Sampling Methods

- For most prob. models, exact inference is intractable.
  - * VB one approach

Monte Carlo approaches today.

Note: though post. dist itself may be of interest, usually expectations w.r.t. to posterior dist. are really of interest

Goal:

Compute

if z discrete, sum instead.

$$\mathbb{E}[f] = \int f(z)p(z)\,dz$$

Examples: $f(z) = z \rightarrow$ posterior mean
$f(z) = (z - \mathbb{E}[z])^2 \rightarrow$ post. variance
$f(z) = \mathbb{I}(a \leq z_i \leq b) \rightarrow$ post. reg. of conf.

etc.

Sampling: general idea;

1) Draw samples $z^{(l)}$, $l = 1 \ldots L$ iid $\sim p(z)$

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

$$\mathbb{E}[f] = \int f(z)\,p(z)dz \approx \frac{1}{L}\sum_{l=1}^{L} f(z^{(l)})$$

Note, this estimator is unbiased as

$$\mathbb{E}[\hat{f}] = \mathbb{E}[f]$$

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

$$\cancel{\mathbb{E}[\hat{f}] = \int p(z)\frac{1}{L}\sum_{l=1}^{L} f(z^{l})\,dz}$$

$$\cancel{= \frac{1}{L}\sum_{l=1}^{L}\int p(z)\,f(z^{(l)})\,dz}$$

$$\mathbb{E}[\hat{f}] = \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\, f(z^{(l)})$$

but $z^{(l)} \sim$ iid p's
so $\mathbb{E}\, f(z^{i}) = \mathbb{E}\, f(z^{j})$

$$\Rightarrow \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\, f(z)$$

$$= \mathbb{E}\, f(z)$$

$$\cancel{\mathrm{var}[\hat{f}] = \mathbb{E}\left[\left(\hat{f} - \mathbb{E}[\hat{f}]\right)^2\right]}$$

$$\cancel{= \mathbb{E}[\hat{f}^2] - 2\,\mathbb{E}[\hat{f}]\,\mathbb{E}[f] + \mathbb{E}[\hat{f}]^2}$$

$$\cancel{= \mathbb{E}[\hat{f}^2] - \mathbb{E}[f]^2}$$

$$\cancel{= \mathbb{E}\left[\left(\frac{1}{L}\sum_{l=1}^{L} f(z^{(l)})\right)^2\right] - \mathbb{E}[f]^2}$$

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

$$= \frac{1}{L^2}\sum_{l=1}^{L}\mathrm{var}\left[f(z^{(l)})\right]$$

z^{l} iid !
variances of iid R.V.'s add.

$$= \frac{1}{L}\,\mathrm{var}[f(z)]$$

$$= \frac{1}{L}\,\mathbb{E}\left[\left(f - \mathbb{E}[f]\right)^2\right]$$

And the variance of the estimator

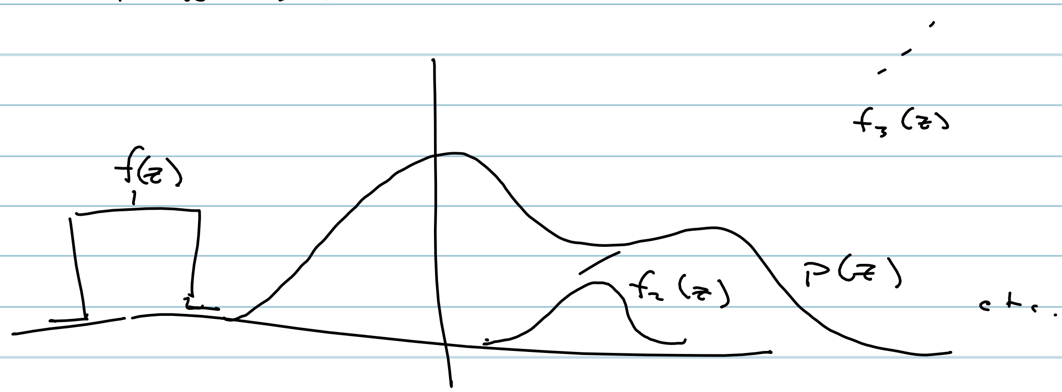
$$\mathrm{Var}[\hat{f}] = \frac{1}{L}\mathbb{E}[(f - \mathbb{E}(f))^2]$$

needs z_ℓ iid

is the variance of the function f and independent of the dimensionality of f!

- implication: relatively small number of samples can do a good job of approximating this expectation if the function f is low variance.

- Problems
  1) $f(z)$ might be small where $p(z)$ is large & vice versa
  2) $z^{(\ell)}$'s might not truly be independent yielding an effective sample size that is too small



- Sampling in Graphical Models

If $p(z)$ given by G.M. (directed) and no variables are observed then
- ancestral sampling works

$$p(z) = \prod_{i=1}^{M} p(z_i \mid pa_i)$$

Pass through graph sampling parents first.

What if nodes are observed?

- Inefficient but intuitive approach: sample all vars up to an observed z_i, if when sampling z_i the sampled value matches the observed value, keep the whole sample, otherwise discard everything and start over (a form of importance sampling)
  * This approach draws samples from the posterior because it samples from the joint and discards those that disagree with the observed data
  * This approach is highly inefficient in most cases (large model, high dimensions, few observations at leaf nodes)

- Undirected graphs? : no 1-pass sampling alg. Gibbs must be employed.

Important { Sampling from a marginal dist: if we can sample from a joint dist. $p(u,v)$ and need samples from $p(u)$ it suffices to sample the joint and discard v parts.

Basic Sampling Algs

In order to sample from various distributions (complicated ones) we need to be able to first sample from simple ones. To do this we will use transformations and other tricks to generate pseudo-random numbers starting from $U(0,1)$
" " "

$$\frac{d}{dy} \exp(-\lambda y) = -\lambda \exp(-\lambda y)$$

$$\Rightarrow \int \lambda \exp(-\lambda y)\, dy = -\exp(-\lambda y)$$

$$\int_0^{10} x^2\, dx = \frac{1}{3} x^3 \Big|_0^{10} = \frac{1000}{3} =$$

$$z = F(y) = 1 - \exp(-\lambda y) \quad \Rightarrow \quad z - 1 = -\exp(-\lambda y)$$

$$\cancel{\ln(z-1)}$$

$$1 - z = \exp(-\lambda y)$$

$$\ln(1-z) = -\lambda y$$

$$y = -\frac{\ln(1-z)}{\lambda} = F^{-1}(z)$$

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

$$\frac{dy}{dz} = +\frac{1}{\lambda(1-z)}$$

$U(0,1)$ pseudo-random #'s are generally available on all OS's and in most software packages and generally derive from the linear congruential generator

$$X_{n+1} = (a X_n + b) \bmod m$$

where m is the maximum # of random numbers that can be generated. $\exists$ good choices for a & b.

Fast and improved PRNG's exist and include the Mersenne twister, etc.

Starting with $z \sim U(0,1]$ we can transform z using $f(\cdot)$ s.t. $y = f(z)$. The dist of y is given by the transformation rule:

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

where, of course, here $p(z) = 1$

Goal: choose f s.t. the resulting y have the "correct" dist. $p(y)$

Good choice of transformation: inv-CDF

Example: (Exponential)

$$p(y) = \lambda \exp(-\lambda y)$$

$$F(y) = \int_0^y p(y')dy' = -\exp(-\lambda y)\Big|_0^y$$

$$= 1 - \exp(-\lambda y)$$

Let $z = F(y) = 1 - \exp(-\lambda y)$

$z \in [0,1]$ because $F(y)$ is CDF of y

$$y_1 = z_1 \left(\frac{-2 \ln z_1}{r^2} \right)^{1/2}$$

$$-y_1^2 = z_1 \left(\frac{-2 \ln z_1}{r^2} \right)$$

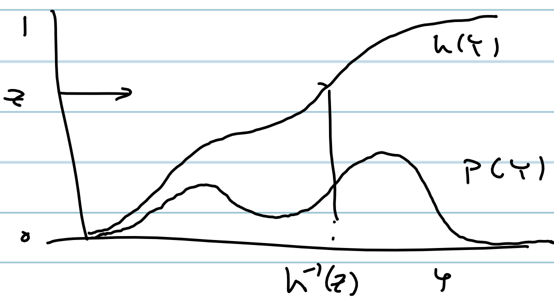
$$-\frac{y_1^2}{2} = \frac{z_1 \ln z_1}{r^2} \quad , \quad -\frac{y_1^2}{2} = \ln \left(z_1^{\frac{z_1}{r^2}} \right)$$

$$\exp \frac{-y_1^2}{2} = z_1^{\frac{z_1}{r^2}}$$

Solve for $y = F^{-1}(z) = -\lambda^{-1}\ln(1-z)$ and check

$$p(y) = p(z)\left|\frac{dz}{dy}\right| = 1 \cdot -\lambda \cdot (-\exp -\lambda y)$$
$$= \lambda \exp -\lambda y$$

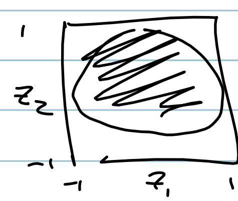
means: choosing $z \sim U(0,1)$ and transforming z via $-\lambda^{-1}\ln(1-z)$ yields and $\lambda \exp(-\lambda y)$ R.V. distributed.



Box-Muller for Gaussian R.V.'s

Recipe:

- Generate 2 r.v.'s $z_1, z_2 \in [-1, 1]$ i.i.d.
  - i.e. generate 2 $U(0,1)$ R.V.'s, $(*2 - 1)$
- Marginalize joint to



$$p(z_1, z_2) = 1/\pi$$

by rejecting samples outside $z_1^2 + z_2^2 \le 1$

By transforming

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

$$p(y_1, y_2) = p(z_1, z_2)\left|\frac{\partial(z_1, z_2)}{\partial(y_1, y_2)}\right|$$
$$= \left[\frac{1}{\sqrt{2\pi}}\exp\left(-y_1^2/2\right)\right]\left[\frac{1}{\sqrt{2\pi}}\exp\left(-y_2^2/2\right)\right]$$

i.e. 2 $N(0,1)$ R.V.'s !!

Remember: if $y \sim N(0, 1)$
$\sigma y + \mu \sim N(\mu, \sigma^2)$

because
$\mathbb{E}[\sigma y + \mu] = \mathbb{E}[\sigma y] + \mu = \mu$

and
$Var[\sigma y + \mu] = \sigma^2 Var[y] = \sigma^2$

So $N(\mu, \sigma^2)$ RV's can be sampled easily.

As well if $\vec{z} \sim N(\vec{0}, I)$ then

$$\vec{y} = \vec{\mu} + L\vec{z} \quad \text{where} \quad \Sigma = LL^T$$

$$\Rightarrow \vec{y} = \vec{\mu} + L\vec{z} \sim N(\vec{\mu}, \Sigma) \quad \text{because}$$

$$\mathbb{E}[\vec{\mu} + L\vec{z}] = \vec{\mu}$$

$$Cov[\vec{\mu} + L\vec{z}] = L\, Cov(\vec{z})\, L^T = LL^T = \Sigma$$

So Multivariate Gaussian RV's can be generating easily starting with uniform $(0,1)$ RV's

Transformation approach limited to analytically tractable CDF's and analytic inversions.

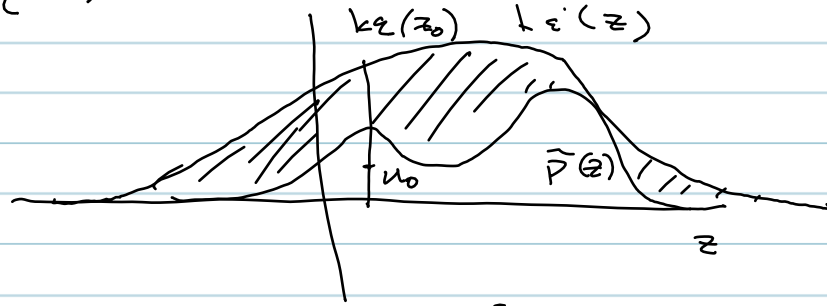
<u>Rejection Sampling</u> (very general, efficient (usually) in low-D)

Suppose we want to sample from $p(z)$, but, like usual, we only know $p(z)$ up to a normalizing constant

$$p(z) = \frac{1}{Z_p} \tilde{p}(z)$$

where $\tilde{p}(z)$ is easily evaluated but Z_p is unknown.

Rejection sampling involves a simpler "proposal dist" $q(z)$



which is easy to sample from. Also a constant k must be found s.t. $kq(z) \geq \tilde{p}(z) \; \forall z$.

Recipe:
- draw z_0 from $q(z)$
- draw u_0 from $U[0, kq(z_0)]$
- keep sample z_0 if $u_0 \leq \tilde{p}(z_0)$ otherwise reject z_0
- repeat

The probability of accepting a sample z is

$$p(\text{accept}) = \int \underbrace{\{\tilde{p}(z)/kq(z)\}}_{\text{P accepting}} \underbrace{q(z)}_{\text{prob choosing } z} dz$$

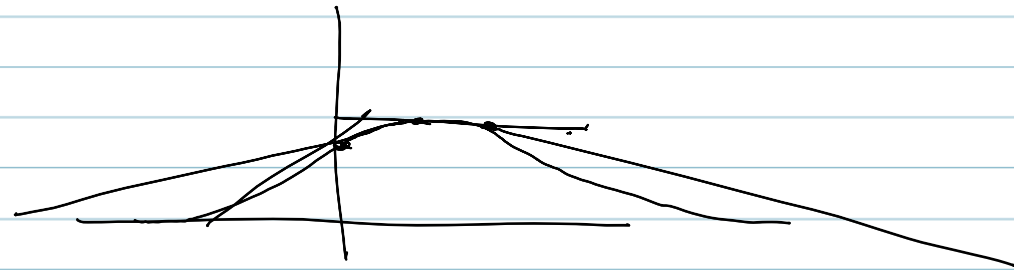
$$= \frac{1}{k} \int \tilde{p}(z) \, dz$$

Which means that we want to make k as small as possible

Canonical Example
Sampling from Gamma dist using Cauchy proposal.

Extensions Adaptive Rejection Sampling

ARS: if $p(z)$ is log concave then an adaptive shell consisting of piecewise linear functions can be constructed



Every time $p(z)$ is evaluated a new point is added to the piecewise linear envelope.

<u>Rejection</u> <u>sampling</u> suffers in <u>high dim.</u>

Illustrative example
Sample from $z \sim \mathcal{N}(\vec{0}, \sigma_p^2 I)$
Use
$q(z) = \mathcal{N}(\vec{0}, \sigma_q^2 I)$ as proposal
(ie. well matched distribution)

Clearly $\sigma_q^2 > \sigma_p^2$ for rejection sampling, and, $k\,q(z) \geq p(z) \Rightarrow k^{opt} = (\sigma_q/\sigma_p)^D$ in D-dim case (0 max, ratio of det's)

Unfortunately σ_q/σ_p is to the power D so even a small and the acceptance rate goes $O(1/k)$ so the acceptance rate goes $O(\exp(-D))$ which is bad.

Importance Sampling

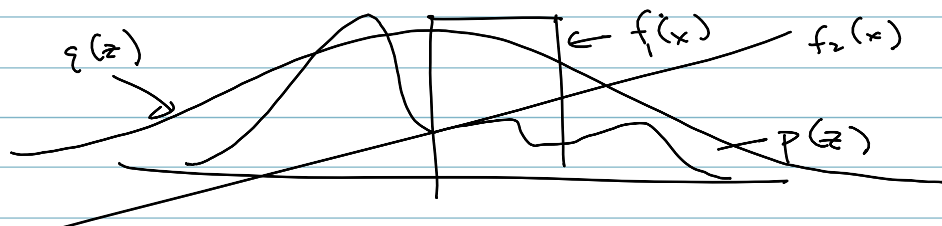
So far we have been interested in sampling but usually we are interested in integrating. What if we skip sampling and directly integrate? Assume $p(z)$ is hard to sample from but easy to evaluate.

Intuition: grid the space of z uniformly and evaluate

$$\mathbb{E}[f] = \int f(z)\,p(z)\,dz \approx \sum_{\ell=1}^{L} p(z^{(\ell)})\,f(z^{(\ell)})$$

High dimensions require an exponential number of z's. Many will be in regions where $p(z)$ is small and thus they are largely irrelevant to approx. $\mathbb{E}[f]$.

Instead what if we sample from $q(z)$ from which samples are easy to draw?



with $\{z^{(\ell)}\} \sim q$ we can write

$$\mathbb{E}[f] = \int f(z)\,p(z)\,dz$$
$$= \int f(z)\,\frac{p(z)}{q(z)}\,q(z)\,dz$$
$$\approx \frac{1}{L}\sum_{\ell=1}^{L} \frac{p(z^{(\ell)})}{q(z^{(\ell)})}\,f(z^{(\ell)})$$

where $r_\ell = \dfrac{p(z^{(\ell)})}{q(z^{(\ell)})}$ are called "importance weights".