ECON 512: Empirical Methods **Fall 2018/Spring 2019**

Lecture 5: Markov Chain Monte Carlo Methods

Instructor: Paul Grieco *Date: Feb 18/25*

These lecture notes are for my personal use. If you are reading them, I decided to distribute them as an experiment. There are typos and probably outright errors, if you find them please accept my apologies and report them to me.

These notes will introduce sampling methods leading to MCMC methods. These methods allow us to integrate over complicated distributions via simulation. They are most closely associated with Bayesian estimation but can be used in frequentist estimation as well.

- Cameron and Trivedi, Microeconometrics, Ch 13.
- Rossi, Allenby, MucCulloch, Bayesian Statistics and Marketing.
- Robert and Casella, Monte Carlo Statistical Methods.
- Chernozhukov and Hong (2003, Journal of Econometrics) for frequentist interpretation.

5.1 Quick Overview of Bayesian Estimation

In parametric estimation, we wish to know about an unknown parameter $\theta^0 \in R^N$, we have some data y, and we know that $L(y|\theta)$ is the likelihood of y given that $\theta = \theta^0$.

A frequentist approach is to derive an estimator (e.g., maximum likelihood) and analyze the statistical properties of that estimator.

$$\hat{\theta} = \max_{\theta} L(y|\theta).$$

Bayesians, on the other hand start with a prior belief $p(\theta)$ on the identity of the true parameter θ^0. Then the use the data to update their belief to a posterior distribution using Bayes rule. Note that the key ingredient here is the likelihood:

$$\pi(\theta|y) = \frac{L(y|\theta)p(\theta)}{f(y)}$$

where $f(y) = \int L(y|\theta)\pi(\theta)d\theta$ is the marginal distribution of y.

There is nothing fancy here, its just:

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)} = \frac{Pr(B|A)Pr(A)}{Pr(B)}.$$

For Bayesians, $\rho(\cdot)$ is the outcome of estimation, it summarizes everything we know about where θ is, and typically we report its moments. Unfortunately, except in special cases, $\rho(\cdot)$ is typically not tractable, so MCMC estimation steps in as a way to simulate draws from ρ that we can analyze.

Frequentists have many complaints about Bayesianism, largely because it is subjective, as exemplified by the presence of a prior. For a long time, priors had to be chosen very carefully so that $\rho(\cdot)$ was tractable.

However, even if you are a committed frequentist, Laplace estimators allow you to use MCMC techniques. They convert your objective function (which you would maximize) into a distribution (sometimes called a quasi-posterior) where your estimate is the mean of this distribution.

So we can use MCMC estimators without really fighting about Bayesian vs. Frequentist estimation, although they have pluses and minuses computationally:

- Good: They avoid the need to solve a complicated non-convex optimization problem.
- Bad: they necessitate a complex integration.

5.2 An Analytically Tractable Example

Adapted from (Cameron & Trivedi, p 422):

Here is a simple example of a Bayesian estimaton to show that there is nothing about a posterior per se that requires MCMC. Suppose you observe N draws from a normal distribution with mean θ and variance σ^2, e.g.,

$$y_i \sim N(\theta, \sigma^2).$$

Suppose further σ^2 is known, but you want to estimate θ. A frequentist might use maximum likelihood estimation. The likelihood is:

$$\begin{aligned}
L(y|\theta) &= \prod_{i=1}^{N} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{(y_i-\theta)^2}{2\sigma^2}\right\} \\
&= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\sum_{i=1}^{N}\frac{(y_i-\theta)^2}{2\sigma^2}\right\} \\
&\propto \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\bar{y}+\bar{y}-\theta)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\bar{y})^2\right\}\exp\left\{-\frac{N}{2\sigma^2}(\bar{y}-\theta)^2\right\} \\
&\propto \exp\left\{-\frac{N}{2\sigma^2}(\bar{y}-\theta)^2\right\}
\end{aligned}$$

This is clearly maximized at $\bar{y}$ which would be the frequentist's MLE estimate. However, a Bayesian would have a prior belief for θ, suppose that belief is normally distributed, with mean μ and variance τ^2, then we have a prior density:

$$p(\theta) = (2\pi\tau^2)^{-\frac{1}{2}} \exp\left\{-\frac{(\theta-\mu)^2}{2\tau^2}\right\}.$$

Following Bayes's rule, the Posterior density is proportional to:

$$\begin{aligned}\pi(\theta|y) &\propto L(y|\theta)p(\theta) \\ &\propto \exp\left\{-\frac{N}{2\sigma^2}(\bar{y}-\theta)^2\right\}\exp\left\{-\frac{(\theta-\mu)^2}{2\tau^2}\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{(\bar{y}-\theta)^2}{N^{-1}\sigma^2}+\frac{(\theta-\mu)^2}{\tau^2}\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{(\theta-\tilde{\mu})^2}{\tilde{\tau}^2}\right]\right\}\end{aligned}$$

Where,

$$\tilde{\mu} = \tilde{\tau}^2\left(\frac{N}{\sigma^2}\bar{y}+\frac{1}{\tau^2}\mu\right)$$

$$\tilde{\tau}^2 = \left(\frac{N}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}$$

The final line is the kernel of a normal distribution.[1] So we have found that the posterior is normally distributed with mean $\tilde{\mu}$ which is a weighted sum of $\bar{y}$ and the prior mean μ where weights relate to the precision of the prior and MLE.

Since the posterior is normal, it's easy to compute its moments. However, in general, we won't have such a clean result. So even though we can write down the density of π fairly easily, calculating its moments will require simulation.

But how? The rest of these notes will focus on this: How one might simulate a sample of draws from a distribution that does not have a simple closed form.

5.3 Monte Carlo Integration and Importance Sampling

5.3.1 Monte Carlo Integration using $\rho(\cdot)$ directly

The point of monte carlo integration is to use draws from a distribution to calculate the moments of $\rho(\theta|y)$. If $\rho(\cdot)$ is "easy" to draw from (say, uniform or normal) than we can use traditional monte carlo integration techniques:

$$E[m(\theta)] = \int_\Theta m(\theta)\rho(\theta|y)d\theta \approx \frac{1}{S}\sum_{s=1}^{S} m(\theta_s)$$

- $m(\cdot)$ is an arbitrary function, say identity if we want the mean. We need to assume this expectation exists (of course).
- θ_s is a draw from $\rho(\theta|y)$.

However, this isn't helpful if we don't know how to generate draws from $\rho(\cdot)$ and if we did, we could probably just integrate it directly.

[1] It can be obtained by completing the square of the second to last line, it's not pleasant and the full derivation is on Cameron & Trividi, p 443.

5.3.2 Importance Sampling

Let's say that $\rho(\cdot)$ is hard to draw from, however there is some distribution $g(\cdot)$ that is easy to draw from, and is close to ρ. Let's further say that g is absolutely continuous with respect to ρ: $g(\cdot) > 0$ on the support of ρ. Then,

$$E[m(\theta)] = \int_\Theta \frac{m(\theta)\rho(\theta|y)}{g(\theta)} g(\theta) d\theta \approx \frac{1}{S} \sum_{i=s}^{S} m(\theta_s) \frac{\rho(\theta_s|y)}{g(\theta_s)}$$

where θ_s is a draw from $g(\theta)$.

Importance sampling simply draws from $g(\cdot)$ and then re-weights draws according to relative density of the posterior and g. It will work great *as long as g is a good approximation to ρ.*

- If $\frac{\rho}{g}$ is low, then we down weight these draws, leading to inefficiency of the simulator.
- If $\frac{\rho}{g}$ is high, then we upweight when we see these draws, but we won't see them very often also leading to inefficiency.

The "ideal" g would simply be ρ, but the whole point is that drawing from ρ isn't tractable.

In practice, particularly when θ is high dimensional, finding a "good" candidate for g is hard, if different g's generate different results, we know we have a problem. This is where MCMC methods can help.

5.4 Markov Chain Monte Carlo

So far, we've only considered methods which take *iid* draws from some distribution. MCMC methods draw from a Markov chain instead. We use this when:

- Analytic solutions aren't tractable.
- IID sampleing doesn't give adequate covergage (perhaps dimension is too high or good approximation of ρ is unknown).

The goal becomes constructing an ergodic Markov Chain F (so that the stationary distribution exists) such that the stationary distribution is exactly ρ. If we do this then we can generate moments of ρ from

$$E[m(\theta)] \approx \frac{1}{S} \sum_{i=1}^{S} m(\theta_i)$$

where $\theta_i \sim F(\cdot|\theta_{i-1})$.

5.5 A Little Markov Chain Theory

To keep things simple, let the state space for θ be discrete, $\Theta = \{\theta^{(1)}, \ldots, \theta^{(K)}\}$. Of course this isn't reasonable for estimation but it allows me to skip over a bunch of measure theory. Let our chain be defined by

$$P(\theta_{r+1} = \theta^{(j)} | \theta_r = \theta^{(i)}) = p_{ij}$$

So the Markov transition matrix is,

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & & & \vdots \\ p_{K1} & p_{K2} & \cdots & p_{KK} \end{bmatrix}$$

5.5.1 Stationarity

Let π_0 be an initial distribution over states (a $1 \times K$ vector). Then the distribution over states after 1 period will be:

$$Pr(\theta_1 = \theta^{(j)}) = \sum_{i=1}^{K} Pr(\theta_0 = \theta^{(i)}) p_{ij} = \sum_{i=1}^{K} \pi_{0i} p_{ij}$$

Or in matrix notation for the entire distribution,

$$\pi_1 = \pi_0 P$$

If for all i, j: $p_{ij} > 0$, then every state will be visited infinitely often and then chain is irreducible. In the discrete case, irreducibility implies the chain is ergodic, which means a stationary distribution exists and is unique:[2]

$$\lim_{r \to \infty} \pi_0 P^r = \pi$$

for any π_0 and of course,

$$\pi = \pi P$$

The stationary distribution π is sometimes called the invariant distribution.

5.5.2 Time Reversibility

Time reversibility is a symmetry property of a Markov chain. We'll need it because it gives us another way to prove a chain is generating the stationary distribution we want. It is at the heart of the Metropolis-Hastings algorithm.

Definition 5.1 *A chain is time reversible with respect to π if it has the same behavior backwards and forwards starting from π. That is if the chance of seeing a transition from i to j is the same as seeing a transition from j to i:*

$$\pi_i p_{ij} = \pi_j p_{ji}$$

First, what does it mean to "run a Markov Chain backwards?", it's not even clear this sequence is a Markov chain, but it turns out that it is:

$$\begin{aligned} Pr(\theta_r | \theta_{r+1}, \theta_{r+2}, \ldots) &= \frac{Pr(\theta_r, \theta_{r+1}, \theta_{r+2}, \ldots)}{Pr(\theta_{r+1}, \theta_{r+2}, \ldots)} \\ &= \frac{Pr(\theta_r) Pr(\theta_{r+1} | \theta_r) Pr(\theta_{r+2}, \ldots | \theta_r, \theta_{r+1})}{Pr(\theta_{r+1}) Pr(\theta_{r+2}, \ldots | \theta_{r+1})} \\ &= \frac{Pr(\theta_r) Pr(\theta_{r+1} | \theta_r)}{Pr(\theta_{r+1}))} \end{aligned}$$

[2] This condition is stronger than we need, but is typically satisfied given that we get to design the chain.

Where the last equality follows from the fact that $Pr(\theta_{r+2}, \ldots | \theta_r, \theta_{r+1}) = Pr(\theta_{r+2}, \ldots | \theta_{r+1})$ because the sequence θ_r is a generated by a Markov chain going forward. This shows you the chain is also a markov chain going backwards since θ_r depends only on θ_{r+1}.

This backwards chain has transition probabilities:

$$p^*_{ij} = \frac{\pi_j p_{ji}}{\pi_i}$$

Time reversibility with respect to π is the property that for a given distribution of states $p^*_{ij} = p_{ij}$ which implies:

$$\pi_i p_{ij} = \pi_j p_{ji}$$

Theorem 5.2 *If a chain is time reversible with respect to θ, then θ is the stationary distribution of the chain.*

Proof: Time reversibility means,

$$\pi_i p_{ij} = \pi_j p_{ji}$$

Sum these equations over i:

$$\sum_i \pi_i p_{ij} = \pi_j \sum_i p_{ji}$$

$$\pi P = \pi.$$

All of this extends to continuous states, see Rossi, page 62 or other sources for details.

Notice that this does not imply that all stationary distributions are time reversible. We will be designing time reversible chains, so this is all we need.

5.6 Gibbs Sampling

A Gibbs sampler is a Markov chain constructed by cycling through conditional distributions related to π. It is used to deal with problems drawing from a high dimensional space.

Suppose θ can be divided into subvectors $\theta = (\theta^1, \ldots \theta^P)$. While it is difficult to draw from the joint distribution of θ, conditional distributions, such as $f_1(\theta^1 | \theta^2, \ldots, \theta^P)$ are tractable. Then the Gibbs sampling algorithm suggests:

1. Given initial state θ_0
2. For $1 \leq s \leq S$:
   (a) Draw θ^1_r from $f_1(\theta^1_r | \theta^2_{r-1}, \ldots, \theta^P_{r-1})$.
   (b) Draw θ^2_r from $f_2(\theta^2_r | \theta^1_r, \theta^3_{r-1}, \ldots, \theta^P_{r-1})$.
   (c) $\vdots$
   (d) Draw θ^P_r from $f_2(\theta^P_r | \theta^1_r, \ldots, \theta^{P-1}_r)$.
3. Compute:

$$E[m(\theta)] = \frac{1}{S} \sum_{s=B}^{S} m(\theta_s)$$

   Where $B < S$ is a suitable burn-in period to ensure chain has converged.

5.6.1 Why it works:

- Gibbs sampler is clearly a Markov Chain, ergodic as long as conditional distributions have full support.
- Invariant distribution is the joint distribution of θ:
  - Assume θ_{r-1} is a draw from π, then θ^i_{r-1} is a draw from the marginal distribution
$$\pi_i(\theta_i) = \int \pi(\theta_i, \theta_{-i}) d\theta_{-i}.$$
  - So then in our first draw we have:
$$\theta^1_r \sim \int f_1(\theta_1|\theta_{-1})\pi_{-i}(\theta_{-i})d\theta_{-i} = \pi_1(\theta_1)$$
    So we have a draw from the marginal distribution of θ^1.
  - Repeating for $2, \dots P$ gives us a new draw from the joint distribution.

5.6.2 Where is it useful?

There are actually quite a few cases where conditional distributions are conjugate but joints are not:[3]

- Linear regression with unknown variance.
- Hierarchical Linear Models.
- Latent Variable Models (e.g., Probit)

Here let's think about a Probit model, which is the simplest form of data augmentation. Given the model:

$$z_i = x_i\beta + \epsilon_i$$
$$y_i = \begin{cases} 0 & z_i \leq 0 \\ 1 & z_i > 0 \end{cases}$$
$$\epsilon_i \sim N(0,1)$$

We observe a random sample of (y_i, x_i) and want to estimate β.

Suppose we have a prior $\beta \sim N(\bar{\beta}, A^{-1})$. If we observed z_i then the posterior would be normal (normal is the conjugate prior of normal). However, we don't so we use a Gibbs sampler with two blocks (z_i, β):

1. Given β_{r-1}, draw z_i by drawing from a truncated normal:
$$z_{i,r}|\beta_{r-1}, y_i, x_i \sim \text{TruncatedNormal}^b_a(-x_i\beta_{r-1}, 1)$$
   Where bounds are $a = 0, b = \infty$ if $y_i = 1$ and $a = -\infty, b = 0$ if $y_i = 0$
2. Draw $\beta_r|z_{i,r}, x_i$ from the posterior of a regression of z on x:
$$\beta_r \sim N(\tilde{\beta}, (X'X + A)^{-1})$$
   where $\tilde{\beta} = (X'X + A)^{-1}(X'z + A\bar{\beta})$.
3. After many draws, we have a sample of β_r which we use as draws from the stationary distribution.

[3] A conjugate prior is one which, when multiplied against a likelihood, produces a tractable distribution.

5.6.3 Example

The file `simpleGibbs.m` implements Gibbs sampling to draw from a bivariate normal:

$$(y_1, y_2)' \sim N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$$

This trivially implies conditional distributions:

$$y_1|y_2 \sim N(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2), \sigma_1^2(1-\rho^2))$$

Let's go through the code and see how it works: Items to consider:

- See what happens when you start from (50, 50).
- See what happens when you raise ρ to .95.
- Choosing a burn-in period.
- The relationship between correlation and convergence.
- Checking autocorrelation of the chain:

$$s_{\theta_i}(k) = \frac{\sum_{r=k+1}^{R}(\theta_r - \bar{\theta})(\theta_{r-k} - \bar{\theta})}{\sum_r = 1^R(\theta_r - \bar{\theta})^2}$$

- Can also check for convergence by computing moments from different subsamples of the chain and making sure they are robust.

5.6.4 Has my chain converged?

While MCMC measures have nice convergence properties in the limit, it's hard to know how long to run the chain. The more complicated the chain (say in the dimensionality of the state space, the harder it is to make sure the chain has converged.

Some rules of thumb:

- Use a burn in period- Throw away the initial B draws of the chain to try to eliminate the effect of the starting point.
- Plot a time series of the draws along key dimensions to make sure they do not have a trend. (One could even regress draws against a time trend).
- Run chain from several start points and compare results.
- Take different subsamples of the same chain draws and compare results.
- Compute the autocorellation function:

$$s_{\theta_i}(k) = \frac{\sum_{r=k+1}^{R}(\theta_r - \bar{\theta})(\theta_{r-k} - \bar{\theta})}{\sum_r = 1^R(\theta_r - \bar{\theta})^2}$$

  To make sure correlation is dying as time between draws increases.

At the end of the day, there is no "proof" of convergence. This is analogous to finding the global optimium of a non-convex function. Sadly, MCMC is *not* a free lunch.

5.7 Metropolis-Hastings Algorithm

Gibbs sampling is great if θ can be segmented into easy-to-draw from conditionals. However, that is not always the case. Metropolis-hastings gives us a general way to operationalize MCMC.

Idea from importance sampling: draw from a known distribution and re-weight.

For simplicity, let's again do everything with discrete states. Suppose we want to draw from π. But we only know how to draw from some candidate distribution q_i (a known density that is potentially conditional on current state). That is:

$$q_{ij} = Pr(\theta^{(j)}|\theta^{(i)})$$

The MH algorithm generates a Time-Reversible Markov chain with respect to π by following:

1. Initialize θ_0 to some initial state.
2. For $t = 1 : T$:
   (a) Let i be such that $\theta_{t-1} = \theta^{(i)}$, draw a candidate state $\tilde{\theta}$ from q_i,
   (b) Let j be such that $\tilde{\theta} = \theta^{(j)}$, compute:

$$\alpha_{ij} = \min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\}$$

   (c) Draw:

$$\theta_t = \begin{cases} \tilde{\theta} & \text{w.p. } \alpha_{ij} \\ \theta_{t-1} & \text{w.p. } 1 - \alpha_{ij} \end{cases}$$

Theorem 5.3 *The Markov Chain $P = [p_{ij}]$ generated by the Metropolis-Hastings algorithm is time reversible with respect to π.*

To see why:

$$p_{ij} = q_{ij}\alpha_{ij} = q_{ij} \min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\}$$

Therefore

$$\pi_i p_{ij} = \pi_i q_{ij} \min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\} = \min\left\{\pi_i q_{ij}, \pi_j q_{ji}\right\}$$

Which equals,

$$\pi_j p_{ji} = \pi_j q_{ji} \min\left\{1, \frac{\pi_i q_{ij}}{\pi_j q_{ji}}\right\} = \min\left\{\pi_j q_{ji}, \pi_i q_{ij}\right\}$$

Continuous state version just replaces probabilities with densities and deals with measure theory issues.

Usually, we will use symmetric proposals so that $q_{ij} = q_{ji}$, then notice that the probability of choosing a state is related directly to the ratio of posterior density. We always transition to higher equal or higher density points, but sometimes transition to lower density points.

5.7.1 Gibbs Sampling as a Special Case (optional)

It turns out that Gibbs Sampling is a special case of the MH algorithm where the conditional distributions are the proposals. To see, this, let $\pi(\theta_1, \theta_2)$ be the joint distribution of (θ_1, θ_2) while $m(\theta_i)$ is the marginal and $c(\theta_i|\theta_{-i})$ is the conditional distribution of θ_i given θ_{-i}. Then given θ_2, Gibbs sampling will draw:

$$\tilde{\theta}_1 \sim c(\cdot|\theta_2)$$

and the new point will be $\tilde{\theta} = (\tilde{\theta}_1, \theta_2)$. Following the MH example, compute α:

$$\alpha = \frac{\pi(\tilde{\theta}_1, \theta_2)c(\theta_1|\theta_2)}{\pi(\theta_1, \theta_2)c(\tilde{\theta_1}|\theta_2)} = \frac{c(\tilde{\theta}_1|\theta_2)m(\theta_2)c(\theta_1|\theta_2)}{c(\theta_1|\theta_2)m(\theta_2)c(\tilde{\theta_1}|\theta_2)} = 1$$

So acceptance rate is always 1, which is why we didn't have to deal with it. The first equality follows because a transition from $\tilde{\theta}$ to θ occurs when θ_2 is fixed in the Gibbs sampler and θ_1 is drawn. The second equality applies the fact that the joint distribution is the conditional times the marginal.

5.8 Metropolis Hastings Estimation Example

Let's do a very simple MCMC estimation of a mean and variance.

5.8.1 Setup:

Need to clean up notation here.

Suppose we have data $\{y_1, \ldots, y_N\}$ from a DGP:

$$y_i \sim N(\mu, \sigma^2).$$

We wish to estimate (μ, σ). The likelihood is,

$$\begin{aligned} L(y|\mu, \sigma) &= \prod_{i=1}^{N} f(y_i|\mu, \sigma) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{y_i - \mu}{\sigma}\right)^2\right\} \\ &= \prod_{i=1}^{N} \phi\left(\frac{y_i - \mu}{\sigma}\right) \end{aligned}$$

Let's assume we have a prior belief on the parameters:

$$\begin{pmatrix}\mu \\ \sigma\end{pmatrix} \sim N\left(\begin{pmatrix}\bar{\mu} \\ \bar{\sigma}\end{pmatrix}, \Omega\right)$$

It's an odd prior since we believe it is possible σ is negative, but people have weird beliefs these days. This implies the prior density is:

$$f(\mu, \sigma) = (2\pi)^{-1}|\Omega|^{-\frac{1}{2}} \exp\left\{\begin{pmatrix}\mu - \bar{\mu} \\ \sigma - \bar{\sigma}\end{pmatrix}' \Omega^{-1} \begin{pmatrix}\mu - \bar{\mu} \\ \sigma - \bar{\sigma}\end{pmatrix}\right\}$$

And the posterior is,

$$\pi(\mu, \sigma|y) \propto L(y|\mu, \sigma) f(\mu, \sigma).$$

Where $\propto$ means "proportionate to" we know there is a constant of integration that depends only on y, but we don't need to worry about it.

This is a well-defined distribution, but it is not normal, so we don't really know how to draw from it. Moreover, the conditional $\sigma|\beta, y$ does not have a closed form either.

5.8.2 Algorithm

We will draw from this posterior using a random walk distribution Metropolis-Hastings algorithm. Let $\theta = (\mu, \sigma)'$, then our proposal is:

$$\tilde{\theta}_t \sim \theta_{t-1} + \xi_t$$

where

$$\xi_t \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma_p\right)$$

Notice that this proposal is symmetric, that will be convenient.

As usual, the likelihood itself is going to be really small, so we will use the log likelihood everywhere. Our algorithm is:

1. Start with θ_0
2. Compute the log posterior of θ_0:

$$\pi(\theta_0) = \sum_{i=1}^{N} \log \phi\left(\frac{y_i - \mu_0}{\sigma_0}\right) + \log f(\theta_0)$$

3. For $1 \leq t \leq T$:
   (a) Draw ξ_t, compute $\tilde{theta}_t = \theta_{t-1} + \xi_t$.
   (b) Compute $\pi(\tilde{\theta}_t)$ and $\log \alpha = \pi(\tilde{\theta}_t) - \pi(\theta_0)$
   (c) Draw $u \sim Unif(0, 1)$
   (d) Define

$$\theta_t = \begin{cases} \tilde{\theta}_t & \log(u) < \log \alpha \\ \theta_{t-1} & \text{otherwise} \end{cases}.$$

4. Compute moments using the sample $\{\theta_B, \ldots, \theta_T\}$ where B is the burn in period.

5.8.3 Go over Code

This algorithm is implemented in `metroHastMain.m` in the class repository. Some things to keep in mind:

- Look at the acceptance rate, but you don't want to to be too close to 0 or 1. If it is 0, you have perfect auto-correlation. If it is 1 you probably haven't converged.
- There are "rules of thumb" that say acceptance should be between 20 and 70 percent. Your milage may vary.

- Proposal densities will also affect performance.
- As will start point.