Math 650 5/9/17

Recall: $X_1, \dots, X_n \mid \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$

$\theta \sim N(\mu_0, \tau^2)$

$\sigma^2 \sim \text{Invgamma}\left(\frac{n_0}{2}, \frac{n_0 \sigma_0^2}{2}\right)$

$\Rightarrow p(\theta, \sigma^2 \mid X) \propto N(\theta_p, \sigma_p^2) K(\sigma^2 \mid X)$

Recall: $P(\theta \mid \sigma^2, X) = N(\theta_p, \sigma_p^2)$, $P(\sigma^2 \mid \theta, X) = \text{Invgamma}\left(\frac{n_0+n}{2}, \frac{n\hat{\sigma}^2_{MLE} + n_0\sigma_0^2}{2}\right)$ } Can we use these 2 conditionals to tells us about $P(\theta, \sigma^2 \mid X)$ = ?

No. You need to know one of the marginals. Not just the conditionals.

- $P(AB) = P(A \mid B) P(B) = P(B \mid A) \cdot P(A)$

$P(\theta, \sigma^2 \mid X) = P(\theta \mid \sigma^2, X) P(\sigma^2 \mid X) = P(\sigma^2 \mid \theta, X) P(\theta \mid X)$

What if we used an iterative approach?

1. ~~Begin at~~ Assume an arbitrary θ_0
2. Draw σ_0^2 from $P(\sigma^2 \mid \theta = \theta_0, X) = \text{Invgamma}$
3. Draw θ_1 from $P(\theta \mid \sigma^2 = \sigma_0^2, X)$
4. Draw σ_1^2 from $P(\sigma_1^2 \mid \theta = \theta_1, X)$
5. Repeat 2, 3 until "convergence". like Newton-Rafson / EM.

→ will return whole posterior.

EM ≠ return whole posterior

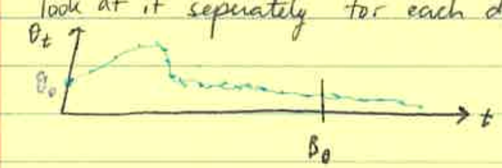
Algorithm Called Gibbs sampling or Gibbs sampler. = return $\hat{\theta}_{max}$

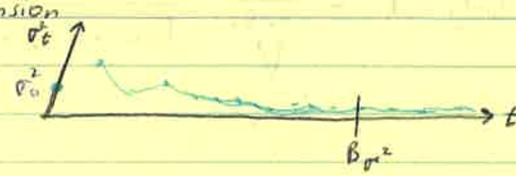
"Edward Teller" H-Bomb

The output of a Gibbs sample looks like

$\left\langle \begin{bmatrix}\theta_0 \\ \sigma_0^2\end{bmatrix}, \begin{bmatrix}\theta_1 \\ \sigma_1^2\end{bmatrix}, \dots, \begin{bmatrix}\theta_t \\ \sigma_t^2\end{bmatrix}, \dots \right\rangle$

look at it sepuratly for each dimension

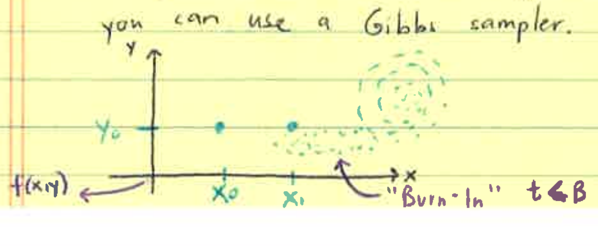




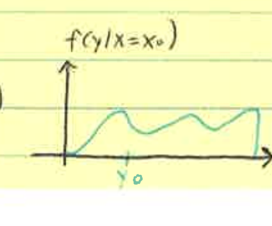
"Burn-in point" Point where operates the way it's supposed to normally. (Buying new car)

For the Gibbs ~~sampler~~ chain the burn-in is the max B_j for all θ_j's. For $t \geq B$, the chain is "burned-in".

If you seek $f(x,y)$ but only know $f(x \mid y)$ and $f(y \mid X)$ you can use a Gibbs sampler.



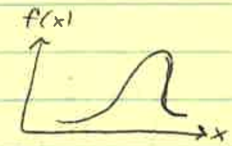
1. Begin at arbitrary x_0
2. Draw y_0 from $f(y \mid X = x_0)$
3. Draw x_1 from $f(x \mid y = y_0)$



What if you want $f(x)$?

$$f(x) = \int_{\mathrm{Supp}[Y]} f(x,y)\,dy$$

collapsing y values
get all x-values.
get one-dimensional hill.

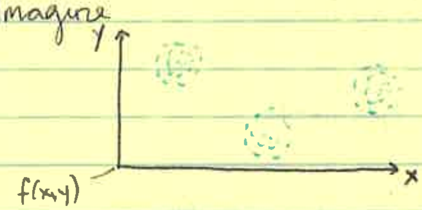


Main Problems with Gibb Sampler

Good ✓ "Mixing"

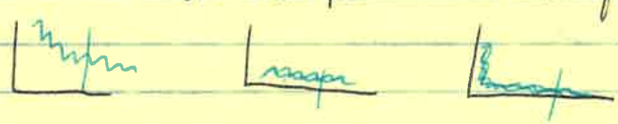
? Don't know what's going on.

Imagine



⇒ Sampler gets hopelessly stuck in a piece of the distribution.

A possible solution: Use multiple Gibbs chains from different X_0's.



A smaller problem which is more fixable is as follows:

Is θ_1 related to θ_0? YES.

$$\text{Correlation}[X,Y] := \frac{\mathrm{Cov}[X,Y]}{SE[X]\,SE[Y]} = \frac{E([X-\mu_x][Y-\mu_y])}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

$$\mathrm{Cov}[\theta_1, \theta_0] \neq 0$$

estimates covariance

$$r := \frac{s_{xy}}{s_x s_y} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2}\sqrt{\Sigma(y_i-\bar{y})^2}}$$

estimators for stand. errors.

Is $\theta_{1,000}$ related to θ_{999}? Yes. ⇒ So Burn-In doesn't help.

Auto = Self

"Auto-correlation". for "lag" one.

$$\hookrightarrow r_{a,1} := \frac{\sum_{t=B}^{B+S-1} (\theta_t - \bar{\theta})(\theta_{t+1} - \bar{\theta})}{\sum_{t=B}^{B+S} (\theta_t - \bar{\theta})^2}$$

$\bar{\theta}$ = average of all θ's
$= \frac{1}{S} \sum_{t=B}^{S} \theta_t$

B = Burn In

Autocorrelation of lag 2 : $r_{a,2} := \dfrac{\sum_{t=B}^{B+S-2} (\theta_t - \bar{\theta})(\theta_{t+2} - \bar{\theta})}{\sum_{t=B}^{B+S} (\theta_t - \bar{\theta})^2}$

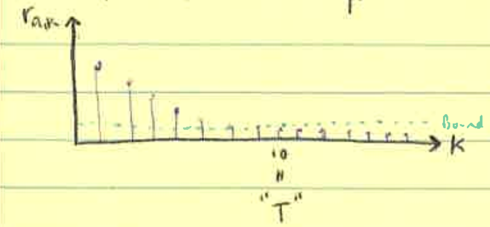
$B+S$ = the total # of iterations.

$\vdots$

Autocorrelation of Lag k : $r_{a,k} := \dfrac{\sum_{t=B}^{B+S-k} (\theta_t - \bar{\theta})(\theta_{t+k} - \bar{\theta})}{\sum_{t=B}^{B+S} (\theta_t - \bar{\theta})^2}$

when does correlation not matter anymore?

Look at "autocorrelation plot"



takes 10 iterations to say not correlation doesn't matter anymore. (correlation so minimal)

$\Downarrow$

$\left\langle \begin{bmatrix} \theta_B \\ \sigma^2_B \end{bmatrix}, \begin{bmatrix} \theta_{B+T} \\ \sigma^2_{B+T} \end{bmatrix}, \begin{bmatrix} \theta_{B+2T} \\ \sigma^2_{B+2T} \end{bmatrix} \dots \right\rangle$ good to have ≥ 1000. length L

$\Downarrow$

Now your Burned & thinned chain.

$\hat{\theta}_{MMSE} = E[\theta | X] \approx \bar{\theta} = \frac{1}{L} \sum_{\ell=1}^{L} \theta_\ell$ average out

$\hat{\theta}_{MAE} = Med[\theta | X] \approx$ order θ's from lowest to highest

$\theta_1, \theta_2, \dots, \theta_\ell$

and take $\theta_{(\frac{L}{2})}$ middle value.

mode estimations are not easy ...

$P(X^* | X) = \int_{\Theta} P(X^* | \theta) P(\theta | X) \, d\theta$ → explain process (in Final?) Sample θ from posterior Sample X.

How do we ~~approximate~~ sample from this Distribution?

1. Pick $\ell \in \{1, \dots, L\}$
2. Sample X^* from $P(X^* | \theta = \theta_\ell)$
3. Repeat 1-2.