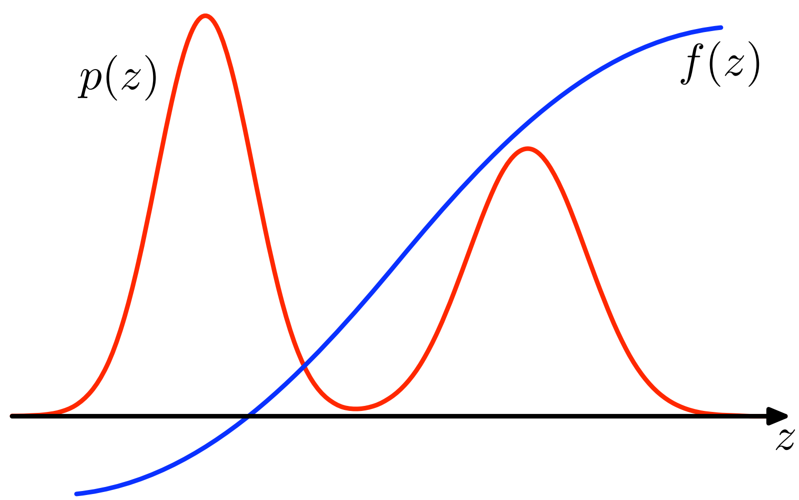
11. SAMPLING

J. Elder

Motivation

- For many probabilistic models, exact inference is intractable.
- In such cases, approximate solutions can often be obtained by sampling.
- We will focus on estimating expectations of functions of the hidden variables **z**, i.e.,

$$E[f] = \int f(\mathbf{z})p(\mathbf{z})\,d\mathbf{z}$$



Goal

$$E[f] = \int f(\mathbf{z})p(\mathbf{z})\,d\mathbf{z}$$

- Basic idea:

Draw L independent samples $\mathbf{z}^{(l)}$ from the distribution $p(\mathbf{z})$.

Then $E[f]$ can be approximated by:

$$\hat{f} = \frac{1}{L}\sum_{l=1}^{L} f\left(z^{(l)}\right)$$

Goal

- You can verify that:

$$E\left[\hat{f}\right] = E\left[f\right]$$

- and

$$\mathrm{var}\left[\hat{f}\right] = \frac{1}{L} E\left[\left(f - E\left[f\right]\right)^2\right]$$

- Thus the accuracy does not depend upon the dimensionality of **z**!

Sampling methods

- Directed graphical models
  - Can use ancestral sampling.
- Markov random fields
  - No one-pass method.
  - Can use Gibbs sampling.

Outline

- Basic sampling algorithms
- Markov Chain Monte Carlo (MCMC)
- Gibbs Sampling

BASIC SAMPLING ALGORITHMS

Standard Distributions

Standard distributions

- Suppose that we have a good method for generating (pseudo-)random uniformly distributed numbers z over [0, 1].
  - e.g., MATLAB's **unifrnd()**.
- Suppose that we wish to generate samples from a standard distribution $p(y)$.
- We would like to find a deterministic function $f(z)$ that will transform each sample z to a sample y such that y is distributed according to $p(y)$.

Standard distributions

- Recall that:

$$p(y) = p(z)\left|\frac{dz}{dy}\right|$$

Without loss of generality, we choose
$y = f(z)$ to be an increasing function of z.

Then $z = f^{-1}(y) \triangleq h(y)$ will be an increasing function of y.

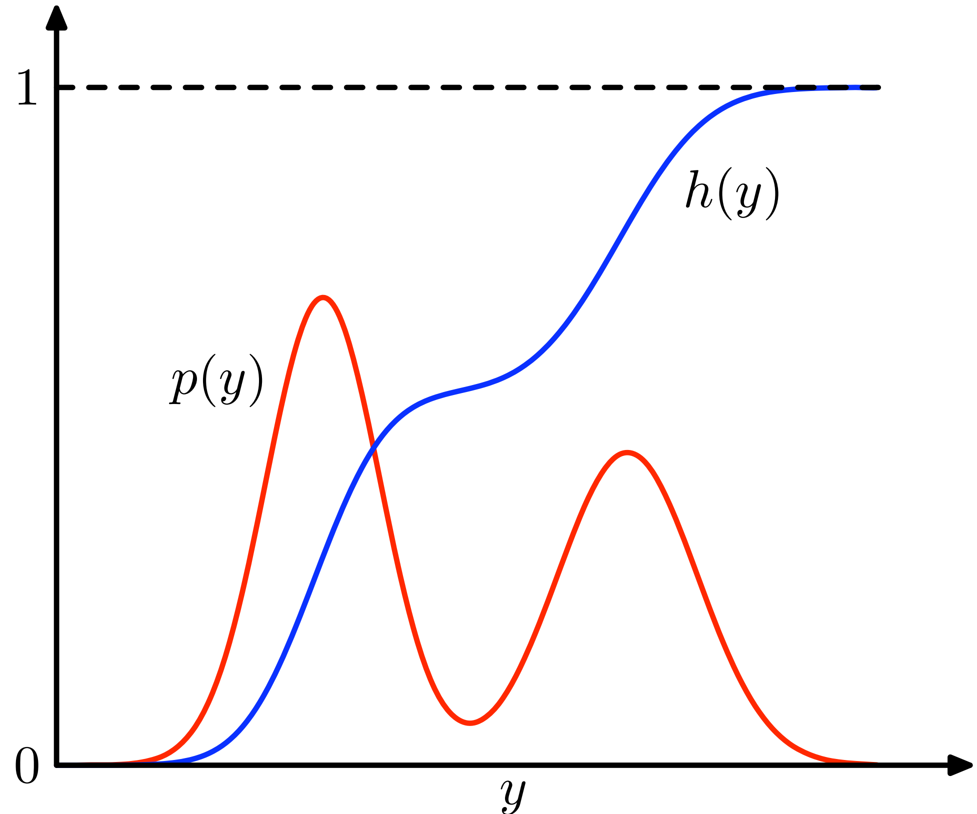
Thus $dz = p(y)dy \rightarrow z = h(y) = \int_{-\infty}^{y} p(\hat{y})\,d\hat{y}$

Standard distributions

- Thus to sample from *p(y)*, we generate random uniformly distributed numbers *z*, then transform them according to

$$y = h^{-1}(z),$$

$$\text{where } h(y) = \int_{-\infty}^{y} p(\hat{y})\, d\hat{y}$$



Generalization to multivariate distributions

$$p\left(y_1,\ldots y_M\right) = p\left(z_1,\ldots z_M\right)\left|\frac{\partial\left(z_1,\ldots z_M\right)}{\partial\left(y_1,\ldots y_M\right)}\right|$$

where $\left|\frac{\partial\left(z_1,\ldots z_M\right)}{\partial\left(y_1,\ldots y_M\right)}\right|$ is the Jacobian of *h*.

Example 1

- The exponential distribution

$$p(y) = \lambda \exp\left(-\lambda y\right), \quad y \geq 0$$

Example 2

- The Cauchy distribution

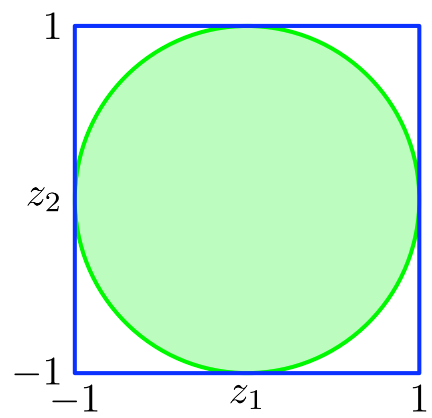
$$p(y) = \frac{1}{\pi} \frac{1}{1 + y^2}$$

Example 3

- Box-Muller method for generating i.i.d. Gaussian samples

1. Generate samples from two i.i.d. uniformly-distributed rv's $z_1, z_2 \in [-1,1]$

2. Reject samples lying outside unit circle.



3. Now transform to samples y_1, y_2 according to:

$$y_1 = z_1 \left(\frac{2 \log r^2}{r^2} \right)^{1/2} \qquad y_2 = z_2 \left(\frac{2 \log r^2}{r^2} \right)^{1/2}$$

It can be shown that y_1, y_2 are i.i.d. standard normal variables (0-mean, unit variance).

To generate i.i.d. Gaussian rv's with mean μ and std deviation σ, transform according to $\sigma y + \mu$.

Example 4

- Multivariate normal distributions

Use **Cholesky decomposition** $\Sigma = LL^t$

Then if $\mathbf{z}$ is a standard normal random vector,

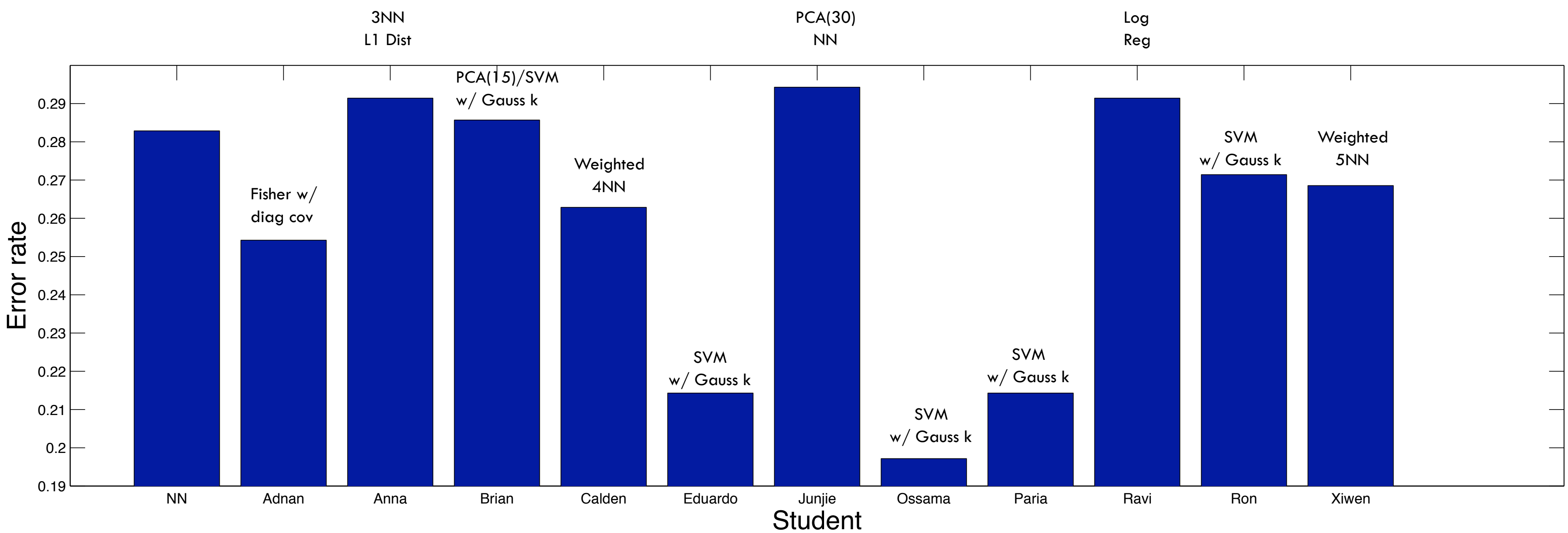
$\mathbf{y} = \mu + L\mathbf{z}$ will generate samples from $N\left(\mathbf{y}; \mu, \Sigma\right)$.

Limitations of the standard method

- Often the integration of $p(y)$ and/or inverse to generate $h(z)$ is not tractable.

END OF LECTURE DEC 6, 2010

Assignment 2 Competition Results



Rejection Sampling

Motivation

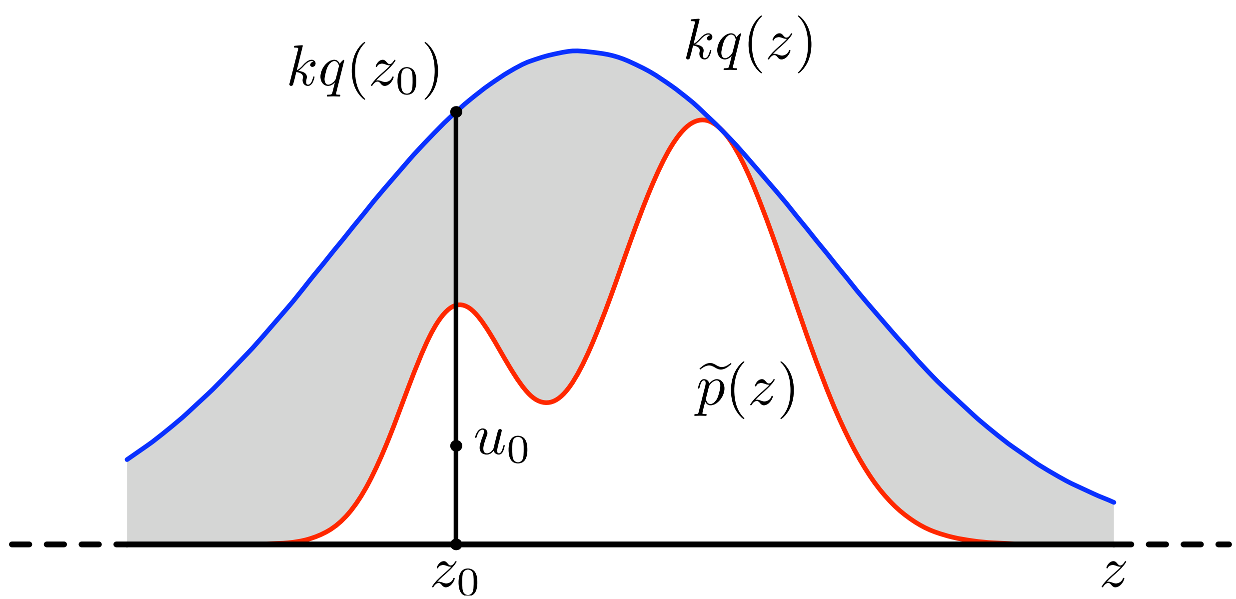
- Though it may be difficult to sample fairly from *p(**z**)* directly, it is often the case that *p(**z**)* can easily be evaluated for any given ***z*** (at least up to a normalizing constant *Z*).

$$\text{i.e., } p(\mathbf{z}) = \frac{1}{Z_p}\tilde{p}(\mathbf{z}), \text{ where } \tilde{p}(\mathbf{z}) \text{ can readily be evaluated.}$$

Main Idea

- Consider first the univariate case.
- Suppose we have a simpler distribution *q(z)* from which we can readily draw fair samples.
- Suppose further we can find a constant *k* such that:

$$kq(z) \geq \tilde{p}(z).$$



Algorithm

1. Generate a sample z_0 from $q(z)$.

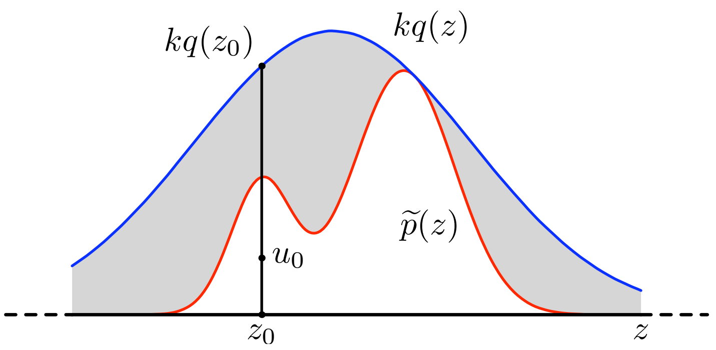
2. Generate a number u_0 from the uniform distribution on $\left[0, kq\left(z_0\right)\right]$.

   (Note that (z_0, u_0) is uniformly distributed under the curve $kq(z)$.)

3. If $u_0 > \tilde{p}\left(z_0\right)$, reject the sample, otherwise retain.

The retained pairs $\left(z_0, u_0\right)$ will have a uniform distribution under $\tilde{p}(z)$.

Thus the corresponding z values will be fair samples from $p(z)$.

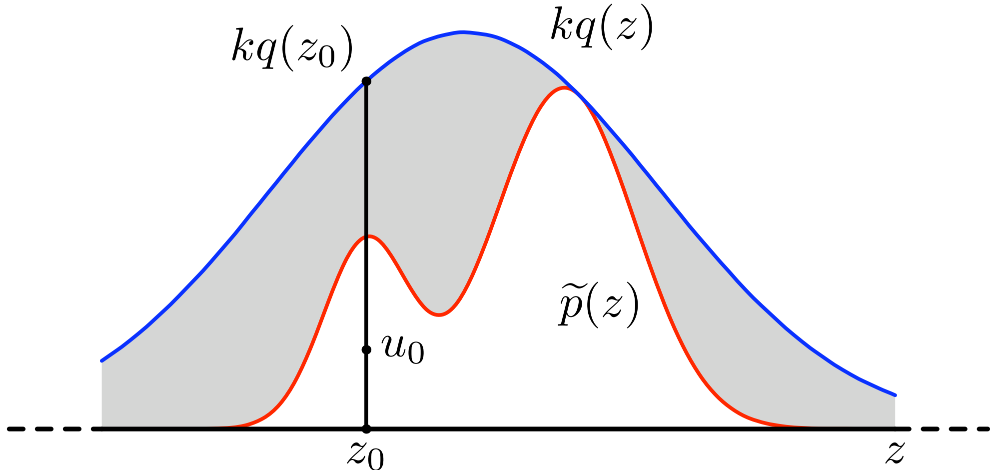


Efficiency

- The probability that a proposal is accepted is given by

$$p(\text{accept}) = \int \left\{ \tilde{p}(z) / kq(z) \right\} q(z)\, dz = \frac{1}{k} \int \tilde{p}(z)\, dz.$$

- Thus we want k to be as small as possible.



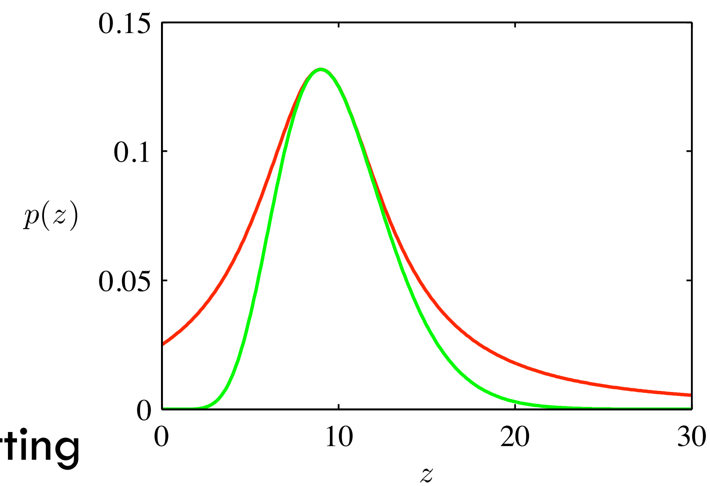
Example

- Suppose we wish to sample from the gamma distribution:

$$Gam(z \mid a,b) = \frac{b^a z^{a-1} \exp(-bz)}{\Gamma(a)}$$

- We know we can sample from the Cauchy distribution. We generalize slightly, and transform uniform random variables y according to

$$z = b \tan y + c$$



- which yields

$$q(z) = \frac{k}{1 + (z-c)^2 / b^2}$$

- The minimum rejection rate is obtained by setting

$$c = a - 1, \quad b^2 = 2a - 1$$

Limitations

- Can be hard to find a good bound $kq(z)$.
- Acceptance rate declines exponentially with dimensionality

Importance Sampling

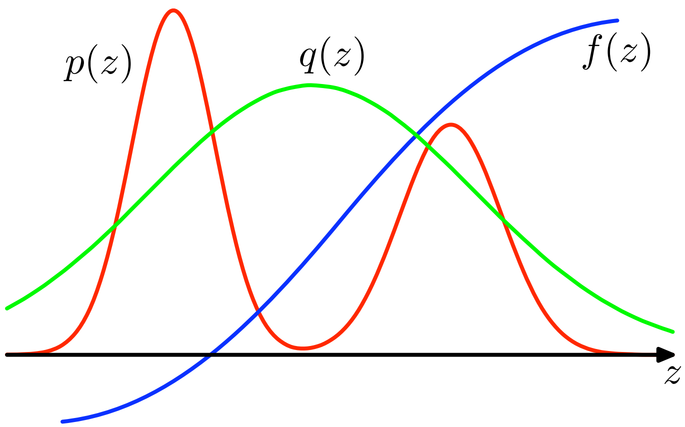
Importance Sampling

- Rather than trying to sample fairly from $p(\mathbf{z})$, let's just try to estimate the expectation E[f] directly.

$$E[f] = \int f(\mathbf{z})p(\mathbf{z})\,d\mathbf{z}$$

$$= \int f(\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})}q(\mathbf{z})\,d\mathbf{z}$$

$$\simeq \frac{1}{L}\sum_{l=1}^{L}\frac{p\left(\mathbf{z}^{(l)}\right)}{q\left(\mathbf{z}^{(l)}\right)}f\left(\mathbf{z}^{(l)}\right)$$



where the importance weights $r_l = \dfrac{p\left(\mathbf{z}^{(l)}\right)}{q\left(\mathbf{z}^{(l)}\right)}$

correct the bias introduced by sampling the wrong distribution.

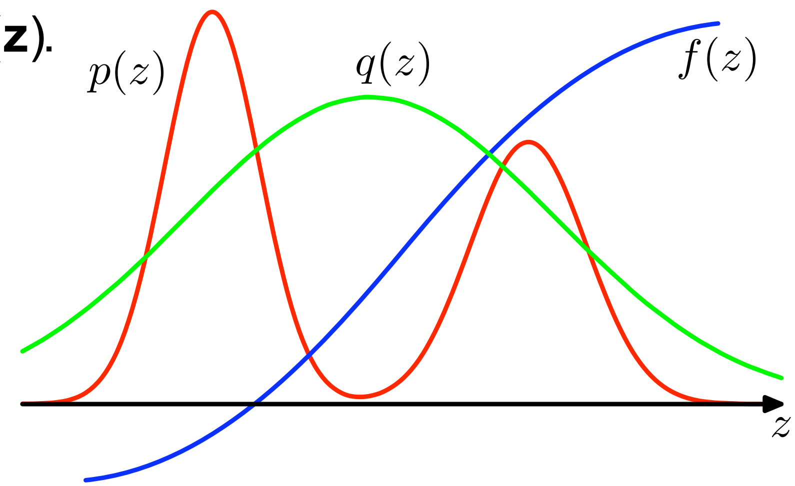
- Note that all samples can be retained.

Importance Sampling

- Suppose that $p(\mathbf{z})$ and $q(\mathbf{z})$ can only be evaluated up to a constant,

  i.e., we can sample from $\tilde{q}(\mathbf{z})$, and can calculate $\tilde{p}(\mathbf{z})$,

  where $p(\mathbf{z}) = \frac{1}{Z_p}\tilde{p}(\mathbf{z})$, and $q(\mathbf{z}) = \frac{1}{Z_q}\tilde{q}(\mathbf{z})$.



- Then we have

$$E[f] = \int f(\mathbf{z})p(\mathbf{z})\,d\mathbf{z}$$

$$= \frac{Z_q}{Z_p}\int f(\mathbf{z})\frac{\tilde{p}(\mathbf{z})}{\tilde{q}(\mathbf{z})}q(\mathbf{z})\,d\mathbf{z}$$

$$\simeq \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L}\tilde{r}_l f\left(\mathbf{z}^{(l)}\right)$$

where $\tilde{r}_l = \frac{\tilde{p}\left(\mathbf{z}^{(l)}\right)}{\tilde{q}\left(\mathbf{z}^{(l)}\right)}$.

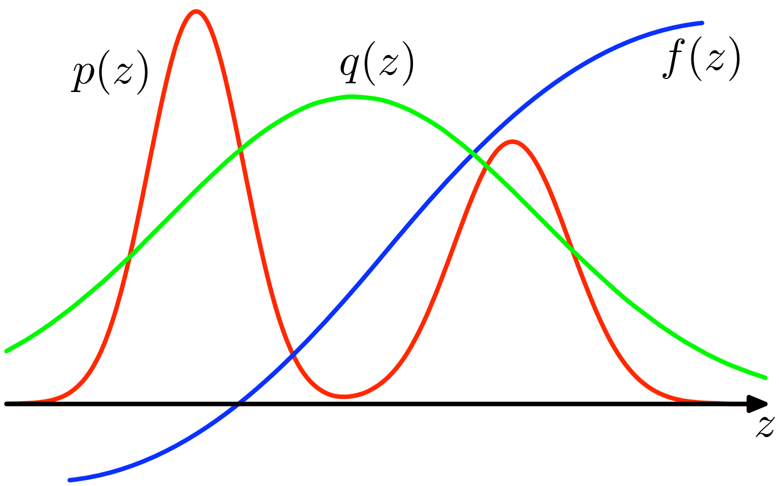
Importance Sampling

- Furthermore,

$$\frac{Z_p}{Z_q} = \frac{1}{Z_q}\int \tilde{p}(\mathbf{z})\,d\mathbf{z} = \int \frac{\tilde{p}(\mathbf{z})}{\tilde{q}(\mathbf{z})} q(\mathbf{z})\,d\mathbf{z} \simeq \frac{1}{L}\sum_{l=1}^{L} \tilde{r}_l$$

Thus $E[f] \simeq \sum_{l=1}^{L} w_l f\left(\mathbf{z}^{(l)}\right)$

where $w_l = \frac{\tilde{r}_l}{\sum_m \tilde{r}_m} = \frac{\tilde{p}\left(\mathbf{z}^{(l)}\right)/q\left(\mathbf{z}^{(l)}\right)}{\sum_m \tilde{p}\left(\mathbf{z}^{(m)}\right)/q\left(\mathbf{z}^{(m)}\right)}.$



Limitations

- Requires a good proposal distribution $q(\mathbf{z})$.

Likelihood Weighted Sampling

- A form of importance sampling can be applied to directed graphical models when some of the nodes have been observed.

Let the evidence set e represent the subset of variables that have been observed.

The algorithm is a modification of ancestral sampling in which:

1. If $\mathbf{z} \in e$, set $\mathbf{z}$ to its observed value.
2. Otherwise, sample from $p\left(\mathbf{z}_i \mid pa_i\right)$.

The resulting sample $\mathbf{z}$ is then assigned the weight

$$r(\mathbf{z}) = \prod_{\mathbf{z}_i \notin e} \frac{p\left(\mathbf{z}_i \mid pa_i\right)}{p\left(\mathbf{z}_i \mid pa_i\right)} \prod_{\mathbf{z}_i \in e} \frac{p\left(\mathbf{z}_i \mid pa_i\right)}{1} = \prod_{\mathbf{z}_i \in e} p\left(\mathbf{z}_i \mid pa_i\right)$$

Extensions

- Sampling-importance-resampling
  - Uses proposal distribution $q(\mathbf{z})$ to generate sample $\mathbf{z}$ with distribution that approximates $p(\mathbf{z})$.
  - Two-stage sampling process
  - Unlike rejection sampling, all samples are retained.
- Monte Carlo EM
  - Approximate E-step by sampling

MARKOV CHAIN MONTE CARLO METHODS

J. Elder

Motivation

- Rejection sampling and importance sampling do not scale well to high dimension.
- MCMC can potentially do better in higher dimensions, by staying in higher probability regions of the variable space

Basic Idea

Instead of sampling independently, each sample depends upon the previous sample through a conditional proposal distribution $q\left(\mathbf{z} \mid \mathbf{z}^{(\tau)}\right)$, forming a Markov chain of samples $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots$

Example: Metropolis Algorithm

Requires symmetric proposal distribution:

$$q(\mathbf{z}_A \mid \mathbf{z}_B) = q(\mathbf{z}_B \mid \mathbf{z}_A)$$

Sample is then accepted with probability

$$A(z^*, z^{(\tau)}) = \min\left(1, \frac{\tilde{p}(z^*)}{\tilde{p}(z^{(\tau)})}\right)$$

Note that samples that increase the probability are always kept.

If candidate sample accepted, then $\mathbf{z}^{(\tau+1)} \leftarrow \mathbf{z}^*$.

Otherwise, $\mathbf{z}^{(\tau+1)} \leftarrow \mathbf{z}^{(\tau)}$.

This leads to multiple copies of higher probability samples.

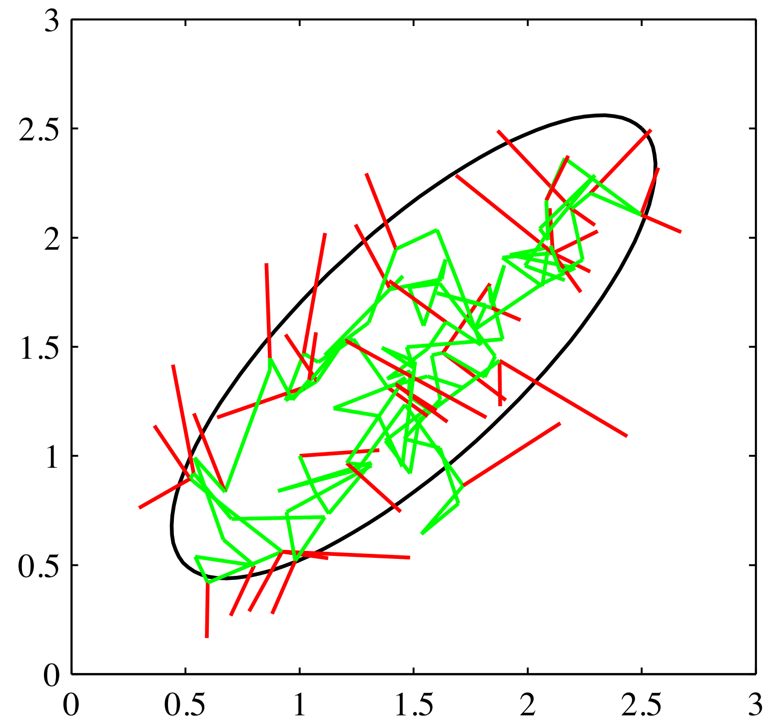
Metropolis Algorithm: Properties

If $q\left(\mathbf{z}_A \mid \mathbf{z}_B\right) > 0 \forall \mathbf{z}_A, \mathbf{z}_B$

Then the distribution of $\mathbf{z}^{(\tau)} \rightarrow p(\mathbf{z})$ as $\tau \rightarrow \infty$.

Note that the $\mathbf{z}^{(\tau)}$ are not independent.

Example



GIBBS SAMPLING

J. Elder

CSE 6390/PSYC 6225 Computational Modeling of Visual Perception

Gibbs Sampling

- Gibbs Sampling is a particularly simple form of MCMC algorithm.
- It's applicable to multivariate distributions for which the conditional distributions of the individual variables can be readily computed (e.g., MRFs).
- Each step involves replacing the value of one variable by a value drawn from the distribution of that variable conditioned on the current values of the remaining variables.

Gibbs Sampling: Algorithm

1. Initialize $\left\{ z_i^{(0)} \right\}$
2. Repeat until convergence
   a. Select a z_i
   b. Sample $z_i^{(\tau+1)} \sim p\left(z_i \mid \mathbf{z}^{(\tau)} \setminus z_i^{(\tau)} \right) = p\left(z_i \mid \text{ne}\left(z_i^{(\tau)} \right) \right)$

As long as $p\left(z_i \mid \mathbf{z}^{(\tau)} \setminus z_i^{(\tau)} \right) > 0 \forall \mathbf{z}, i$

Then the distribution of $\mathbf{z}^{(\tau)} \rightarrow p(\mathbf{z})$ as $\tau \rightarrow \infty$.