Introduction to predictive models / MCMC

- Say I have a measurement, Y_{obs}, ~~Model M~~ and the measurement has ~~known~~ error, ε, with known pdf we can think of Y_{obs} as a R.V.

$$Y_{obs}(\vec{X}) = Y_{true}(\vec{X}) + \varepsilon \qquad \vec{X} \text{ are inputs}$$

- Also if we have a simulation that takes in inputs plus numerical parameters and constants of nature, other params

$$Y_{true}(\vec{X}) = Y_{sim}(\vec{X}, M, \Theta) + \delta(\vec{X})$$

uncertain consts of nature

discrepancy

- ~~We develop the p~~

Predictive model

$$Y_{obs}(\vec{X}) = Y_{sim}(\vec{X}, M, \Theta) + \delta(\vec{X}) + \varepsilon$$

How do we find discrepancy

- Note I have defined discrepancy ideally, it does not depend on mesh parameters or Θ
- Two parts: first tune Θ within reason

find $\Theta(x)$ that minimizes $\dfrac{Y_{obs} - Y_{sim}}{\varepsilon}$

- Then parameterize δ as a function of x

First part can be handled in a Bayesian way

Bayesian Inference for Θ

- Say we believe Θ has some distribution (prior) $\pi(\Theta)$, given a particular observation we want to change the distribution to account for more information
- Since we want to minimize $Y_{obs}(\vec{x}) - Y_{sim}(\vec{x}, \Theta) = D$

we can think of $p(D|\Theta) \sim N(0, \sigma)$

- σ can be the std of ε but it doesn't have to be.
- Don't have to use normal either
- IDEA if $|D|$ is large, that value of Θ is unlikely (not what we want)
- Bayes' Thm says

$$P(\Theta|D) = \frac{P(D|\Theta)\,\pi(\Theta)}{\int P(D|\Theta)\,\pi(\Theta)\,d\Theta}$$

- Information from observation changes our distribution
- Ok, so we have an observation how can we find new dist for ~~θ~~ Θ? Answer MCMC
- Before we do that lets remember what we want
  1. if some likely value of Θ can match the ~~simulation~~ obs. we want $p(\theta)$ to tell us that
  2. if no likely value of Θ can and all are equally bad, tell us that too

Basic MCMC

Recall in Monte Carlo we sample a dist and then compute
$E[f(X)] \cong \frac{1}{n} \sum_{t=1}^{n} f(X_t)$ for n samples

Markov chain

- ~~Given a~~ A sequence of rand. vars. $\{X_0, X_1, \dots X_t\}$ such that each time $t \geq 0$, the next state, X_{t+1} is a sample from $P(X_{t+1} | X_t)$, i.e. ~~it is only the~~ X_{t+1} only depends on X_t, is known as a Markov Chain and $P(X_{t+1}|X_t)$ is the transition probability.

- for our purposes $P(X_{t+1}|X_t)$ is independent of t,
- Under almost all circumstances X_t will be independent of X_0 (the chain forgets it's initial state)
- The distribution that X_t are samples of for $t \gg 0$ we call the stationary dist.
- After a long enough "burn-in" we can estimate expectation values from the stat. dist.

$$E[f(X)] \cong \frac{1}{n-m} \sum_{t=m+1}^{n} f(X_t) \qquad m = \text{burn-in time}$$

Metropolis-Hastings Algorithm

- We want ~~a~~ to construct a Markov chain with stationary dist that is the posterior from Bayes' Thm, $P(\theta|D)$

- This is actually pretty easy, which is surpising

- Only need $\pi(\theta)$ ~~$p(\theta)$~~ $\cdot\, P(D|\theta)$ (not normalization)

Idea: 1. take any proposal dist. $q(\cdot|X_t)$, usually multivariate normal, sample a value Y

2. Rejection method: ~~accepted~~ we accept proposals with probability

$$\alpha(X,Y) = \min\left(1, \frac{\hat{p}(Y)\,q(X|Y)}{\hat{p}(X)\,q(Y|X)}\right)$$

(if Y is more likely than X we are more likely to accept it)

3. if candidate is accepted $X_{t+1} = Y$, otherwise we don't move, $X_{t+1} = X_t$

Algorithm

Pick X_0,

for ~~θ in~~ $t = 0..T$
  Sample $Y \sim q(\cdot|X_t)$
  Pick $U \equiv \text{rand}(0,1)$
  if $U \le \alpha(X_t, Y)$ then $X_{t+1} = Y$
  else $X_{t+1} = X_t$
end for

- You can prove that if $\hat{p}(X) = P(D|X)\pi(X)$ (numerator from Bayes thm) the stationary dist is $p(\theta|D)$.

- Also ~~if~~ once $X_t \in p(X|D)$ all subsequent samples will also be

This makes (**)

$$\hat{p}(X_t)\, P(X_{t+1} | X_t) = \hat{p}(X_{t+1})\, P(X_t | X_{t+1})$$

This is called the detailed balance equation

If we integrate over all X_t, we get

$$\int dX_t\, \hat{p}(X_t)\, P(X_{t+1} | X_t) = \hat{p}(X_{t+1}) \cancelto{1}{\int dX_t\, P(X_t | X_{t+1})}$$

What this equation says - what is the probability of transitioning to X_{t+1} given that X_t is from $\hat{p}(X_t)$.

Therefore, once a sample from $\hat{p}(\cdot)$ has been obtained all subsequent samples will be from $\hat{p}(\cdot)$

It also shows that from our algorithm, the stationary dist of the Markov Chain is $\hat{p}(\cdot)$.

Therefore, after a long-enough burn-in our Markov Chain will consist of samples from

$$\hat{p}(X_{t+1}) = \frac{P(D | X_{t+1})\, \pi(X_{t+1})}{\int dX_{t+1}\, p(D | X_{t+1})\, \pi(X_{t+1})} = P(X_{t+1} | D)$$

Let's do a very simple example.

$\hat{p}(X) \sim N(0,1)$ and we pick $g(\cdot | X) = N(X, \sigma^2)$

Let $$\hat{p}(X) = \frac{p(D|X)\pi(X)}{\int dx\, p(D|X)\pi(X)} = P(X|D)$$

This makes

$$\alpha(X,Y) = \min\left(1, \frac{P(D|Y)\pi(Y)\,q(X|Y)}{p(D|X)\pi(X)\,q(Y|X)}\right)$$

Note we "cancelled" denominator

$$= \min\left(1, \frac{\hat{P}(Y)\,q(X|Y)}{\hat{P}(X)\,q(Y|X)}\right) \quad (*)$$

Manipulating $(*)$ and set $X \to X_t$, $Y \to X_{t+1}$ or X_t

$$\hat{P}(X_t)\,q(X_{t+1}|X_t)\,\alpha(X_t,X_{t+1}) = \hat{P}(X_{t+1})\,q(X_t|X_{t+1})\,\alpha(X_{t+1},X_t) \quad (**)$$

by manipolating $(*)$

Now notice that

$q(X_{t+1}|X_t)\,\alpha(X_t,X_{t+1})$ = prob. dens of switching from X_t to X_{t+1}

$$= P(X_{t+1}|X_t)$$

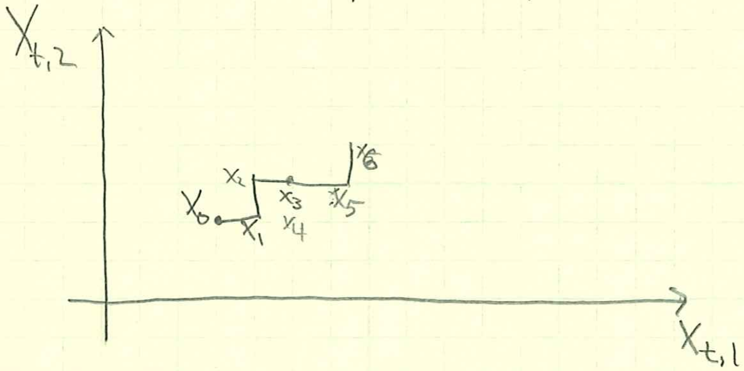
How likely is X_{t+1} given X_t times prob. of acceptance

Other Considerations

- Original Metropolis alg assumed $g(Y|X) = g(X|Y)$ (easy to do)
  then $\alpha(X,Y) = \min\left(1, \frac{\hat{p}(Y)}{\hat{p}(X)}\right)$

- X does not have to be of constant length

- Single-Component MH updates ~~each~~ one component of X in each step example



Can be a lot simpler than updating all comps

- Chains can be run in parallel (and this can be a good ide)
  Check burn-in, convergence

- Burn-in iterations no general rule, depends on how close transition prob is to stationary prob
  - it has been suggested that it be 1% - 2% of total samples

Example (Show pg 6 of Gilks, et al.)

Example for Calibration

- Modeling an object falling from rest.
- Our simulation will give

$V_{sim} = gt$ where t is time object has fallen

we also know $g \in [9.79, 9.82]$ and uniformly dist.

- We will also have exp. data V_{obs}
- For this example $V_{obs} = 9.81(t - 0.5\sin(\pi t))$

- To find best value of g we use

$$\hat{p}(t,g) = \begin{cases} \phi(V_{obs}(t) - V_{sim}(t,g) \mid \mu = 0, \sigma = 1.0 \times 10^{-3}) & g \in [9.79, 9.82] \\ 0 & \text{otherwise} \end{cases}$$

$$q(Y|X) \sim N(X, \sigma_q = 0.05)$$

We observe at $t = 0.1, 1, 1.5, 2, 5, 10$ | do 1000 MCMC samples at each point, 100 burnin

t	g_{ex}	$\mu(g_{post})$	$\sigma^2(g_{post}) \times 10^{-5}$
0.1	6.7785	9.8051	7.673
1	9.81	9.8101	0.094461
1.5	10.464	9.8049	7.0692
2	9.81	9.8099	0.027825
5	9.81	9.8100	0.00327
10	9.81	9.8101	0.001971

- Mean of all samples 9.8084
- Show all samples

- Can do better if we pick a different $\hat{p}$

$$\hat{p}(t, g) = \begin{cases} \phi(g_{exact} - g \mid \mu = 0, \sigma = 10^{-3}) & g \in [9.79, 9.82] \\ \phi(9.79 - g \mid \mu = 0, \sigma = 10^{-3}) & g < 9.79 \\ \phi(9.82 - g \mid \mu = 0, \sigma = 10^{-3}) & g > 9.82 \end{cases}$$

$$g_{exact}(t) = \frac{V_{obs}}{t}$$

This says if g is out of bounds, gravitate to the closest value

Now:

t	$\mu(g)$
0.1	9.7907
1	9.81
1.5	9.819
2	9.8099
5	9.8100
10	9.8100

mean of all: 9.8083