Recall

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

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

$$\sigma^2 \sim \text{Inv-}\chi^2\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

$$P(\theta, \sigma^2 \mid x) \propto K(\theta, \sigma^2 \mid x) \quad \text{non-conjugate}$$

but

$$P(\theta \mid x, \sigma^2) = N(\theta_p, \sigma_p^2)$$

$$P(\sigma^2 \mid x, \theta) = \text{Inv-}\chi^2\left(\frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + n\hat{\sigma}^2}{2}\right)$$

Can you use $P(\theta \mid x, \sigma^2)$ & $P(\sigma^2 \mid x, \theta)$ to solve for $P(\theta, \sigma^2 \mid x)$?

$$P(A \cap B) = P(A \mid B) P(B) = P(B \mid A) P(A)$$

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

not possible unless either $P(\theta \mid x)$ or $P(\sigma^2 \mid x)$

are also as not possible ... so no!

However ... what if you use an iterative algorithm?

1. Begin at θ_0
2. Draw σ_0^2 from $P(\sigma^2 \mid x, \theta = \theta_0)$
3. Draw θ_1 from $P(\theta \mid x, \sigma^2 = \sigma_0^2)$
4. Draw σ_1^2 from $P(\sigma^2 \mid x, \theta = \theta_1)$

Until "convergence"

AKA "Gibbs sampling" or the "Gibbs sampler"!

This is different than the N-R and E-M algs. Why?

Newton's method:

solve for $f(x) = 0$ one value

E-M

solve for $\hat{\theta}_{MAP}$ is one value (or vector)

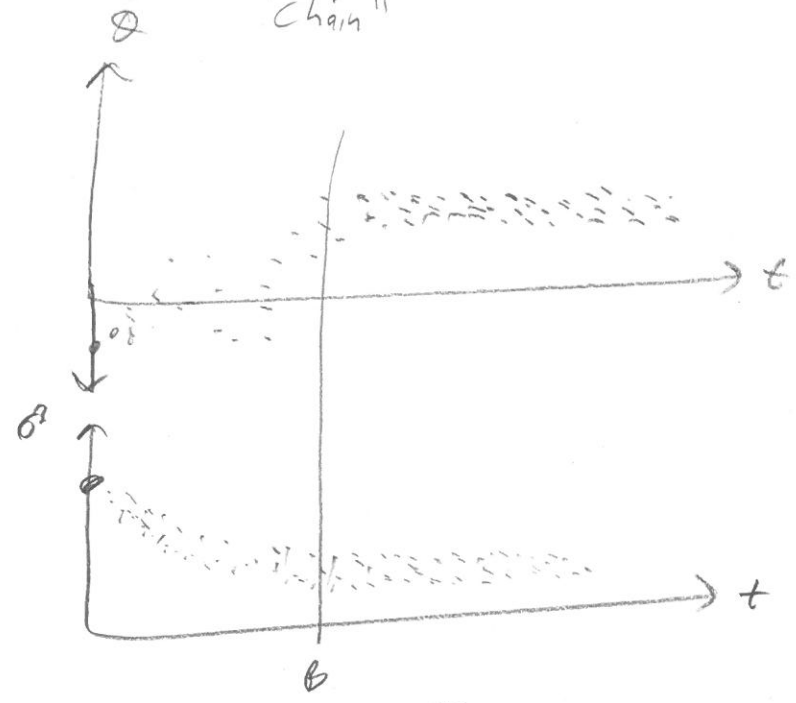
Here:

$p(\theta, \sigma^2 | x)$ ... entire posterior!!

It returns look like:

$\langle \begin{bmatrix} \theta_0 \\ \sigma_0^2 \end{bmatrix}, \begin{bmatrix} \theta_1 \\ \sigma_1^2 \end{bmatrix}, \begin{bmatrix} \theta_2 \\ \sigma_2^2 \end{bmatrix}, \ldots \begin{bmatrix} \theta_t \\ \sigma_t^2 \end{bmatrix} \ldots \rangle$ where t is iteration #

"chain"

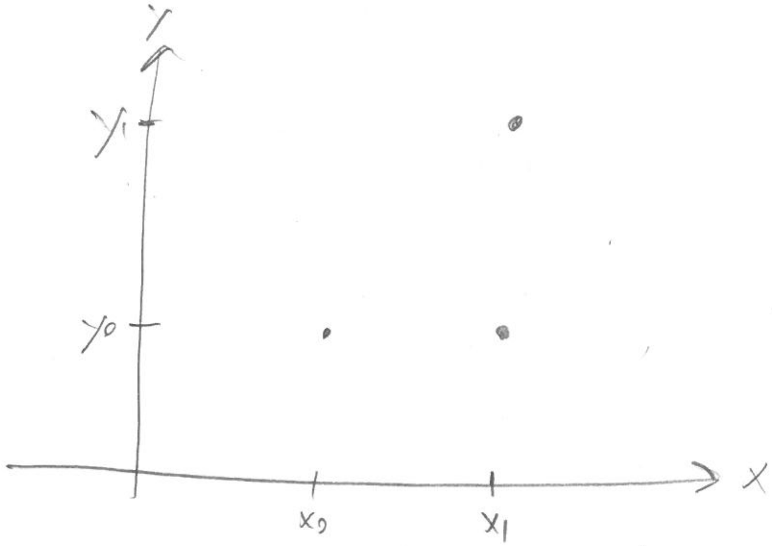


let $B = \max_j B_j$
s.t. B_j is the convergent pt of θ_j

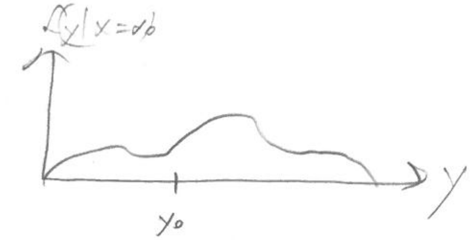
Where did algorithm converge?

We call $t = B$ the burn-in point. Kind of like E M N-R or E-M

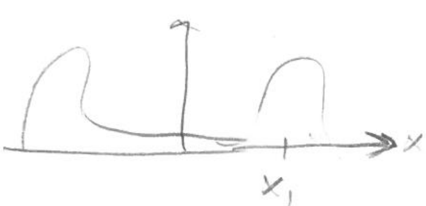
Put another way. You seek $f(x,y)$ but you only know $f(x|y)$ & $f(y|x)$.



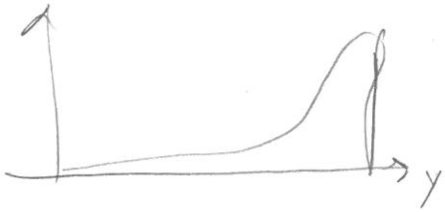
Begin with x_0. Draw y_0 from $f(y|x=x_0)$



then draw x_1 from $f(x|y=y_0)$



then draw y_1 from $f(y|x=x_1)$



⋮

Goal

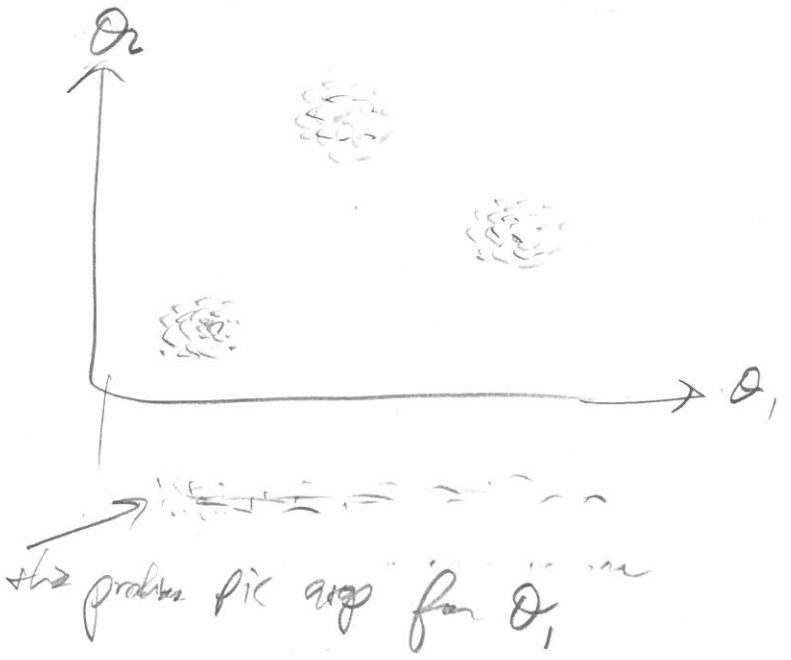
If you only care about $f(\alpha)$, you collapse all y's by just deleting the second dimension

The Basic problem with this type of sampler

① Bad mixing

Lacks ability to traverse $\text{supp}[\vec{\theta}]$ well

$\vec{\theta}$ may be pair of a set of Normal distr's with multiple modes



The sampler will get stuck in any of these modes.

Sol: make many chains! Start from all different starting pts...

problematic with big dim($\vec{\theta}$)!

⇒ BIG Problems...

unsure of its Sol'n adequacy

A smaller (but feasible) problem is as follows

Init σ_0^2

draw θ_0 from $P(\theta | X, \sigma^2 = \sigma_0^2)$

draw σ_1^2 from $P(\sigma^2 | X, \theta = \theta_0)$

draw θ_1 from $P(\theta | X, \sigma^2 = \sigma_1^2)$

Is θ_1 related to θ_0? Yes...

Is θ_{1000} related to θ_{999}? Yes... After Burn-in (B) still!!

thus θ_{1000} and θ_{999} are not "independent samples"! since $\text{Corr}[\theta_{1000}, \theta_{999}] \neq 0$

recall $\text{Corr}[X, Y] = \frac{\text{Cov}(X,Y)}{SE(X)SE(Y)} = \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$

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

we can look at the auto-correlation

auto = self

autocorrelation for lag 1 estimates $\text{corr}[\theta_t, \theta_{t+1}]$

$$r_{a1} := \frac{\sum_{t=B}^{B+S-1} (\theta_t - \bar{\theta})(\theta_{t+1} - \bar{\theta})}{\sum_{t=B}^{B+S} (\theta_t - \bar{\theta})^2} \quad \text{s.t. } \bar{\theta} = \frac{1}{S} \sum_{t=B}^{B+S} \theta_t$$

Autocorrelation for lag 2 is:

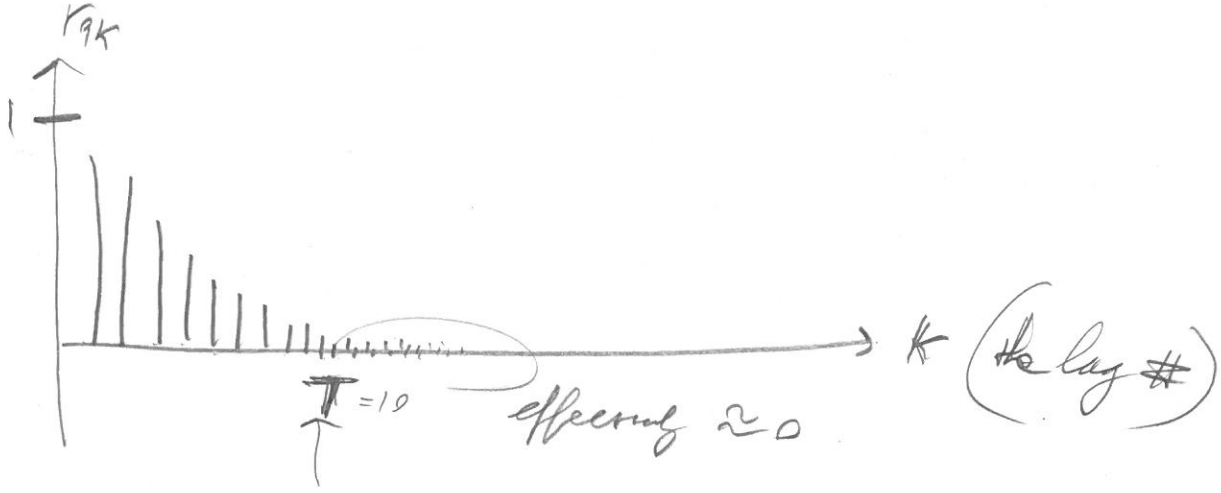
$$r_{y2} := \frac{\sum_{t=B}^{B+S-2} (\theta_t - \bar{\theta})(\theta_{t+2} - \bar{\theta})}{\sum_{t=B}^{B+S} (\theta_t - \bar{\theta})^2}$$

$$\vdots$$

$$r_{yk} := \frac{\sum_{t=B}^{B+S-k} (\theta_t - \bar{\theta})(\theta_{t+k} - \bar{\theta})}{\sum_{t=B}^{B+S} (\theta_t - \bar{\theta})^2}$$

At some k, $r_{yk} \simeq 0$. Why? Eventually the dependence is lost...

How to see? Autocorrelation plot



At around T iterations ago, the draws effectively are independent.

In order to make the chain represent all independent samples from posterior, we need to throw out all samples except those that are multiple of T after B. This is known as "thinning".

$$\left\{ \begin{pmatrix} \theta_B \\ \sigma^2_B \end{pmatrix}, \begin{pmatrix} \theta_{B+T} \\ \sigma^2_{B+T} \end{pmatrix}, \begin{pmatrix} \theta_{B+2T} \\ \sigma^2_{B+2T} \end{pmatrix}, \ldots \right\} \rightarrow \text{the burned and thinned chain.}$$

let $l = 1 \ldots L$ be the index on the burned and thinned chain. This is almost as good as having $P(\theta|x)$ directly:

How to get $\hat{\theta}_{MMSE} = E(\theta|x) \approx \bar{\theta} = \frac{1}{L}\sum_{l=1}^{L} \theta_l$

$\hat{\theta}_{MAE} = \text{Med}(\theta|x)$ = order from smallest to largest $\theta_{(1)}, \ldots \theta_{(L)}$ and take $\theta_{\left(\frac{L}{2}\right)}$ (round)

$CR_{\theta, 1-\alpha}$ = order from smallest to largest

$$\left[\theta_{\left(\frac{\alpha}{2}L\right)}, \theta_{\left(\left(1-\frac{\alpha}{2}\right)L\right)}\right]$$

round ... round ...

$$\text{pval} = P(H_0|x) = P(\theta \in \Theta_0|x) \approx \frac{1}{L}\sum_{l=1}^{L} \mathbb{1}_{\theta_l \in \Theta_0}$$

e.g. $P(\theta > 3|x) \approx \frac{1}{L}\sum_{l=1}^{L} \mathbb{1}_{\theta_l > 3}$ i.e. prop. of L s.t. $\theta > 3$

$$P(X^*|x)\ ? = \int_{\Theta} P(X^*|\theta) P(\theta|x)\, d\theta$$

to sample from this ...

① Pick $l \in \{1, \ldots, L\}$

② Draw X^* from $P(X^*|\theta = \theta_l)$

③ Repeat steps 1, 2 over and over ...

Def: Systematic Sweep Gibbs Sampler

Assume posterior $P(\theta_1, \ldots, \theta_p|X)$ is unknown but

$P(\theta_j | \theta_{-j}, X)$ s.t. $\theta_{-j} = \{\theta_1, \ldots \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p\}$ i.e. all θ's except θ_j

is known $\forall j$.