Bayesian Statistics and Markov-Chain Monte Carlo

Mark M. Fredrickson (mfredric@umich.edu)

Computational Methods in Statistics and Data Science (Stats 406)

Baysian Statistics

Definitions of Probability

There are generally **two approaches** to defining probability:

Frequentist The probability that event A is true is the proportion of an infinitely repeated series of A that are true.

Definitions of Probability

There are generally **two approaches** to defining probability:

Frequentist The probability that event A is true is the proportion of an infinitely repeated series of A that are true.
Parameters are fixed; data are random.

Definitions of Probability

There are generally **two approaches** to defining probability:

Frequentist The probability that event A is true is the proportion of an infinitely repeated series of A that are true.
Parameters are fixed; data are random.

Bayesian Probabilities represent subjective beliefs that range between zero ("cannot occur") and one ("must occur").

Definitions of Probability

There are generally **two approaches** to defining probability:

Frequentist The probability that event A is true is the proportion of an infinitely repeated series of A that are true.
Parameters are fixed; data are random.

Bayesian Probabilities represent subjective beliefs that range between zero ("cannot occur") and one ("must occur").
Belief about parameters can be expressed by a random variable; the data we see are fixed.

Bayes's Rule

Deriving Bayes rule (for events A and B):

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} \qquad \text{(definition of cond. prob.)}$$

Deriving Bayes rule (for events A and B):

$$\begin{aligned} P(B \mid A) &= \frac{P(A \text{ and } B)}{P(A)} && \text{(definition of cond. prob.)} \\ &= \frac{P(A \mid B) P(B)}{P(A)} && \text{(again, using def. cond. prob.)} \end{aligned}$$

Bayes's Rule

Deriving Bayes rule (for events A and B):

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} \qquad \text{(definition of cond. prob.)}$$

$$= \frac{P(A \mid B)P(B)}{P(A)} \quad \text{(again, using def. cond. prob.)}$$

Observe: Frequentist statistical approaches often use Bayes's rule when, e.g., predicting B after observing A.

Putting the “Bayes” in “Bayesian”

If Bayes’ Rule is uncontroversial, so why “Bayesian statistics?”

Putting the “Bayes” in “Bayesian”

If Bayes’ Rule is uncontroversial, so why “Bayesian statistics?”

If θ were a **random variable**, then Bayes’ Rule states:

$$\pi(\theta \,|\, x) = \frac{f(x \,|\, \theta)\, p(\theta)}{\int f(x \,|\, \theta)\, p(\theta)\, d\theta}$$

Putting the “Bayes” in “Bayesian”

If Bayes’ Rule is uncontroversial, so why “Bayesian statistics?”

If θ were a **random variable**, then Bayes’ Rule states:

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\, p(\theta)}{\int f(x \mid \theta)\, p(\theta)\, d\theta}$$

- $p(\theta)$: The **prior distribution** of θ

Putting the "Bayes" in "Bayesian"

If Bayes' Rule is uncontroversial, so why "Bayesian statistics?"

If θ were a **random variable**, then Bayes' Rule states:

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\, p(\theta)}{\int f(x \mid \theta)\, p(\theta)\, d\theta}$$

- $p(\theta)$: The **prior distribution** of θ
- $\pi(\theta \mid x)$: The **posterior distribution** of θ

Putting the "Bayes" in "Bayesian"

If Bayes' Rule is uncontroversial, so why "Bayesian statistics?"

If θ were a **random variable**, then Bayes' Rule states:

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\, p(\theta)}{\int f(x \mid \theta)\, p(\theta)\, d\theta}$$

- $p(\theta)$: The **prior distribution** of θ
- $\pi(\theta \mid x)$: The **posterior distribution** of θ
- $f(x \mid \theta)$: The **likelihood of** x (given θ)

Putting the "Bayes" in "Bayesian"

If Bayes' Rule is uncontroversial, so why "Bayesian statistics?"

If θ were a **random variable**, then Bayes' Rule states:

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\, p(\theta)}{\int f(x \mid \theta)\, p(\theta)\, d\theta}$$

- $p(\theta)$: The **prior distribution** of θ
- $\pi(\theta \mid x)$: The **posterior distribution** of θ
- $f(x \mid \theta)$: The **likelihood of** x (given θ)
- $\int f(x \mid \theta)\, p(\theta)\, d\theta$: A **normalizing constant** (also known as the marginal likelihood of X, $f(x)$)

Using Bayesian Statistics

Suppose we are willing to pick a model for the data ($f(x \mid \theta)$) and a prior for θ ($p(\theta)$).

Using Bayesian Statistics

Suppose we are willing to pick a model for the data ($f(x \mid \theta)$) and a prior for θ ($p(\theta)$).

After observing x, **posterior distribution of θ** answers:

- What is the most likely value of θ? ($\sup \pi(\theta \mid x)$, "maximum a posteriori (MAP) estimator")

Using Bayesian Statistics

Suppose we are willing to pick a model for the data ($f(x \mid \theta)$) and a prior for θ ($p(\theta)$).

After observing x, **posterior distribution of θ** answers:

- What is the most likely value of θ? ($\sup \pi(\theta \mid x)$, "maximum a posteriori (MAP) estimator")
- What value of $\hat{\theta}$ would minimize MSE? ($\hat{\theta} = E(\theta \mid x)$, "Bayes estimator")

Using Bayesian Statistics

Suppose we are willing to pick a model for the data ($f(x \mid \theta)$) and a prior for θ ($p(\theta)$).

After observing x, **posterior distribution of θ** answers:

- What is the most likely value of θ? ($\sup \pi(\theta \mid x)$, "maximum a posteriori (MAP) estimator")
- What value of $\hat{\theta}$ would minimize MSE? ($\hat{\theta} = E(\theta \mid x)$, "Bayes estimator")
- What is the probability that θ is positive? ($P(\theta > 0 \mid x)$)

Using Bayesian Statistics

Suppose we are willing to pick a model for the data ($f(x \mid \theta)$) and a prior for θ ($p(\theta)$).

After observing x, **posterior distribution of θ** answers:

- What is the most likely value of θ? ($\sup \pi(\theta \mid x)$, "maximum a posteriori (MAP) estimator")
- What value of $\hat{\theta}$ would minimize MSE? ($\hat{\theta} = E(\theta \mid x)$, "Bayes estimator")
- What is the probability that θ is positive? ($P(\theta > 0 \mid x)$)
- What is the smallest interval for θ with probability $1 - \alpha$? ("credible interval")

Using Bayesian Statistics

Suppose we are willing to pick a model for the data ($f(x \mid \theta)$) and a prior for θ ($p(\theta)$).

After observing x, **posterior distribution of θ** answers:

- What is the most likely value of θ? ($\sup \pi(\theta \mid x)$, "maximum a posteriori (MAP) estimator")
- What value of $\hat{\theta}$ would minimize MSE? ($\hat{\theta} = E(\theta \mid x)$, "Bayes estimator")
- What is the probability that θ is positive? ($P(\theta > 0 \mid x)$)
- What is the smallest interval for θ with probability $1 - \alpha$? ("credible interval")

Suppose we are willing to pick a model for the data ($f(x \mid \theta)$) and a prior for θ ($p(\theta)$).

After observing x, **posterior distribution of θ** answers:

- What is the most likely value of θ? ($\sup \pi(\theta \mid x)$, "maximum a posteriori (MAP) estimator")
- What value of $\hat{\theta}$ would minimize MSE? ($\hat{\theta} = E(\theta \mid x)$, "Bayes estimator")
- What is the probability that θ is positive? ($P(\theta > 0 \mid x)$)
- What is the smallest interval for θ with probability $1 - \alpha$? ("credible interval")

With two priors, we can compare posterior distributions to get **Bayes Factors** (Bayesian hypothesis tests).

Integrals of posterior

With the exception of MAP estimator from the previous slide, all of those ideas require **integrating the posterior π**:

$$E(g(\theta) \,|\, x) = a$$

Integrals of posterior

With the exception of MAP estimator from the previous slide, all of those ideas require **integrating the posterior** π:

$$E(g(\theta) \mid x) = a$$

Naturally, we can estimate a using Monte Carlo techniques if we can **draw from the posterior**.

With the exception of MAP estimator from the previous slide, all of those ideas require **integrating the posterior** π:

$$E(g(\theta) \,|\, x) = a$$

Naturally, we can estimate a using Monte Carlo techniques if we can **draw from the posterior**.

We'll see an immediate example and then revisit for more complicated examples

Inference for binomial θ

You are tutoring a student. If the student can learn 75% of the material, you will be happy with the results.

You are tutoring a student. If the student can learn 75% of the material, you will be happy with the results.

You adminsiter a test of 30 true/false questions. The student scores

```
[1] 27
```

correctly.

Assumptions and Likelihood

Let us assume that

- All questions are answered with probability $\theta \in (0, 1)$.
- All questions are independent.
- You were successful at teaching if $\theta > 0.75$.

Assumptions and Likelihood

Let us assume that

- All questions are answered with probability $\theta \in (0, 1)$.
- All questions are independent.
- You were successful at teaching if $\theta > 0.75$.

By these assumptions, we get a **binomial likelihood** for the total X:

$$f(x \mid \theta) = \binom{30}{x} \theta^x (1 - \theta)^{30-x}$$

Prior Distribution

For the parameter θ, we need a **prior distribution** that captures our beliefs about how well we did at teaching (i.e., **a distribution for θ**).

Prior Distribution

For the parameter θ, we need a **prior distribution** that captures our beliefs about how well we did at teaching (i.e., **a distribution for θ**).

A common model for random variables in (0, 1) is the **beta distribution**:

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

It has parameters α and β, which we will pick to capture our beliefs.

Prior Distribution

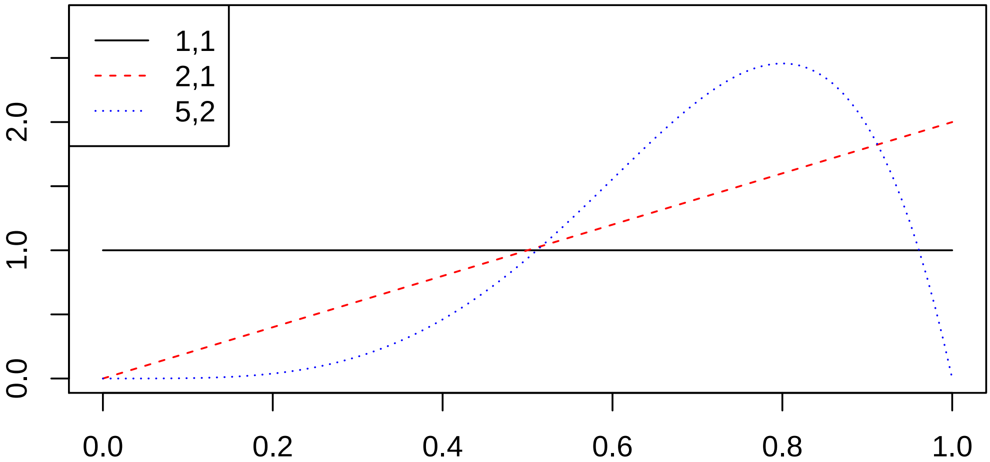
For the parameter θ, we need a **prior distribution** that captures our beliefs about how well we did at teaching (i.e., **a distribution for θ**).

A common model for random variables in (0, 1) is the **beta distribution**:

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

It has parameters α and β, which we will pick to capture our beliefs.

We think that we did a decent job teaching (θ probably close to 0.75), but we want to leave the possibility that we did not do a good job.


1,1
2,1
5,2
0.0
1.0
2.0
0.0
0.2
0.4
0.6
0.8
1.0

Computing posteriors

Recall **Bayes Rule** states

$$\pi(\theta \,|\, x) = \frac{f(x \,|\, \theta)\, p(\theta)}{\int f(x \,|\, \theta)\, p(\theta)\, d\theta}$$

Computing posteriors

Recall **Bayes Rule** states

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\, p(\theta)}{\int f(x \mid \theta)\, p(\theta)\, d\theta}$$

The **marginal likelihood** ($\int f(x \mid \theta)\, p(\theta)\, d\theta$) is often **difficult to compute**

Computing posteriors

Recall **Bayes Rule** states

$$\pi(\theta \,|\, x) = \frac{f(x \,|\, \theta)\, p(\theta)}{\int f(x \,|\, \theta)\, p(\theta)\, d\theta}$$

The **marginal likelihood** ($\int f(x \,|\, \theta)\, p(\theta)\, d\theta$) is often **difficult to compute**

It is often helpful to consider the portion of the RHS that **only depends on θ**:

$$\pi(\theta \,|\, x) \propto f(x \,|\, \theta)\, p(\theta) \quad \text{(proportional to)}$$

Computing posteriors

Recall **Bayes Rule** states

$$\pi(\theta \,|\, x) = \frac{f(x \,|\, \theta)\, p(\theta)}{\int f(x \,|\, \theta)\, p(\theta)\, d\theta}$$

The **marginal likelihood** ($\int f(x \,|\, \theta)\, p(\theta)\, d\theta$) is often **difficult to compute**

It is often helpful to consider the portion of the RHS that **only depends on** θ:

$$\pi(\theta \,|\, x) \propto f(x \,|\, \theta)\, p(\theta) \quad \text{(proportional to)}$$

We would still like to know the full posterior π.

Proportionality

We can often infer π from the **kernel** of $f(x \mid \theta)p(\theta)$.

Proportionality

We can often infer π from the **kernel** of $f(x \mid \theta)p(\theta)$.

Let $f^*(x)$ be a **kernel (unnormalized PDF)** such that

$$f^*(x) \geq 0, \int_{-\infty}^{\infty} f^*(x)\, dx = \frac{1}{c}$$

Proportionality

We can often infer π from the **kernel** of $f(x \mid \theta)p(\theta)$.

Let $f^*(x)$ be a **kernel (unnormalized PDF)** such that

$$f^*(x) \geq 0, \int_{-\infty}^{\infty} f^*(x)\, dx = \frac{1}{c}$$

Suppose that $g(x)$ is a **proper PDF** and

$$g(x) \propto f^*(x)$$

Proportionality

We can often infer π from the **kernel** of $f(x \mid \theta)p(\theta)$.

Let $f^*(x)$ be a **kernel (unnormalized PDF)** such that

$$f^*(x) \geq 0, \int_{-\infty}^{\infty} f^*(x)\, dx = \frac{1}{c}$$

Suppose that $g(x)$ is a **proper PDF** and

$$g(x) \propto f^*(x)$$

Then $g(x) = cf^*(x)$.

Posterior for beta prior and binomial likelihood

In our case, we have:

$$\pi(\theta \mid x) \propto \binom{30}{x} \theta^x (1-\theta)^{30-x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

Posterior for beta prior and binomial likelihood

In our case, we have:

$$\begin{aligned}\pi(\theta \mid x) &\propto \binom{30}{x}\theta^{x}(1-\theta)^{30-x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\\ &\propto \theta^{x}(1-\theta)^{30-x}\theta^{\alpha-1}(1-\theta)^{\beta-1}\end{aligned}$$

Posterior for beta prior and binomial likelihood

In our case, we have:

$$\begin{aligned}
\pi(\theta \mid x) &\propto \binom{30}{x}\theta^{x}(1-\theta)^{30-x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&\propto \theta^{x}(1-\theta)^{30-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \theta^{(\alpha+x)-1}(1-\theta)^{(\beta+30-x)-1}
\end{aligned}$$

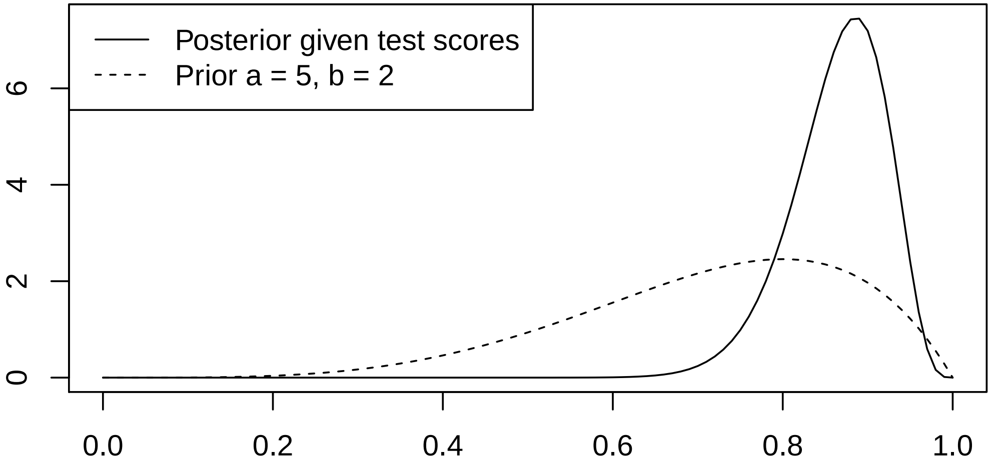
Posterior for beta prior and binomial likelihood

In our case, we have:

$$\begin{aligned}\pi(\theta \mid x) &\propto \binom{30}{x}\theta^x(1-\theta)^{30-x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\\ &\propto \theta^x(1-\theta)^{30-x}\theta^{\alpha-1}(1-\theta)^{\beta-1}\\ &= \theta^{(\alpha+x)-1}(1-\theta)^{(\beta+30-x)-1}\end{aligned}$$

Key insight: since $\pi(\theta \mid x) \propto \text{Beta}(\alpha+x, \beta+30-x)$ it **must also be Beta distributed** (the beta distribution is **conjugate** for the binomial distribution).

Posterior



Using the posterior

What is the probability that you were successful at teaching?

```
> 1 - pbeta(0.75, 5 + test_score, 32 - test_score)

[1] 0.9658
```

Using the posterior

What is the probability that you were successful at teaching?

```
> 1 - pbeta(0.75, 5 + test_score, 32 - test_score)

[1] 0.9658
```

How much did the scores reduce the variance in your uncertainty?

```
> 1 - var(rbeta(1000, 5 + test_score, 32 - test_score)) /
+   var(rbeta(1000, 5, 2)) ## MC estimates of variance

[1] 0.8925
```

More complex problems

We were able to get a **closed form solution** because we picked our likelihood and prior carefully, but **not every problem can be expressed as conjugates**.

More complex problems

We were able to get a **closed form solution** because we picked our likelihood and prior carefully, but **not every problem can be expressed as conjugates**.

General issues:

- Normalizing constant needed for **inversion method**, often difficult.

More complex problems

We were able to get a **closed form solution** because we picked our likelihood and prior carefully, but **not every problem can be expressed as conjugates**.

General issues:

- Normalizing constant needed for **inversion method**, often difficult.
- Even if we can figure out the posterior, might **unable to draw from it directly**.

More complex problems

We were able to get a **closed form solution** because we picked our likelihood and prior carefully, but **not every problem can be expressed as conjugates**.

General issues:

- Normalizing constant needed for **inversion method**, often difficult.
- Even if we can figure out the posterior, might **unable to draw from it directly**.
- **Multiple parameters** make life even harder.

Markov Chain Monte Carlo

Markov Chain Monte Carlo

We already know what **Monte Carlo** means.

We already know what **Monte Carlo** means.

The **Markov Chain** comes from the fact that we will draw samples from a **stochastic process**:

$$\theta(t) \,|\, \theta(t-1), \theta(t-2), \ldots, \theta(0) \sim \theta(t) \,|\, \theta(t-1)$$

(i.e., observation t only depends on observation $t-1$. We suppress dependence on x for simplicity).

We already know what **Monte Carlo** means.

The **Markov Chain** comes from the fact that we will draw samples from a **stochastic process**:

$$\theta(t) \mid \theta(t-1), \theta(t-2), \ldots, \theta(0) \sim \theta(t) \mid \theta(t-1)$$

(i.e., observation t only depends on observation $t-1$. We suppress dependence on x for simplicity).

Such a stochastic process is called a **(discrete) Markov Chain**.

We already know what **Monte Carlo** means.

The **Markov Chain** comes from the fact that we will draw samples from a **stochastic process**:

$$\theta(t) \,|\, \theta(t-1), \theta(t-2), \ldots, \theta(0) \sim \theta(t) \,|\, \theta(t-1)$$

(i.e., observation t only depends on observation $t-1$. We suppress dependence on x for simplicity).

Such a stochastic process is called a **(discrete) Markov Chain**.

Goal: Markov chains that lead to a law of large numbers:

$$\frac{1}{B}\sum_{b=1}^{B} g(\theta(b)) \overset{\text{a.s.}}{\to} E(g(\theta) \,|\, x)$$

Achieving a SLLN

To achieve the desired result, we require Markov Chains that the following properties:

- **Stationary distribution:** A chain is stationary if, when $\theta(t-1) \sim \pi$, $\theta(t) \sim \pi$.

Achieving a SLLN

To achieve the desired result, we require Markov Chains that the following properties:

- **Stationary distribution:** A chain is stationary if, when $\theta(t-1) \sim \pi$, $\theta(t) \sim \pi$.
- **Irreproducible**: No matter the start of the chain $\theta(0)$, there positive probability of visiting any region in the support of θ.

Achieving a SLLN

To achieve the desired result, we require Markov Chains that the following properties:

- **Stationary distribution:** A chain is stationary if, when $\theta(t-1) \sim \pi$, $\theta(t) \sim \pi$.
- **Irreproducible**: No matter the start of the chain $\theta(0)$, there positive probability of visiting any region in the support of θ.
- **Aperiodic**: There is no region such that if $\theta(t) \in \mathcal{R}_1$ we cannot reach $\mathcal{R}_2$.

Achieving a SLLN

To achieve the desired result, we require Markov Chains that the following properties:

- **Stationary distribution:** A chain is stationary if, when $\theta(t-1) \sim \pi$, $\theta(t) \sim \pi$.
- **Irreproducible**: No matter the start of the chain $\theta(0)$, there positive probability of visiting any region in the support of θ.
- **Aperiodic**: There is no region such that if $\theta(t) \in \mathcal{R}_1$ we cannot reach $\mathcal{R}_2$.

Achieving a SLLN

To achieve the desired result, we require Markov Chains that the following properties:

- **Stationary distribution:** A chain is stationary if, when $\theta(t-1) \sim \pi$, $\theta(t) \sim \pi$.
- **Irreproducible**: No matter the start of the chain $\theta(0)$, there positive probability of visiting any region in the support of θ.
- **Aperiodic**: There is no region such that if $\theta(t) \in \mathcal{R}_1$ we cannot reach $\mathcal{R}_2$.

(The last two imply a condition called **ergodicity**, which you may see elsewhere.)

Achieving a SLLN

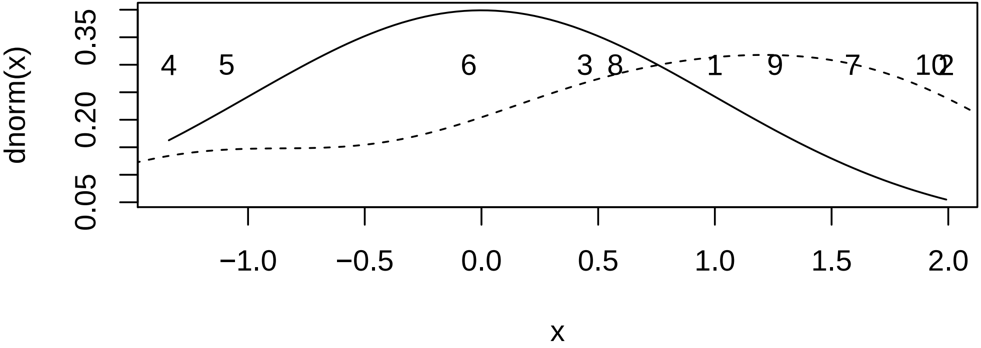
To achieve the desired result, we require Markov Chains that the following properties:

- **Stationary distribution:** A chain is stationary if, when $\theta(t-1) \sim \pi$, $\theta(t) \sim \pi$.
- **Irreproducible**: No matter the start of the chain $\theta(0)$, there positive probability of visiting any region in the support of θ.
- **Aperiodic**: There is no region such that if $\theta(t) \in \mathcal{R}_1$ we cannot reach $\mathcal{R}_2$.

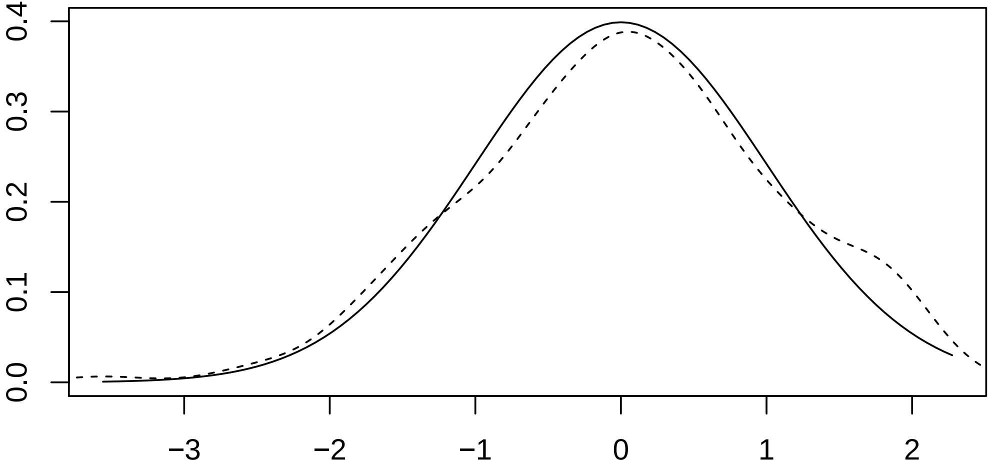
(The last two imply a condition called **ergodicity**, which you may see elsewhere.)

Luckily for us, we can use algorithms that already have these properties established!

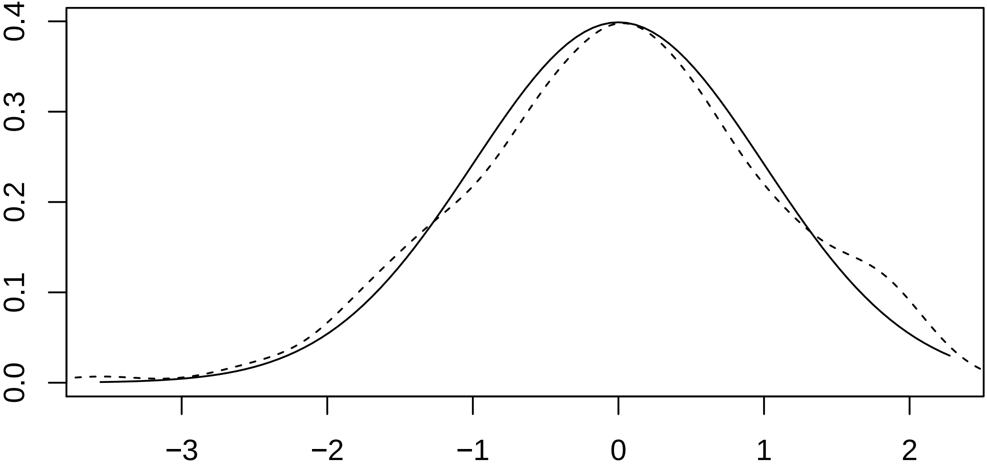
Visual Interpretation: Starting the chain at 1



Visual Interpretation: After many samples



Drop Burn In



Metropolis-Hastings

Accept-reject Review

Recall the **accept-reject algorithm** for IID random variables:

- We wish to draw X from **target density** f but doing so is difficult.

Recall the **accept-reject algorithm** for IID random variables:

- We wish to draw X from **target density** f but doing so is difficult.
- We pick **candidate density** g that we can draw from.

Accept-reject Review

Recall the **accept-reject algorithm** for IID random variables:

- We wish to draw X from **target density** f but doing so is difficult.
- We pick **candidate density** g that we can draw from.
- We find a c such that $f(x)/cg(x) < 1$, for any x in the support of X

Accept-reject Review

Recall the **accept-reject algorithm** for IID random variables:

- We wish to draw X from **target density** f but doing so is difficult.
- We pick **candidate density** g that we can draw from.
- We find a c such that $f(x)/cg(x) < 1$, for any x in the support of X
- Draw Y from g and U from $U(0, 1)$

Recall the **accept-reject algorithm** for IID random variables:

- We wish to draw X from **target density** f but doing so is difficult.
- We pick **candidate density** g that we can draw from.
- We find a c such that $f(x)/cg(x) < 1$, for any x in the support of X
- Draw Y from g and U from $U(0,1)$
- Accept $X = Y$ if $U < f(Y)/cg(Y)$, reject and repeat otherwise.

Metropolis-Hastings: AR for MCMC

The **Metropolis-Hastings algorithm** applies the accept reject concept to **Markov Chains**.

Metropolis-Hastings: AR for MCMC

The **Metropolis-Hastings algorithm** applies the accept reject concept to **Markov Chains**.

MH draws from $\theta^*|\theta(t-1)$ and either

1. Sets $\theta(t) = \theta^*$ if

$$U \leq \frac{\pi(\theta^*)}{\pi(\theta(t-1))} \frac{g(\theta(t-1) \mid \theta^*)}{g(\theta^* \mid \theta(t-1))}$$

Metropolis-Hastings: AR for MCMC

The **Metropolis-Hastings algorithm** applies the accept reject concept to **Markov Chains**.

MH draws from $\theta^*|\theta(t-1)$ and either

1. Sets $\theta(t) = \theta^*$ if
$$U \leq \frac{\pi(\theta^*)}{\pi(\theta(t-1))} \frac{g(\theta(t-1) \mid \theta^*)}{g(\theta^* \mid \theta(t-1))}$$
2. Sets $\theta(t) = \theta(t-1)$, otherwise.

Example: Rayleigh Density

We'll start with an example that is not specifically Bayesian: drawing from the Rayleigh density:

$$f(x) = \frac{x}{\sigma^2} \exp\left\{\frac{-x^2}{2\sigma^2}\right\}, \quad x \geq 0, \sigma^2 > 0$$

Example: Rayleigh Density

We'll start with an example that is not specifically Bayesian: drawing from the Rayleigh density:

$$f(x) = \frac{x}{\sigma^2} \exp\left\{\frac{-x^2}{2\sigma^2}\right\}, \quad x \geq 0, \sigma^2 > 0$$

To be clear: we want a sequence $X(0), X(1), \ldots$ that converges to f (we'll fix σ^2 to a constant).

Example: Rayleigh Density

We'll start with an example that is not specifically Bayesian: drawing from the Rayleigh density:

$$f(x) = \frac{x}{\sigma^2} \exp\left\{\frac{-x^2}{2\sigma^2}\right\}, \quad x \geq 0, \sigma^2 > 0$$

To be clear: we want a sequence $X(0), X(1), \ldots$ that converges to f (we'll fix σ^2 to a constant).

The main requirement we have for the **candidate distribution** is that it have the **same support** as the target.

Example: Rayleigh Density

We'll start with an example that is not specifically Bayesian: drawing from the Rayleigh density:

$$f(x) = \frac{x}{\sigma^2} \exp\left\{\frac{-x^2}{2\sigma^2}\right\}, \quad x \geq 0, \sigma^2 > 0$$

To be clear: we want a sequence $X(0), X(1), \ldots$ that converges to f (we'll fix σ^2 to a constant).

The main requirement we have for the **candidate distribution** is that it have the **same support** as the target.

It should also be **conditional on** $X(t-1)$.

Example: Rayleigh Density

We'll start with an example that is not specifically Bayesian: drawing from the Rayleigh density:

$$f(x) = \frac{x}{\sigma^2} \exp\left\{\frac{-x^2}{2\sigma^2}\right\}, \quad x \geq 0, \sigma^2 > 0$$

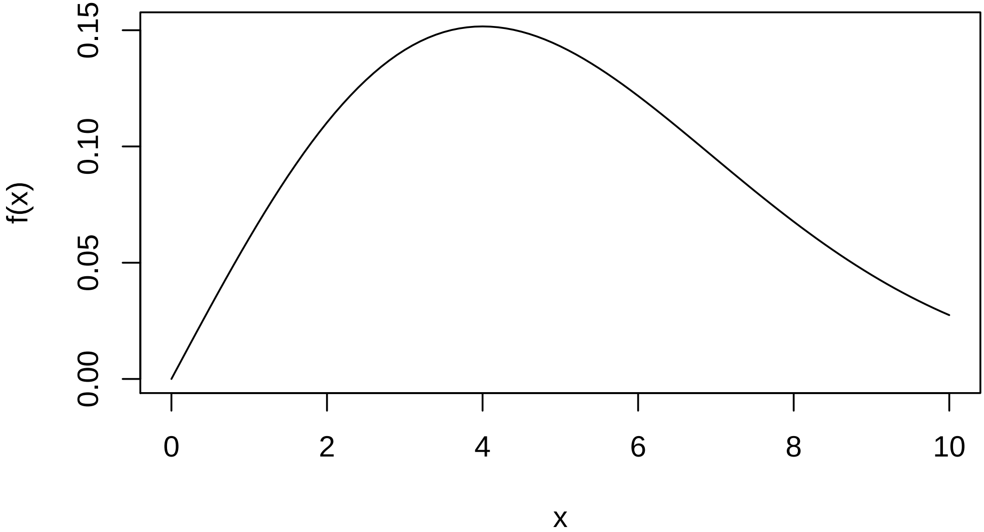
To be clear: we want a sequence $X(0), X(1), \ldots$ that converges to f (we'll fix σ^2 to a constant).

The main requirement we have for the **candidate distribution** is that it have the **same support** as the target.

It should also be **conditional on** $X(t-1)$.

We'll use a χ^2 distribution with $X(t-1)$ degrees of freedom.

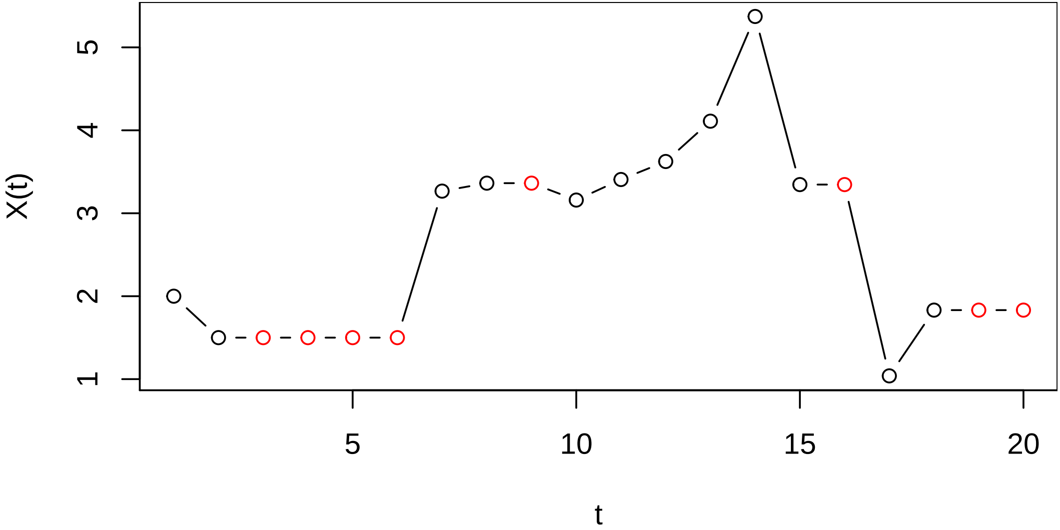
```r
> # we'll fix sigma at 4
> f <- function(x) {
+   (x / 16) * exp(-x^2 / 32)
+ }
> B <- 10000
> xs <- numeric(B)
> xs[1] <- 2 # arbitrary starting point
> # we'll log rejects
> rejected <- logical(B)
> rejected[1] <- FALSE
```



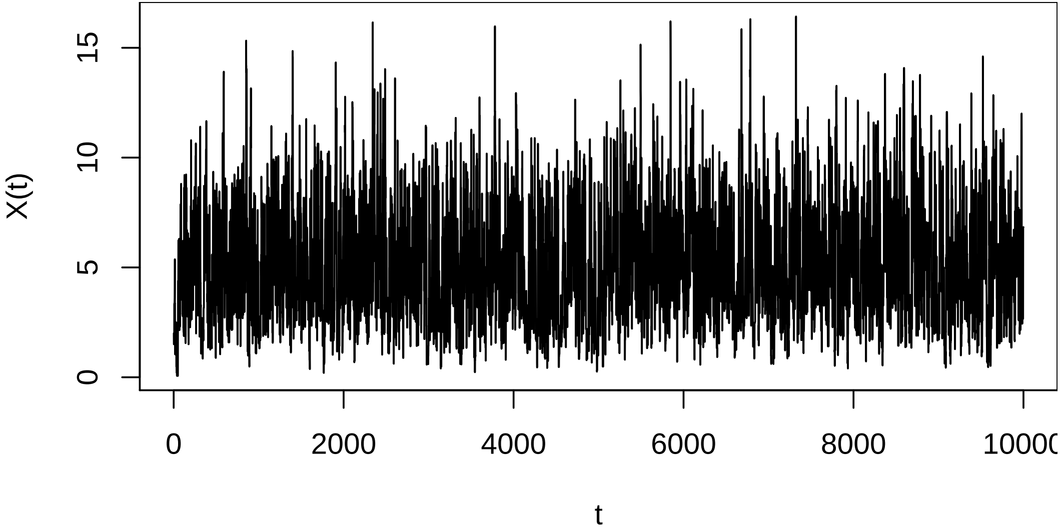
0.00
0.05
0.10
0.15
f(x)
0
2
4
6
8
10
x

```
> # starting i = 2, apply MH
> for (i in 2:B) {
+   x <- xs[i - 1]
+   xstar <- rchisq(1, df = x)
+   ratio <- f(xstar) * dchisq(x, df = xstar) /
+            (f(x) * dchisq(xstar, df = x))
+   u <- runif(1)
+   if (u <= ratio) {
+     xs[i] <- xstar
+     rejected[i] <- FALSE
+   } else {
+     xs[i] <- x
+     rejected[i] <- TRUE
+   }
+ }
```

Start of Chain



Full Chain



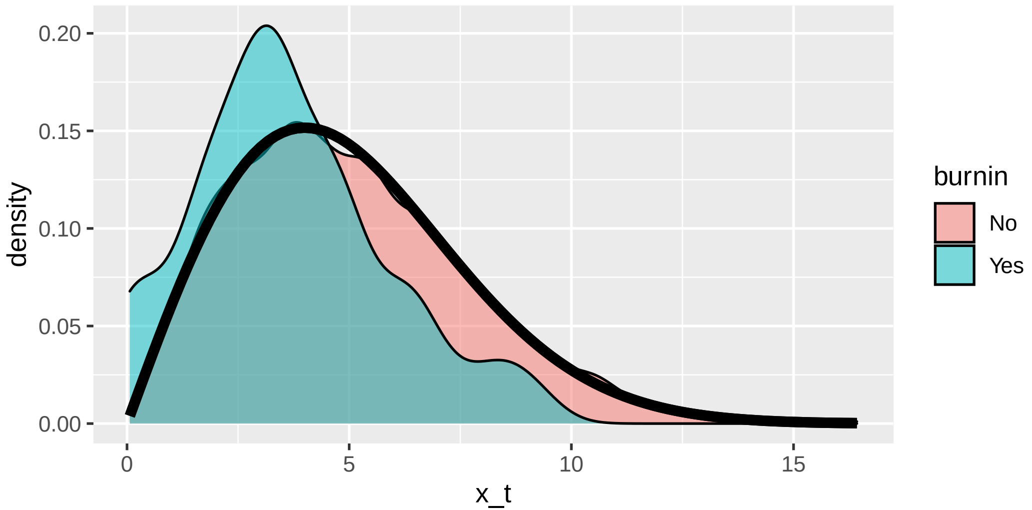
Burn In

Recall, guarantees for MCMC only state that **the chain converges to π** (or f).

Burn In

Recall, guarantees for MCMC only state that **the chain converges to π** (or f).

We often ignore the early portion of the chain (**burn in**).


0.20
0.15
0.10
0.05
0.00
density
0
5
10
15
x_t
burnin
No
Yes

Normalizing constants

We saw with **accept-reject** and **importance sampling** we could often ignore **normalizing constants**.

Normalizing constants

We saw with **accept-reject** and **importance sampling** we could often ignore **normalizing constants**.

This holds true in MH as well since

$$\frac{\pi(\theta^*)}{\pi(\theta(t-1))}$$

Normalizing constants

We saw with **accept-reject** and **importance sampling** we could often ignore **normalizing constants**.

This holds true in MH as well since

$$\frac{\pi(\theta^*)}{\pi(\theta(t-1))} = \frac{c\pi^*(\theta^*)}{c\pi^*(\theta(t-1))}$$

Normalizing constants

We saw with **accept-reject** and **importance sampling** we could often ignore **normalizing constants**.

This holds true in MH as well since

$$\frac{\pi(\theta^*)}{\pi(\theta(t-1))} = \frac{c\pi^*(\theta^*)}{c\pi^*(\theta(t-1))} = \frac{\pi^*(\theta^*)}{\pi^*(\theta(t-1))}$$

Normalizing constants

We saw with **accept-reject** and **importance sampling** we could often ignore **normalizing constants**.

This holds true in MH as well since

$$\frac{\pi(\theta^*)}{\pi(\theta(t-1))} = \frac{c\pi^*(\theta^*)}{c\pi^*(\theta(t-1))} = \frac{\pi^*(\theta^*)}{\pi^*(\theta(t-1))}$$

As we saw, this is useful because it is often much easier to calculate:

$$\pi^*(\theta \mid x) = f(x \mid \theta)\, p(\theta)$$

Binomial θ example, again

We only need **terms that contain θ**. What are they?

Binomial θ example, again

We only need **terms that contain θ**. What are they?

Binomial likelihood:

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

Binomial θ example, again

We only need **terms that contain θ**. What are they?

Binomial likelihood:

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad \propto \theta^x (1-\theta)^{n-x}$$

Binomial θ example, again

We only need **terms that contain θ**. What are they?

Binomial likelihood:

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad \propto \theta^x (1-\theta)^{n-x}$$

Beta prior:

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

Binomial θ example, again

We only need **terms that contain θ**. What are they?

Binomial likelihood:

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad \propto \theta^x (1-\theta)^{n-x}$$

Beta prior:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \quad \propto \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

Binomial θ example, again

We only need **terms that contain θ**. What are they?

Binomial likelihood:

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad \propto \theta^x (1-\theta)^{n-x}$$

Beta prior:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \quad \propto \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

Posterior is proportional to the product of prior and likelihood:

$$\pi^*(\theta \mid x) = \theta^x (1-\theta)^{n-x} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

Binomial θ example, again

We only need **terms that contain θ**. What are they?

Binomial likelihood:

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad \propto \theta^x (1-\theta)^{n-x}$$

Beta prior:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Posterior is proportional to the product of prior and likelihood:

$$\pi^*(\theta \mid x) = \theta^x (1-\theta)^{n-x} \theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{\alpha+x-1}(1-\theta)^{\beta+(n-x)-1}$$

Binomial θ example, again

We only need **terms that contain θ**. What are they?

Binomial likelihood:

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad \propto \theta^x (1-\theta)^{n-x}$$

Beta prior:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Posterior is proportional to the product of prior and likelihood:

$$\pi^*(\theta \mid x) = \theta^x (1-\theta)^{n-x} \theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{\alpha+x-1}(1-\theta)^{\beta+(n-x)-1}$$

As we saw, **π is be Beta** with $\alpha + x$ and $\beta + n - x$.

MCMC for π^*

We need to **pick a candidate distribution** for θ^*.

MCMC for π^*

We need to **pick a candidate distribution** for θ^*.

Since $\theta^* \in (0, 1)$, a candidate **uniform distribution** can be selected.

MCMC for π^*

We need to **pick a candidate distribution** for θ^*.

Since $\theta^* \in (0, 1)$, a candidate **uniform distribution** can be selected.

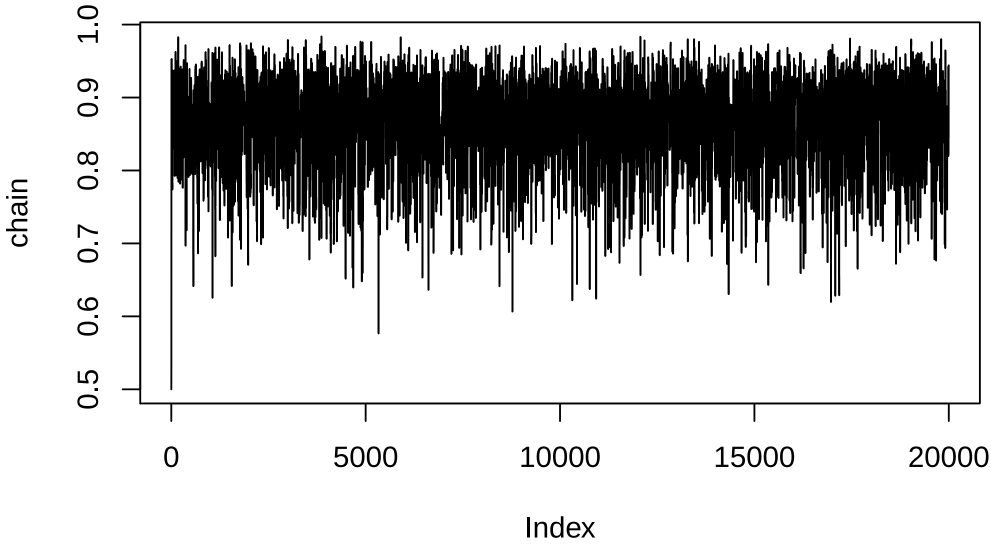
The candidate should be based on $\theta(t-1)$ in some way. We'll do something simple:

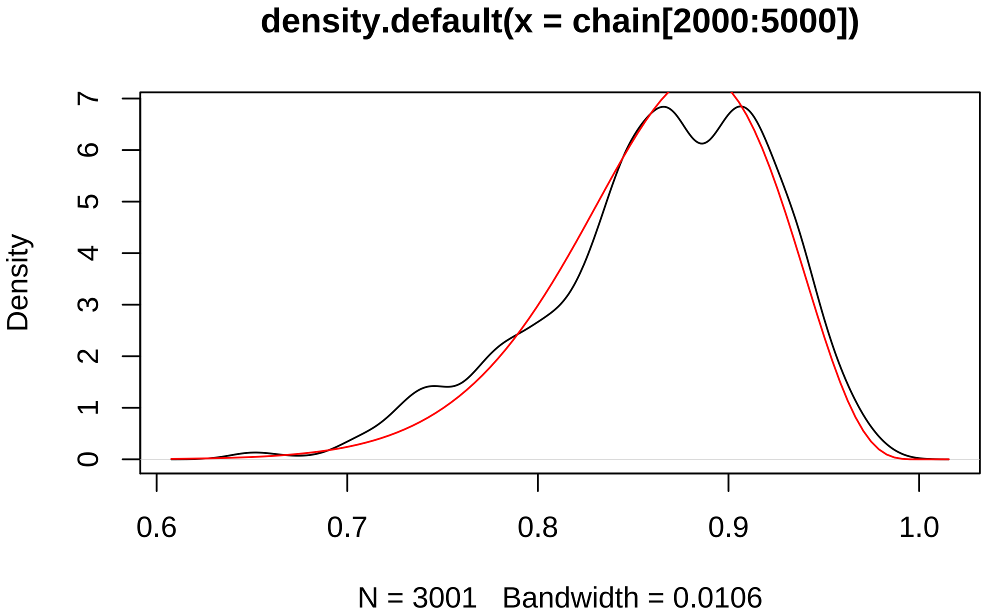
- If $\theta(t-1) < 0.5$, we'll draw from $U(0, 0.6)$
- If $\theta(t-1) \geq 0.5$, we'll draw from $U(0.4, 1)$

Note: the candidate always has density $5/3$.

```
> pi_star <- function(theta) {
+   theta^(5 + test_score - 1) *
+   (1 - theta)^(32 - test_score - 1)
+ }
```

```
> B <- 20000; chain <- numeric(B) ; chain[1] <- 0.5; rejects <- 0
> for (i in 2:B) {
+     candidate <- ifelse(chain[i - 1] < 0.5,
+                         runif(1, 0, 0.6),
+                         runif(1, 0.4, 1))
+     ratio <- pi_star(candidate) * (5/3) /
+         (pi_star(chain[i - 1]) * (5/3))
+     if (runif(1) <= ratio) {
+         chain[i] <- candidate
+     } else {
+         chain[i] <- chain[i - 1]
+         rejects <- rejects + 1
+     }
+ }
> reject_rate <- rejects / B
```


chain
0.5
0.6
0.7
0.8
0.9
1.0
0
5000
10000
15000
20000
Index


density.default(x = chain[2000:5000])
Density
0
1
2
3
4
5
6
7
0.6
0.7
0.8
0.9
1.0
N = 3001 Bandwidth = 0.0106

Independent MH

A special case of the proposal density is to **pick candidate values independently** of the previous value in the chain.

$$g(\theta^* \mid \theta(t-1)) = g(\theta^*)$$

Independent MH

A special case of the proposal density is to **pick candidate values independently** of the previous value in the chain.

$$g(\theta^* \mid \theta(t-1)) = g(\theta^*)$$

Then we have the ratio:

$$\frac{1}{\pi(\theta(t-1))/g(\theta(t-1))} \frac{\pi(\theta^*)}{g(\theta^*)}$$

Independent MH

A special case of the proposal density is to **pick candidate values independently** of the previous value in the chain.

$$g(\theta^* \mid \theta(t-1)) = g(\theta^*)$$

Then we have the ratio:

$$\frac{1}{\pi(\theta(t-1))/g(\theta(t-1))} \frac{\pi(\theta^*)}{g(\theta^*)}$$

This notation is to highlight the connection to **accept-reject** where we had

$$\frac{1}{c} \frac{f(Y)}{g(Y)}$$

Independent MH

A special case of the proposal density is to **pick candidate values independently** of the previous value in the chain.

$$g(\theta^* \mid \theta(t-1)) = g(\theta^*)$$

Then we have the ratio:

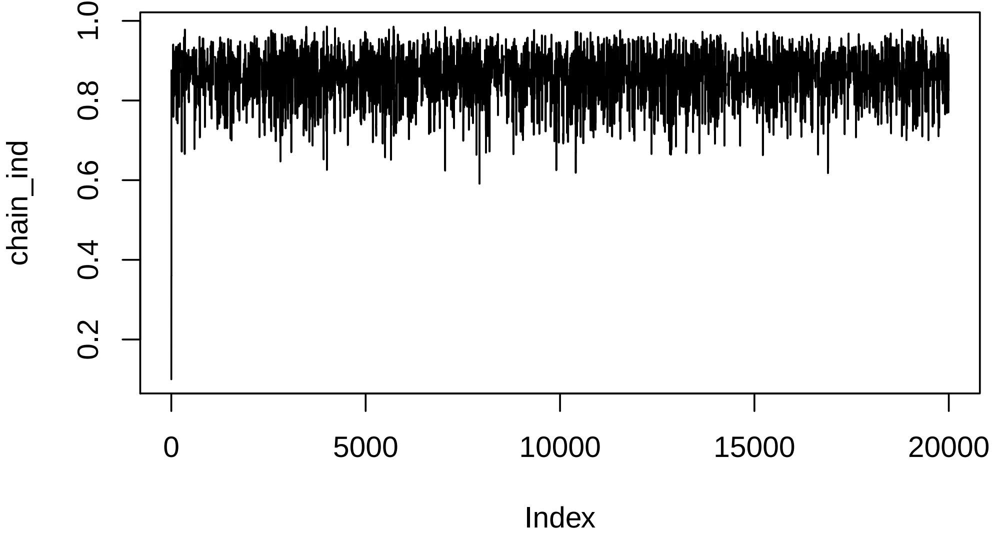
$$\frac{1}{\pi(\theta(t-1))/g(\theta(t-1))} \frac{\pi(\theta^*)}{g(\theta^*)}$$

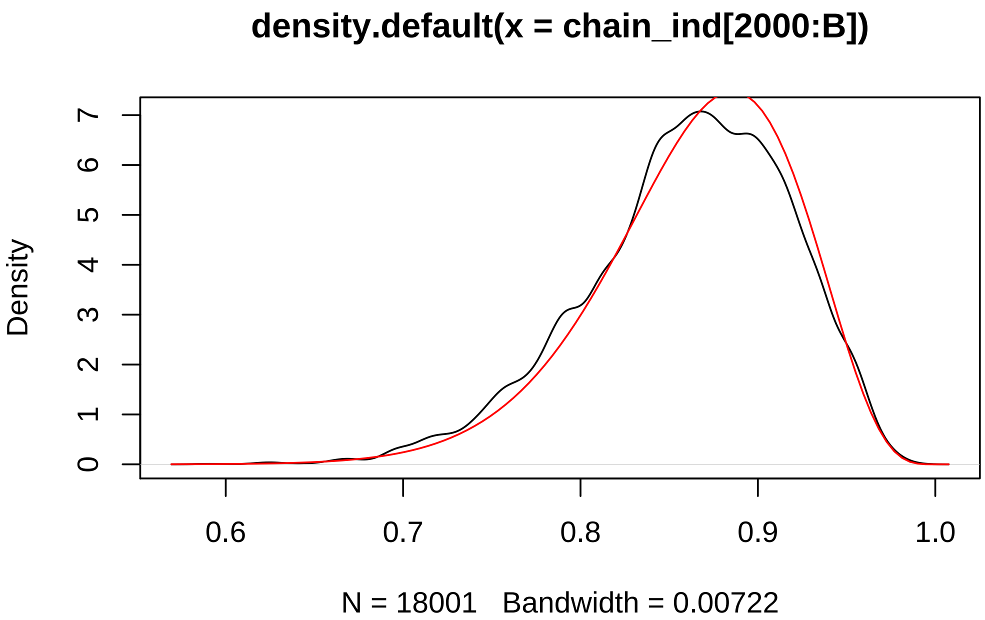
This notation is to highlight the connection to **accept-reject** where we had

$$\frac{1}{c} \frac{f(Y)}{g(Y)}$$

An advantage of MH is we don't need to find a c.

```
> chain_ind <- numeric(B) ; chain_ind[1] <- 0.1 ; rejects_ind <- 0
> for (i in 2:B) {
+     candidate <- runif(1)
+     ratio <- pi_star(candidate) / pi_star(chain_ind[i - 1])
+     if (runif(1) <= ratio) {
+         chain_ind[i] <- candidate
+     } else {
+         chain_ind[i] <- chain_ind[i - 1]
+         rejects_ind <- rejects_ind + 1
+     }
+ }
> reject_rate_ind <- rejects_ind / B
```


1.0
0.8
0.6
0.4
0.2
chain_ind
0
5000
10000
15000
20000
Index


density.default(x = chain_ind[2000:B])
Density
0
1
2
3
4
5
6
7
0.6
0.7
0.8
0.9
1.0
N = 18001   Bandwidth = 0.00722

Comparing Methods

The independent sampler was easier to implement, does it perform as well?

Comparing Methods

The independent sampler was easier to implement, does it perform as well?

Fewer rejects means that we have more **unique samples** in the chain (closer to independent).

```
> reject_rate
[1] 0.7178
> reject_rate_ind
[1] 0.8288
```

Symmetric MH

One nice feature of independent MH is that

$$g(\theta * \mid \theta(t-1)) = g(\theta(t-1) \mid \theta *)$$

so that the ratio reduced to:

$$\frac{\pi(\theta^*)}{\pi(\theta(t-1))}$$

Symmetric MH

One nice feature of independent MH is that

$$g(\theta * \mid \theta(t-1)) = g(\theta(t-1) \mid \theta *)$$

so that the ratio reduced to:

$$\frac{\pi(\theta^*)}{\pi(\theta(t-1))}$$

There are many cases when g is not uniform, but this property (**symmetry**) holds.

Random Walks

An example of symmetry,

$$\theta^* = \theta(t-1) + \epsilon$$

where ϵ is symmetric about 0.

An example of symmetry,

$$\theta^* = \theta(t-1) + \epsilon$$

where ϵ is symmetric about 0.

Here the **proposals are a random walk**, though the chain itself is not (why?).

Example: $N(0, 1)$

Suppose we are trying to generate $N(0, 1)$ using a Markov Chain.

Example: $N(0,1)$

Suppose we are trying to generate $N(0,1)$ using a Markov Chain.

As a proposal, we will use

$$\theta^* = \theta(t-1) + U(-\delta, \delta)$$

Example: $N(0,1)$

Suppose we are trying to generate $N(0,1)$ using a Markov Chain.

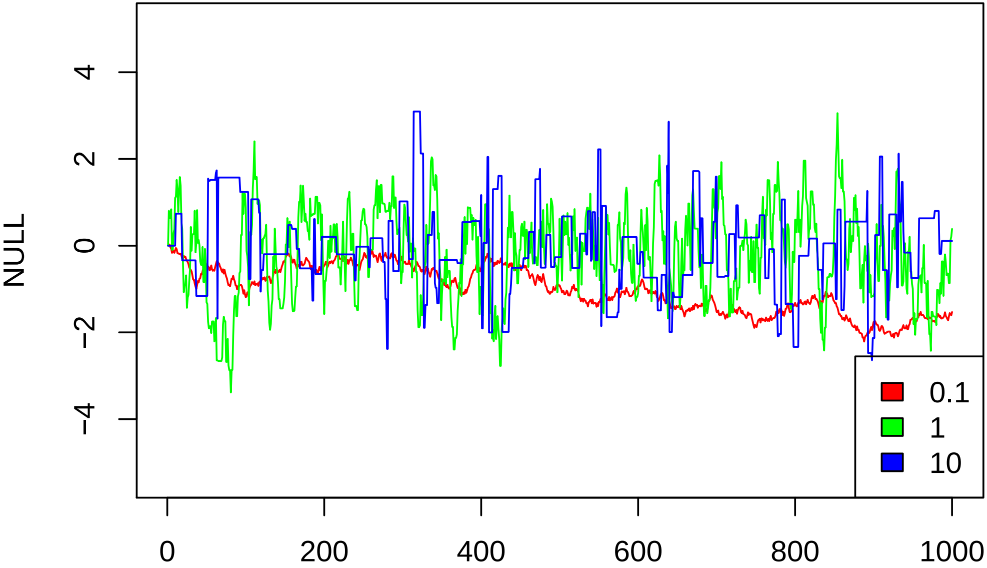
As a proposal, we will use

$$\theta^* = \theta(t-1) + U(-\delta, \delta)$$

We will try a few different versions of δ to see how it changes the chain behavior.

```
> unif_chain <- function(delta, B = 5000) {
+   chain <- numeric(B); chain[1] <- 0 ; rejects <- 0
+   for (i in 2:B) {
+     candidate <- chain[i - 1] + runif(1, -delta, delta)
+     ratio <- dnorm(candidate) / dnorm(chain[i - 1])
+     if (runif(1) <= ratio) {
+         chain[i] <- candidate
+     } else {
+         chain[i] <- chain[i - 1]
+         rejects <- rejects + 1
+     }
+   }
+   list(reject_rate = rejects / B, chain = chain)
+ }
```

```
> n01_chain_0.1 <- unif_chain(0.1)
> n01_chain_1 <- unif_chain(1)
> n01_chain_10 <- unif_chain(10)
```


NULL
4
2
0
−2
−4
0
200
400
600
800
1000
0.1
1
10

```
> n01_chain_0.1$reject_rate
[1] 0.0254
> n01_chain_1$reject_rate
[1] 0.193
> n01_chain_10$reject_rate
[1] 0.843
```

Summary

- Bayesian statistics **treat parameters as random variables** with distributions (prior, posterior).

Summary

- Bayesian statistics **treat parameters as random variables** with distributions (prior, posterior).
- **Inference** frequently requires **integrals on posteriors**.

Summary

- Bayesian statistics **treat parameters as random variables** with distributions (prior, posterior).
- **Inference** frequently requires **integrals on posteriors**.
- In some cases, we can deduce posteriors (or something **proportional to the posterior**)

Summary

- Bayesian statistics **treat parameters as random variables** with distributions (prior, posterior).
- **Inference** frequently requires **integrals on posteriors**.
- In some cases, we can deduce posteriors (or something **proportional to the posterior**)
- More complicated cases require **Markov Chain Monte Carlo**: drawing from a **Markov Chain** with a **stationary target distribution**

Summary

- Bayesian statistics **treat parameters as random variables** with distributions (prior, posterior).
- **Inference** frequently requires **integrals on posteriors**.
- In some cases, we can deduce posteriors (or something **proportional to the posterior**)
- More complicated cases require **Markov Chain Monte Carlo**: drawing from a **Markov Chain** with a **stationary target distribution**
- Algorithms guarantee (asympotic) convergence: Metropolis-Hastings (regular, independent, random walk), more next time.

Summary

- Bayesian statistics **treat parameters as random variables** with distributions (prior, posterior).
- **Inference** frequently requires **integrals on posteriors**.
- In some cases, we can deduce posteriors (or something **proportional to the posterior**)
- More complicated cases require **Markov Chain Monte Carlo**: drawing from a **Markov Chain** with a **stationary target distribution**
- Algorithms guarantee (asympotic) convergence: Metropolis-Hastings (regular, independent, random walk), more next time.
- Often a tradeoff between **amount of rejection** and **exploring the posterior**