LECTURE 21

Recap: In lecture 19, we wanted to find $\hat{\theta}_{MAP} = \text{argmax}\{p(\theta|x)\} = \text{argmax}\{k(\theta|x)\}$ which involved finding $k'(\theta|x)$. We introduced the Newton Raphson Algorithm for this. In lecture 20, we wanted to find good estimates of our parameters, so we wanted to find the MLE. We introduced the Expectation-Maximization Algorithm for this.

Both start with a guess & then iterate until you get close enough.

Today we introduce another algorithm, called Gibbs sampling. Our motivation is as follows:

Recall in Lecture 15 we ran into the following problem. For $X_1, \dots, X_n$ each $N(\theta, \sigma^2)$, we wanted to look at the posterior $p(\theta, \sigma^2|X)$ where our priors were "independent", i.e.

$$P(\theta, \sigma^2) = \underbrace{p(\theta)}_{N(\mu_0, \tau^2)} \; \underbrace{p(\sigma^2)}_{\text{InvGamma}\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right)}$$

Going through the math, we saw

$$p(\theta, \sigma^2|X) \propto N(\theta_p, \sigma_p^2)\underbrace{(\sigma^2)^{-\frac{n_0+n}{2}-1} e^{-\frac{\sum x_i + n_0\sigma_0^2}{2\sigma^2}}}_{\text{Kernel of Inv Gamma}} \underbrace{e^{-Q}\sqrt{\frac{2\pi}{\frac{n^2}{\sigma^2}+\frac{1}{\tau^2}}}}_{\text{other stuff}}$$

where $Q := \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\left(\frac{\frac{\mu_0}{\tau^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right)^2$

$k(\sigma^2|X) \propto$ UNKNOWN DISTRIBUTION

So we did not know what our posterior was. To get around this we introduced GRID SAMPLING which approximated a distribution of $k(\sigma^2|X)$. With this approximation, we could draw samples from the posterior.

We now attempt to get samples from $p(\theta, \sigma^2|X)$ by means of Gibb Sampling. The idea behind Gibbs sampling is that we know the following conditionals:
We know $p(\theta|X, \sigma^2) = N(\theta_p, \sigma_p^2)$ and we know
$p(\sigma^2|X, \theta) = \text{InvGamma}\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2 + n\hat{\sigma}^2}{2}\right)$

Now $\theta|X, \sigma^2$ & $\sigma^2|X, \theta$ cannot give us $\theta, \sigma^2|X$, BUT we can do the

following:

STEP ONE: Guess σ^2, call it σ_0^2. (A good guess for σ^2 is the sample variance, S^2.)
Guess θ, call it θ_0. (A good guess for θ is $\bar{X}$.)

STEP TWO: Draw θ_1 from $P(\theta \mid X, \sigma^2 = S^2 = \sigma_0^2) = N\left(\frac{\frac{n\bar{X}}{\sigma_0^2} + \frac{\mu_0}{\tau^2}}{\frac{n}{\sigma_0^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma_0^2} + \frac{1}{\tau^2}}\right)$

Using this θ_1, draw σ_1^2 from
$P(\sigma^2 \mid X, \theta = \theta_1) = \text{InvGamma}\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2 + \sum_{i=1}^{n}(x_i - \theta_1)^2}{2}\right)$

We get an ordered pair (θ_1, σ_1^2)

STEP THREE: Repeat Step Two until "convergence."
↳ when we say convergence we don't mean the usual "REAL ANALYSIS/CALCULUS" Definition of convergence like in N-R or EM Algorithm. That is, we do not mean after some integer $N > 0$, $|\theta_N - \theta_{N+1}| < \varepsilon$ & $|\sigma_N^2 - \sigma_{N+1}^2| < \varepsilon$. INSTEAD we mean CONVERGENCE IN DISTRIBUTION! NOT CONVERGENCE TO A REAL NUMBER. That is, after some point, the θ's we are drawing in step two are samples from the distribution of $p(\theta|x)$. Similarly, the σ^2's we are drawing (after some point) are samples from $p(\sigma^2|x)$. Visually, we have the following ordered set:

$$\left\langle \begin{bmatrix}\theta_0\\ \sigma_0^2\end{bmatrix}, \begin{bmatrix}\theta_1\\ \sigma_1^2\end{bmatrix}, \begin{bmatrix}\theta_2\\ \sigma_2^2\end{bmatrix}, \cdots\cdots, \begin{bmatrix}\theta_T\\ \sigma_T^2\end{bmatrix} \right\rangle$$

Let $t = B$ be the iteration number of convergence

Then $\{\theta_B, \theta_{B+1}, \theta_{B+2}, \cdots, \theta_{B+N}\}$ are samples from $p(\theta|x)$ &
$\{\sigma^2_B, \sigma^2_{B+1}, \sigma^2_{B+2}, \cdots, \sigma^2_{B+N}\}$ are samples from $p(\sigma^2|x)$,

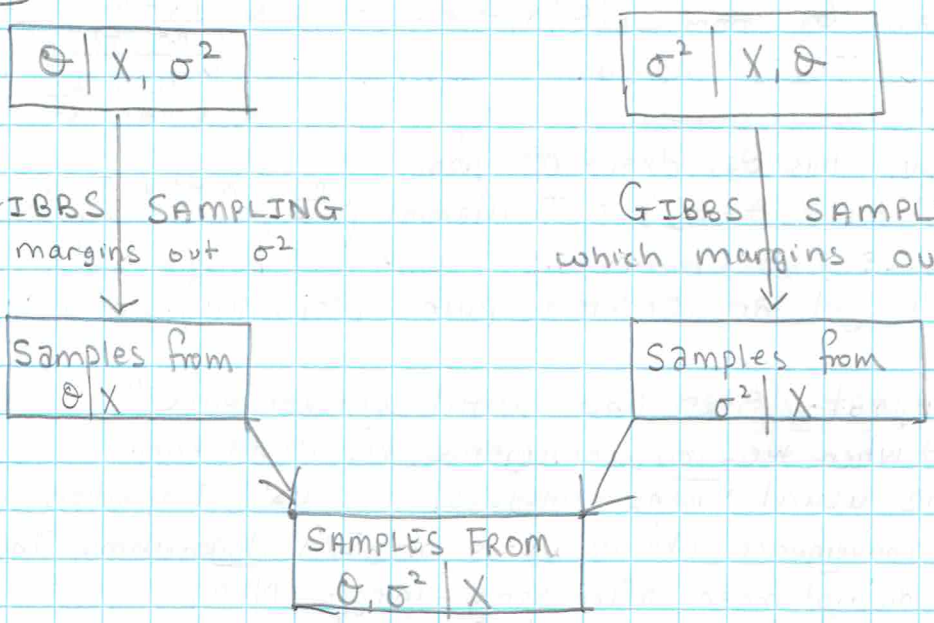
AND HENCE,

$$\left\langle \begin{bmatrix}\theta_B\\ \sigma_B^2\end{bmatrix}, \begin{bmatrix}\theta_{B+1}\\ \sigma_{B+1}^2\end{bmatrix}, \cdots, \begin{bmatrix}\theta_{B+N}\\ \sigma_{B+N}^2\end{bmatrix} \right\rangle$$ are samples from $p(\theta, \sigma^2 \mid X)$ which is exactly what we wanted!

Note that we call $\langle \theta_0, \theta_1, \theta_2, \ldots \rangle$ and $\langle \sigma_0^2, \sigma_1^2, \sigma_2^2, \ldots \rangle$ CHAINS.

We call $\langle \theta_B, \theta_{B+1}, \theta_{B+2}, \ldots \rangle$ and $\langle \sigma_B^2, \sigma_{B+1}^2, \sigma_{B+2}^2, \ldots \rangle$ the CHAINS after convergence.

Visually what happened was this:



FROM $\{\theta_B, \ldots, \theta_{B+N}\}$ & $\{\sigma_B^2, \ldots, \sigma_{B+N}^2\}$ we get

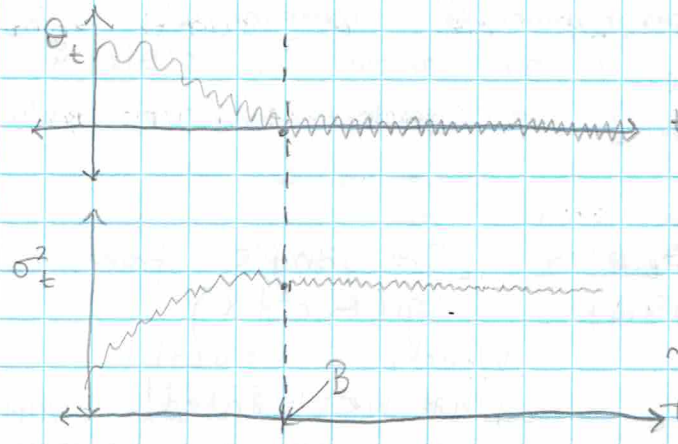
$$\hat{\theta}_{mmse} \approx \frac{1}{N+1} \sum_{t=B}^{B+N} \theta_t \qquad \hat{\sigma}^2_{mmse} \approx \frac{1}{N+1} \sum_{t=B}^{B+N} \sigma_t^2$$

$\hat{\theta}_{MAE} \approx$ middle θ_t $\qquad \hat{\sigma}^2_{MAE} \approx$ middle σ_t^2

Now we don't usually graph the 2 dimensional chain, i.e. we do not plot σ_t^2 vs θ_t

Instead we plot θ_t vs t & σ_t^2 vs t

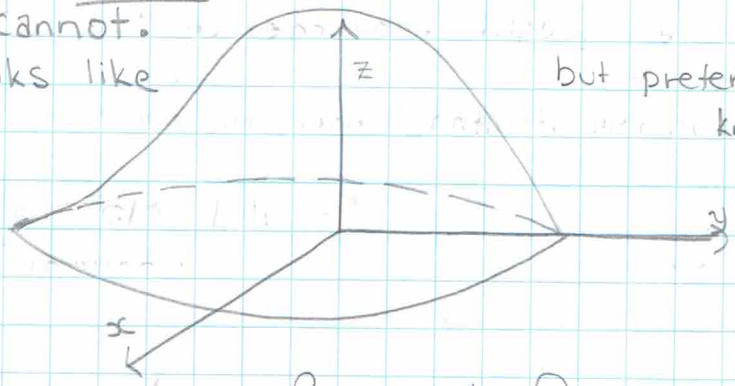
That is, we plot each chain individually. We see the following



After the point B, we have converged so we are drawing samples from a distribution. Samples are not always equal. (i.e. if I draw from $N(5,2)$ I would not always get 4.7.) This explains the wiggle in the line after convergence.

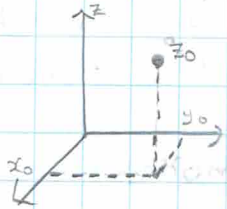
Without the notation, this is what is essentially happening:

- $f(x,y)$ is a density which we wish to sample from but we cannot.
  $f(x,y)$ looks like



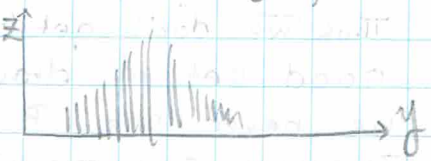
but pretend we don't know this.

Instead, we know $f(x|y)$ & $f(y|x)$.
So beginning with y_0, you sample from $f(x|y=y_0)$ to get x_0. So we get the point x_0, y_0 & evaluate $f(x_0,y_0)$. We plot this point



- We then iterate, i.e. $y_1 = f(y|x=x_0)$ & $x_1 = f(x|y=y_1)$ etc. Eventually we get a sample of the whole space:

If we forget about the x's and just look at our y's, then we have an estimate of $f(y)$:



Similarly we can drop our y's and just look at our x's to get an estimate of $f(x)$.

Going back to our $\{\theta_B, \ldots, \theta_{B+N}\}$, we can compute the following

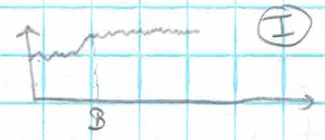
1) Quantile $[\theta|p] \approx$ Sort $\langle \theta_{SMALL}, \ldots \theta_{BIG} \rangle$ & take the % you want.

2) $\theta_{95\%\,CR}$ = [2.5% quantile, 97.5% quantile]

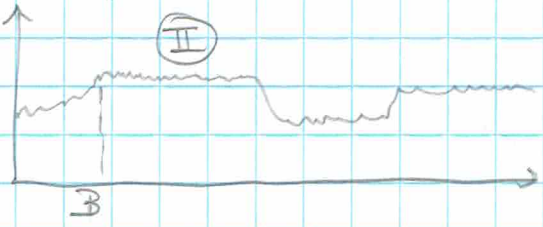
3) p-value: if $H_0: \theta > \theta_0$, $P(\theta|H_0) \approx \dfrac{\sum_t \mathbb{1}_{\theta_t > \theta_0}}{N}$

PROBLEMS OF GIBBS SAMPLER

① When did it converge? If we saw we might have thought convergence happened after B.



But if we run the iteration longer, we might see



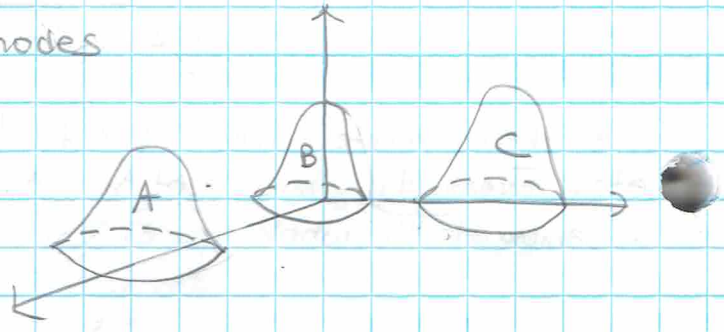
it did NOT actually converge!

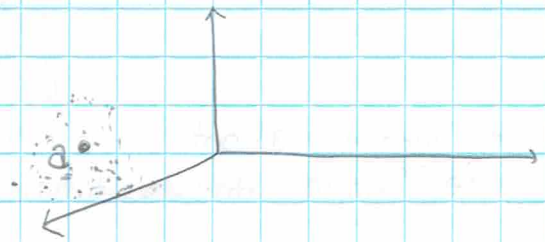
We say graph I has good mixing & graph II has bad mixing.

Def: Mixing → ability of the chain to effectively traverse the parameter space, i.e. the support.

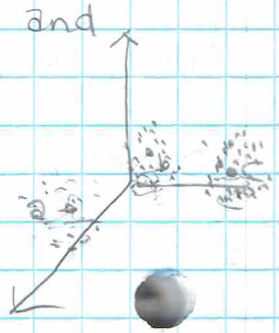
② Getting stuck in local modes

If our density looks like this, when we sample, we might get locked in one region. Ex: If we run Gibbs starting at the point a, then when we Gibbs sample, we only get draws from A. Thus we don't get back a good set of draws since we never see B & C.





PARTIAL SOLUTION: start at different places. So start at a & we get the above picture. Then start at b and runs Gibbs & then start at C and run Gibbs



③ AUTOCORRELATION

θ_{t+1} depends on θ_t
θ_t depends on θ_{t-1}
θ_{t-1} depends on θ_{t-2}

} SAMPLES ARE NOT independence.

Since we are not drawing independent samples, this affects quantiles & future samples. We will try & solve this next time.