Sampling based inference

- Resources.
  - Bishop, chapter 11
  - Koller and Friedman, chapter 12
  - Andrieu et al. (linked on lecture page).
- Koller and Friedman uses “particles” terminology instead of “samples”.

Sampling based inference

- We have studied two themes in inference.
  - Marginalization / expectation / summing out or integration
  - Optimization
- Two flavors of activities
  - Fitting (inference using a model)
  - Learning (inference to find a model)
- These activities are basically the same in the generative modeling approach.

Motivation for sampling methods

- Real problems are typically complex and high dimensional.
- Example, images as evidence for stuff in the world

Motivation for sampling methods

- Real problems are typically complex and high dimensional.
- Suppose that we *could* generate samples from a distribution that is proportional to one we are interested in.

Typical case we are often interested in is $p(\theta|D)$

$$p(\theta|D)=\frac{p(\theta)p(D|\theta)}{p(D)}$$

Consider $\tilde{p}(z)=p(\theta)p(D|\theta)$

Motivation for sampling methods

- Generally, θ lives in a very high dimensional space.
- Generally, regions of high $\tilde{p}(z)$ is very little of that space.
- IE, the probability mass is very localized.
- Watching samples from $\tilde{p}(z)$ should provide a good maximum (one of our inference problems)

Motivation for sampling methods (II)

- Now consider computing the expectation of a function $f(z)$ over $p(z)$.
- Recall that this looks like $E_{p(z)}[f] = \int_z f(z)p(z)dz$
- A bad plan for computing E:

  Discretize the space where z lives into L blocks

  Then compute $E_{p(z)}[f] \cong \frac{1}{L}\sum_{l=1}^{L} p(z)f(z)$

Motivation for sampling methods (II)

- Now consider computing the expectation of a function $f(z)$ over $p(z)$.
- Recall that this looks like $E_{p(z)}[f] = \int_z f(z)p(z)dz$
- A better plan, assuming we can sample $\tilde{p}(z)$

  Given independant samples $z^{(l)}$ from $\tilde{p}(z)$

  Estimate $E_{p(z)}[f] \cong \frac{1}{L}\sum_{l=1}^{L} f(z)$

Challenges for sampling

In real problems sampling $p(z)$ is very difficult.

We typically do not know the normalization constant, Z. (So we need to use $\tilde{p}(z)$).

Even if we can draw samples, it is hard to know if (when) they are good, and if we have enough of them.

Evaluating $\tilde{p}(z)$ is generally much easier (although, it can also be quite involved).

Sampling framework

We assume that sampling from $\tilde{p}(z)$ is hard, but that evaluating $\tilde{p}(z)$ is relatively easy.

We also assume that the dimension of z is high, and that $\tilde{p}(z)$ may not have closed form (but we can evaluate it).

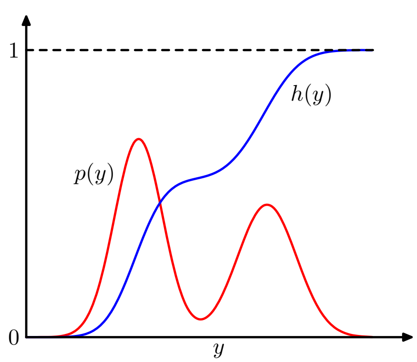
We will develop the material in the context of computing expections, but sampling also supports picking a good answer, such as a MAP estimate of parameters.

Basic Sampling (so far)

- Uniform sampling (everything builds on this)
- Sampling from a multinomial
- Sampling for selected other distributions (e.g., Gaussian)
  - At least, Matlab knows how to do it.
- Sampling univariate distributions using the inverse of the cumulative distribution (recall from HW 2).

Basic Sampling (so far)

- Sampling univariate distributions using the inverse of the cumulative distribution.



Basic Sampling (so far)

- Sampling directed graphical models using ancestral sampling.

Rejection Sampling

Assume that we have an easy to sample function, ,
and a constant, k, where we know that $p(z) \leq k \bullet q(z)$.

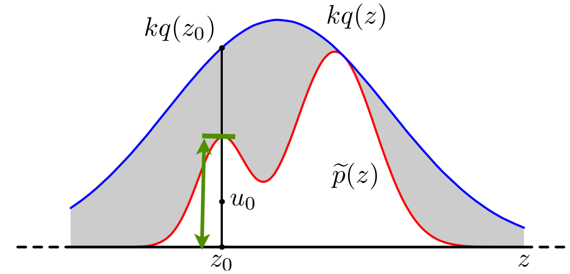
1) Sample $q(z)$

2) Keep samples in proportion to $\frac{p(z)}{k \bullet q(z)}$ and reject the rest.

Rejection Sampling

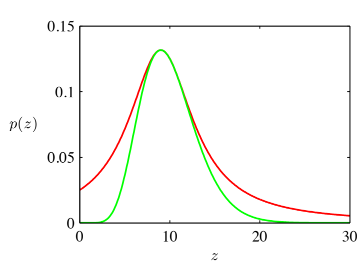
1) Sample $q(z)$

2) Keep samples in proportion to $\frac{p(z)}{k \bullet q(z)}$ and reject the rest.



Rejection Sampling

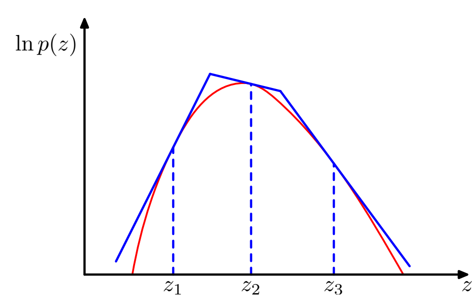
- Rejection sampling is hopeless in high dimensions, but is useful for sampling low dimensional "building block" functions.
- E.G., the Box-Muller method for generating samples from a Gaussian uses rejection sampling.



A second example where a gamma distribution is approximated by a Cauchy proposal distribution.

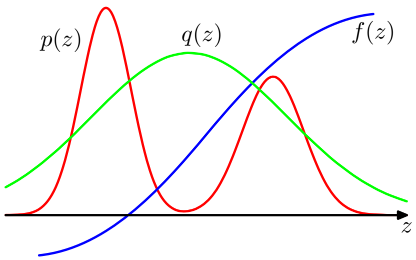
Rejection Sampling

- For complex functions, a good *q()* and *k* may not be available.
- One attempt to adaptively find a good *q()* (see Bishop 11.1.3)



(For log concave functions)

Importance Sampling



Rewrite $E_{p(z)}[f] = \int f(z)\, p(z)\, dz$

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

$$\cong \frac{1}{L} \sum_{l=1}^{L} \frac{p(z^{(l)})}{q(z^{(l)})} f(z^{(l)}) \qquad \text{where samples come from } q(z)$$

Importance Sampling (unnormalized)

$$p(z) = \frac{\tilde{p}(z)}{Z_p} \quad \text{and} \quad q(z) = \frac{\tilde{q}(z)}{Z_q}$$

$$E_{p(z)}[f] \cong \frac{1}{L} \sum_{l=1}^{L} \frac{p(z^{(l)})}{q(z^{(l)})} f(z^{(l)}) \qquad \text{(samples from } q(z^{(l)}) \text{, equivalently, } \tilde{q}(z^{(l)}))$$

$$\cong \frac{Z_q}{Z_p} \frac{1}{L} \sum_{l=1}^{L} \frac{\tilde{p}(z^{(l)})}{\tilde{q}(z^{(l)})} f(z^{(l)})$$

$$= \frac{Z_q}{Z_p} \frac{1}{L} \sum_{l=1}^{L} \tilde{r}_l f(z^{(l)}) \qquad \text{(introducing } \tilde{r}_l = \frac{\tilde{p}(z^{(l)})}{\tilde{q}(z^{(l)})})$$

Importance Sampling (unnormalized)

$$Z_p = \int \tilde{p}(z)\, dz$$

$$\frac{Z_p}{Z_q} = \int \frac{\tilde{p}(z)}{Z_q} dz = \int \frac{\tilde{p}(z)}{\tilde{q}(z)} q(z)\, dz \qquad \text{(because } Z_q = \frac{\tilde{q}(z)}{q(z)})$$

$$\cong \frac{1}{L} \sum_{l=1}^{L} \tilde{r}_l \qquad \text{(samples coming from } \tilde{q}(z^{(l)}))$$

Importance Sampling (unnormalized)

$$E_{p(z)}[f] \cong \frac{Z_q}{Z_p} \frac{1}{L} \sum_{l=1}^{L} \tilde{r}_l f(z^{(l)}) \qquad \text{(samples coming from } \tilde{q}(z^{(l)}))$$

and $\displaystyle \frac{Z_p}{Z_q} \cong \frac{1}{L} \sum_{l=1}^{L} \tilde{r}_l$ (samples coming from $\tilde{q}(z^{(l)})$)

so $\displaystyle E_{p(z)}[f] \cong \frac{\sum_{l=1}^{L} \tilde{r}_l f(z^{(l)})}{\sum_{l=1}^{L} \tilde{r}_l}$ (samples coming from $\tilde{q}(z^{(l)})$)

where $\displaystyle \tilde{r}_l = \frac{\tilde{p}(z^{(l)})}{\tilde{q}(z^{(l)})}$

(from Koller and Friedman)

Importance sampling for graphical models

We know how to sample from directed graphical models
where no variables are observed or conditioned on.

Suppose we want to use sampling to compute $p(Y = y)$.

$$p(Y = y) \cong \frac{1}{L}\sum_{l} I\left(y^{(l)}, y\right) \quad \text{(samples from } p(y)\text{)}$$

$$\text{where } I\left(y^{(l)}, y\right) = \begin{cases} 1 & \text{if } \; y^{(l)} = y \\ 0 & \text{otherwise} \end{cases}$$

(from Koller and Friedman)

Importance sampling for graphical models

We know how to sample from directed graphical models
where no variables are observed or conditioned on.

What about the case of a particular value of a subset of the variables.

EG, we might want to sample: $p\left(Y \mid E = e\right)$
or, we might want to evaluate: $p\left(y = Y \mid E = e\right)$

(from Koller and Friedman)

Importance sampling for graphical models

EG, we might want to sample: $p\left(Y \mid E = e\right)$
or, we might want to evaluate: $p\left(y = Y \mid E = e\right)$

A fool-proof plan is to sample $p(y, e)$, and reject $\text{e} \neq \text{E}$

(Potentially very expensive!)

(from Koller and Friedman)

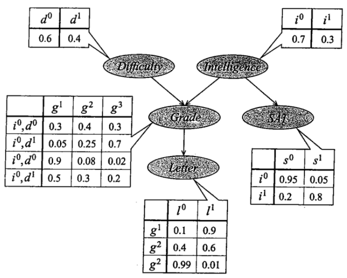
Importance sampling for graphical models

A natural idea is to use ancestral sampling on the graph,
where we set E=e.

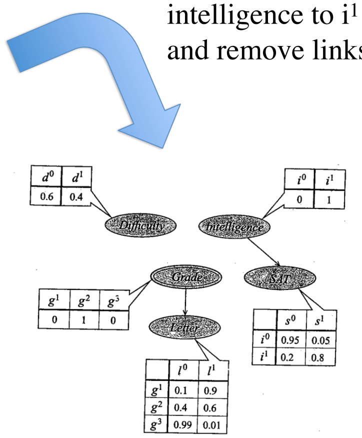
Kollar and Friedman develop this as sampling from the
"mutilated" Bayesian network.

(from Koller and Friedman)

Mutilating graphical models



Set grade to g^2 and intelligence to i^1, and remove links.



(from Kollar and Friedman)

Importance sampling for graphical models

A natural idea is to use ancestral sampling on the graph, where we set E=e.

However, when E=e, this can influence the correct sampling of Y, and we have ignored this!

Instead, we use samples from the mutilated network for the proposal distribution in importance sampling .

(from Koller and Friedman)

Importance sampling for graphical models

$$\frac{p(y|e)}{q(y|e)} = \frac{P_{BN}(y|e)}{P_{MBN}(y|e)} = \frac{P_{BN}(y,e)}{P_{MBN}(y,e)}$$

$$p(y|e) \cong \frac{1}{L}\sum_{l}\frac{P_{BN}(y,e)}{P_{MBN}(y,e)} I(Y=y) \quad \text{(samples from } P_{MBN}(Y,e)\text{)}$$