**Sampling Methods**

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

**Monte Carlo approaches today.**

Note: though posterior itself may not be of interest, usually expectations w.r.t. posterior dist. are really of interest.

**Goal:** Compute if a discrete, sample

\[ E[f] = \int f(z) p(z) \, dz \]

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

etc.

**Sampling:** general idea:

1) Draw samples \( z^{(1)}, \ldots, z^{(L)} \sim p(z) \)

2) Approximate

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

\[ E[f] \approx \frac{1}{L} \sum_{l=1}^{L} f(z^{(l)}) \]

Note, this estimator is unbiased as

\[ E[\hat{f}] = E[f] \]
\[ \tilde{f} = \frac{1}{L} \sum_{i=1}^{L} f(z^{(i)}) \]

\[ \mathbb{E} [\tilde{f}] = \frac{1}{L} \sum_{i=1}^{L} \mathbb{E} [f(z^{(i)})] \]

so \( \mathbb{E} [\tilde{f}] = \mathbb{E} [f(z)] \)

\[ \mathbb{V} [\tilde{f}] = \mathbb{V} [\frac{1}{L} \sum_{i=1}^{L} f(z^{(i)})] \]

\[ = \frac{1}{L^2} \sum_{i=1}^{L} \mathbb{V} [f(z^{(i)})] \]

\[ = \frac{1}{L} \mathbb{V} [f(z)] \]

\[ = \frac{1}{L} \mathbb{E} [(z - \mathbb{E}[z])^2] \]
And the variance of the estimator

$$\text{Var}[\hat{f}] = \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - \hat{f})^2$$

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) $f(x)$ might be small where $p(x)$ is large or vice versa
  2) $x_i$'s might not truly be independent yielding an effective sample size that is too small

- Sampling in Graphical Models

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

$$p(x) = \prod_{i=1}^{N} p(x_i | \text{parents of } x_i)$$

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 is worst cases (large models, high dimensions, few observations at leaf nodes)

- Undirected graphs? no 2-pass sampling alg. Gibbs must be employed.

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

Basic Sampling Alg.

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 \( U(0,1) \)
\[ U(0,1) \text{ pseudo-random numbers are generally available on all OSs and in most software packages and generally derive from the linear congruential generator:} \]

\[ X_{n+1} = (aX_n + b) \mod m \]

where \( m \) is the maximum # of random numbers that can be generated, \( a \) and \( b \) are choices for \( a \neq b \). Fast and improved PRNG's exist - such as the Mersenne Twister, etc.

Starting with \( z \sim U(0,1) \) we transform \( z \) using \( f(z) \) 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)

\[ p(y) = 2 \exp(-2y) \]

\[ F(y) = \int_0^y p(x)dx = \exp(-2x) \bigg| _0^y = 1 - \exp(-2y) \]

Let \( z = F(y) = 1 - \exp(-2y) \)

\( z \in [0,1] \) because \( F(y) \) is CDF of \( Y \)
\[ \gamma_1 = z_1 \left( -2 \frac{1 - z_1}{r^2} \right)^{1/2} \]

\[ \gamma_2 = z_1 \left( \frac{-2 \left( 1 - z_1 \right)}{r^2} \right) \]

\[ -\frac{\gamma_1}{2} = \frac{z_1 \left( 1 - z_1 \right)}{r^2} \]

\[ -\frac{\gamma_2}{2} = \ln \left( z_1 \frac{z_2}{r^2} \right) \]

\[ \exp \left( -\frac{\gamma_1}{2} - \frac{\gamma_2}{2} \right) = z_1 \frac{z_2}{r^2} \]
Solve for $y = F^{-1}(x) = -\ln(x(1-x))$ and check

$p(y) = p(z) \left| \frac{dz}{dy} \right| = 1 \cdot x \cdot (-e^{-x^2}) = x \exp(-x^2)$

Near choosing $z \sim U(0,1)$ and transforming $z \sim -\frac{1}{2} \ln(1-x)$ yields and $2 \exp(-x^2)$ R.U. distribution.

\[ \text{Box-Muller for Gaussian R.U.'s} \]

Recipe:
- Generate $z \sim U(0,1)$ R.U.'s \( z_1, z_2 \in [0,1] \)
- is. generate $z \sim U(0,1)$ R.U.'s \(*z^2 - 1*\)
- Maximum size joint to

\[ p(z_1, z_2) = \frac{z_1}{\pi} \]

by rejecting samples outside \( z_1^2 + z_2^2 \leq 1 \)

\[ B_1 \text{ transform:} \]

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

\[ y_2 = z_2 \left( -\frac{2 \ln z_2}{z_2^2} \right)^{1/2} \]

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

i.e. $z \sim U(0,1)$ R.U.'s!!
\[ \text{Remember: if } \gamma \sim N(0, I), \]
\[ \sigma \gamma + \mu \sim N(\mu, \sigma^2) \]
\[ \text{because } E[\sigma \gamma + \mu] = E[\sigma \gamma] + \mu = \mu \]
\[ \text{and } \text{Var}[\sigma \gamma + \mu] = \sigma^2 \text{Var}[\gamma] = \sigma^2 \]
\[ \therefore N(\mu, \sigma^2) \text{ RVs can be sampled easily.} \]

As well if \( z \sim \mathcal{N}(0, I) \) then
\[ y = \mu + Lz \quad \text{where} \quad z = LL^T \]
\[ \Rightarrow y \sim \mathcal{N}(\mu, \Sigma) \quad \text{because} \]
\[ E[\mu + Lz] = \mu \]
\[ \text{Cov}[\mu + Lz] = L \text{Cov}(z) L^T = LL^T = \Sigma \]

So Multivariate Gaussian RVs can be generated easily starting with uniform \((0, 1)\) RVs.

Transformation approach limited to analytically tractable CDFs and analytic inversion.

**Rejection Sampling** (very general, efficient (usually) in low \( D \)),

Suppose we want to sample from \( p(z) \), but, like usual, we only know \( \hat{p}(z) \) up to a normalising constant
\[ p(z) = \frac{1}{Z_p} \hat{p}(z) \]
\[ w \text{ we } \hat{p}(z) \text{ is easily evaluated but } Z_p \text{ is unknown.} \]
Rejection sampling involves a simpler proposal dist $q(z)$ which is easy to sample from. Also a constant $k$ must be found such that $k q(z) \geq \tilde{p}(z) \forall z$.

**Recipe:**
- Draw $z_0$ from $q(z)$
- Draw $u_0$ from $U[0, k q(z_0)]$
- Keep sample $z_0$ if $u_0 \leq \tilde{p}(z_0)$ otherwise reject $z_0$
- Repeat

The probability of accepting a sample $z_0$ is

$$p(\text{accept}) = \frac{\int p(z) q(z_0) \, d\bar{z}}{\int q(z) \, d\bar{z}} = \frac{1}{k} \int \tilde{p}(z) \, d\bar{z}$$

which means that we want to make $k$ as small as possible.

**Canonical Example**
Supposing $\text{Gamma dist using Canonical proposal.}$

**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.

Rejection sampling suffers in high dim.

Illustrative example

Sample from $z \sim N(\hat{\sigma}, \sigma^2 I)$

Use

$q(z) = N(\hat{\sigma}, \sigma^2 I)$ as proposal (i.e., well smoothed distribution)

Clearly $\sigma^2 > \sigma^2_p$ for rejection sampling, so

$k = (\sigma^2 / \sigma^2_p)^D$ in D-site case (ratio of det's)

Unfortunately $\sigma^2 / \sigma^2_p$ 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

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

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

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

![Graph showing q(z) and p(z) with importance sampling]

With \( \{z^{(1)}\} \sim q \) we can write

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

where \( z^{(1)} = q(z^{(1)}) \) are called "importance weights".