Monte Carlo Methods

Deep Learning Book Club

Don M. Dini
New Cortex

9/25/2017

Today

- Bayesian inference
- Monte Carlo for integration (but especially Bayesian inference)
- Markov Chain Monte Carlo
- Applications

Why do people care about Bayesian Inference?

Bayesian Inference

- Suppose we have a coin, which we strongly suspect is fair.
- We flip it 10 times - and we get:

H, T, T, T, H, T, H, T, T, T

H	3
T	7

- We hypothesize this data is produced by the Bernoulli model:

$$Pr(X; p) = p^x(1-p)^{1-x}$$

- We would like to estimate p from the observed data.

One method: Maximum Likelihood

Maximize:

$$Pr(x_1, \ldots, x_{10}) = Pr(x_1)Pr(x_2)\ldots Pr(x_{10})$$

$$p = \frac{s}{n}$$

One method: Maximum Likelihood

Maximize:

$$Pr(x_1, \ldots, x_{10}) = Pr(x_1)Pr(x_2)\ldots Pr(x_{10})$$

$$p = \frac{s}{n}$$

(In this example - p=0.3)

Is this any *good*?

- If we flipped it *many* times, and *still* got 0.3, we might believe it.

- But we haven’t. So we’re suspicious.

2nd method: Bayesian inference

Estimate:

$$f(p \mid x^n)$$

2nd method: Bayesian inference

Estimate:

$$f(p \mid x^n)$$

$$\begin{aligned} f(p \mid x^n) &= cPr(x^n \mid p) f(p) \\ &= cp^s(1-p)^{n-s} * 1 \\ &= Beta(s+1, n-s+1) \end{aligned}$$

Where $Beta(p; a, b) = Kp^{a-1}(1-p)^{b-1}$

2nd method: Bayesian inference

Estimate**:**
$f(p|x^n)$

Estimating **p amounts to evaluating:**
$E[\, p|x^n \,]$

$$= \frac{s+1}{n+2}$$

2nd method: Bayesian inference

Estimate**:**
$f(p|x^n)$

Suppose we do the same calculation - but with a prior of Beta(a,b)

$f(p|x^n) = c * Pr(x^n|p)*f(p)$

2nd method: Bayesian inference

Estimate**:**
$f(p|x^n)$

Suppose we do the same calculation - but with a prior of Beta(a,b)

$f(p|x^n) = c * Pr(x^n|p)*f(p)$

$= c * p^s*(1-p)^{n-s} * Beta(a,b)$

$= Beta(a+s, b+n-s)$

2nd method: Bayesian inference

Estimate**:**
$f(p|x^n)$

The new estimate for *p* is:
E[p]

$$= \frac{s + a}{a + b + n}$$

2nd method: Bayesian inference

Estimate**:**
$f(p|x^n)$

The new estimate for *p* is:
E[p]

$$= \frac{s + a}{a + b + n}$$

This is the same as the MLE estimate - *if* there were also *a* heads and *b* tails as additional data.

It's as if a and b are *hidden counts*.

Estimate *p* after observing 3 heads and 7 tails

Estimate *p* after observing 3 heads and 7 tails

<u>Case 1: Prior of Beta(1,1)</u>

$$p = a + s / a + b + n$$

$$= 1 + 3 / 1 + 1 + 10$$

$$= 0.33$$

Estimate *p* after observing 3 heads and 7 tails

<u>Case 1: Prior of Beta(1,1)</u>

$$p = a + s / a + b + n$$

$$= 1 + 3 / 1 + 1 + 10$$

$$= 0.33$$

Close to MLE

NEW CORTEX

Estimate *p* after observing 3 heads and 7 tails

<u>Case 1: Prior of Beta(1,1)</u>

$$p = a + s / a + b + n$$

$$= 1 + 3 / 1 + 1 + 10$$

$$= 0.33$$

Close to MLE

<u>Case 2: Prior of Beta(10,10)</u>

$$p = a + s / a + b + n$$

$$= 10 + 3 / 10 + 10 + 10$$

$$= 0.43$$

Estimate *p* after observing 3 heads and 7 tails

Case 1: Prior of Beta(1,1)

$$p = a + s / a + b + n$$

$$= 1 + 3 / 1 + 1 + 10$$

$$= 0.33$$

Close to MLE

Case 2: Prior of Beta(10,10)

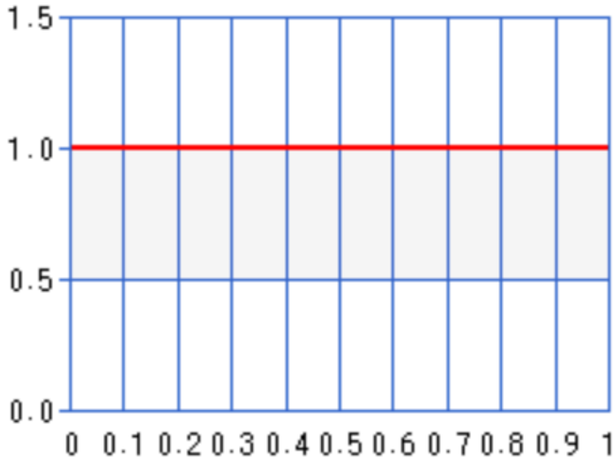
$$p = a + s / a + b + n$$

$$= 10 + 3 / 10 + 10 + 10$$

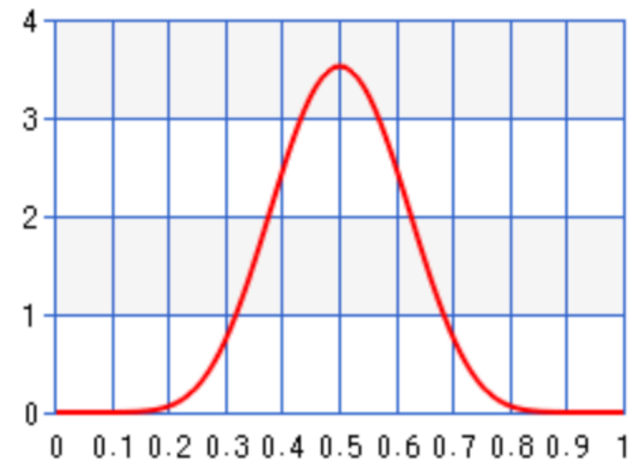
$$= 0.43$$

Not quite as far from “fair”

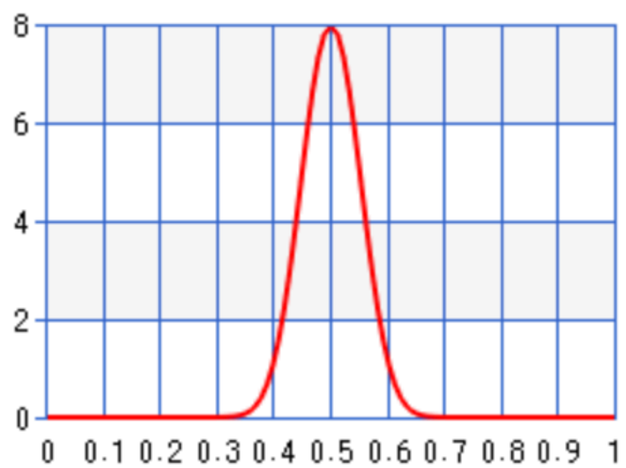
Beta(1,1)



Beta(10,10)



Beta(50,50)



Greater certainty is captured as a sharper peak in the prior distribution.

Monte Carlo techniques for integration

- We must integrate the posterior to get quantities we are interested in.

$$\hat{p} = \int p * f(p \mid x^n) dp$$

- We must integrate the posterior to get quantities we are interested in.

$$\hat{p} = \int p * f(p \mid x^n) dp$$

- Sometimes that's not so easy:

$$\hat{\delta} = \int \delta(p_1, p_2, \dots) * f(p \mid x, y, \dots) dp$$

- Sometimes it’s impossible

$$\hat{p} = \int p * \underbrace{f(p \mid x^n)}_{\text{Represented by an algorithm (e.g. NN)}} dp$$

Monte Carlo Methods

Monte Carlo Methods

$$I = \int_a^b h(x) d\,x$$

Monte Carlo Methods

$$I = \int_a^b h(x) d\,x$$

Reformulate this as:

$$I = \int_a^b w(x) * f(x) d\,x$$

Monte Carlo Methods

$$I = \int_a^b h(x) d\,x$$

Reformulate this as:

$$I = \int_a^b w(x) * \underbrace{f(x)}_{\text{Interpreted as probability distribution}} d\,x$$

Monte Carlo Methods

$$I = \int_a^b w(x) * f(x) d\,x$$

If **you can do that - then this can be evaluated in the following way:**

$$x_1, x_2, \ldots, x_n \sim f(x)$$

Average: $w(x_1), w(x_2), \ldots w(x_n)$

Simple example

$$I = \int_0^1 x^3 d\,x$$

$$I = \int_a^b w(x) * f(x) d\,x$$

Simple example

$$I = \int_0^1 x^3 d\,x$$

$$I = \int_a^b w(x) * f(x) d\,x$$

$$w(x) = x^3 * (1 - 0)$$

$$f(x) = \frac{1}{1 - 0}$$

Simple example

$$I = \int_0^1 x^3 d\,x$$

$$I = \int_a^b w(x) * f(x) d\,x$$

$$w(x) = x^3 * (1 - 0)$$

$$f(x) = \frac{1}{1 - 0}$$

$$I = \int_0^1 x^3 * (1 - 0) * \frac{1}{1 - 0} d\,x$$

Simple example

$$I = \int_0^1 x^3 d\,x$$

$$w(x) = x^3 * (1 - 0)$$

$$f(x) = \frac{1}{1 - 0}$$

We've changed the above integral, to finding the expectation:

$$E\,[x^3 * (1 - 0)]$$

Where *x* is uniformly distributed between [0,1]

Simple example

$$I = \int_0^1 x^3 d\,x$$

$$w(x) = x^3 * (1 - 0)$$

$$f(x) = \frac{1}{1 - 0}$$

$$x_1, x_2, \ldots, x_n \sim Uniform(0,1)$$

$$\hat{I} = \frac{1}{n} * ((x_1)^3 + (x_2)^3 + \ldots + (x_n)^3)$$

More interesting example

$$f(x) = N(0,1) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

We would like to know P(Z > 3)

More interesting example

$$f(x) = N(0,1) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

We would like to know P(Z > 3)

$$I = \int_3^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

$$I = \int_3^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Reformulate as $I = \int_a^b w(x) * f(x) dx$

$$w(x) = \begin{cases} \text{1 if x>3} \\ \text{0 else} \end{cases} \qquad f(x) = N(0,1)$$

NEW CORTEX

$$I = \int_3^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Reformulate as $I = \int_a^b w(x) * f(x) dx$

$$w(x) = \begin{cases} 1 \text{ if } x>3 \\ 0 \text{ else} \end{cases} \qquad f(x) = N(0,1)$$

Generate: $x_1, x_2, \ldots, x_n \sim N(0,1)$

Evaluate average: $\hat{I} = \frac{1}{n} * (w(x_1) + w(x_2) + \ldots + w(x_n))$

$\hat{I} \approx .0015$ (For N = 100)

Importance Sampling

$$I = \int_a^b w(x) * f(x) d\,x$$

There are cases where we may, for many reasons, wish to sample from a different distribution than *f(x)*, in order to find the expectation of *w(x)*.

f(x) might be difficult to sample from. Or it might be more convenient to sample from a different distribution.

Consider again

$$I = \int_3^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

We evaluated above by reformulating as

$$w(x) = \begin{cases} \text{1 if x>3} \\ \text{0 else} \end{cases} \qquad f(x) = N(0,1)$$

Then generating: $x_1, x_2, \ldots, x_n \sim N(0,1)$

Consider again

$$I = \int_3^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

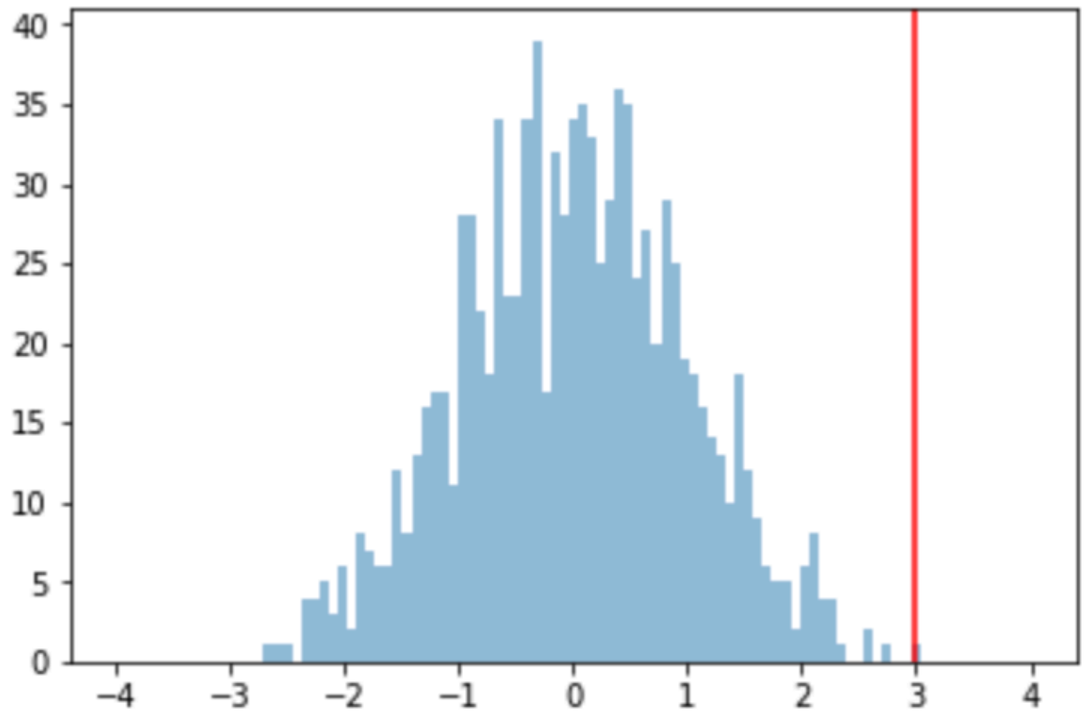
We evaluated above by reformulating as

$$w(x) = \begin{cases} 1 \text{ if } x>3 \\ 0 \text{ else} \end{cases} \qquad f(x) = N(0,1)$$

Then generating: $x_1, x_2, \ldots, x_n \sim N(0,1)$

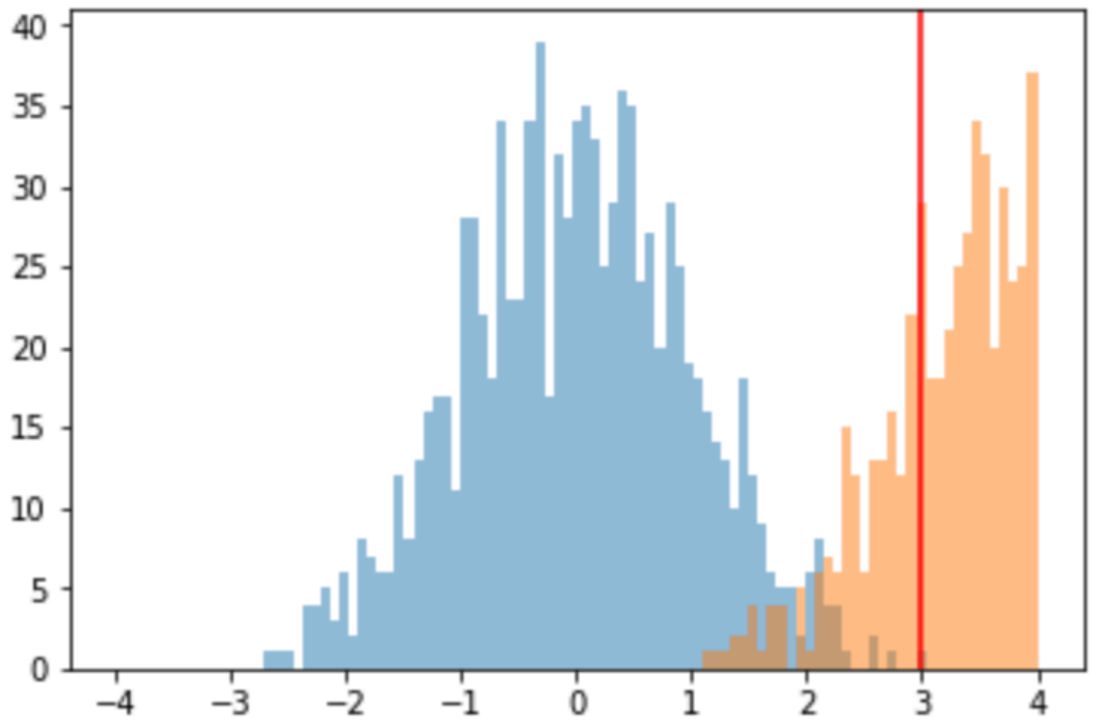
Most of these are less than *3*.

$$I = \int_3^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$



Most of the points are “wasted”

$$I = \int_3^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} d\,x$$



We'd have a higher quality sample if the data points came from N(4, 1), instead.

Importance Sampling

$$I = \int_a^b w(x) * f(x) d\,x$$

You can change this to instead sample from a preferred, g(x):

$$I = \int_a^b \frac{w(x) f(x)}{g(x)} g(x) d\,x$$

Importance Sampling

$$I = \int_a^b w(x) * f(x) d\,x$$

You can change this to instead sample from a preferred, g(x):

$$I = \int_a^b \frac{w(x) f(x)}{g(x)} g(x) d\,x$$

Now, you're calculating the expectation of: $\frac{w(x) f(x)}{g(x)}$

using samples drawn from *g(x)*

$$I = \int_3^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Reformulate as $I = \int_a^b \frac{w(x)f(x)}{g(x)} g(x) dx$

$$w(x) = \begin{cases} 1 \text{ if } x>3 \\ 0 \text{ else} \end{cases} \qquad f(x) = N(0,1) \qquad g(x) = N(4,1)$$

$$I = \int_3^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Reformulate as $I = \int_a^b \frac{w(x)f(x)}{g(x)} g(x) dx$

$$w(x) = \begin{cases} 1 \text{ if } x>3 \\ 0 \text{ else} \end{cases} \qquad f(x) = N(0,1) \qquad g(x) = N(4,1)$$

Generate: $x_1, x_2, \ldots, x_n \sim N(4,1)$

Evaluate average: $\hat{I} = \frac{1}{n} * (\frac{w(x_1)f(x_1)}{g(x_1)} + \frac{w(x_2)f(x_2)}{g(x_2)} + \ldots + \frac{w(x_3)f(x_3)}{g(x_3)})$

$$\hat{I} \approx .0011$$

$$I = \int_3^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Reformulate as $I = \int_a^b \frac{w(x)f(x)}{g(x)} g(x) dx$

$$w(x) = \begin{cases} 1 \text{ if } x>3 \\ 0 \text{ else} \end{cases} \qquad f(x) = N(0,1) \qquad g(x) = N(4,1)$$

Generate: $x_1, x_2, \ldots, x_n \sim N(4,1)$

Evaluate average: $\hat{I} = \frac{1}{n} * (\frac{w(x_1)f(x_1)}{g(x_1)} + \frac{w(x_2)f(x_2)}{g(x_2)} + \ldots + \frac{w(x_3)f(x_3)}{g(x_3)})$

$\hat{I} \approx .0011$

$V(\hat{I})$ ***went from .0039 to .0002!***

Where we are so far

Where we are so far

- Bayesian inference is super important.

- Integrals that are hard to evaluate appear frequently in Bayesian inference.

- If you can find a way to recast as:

$$I = \int_a^b w(x) * f(x) d\,x$$

then you can sample *f*, and evaluate average of *w(x)*.

- Sometimes sampling *f* is tough, however.

Markov Chain Monte Carlo

NEW CORTEX

MCMC is a mechanism for generating data points:

$$x_1, x_2, x_3 \ldots$$

such that their distribution settles on a specified *f*

Broad outline of Metropolis-Hastings algorithm

- Generate the first point, X_0.
- Suppose we have generated i previous points $X_0,\ldots X_i$.

1. Generate proposal for a new value, Y

2. r - probability Y is accepted as the next value

3. If Y is rejected, use the previous value as the next one. Otherwise, use Y as the new Xi+1

1. Draw proposal for new value, Y:

$$Y \sim q(y \mid X_i)$$

A common choice for q is: $N(x, b^2)$

i.e. a Normal, centered around the previous value in the sequence.

1. Draw proposal for new value, Y:

$$Y \sim q(y \mid X_i)$$

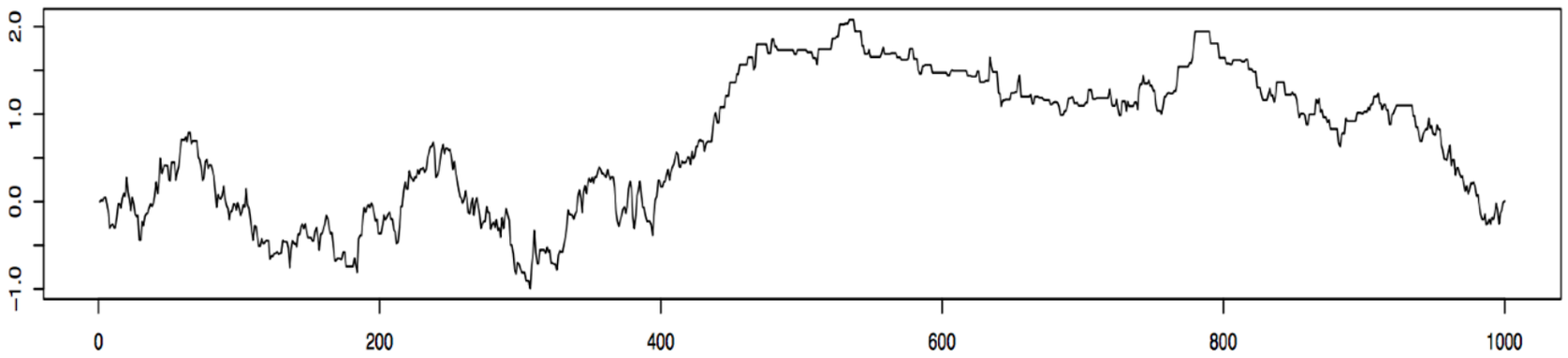
A common choice for q is: $N(x, b^2)$

i.e. a Normal, centered around the previous value in the sequence.

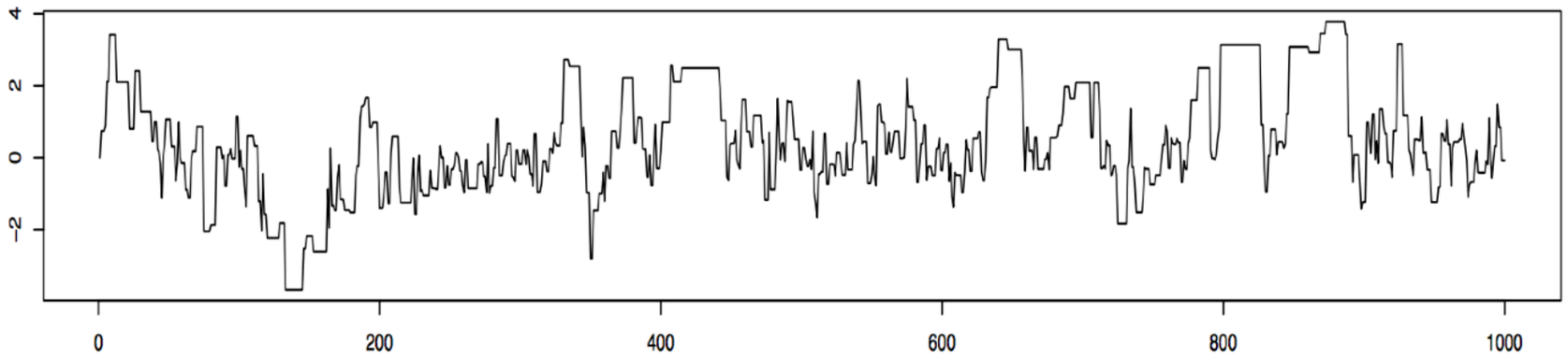
2. Form acceptance probability, r

$$r = \min \frac{f(Y)}{f(X_i)}, 1 \qquad \text{for q as } N(x, b^2)$$

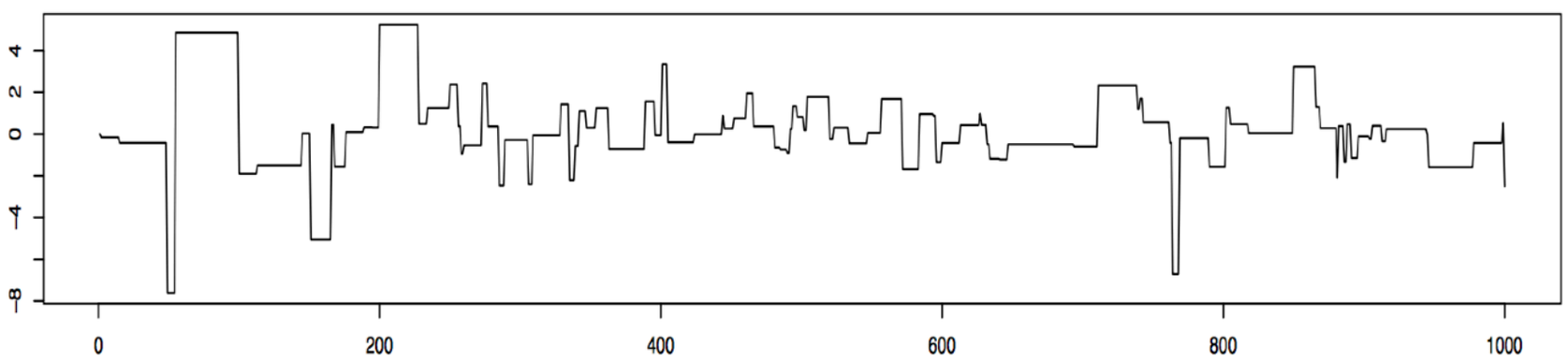
Intuitively - b controls how far to branch out from current part of space. Higher b means proposals are more "daring" - but are less likely to be accepted.


2.0
1.0
0.0
−1.0
0
200
400
600
800
1000

b=0.1


4
2
0
−2
0
200
400
600
800
1000

b=1


4
2
0
−4
−8
0
200
400
600
800
1000

b=10

Many flavors of MCMC

- Gibbs samling
- Random walk - Metropolis
- …

Applications

Applications

Inference in large graphical models

```
 100 ER ENOHDLAE OHDLO UOZEOUNORU O UOZEO HD OITO HEOQSET IUROFHE HENO ITORUZAEN
 200 ES ELOHRNDE OHRNO UOVEOULOSU O UOVEO HR OITO HEOQAET IUSOPHE HELO ITOSUVDEL
 300 ES ELOHANDE OHANO UOVEOULOSU O UOVEO HA OITO HEOQRET IUSOFHE HELO ITOSUVDEL
 400 ES ELOHINME OHINO UOVEOULOSU O UOVEO HI OATO HEOQRET AUSOWHE HELO ATOSUVMEL
 500 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
 600 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
 900 ES ELOHANME OHANO UODEOULOSU O UODEO HA OITO HEOQRET IUSOWHE HELO ITOSUDMEL
1000 IS ILOHANMI OHANO RODIORLOSR O RODIO HA OETO HIOQUIT ERSOWHI HILO ETOSRDMIL
1100 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1200 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1300 ISTILOHARMITOHAROT ODIO LOS TOT ODIOTHATOENOTHIOQUINTE SOWHITHILOTENOS DMIL
1400 ISTILOHAMRITOHAMOT OFIO LOS TOT OFIOTHATOENOTHIOQUINTE SOWHITHILOTENOS FRIL
1600 ESTEL HAMRET HAM TO CE OL SOT TO CE THAT IN THE QUENTIOS WHETHEL TIN SOCREL
1700 ESTEL HAMRET HAM TO BE OL SOT TO BE THAT IN THE QUENTIOS WHETHEL TIN SOBREL
1800 ESTER HAMLET HAM TO BE OR SOT TO BE THAT IN THE QUENTIOS WHETHER TIN SOBLER
1900 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
2000 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
```

References

- *All of Statistics* - Larry Wasserman
- *Probabilistic Graphical Models* - Koller and Friedman
- Markov Chain Monte Carlo Revolution - Diaconis

Thanks!

don@newcortex.ai

@dondini

NEW CORTEX