Lightning Quick Recap of Bayesian Inference

Shozen Dan
24/03/2025

Agenda

How to represent and rationally update our beliefs

"Bayes Theorem: A mathematicians idea of how a rational human thinks."

Beliefs, Uncertainty, and Probability

We use probability to express information and beliefs about unknown quantities

- We use probabilities to informally express our information and beliefs about unknown quantities
- It can be shown that probability can formally be used to represent a set of rational beliefs
- Bayes' rule is a rational method of updating beliefs in light of new information
- The process of inductive learning through Bayes' rule is referred to as **Bayesian Inference**

Recap of Bayes' rule
How to rationally update one's belief

- Let $\mathcal{Y}$ be the sample space: the set of all possible outcomes.
- Let Θ be the parameter space: the set of all possible parameter values.

Prior distribution

For each value $\theta \in \Theta$, the prior distribution $p(\theta)$ describes our belief that θ represents the true population dynamics.

Sampling model

For every $\theta \in \Theta$ and $y \in \mathcal{Y}$, the sampling model $p(y|\theta)$ describes our belief that the outcome would be y if θ was the true value

Recap of Bayes' rule
How to rationally update one's belief

Once we have obtained the data y, we update our beliefs about θ.

Posterior distribution

For every $\theta \in \Theta$, the posterior distribution $p(\theta|y)$ is the degree in which we believe θ is the true value after observing the data y.

The posterior distribution is given by Bayes's rule after defining the prior and sampling model.

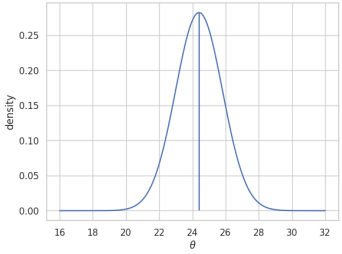
$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_{\Theta} p(y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}$$

Important

Bayes' rule does not state how the posterior should be but rather how the posterior should be updated after we obtain new information.

Bayes' rule in action

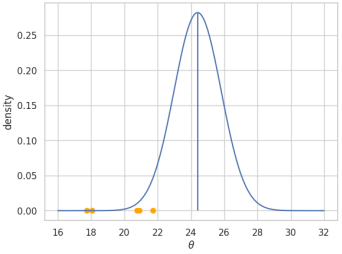
A simple univariate Gaussian example



Specifying the prior

We use the prior to express our knowledge and uncertainty about a population parameter θ (*e.g.*, average BMI) of interest.

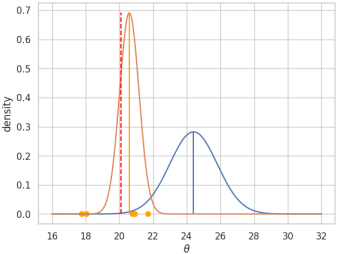
Here $\theta_0 = 24.4$ and $\sigma_0^2 = 2$.



Observe samples

We observe samples obtained from the true data generating process.

$p(y|\theta)$ is assumed to be Gaussian with $\sigma^2 = 2$.



Belief update

We update our belief about θ based on the observations.

Here the true θ is 20.1 (red dashed line) and the posterior (orange) mean $\theta_n \approx 20.6$.

Deriving posteriors: Conjugacy

"Conjugate priors: Making life easier for Bayesian statisticians since forever."

Conjugacy

When the posterior is in the same class as the prior

Definition: Conjugacy

If the class $\mathcal{P}$ of the prior distribution of θ is conjugate w.r.t. the sampling model $p(y|\theta)$ then

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|y) \in \mathcal{P}$$

Some examples of conjugate priors

- Beta prior, Binomial sampling model $\rightarrow$ Beta posterior
- Gamma prior, Poisson sampling model $\rightarrow$ Gamma posterior
- Gaussian prior, Gaussian sampling model $\rightarrow$ Gaussian posterior
- Dirichlet prior, Multinomial sampling model $\rightarrow$ Dirichlet posterior

Working with Binary data

Binomial distribution

Let $Y_i \in \{0, 1\}$ is a binary random variable. $Y_1, Y_2, \ldots, Y_n|\theta$ is i.i.d. according to a Bernoulli distribution with parameter θ. Suppose we obtain data $y_1, y_2, \ldots, y_n$ from this process. We are interested in the inference of θ.

We recall that when $Y_1, Y_2, \ldots, Y_n|\theta$ is i.i.d. w.r.t. to a Bernoulli distribution with parameter θ then the summary statistic $Y = \sum_{i=1}^{n}$ follows a binomial distribution with parameters (n, θ).

Binomial distribution

For $Y = \sum_{i=1}^{n} Y_i$ where Y_i is distributed i.i.d. w.r.t. Bernoulli(θ) $(\theta \in [0, 1])$

$$p(Y = y|\theta) = p(y|\theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}$$

We need to specify a prior for $\theta \in [0, 1]$.

Working with Binary data
the Beta distribution

We want to constrain θ to the unit interval $[0, 1]$. The Beta distribution is a convenient choice.

Beta distribution

For $\theta \in [0, 1]$, $\alpha > 0$, and $\beta > 0$, the probability density function of the Beta distribution is given as

$$p(\theta; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

- $E[\theta] = \alpha/(\alpha + \beta)$
- $\mathrm{Var}[\theta] = \alpha\beta/[(\alpha + \beta + 1)(\alpha + \beta)^2]$

(Above) probability density function of the Beta distribution for different combinations of α and β, courtesy of Wikipedia. https://en.wikipedia.org/wiki/Beta_distribution.

Working with Binary data
Deriving the posterior

To derive the posterior, we apply Bayes' rule

$$
\begin{aligned}
p(y|\theta) &= \frac{p(y|\theta)p(\theta)}{\int_0^1 p(y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}} = \frac{p(y|\theta)p(\theta)}{p(y)} \\
&= \frac{1}{p(y)} \times \binom{n}{y} \theta^y (1-\theta)^{n-y} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= c(n, y, \alpha, \beta) \times \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1} \\
&= \text{dbeta}(\theta, \alpha+y, \beta+n-y)
\end{aligned}
$$

where $c(n, y, \alpha, \beta)$ is a constant dependent on n, y, α, β and dbeta is used to denote the density function of a Beta distribution.

For details on how the normalising constant is calculated, refer to the Appendix.

Working with Count data

Poisson distribution

Let $Y_i \in \mathbb{N}_0$ be a count random variable. Assume $Y_1, Y_2, \ldots, Y_n|\theta$ is i.i.d. according to a Poisson distribution with parameter θ. Suppose we obtain data $y_1, y_2, \ldots, y_n$ from this process. We are interested in the inference of θ.

Poisson distribution

For $\theta > 0$ and $y \in \mathbb{N}_0$ the density function of the Poisson distribution is given as

$$p(Y = y|\theta) = p(y|\theta) = \frac{1}{y!}\theta^y e^{-\theta}$$

If $Y_1, \ldots, Y_n|\theta \sim \text{Poisson}(\theta)$, then the joint density is

$$p(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} p(y_i|\theta) = c(y_1, \ldots, y_n)\theta^{\sum y_i} e^{-n\theta}.$$

Working with Count data
the Gamma distribution

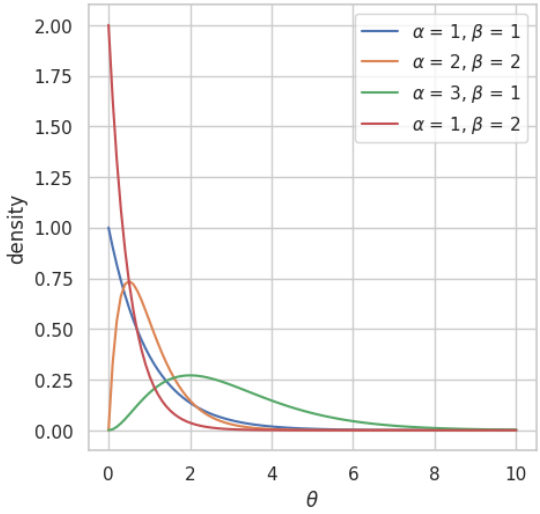
The conjugate prior for the Poisson model is the Gamma distribution.

Gamma distribution

For $\alpha > 0$ and $\lambda > 0$, the probability density function of the Gamma distribution is given as

$$p(\theta; \alpha, \lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\lambda\theta}$$

- $E[\theta] = \alpha/\lambda$
- $\text{Var}[\theta] = \alpha/\lambda^2$



Working with Count data
Deriving the posterior

To derive the posterior, we apply Bayes' rule

$$
\begin{aligned}
p(y_1, \ldots, y_n|\theta) &= \frac{p(y_1, \ldots, y_n|\theta)p(\theta)}{p(y_1, \ldots, y_n)} \\
&= c(y_1, \ldots, y_n, \alpha, \lambda) \times \left\{\theta^{\sum y_i} e^{-n\theta}\right\} \times \left\{\theta^{\alpha-1} e^{-\lambda\theta}\right\} \\
&= c(y_1, \ldots, y_n, \alpha, \lambda) \times \left\{\theta^{\alpha+\sum y_i-1} e^{-(\lambda+n)\theta}\right\} \\
&= \text{dgamma}(\theta, \alpha + \sum y_i, \lambda + n)
\end{aligned}
$$

where $c(n, y, \alpha, \lambda)$ is a constant dependent on n, y, α, λ and dgamma is used to denote the density function of the Gamma distribution.

Monte Carlo approximations

"Monte Carlo: Turning complex integrals into glorified coin flips."

Motivation

Estimating various quantities of interest

When we have an analytical expression for the posterior, it is easy to derive basic summary statistics such as the mean and variance.

However, we often would like to estimate statistics that are difficult or impossible to compute analytically. For instance

- $p(\theta \in A | y_1, \ldots, y_n)$ for arbitrary A
- $E[g(\theta)]$ or $\text{Var}[g(\theta)]$ for some function of θ
- Function involving multiple parameters: $\theta_1 - \theta_2$, θ_1/θ_2, $\max\{\theta_1, \ldots, \theta_m\}$

Monte Carlo approximation

Computationally approximate a given distribution

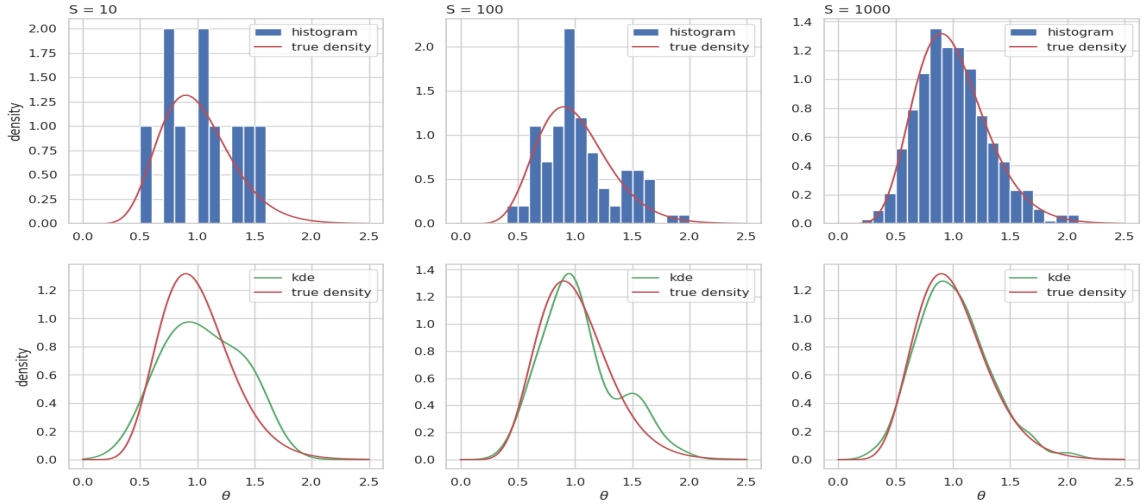
Monte Carlo approximation

Let $y_1, \ldots, y_n$ be samples from the distribution $p(y_1, \ldots, y_n|\theta)$ where θ is the parameter of interest. Suppose we can obtain S independent random samples from the posterior

$$\theta^{(1)}, \ldots, \theta^{(S)} \sim \text{ i.i.d. } p(\theta|y_1, \ldots, y_n)$$

The empirical distribution of $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ is called the Monte Carlo approximation of $p(\theta|y_1, \ldots, y_n)$.

Monte Carlo approximation



Monte Carlo approximation

With enough samples, we can estimate many quantities of interest

If $g(\theta)$ is an arbitrary function of θ and $\theta^{(1)}, \ldots, \theta^{(S)}$ are i.i.d. samples from $p(\theta|y_1, \ldots, y_n)$ then the law of large numbers ensures the following

$$\frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \to E[g(\theta)|y_1, \ldots, y_n]$$
$$= \int g(\theta)p(\theta|y_1, \ldots, y_n)d\theta \quad (S \to \infty)$$

The following is holds when $S \to \infty$

- $\bar{\theta} = \sum_{s=1}^{S} \theta^{(s)}/S \to E[\theta|y_1, \ldots, y_n]$,
- $\sum_{s=1}^{S} (\theta^{(s)} - \bar{\theta})^2/(S-1) \to \text{Var}[\theta|y_1, \ldots, y_n]$,
- $\#(\theta^{(s)} \leq c)/S \to \Pr(\theta \leq c|y_1, \ldots, y_n)$,
- The empirical distribution of $\left\{\theta^{(1)}, \ldots, \theta^{(S)}\right\} \to p(\theta|y_1, \ldots, y_n)$,
- The median of $\left\{\theta^{(1)}, \ldots, \theta^{(S)}\right\} \to \theta_{1/2}$,
- The α quantile of $\left\{\theta^{(1)}, \ldots, \theta^{(S)}\right\} \to \theta_{\alpha}$

Gibbs Sampling

"Finding the full conditionals is half the battle. The other half is debugging the resulting code."

Posterior inference with multiple parameters

Gaussian with unknown mean and variance

Conceptually, Bayesian inference of two or more parameters is similar to the univariate case. Suppose $Y_1, \ldots, Y_n | \theta, \sigma^2 \sim$ *i.i.d.* normal(θ, σ^2) where θ and σ^2 are unknown. We place some prior $p(\theta, \sigma^2)$ on the unknown parameters and proceed to write down Bayes' rule

$$\begin{aligned} p(\theta, \sigma^2 | y_1, \ldots, y_n) &= \frac{p(y_1, \ldots, y_n | \theta, \sigma^2) p(\theta, \sigma^2)}{p(y_1, \ldots, y_n)} \\ &\propto p(y_1, \ldots, y_n | \theta, \sigma^2) p(\theta, \sigma^2) \end{aligned}$$

However, we quickly realise that there are no conjugate priors for the **joint** posterior thus we cannot obtain an analytical expression of the joint posterior. Fortunately, we may still be able to sample from the **full-conditional distribution** of each parameter.

Conditional conjugacy

Conjugate conditioned on the data and other parameters

First we need to introduce the concept of conditional-conjugacy.

Semi-conjugacy

If $\mathcal{Q}$ is a class of sampling distributions $p(y|\theta, \phi)$ and $\mathcal{P}$ is a class of prior distributions for θ conditional on ϕ, then the class $\mathcal{P}$ is semi-conjugate if

$$p(\theta|y, \phi) \in \mathcal{P} \text{ for all } p(y|\theta, \phi) \in \mathcal{Q} \text{ and } p(\theta|\phi) \in \mathcal{P}.$$

Conditional conjugacy

Gaussian with unknown mean and variance

Suppose $y_1, y_2, \ldots, y_n | \theta, \sigma^2$ are i.i.d. according to a Gaussian with unknown mean θ and known variance σ^2. Let the prior distribution of θ also be a Gaussian distribution with mean θ_0 and variance σ_0^2. Derive the posterior distribution of θ: $p(\theta | y_1, \ldots, y_n)$.

Guiding questions

1. Write down the expression for the prior $p(\theta)$.
2. Derive the joint likelihood $p(y_1, \ldots, y_n | \theta)$.
3. Apply Bayes' rule: $p(\theta | y_1, \ldots, y_n) = p(y_1, \ldots, y_n | \theta) p(\theta) / p(y_1, \ldots, y_n)$
   - Hint 1: There is no need to calculate the normalising constant
   - Hint 2: One may ignore all the constant terms

Derivations

The expression of the prior is given as:

$$p(\theta) = (2\pi\tau_0^2)^{-1/2} \exp\left\{ -\frac{1}{2}\left(\frac{\theta - \mu_0}{\tau_0} \right)^2 \right\}$$

The joint density function is given as:

$$\begin{aligned} p(y_1, \ldots, y_n | \theta, \sigma^2) &= \prod_{i=1}^{n} p(y_i | \theta, \sigma^2) \\ &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{1}{2}\left(\frac{y_i - \theta}{\sigma} \right)^2 \right\} \\ &= (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(\frac{y_i - \theta}{\sigma} \right)^2 \right\} \end{aligned}$$

Derivations: Continued

Applying Bayes' rule, we have

$$\begin{aligned} p(\theta|y_1,\ldots,y_n,\sigma^2) &= p(\theta|\sigma^2)p(y_1,\ldots,y_n|\theta,\sigma^2)/p(y_1,\ldots,y_n|\sigma^2) \\ &\propto p(\theta|\sigma^2)p(y_1,\ldots,y_n|\theta,\sigma^2) \\ &\propto \exp\left\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right\}\exp\left\{-\frac{1}{2\sigma^2}\sum(y_i-\theta)^2\right\}. \end{aligned}$$

Expanding the quadratics and calculating the sum in the exponential while ignoring the $-1/2$ for the time being gives us the following:

$$\frac{1}{\tau_0^2}(\theta^2-2\theta\mu_0+\theta^2)+\frac{1}{\sigma^2}(\sum y_i^2-2\theta\sum y_i+n\theta^2) = a\theta^2+2b\theta+c$$

where

$$a=\frac{1}{\tau_0^2}+\frac{n}{\sigma^2},\quad b=\frac{\mu}{\tau_0^2}+\frac{\sum y_i}{\sigma^2},\quad c=c(\mu_0,\tau_0^2,\sigma^2,y1,\ldots,y_n)$$

Derivations: Continued

This tells us that

$$
\begin{aligned}
p(\theta|\sigma^2, y_1, \ldots, y_n) &\propto \exp\left\{-\frac{1}{2}(a\theta^2 - 2b\theta)\right\} \\
&= \exp\left\{-\frac{1}{2}a(\theta^2 - 2b\theta/a + b^2/a^2) + \frac{1}{2}b^2/a\right\} \\
&\propto \exp\left\{-\frac{1}{2}a(\theta - b/a)^2\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{\theta - b/a}{1/\sqrt{a}}\right)\right\}
\end{aligned}
$$

The form of this function is equivalent to a Gaussian with mean b/a and standard deviation $1/\sqrt{a}$. Thus the posterior of θ is a Gaussian distribution with parameters

$$
\tau_n^2 = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)^{-1}, \quad \mu_n = \tau_n^2\left(\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}\right)
$$

Conditional conjugacy

Inverse Gamma prior for unknown variance

Let $Y_1, \ldots, Y_n | \theta, \sigma^2 \sim$ *i.i.d* normal(θ, σ^2) and suppose that θ is given. We want to choose a prior for σ^2. This prior should have support on $(0, \infty)$. The gamma distribution fits the criteria but unfortunately it is not conjugate. However, it is conjugate for the precision: $1/\sigma^2$

Inverse-gamma prior

- If the precision $1/\sigma^2 \sim$ gamma(a, b) then,
- the variance $\sigma^2 \sim$ inverse-gamma(a, b)

For interpretability later on, we parameterise this prior as

$$1/\sigma^2 \sim \text{gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right)$$

Conditional conjugacy
Inverse gamma prior for unknown variance

Let us denote $\gamma = 1/\sigma^2$. Applying Bayes' rule, we have

$$p(\gamma|\theta, y_1, \ldots, y_n) \propto p(y_1, \ldots, y_n, \theta, \gamma) = p(y_1, \ldots, y_n|\theta, \gamma)p(\theta|\gamma)p(\gamma).$$

If we assume that θ and γ are independent, then $p(\theta|\gamma) = p(\theta)$ and

$$\begin{aligned} p(\gamma|\theta, y_1, \ldots, y_n) &\propto p(y_1, \ldots, y_n|\theta, \gamma)p(\theta)p(\gamma) \\ &\propto (\gamma)^{n/2} \exp\left\{-\gamma \sum_{i=1}^{n}(y_i - \theta)^2/2\right\} \times (\gamma)^{\nu_0/2-1} \exp\left\{-\gamma\nu_0\sigma_0^2/2\right\} \\ &= \gamma^{(\nu_0+n)/2-1} \times \exp\left\{-\gamma \times \left[\nu_0\sigma_0^2 + \sum(y_i - \theta)^2\right]/2\right\} \end{aligned}$$

$\{\gamma|\theta, y_1, \ldots, y_n\} \sim \text{gamma}(\nu_n/2, \nu_n\sigma_n^2(\theta)/2)$ where

$$\nu_n = \nu_0 + n, \quad \sigma_n^2(\theta) = \frac{1}{\nu_n}[\nu_0\sigma_0^2 + ns_n^2(\theta)]$$

with $s_n^2(\theta) = \sum(y_i - \theta)/n$, the unbiased estimate of σ^2 if θ were known.

Sampling from full-conditionals

Can we use full-conditionals to obtain samples from the joint posterior?

We possess the ability to sample from $p(\theta|\sigma^2, y_1, \ldots, y_n)$ as well as $p(\sigma^2|\theta, y_1, \ldots, y_n)$. Suppose we were given $\sigma^{2(1)}$, a single sample from the marginal posterior $p(\sigma^2|y_1, \ldots, y_n)$.

Outline of the sampling algorithm

1. Sample
$$\theta^{(1)} \sim p(\theta|\sigma^{2(1)}, y_1, \ldots, y_n)$$
and $\{\theta^{(1)}, \sigma^{2(1)}\}$ would be a sample from the joint posterior of $\{\theta, \sigma^2\}$. Additionally $\theta^{(1)}$ can be considered a sample from the marginal distribution $p(\theta|y_1, \ldots, y_n)$.
2. From this θ-value we generate
$$\sigma^{2(2)} \sim p(\sigma^2|\theta^{(1)}, y_1, \ldots, y_n)$$
then $\{\theta^{(1)}, \sigma^{2(2)}\}$ is also a sample from the joint distribution of $\{\theta, \sigma^2\}$. This in turn means that $\sigma^{2(2)}$ is a sample from the marginal distribution $p(\sigma^2|y_1, \ldots, y_n)$, which can be used to generate $\theta^{(2)}$.

Sampling from full-conditionals
The Gibbs Sampler

Summary of the sampling algorithm

Given the current sate $\phi^{(s)} = \{\theta^{(s)}, \gamma^{(s)}\}$, we generate a new state as follows:

1. sample $\theta^{(s+1)} \sim p(\theta|\sigma^{2(s)}, y_1, \ldots, y_n)$;
2. sample $\gamma^{(s+1)} \sim p(\gamma|\theta^{(s+1)}, y_1, \ldots, y_n)$
3. let $\phi^{(s+1)} = \{\theta^{(s+1)}, \gamma^{(s+1)}\}$

Gibbs Sampler

This algorithm is called a **Gibbs sampler** and it generates a *dependent* sequence of states

$$\{\phi^{(1)}, \phi^{(2)}, \ldots, \phi^{(S)}\}.$$

Visualising the Gibbs Sampler

We can visualise the trajectory of a simple Gibbs sampler in two dimensions.

- Gibbs Sampler Animation 1
- Gibbs Sampler Animation 2

Note how the sampler explores the joint distribution much more slowly when the parameters are highly correlated.

General properties of the Gibbs Sampler
A more general description of the Gibbs sampler

Suppose you have a vector of parameters

$$\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_p\}$$

Information about $\boldsymbol{\phi}$ is measured with $p(\boldsymbol{\phi}) = p(\phi_1, \ldots, y\phi_p)$.

Given a starting point $\boldsymbol{\phi}^{(0)} = \{\phi_1^{(0)}, \ldots, \phi_p^{(0)}\}$, the Gibbs sampler generates $\boldsymbol{\phi}^{(s)}$ from $\boldsymbol{\phi}^{(s-1)}$:

1. sample $\phi_1^{(s)} \sim p(\phi_1 | \phi_2^{(s-1)}, \ldots, \phi_p^{(s-1)})$
2. sample $\phi_2^{(s)} \sim p(\phi_2 | \phi_1^{(s)}, \phi_3^{(s-1)}, \ldots, \phi_p^{(s-1)})$

   $\vdots$

p. $\phi_p^{(s)} \sim p(\phi_p | \phi_1^{(s)}, \phi_2^{(s)}, \ldots, \phi_{p-1}^{(s)})$

The Markov property

Each state only depends on the previous state

The Gibbs sampling algorithm generates a dependent sequence of vectors

$$\begin{aligned}
\boldsymbol{\phi}^{(1)} &= \left\{\phi_1^{(1)}, \ldots, \phi_p^{(1)}\right\} \\
\boldsymbol{\phi}^{(2)} &= \left\{\phi_1^{(2)}, \ldots, \phi_p^{(2)}\right\} \\
&\vdots \\
\boldsymbol{\phi}^{(S)} &= \left\{\phi_1^{(S)}, \ldots, \phi_p^{(S)}\right\}
\end{aligned}$$

$\boldsymbol{\phi}^{(S)}$ depends on $\boldsymbol{\phi}^{(0)}, \ldots, \boldsymbol{\phi}^{(s-1)}$ only through $\boldsymbol{\phi}^{(s-1)}$, i.e., $\boldsymbol{\phi}^{(S)}$ is conditionally independent of $\boldsymbol{\phi}^{(0)}, \ldots, \boldsymbol{\phi}^{(s-1)}$ given $\boldsymbol{\phi}^{(s-1)}$. This is called the **Markov property**, thus the sequence is called a **Markov chain**

Important properties of Markov chains

Under certain condition, for which all models we teach in this course will satisfy, the following holds:

$$\Pr(\phi^{(s)} \in A) \to \int_A p(\phi) d\phi \quad \text{as} \quad S \to \infty$$

More importantly, for functions of interest

$$\frac{1}{S} \sum_{s=1}^{S} g(\phi^{(s)}) \to E[g(\phi)] = \int g(\phi) p(\phi) d\phi \quad \text{as} \quad S \to \infty.$$

This means we can approximate $E[g(\phi)]$ with the sample average of $\{g(\phi^{(1)}), \ldots, g(\phi^{(S)})\}$ just as in Monte Carlo approximation.

For this reason, we call such approximations **Markov chain Monte Carlo (MCMC) approximations** and the procedure an **MCMC algorithm**.

Metropolis Algorithm

"I tried leaving the city, but the Metropolis algorithm kept accepting me back with probability $\min(1, r)$."

Generalised Linear Models

- We often work with count data, binary data, and other non-normal data types
- Generalised linear models (GLMs) are a generalisation of linear regression
- GLMs often involve non-linear link functions

Example: Poisson regression

$$y_i \sim \text{Poisson}(\lambda_i), \quad \log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

Example: Logistic regression

$$y_i \sim \text{Bernoulli}(\pi_i) \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

Key point

In many cases, neither conjugate or semi-conjugate priors are available for the parameters of GLMs. Analytical solutions and Gibbs sampling are out of reach.

Metropolis algorithm
The basic idea

Given two different values θ_a and θ_b we want the following to hold:

basic idea

$$\frac{\#\{\theta^{(s)} \text{ that take value } \theta_a\}}{\#\{\theta^{(s)} \text{ that take value } \theta_b\}} \approx \frac{p(\theta_a|y)}{p(\theta_b|y)}$$

How do we construct such a set? Suppose we have set $\{\theta^{(1)}, \ldots, \theta^{(s)}\}$. We can add a new sample $\theta^{(s+1)}$ to this set according to the following algorithm:

- Generate a candidate θ^* that is close to $\theta^{(s)}$
- If $p(\theta^*|y) > p(\theta^{(s)}|y)$, accept θ^* as $\theta^{(s+1)}$
- If $p(\theta^*|y) \leq p(\theta^{(s)}|y)$, accept θ^* as $\theta^{(s+1)}$ with probability $\frac{p(\theta^*|y)}{p(\theta^{(s)}|y)}$

Metropolis algorithm
Acceptance rule

For a given candidate θ^*, we calculate the following ratio:

$$r = \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y)} \frac{p(y)}{p(y|\theta^{(s)})p(\theta^{(s)})} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(s)})p(\theta^{(s)})}$$

Notice that the $p(y)$ terms cancel out, and we do not need to know the posterior to calculate r.

Acceptance rule

- If $r \geq 1$:
  - Intuition: $\theta^{(s)}$ is already in the set and θ^* is more likely than $\theta^{(s)}$ so we should accept it
  - Action: Accept θ^* as $\theta^{(s+1)}$
- If $r < 1$
  - Intuition: $\theta^{(s)}$ is already in the set and θ^* is less likely than $\theta^{(s)}$ so we should accept it with some probability
  - Action: Accept θ^* or $\theta^{(s)}$ as $\theta^{(s+1)}$ with probability r or $1 - r$ respectively

Metropolis algorithm
How to propose a candidate

We covered how to accept a candidate θ^*. How do we propose a candidate? For the Metropolis algorithm, we use a **symmetric proposal distribution**.

Symmetric proposal distribution

$$J(\theta_b|\theta_a) = J(\theta_a|\theta_b)$$

In other words, the probability of proposing θ_b given θ_a is the same as the probability of proposing θ_a given θ_b.

For example the following distributions are symmetric proposal distributions:

- $J(\theta^*|\theta^{(s)}) = \text{uniform}(\theta^{(s)} - \delta, \theta^{(s)} + \delta)$
- $J(\theta^*|\theta^{(s)}) = \text{normal}(\theta^{(s)}, \delta^2)$

Choosing δ is a very tricky buisness!

Metropolis algorithm
Summary of the algorithm

Let us summarise the Metropolis algorithm:

1. Generate $\theta^* \sim J(\theta|\theta^{(s)})$
2. Calculate the acceptance ratio

$$r = \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(s)})p(\theta^{(s)})}$$

3. Set $\theta^{(s+1)}$ in the following manner:

$$\theta^{(s+1)} = \begin{cases} \theta^* & \text{with probability } \min(1, r) \\ \theta^{(s)} & \text{with probability } 1 - \min(1, r) \end{cases}$$

Visualising the Metropolis algorithm

We can visualise the Metropolis algorithm for a simple bi-variate Gaussian example

- Metropolis algorithm animation 1
- Metropolis algorithm animation 2
- Metropolis algorithm animation 3

Note how the sampler becomes less efficient as the parameters become more correlated. Also note that tuning the variance of the proposal distribution increases acceptance rate but decreases efficiency.

Metropolis Hastings algorithm

Metropolis-Hastings algorithm
Quick recap of Gibbs and Metropolis

Suppose the target distribution is $p_0(u, v|y)$ and we want to sample from it.

Gibbs sampling

Given a current state $x^{(s)} = (u^{(s)}, v^{(s)})$ generate a new state $x^{(s+1)} = (u^{(s+1)}, v^{(s+1)})$ by:

- Update U: $u^{(s+1)} \sim p_0(u|v^{(s)}, y)$
- Update V: $v^{(s+1)} \sim p_0(v|u^{(s+1)}, y)$

Metropolis algorithm

Given a current state $x^{(s)} = (u^{(s)}, v^{(s)})$ generate a new state $x^{(s+1)} = (u^{(s+1)}, v^{(s+1)})$ by:

1. Update U:
   a. Propose $u^* \sim J_u(u|u^{(s)})$
   b. Calculate $r = p_0(u^*, v^{(s)})/p_0(u^{(s)}, v^{(s)})$
   c. Accept u^* as $u^{(s+1)}$ with probability $\min(1, r)$
2. Update V:
   a. Propose $v^* \sim J_v(v|v^{(s)})$
   b. Calculate $r = p_0(u^{(s+1)}, v^*)/p_0(u^{(s+1)}, v^{(s)})$
   c. Accept v^* as $v^{(s+1)}$ with probability $\min(1, r)$

Here, J_u and J_v are symmetric proposal distributions.

Metropolis-Hastings algorithm
A generalisation of Gibbs and Metropolis

- Metroplis algorithm generates proposals from symmetric distributions J_u and J_v, and accepts them with probability $min(1, r)$
- Gibbs sampling generates proposals from the full-conditionals and accepts them with probability 1

Metropolis-Hastings algorithm

The metropolis-hastings algorithm generalises the two algorithms by allowing for arbitrary proposal distributions.

Metropolis-Hastings algorithm
Outline of the algorithm

1. Update U:
   a. Propose $u^* \sim J_u(u|u^{(s)}, v^{(s)})$
   b. Calculate $r = \frac{p_0(u^*, v^{(s)})}{p_0(u^{(s)}, v^{(s)})} \times \frac{J_u(u^{(s)}|u^*, v^{(s)})}{J_u(u^*|u^{(s)}, v^{(s)})}$
   c. Accept u^* as $u^{(s+1)}$ with probability $\min(1, r)$ or set $u^{(s+1)} = u^{(s)}$
2. Update V:
   a. Propose $v^* \sim J_v(v|u^{(s+1)}, v^{(s)})$
   b. Calculate $r = \frac{p_0(u^{(s+1)}, v^*)}{p_0(u^{(s+1)}, v^{(s)})} \times \frac{J_v(v^{(s)}|u^{(s+1)}, v^*)}{J_v(v^*|u^{(s+1)}, v^{(s)})}$
   c. Accept v^* as $v^{(s+1)}$ with probability $\min(1, r)$ or set $v^{(s+1)} = v^{(s)}$

- Metropolis-Hastings resembles the Metropolis algorithm, but has an adjustment factor in the acceptance ratio that accounts for the asymmetry of the proposal distribution.
- It can be shown that the Metropolis algorithm and Gibbs sampling are special cases of the Metropolis-Hastings algorithm.

Intro to MCMC diagnostics

"Convergence? Oh, you mean that mythical beast that MCMC practitioners chase into the twilight?"

Monte Carlo and MCMC
Similar name, different behaviour

The purpose of Monte Carlo or Markov chain Monte Carlo approximation is to obtain a sequence of parameter values $\{\phi^{(1)}, \ldots, \phi^{(S)}\}$ such that

$$\frac{1}{S}\sum_{s=1}^{S} g(\phi^{(s)}) \approx \int g(\phi)p(\phi)d\phi,$$

for any functions g of interest. However, there are differences between the two approaches

Monte Carlo samples

- Samples are independent
- Probability that $\phi^{(s)} \in A$ for any set A is $\int_A p(\phi)d\phi$

MCMC samples

- All we are sure about is

$$\lim_{s\to\infty} \Pr(\phi^{(s)} \in A) = \int_A p(\phi)d\phi$$

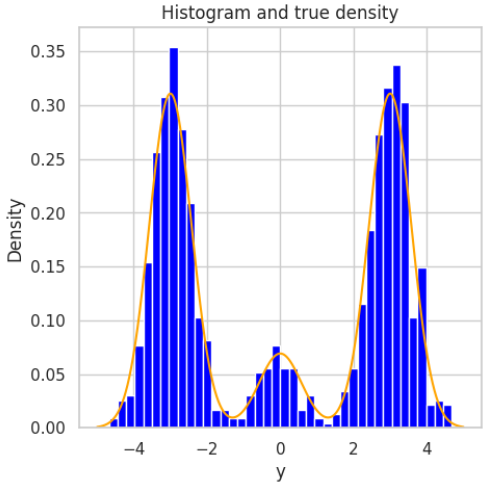
Mixture of Gaussians
Monte Carlo is the "gold standard"

We examine the following mixture of Gaussians

$$p(y) = \sum_{k=1}^{3} p(\theta|\delta = k)p(\delta = k)$$

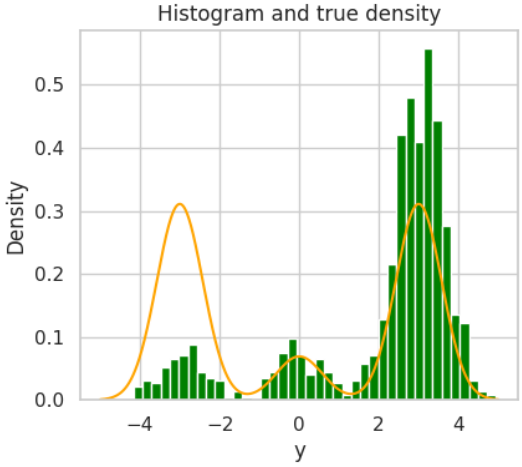
where

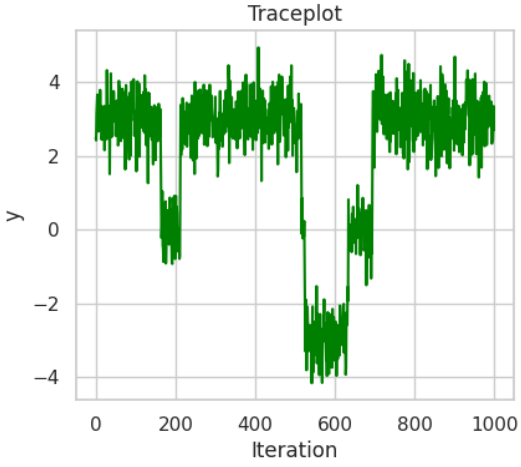
- $p(\delta = 1) = 0.45$, $p(\delta = 2) = 0.1$, $p(\delta = 3) = 0.45$
- $p(\theta|\delta_1 = 1) = \text{dnorm}(-3, 1/3)$, $p(\theta|\delta_1 = 2) = \text{dnorm}(0, 1/3)$, $p(\theta|\delta_1 = 3) = \text{dnorm}(3, 1/3)$



(Above) 1000 MC samples overlayed by the true density

Mixture of Gaussians
When MCMC fails





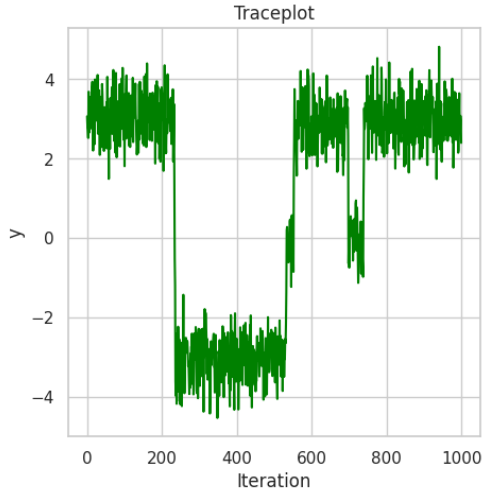
Why MCMC fails sometimes

Autocorrelation

- From the traceplot, we see that the "chain" is stuck in certain regions
- Due to so called **autocorrelation**
- Samples of a Markov chain are dependent on the previous sample, i.e., the samples are correlated

Convergence

- The theory on MCMC tells us that the chain will "eventually" converge to a target distribution (posterior)
- But how long is "eventually"...? The theory doesn't tell.



What do we need to check for and how
Convergence and mixing

Convergence

The MCMC chain(s) needs to have "converged" or "achieved stationarity" (= end up in a fixed region)

How to check

1. Visually: Check the traceplot to see if the chain(s) end up in one area
2. Diagnostic statistics: $\hat{R}$ (Gelman & Rubin, 1992; Brooks & Rubin 1998)

Mixing

The "particle" needs to explore the posterior distribution efficiently. Autocorrelation should be minimised.

How to check

1. Visually: The traceplot should look like a "fat hairy catepillar"
2. Diagnostic statistics: Effective sample size (ESS)

Checking for convergence

It is difficult to check for convergence with only one chain. Recommended approach is to run multiple chains and compare them using:

- **Traceplot**: Visual inspection of the chains. Check whether chains with different starting points converge to the same region.
- **Gelman-Rubin statistic** ($\hat{R}$): Compares the variance within chains to the variance between chains. $\hat{R} < 1.1$ is a commonly used threshold.

Gelman-Rubin statistic $\hat{R}$

- Suppose we have m chains of length n and let θ be a scalar parameter. For example, suppose we have $m = 5$ chains and $n = 500$ samples per chain.
- We split each chains into two halves such that we have $m = 10$ sequences of length $n/2 = 250$.

We first define the following quantities:

$$\bar{\theta}_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} \theta_{ij} \quad \text{(Sample mean of the } j\text{-th sequence)}$$

$$\bar{\theta}_{\cdot \cdot} = \frac{1}{m} \sum_{j=1}^{m} \bar{\theta}_{\cdot j} \quad \text{(Mean of the sample means)}$$

Gelman-Rubin statistic $\hat{R}$

We then define the between-chain variance and within-chain variance as follows:

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\theta}_{\cdot j} - \bar{\theta}_{\cdot\cdot})^2 \quad \text{(Between-chain variance)}$$

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2, \quad s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\theta_{ij} - \bar{\theta}_{\cdot j})^2 \quad \text{(Within-chain variance)}.$$

We can estimate the marginal posterior variance of θ as

$$\widehat{\mathbb{V}(\theta|y)}^{+} = \frac{n-1}{n} W + \frac{1}{n} B.$$

Gelman-Rubin statistic $\hat{R}$

We make the following claims:

- $\widehat{\mathbb{V}(\theta|y)}^{+}$ is an overestimate of the posterior variance.
- It is *unbiased* under stationarity or in the limit $n \to \infty$.

Suppose $\theta_{ij} \sim p(\theta|y)$. Then as $n \to \infty$

$$\bar{\theta}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n}\theta_{ij} \to \mathbb{E}[\theta|y], \quad \bar{\theta}_{\cdot\cdot} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_{\cdot j} \to \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[\theta|y] = \mathbb{E}[\theta|y], \quad W = \frac{1}{m}\sum_{j=1}^{m}s_j^2 \to \mathbb{V}[\theta|y]$$

Thus,

$$\frac{1}{n}B = \frac{1}{m-1}\sum_{j=1}^{m}(\bar{\theta}_{\cdot j} - \bar{\theta}_{\cdot\cdot})^2 \to \frac{1}{m-1}\sum_{j=1}^{m}(\mathbb{E}[\theta|y] - \mathbb{E}[\theta|y])^2 = 0$$

$$\frac{n-1}{n}W \to \mathbb{V}[\theta|y]$$

Gelman-Rubin statistic $\hat{R}$
Definition

For any finite n, W is an *underestimate* because the individual sequences did not have enough time to explore the whore range of the posterior distribution. As a result it will have less variability.

We can assess the convergence of the chains by measuring the factor by which the posterior variance is reduced by running the chains longer.

Gelman-Rubin statistic

$$\hat{R} = \sqrt{\frac{\widehat{\mathbb{V}(\theta|y)}^{+}}{W}}$$

which declines to 1 as $n \to \infty$. If the potential scale reduction is high, then we have reason to believe that continuing the sampling may improve our inference about θ. A commonly accepted threshold is $\hat{R} = 1.1$.

Effective Sample Size (ESS)
Motivation

- The Gelman-Rubin statistic $\hat{R}$ only tells us whether our chain (seems to have) converged to a stationary target distribution.
- It does not tell us how many samples we have to approximate the posterior.
- Recall that MCMC samples are **not** independent. High autocorrelation means that even if you have a long chain, the amount of new information each sample provides is limited.

Effective Sample Size (ESS)

Suppose we want the approximate the integral $\mathbb{E}[\theta] = \int \theta p(\theta) d\theta = \theta_0$ using the emprical distribution $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$. If the $\theta^{(s)}$ are independent Monte Carlo samples, then the variance of the estimator is

$$\mathbb{V}_{\mathsf{MC}}[\bar{\theta}] = \mathbb{E}[(\bar{\theta} - \theta_0)^2] = \frac{\mathbb{V}[\theta]}{S}$$

where $\mathbb{V}[\theta] = \int \theta^2 p(\theta) d\theta - \theta_0^2$.

Thus, we expect the true value of the integral would be contained within the interval $\bar{\theta} \pm 2\sqrt{\mathbb{V}_{\mathsf{MC}}[\bar{\theta}]}$ for roughly 95% of MC approximations.

Effective Sample Size (ESS)

In the case of MCMC, consecutive samples are correlated. Therefore, assuming stationarity has been achieved, the variance of the estimator is

$$
\begin{aligned}
\mathbb{V}_{\text{MCMC}}[\bar{\theta}] &= \mathbb{E}[(\bar{\theta} - \theta_0)^2] \\
&= \mathbb{E}\left[\left\{\frac{1}{S}\sum \theta^{(s)} - \theta_0\right\}^2\right] \\
&= \frac{1}{S^2}\mathbb{E}\left[\sum_{s=1}^{S}(\theta^{(s)} - \theta_0)^2 + \sum_{s\neq t}(\theta^{(s)} - \theta_0)(\theta^{(t)} - \theta_0)\right] \\
&= \frac{1}{S^2}\sum_{s=1}^{S}\mathbb{E}[(\theta^{(s)} - \theta_0)^2] + \frac{1}{S^2}\sum_{s\neq t}\mathbb{E}[(\theta^{(s)} - \theta_0)(\theta^{(t)} - \theta_0)] \\
&= \mathbb{V}_{\text{MC}}[\bar{\theta}] + \frac{1}{S^2}\sum_{s\neq t}\mathbb{E}[(\theta^{(s)} - \theta_0)(\theta^{(t)} - \theta_0)]
\end{aligned}
$$

Effective Sample Size (ESS)

Autocorrelation function

For a stationary sequence $\{\theta^{(s)}\}$, the autocorrelation function at lag h is defined as:

$$\text{acf}_t(\boldsymbol{\theta}) = \frac{\frac{1}{S-t}\sum_{s=1}^{S-t}(\theta^{(s)} - \bar{\theta})(\theta^{(s+1)} - \bar{\theta})}{\frac{1}{S-1}\sum_{s=1}^{S}(\theta^{(s)} - \bar{\theta})^2}$$

where $\bar{\theta} = \frac{1}{S}\sum_{s=1}^{S}\theta^{(s)}$ is the sample mean. The autocorrelation function is a measure of how correlated the samples are at different lags.

Effective Sample Size (ESS)

Definition

Effective Sample Size (ESS)

The effective sample size (ESS) is defined as the number of independent samples that would have the same Monte Carlo variance as the MCMC samples. It is given by

$$S_{\text{eff}} = \frac{S}{1 + 2\sum_{t=1}^{\infty} \text{acf}_t(\boldsymbol{\theta})}.$$

Note that in practice, the infinite sum is truncated when the autocorrelation function is close to zero.

- If the effective sample size of 10,000 MCMC sampels is 100, then the precision of the MCMC approximation to $\mathbb{E}[\theta]$ is only as good as that of 100 independent Monte Carlo samples.
- The ESS you target depends on the precision you require. In Stan, the recommended minimum ESS is 400.

MCMC Remedies

What we can do when things are not working

Convergence

- We wait for eventuality... (run the chain(s) longer)
- Re-examine our model for the following
  - Mathematical mistakes
  - Implementation mistakes
  - Multi-modal posteriors
- Tighten our priors (shrink the parameter space = regularisation). May lead to convergence in local minima.

Mixing

- Increase the number of chains
- Explore various tricks to make sampling more efficient
- Leverage "fancier" MCMC algorithms (e.g., Stan = Hamiltonian Monte Carlo + No U-turn Sampler)

Thank you. Questions?