Sampling based inference

- Resources.
  - Bishop, chapter 11
  - Kollar and Friedman, chapter 12
  - Andrieu et al. (linked to on lecture page)
  - Neal 93 (linked to on lecture page)
- Kollar and Friedman uses "particles" terminology instead of "samples".

Sampling based inference

- Exact inference of complex models in high dimensions is often infeasible.
- Sampling based inference addresses this by drawing samples from the distribution (e.g., posterior over parameters).
- Samples can be used to compute expectations or find the maximum.
- In real-life high dimensional problems, the probability mass is very localized, and samples from the distribution are expected to be from near interesting peaks.

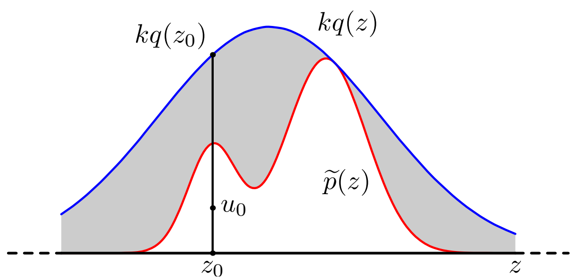
Last time

- Details on the previous slides
- Review of basic sampling already covered
- Rejection sampling
- Started importance sampling

Rejection Sampling

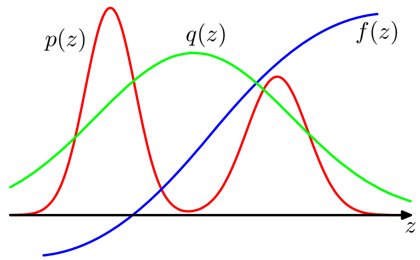
1) Sample $q(z)$

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



Importance Sampling

(computing expectations)



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

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

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

Importance Sampling (unnormalized)

$$p(z) = \frac{\tilde{p}(z)}{Z_p}$$

$$q(z) = \frac{\tilde{q}(z)}{Z_q}$$

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

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

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

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

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

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

Importance Sampling (unnormalized)

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

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

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

(from Kollar 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 $p(Y = y)$.

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

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

(from Kollar 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: $p(Y|E=e)$

or, we might want to evaluate: $p(y=Y|E=e)$

(from Kollar and Friedman)

Importance sampling for graphical models

EG, we might want to sample: $p(Y|E=e)$

or, we might want to evaluate: $p(y=Y|E=e)$

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

(Potentially very expensive!)

(from Kollar and Friedman)

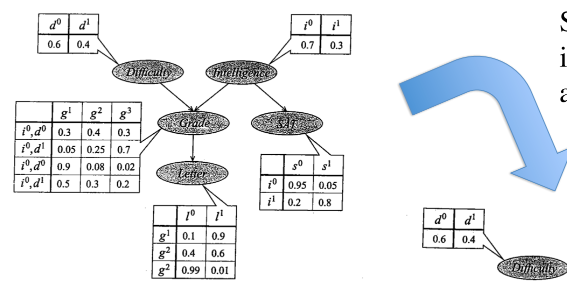
Importance sampling for graphical models

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

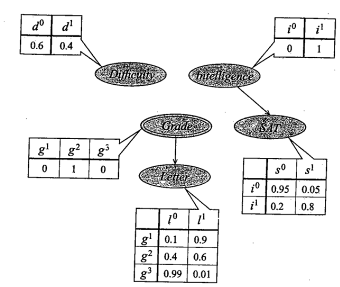
Kollar and Friedman develop this as sampling from the "mutilated" Bayesian network.

(from Kollar 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 Kollar and Friedman)

Importance sampling for graphical models

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

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

Markov chain Monte Carlo methods

- The approximations of expectation so far have assumed that the samples are independent draws.
- This sounds good, but in high dimensions, we do not know how to get **good** independent samples from the distribution.
- MCMC methods drop this requirement.
- Basic intuition
  - If you have **finally** found a region of high probability, stick around for a bit, enjoy yourself, grab some more samples.

Markov chain Monte Carlo methods

- Samples are conditioned on the previous one (this is the Markov chain).
- MCMC is generally a good hammer for complex, high dimensional, problems.
- Main downside is that it is not “plug-and-play”
  - Doing well requires taking advantage to the structure of your problem
  - MCMC tends to be expensive (but take heart---there may not be any other solution, and at least your problem is being solved).

Metropolis Example

We want samples $z^{(1)}, z^{(2)}, \ldots$

Again, write $p(z) = \tilde{p}(z)/Z$

Assume that $q\left(z \middle| z^{(prev)}\right)$ can be sampled easily

Also assume that $q(\ )$ is symmetric, i.e., $q\left(z_A \middle| z_B\right) = q\left(z_B \middle| z_A\right)$

Metropolis Example

While not_bored

{

Sample $q\left(z \middle| z^{(prev)}\right)$

Accept with probability $A\left(z, z^{(prev)}\right) = \min\left(1, \dfrac{\tilde{p}(z)}{\tilde{p}\left(z^{(prev)}\right)}\right)$

If accept, emit z, otherwise, emit $z^{(prev)}$.

}

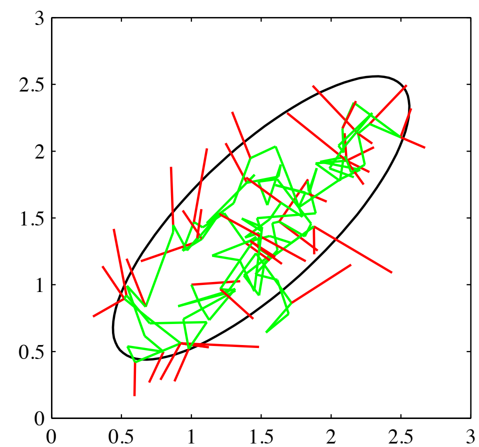
Metropolis Example

Note that

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

We do not need to normalize $p(z)$

Metropolis Example



Green follows accepted proposals
Red are rejected moves.

Markov chain view

Denote an initial probability distribution by $p\left(z^{(1)}\right)$

Define transition probabilities by:

$T\left(z^{(prev)},z\right)=p\left(z\middle|z^{(prev)}\right)$ (a probability distribution)

$T=T_m(\ )$ can change over time, but for now, assume that it it is always the same (homogeneous chain)

A given chain evolves from a sample of $p\left(z^{(1)}\right)$, and is an instance from an essemble of chains.

Stationary Markov chains

- Recall that our goal is to have our Markov chain emit samples from our target distribution.
- This implies that the distribution being sampled at time *t+1* is the same as that of time *t* (stationary).
- If our stationary (target) distribution is *p()*, what do the transition probabilities look like?

Detailed balance

- Detailed balance is defined by:

$$p(z)T(z,z')=p(z')T(z',z)$$

(We assume that $\mathrm{T}(\bullet)>0$)

- Detailed balance is a sufficient condition for a stationary distribution.
- Detailed balance is also referred to as reversibility.

Detailed balance implies stationary

$p(z)=\sum_{z'}p(z')\,T(z',z)$ (marginalization)

If we have detailed balance, then

$$p(z)T(z,z')=p^{(prev)}(z')T(z',z)$$

So,

$$p(z)=\sum_{z'}p^{(prev)}(z')\,T(z',z)=\sum_{z}p^{(prev)}(z)T(z,z')=p^{(prev)}(z')$$

Hence, detailed balance implies the distribution is stationary.