In regression and other settings the primary inference objectives can often be expressed in the form

$$E(f) = \int f(z) \, p(z) \, dz \qquad (1)$$

For instance in regression we might have

$$z = [\beta, \sigma^2] \, , \quad f(z) \sim p(Y_{new} | X_{new}, z)$$

with

$$p(z) = p(z | X, Y) \quad \text{the posterior distribution}$$

over the model parameters, meaning that

$$E(f) = \int p(Y_{new} | X_{new}, z) \, p(z | X, Y) \, dz$$

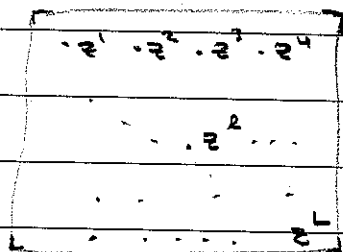is the posterior predictive distribution of Y_{new} given X_{new} and the posti distribution $p(z | X, Y)$

Today: approaches for generating "samples" from $p(z)$ so that (1) can be approximated as

$$E(f) \approx \frac{1}{L} \sum f(z^{(\ell)})$$

with $z^{(\ell)} \sim p(z)$

First intuitive (but bad) approach:
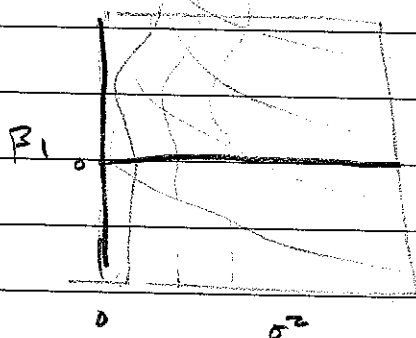
Grid the domain of $p(z)$

consider $z \in \mathbb{R}^2$

and compute

$$\mathbb{E}(f) \approx \sum_{\ell=1}^{L} p(z^{(\ell)}) f(z^\ell)$$

ie. Bayesian regression through origin with prior like:

$$p(z^\ell | \lambda, \mu_{\beta}, \sigma_{\beta}^2) = \lambda \exp(-\gamma \sigma^2)$$
$$\times N(\beta_1 | \mu_{\beta_1}, \sigma_{\beta}^2)$$

But gridding requires the number of points to grow exponentially in the dimension of z

Better: choose points $z^{(\ell)}$ only where $p(z)$ large* ... but how?

* ideally where $f(z) p(z)$ large!

One way : Importance sampling

Assume that we have $q(z)$ which is easy to draw samples from
i.e. $q(z) = N(z | \mu_q, \Sigma_q)$

We can rewrite

$$E(f) = \int f(z) \, p(z) \, dz$$

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

$$\approx \sum_{\ell=1}^{L} \frac{p(z^\ell)}{q(z^\ell)} f(z^\ell)$$

where $z^\ell \sim q$ and
$$r_\ell = \frac{p(z^\ell)}{q(z^\ell)}$$
are known as importance weights

Remember!

1) We're doing this because $p(z)$ is difficult to sample from:
   When does this happen?

   - When parametric form isn't easily transformed or doesn't have an easy analytic description.

   eg. $p(z | X, Y) = \dfrac{p(z | X, Y) \, p(z)}{\int p(z | X, Y) \, p(z) \, dz}$

What if we could get away with
knowing $p(z)$ up to a normalizing
constant only? ie

$$p(z \mid X, Y) \propto p(z \mid X, Y) p(z)$$

then we could, for instance, do posterior
inference without needing to be able to
compute the posterior normalizing constant.
   To be most general let's consider

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

$\tilde{q}$ ie proposal
only known up to a
normalizing constant too

Remember that we can sample from
$q(z)$ even if we only know it up to
a normalizing constant, ie. $q(z) = \frac{\tilde{q}(z)}{Z_q}$,
sampling from $\tilde{q}(z)$ is the same as sampling
from $q(z)$
   In this situation we can write

$$\mathbb{E}(f) = \int f(z) \, p(z) \, dz$$
$$= \int f(z) \frac{\frac{\tilde{p}(z)}{Z_p}}{\frac{\tilde{q}(z)}{Z_q}} q(z) \, dz$$
$$= \frac{Z_p}{Z_q} \int f(z) \frac{\tilde{p}(z)}{\tilde{q}(z)} q(z) \, dz$$

$$\approx \frac{Z_q}{Z_p} \frac{1}{L} \sum_{\ell=1}^{L} \tilde{r}_\ell \, f(z^\ell)$$

where

$$\tilde{r}_\ell = \frac{\tilde{p}(z^\ell)}{\tilde{q}(z^\ell)} \quad \text{and} \quad z^\ell \sim q$$

But now what about $\frac{Z_q}{Z_p}$?

$$\frac{Z_p}{Z_q} = \frac{1}{Z_q} \int \hat{p}(z) \, dz$$

$$Z_q = \frac{\tilde{q}(z)}{q(z)} = \text{is constant} \quad \text{so}$$

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

$$\approx \frac{1}{L} \sum_{\ell=1}^{L} \tilde{r}_\ell$$

So $\quad \mathbb{E}(f) \approx \sum w_\ell f(z^\ell) \quad$ where $z^\ell \sim q$

and $\quad w_\ell = \frac{\tilde{r}_\ell}{\sum_m \tilde{r}_m} = \frac{\tilde{p}(z^\ell)/\tilde{q}(z^\ell)}{\sum_m \tilde{p}(z^m)/\tilde{q}(z^m)}$

So now, we can do posterior inference numerically with out computing any integrals of the Bayesian normalization kind explicitly. What's wrong?

if $\tilde{q}$ doesn't align well (tough to know in advance), most weights will be small and the approximation will be dominated by a few (or worse $\sim_0$ samples) -- no diagnostics to help

A solution : get better samples from $p(z)$ or $\tilde{p}(z)$. One way (not without it's own set of problems) MCMC

Simplest recipe : Metropolis algorithm

Ingredients : $\tilde{p}(z)$, distribution to sample from

$q(z|z^\tau)$ "proposal" distribution

with

$$q(z|z^\tau) = q(z^\tau | z)$$

ie. symmetric

Algorithm

Set $\tau = 0$, pick $z^{\tau=0}$ arbitrarily in domain

Repeat Forever $\Bigg\{$ Propose z^* from $q(z^*|z)$

"Accept" z^* w.p.

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

Set $\tau = \tau + 1$

If z^* "accepted" set

$$z^\tau = z^*$$

otherwise set

$$z^\tau = z^{\tau-1}$$

Note: if ratio is > 1 sample is always accepted

Claim $\{z^\tau\}_{\tau=m}^{T}$ are samples from $\tilde{p}$ and correspondingly from p as well.

Justification of Metropolis sampling

Realize: $z^0, \ldots z^{\tau} \ldots z^T$ $T \to \infty$ forms
a Markov chain with a given transition
function

$$z^{\tau+1} = \begin{cases} z^* & w.p. & A(z^*, z^{\tau}) \\ \\ z^{\tau} & w.p. & 1 - A(z^*, z^{\tau}) \end{cases}$$

Remember Markov chains have "equilibrium"
distributions.

Markov chain (1st order): A first order Markov
chain is defined to be a series of RV's
$z^1, \ldots, z^{\mu}$ s.t.

$$p(z^{\mu+1} | z^1, \ldots, z^{\mu}) = p(z^{\mu+1} | z^{\mu})$$

We can specify a Markov chain by
giving $p(z^0)$ and the transition probabilities

$$T_{\mu}(z^{\mu} | z^{\mu+1}) = p(z^{\mu+1} | z^{\mu})$$

A Markov chain is homogenous if the
transition probabilities are the same for
all μ.

The marginal probability of a particular variable $z^{(m+1)}$ can be expressed in terms of the marginal probability of the previous variable in the chain by

$$p(z^{(m+1)}) = \sum_{z^m} p(z^{(m+1)} | z^m) p(z^m)$$

A dist $p^*(z)$ is invariant or stationary w.r.t the Markov chain if each step in the chain leaves it unchanged

For a homogenous Markov chain with transition probs $T(z', z)$, the dist $p^*(z)$ is invariant if

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

Intuition

$$z = T z \quad \text{is eigen vector of } T \text{ with eigenvalue } 1$$

A sufficient (but not necessary) condition for ensuring that $p^*(z)$ is "invariant" under the Markov chain is to choose transition probabilities to satisfy "detailed balance", defined by

$$p^*(z) \, T(z, z') = p^*(z') \, T(z', z)$$

To show that a transition probability that satisfies detailed balance w.r.t. to a particular dist will leave that dist. invariant note that

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

Starting here

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

we get

$$= p^*(z) \sum_{z'} p(z' | z)$$

$$= p^*(z)$$

Remember the goal: construct a Markov chain s.t. "running" the Markov chain yields samples from a $p^*(z)$ of our choice.

This means picking $T(z', z)$ s.t. $p^*(z)$ is the distribution we're interested in drawing samples from. (at a minimum)

Metropolis Algorithm

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

We have to show that $p(z)$ is the invariant distribution of the Markov chain defined by $T(z^{\tau+1}|z^\tau)$

$$= \begin{cases} z^*, & z^* \sim q(z^*|z^\tau) \quad \text{w.p. } \min\left(1, \frac{\tilde{p}(z^*)}{\tilde{p}(z^\tau)}\right) \\ z^\tau & \text{w.p. } 1 - \min\left(1, \frac{\tilde{p}(z^*)}{\tilde{p}(z^\tau)}\right) \end{cases}$$

We can do this by demonstrating that the Markov chain defined by the Metropolis algorithm satisfies detailed balance, ie.

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

$$p^*(z) \, q(z'|z) \, A(z'|z)$$

$$= \min\left(p^*(z) \, q(z'|z), \frac{p(z) q(z'|z) p(z')}{p(z)}\right)$$

$$= \min\left(p(z) q(z'|z), q(z'|z) p(z')\right)$$

$$= \min\left(p(z) q(z'|z), p(z) q(z'|z)\right)$$

but $q(z'|z) = q(z|z')$ by assumption and

$$= p(z') q(z|z') \min\left(1, \frac{p(z)}{p(z')}\right)$$

$$= p(z') \, q(z|z') \, A(z|z') \quad = T(z', z)$$

which shows that $p(z)$ is an invariant dist. of the Metropolis Alg.

$$T(z, z') \qquad P^*(z)$$

$$z$$

$$z' \quad \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$$

$$T(z', z) \qquad P^*(z')$$

$$z'$$

$$= \begin{bmatrix} a & c \\ b & d \end{bmatrix} \qquad \begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$$

$$P_1 a + P_2 b = P_1 a + P_2 c$$

$$P_1 c + P_2 d = P_1 b + P_2 d$$

$$a + c = 1 \qquad a + b = 1$$

$$b + d = 1 \qquad c + d = 1$$

$$b = c \quad , \quad c = b \qquad /$$

$$T(z,z') \qquad \vec{p}(z)$$

$$z' \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \\ P_3 \end{bmatrix}$$

$$= z \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \\ P_3 \end{bmatrix} \qquad s.t. \qquad \begin{array}{l} a+b+c=1 \\ d+e+f=1 \\ g+h+i=1 \end{array}$$

$$\begin{array}{c} a-d \\ a+d+g=1 \\ b+e+h=1 \\ c+f+i=1 \end{array}$$

$$aP_1 + bP_2 + cP_3 = aP_1 + dP_2 + gP_3$$
$$dP_1 + eP_2 + fP_3 = bP_1 + eP_2 + hP_3$$
$$gP_1 + hP_2 + iP_3 = cP_1 + fP_2 + iP_3$$

$$bP_2 + cP_3 + dP_1 = bP_1 + dP_2 + gP_3$$

$$b(P_2 - P_1) = dP_2 + gP_3 + cP_3 + dP_1$$

$$b = \frac{dP_2 + gP_3 + cP_3 + dP_1}{P_2 - P_1}$$

$$T_p = T'_p$$

if T orthogonal
and symmetric the—
$$T^{-1} = T' \qquad \text{and} \qquad T^{-1} T_p = T^{-1} T_p$$
$$p = p$$

We also require that for $m \to \infty$, the dist $p(z^{(m)})$ converges to the desired dist $p^*(z)$ irrespective of choice of $p(z^{(0)})$

This prop. is called ergodicity and the invariant dist of an ergodic Markov chain is known as its equilibrium dist. An ergodic Markov chain can have only one equilibrium dist.

It can be shown that a homogeneous Markov chain will be ergodic subject to weak restrictions on the invariant dist. and the transition probs.

R — Fill

The Metropolis Algorithm is
1) homogeneous
2) ergodic
3) and has p^* as its invariant dist.

It is a general Markov chain transition rule that is parameterized by the distribution of interest $p(z)$ and results in an ergodic Markov chain whose equilibrium dist is $p(z)$.

Uses: sample from posterior dist for use in Monte Carlo integrals like posterior predictive and/or posterior inference.

Metropolis Hastings : q not symmetric

$$p(z) \, q(z'|z) \, A(z',z) = p(z) \, q(z'|z) \min\left(1, \frac{p(z')p(z|z')}{p(z)q(z'|z)}\right)$$

$$= \min\left(p(z) \, q(z'|z) \, , \, p(z') \, p(z|z')\right)$$

$$= \min\left(p(z|z') \, p(z') \, , \, p(z) \, q(z'|z)\right)$$

$$= p(z') \, q(z|z') \min\left(1, \frac{p(z) \, q(z'|z)}{p(z') \, q(z|z')}\right)$$

$$= p(z') \, q(z|z') \, A(z|z') \qquad \checkmark$$