q should not be small or zero in regions where the target distribution is significant.
- Same sample (generated from q) can be used repeatedly, not only for different functions h, but also for different densities f

Summary: Choice of Importance Function

- q can be any density but some choices better than others
- Finite variance only when

$$\mathbb{E}_f\left[h^2(x)\frac{f(x)}{q(x)}\right] = \int_{\mathcal{X}} h^2(x)\frac{f^2(x)}{q(x)}dx < \infty$$

- Instrumental distributions with tails lighter than those of f (that is, with sup $f/q = \infty$) not appropriate, because weights $f(x_j)/q(x_j)$ vary widely, giving too much importance to a few values x_j.
- If $\sup f/q = M < \infty$, the accept-reject algorithm can be used as well to simulate f directly.
- IS suffers from the curse of dimensionality

Discussion

- Importance sampling is useful for a few non-standard distributions but does not work for most other problems.
- The key problem is the design of a proper proposal distribution.
- Sequential MC will be discussed later addressing this last problem.

Additional Readings on Importance Sampling:

- John Geweke, Bayesian Inference in Econometric Models using MC Integration, *Econometrica*, Vol. 57, No. 6 (Nov., 1989), pp. 1317-1339.
- Herman K. Van Dijk, J. Peter Hop and Adri S. Louter, An Algorithm for the Computation of Posterior Moments and Densities Using Simple Importance Sampling, *Journal of the Royal Statistical Society. Series D (The Statistician)*, Vol. 36, No. 2/3, (1987) pp. 83-90.
- Art Owen and Yi Zhou, Safe and effective importance sampling, *Journal of the American Statistical Association*, Vol. 95, No. 449 (Mar., 2000), pp. 135-143.