MANCHESTER 1824

The University of Manchester

M\cr NA
Manchester Numerical Analysis

The Random Walk Metropolis–Hastings Algorithm

James Rynn
School of Mathematics
The University of Manchester

`james.rynn@postgrad.manchester.ac.uk`

NA-UQ Reading Group, 16/03/16

Outline

Background

Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A discrete-time **stochastic process** is a set of random variables $\{X_k \in \mathbb{R} \mid k \in \mathbb{N}_0\}$.
We often simplify this to X_k but it is important not to forget the underlying probability space, i.e., that in truth $X_k = X_k(\omega)$ for all $\omega \in \Omega,\ k \in \mathbb{N}_0$.

Background

Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A stochastic process $X = (X_k)_{k \in \mathbb{N}_0}$ with values in a set S is a (discrete time) **Markov chain**, if it satisfies the Markov (or 'memoryless') property:

Background

Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A stochastic process $X = (X_k)_{k \in \mathbb{N}_0}$ with values in a set S is a (discrete time) **Markov chain**, if it satisfies the Markov (or 'memoryless') property:

$$\mathbb{P}(X_k \in A_k \mid X_{k-1} \in A_{k-1}, \ldots, X_0 \in A_0)$$

$$= \mathbb{P}(X_k \in A_k \mid X_{k-1} \in A_{k-1}),$$

for all $A_0, A_1, \ldots, A_k \subseteq S$ and all $k \in \mathbb{N}$.

Background

Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A stochastic process $X = (X_k)_{k \in \mathbb{N}_0}$ with values in a set S is a (discrete time) **Markov chain**, if it satisfies the Markov (or 'memoryless') property:

$$\mathbb{P}(X_k \in A_k \mid X_{k-1} \in A_{k-1}, \ldots, X_0 \in A_0)$$

$$= \mathbb{P}(X_k \in A_k \mid X_{k-1} \in A_{k-1}),$$

for all $A_0, A_1, \ldots, A_k \subseteq S$ and all $k \in \mathbb{N}$.
That is, if the distribution of X_k depends on $X_0, X_1, \ldots, X_{k-1}$ only through X_{k-1}. The set S is called the **state space** of X (with $X_k,\ k \in \mathbb{N}_0$ the **state at time** k), the distribution of X_0 is called the **initial distribution** of X and we interpret the index k as time.

Background

Definition

A **transition density** is a map $\nu : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that:

a) $\nu(x, y) \geq 0 \; \forall \; x, y \in \mathbb{R}$; and

b) $\int_{\mathbb{R}} \nu(x, y) \mathrm{d}y = 1 \; \forall \; x \in \mathbb{R}$.

Background

Definition

A **transition density** is a map $\nu : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that:

a) $\nu(x, y) \geq 0 \; \forall \; x, y \in \mathbb{R}$; and

b) $\int_{\mathbb{R}} \nu(x, y) \mathrm{d}y = 1 \; \forall \; x \in \mathbb{R}$.

Definition

A probability density $\pi : \mathbb{R} \to [0, \infty)$ is a **stationary density** for a Markov chain on the state space $S \subseteq \mathbb{R}$ with transition density ν, if it satisfies

$$\int_S \pi(x)\nu(x, y)\mathrm{d}x = \pi(y),$$

for all $y \in \mathbb{R}$.

Aims

We wish to investigate models involving uncertainty, e.g., PDEs with uncertain data.

We do this by finding probability distributions for the uncertain parameters, given (indirect) observations of the data.

This allows us to quantify uncertainty.

Method

We will sample from the posterior distribution of the parameters given the observations.

Ideally we would like to generate i.i.d. samples from this distribution - but this is difficult - not least because we don't actually know the distribution!

We instead choose to generate the next best thing - Markov chains.

This is the main idea behind Markov chain Monte Carlo (MCMC) methods.

We treat the states of the Markov chain as samples from the posterior distribution and use these to build a picture of the distribution.

Metropolis–Hastings Algorithm

This algorithm is used to produce a Markov chain with stationary density, π, that of the distribution we wish to sample from.

Main Idea:

1. assume the current state is $X_{k-1} = x$,
2. generate proposed value for the next state of the chain, $Y_k = y$ (in a clever way)
3. compute the ***acceptance probability***,

$$\alpha(x, y) := \min\left(\frac{\pi(y)\nu(y, x)}{\pi(x)\nu(x, y)}, 1\right),$$

4. accept the proposed value with probability α, or reject and stay at the current state,
5. repeat until enough states (samples) have been generated.

Metropolis–Hastings Algorithm

Metropolis–Hastings Algorithm:

1: Set initial state X_0
2: **for** $k = 1, 2, 3, \ldots, N$
3: generate Y_k with density $\nu(X_{k-1}, \cdot)$
4: generate $U_k \sim \mathcal{U}[0, 1]$
5: **if** $U_k \leq \alpha(X_{k-1}, Y_k)$
6: $X_k \leftarrow Y_k$
7: **else**
8: $X_k \leftarrow X_{k-1}$
9: **end if**
10: output X_k
11: **end for**

Metropolis–Hastings Algorithm

Metropolis–Hastings Algorithm:

1: Set initial state X_0
2: **for** $k = 1, 2, 3, \ldots, N$
3: generate Y_k with density $\nu(X_{k-1}, \cdot)$
4: generate $U_k \sim \mathcal{U}[0, 1]$
5: **if** $U_k \leq \alpha(X_{k-1}, Y_k)$
6: $X_k \leftarrow Y_k$
7: **else**
8: $X_k \leftarrow X_{k-1}$
9: **end if**
10: output X_k
11: **end for**

Key Observation: We only need to know π up to a constant of proportionality!

Acceptance Probability

Why choose $\alpha(x, y) = \min\left(\frac{\pi(y)\nu(y,x)}{\pi(x)\nu(x,y)}, 1\right)$?

We want α to represent a probability, so it must be bounded above by 1.

It can be shown that this choice of α results in a chain with stationary density π.

Random Walk Metropolis–Hastings Algorithm

This is a specific case of the Metropolis–Hastings algorithm, where the proposals Y_k are constructed as

$$Y_k = X_{k-1} + \epsilon_k,$$

where the ϵ_k are chosen to be i.i.d. with a symmetric distribution.

We will use

$$\epsilon_k \sim \mathcal{N}(0, \beta^2),$$

so $Y_k \sim \mathcal{N}(X_{k-1}, \beta^2)$

The choice of proposal variance β^2 is very important, the optimal value is problem dependent.

Random Walk Metropolis–Hastings Algorithm

Random Walk Metropolis–Hastings Algorithm:

1: Set initial state X_0
2: **for** $k = 1, 2, 3, \ldots, N$
3: generate ϵ_k
4: let $Y_k \leftarrow X_{k-1} + \epsilon_k$
5: generate $U_k \sim \mathcal{U}[0, 1]$
6: **if** $U_k \leq \alpha(X_{k-1}, Y_k)$
7: $X_k \leftarrow Y_k$
8: **else**
9: $X_k \leftarrow X_{k-1}$
10: **end if**
11: output X_k
12: **end for**

Biased Coin Example

We have a (possibly biased) coin and wish to determine the probability of throwing a "head", i.e., find

$$p := \mathbb{P}(\{\mathsf{H}\}).$$

We toss the coin 10 times and observed the event

$$A = \{\mathsf{H}, \mathsf{H}, \mathsf{T}, \mathsf{H}, \mathsf{H}, \mathsf{H}, \mathsf{H}, \mathsf{T}, \mathsf{T}, \mathsf{H}\}, \quad \Big(7\mathsf{H}, 3\mathsf{T}\Big).$$

We aim to characterize the posterior probability distribution, $\mathbb{P}(p|A)$.

Prior Beliefs

Informally:

$$p \sim \mathcal{N}\left(\frac{1}{2}, \left(\frac{1}{10}\right)^2\right)\Big|_{[0,1]}.$$

Prior Beliefs

Informally:

$$p \sim \mathcal{N}\left(\frac{1}{2}, \left(\frac{1}{10}\right)^2\right)\Big|_{[0,1]}.$$

Formally: Let p be a random variable with (Lebesgue) density $\rho_0(p)$ (the ***prior***), given by

$$\rho_0(p) = c \cdot \frac{1}{\frac{1}{10}\sqrt{2\pi}} \exp\left(\frac{-(p-\frac{1}{2})^2}{2(\frac{1}{10})^2}\right) \mathbb{1}_{[0,1]}(p),$$

where c is a constant ensuring that $\rho_0(p)$ is a pdf.

Likelihood

The ***likelihood***, denoted $\rho(A|p)$, is the density of the random variable $A|p$.

#H's $\sim$ Binomial$(10, p)$.

Hence,

$$\rho(A|p) = \binom{10}{7} p^7(1-p)^3 = 120\ p^7(1-p)^3.$$

Bayes' Theorem

Theorem

(Bayes' Theorem)
Assume that

$$Z := \int_{\mathbb{R}} \rho(A|p)\rho_0(p)dp > 0.$$

Then, $p|A$ is a random variable with (Lebesgue) density $\pi(p|A)$ given by

$$\pi(p|A) = \frac{1}{Z}\rho(A|p)\rho_0(p).$$

Bayes' Theorem

Theorem

(Bayes' Theorem)
Assume that

$$Z := \int_{\mathbb{R}} \rho(A|p)\rho_0(p)dp > 0.$$

Then, $p|A$ is a random variable with (Lebesgue) density $\pi(p|A)$ given by

$$\pi(p|A) = \frac{1}{Z}\rho(A|p)\rho_0(p).$$

- $\pi(p|A)$ is called the ***posterior density***

Bayes Theorem

Key Observation:

$$\pi(p|A) \propto \rho(A|p)\rho_0(p)$$

Bayes Theorem

Key Observation:

$$\pi(p|A) \propto \rho(A|p)\rho_0(p)$$

In our example:

$$\begin{aligned}\pi(p|A) &\propto \rho(A|p)\rho_0(p) \\ &= 120\, p^7(1-p)^3 \cdot \frac{10c}{\sqrt{2\pi}} \exp\left(-50\left(p-\frac{1}{2}\right)^2\right) \mathbb{1}_{[0,1]}(p) \\ &= \frac{1200c}{\sqrt{2\pi}}\, p^7(1-p)^3 \exp\left(-50\left(p-\frac{1}{2}\right)^2\right) \mathbb{1}_{[0,1]}(p),\end{aligned}$$

Analytic Posterior

In this simple example, we can show

$$\pi(p|A) = \tilde{C} p^7 (1-p)^3 \exp\left(-50\left(p - \frac{1}{2}\right)^2\right) \mathbb{1}_{[0,1]}(p),$$

where

$$\tilde{C} := \left[\frac{65023}{625000000}\sqrt{2\pi}\,\mathrm{erf}\left(\frac{5}{\sqrt{2}}\right) + \frac{2577}{62500000}e^{-\frac{25}{2}}\right]^{-1},$$

and

$$\mathrm{erf}(z) := \frac{2}{\pi}\int_0^z e^{-t^2}\mathrm{d}t.$$

Experiment 1: Varying the Proposal Variance, β^2

β too small

$\Longrightarrow$ high proportion of proposals accepted,

$\Longrightarrow$ slow movement due to small jump size.

Experiment 1: Varying the Proposal Variance, β^2

β too small

$\Longrightarrow$ high proportion of proposals accepted,

$\Longrightarrow$ slow movement due to small jump size.

β too large

$\Longrightarrow$ low proportion of proposals accepted,

$\Longrightarrow$ slow movement due to long time between jumps.

Experiment 1: Varying the Proposal Variance, β^2

β too small

$\implies$ high proportion of proposals accepted,

$\implies$ slow movement due to small jump size.

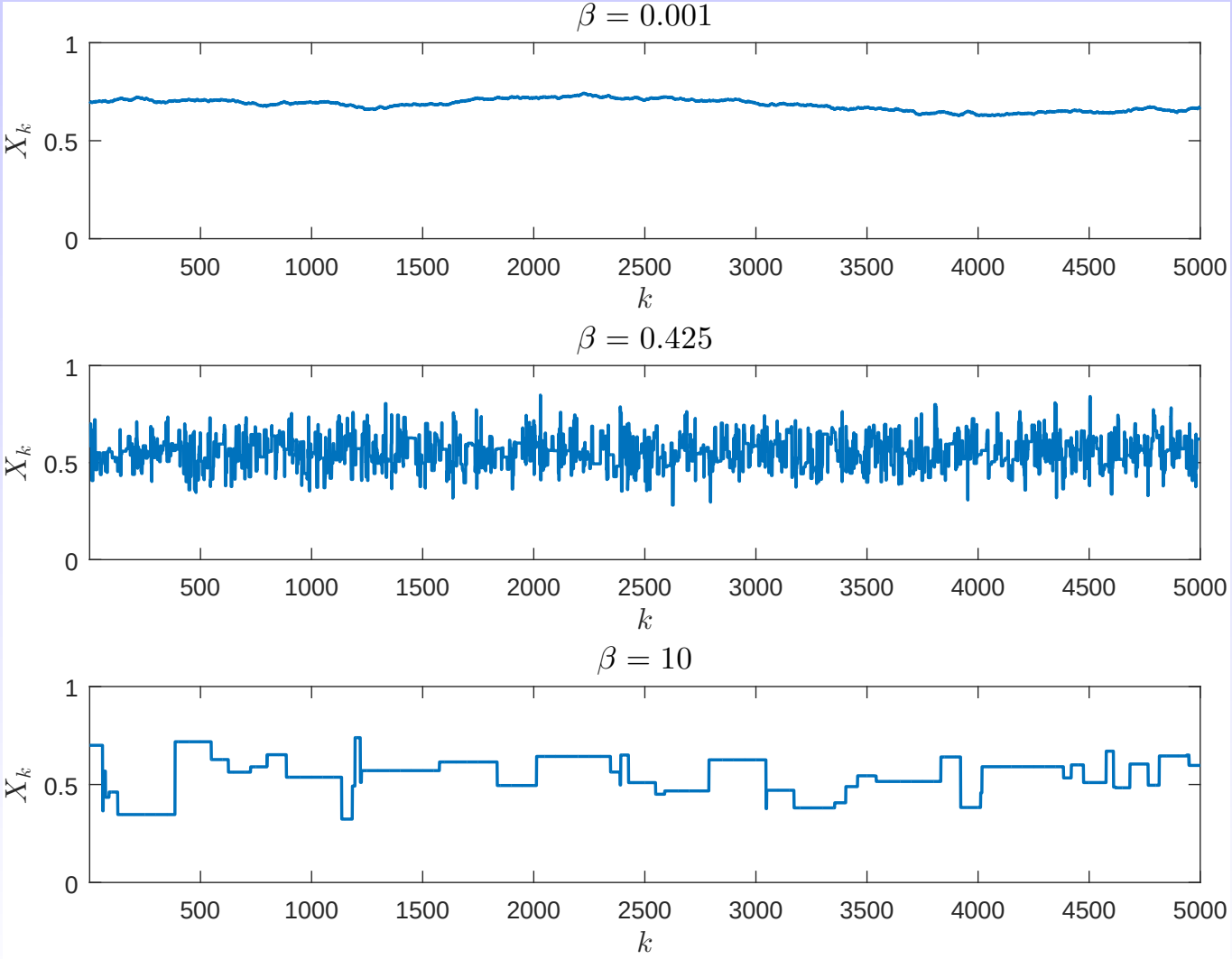
β too large

$\implies$ low proportion of proposals accepted,

$\implies$ slow movement due to long time between jumps.

The optimal value of β lies between these two extremes.

Experiment 1: Varying the Proposal Variance, β^2



Experiment 2: Varying Sample Size, N

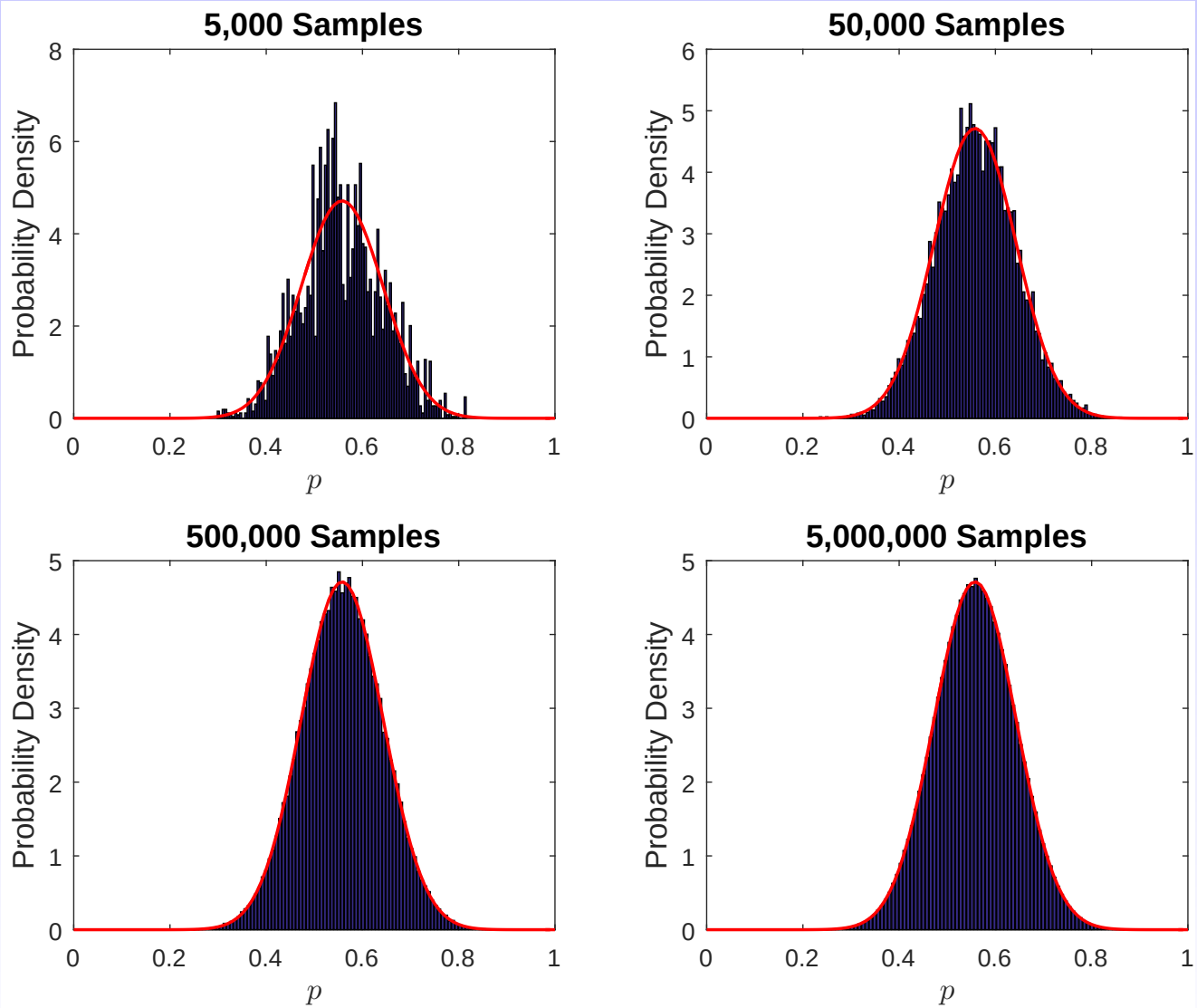
Plotting a normalised histogram of the states of the Markov chain output by the algorithm gives an approximation to the true posterior density π.

Experiment 2: Varying Sample Size, N

Plotting a normalised histogram of the states of the Markov chain output by the algorithm gives an approximation to the true posterior density π.

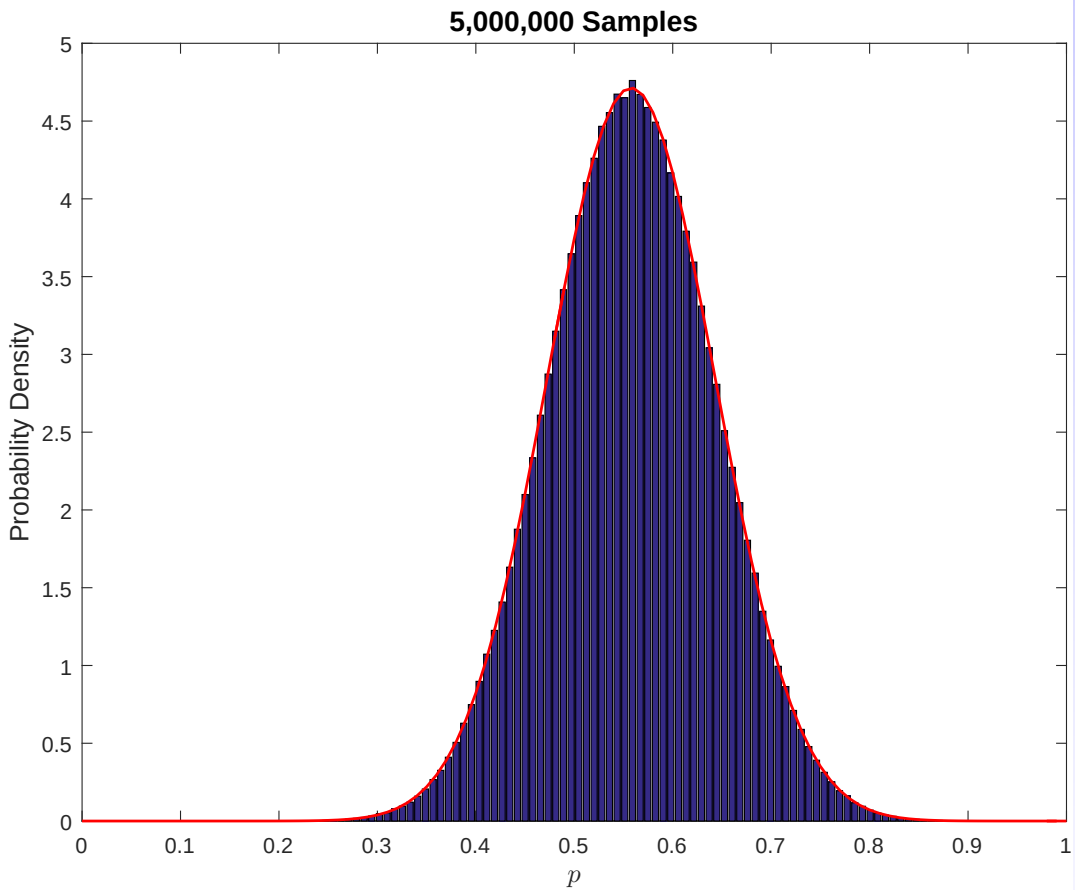
But how many samples do we need for a sufficient approximation?

Experiment 2: Varying Sample Size, N

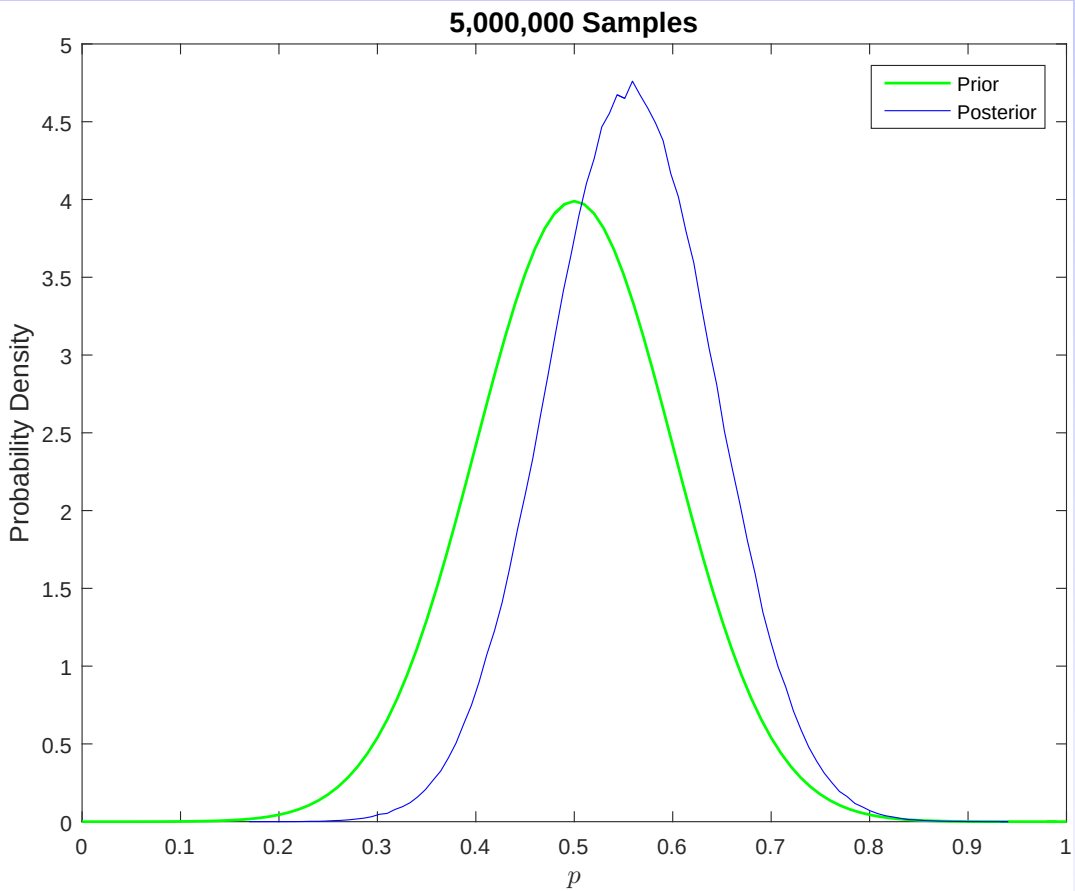


Extracting Results

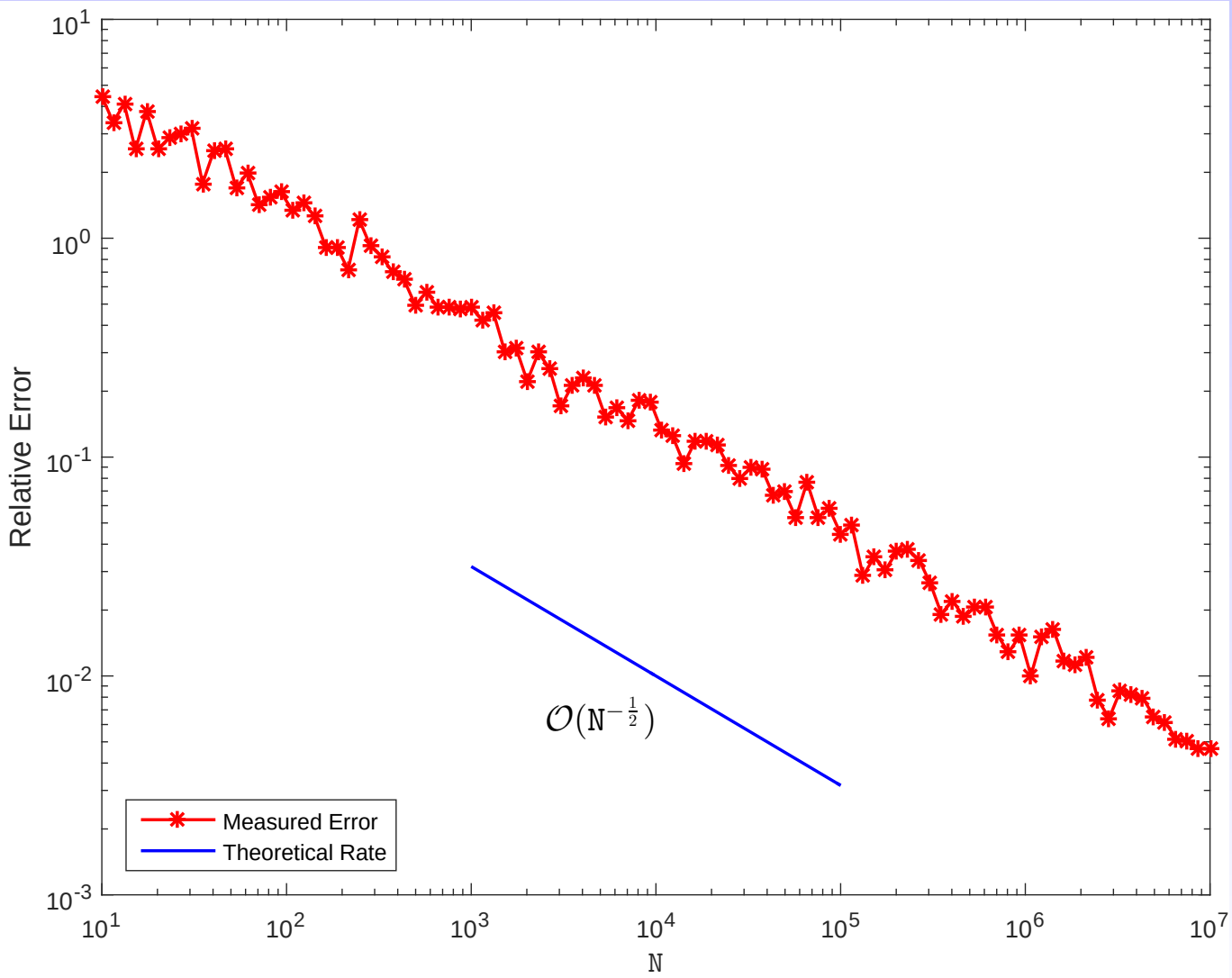
What does a converged plot tell us?



Extracting Results



Error Analysis



Experiment 3: Increasing the Number of Observations

Suppose we have more observations, what effect does this have on the resulting posterior?

Experiment 3: Increasing the Number of Observations

Suppose we have more observations, what effect does this have on the resulting posterior?

More observations

Experiment 3: Increasing the Number of Observations

Suppose we have more observations, what effect does this have on the resulting posterior?

More observations = more data

Experiment 3: Increasing the Number of Observations

Suppose we have more observations, what effect does this have on the resulting posterior?

More observations = more data = more knowledge.

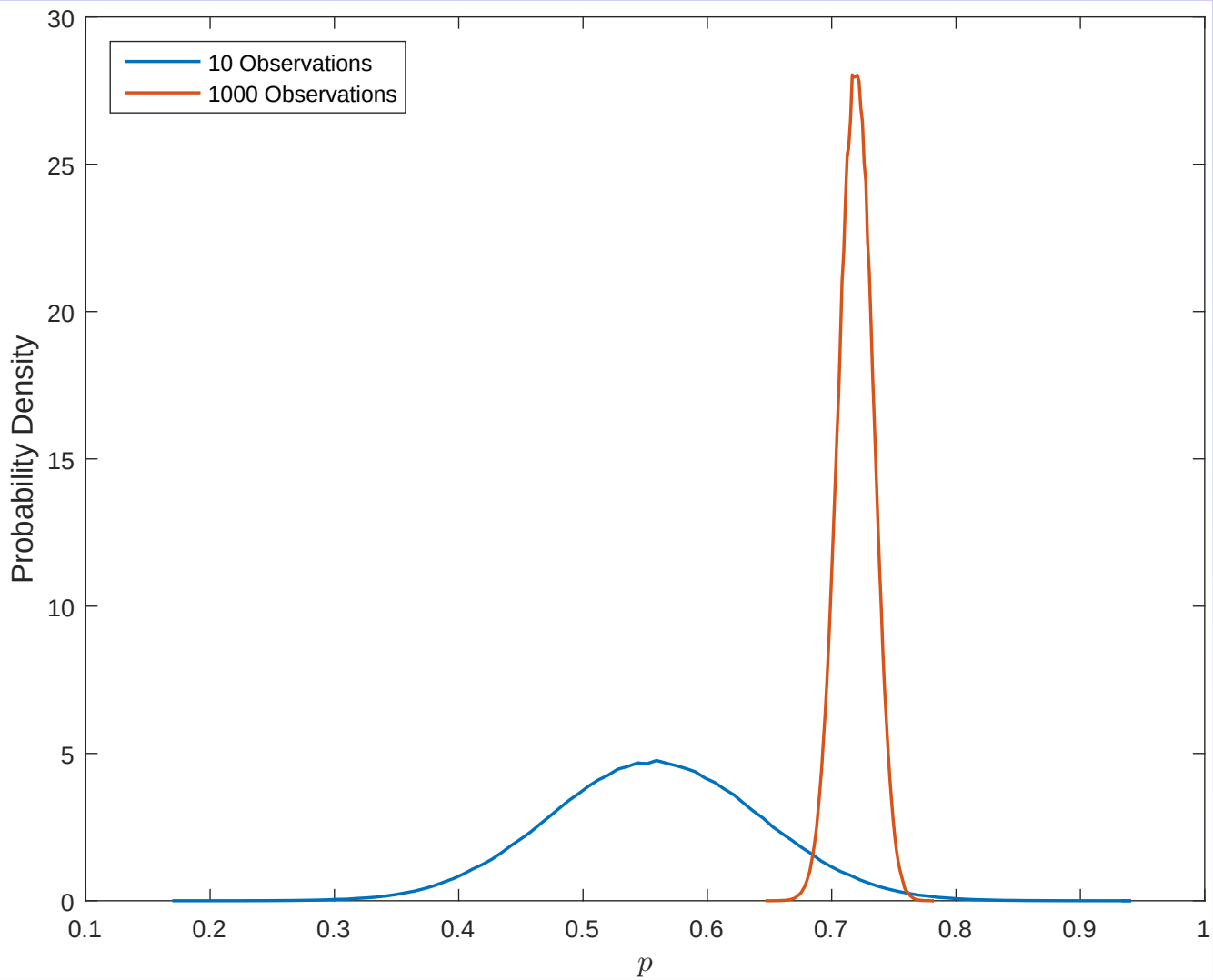
Experiment 3: Increasing the Number of Observations

Suppose we have more observations, what effect does this have on the resulting posterior?

More observations = more data = more knowledge.

The posterior will reflect this.

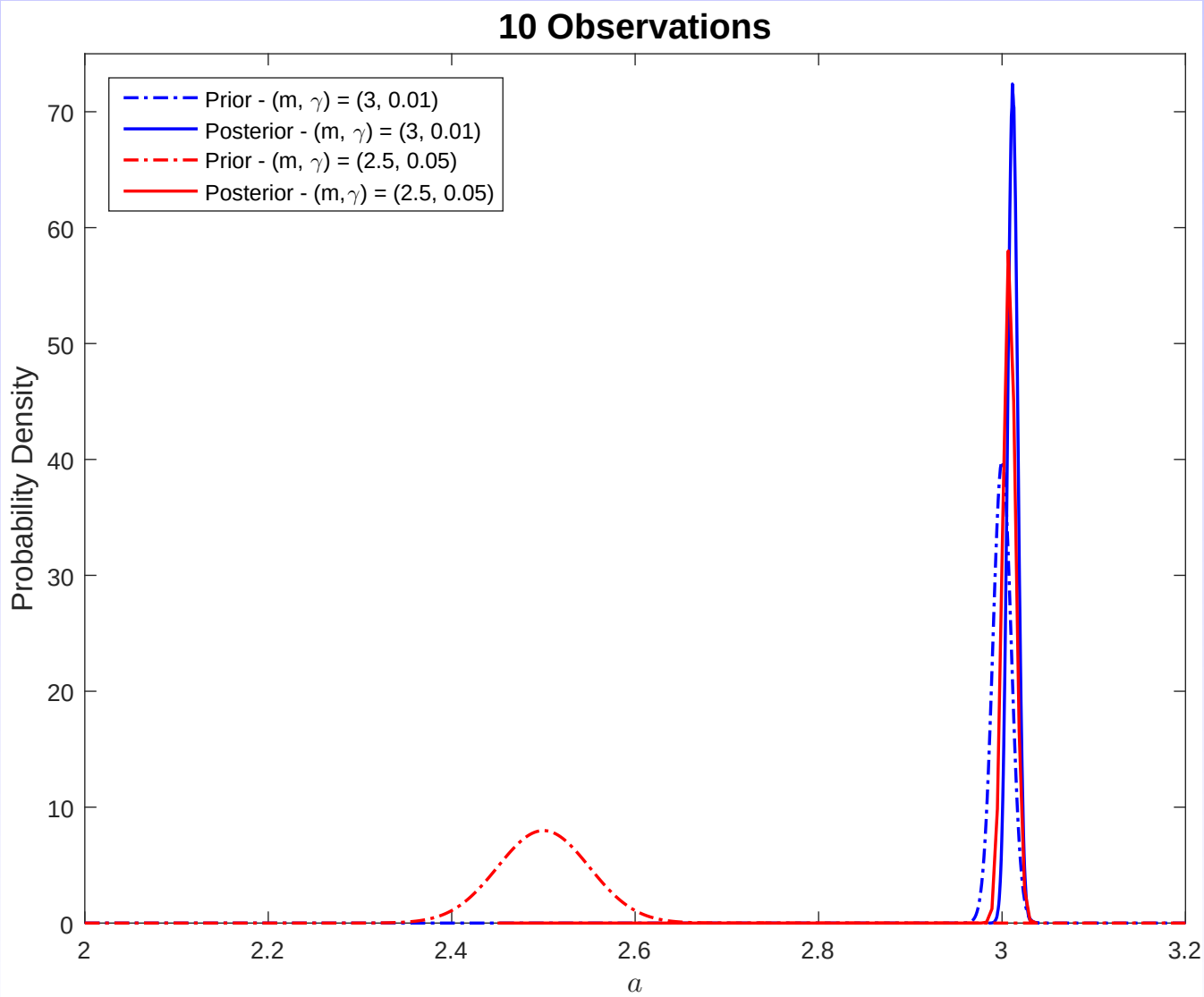
Experiment 3: Increasing the Number of Observations



Questions

Thank you for listening.

Experiment 4: Varying the Prior



Experiment 4: Varying the Prior