Probabilistic Graphical Models

Brown University CSCI 2950-P, Spring 2013
Prof. Erik Sudderth

Lecture 14:
Monte Carlo Methods,
Rejection Sampling, Importance Sampling

Monte Carlo Methods

$$\mathbb{E}[f] = \int f(x)p(x)\ dx \approx \frac{1}{L}\sum_{\ell=1}^{L} f(x^{(\ell)}) \qquad x^{(\ell)} \sim p(x)$$

Estimation of expected model properties via simulation

Provably good if *L* sufficiently large:

- Unbiased for any sample size
- Variance inversely proportional to sample size (and independent of dimension of space)
- Laws of large numbers, central limit theorem, …

PROBLEM: Sampling from complex distributions

- Exact sampling: Closed form and iterative methods
- Importance sampling
- Sequential importance sampling & particle filters
- Markov chain Monte Carlo (MCMC)

Monte Carlo Estimators

$$\mu \triangleq \mathbb{E}[f] = \int f(x)p(x)\, dx \approx \frac{1}{L}\sum_{\ell=1}^{L} f(x^{(\ell)}) \triangleq \hat{f}_L$$

- Expectation estimated from *empirical distribution* of L samples:

$$\hat{p}_L(x) = \frac{1}{L}\sum_{\ell=1}^{L} \delta_{x^{(\ell)}}(x) \qquad\qquad x^{(\ell)} \sim p(x)$$

- The *Dirac delta* function is only well-defined within integrals:

$$\int_{\mathcal{X}} \delta_{\bar{x}}(x) f(x)\, dx = f(\bar{x}) \qquad\qquad \int_{A} \delta_{\bar{x}}(x)\, dx = \mathbb{I}(\bar{x} \in A)$$

- For any *L* this estimator, a random variable, is *unbiased*:

$$\mathbb{E}[\hat{f}_L] = \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}[f(x^{(\ell)})] = \mathbb{E}[f]$$

Monte Carlo Asymptotics

$$\mu \triangleq \mathbb{E}[f] = \int f(x)p(x)\,dx \approx \frac{1}{L}\sum_{\ell=1}^{L} f(x^{(\ell)}) \triangleq \hat{f}_L$$

- Variance is inversely proportional to the number of samples:

$$\mathrm{Var}[\hat{f}_L] = \frac{1}{L}\mathrm{Var}[f] = \frac{1}{L}\mathbb{E}[(f(x)-\mu)^2]$$

- Even if true variance is infinite, have *laws of large numbers*:

Weak Law

$$\lim_{L\to\infty} \Pr(|\hat{f}_L - \mu| < \epsilon) = 1, \text{ for any } \epsilon > 0$$

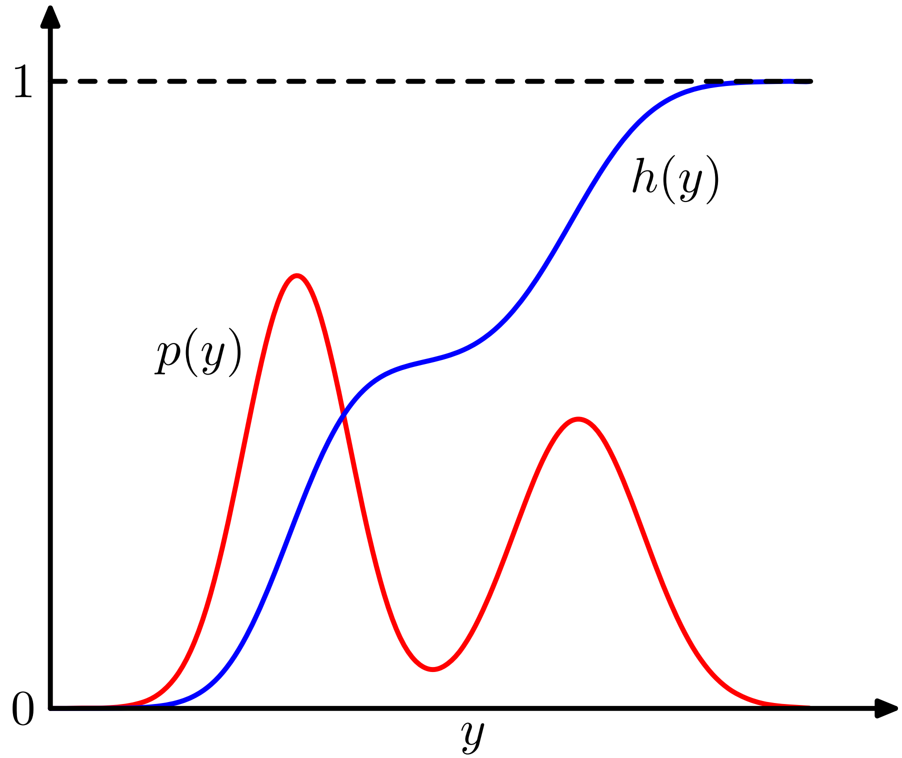
Strong Law

$$\Pr\left(\lim_{L\to\infty} \hat{f}_L = \mu\right) = 1$$

- If the true variance is finite, also have *central limit theorem*:

$$\sqrt{L}\left(\hat{f}_L - \mu\right) \underset{L\to\infty}{\Longrightarrow} \mathcal{N}(0, \mathrm{Var}[f])$$

Random Number Generation



Cumulative Distribution:

$$h(x) = \int_{-\infty}^{x} p(z)\, dz$$

Applying Inverse CDF:

$$u^{(\ell)} \sim \text{Unif}(0, 1)$$
$$x^{(\ell)} = h^{-1}(u^{(\ell)})$$

$$\Rightarrow \quad x^{(\ell)} \sim P(x)$$

- Chaotic dynamical systems are used to generate sequences of pseudo-random numbers approximately distributed uniformly on [0,1]
- Simplest examples are *linear congruential generators*, but try to *use more sophisticated methods!*

$$\bar{u}^{(\ell+1)} = (a\bar{u}^{(\ell)} + c) \bmod m \qquad u^{(\ell)} = \frac{1}{m}\bar{u}^{(\ell)}$$

Among other conditions, *c* and *m* should be relatively prime

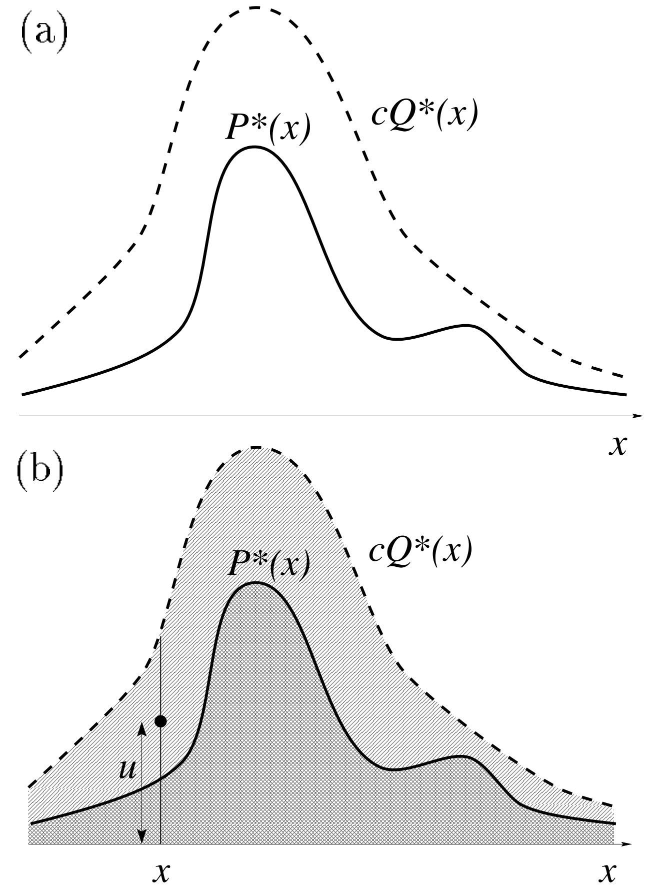
Rejection Sampling

Target Distribution:

$$p(x) = \frac{1}{Z} p^*(x)$$

Proposal Distribution:

$$q(x) = \frac{1}{Z'} q^*(x)$$



- Can sample by drawing uniformly from region under a density function:

$$p(x, u) = p(x)p(u \mid x) = p(x)\text{Unif}(u \mid 0, p^*(x))$$

- A rejection sampler requires an *envelope* proposal distribution:

$$cq^*(x) > p^*(x) \text{ for all } x$$

- The rejection sampling algorithm is:

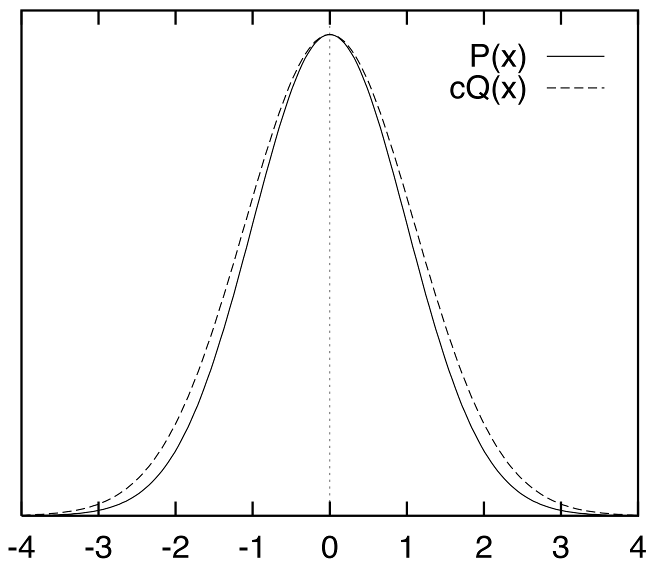
$$x \sim q(x), u \sim \text{Unif}(0, 1)$$

  - Accept this sample if $u < \frac{p^*(x)}{cq^*(x)}$
  - Otherwise, reject and repeat until a sample is accepted

Always produces a valid sample, but running time is a random variable. The constant c must be known!

High-Dimensional Rejection Sampling

- Consider N-dimensional multivariate Gaussian distributions:



Target Distribution:

$$p(x) = \mathcal{N}(x \mid 0, \sigma_P^2 I_N)$$

Proposal Distribution:

$$q(x) = \mathcal{N}(x \mid 0, \sigma_Q^2 I_N)$$

$$\sigma_Q > \sigma_P$$

- The tightest envelope matches densities at the origin:

$$c = \frac{(2\pi\sigma_Q^2)^{N/2}}{(2\pi\sigma_P^2)^{N/2}} = \exp\left(N \log \frac{\sigma_Q}{\sigma_P}\right)$$

- Even small mismatch can lead to tiny acceptance probabilities:

$$\frac{\sigma_Q}{\sigma_P} = 1.01 \qquad N = 1000$$

$$c = \exp(10) \simeq 20,000$$

For normalized densities, acceptance probability is c^{-1}

Importance Sampling

Target Distribution: *Proposal Distribution:*

$$p(x) = \frac{1}{Z} p^*(x) \qquad q(x) = \frac{1}{Z'} q^*(x) \qquad q(x) > 0 \text{ where } p(x) > 0$$

$$\mathbb{E}[f] = \int f(x)p(x)\;dx = \int f(x)w(x)q(x)\;dx \qquad w(x) = \frac{p(x)}{q(x)}$$

- Estimate target moments via *importance weighted* samples:

$$\hat{f}_L = \frac{1}{L}\sum_{\ell=1}^{L} f(x^{(\ell)})w(x^{(\ell)}) \qquad x^{(\ell)} \sim q(x)$$

- Alternative estimator when normalization constants unknown:

$$\mathbb{E}[f] = \int f(x)p(x)\;dx = \frac{Z'}{Z}\int f(x)w^*(x)q(x)\;dx \qquad w^*(x) = \frac{p^*(x)}{q^*(x)}$$

$$\hat{f}_L = \sum_{\ell=1}^{L} w_\ell f(x^{(\ell)}) \qquad x^{(\ell)} \sim q(x) \qquad w_\ell = \frac{w^*(x^{(\ell)})}{\sum_{m=1}^{L} w^*(x^{(m)})}$$

Optimal Proposal Distributions

Target Distribution: $$p(x) = \frac{1}{Z} p^*(x)$$

Proposal Distribution: $$q(x) = \frac{1}{Z'} q^*(x)$$

$$q(x) > 0 \text{ where } p(x) > 0$$

$$\hat{f}_L = \frac{1}{L} \sum_{\ell=1}^{L} f(x^{(\ell)}) w(x^{(\ell)}) \qquad x^{(\ell)} \sim q(x) \qquad w(x) = \frac{p(x)}{q(x)}$$

- This estimator is always *unbiased*: $\mathbb{E}_q[\hat{f}_L] = \mathbb{E}_p[f] \triangleq \mu$
- We can choose proposal distribution to minimize variance:

$$\mathrm{Var}_q[f(x)w(x)] = \mathbb{E}_q[f^2(x)w^2(x)] - \mu^2$$

$$\mathbb{E}_q[f^2(x)w^2(x)] \geq \left(\mathbb{E}_q[|f(x)|w(x)]\right)^2 = \left(\int |f(x)|p(x)\, dx\right)^2$$

Jensen's Inequality

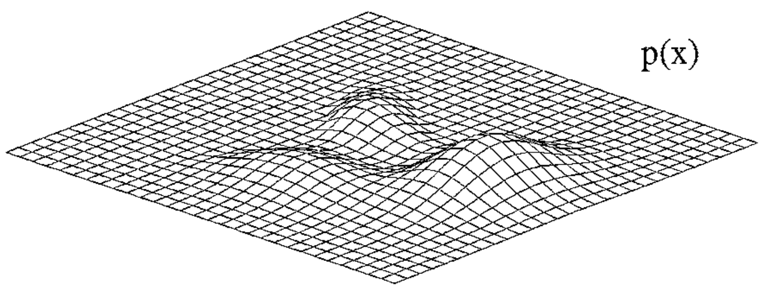
- Applying similar analysis with unknown normalization constants:
  - The importance estimator is *asymptotically unbiased*
  - The optimal, *minimum variance proposal* distribution is

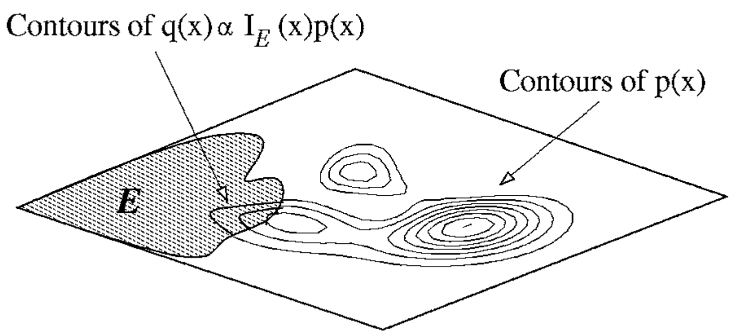
$$\hat{q}^*(x) = |f(x)|p(x) \qquad \hat{q}(x) \propto |f(x)|p(x)$$

Rare Event Simulation

Standard Monte Carlo: $$\hat{e}_L = \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{I}_E(x^{(\ell)}) \qquad x^{(\ell)} \sim p(x)$$

Simulate usual system execution, and count the number of "extreme" events. But what if such events are very rare?





Importance Sampling:

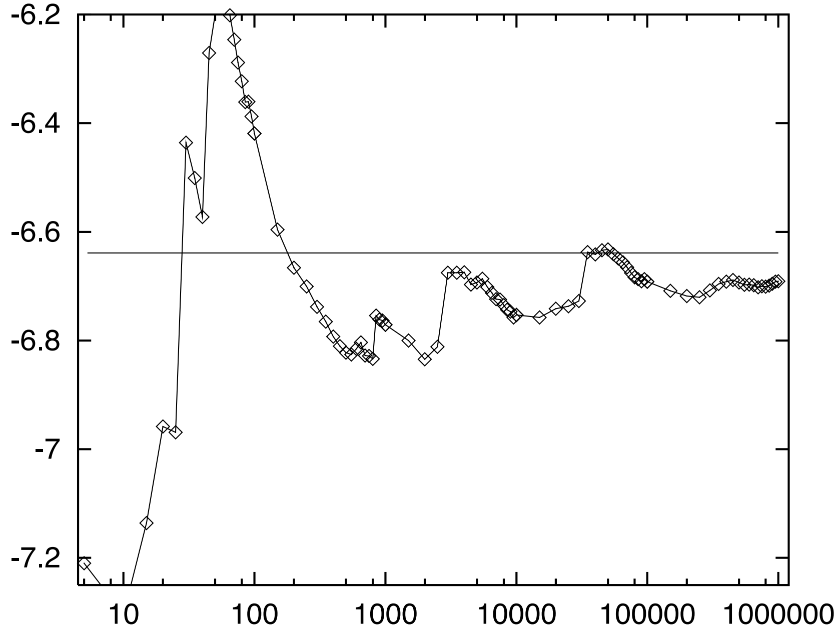
$$\hat{e}_L = \sum_{\ell=1}^{L} w_\ell \mathbb{I}_E(x^{(\ell)}) \qquad x^{(\ell)} \sim q(x) \qquad w_\ell = \frac{w^*(x^{(\ell)})}{\sum_{m=1}^{L} w^*(x^{(m)})}$$

Bias simulations towards extreme events, but use importance weights to correct probabilities.

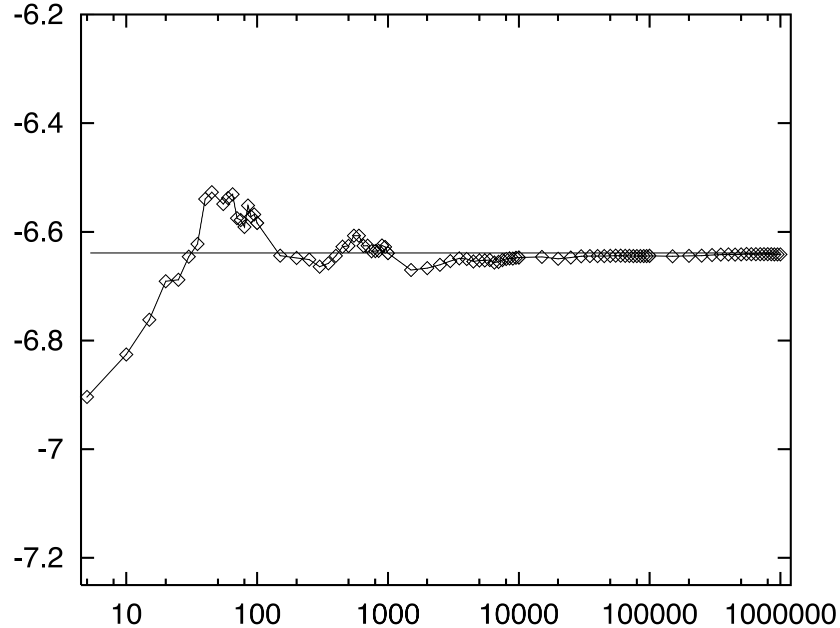
$$w^*(x) = \frac{p^*(x)}{q^*(x)}$$

Selecting Proposal Distributions

- For a toy one-dimensional, heavy-tailed target distribution:



Gaussian Proposal



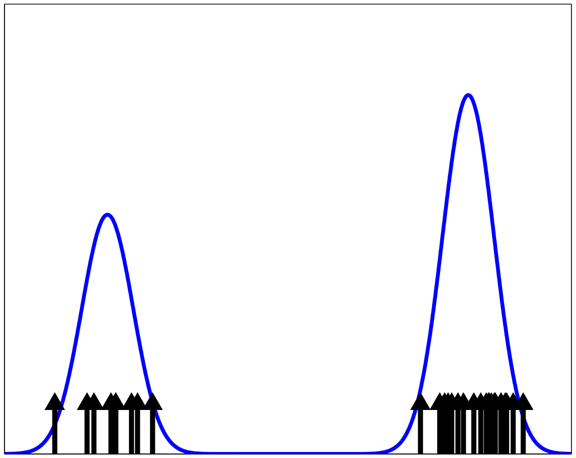
Cauchy (Student's-t) Proposal

Empirical variance of weights may not predict estimator variance

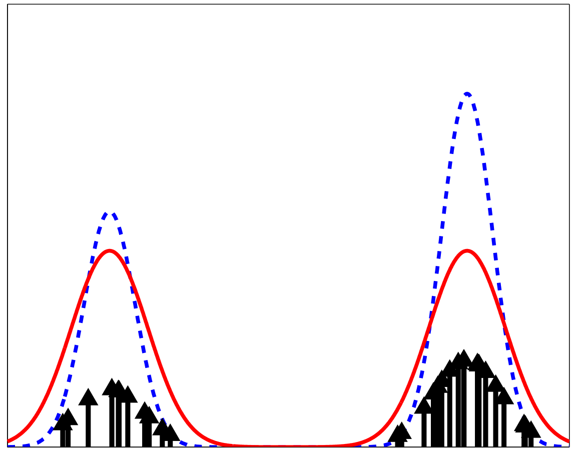
- Always (asymptotically) unbiased, but variance of estimator can be enormous unless weight function bounded above:

$$\mathbb{E}_q[\hat{f}_L] = \mathbb{E}_p[f] \qquad \mathrm{Var}_q[\hat{f}_L] = \frac{1}{L}\mathrm{Var}_q[f(x)w(x)] \qquad w(x) = \frac{p(x)}{q(x)}$$

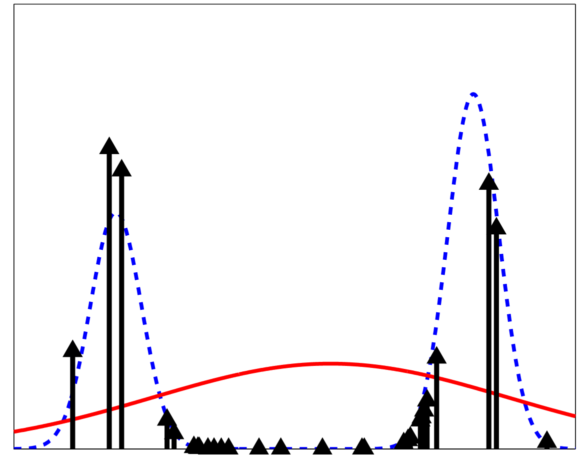
Selecting Proposal Distributions



Target Distribution



Good Proposal



Poor Proposal

Kernel or Parzen window estimators interpolate for nonparametric density prediction

$$\hat{p}(x) = \sum_{\ell=1}^{L} w^{(\ell)} \mathcal{N}(x; x^{(\ell)}, \Lambda)$$