(1)

Lec 19   Math 341   4/30/18

Previously... we needed to sample from $k(\theta|X)$ so we used grid sampling. This involved choices: θ_{min}, θ_{max}, $\Delta\theta$ to make g grid pts. We then calculated $\langle k(\theta_{min}|x), k(\theta_{min}+\Delta\theta|x), \ldots, k(\theta_{min}+(g-1)\Delta\theta|x), k(\theta_{max}|\theta)\rangle$

then we approximated $\int k(\theta|x) \approx \sum k(\theta_g|x)\,\Delta\theta = \frac{1}{c}$

and now we have $\langle P(\theta_{min}|x), \ldots, P(\theta_{max}|x)\rangle$

$$P(\theta_{min}|x) = c\,k(\theta_{min}|x)$$

and we can sample from $\theta|x$ using the discrete r.v. sampling technique

Grid Sampling Disadvantages

(1) Numerically unstable.

Computers have minimum values of #'s / max. value of #'s.

(2) How to pick θ_{min}, θ_{max}, $\Delta\theta$?

Bad decision for θ_{min}, θ_{max} $\Rightarrow$ You miss a part of the support of the parameter!

Bad decision for $\Delta\theta$ $\Rightarrow$ Bad resolution $\Rightarrow$ non-relevant samples

(3) Let's say $\theta_{min} = 0$, $\theta_{max} = 1$, $\Delta\theta = 0.0001$, $|G| = 10{,}000 = 10^5$

What if θ had 10 dimensions? $\Rightarrow$ $|G| = 10^{5^{10}} = 10^{50}$ $\Rightarrow$ IMPOSSIBLE for a computer!

$\Rightarrow$ Grid sampling only good in low dimensions if you know the effective support of θ (i.e. where most of the support lies) and if you know the shape so you can pick a reasonable $\Delta\theta$.

It would be nice to fix these problems with a new method...

Recall

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

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

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

$$p(\theta, \sigma^2 \mid X) \propto k(\theta, \sigma^2 \mid X) \quad \text{non-conjugate}$$

but

$$p(\theta \mid X, \sigma^2) = N(\theta_p, \sigma_p^2)$$

$$p(\sigma^2 \mid X, \theta) = \text{InvG}\left(\frac{n_0 + n}{2}, \frac{n_0 \sigma_0^2 + n \hat{\sigma}^2}{2}\right)$$

Can you use $p(\theta \mid X, \sigma^2)$ and $p(\sigma^2 \mid X, \theta)$ to solve for $p(\theta, \sigma^2 \mid X)$?

$$p(AB) = p(A \mid B)\, p(B) = p(B \mid A)\, p(A)$$

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

not possible unless either $p(\theta \mid X)$ or $p(\sigma^2 \mid X)$

and those are 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 $p(\sigma^2 \mid X, \theta = \theta_1)$

until "converge"

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

This is different than the N-R and E-M alg's, why?

~~Newton's method~~

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

~~E-M~~

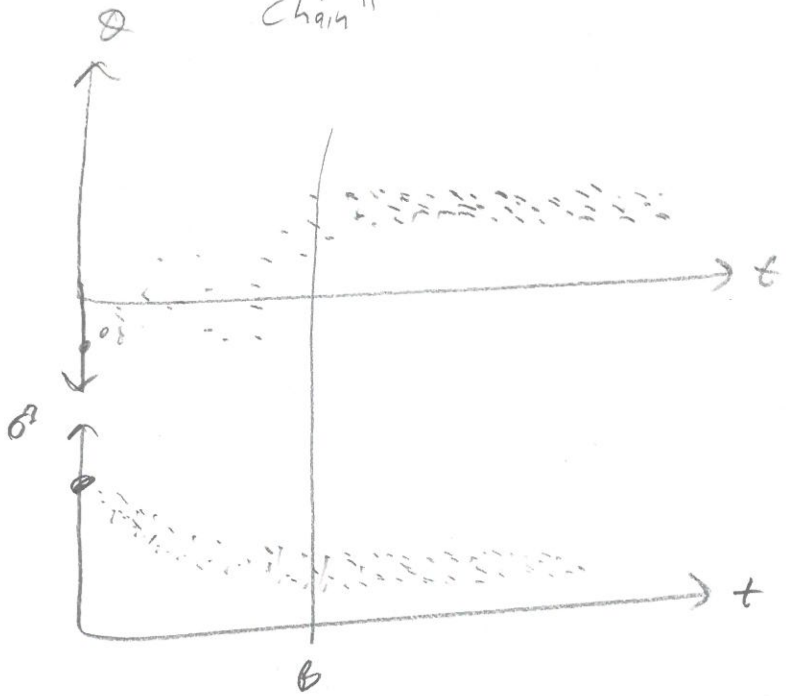
~~solve for $\hat{\theta}_{MAP}$ is one value (or vector)~~

~~Here:~~

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

Iterations look like:

$$\underbrace{\left\langle \begin{bmatrix}\theta_0\\ \sigma_0^2\end{bmatrix}, \begin{bmatrix}\theta_1\\ \sigma_1^2\end{bmatrix}, \begin{bmatrix}\theta_2\\ \sigma_2^2\end{bmatrix}, \ldots \begin{bmatrix}\theta_t\\ \sigma_t^2\end{bmatrix} \right\rangle}_{\text{"Chain"}}$$ where t is iteration #



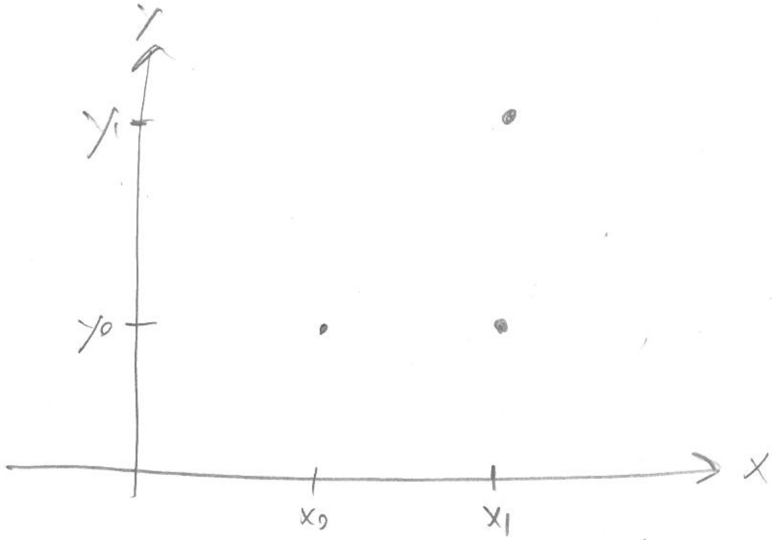
let $B = \max\limits_j B_j$

s.t. B_j is the convergent pt of θ_j

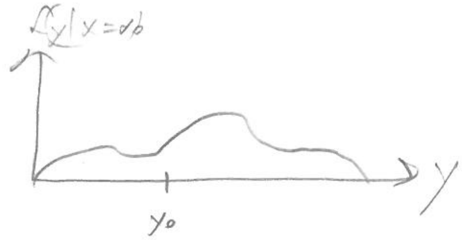
When did algorithm converge?

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

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

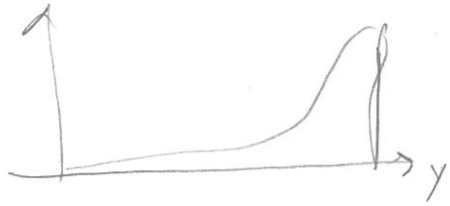


Begin with x_0. Draw y_0 from $f(y|x=x_0)$



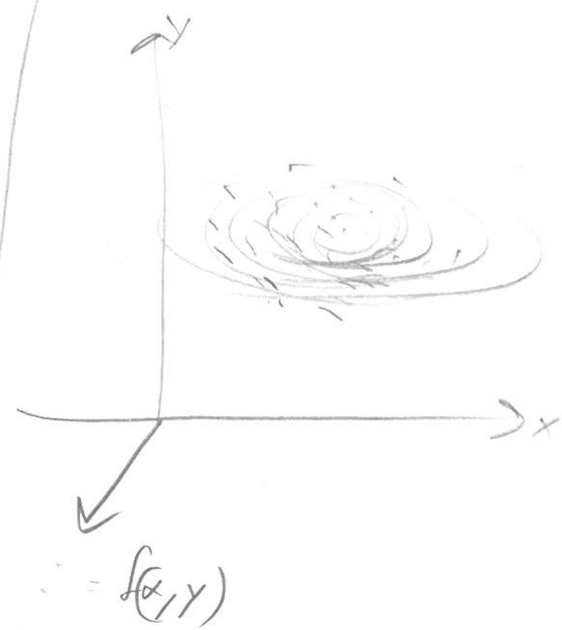
then draw x_1 from $f(x|y=y_0)$

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



⋮

Goal



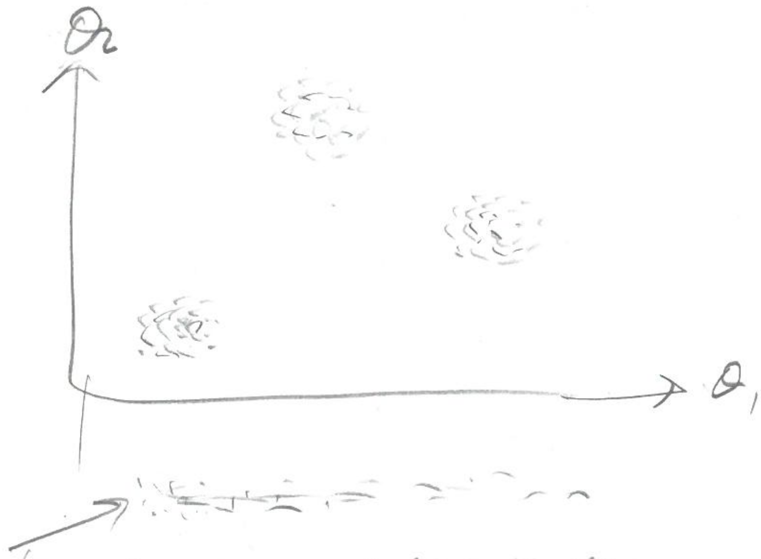
If you only care about $f(x)$, you collapse all y's by just deleting the second dimension

The Main problems with this type of sampling

① Bad mixing

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

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



this problem pic agap from θ_1

The sampler will get stuck in any of these modes.
Sol: make many chains! Start from all different starting pts...

problematic with big $dim(\theta)$!

$\Rightarrow$ BIG problems...
... some of it's solved adequately

A smaller (but feasible) problem is as follows

Init σ_0^2

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

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

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

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

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

we can look at the autocorrelation

$\underbrace{\text{auto}}_{\text{auto} = \text{self}}$

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

$$r_{q1} := \frac{\sum_{t=B}^{B+S-1} (\theta_t - \bar{\theta})(\theta_{t+1} - \bar{\theta})}{\sum_{t=B}^{B+S} (\theta_t - \bar{\theta})^2} \quad \text{s.t. } \bar{\theta} = \frac{1}{S} \sum_{t=B}^{B+S} \theta_t$$