# Gibbs sampling

## Bayesian Methodology in Biostatistics (BST 249)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

# Outline

# Outline

# Introduction

- In many real-world applications, we have to deal with complex distributions on complicated high-dimensional spaces.
- On rare occasions, it is possible to sample exactly from the distribution of interest, but typically exact sampling is not feasible.
- Further, high-dimensional distributions are hard to visualize, making it difficult to even guess where the regions of high probability are located.
- As a result, it may be challenging to even design a reasonable proposal distribution to use with importance sampling.

# Introduction

- Markov chain Monte Carlo (MCMC) is a sampling technique that works remarkably well in many situations like this.

- MCMC constructs a sequence of correlated samples $X_1, X_2, \ldots$ that meander through the region of high probability by making a sequence of incremental movements.

- Even though the samples are not independent, it turns out that when constructed properly,

$$\mathrm{E}h(X) \approx \frac{1}{N} \sum_{i=1}^{N} h(X_i)$$

  as in the case of simple Monte Carlo approximation.

- By a powerful result called the ergodic theorem, these approximations are guaranteed to converge to the true value.

# Introduction

- Advantages of MCMC:
  - applicable even when we can't directly draw samples.
  - works for complicated distributions in high-dimensional spaces.
  - relatively easy to implement.
  - fairly reliable.

- Disadvantages of MCMC:
  - slower than simple Monte Carlo or importance sampling, i.e., more samples are often needed to attain the same accuracy.
  - can be very difficult to assess accuracy and convergence.

- Since it is easy to use, MCMC is often used out of convenience, even when better methods exist.

- MCMC opens up a world of possibilities, allowing us to work with far more interesting and realistic models than we could without it.

# Introduction

- The two main ways of constructing MCMC algorithms are:
  1. Gibbs sampling, and
  2. the Metropolis–Hastings algorithm.

- We'll start with Gibbs sampling (Geman & Geman, 1984) since it is easiest to understand.

- Later, we will also consider more advanced MCMC algorithms.

- We'll illustrate with some examples involving Gibbs sampling:
  - Normal with semi-conjugate priors,
  - Censored data / missing data,
  - Hyperpriors and hierarchical models.

# Outline

# Gibbs sampling with two variables

- Suppose $p(x, y)$ is difficult to sample from directly.
- Suppose, though, that we can easily sample from the conditional distributions $p(x|y)$ and $p(y|x)$.
- *Gibbs sampling*: Initialize $x$ and $y$ and iteratively repeat
  1. update $x$ by sampling from $x|y$, and
  2. update $y$ by sampling from $y|x$.
- Each iteration through all variables ($x$ and $y$, in this case) is referred to as a *sweep* or *scan*.
- When updating a variable, we always use the most recent value of the other variables.

# Gibbs sampling with two variables

- This algorithm generates a sequence of pairs of r.v.s

$$(X_0, Y_0), (X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots$$

  that is a *Markov chain*: the distribution of $(X_i, Y_i)$ given all of the previous values depends only on $(X_{i-1}, Y_{i-1})$.

- Under quite general conditions, the *ergodic theorem* guarantees that for any $h(x, y)$ such that $\mathrm{E}|h(X, Y)| < \infty$,

$$\frac{1}{N} \sum_{i=1}^{N} h(X_i, Y_i) \longrightarrow \mathrm{E}h(X, Y)$$

  as $N \to \infty$, with probability 1, where $(X, Y) \sim p(x, y)$.

- This justifies the use of $\frac{1}{N} \sum_{i=1}^{N} h(X_i, Y_i)$ as an approximation to $\mathrm{E}h(X, Y)$, like a simple Monte Carlo approximation, even though $(X_1, Y_1), (X_2, Y_2), \ldots$ are not i.i.d.

# Burn-in period

- If the starting point $(x_0, y_0)$ is far from the region of high probability under $p(x, y)$, it may take a while for the chain to get to a good place.

- During this *burn-in period*, the distribution of the samples $(X_i, Y_i)$ does not approximate $p(x, y)$.

- Thus, it is recommended to run the chain for a while before starting to compute sample averages.

- In other words, discard $(X_1, Y_1), \ldots, (X_B, Y_B)$ and only use $(X_{B+1}, Y_{B+1}), \ldots$ for inference. For example,

$$\frac{1}{N-B} \sum_{i=B+1}^{N} h(X_i, Y_i).$$

- How to choose $B$? Traceplots and running averages (see below) are useful for assessing the burn-in.

# Mixing

- Roughly speaking, the performance of an MCMC algorithm—that is, how quickly the sample averages $\frac{1}{N}\sum_{i=1}^{N} h(X_i, Y_i)$ converge—is referred to as the *mixing rate*.
- An algorithm with good performance is said to "have good mixing" or to "mix well".
- There are various empirical diagnostics for assessing burn-in and mixing, but none are foolproof.
- Sometimes, a chain may appear to be mixing well, but if you ran it longer you would see that it was actually doing poorly.

# Toy example

- Suppose we need to sample from

$$p(x, y) \propto e^{-xy} \mathrm{I}(x, y \in (0, c))$$

where $c > 0$, and $(0, c)$ is the open interval between $0$ and $c$.
  - This example is due to Casella & George, 1992.

- Gibbs sampling: Iteratively sample from $p(x|y)$ and $p(y|x)$.

- Let's look at $p(x|y)$:

$$p(x|y) \underset{x}{\propto} ???.$$

(Whiteboard activity)

# Toy example

- Suppose we need to sample from

$$p(x, y) \propto e^{-xy} \mathrm{I}(x, y \in (0, c))$$

where $c > 0$, and $(0, c)$ is the open interval between $0$ and $c$.
  - This example is due to Casella & George, 1992.

- Gibbs sampling: Iteratively sample from $p(x|y)$ and $p(y|x)$.

- Let's look at $p(x|y)$:

$$p(x|y) \underset{x}{\propto} p(x, y) \underset{x}{\propto} e^{-xy}\, \mathrm{I}(0 < x < c) \underset{x}{\propto} \mathrm{Exp}(x|y)\mathrm{I}(x < c).$$

- So, $p(x|y)$ is a truncated $\mathrm{Exp}(y)$ distribution, $\mathrm{TExp}(y, (0, c))$.

- By symmetry, $p(y|x) = \mathrm{TExp}(x, (0, c))$.

# Toy example

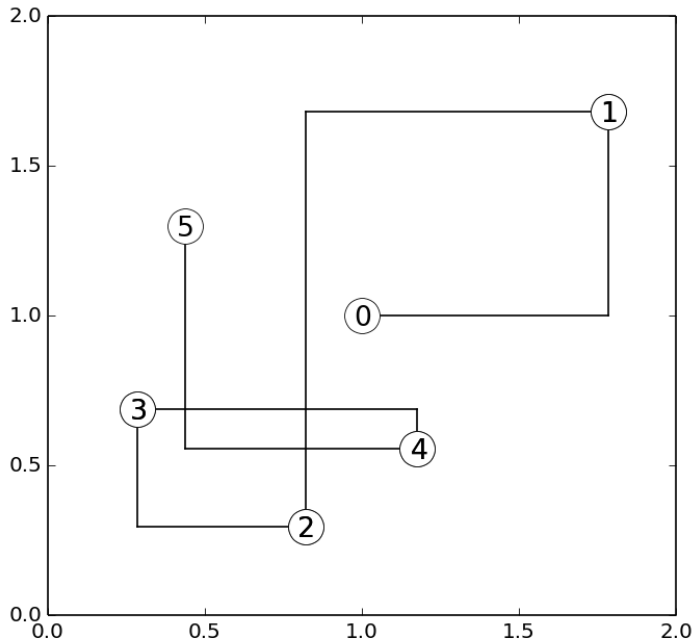
- Denote $S = (0, c)$ for brevity.

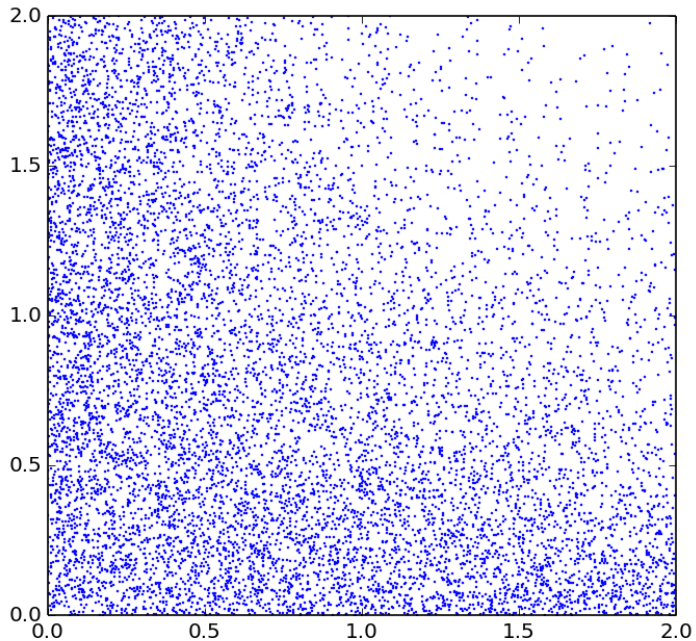
- Gibbs sampler algorithm:
  0. Initialize $x_0, y_0 \in S$.
  1. Sample $x_1 \sim \mathrm{TExp}(y_0, S)$, then sample $y_1 \sim \mathrm{TExp}(x_1, S)$.
  2. Sample $x_2 \sim \mathrm{TExp}(y_1, S)$, then sample $y_2 \sim \mathrm{TExp}(x_2, S)$.

     $\vdots$

  $N$. Sample $x_N \sim \mathrm{TExp}(y_{N-1}, S)$, then sample $y_N \sim \mathrm{TExp}(x_N, S)$.

# Toy example




- Demonstration with $c = 2$ and initial point $(x_0, y_0) = (1, 1)$.
- (Left plot) First 5 Gibbs sampling iterations/sweeps/scans.
- (Right plot) Scatterplot of $10^4$ Gibbs sampling iterations.

# Toy example

- How to sample from a truncated exponential distribution?
- Here's an easy way based on the inverse c.d.f. method.
- The c.d.f. and inverse c.d.f. of $\mathrm{Exp}(\theta)$ are:

$$F(x|\theta) = 1 - e^{-\theta x}$$
$$F^{-1}(u|\theta) = -(1/\theta)\log(1-u)$$

  for $x > 0$ and $u \in (0, 1)$.

- Let $U \sim \mathrm{Uniform}(0, F(c|\theta))$ and let $Z = F^{-1}(U|\theta)$.
- Then $Z \sim \mathrm{TExp}(\theta, (0, c))$.

# Group activity: Check your understanding

Go to breakout rooms and work together to answer these questions:
https://forms.gle/nmFvWuSZJW3P9nfX9

(Three people per room, randomly assigned. 15 minutes.)

# Outline

# Example: Normal with semi-conjugate prior

- Consider an i.i.d. multivariate normal model

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \Lambda^{-1}).$$

- Assume conditionally conjugate (a.k.a., semi-conjugate) priors:

$$\boldsymbol{\mu} \sim \mathcal{N}(m, L^{-1}) \qquad \boldsymbol{\Lambda} \sim \text{Wishart}(S^{-1}, \nu)$$

  independently.

- A Gibbs sampler for $\mu, \Lambda \mid x_{1:n}$ is to iteratively:
  1. update $\mu$ by sampling from $\boldsymbol{\mu} \mid \Lambda, x_{1:n}$, and
  2. update $\Lambda$ by sampling from $\boldsymbol{\Lambda} \mid \mu, x_{1:n}$.

- From before, we know that

$$\boldsymbol{\mu} | \Lambda, x_{1:n} \sim \mathcal{N}(m_n, L_n^{-1})$$
$$\boldsymbol{\Lambda} | \mu, x_{1:n} \sim \text{Wishart}(S_n^{-1}, \nu_n)$$

  with $m_n, L_n, S_n$, and $\nu_n$ as in the slides on Gaussian models.

# Gibbs sampling with multiple variables

- More generally, Gibbs sampling is done by updating each variable in turn, given everything else.
- In each update, we always use the most recent values of all other variables.
- The conditional distribution of a variable given everything else is referred to as the *full conditional*.
- $\theta|\cdots$ denotes the full conditional of a variable $\theta$.
- The order in which variables are updated is usually fixed ("fixed scan") but more general schemes are also possible, e.g., random scan.

# Outline

# Example: Censored data

- Often, some data is missing or partially obscured.
- Gibbs sampling provides a natural Bayesian method for dealing with missing data:
  - Treat missing variables just like an unknown parameter, and update them in each Gibbs iteration.
  - As a side benefit, this also allows us to infer the missing data.
- Censoring is one way in which data can be partially obscured.
  - *Censoring* occurs when we know a data point lies in a particular interval, but we don't get to observe it exactly.
  - For instance, in medical research, some patients may be lost to follow-up during the study.
  - Another example: Measurements may exceed the lower/upper limits of the instrument being used.

# Censoring example: Data

- Suppose researchers are studying the length of life (survival) following a new medical intervention.

- In a study of 12 patients, the survival times (in years) are

$$3.4, 2.9, 1.2+, 1.4, 3.2, 1.8, 4.6, 1.7+, 2.0+, 1.4+, 2.8, 0.6+$$

  where $x+$ indicates that the patient was alive after $x$ years, but the researchers lost contact with the patient at that point.

- Usually, there would be a control group too, but let's focus on one group to keep things simple.

# Censoring example: Model

- Consider the following model:

$$\boldsymbol{\theta} \sim \text{Gamma}(a, b)$$

$$Z_1, \ldots, Z_n | \theta \overset{\text{iid}}{\sim} \text{Gamma}(r, \theta)$$

$$X_i = \begin{cases} Z_i & \text{if } Z_i \leq c_i \\ * & \text{if } Z_i > c_i. \end{cases}$$

where $a$, $b$, and $r$ are known, and $*$ is a special value to indicate that censoring has occurred. The interpretation is:

  - $\theta$ is the parameter of interest—the rate parameter for the survival distribution.
  - $Z_i$ is the survival time for patient $i$, however, this is not directly observed.
  - $c_i$ is the censoring time for patient $i$, which is fixed, but known only if censoring occurs.
  - $X_i$ is the observation—if the survival time is less than $c_i$ then we get to observe it ($X_i = Z_i$), otherwise all we know is that the survival time is greater than $c_i$ ($X_i = *$).

# Censoring example: The posterior is complicated

- Unfortunately, the posterior $p(\theta|x_{1:n}) \propto p(x_{1:n}|\theta)p(\theta)$ does not reduce to a simple form that we can easily work with.

- The reason is that the $p(x_{1:n}|\theta)$ involves the distribution of the observations $x_i$ given $\theta$, integrating out the $z_i$'s.

- In the case of censored observations $x_i = *$, we have

$$p(x_i|\theta) = \mathbb{P}(X_i = * \mid \theta) = \mathbb{P}(Z_i > c \mid \theta),$$

  which involves the incomplete gamma function.

- Also, $p(z_{1:n}|x_{1:n})$ (the posterior on the $z_i$'s, with $\theta$ integrated out) looks a bit nasty as well.

- Thus, it is not obvious how to sample directly (i.i.d.) from this posterior.

# Censoring example: Gibbs sampler

- Meanwhile, the Gibbs sampling approach is a cinch.
- We cycle through the full conditional distributions,

$$\begin{aligned}
\theta \mid z_{1:n}, x_{1:n} \\
z_1 \mid \theta, z_{-1}, x_{1:n} \\
z_2 \mid \theta, z_{-2}, x_{1:n} \\
\vdots \\
z_n \mid \theta, z_{-n}, x_{1:n}
\end{aligned}$$

  sampling from each in turn. (Recall: $z_{-j}$ = all $z$'s except $z_j$.)

- The full conditionals are easy to calculate. First, consider $\theta$:

$$p(\theta | \cdots) \propto p(x_{1:n}, z_{1:n}, \theta) \underset{\theta}{\propto} \text{Gamma}\left(\theta \,\middle|\, a + nr,\, b + \textstyle\sum_{i=1}^n z_i\right).$$

# Censoring example: Gibbs sampler

- Now, consider $z_i | \cdots$
  - If $x_i \neq *$ then $z_i$ is forced to be equal to $x_i$.
  - Otherwise,

$$\begin{aligned} p(z_i | \cdots) &\propto p(x_{1:n}, z_{1:n}, \theta) \\ &\underset{z_i}{\propto} p(x_i | z_i) p(z_i | \theta) \\ &= \mathrm{I}(z_i > c_i) \, \mathrm{Gamma}(z_i \mid r, \theta) \\ &\underset{z_i}{\propto} \mathrm{TGamma}(z_i \mid r, \theta, (c_i, \infty)) \end{aligned}$$

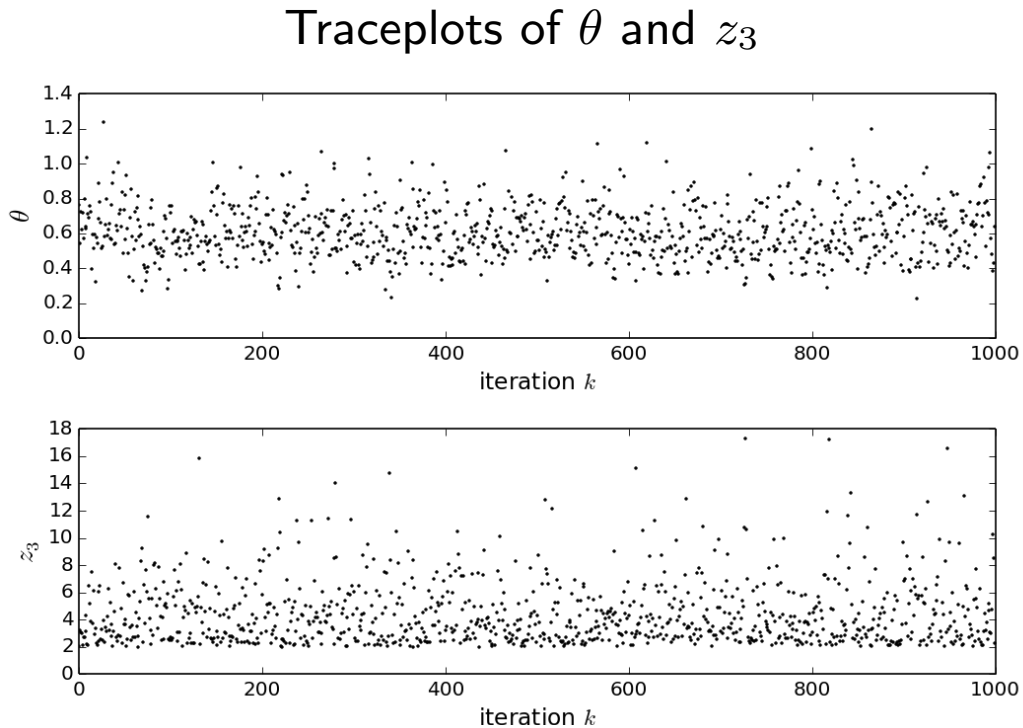
where $\mathrm{TGamma}$ is the truncated Gamma distribution.

- We can sample from $\mathrm{TGamma}(r, \theta, (c, \infty))$ with the same technique we used for the truncated exponential:
  - Let $F(x|r, \theta)$ denote the $\mathrm{Gamma}(r, \theta)$ c.d.f.
  - Let $U \sim \mathrm{Uniform}(F(c|r, \theta), 1)$, and let $V = F^{-1}(U|r, \theta)$.
  - Then $V \sim \mathrm{TGamma}(r, \theta, (c, \infty))$.

# Censoring example: Results

- For the hyperparameters, let's assume $a = b = 1$ and $r = 2$.
- To illustrate, let's run the sampler for $N = 10^3$ iterations.
- For the starting values, let's set $\theta = 1$ and $z_i = c_i + 1$ for those $i$'s that were censored. (I chose these pretty much arbitrarily.)
- Some diagnostic plots are useful to visualize the MCMC run:
  - Traceplots
  - Running averages
  - Histograms

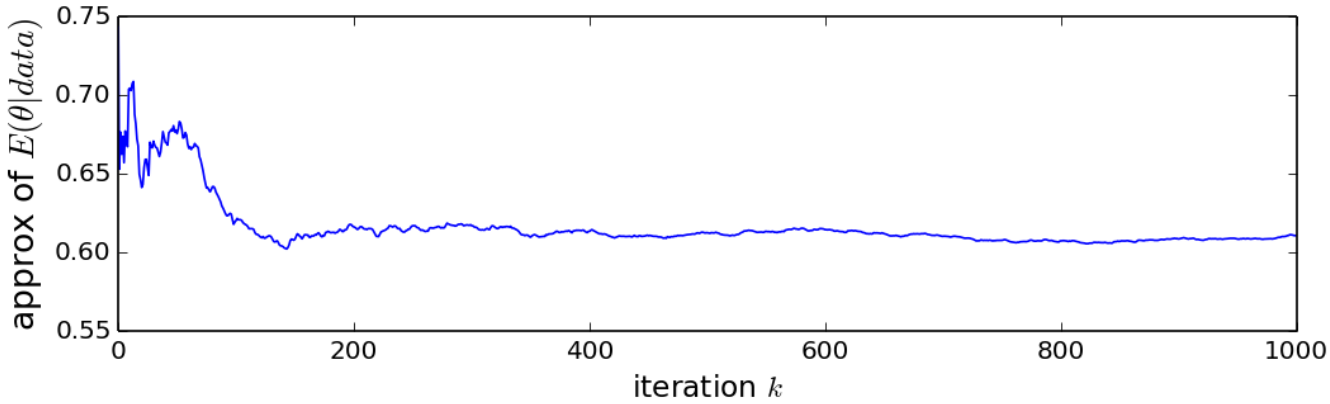
# Censoring example: Traceplots




- A *traceplot* simply shows the sequence of samples, for instance, $\theta_i$ versus $i$.
- Traceplots are a simple but very useful way to visualize how the sampler is behaving.
- The traceplots above look good — the sampler doesn't appear to be getting stuck anywhere.

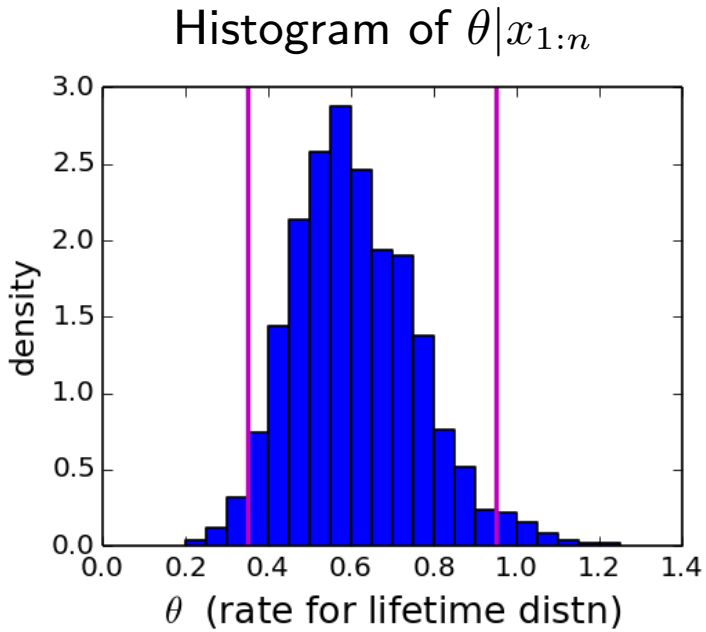
# Censoring example: Running averages

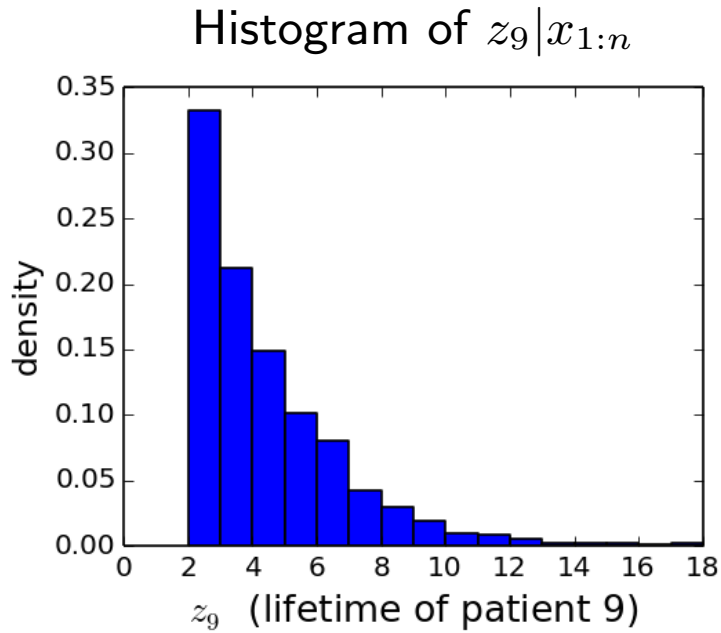
Running averages $\frac{1}{k}\sum_{i=1}^{k}\theta_i$ for $k = 1, \ldots, N$




- Running averages are another useful heuristic for assessing MCMC convergence.
- Running averages can be expected to drift and wander a bit and then settle down as $k$ increases.
  - Drifting in one direction often indicates that the sampler has not yet "burned in".
  - Wandering back and forth around the same spot indicates that the sampler has not yet "mixed" sufficiently.

# Censoring example: Histograms




- We are primarily interested in the posterior on $\theta$, since it represents the rate parameter for the survival distribution.
- By making a histogram of the samples $\theta_1, \ldots, \theta_N$, we can estimate the posterior density $p(\theta|x_{1:n})$.
- The vertical lines are the lower $(\ell)$ and upper $(u)$ endpoints of a 90% *credible interval*—that is, an interval containing 90% of the posterior probability.

# Individual activity: Critical thinking

- True or false? Diagnostic plots can indicate lack of convergence of the sampler.

- True or false? Diagnostic plots can indicate convergence of the sampler.

# Caution! MCMC diagnostics can be misleading

- Even when heuristics like traceplots and running averages suggest that all is well, things may be going horribly wrong.
- For instance, posteriors are often highly multimodal, and the sampler may get stuck in one mode for many iterations.
- General rule: MCMC diagnostics can tell you when things are bad, but they cannot tell you when things are good.

# Outline

# Example: Hyperpriors and hierarchical models

- Gibbs sampling is especially useful for models involving multiple levels.
- We often want to put priors on the hyperparameters, that is, the parameters of the prior. This is called a *hyperprior*.
- More generally, *hierarchical models* involve hierarchical relationships among the data and latent variables/parameters.
- The full conditionals in a hierarchical model are often relatively simple, making Gibbs sampling very convenient.

# Normal example with hyperprior

- As a simple example, consider the Normal example with a semi-conjugate prior from earlier.

- Let's add a $\text{Gamma}(r, s)$ prior on $\lambda_0$, so the model is now:

$$\begin{aligned}
\boldsymbol{\lambda}_0 &\sim \text{Gamma}(r, s) \\
\boldsymbol{\mu}|\lambda_0 &\sim \mathcal{N}(\mu_0, \lambda_0^{-1}) \\
\boldsymbol{\lambda} &\sim \text{Gamma}(a, b) \\
X_1, \ldots, X_n|\lambda_0, \mu, \lambda &\overset{\text{iid}}{\sim} \mathcal{N}(\mu, \lambda^{-1}).
\end{aligned}$$

- This is actually equivalent to putting a $t$-distribution prior on $\mu$, but since the $t$-distribution is not a conjugate prior, we cannot sample directly from $\mu|\lambda, x_{1:n}$.

- We *can* easily sample from $\mu|\lambda_0, \lambda, x_{1:n}$, though, and this is what we need for Gibbs sampling.

# Normal example with hyperprior: Gibbs sampler

- Sample from each full conditional, in turn:
  - $(\lambda_0 | \cdots)$ Since $\lambda_0$ is conditionally independent of everything else given $\mu$, this is the same as the full conditional for the precision in a Normal model with one datapoint (namely, $\mu$):

$$\boldsymbol{\lambda_0} | \mu, \lambda, x_{1:n} \sim \text{Gamma}\left(r + 1/2,\ s + \tfrac{1}{2}(\mu - \mu_0)^2\right).$$

  - $(\mu | \cdots)$ Since we are conditioning on $\lambda_0$, this is just the same as the full conditional for $\mu$ before, without a hyperprior:

$$\boldsymbol{\mu} | \lambda_0, \lambda, x_{1:n} \sim \mathcal{N}(M, L^{-1})$$

    where $L = \lambda_0 + n\lambda$ and $M = (\lambda_0 \mu_0 + \lambda \sum x_i)/(\lambda_0 + n\lambda)$.

  - $(\lambda | \cdots)$ Since we are conditioning on $\mu$ and $\lambda_0$, this is also just the same as before:

$$\boldsymbol{\lambda} | \lambda_0, \mu, x_{1:n} \sim \text{Gamma}(A, B)$$

    where $A = a + n/2$ and $B = n\hat{\sigma}^2 + n(\bar{x} - \mu)^2$.

# Normal example with hyperprior

- We could just as easily add semi-conjugate priors on $\mu_0$ and $b$.
  - Specifically, a Normal prior on $\mu_0$ and a Gamma prior on $b$.
- We could then easily augment the Gibbs sampling algorithm to update them as well.
- In this simple example, the hyperpriors just make the prior less informative.
- However, in many applications, hierarchical models are used to share statistical strength across groups.
- Even with highly complex hierarchical models, the Gibbs sampling approach allows one to perform Bayesian inference in a remarkably straightforward way.

# References and supplements


- S. Geman, and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741.
- G. Casella, and E.I. George (1992). Explaining the Gibbs sampler. The American Statistician, 46(3), 167-174.

# Individual activity: Exit ticket

Answer these questions individually:
https://forms.gle/iZoWzjAzPDeYPrLn7