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

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

For instance in regression we might have

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

with

$p(z) = p(z \mid X, Y)$ the posterior distribution over the model parameters, meaning that

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

is the posterior predictive distribution of Y_{new} given X_{new} and the post. distribution $p(z \mid X, Y)$

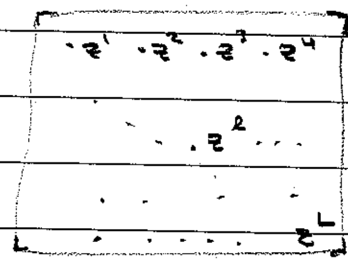
Today: approaches for generating "samples" from $p(z)$ so that (1) can be approximated as

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

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

First intuitive (but bad) approach:

Grid the domain of $p(z)$

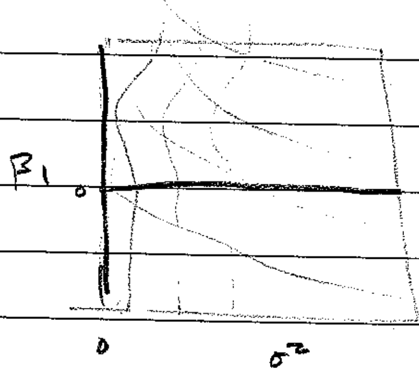


consider $z \in \mathbb{R}^2$

and compute

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

i.e. Bayesian regression through origin with prior like:



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

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. $q(z) = N(z \mid \mu_q, \Sigma_q)$

We can rewrite

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

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

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

when $z^l \sim q$ and

$$r_l = \frac{p(z^l)}{q(z^l)}$$

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.

e.g. $$p(z \mid X, Y) = \frac{p(z \mid X, Y)\, p(z)}{\int p(z \mid X, Y)\, p(z)\, dz}$$

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

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

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

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

i.e. 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, i.e. $q(z) = \frac{\tilde{q}(z)}{Z_q}$, sampling from $\tilde{q}(z)$ is the same as sampling from $q(z)$

In this situation we can write

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

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

where

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

But now what about Z_q / Z_p ?

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

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

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

$$\approx \frac{1}{L} \sum_{l=1}^{L} \tilde{r}_l$$

So $\mathbb{E}(f) \approx \sum w_l f(z^l)$ where $z^l \sim q$

$$\text{and} \quad w_l = \frac{\tilde{r}_l}{\sum_m \tilde{r}_m} = \frac{\tilde{p}(z^l)/\tilde{q}(z^l)}{\sum_m \tilde{p}(z^m)/\tilde{q}(z^m)}$$

So now, we can do posterior inference numerically without 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, one sample) -- 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

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

i.e. symmetric

Algorithm

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

Repeat Forever

- Propose z^* from $q(z^*|z)$
- "Accept" z^* w.p.

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

Note: if ratio is > 1 sample is always accepted

- Set $\tau = \tau + 1$
- If z^* "accepted" set $z^\tau = z^*$
- otherwise set $z^\tau = z^{\tau-1}$

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

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

Remember Markov chains have "equilibrium" distributions.

Markov chain (1st order): A first order Markov chain is defined to be a series of R.V.'s

$z^1, \ldots, z^M$ s.t.

$$p(z^{m+1} \mid z^1, \ldots, z^m) = p(z^{m+1} \mid z^m)$$

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

$$T_m(z^m \mid z^{m+1}) = p(z^{m+1} \mid z^m)$$

A Markov chain is <u>homogenous</u> if the transition probabilities are the same for all m.

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

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

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 homogeneous 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

$$z \begin{bmatrix} .3 \\ .3 \\ .3 \end{bmatrix} = \begin{array}{cc} & z' \\ z & \begin{bmatrix} .5 & .25 & .25 \\ .25 & .5 & .25 \\ .25 & .25 & .5 \end{bmatrix} \end{array} \begin{array}{c} z' \\ \begin{bmatrix} .3 \\ .3 \\ .3 \end{bmatrix} \end{array}$$

$z = Tz$ is eigen vector of T 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

$$\sum_{z'} p^*(z')\, T(z', z) = \sum_{z'} p^*(z)\, T(z, z') \quad \text{1st line from}$$

starting from

where we go

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

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

$$= p^*(z)$$

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 $T(z^{\tau+1}, z^\tau)$

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{array}{cc} & z \\ z' & \begin{bmatrix} a & b \\ c & d \end{bmatrix} \end{array} \qquad \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}$$

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

$$= \begin{array}{cc} & z' \\ z & \begin{bmatrix} a & c \\ b & d \end{bmatrix} \end{array} \qquad \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}$$

$$p_1 a + p_2 b = p_1 a + p_2 c$$

$$p_1 c + p_2 d = p_1 b + p_2 d$$

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

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

$$b = c \quad , \quad c = b$$

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

$$\begin{array}{cc} & z \\ z' & \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \end{array} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}$$

$$= \begin{array}{cc} & z' \\ z & \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix} \end{array} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} \qquad \text{s.t.} \quad \begin{array}{l} a+b+c=1 \\ d+e+f=1 \\ g+h+i=1 \end{array}$$

a–d

$$a+d+g=1$$
$$b+e+h=1$$
$$c+f+i=1$$

$$\cancel{ap_1} + bp_2 + cp_3 = \cancel{ap_1} + dp_2 + gp_3$$
$$dp_1 + \cancel{ep_2} + fp_3 = bp_1 + \cancel{ep_2} + hp_3$$
$$gp_1 + hp_2 + \cancel{ip_3} = cp_1 + fp_2 + \cancel{ip_3}$$

$$bp_2 + cp_3 + dp_1 = bp_1 + dp_2 + gp_3$$

$$b(p_2 - p_1) = dp_2 + gp_3 + cp_3 + dp_1$$

$$b = \frac{dp_2 + gp_3 + cp_3 + dp_1}{p_2 - p_1}$$

$$Tp = T'p$$

if T orthogonal

and symmetric then

$$T^{-1} = T' \quad \text{and} \quad T^{-1}Tp = T^{-1}Tp$$

$$p = p$$

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.

Fill

The Metropolis Algorithm is
1) homogeneous
2) ergodic
3) and has p^* as it's 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

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

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

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

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

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