

Where there is no uncertainty,
there cannot be truth

– Richard Feynman

Where there is no uncertainty,
there cannot be truth

– *Richard Feynman*

There are no routine statistical
questions, only questionable
statistical routines – *Sir David Cox*

Using MCMC for inference using
Hamiltonian Monte Carlo
with Stan

random@yorku.ca

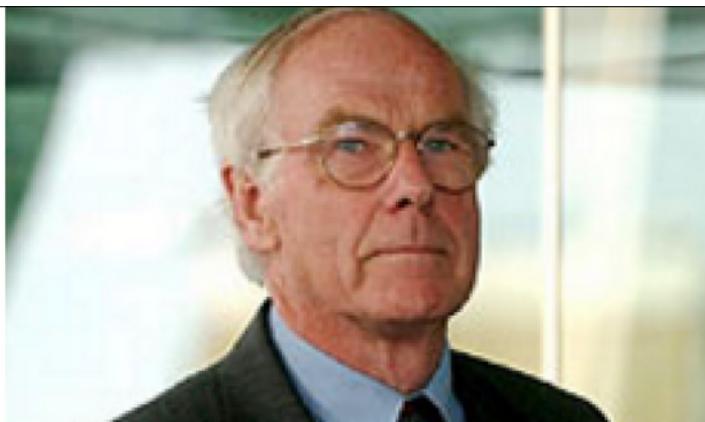
Send me an email message and
I'll add your name to a Piayza
course with useful links.

- Working with uncertainty
 - Frequentist vs. Bayesian inference
- Evolution of MCMC
 - MH, Gibbs, Hamiltonian MC
- Using Stan for HMC
 - non-linear mixed model
 - Simple ODE model



Sally Clark

- ▶ Young lawyer, gives birth to first son in September 1996
- ▶ son dies, apparently of SIDS, at 10 weeks
- ▶ second son born a year later
- ▶ dies, apparently of SIDS, at 8 weeks
- ▶ only evidence of trauma consistent with resuscitation attempts
- ▶ charged with two counts of murder



Sir Roy Meadow

- ▶ distinguished pediatrician
- ▶ as expert witness testifies:
 - ▶ probability of one SIDS death: $\frac{1}{8,500}$
 - ▶ probability of two: $\left(\frac{1}{8,500}\right)^2 = \frac{1}{72,250,000}$
 - ▶ ‘if she’s innocent, the chances of this happening are 1 in 72 million’
- ▶ jury convicts Sally Clark of murder in November 1999
- ▶ first appeal lost in October 2000
- ▶ second appeal succeeds and Sally Clark is released in January 2003
- ▶ she dies in 2007 at the age of 42

H_0 : Sally is innocent

H_0 : Sally is innocent

γ : 2 children die for no apparent cause

H_0 : Sally is innocent

y : 2 children die for no apparent cause

$$P\text{-value} = \Pr(y^+ \mid H_0)$$

H_0 : Sally is innocent

y : 2 children die for no apparent cause

$$P\text{-value} = \Pr(y^+ \mid H_0)$$

Meadow's calculation

$$\approx \frac{1}{8,500} \times \frac{1}{8,500} = \frac{1}{72,250,000}$$

H_0 : Sally is innocent

y : 2 children die for no apparent cause

$$P\text{-value} = \Pr(y^+ \mid H_0)$$

Meadow's calculation

$$\approx \frac{1}{8,500} \times \frac{1}{8,500} = \frac{1}{72,250,000}$$

Criticism :

H_0 : Sally is innocent

y : 2 children die for no apparent cause

$$P\text{-value} = \Pr(y^+ \mid H_0)$$

Meadow's calculation

$$\approx \frac{1}{8,500} \times \frac{1}{8,500} = \frac{1}{72,250,000}$$

Criticism : 1) assumes independence

H_0 : Sally is innocent

y : 2 children die for no apparent cause

$$P\text{-value} = \Pr(y^+ \mid H_0)$$

Meadow's calculation

$$\approx \frac{1}{8,500} \times \frac{1}{8,500} = \frac{1}{72,250,000}$$

- Criticism :
- 1) assumes independence
 - 2) $\frac{1}{8,500}$ too small

H_0 : Sally is innocent

y : 2 children die for no apparent cause

$$P\text{-value} = \Pr(y^+ \mid H_0)$$

Meadow's calculation

$$\approx \frac{1}{8,500} \times \frac{1}{8,500} = \frac{1}{72,250,000}$$

- Criticism :
- 1) assumes independence
 - 2) $\frac{1}{8,500}$ too small

Correct p-value is larger - maybe $\frac{1}{10,000}$!

BUT:

BUT:

Do we really want $P(y^+ | H_0)$?

BUT:

Do we really want $P(Y^+ | H_0)$?

Don't we really want $P(H_0 | Y)$?

BUT:

Do we really want $P(Y^+ | H_0)$?

Don't we really want $P(H_0 | Y)$?

- must be close!?
- Is $P(Y^+ | H_0)$ a good proxy for $P(H_0 | Y)$?

BUT:

Do we really want $P(Y^+ | H_0)$?

Don't we really want $P(H_0 | Y)$?

- must be close!?
- Is $P(Y^+ | H_0)$ a good proxy for $P(H_0 | Y)$?

Does it establish guilt beyond a reasonable doubt?

BUT:

Do we really want $P(Y^+ | H_0)$?

Don't we really want $P(H_0 | Y)$?

= must be close!?

= Is $P(Y^+ | H_0)$ a good

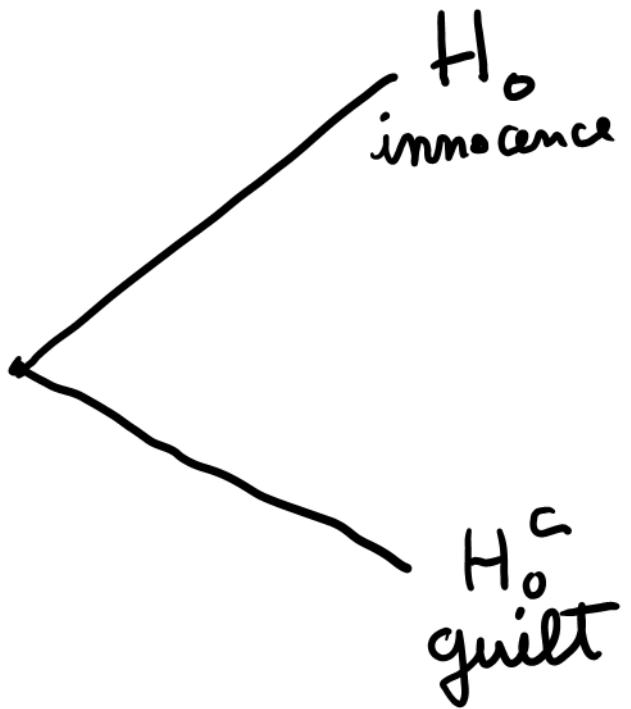
proxy for $P(H_0 | Y)$?

Let's find out:

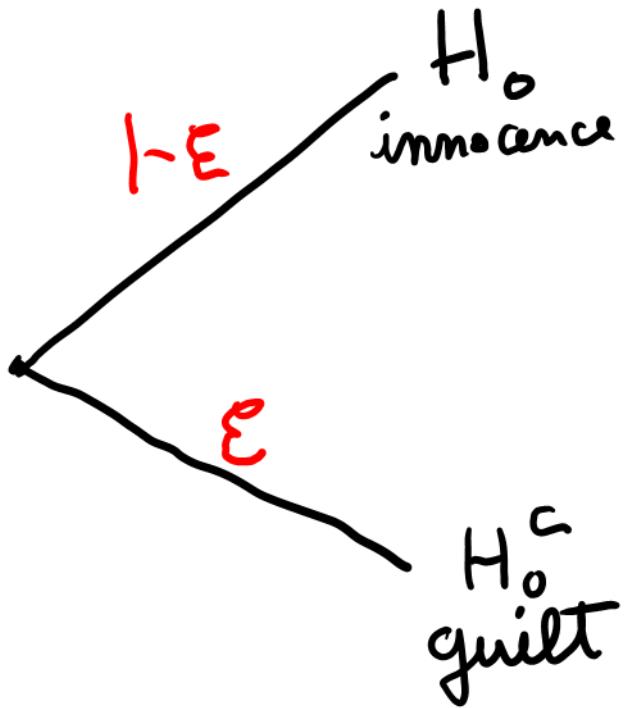
$$P(H_0 | Y) = \frac{P(H_0, Y)}{P(Y)} = \frac{P(Y | H_0) P(H_0)}{P(Y)}$$

Bayesian tree:

Bayesian tree:

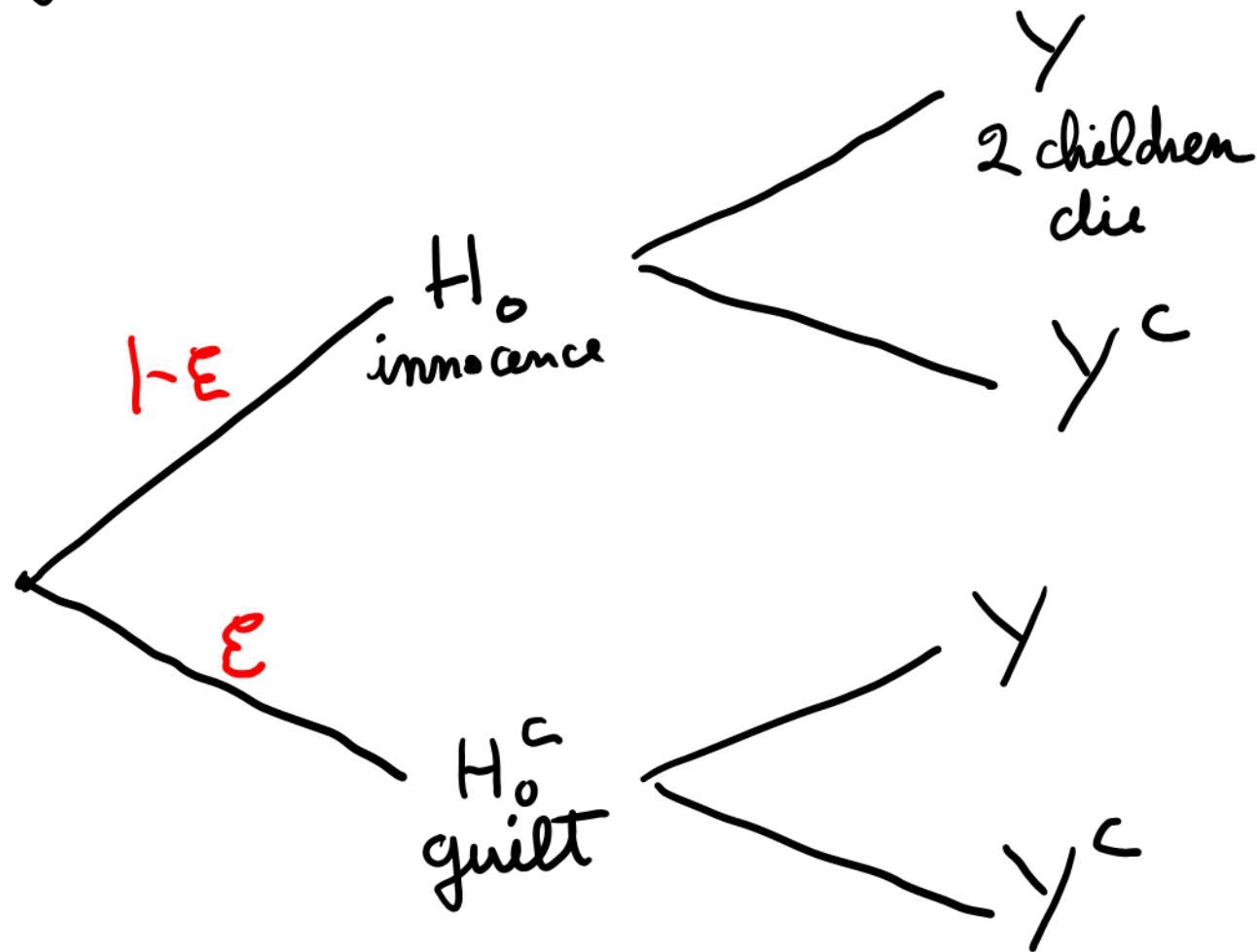


Bayesian tree:

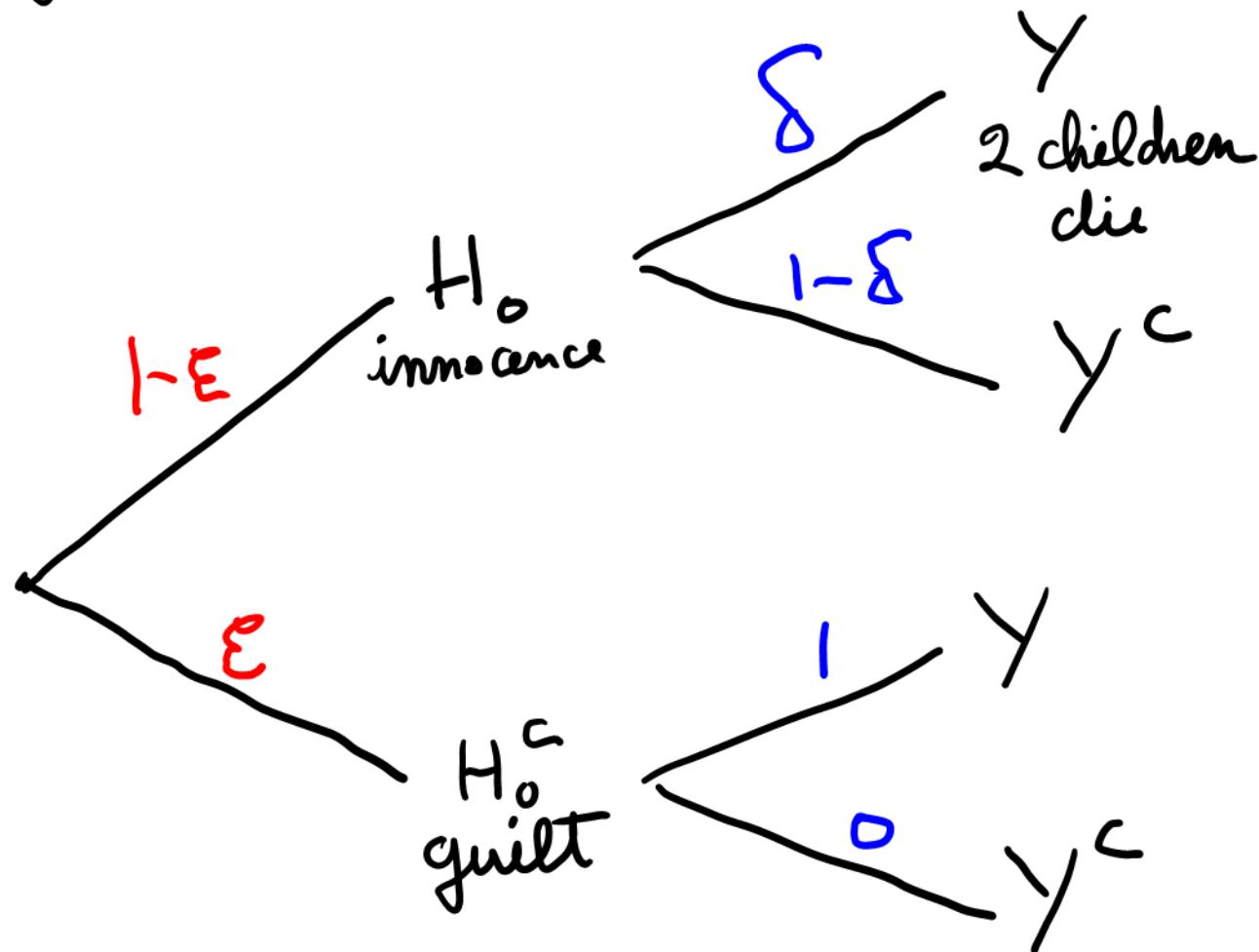


E = very small number

Bayesian tree:

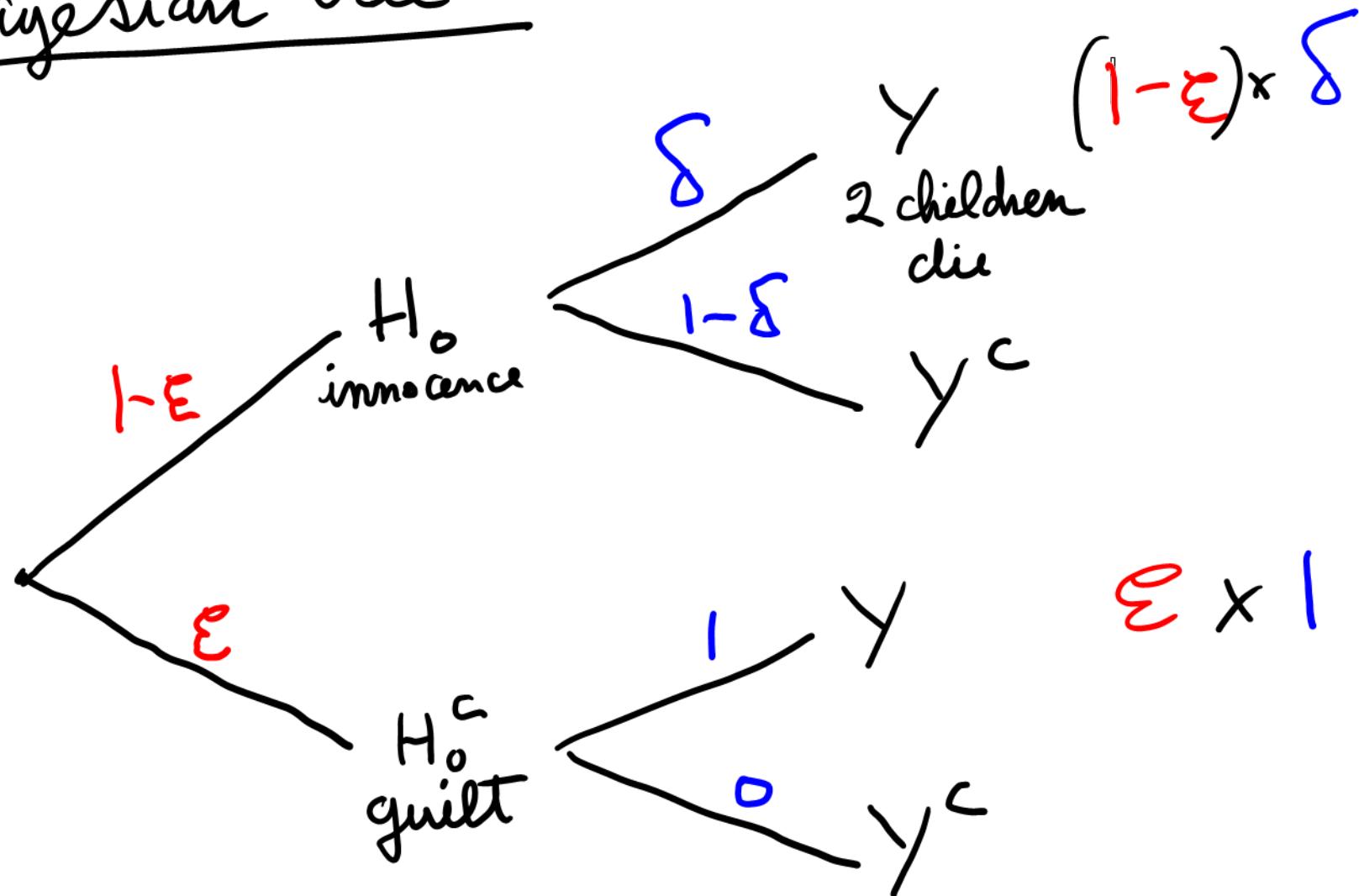


Bayesian tree:



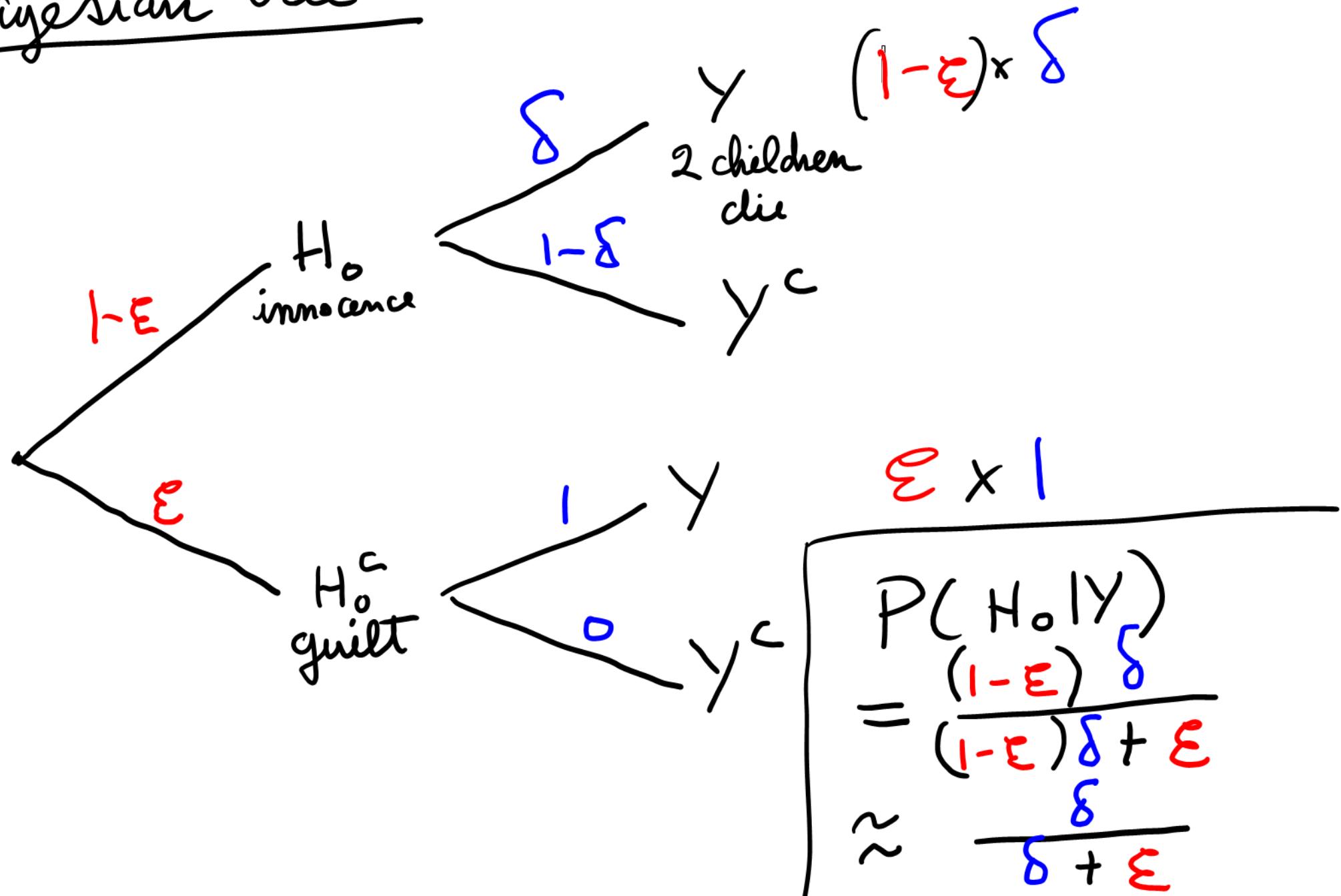
$\delta = \text{very small number}$

Bayesian tree:

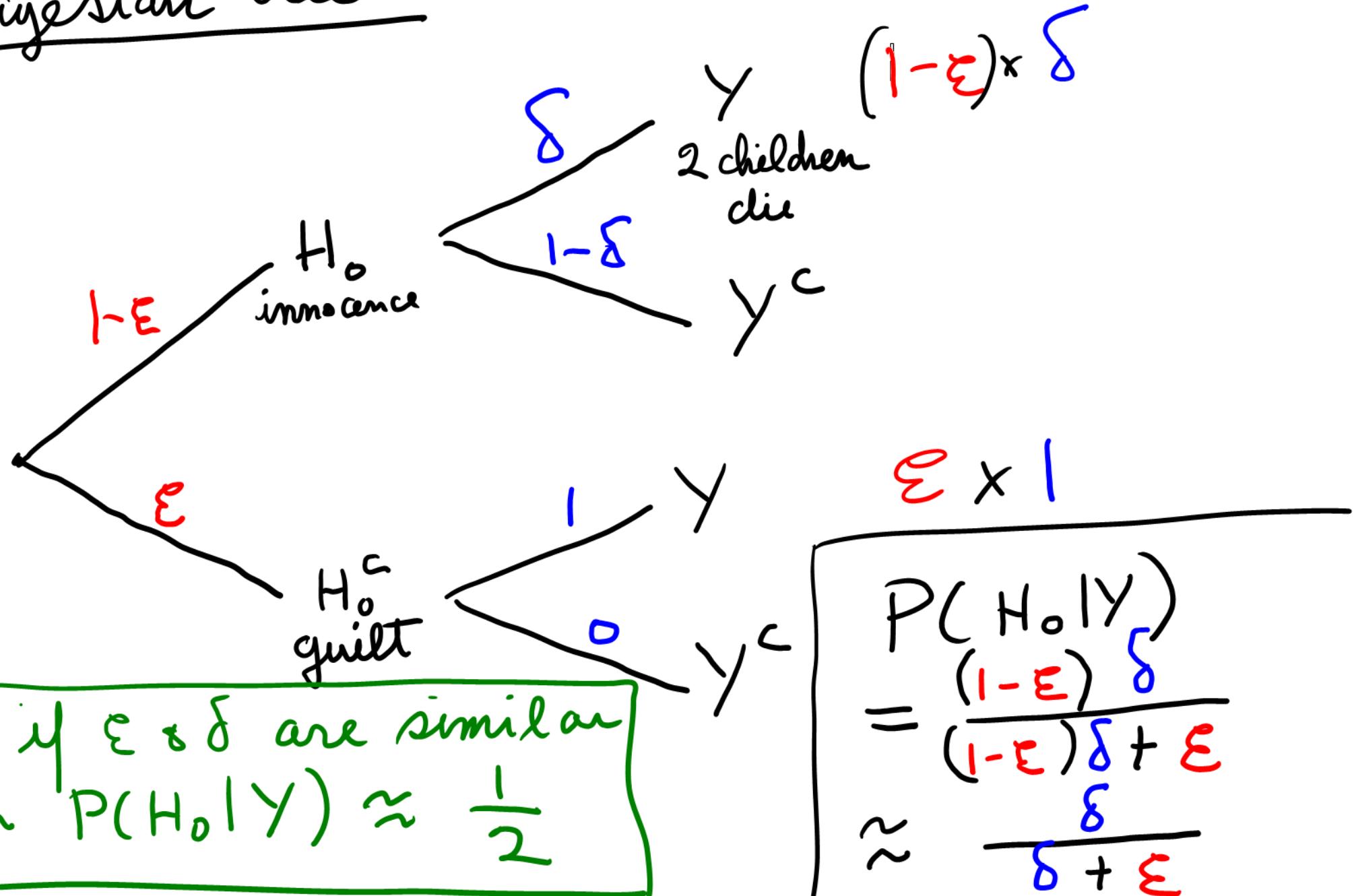


$\delta = \text{very small number}$

Bayesian tree:

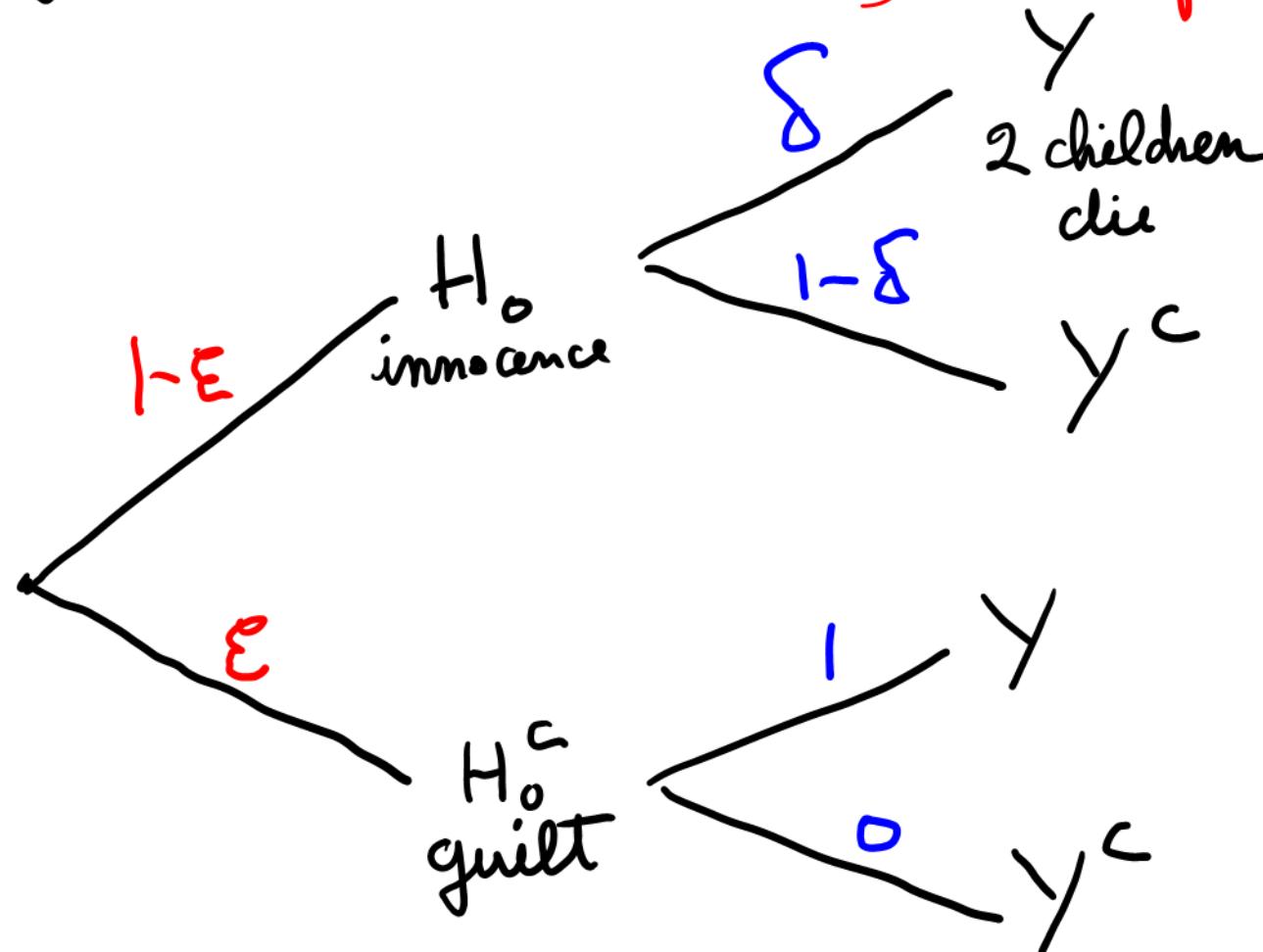


Bayesian tree:



So if $\varepsilon \approx \delta$ are similar
then $P(H_0 | Y) \approx \frac{1}{2}$

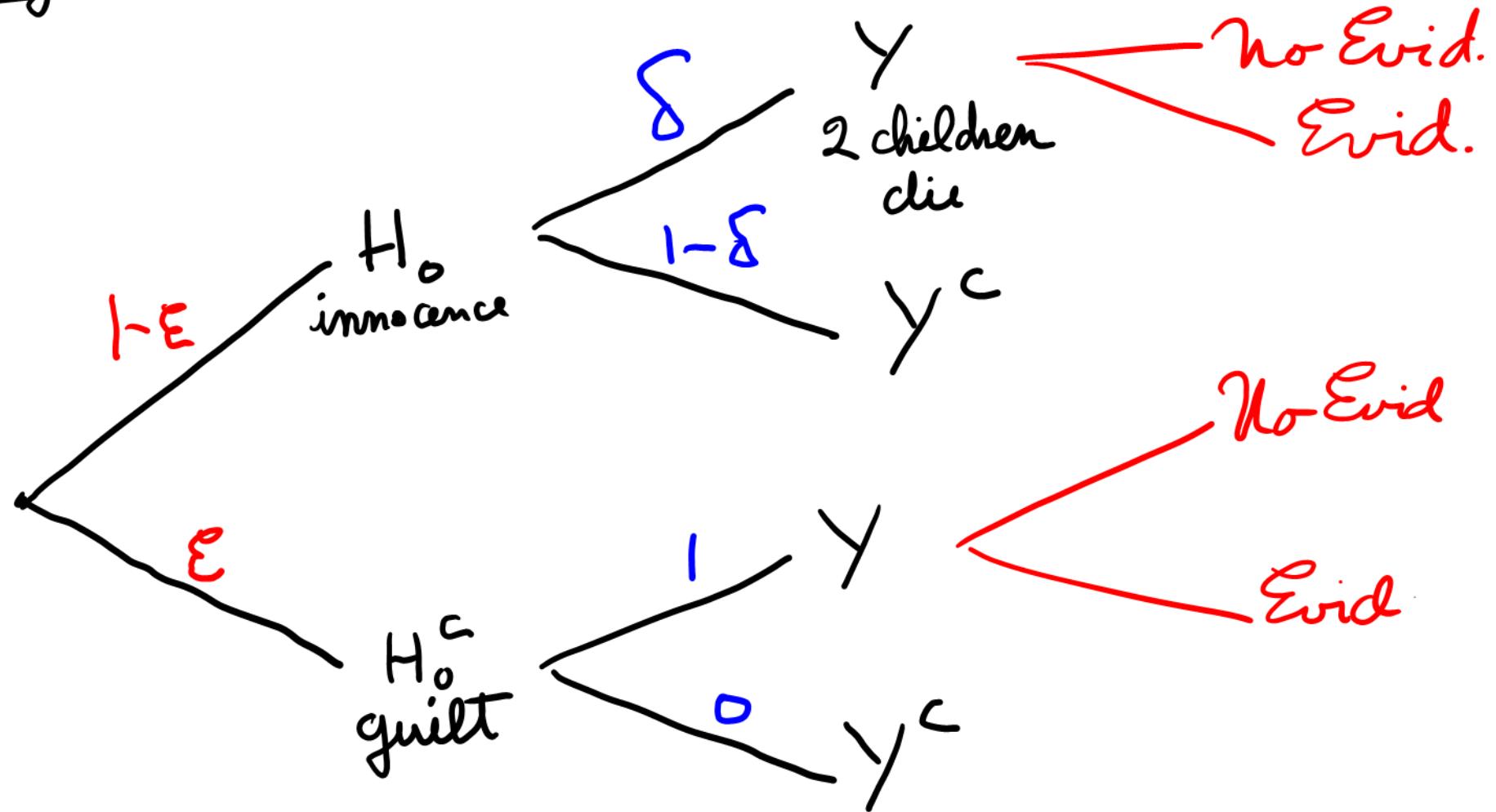
Bayesian tree:



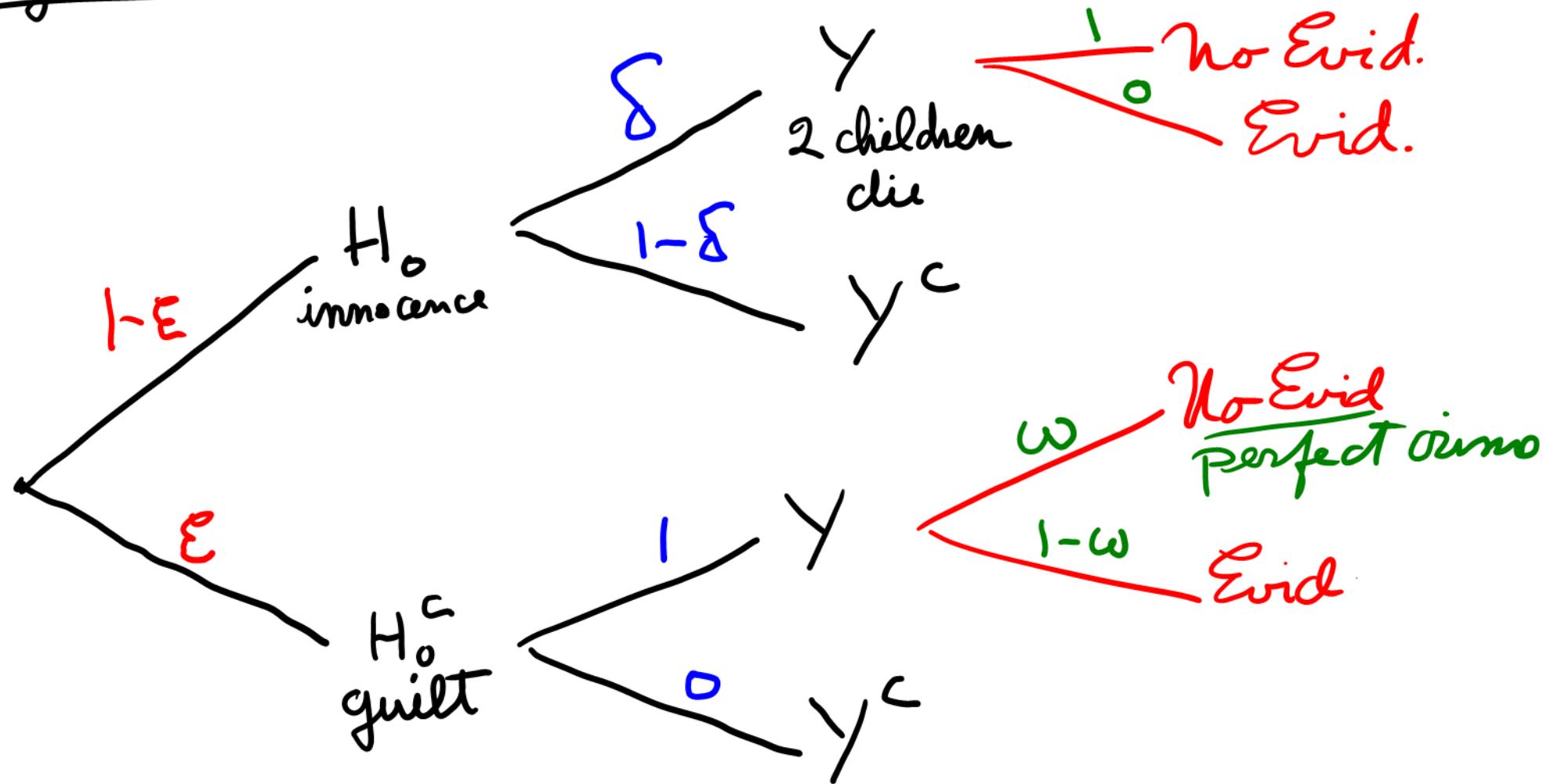
But this ignores
lack of evidence

If it's a
crime, it's
a "perfect"
crime - and,
presumably,
only a
small proportion
of crimes are
perfect crimes.

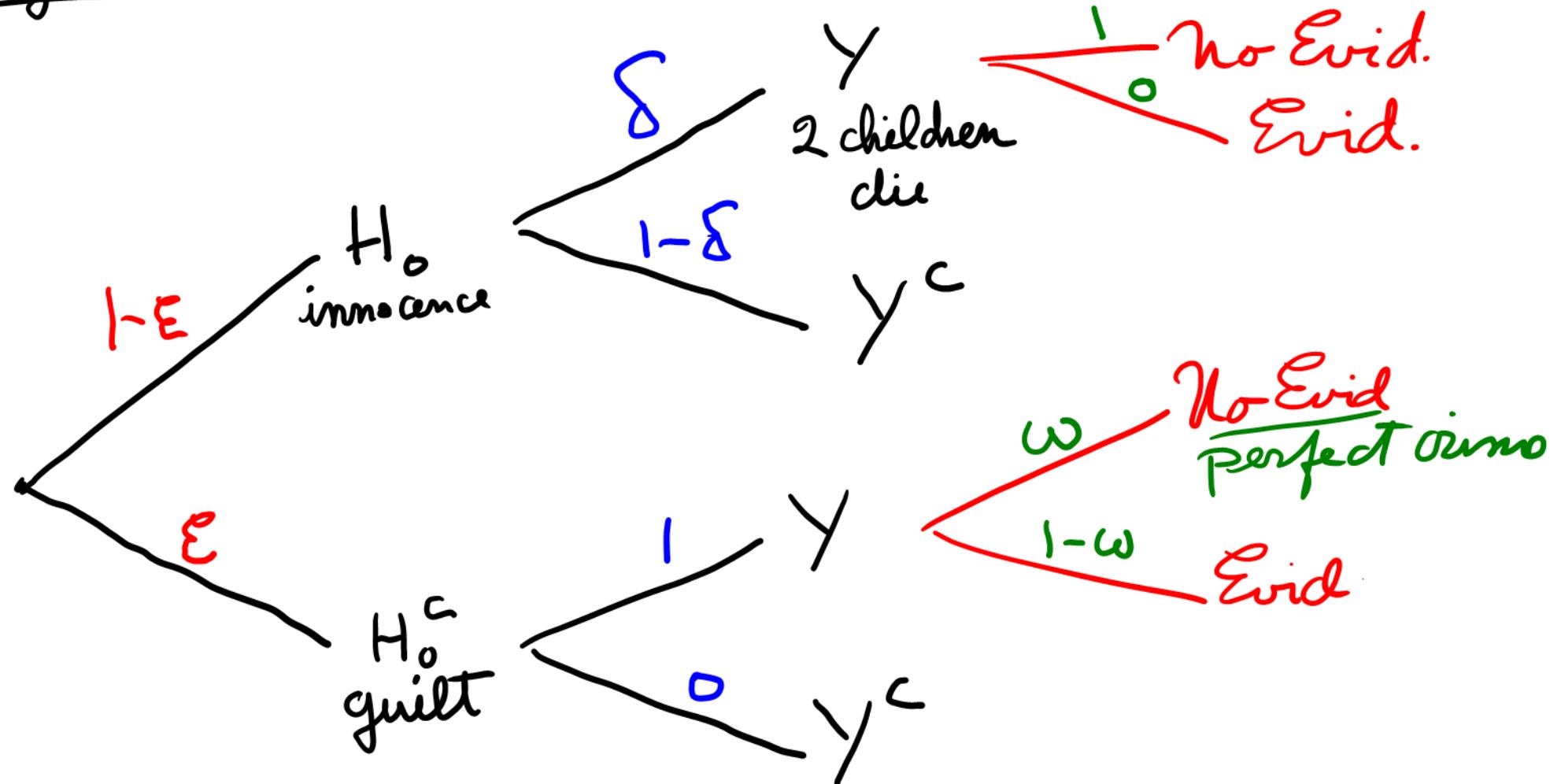
Bayesian tree:



Bayesian tree:



Bayesian tree:



$$Pr(H_0 | y, \text{No Evid}) \approx \frac{\delta}{\delta + \epsilon \omega} \approx \text{close to } 1 !!$$

Sally Clark was
innocent beyond
a reasonable doubt.

$$P(y^+ | H_0)$$

was a very poor proxy for

$$P(H_0 | Y)$$

$$P(Y^+ | H_0)$$

was a very poor proxy for

$$P(H_0 | Y)$$

- But many (most?) scientists only know about $P(Y^+ | H_0)$.
- Many have been trained to be adamantly opposed to using "subjective" Bayesian methods instead of "objective" Frequentist methods.

— Using the p-value as a proxy for $P(H_0 | Y)$
has come to be known as the
"Prosecutor's Fallacy"

- Using the p-value as a proxy for $P(H_0 | Y)$ has come to be known as the "Prosecutor's Fallacy"
- But many (most?) scientists testifying don't know of other ways

— Using the p-value as a proxy for $P(H_0 | Y)$
has come to be known as the
"Prosecutor's Fallacy"

So, how did we get to this?

Why are we using p-values if
they are so bad?

Dilemma

Dilemma

(H) : unknown state of nature

Dilemma

H : unknown state of nature

Y : what's observed

Dilemma

H : unknown state of nature

Y : what's observed

What does Y say about H ?

Dilemma

Θ : unknown state of nature

Y : what's observed

What does Y say about Θ ?

Model : $P(Y|\Theta)$

Dilemma

Θ : unknown state of nature

Y : what's observed

What does Y say about Θ ?

Model : $P(Y|\Theta)$

From this we can get p-values,
confidence intervals, size- α tests.

Dilemma

Θ : unknown state of nature

Y : what's observed

What does Y say about Θ ?

Model : $P(Y|\Theta)$

From this we can get p-values,
confidence intervals, size-& tests.

But not $p(\Theta|Y)$ Posterior

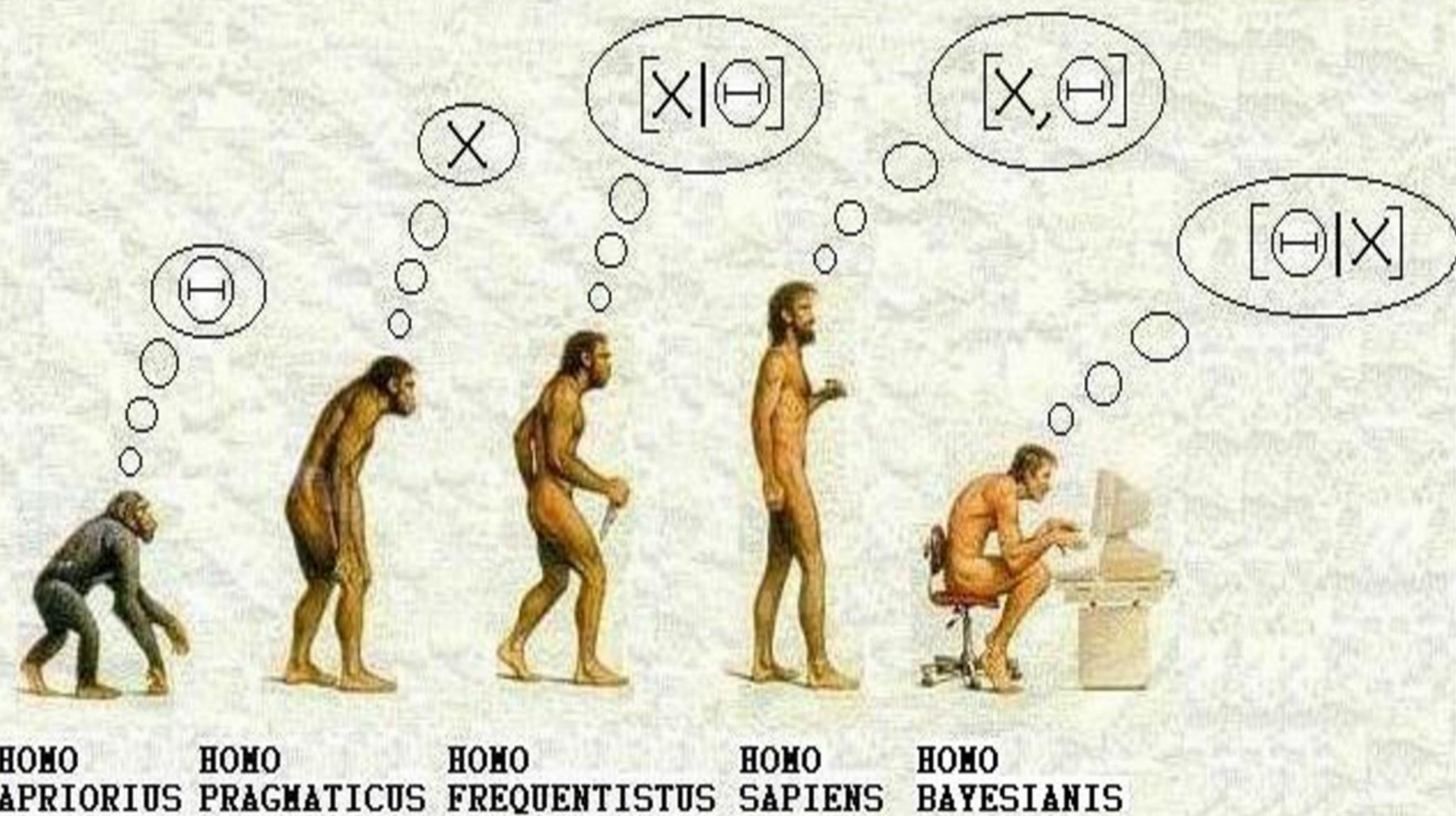
For that we need to start with $p(\Theta)$
Prior

To get what you really want, you
need to pay a price many are reluctant
to pay.

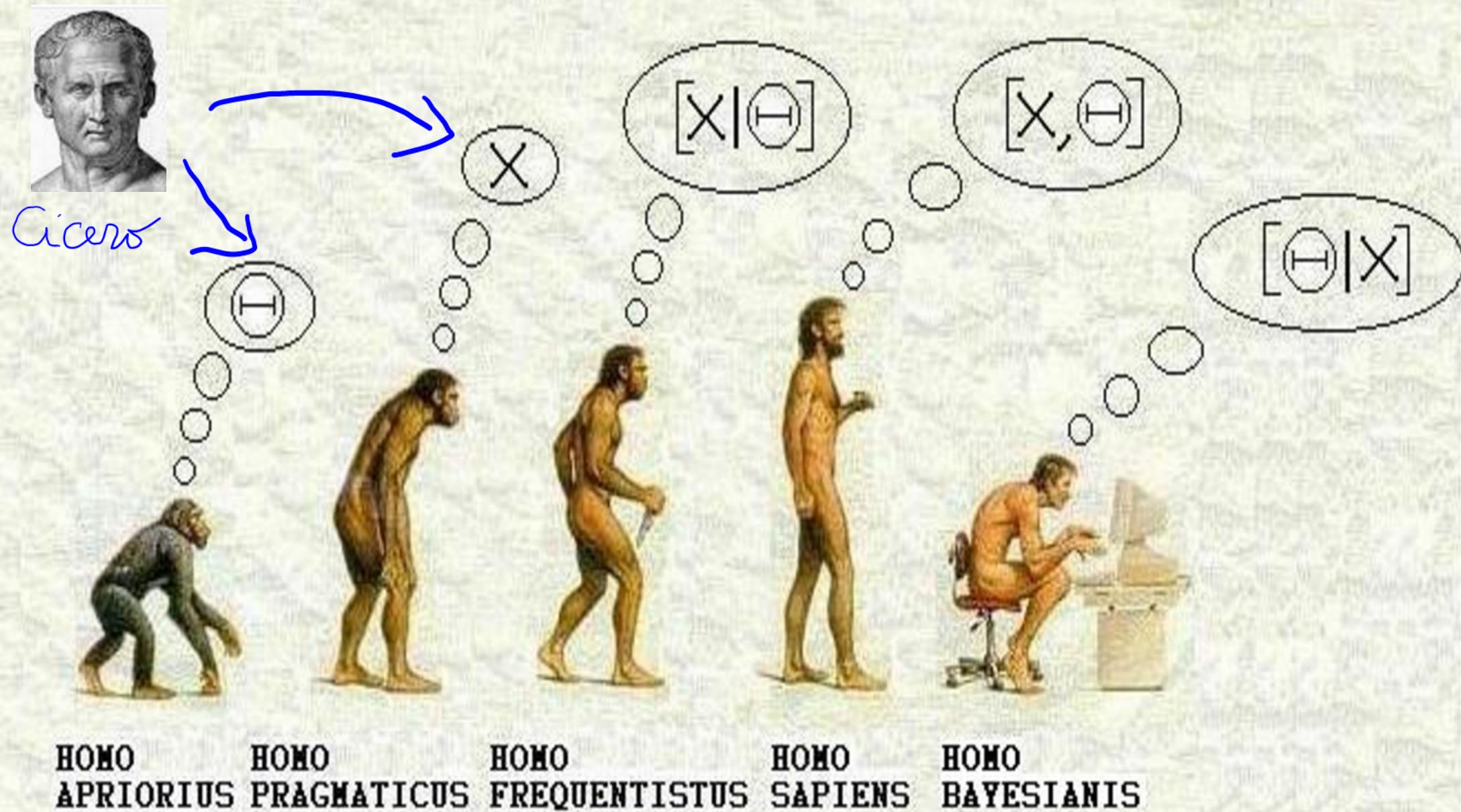
To get what you really want, you
need to pay a price many are reluctant
to pay.

A historical perspective

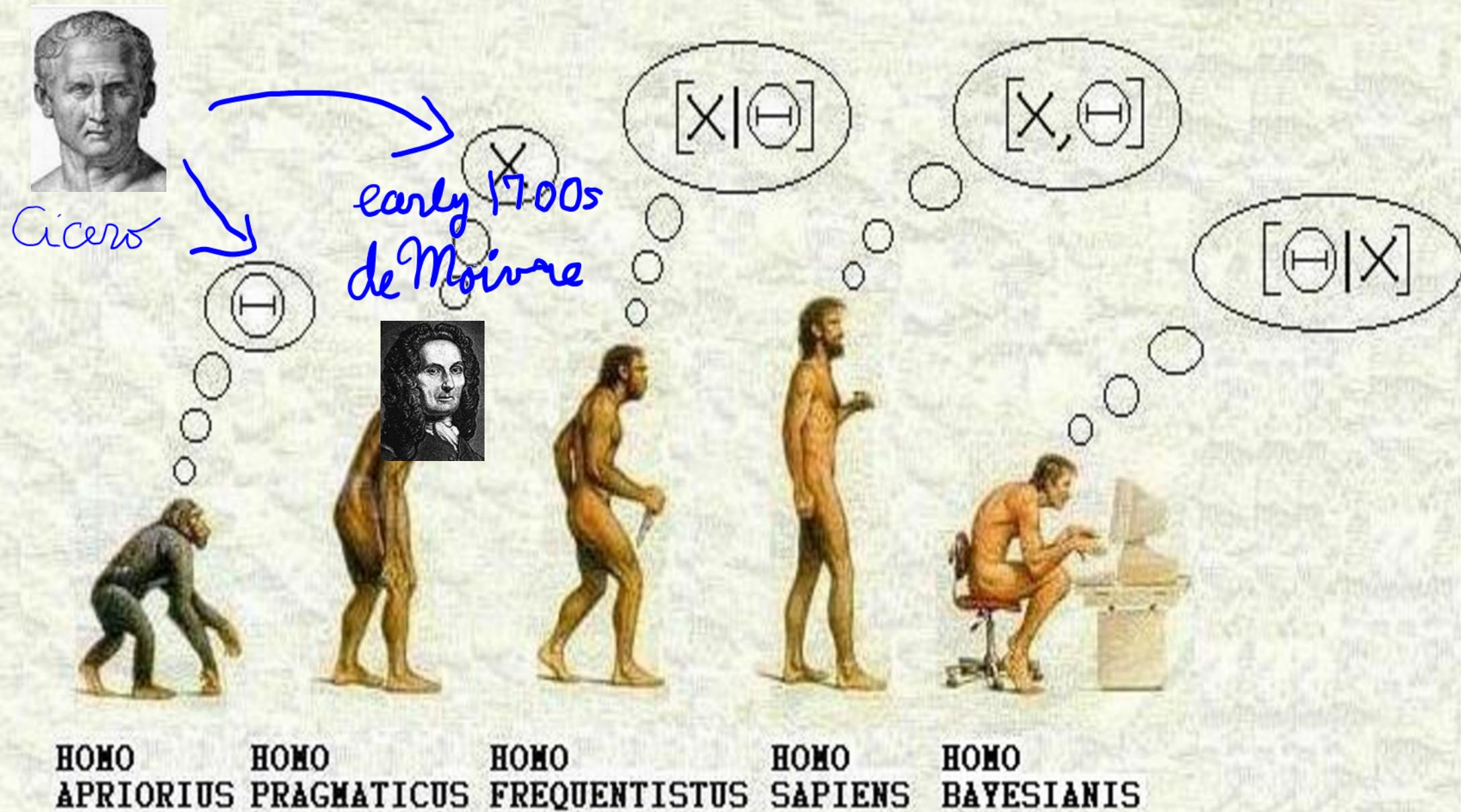
(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...



(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...



(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...



(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...

Bayes
Pascal } $P(\theta|x)$ using uniform $P(\theta)$



Cicero

early 1700s
de Moivre



late
1700s
1800s



HOMO
APRIORIUS



HOMO
PRAGMATICUS

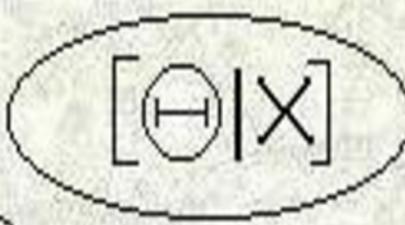
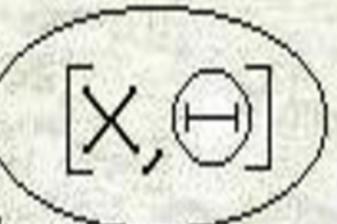
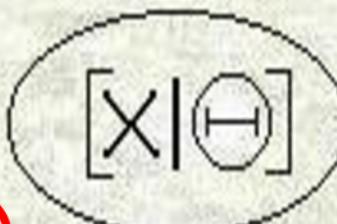
HOMO
FREQUENTISTUS



HOMO
SAPIENS



HOMO
BAYESIANIS



(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...

Bayes

Pascal

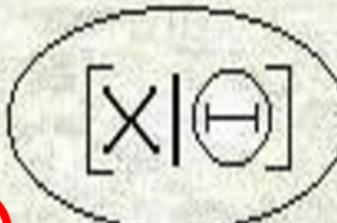
$P(\theta|x)$ using uniform $P(\theta)$



Cicero



early 1700s
de Moivre



HOMO
APRIORIUS



HOMO
PRAGMATICUS



HOMO
FREQUENTISTUS



HOMO
SAPIENS



HOMO
BAYESIANUS

Controversial

Uniform
 $\theta \sim P \in (0,1)$

not same as
Uniform of
 $\theta \sim U(0, \infty)$

odds $w \in (0, \infty)$

(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...

Bayes

Pascal

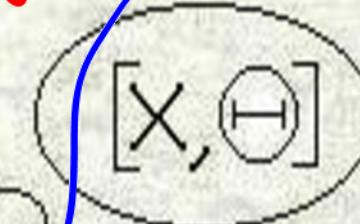
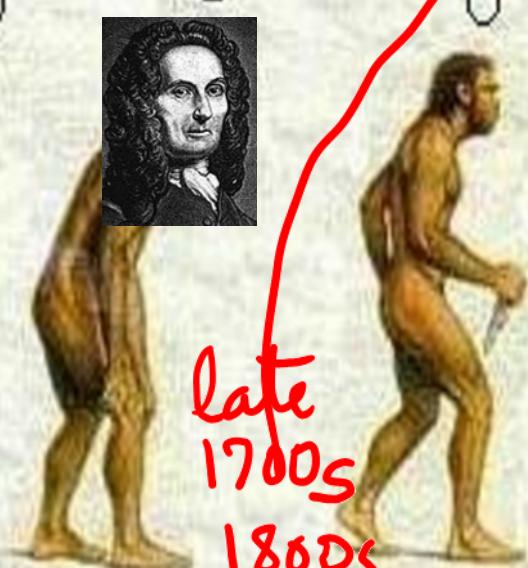
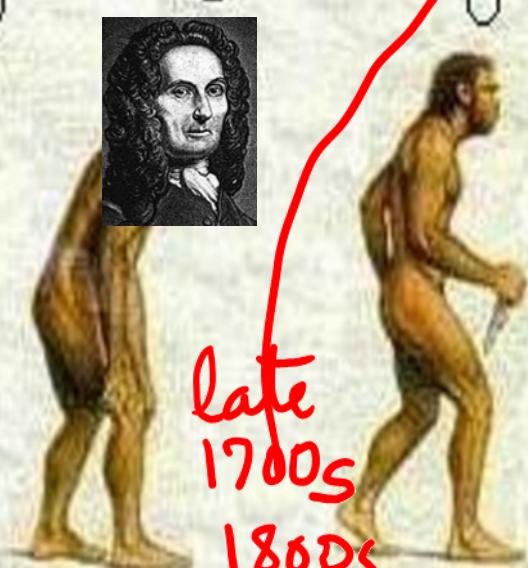
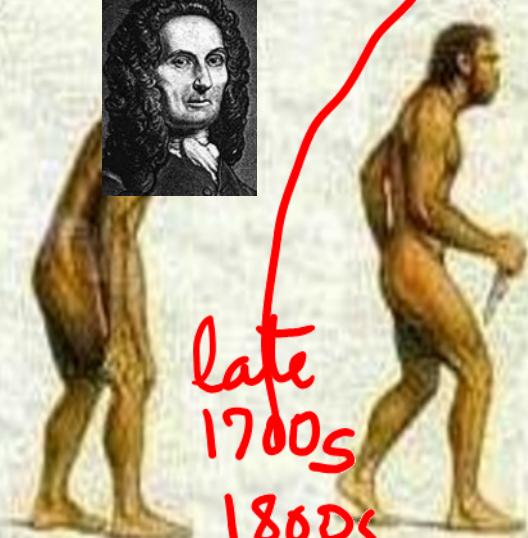
$P(\theta|x)$ using uniform $P(\theta)$



Cicero



early 1700s
de Moivre



Controversial

Uniform
 $\theta \in (0,1)$

not same as

Uniform of
 $\omega \in (0, \infty)$

odds $w \in (0, \infty)$

HOMO
APRIORIUS

HOMO
PRAGMATICUS

HOMO
FREQUENTISTUS

HOMO
SAPIENS

HOMO
BAYESIANUS

- which is not even a probability

(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...

Bayes
Pascal

$P(\theta|x)$ using uniform $P(\theta)$??



Cicero

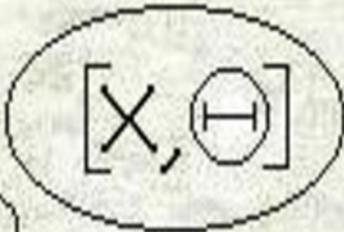


early 1700s
de Moivre



1920
Fisher — p-values

Neyman — decision theory
Pearson — hypothesis testing



HOMO
APRIORIUS



HOMO
PRAGMATICUS

late
1700s
1800s

HOMO
FREQUENTISTUS



HOMO
SAPIENS



HOMO
BAYESIANIS

(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...

Bayes

Pascal

$P(\theta|x)$ using uniform $P(\theta)$



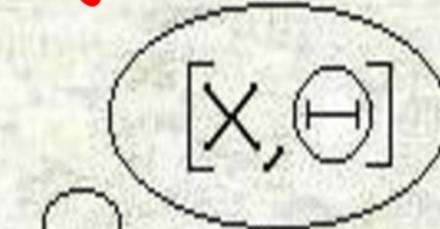
Cicero



early 1700s
de Moivre



1920
Fisher
Neyman
Pearson



1940-60
Fisher - fiducial
Fraser - structural



HOMO
APRIORIUS



HOMO
PRAGMATICUS



HOMO
FREQUENTISTUS



HOMO
SAPIENS



HOMO
BAYESIANUS

late
1700s
1800s

(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...

Bayes
Pascal

$P(\theta|x)$ using uniform $P(\theta)$



Cicero

early 1700s
de Moivre



late
1700s
1800s



HOMO
APRIORIUS



HOMO
PRAGMATICUS

HOMO
FREQUENTISTUS



Fisher
Neyman
Pearson



1940-60
Fisher
Fraser



MCMC
1950s+

[X|θ]

[X, θ]

[θ|X]

(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...

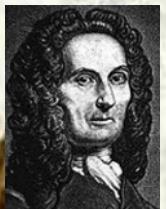
Bayes
Pascal

$P(\theta|x)$ using uniform $P(\theta)$



Cicero

early 1700s
de Moivre



late
1700s
1800s

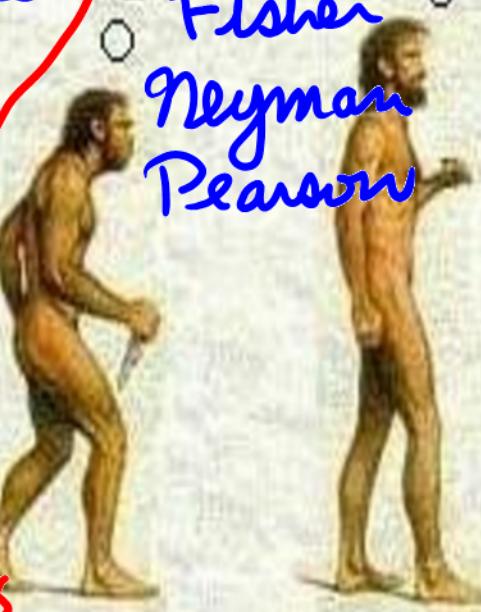


HOMO
APRIORIUS



HOMO
PRAGMATICUS

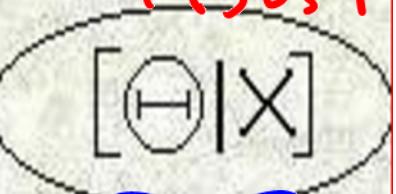
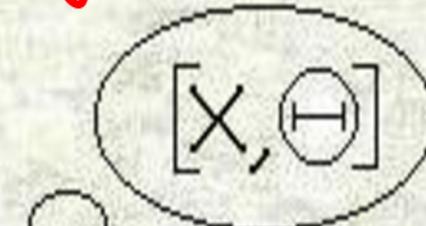
HOMO
FREQUENTISTUS



HOMO
FREQUENTISTUS



HOMO
BAYESIANUS



(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...

Bayes

Pascal

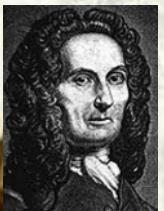
$P(\theta|x)$ using uniform $P(\theta)$



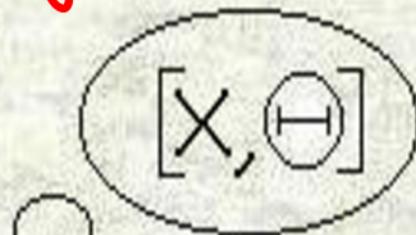
Cicero



early 1700s
de Moivre



1920
Fisher
Neyman
Pearson



1940-60
Fisher
Fraser



$\Theta|X$



HMC

1990s+

Bayesian
Renaissance



HOMO
APRIORIUS



HOMO
PRAGMATICUS

HOMO
FREQUENTISTUS



HOMO

FREQUENTISTUS



HOMO
SAPIENS

HOMO
BAYESIANIS

late
1700s
1800s

Philosophical Basic problem

Philosophical
Basic problem

Given a model $P(x|\theta)$

Philosophical Basic problem

Given a model $P(x|\theta)$

To get $P(\theta|x)$

you need to be willing
to specify $P(\theta)$

Philosophical Basic problem

Given a model $P(X|\theta)$

To get $P(\theta|X)$

you need to be willing
to specify $P(\theta)$

$$\text{Then } P(X,\theta) = P(X|\theta)P(\theta)$$

$$\text{and } P(\theta|X) = \frac{P(X,\theta)}{P(X)}$$

Philosophical Basis problem

Given a model $P(X|\theta)$ ^{Model}

To get $P(\theta|X)$

you need to be willing

to specify $P(\theta)$ ^{Prior}

$$\text{Then } P(X,\theta) = P(X|\theta)P(\theta)$$

$$\text{and } P(\theta|X) = \frac{P(X,\theta)}{P(X)}$$

^{Posterior}

Philosophical Basis problem

Given a model $P(X|\theta)$ *(model)*

To get $P(\theta|X)$

you need to be willing

to specify $P(\theta)$ *(prior)*

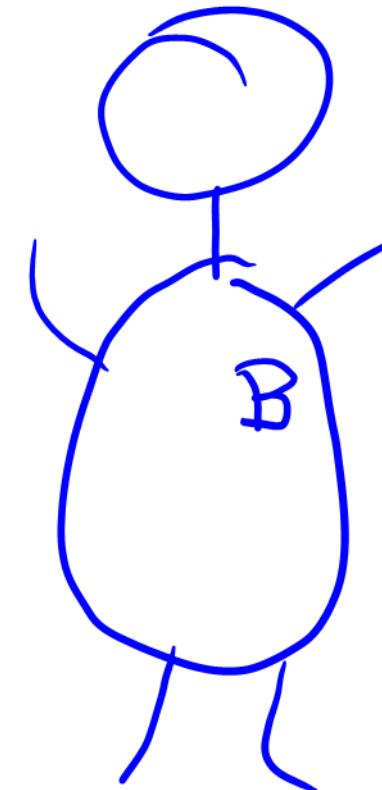
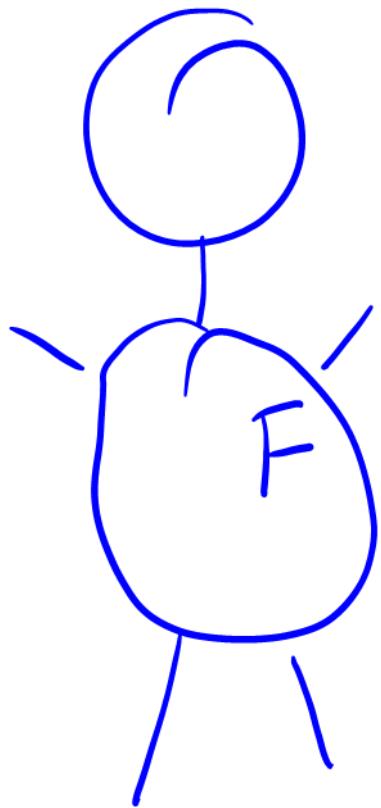
$$\text{Then } P(X,\theta) = P(X|\theta)P(\theta)$$

$$\text{and } P(\theta|X) = \frac{P(X,\theta)}{P(X)}$$

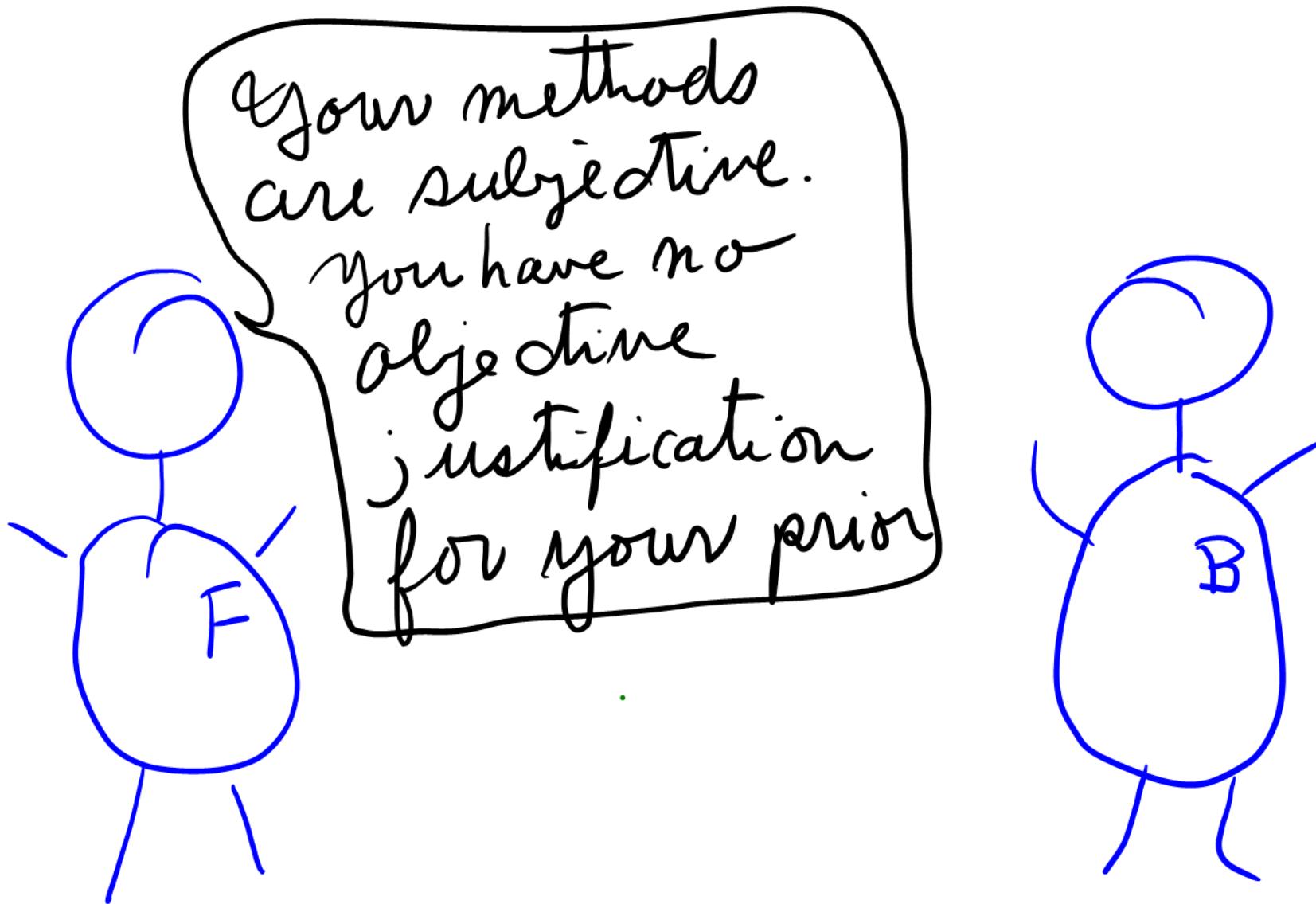
- You need a prior to get a posterior.
- Can we justify a particular prior?

Frequentists only use $P(X|\theta)$
and don't need $P(\theta)$

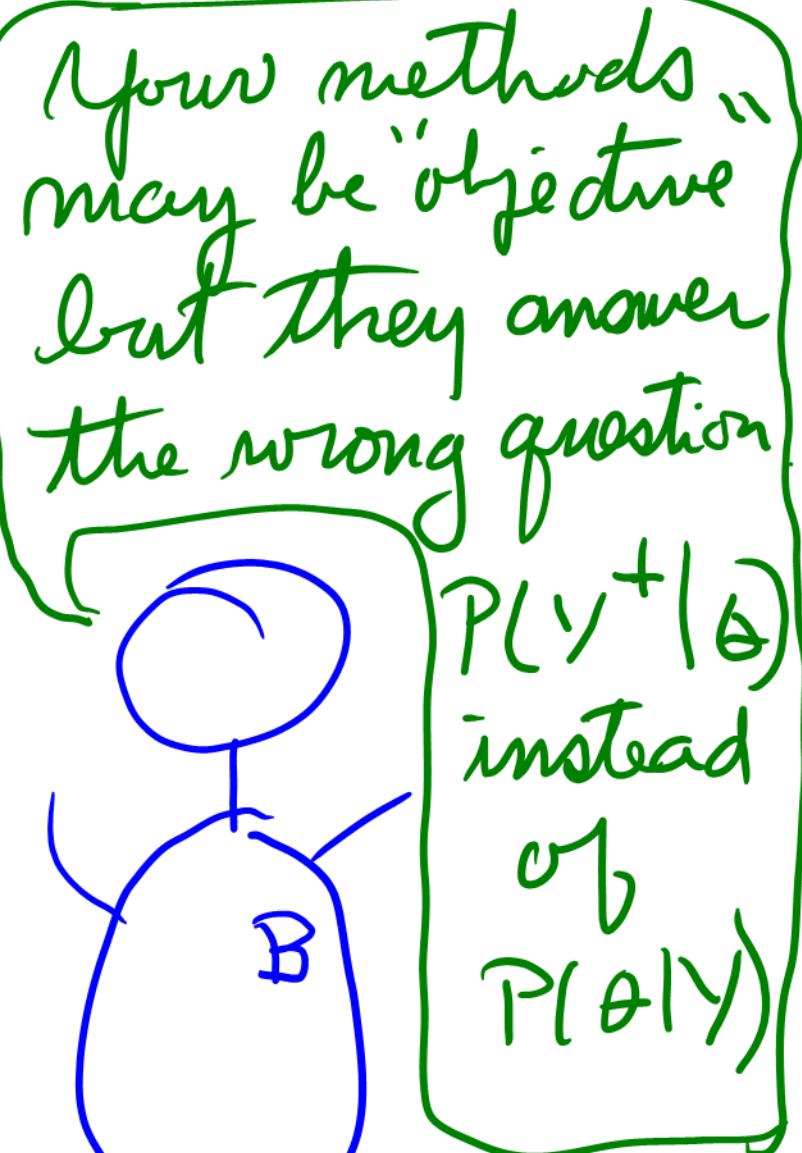
Frequentists only use $P(X|\theta)$
and don't need $P(\theta)$



Frequentists only use $P(X|\theta)$
and don't need $P(\theta)$



Frequentists only use $P(X|\theta)$
and don't need $P(\theta)$



Frequentists only use $P(X|\theta)$
and don't need $P(\theta)$

Your methods
are subjective.
you have no
objective
justification
for your prior



Your methods „
may be ‘objective’
but they answer
the wrong question



$P(Y^+|\theta)$
instead
of
 $P(\theta|Y)$

Practical problem:

$$p(x, \theta) = p(x | \theta) p(\theta)$$

Practical problem:

$$P(X, \theta) = P(X|\theta)P(\theta)$$

$$P(\theta|X) = \frac{P(X, \theta)}{P(X)}$$

Practical problem:

$$P(X, \theta) = P(X|\theta)P(\theta)$$

$$P(\theta|X) = \frac{P(X, \theta)}{P(X)}$$

$$\int_{\theta \in \mathbb{R}^{\text{huge}}} P(X, \theta) d\theta$$

Practical problem:

$$P(X, \theta) = P(X|\theta)P(\theta)$$

$$P(\theta|X) = \frac{P(X, \theta)}{P(X)}$$

$$\int P(X, \theta) d\theta$$

If θ has high dimension
this becomes easily impossible.

Practical problem:

$$P(X, \theta) = P(X|\theta)P(\theta)$$

$$P(\theta|X) = \frac{P(X, \theta)}{P(X)}$$

MCMC (mid 20th C.)

comes to the rescue;

It's possible to sample from
 $P(\theta|X)$ knowing only $P(X, \theta)$

Posteriors without priors?

Fisher - Fiducial inference

Fraser - Structural inference

Objective Bayesian inference

Baking the Bayesian omelette
without breaking the
Bayesian egg.

Emerging practice:

Using weakly informative
priors.

Posteriors without priors?

Fisher - Fiducial inference

Fraser - Structural inference

Objective Bayesian inference

Baking the Bayesian omelette
without breaking the
Bayesian egg.

Emerging practice:

Using weakly informative
priors.

The reason why
thinking of
a 95% CI for
a linear regression
parameter as
a 95% "probability"
interval is harmless.

Posteriors without priors?

Fisher - Fiducial inference

Fraser - Structural inference

Objective Bayesian inference

Baking the Bayesian omelette
without breaking the
Bayesian egg.

Emerging practice:

Using weakly informative
priors.

The reason why thinking of a 95% CI for a linear regression parameter as a 95% "probability" interval is harmless.

But don't take it for granted
Remember Sally Clark

Markov Chain Monte Carlo

use $P(\theta, x) = P(x|\theta)P(\theta)$

Markov Chain Monte Carlo

use $P(\theta, x) = P(x|\theta)P(\theta)$

joint model \times prior

Samples from $P(\theta|x)$ using only $P(\theta, x)$
i.e. no need to find elusive $P(x)$

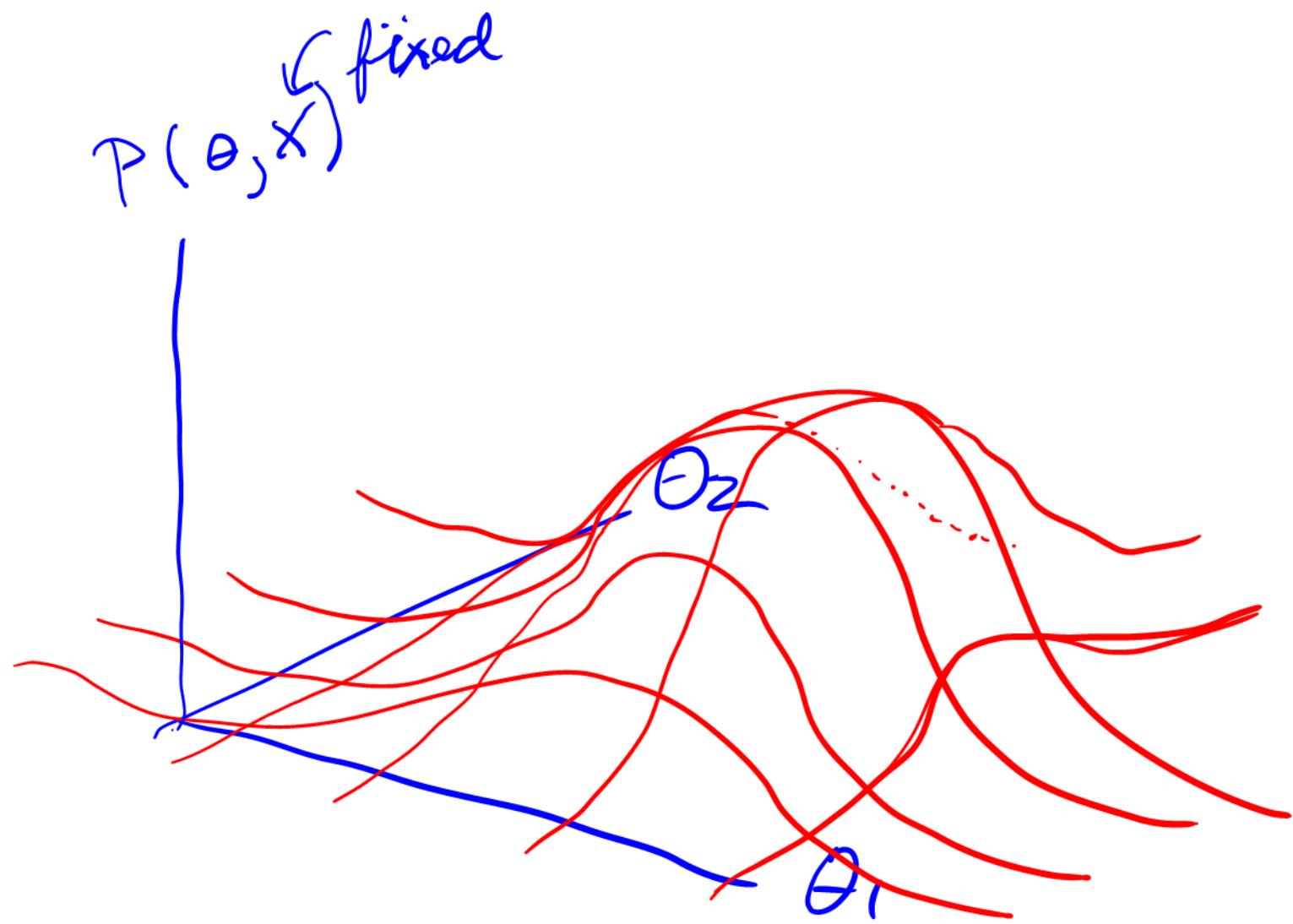
Markov Chain Monte Carlo

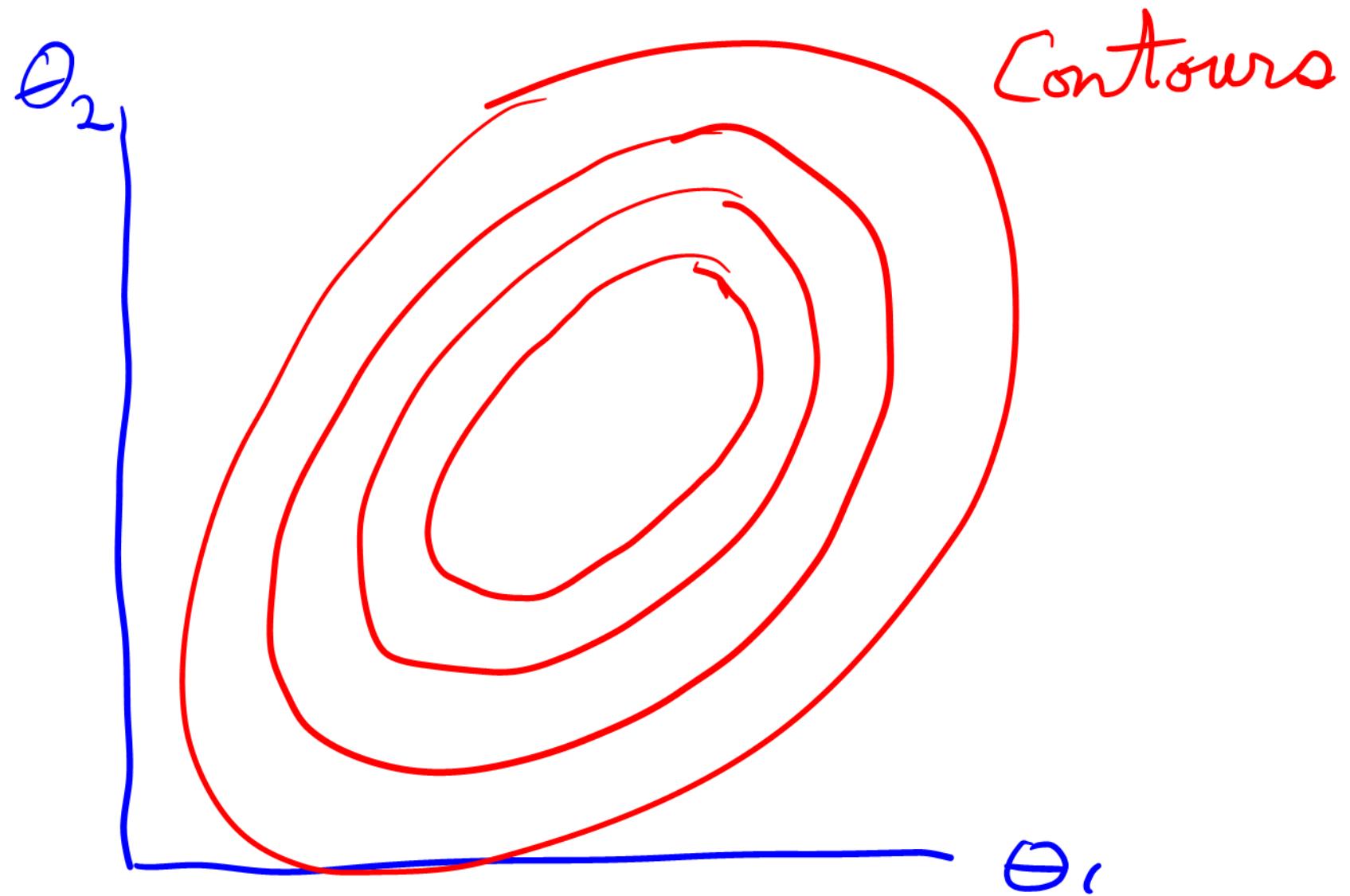
use $P(\theta, x) = P(x|\theta)P(\theta)$

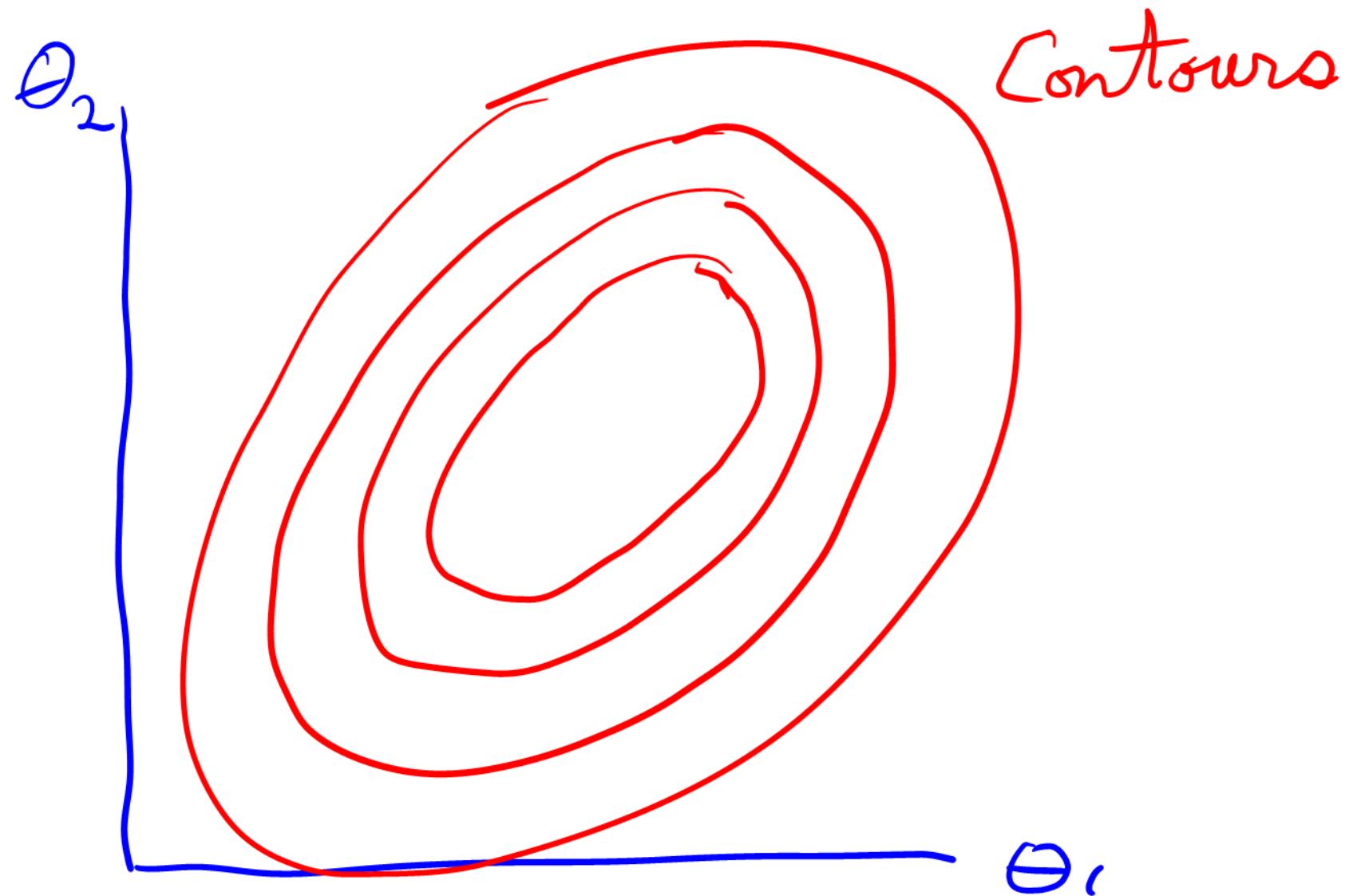
joint model \times prior

Samples from $P(\theta|x)$ using only $P(\theta, x)$
i.e. no need to find elusive $P(x)$

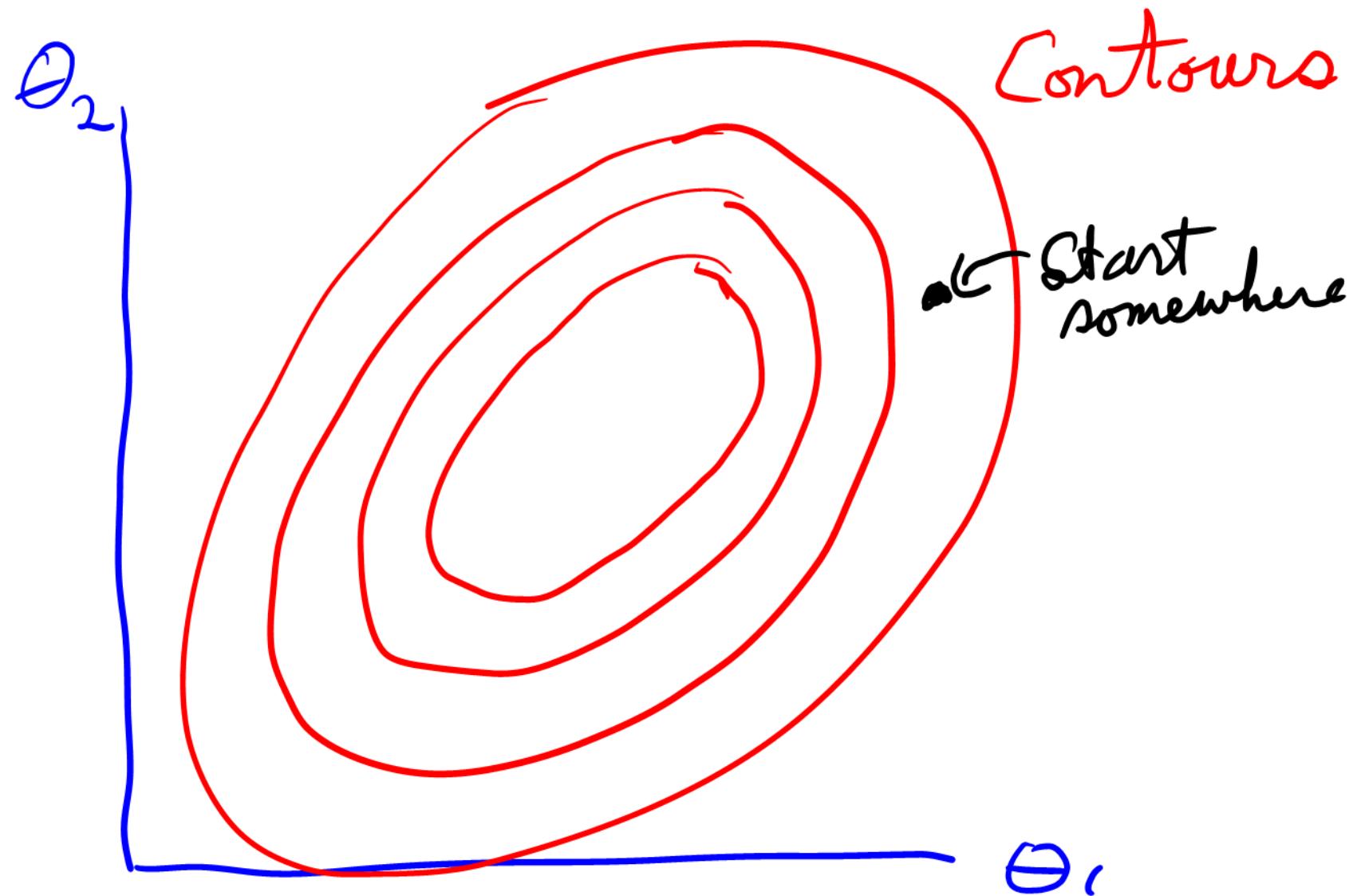
With X fixed, think of $P(\theta, x)$
as defining a mountain over θ space



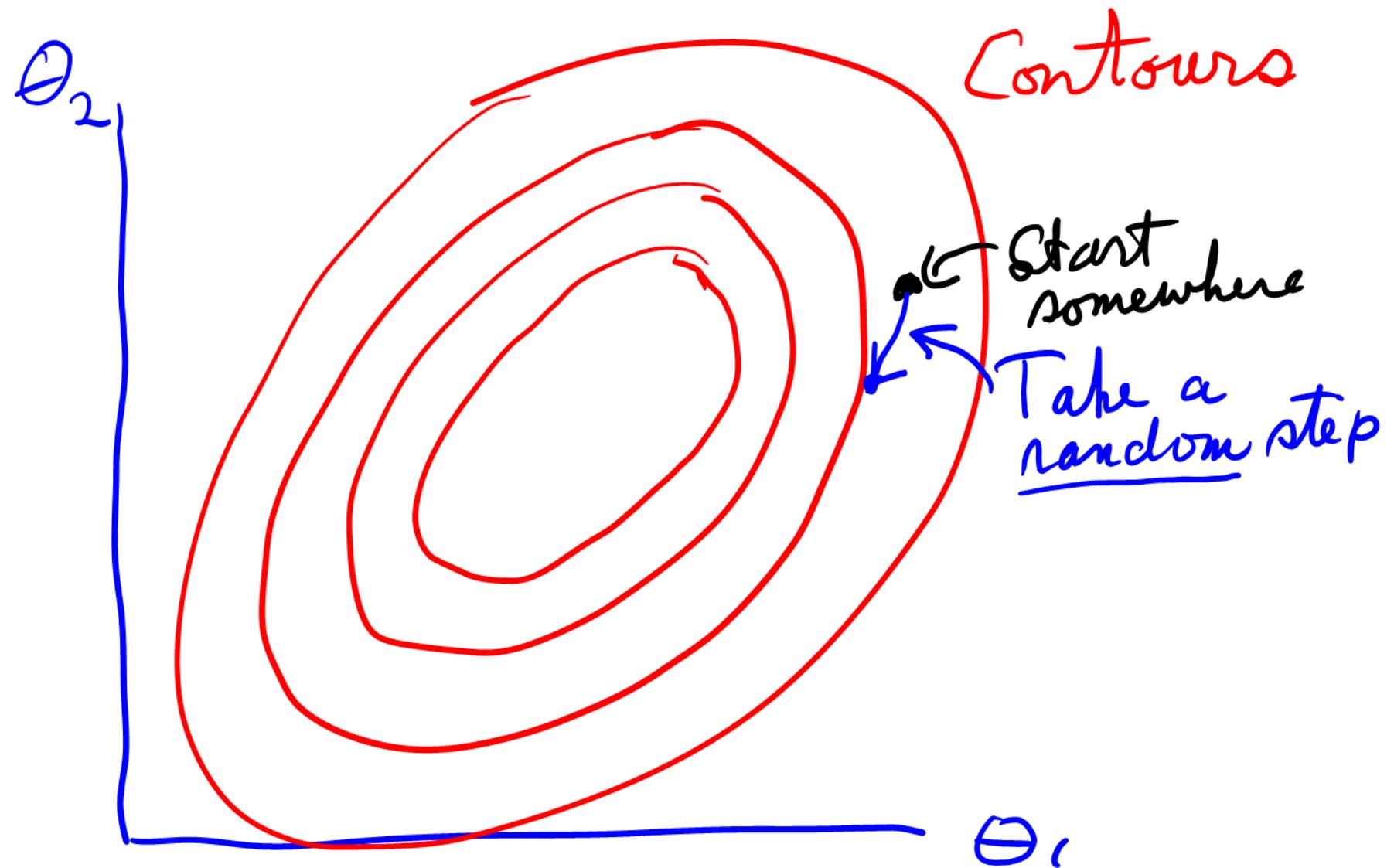




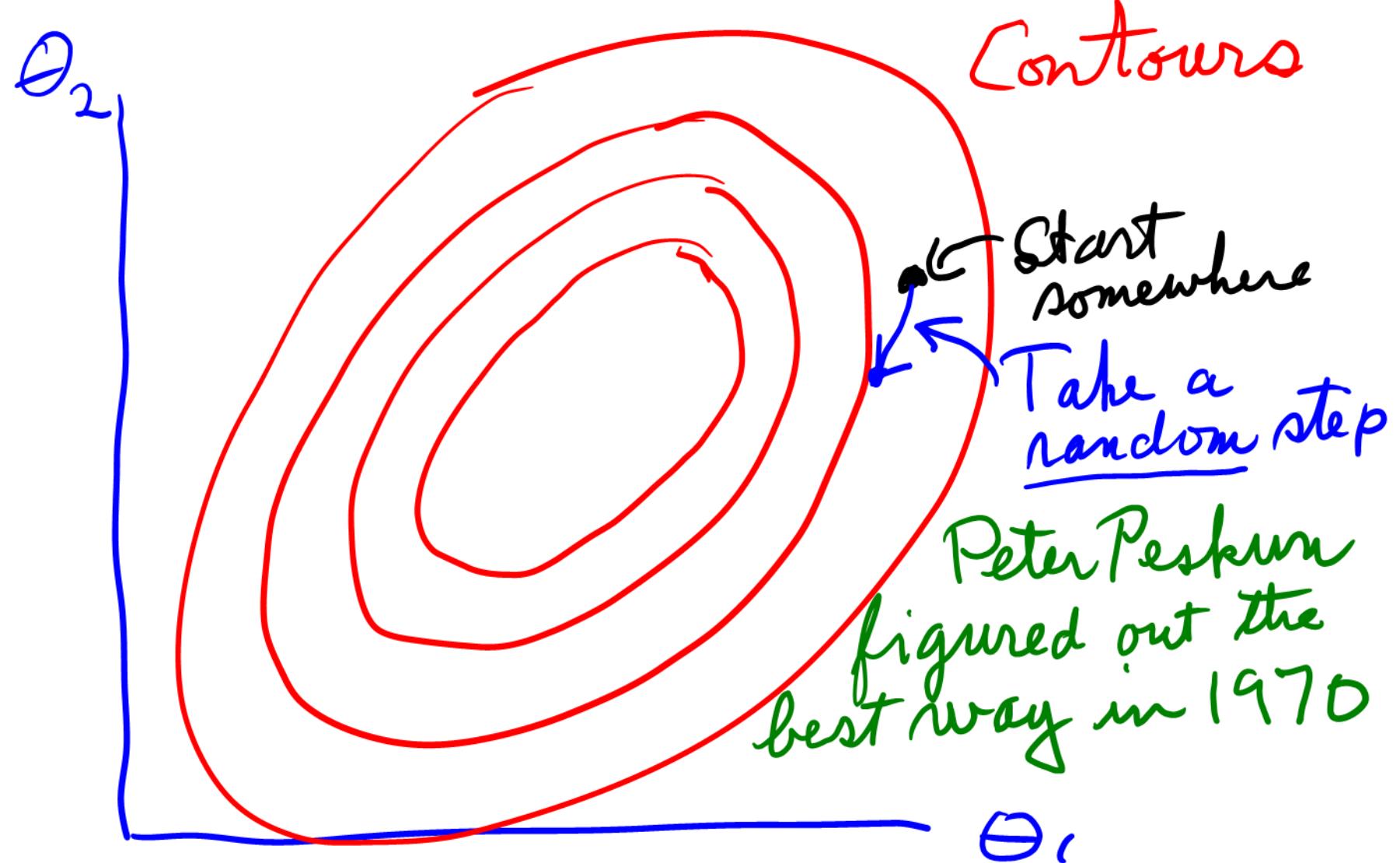
Metropolis-Hastings algorithm



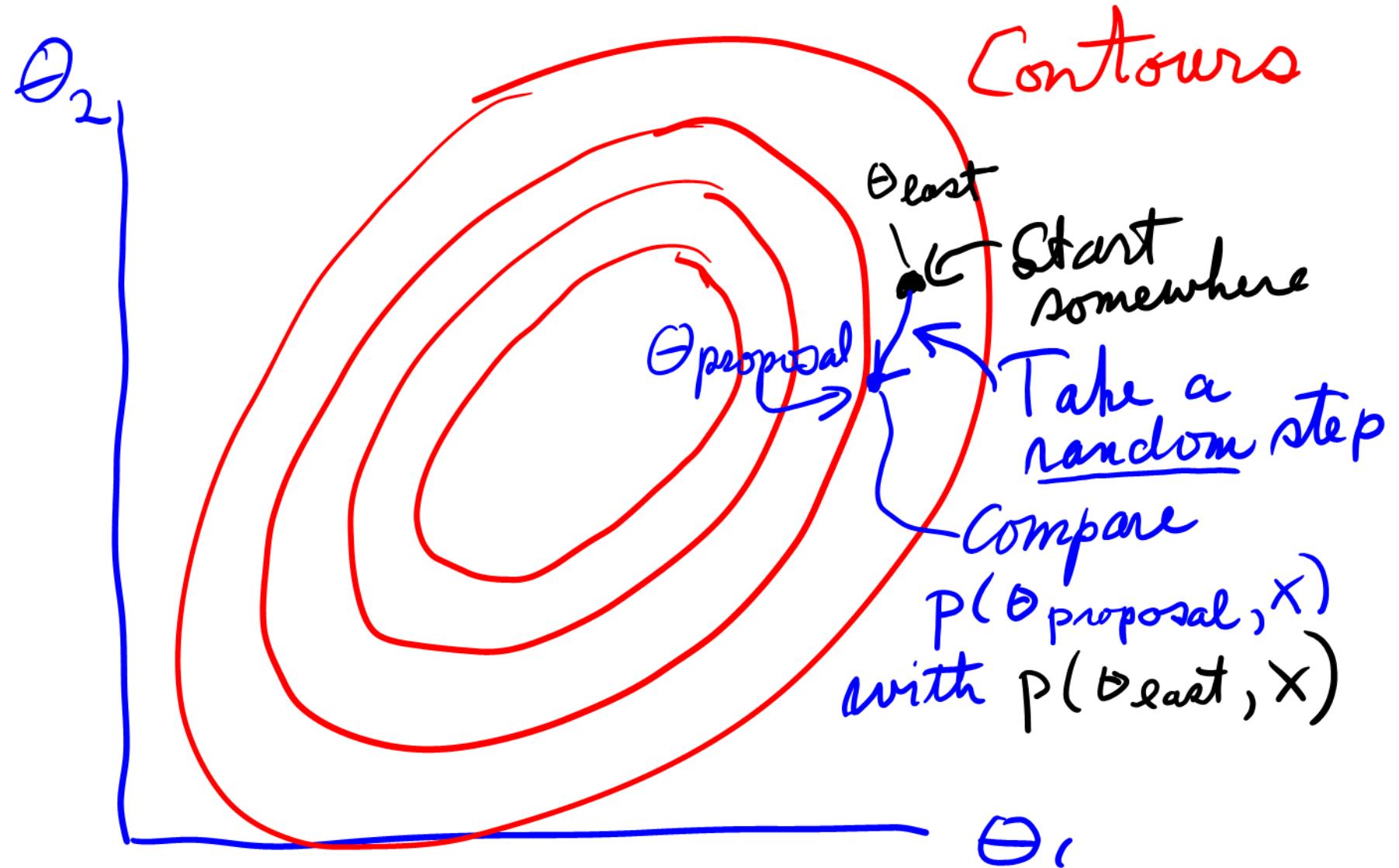
Metropolis-Hastings algorithm



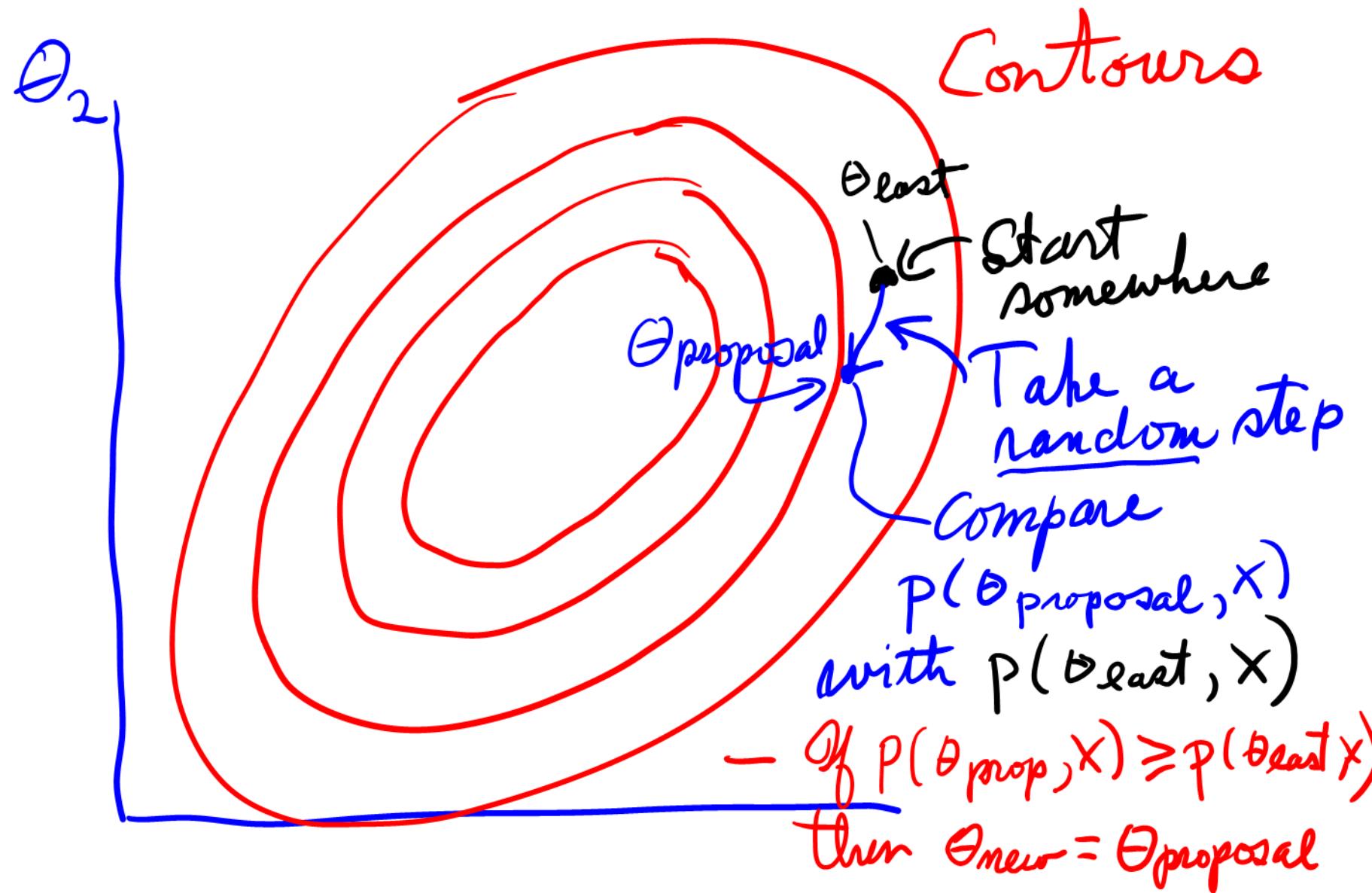
Metropolis-Hastings algorithm

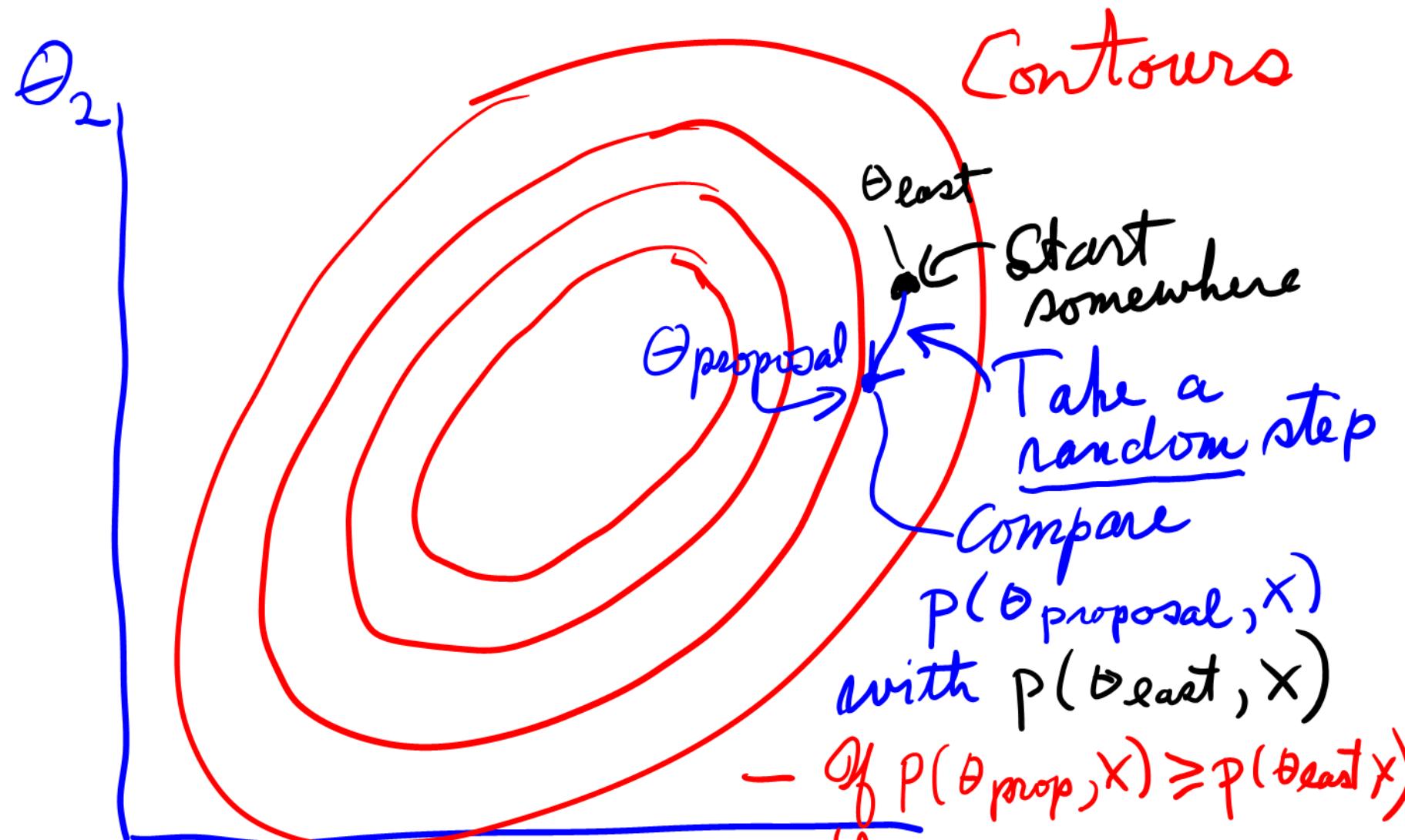


Metropolis-Hastings algorithm

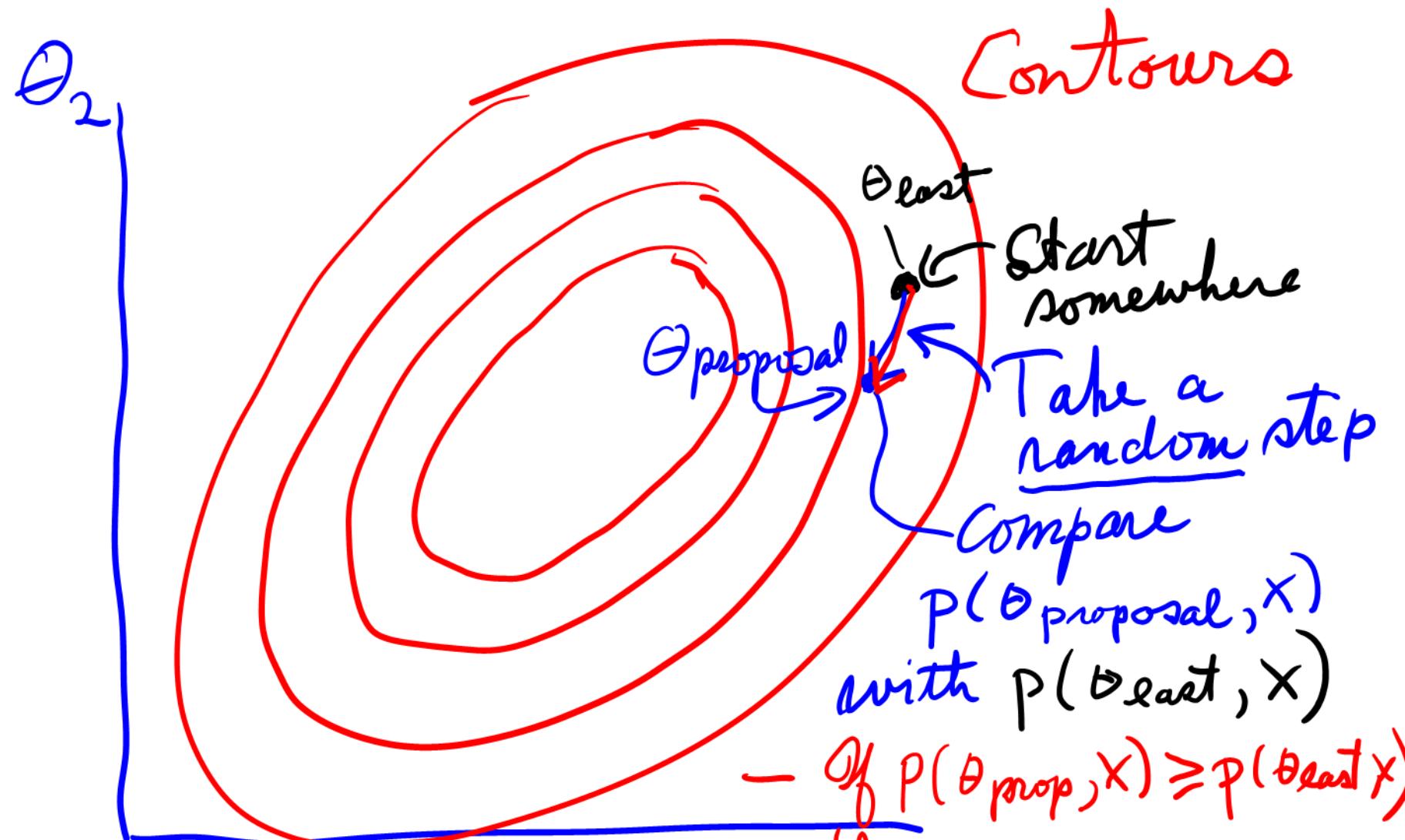


Metropolis-Hastings algorithm

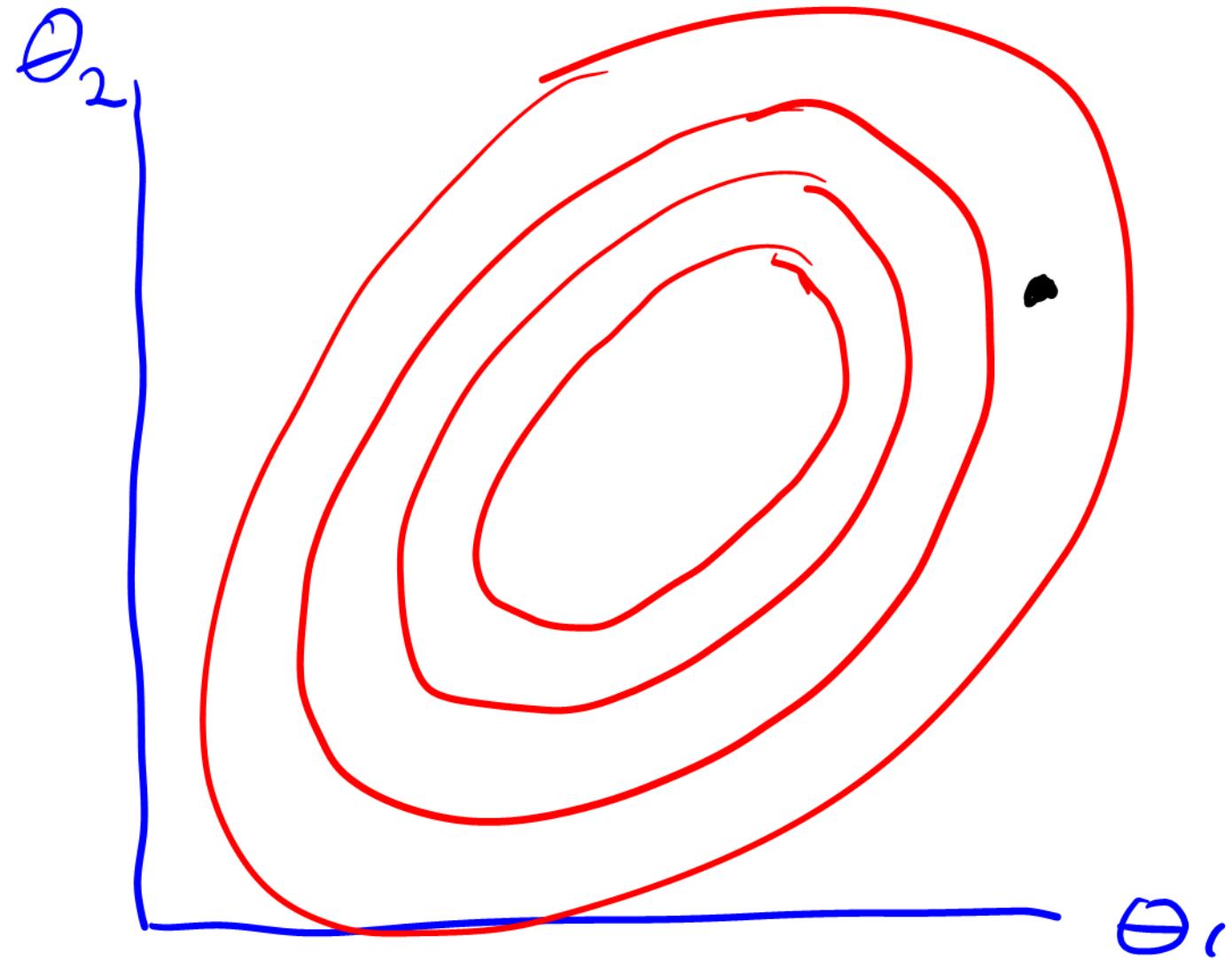


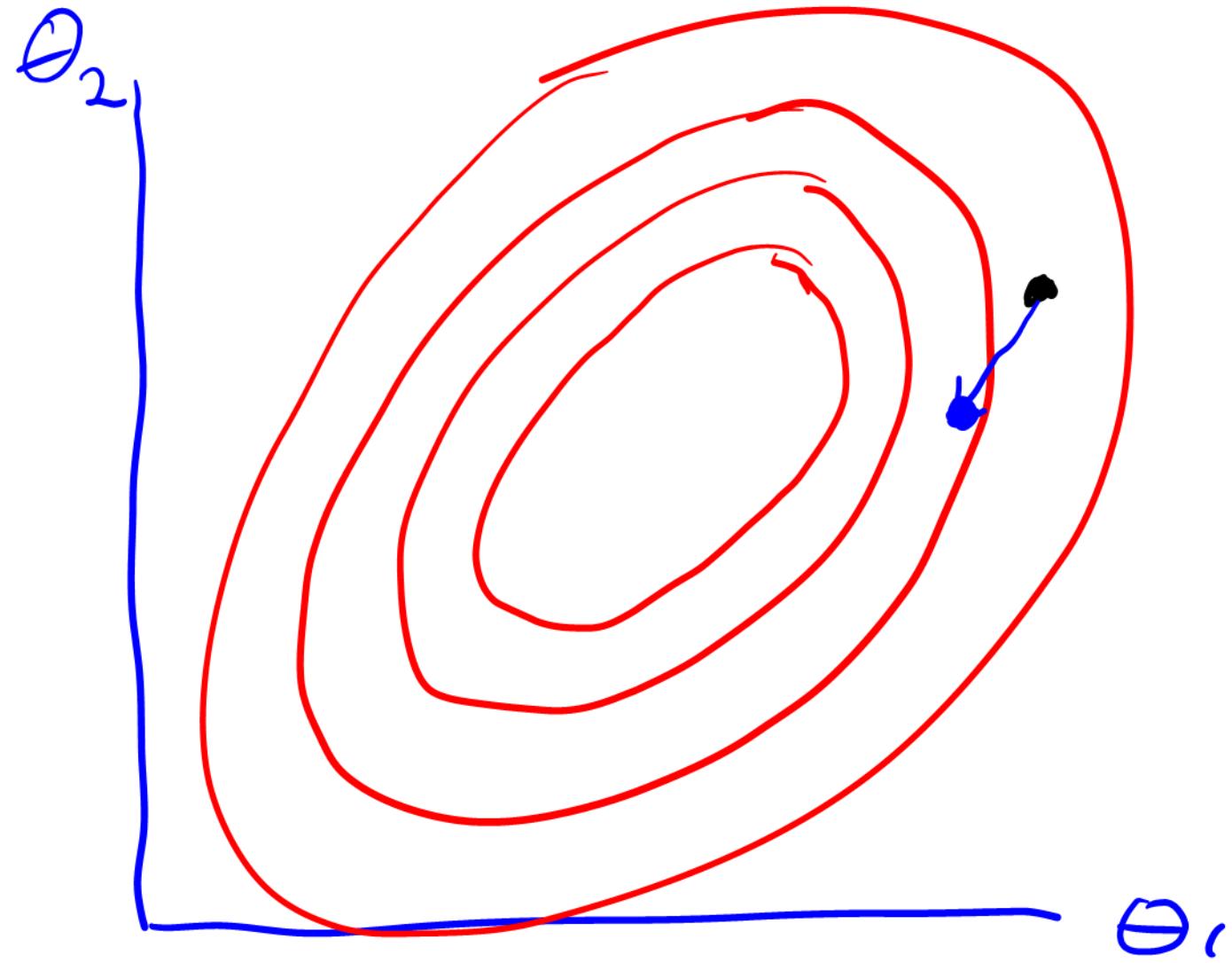


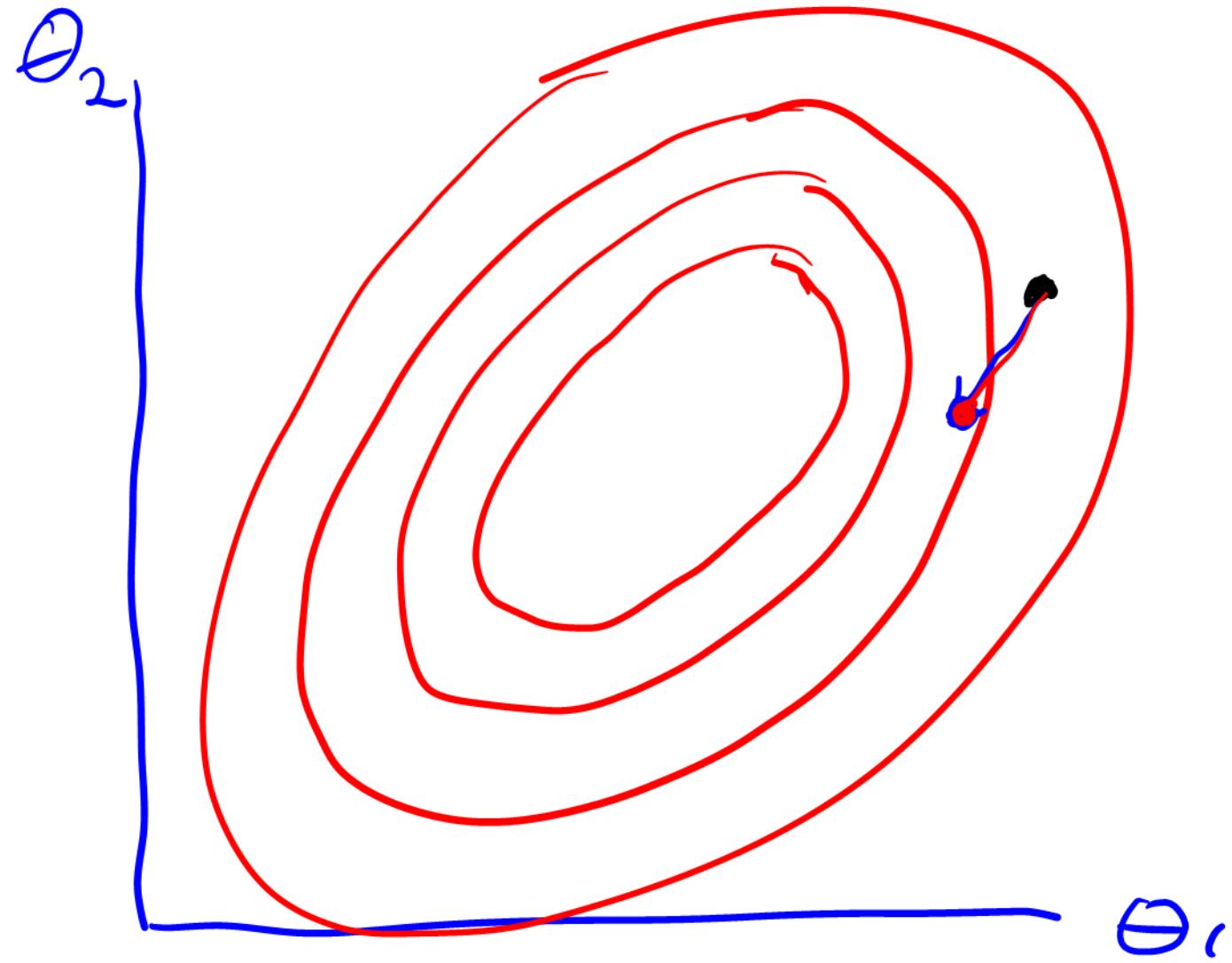
- Otherwise, i.e. if you have moved down in probability, set $\theta_{new} = \theta_{proposal}$ with prob $\frac{p(\theta_{prop}, x)}{p(\theta_{last}, x)}$ else $\theta_{new} = \theta_{last}$

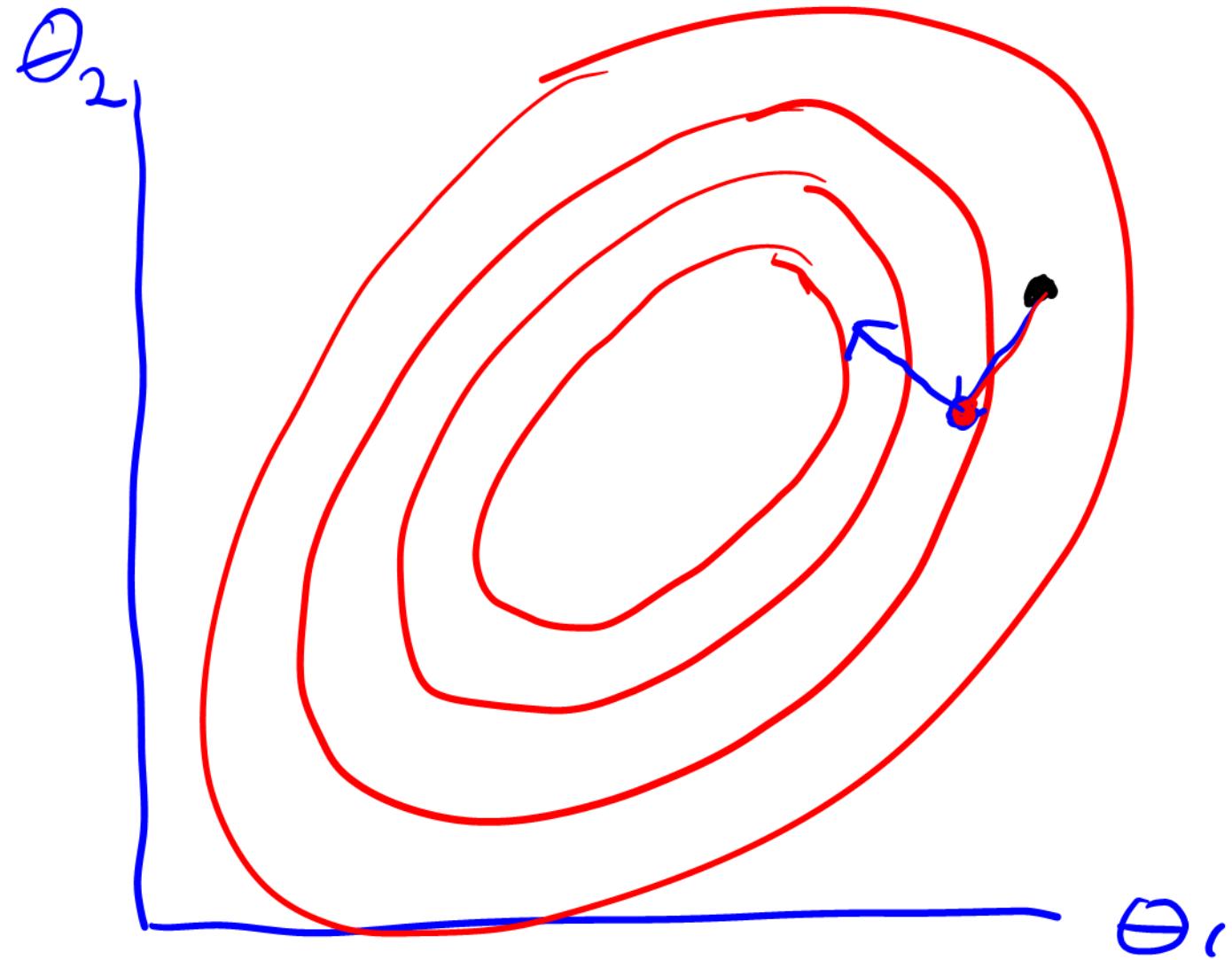


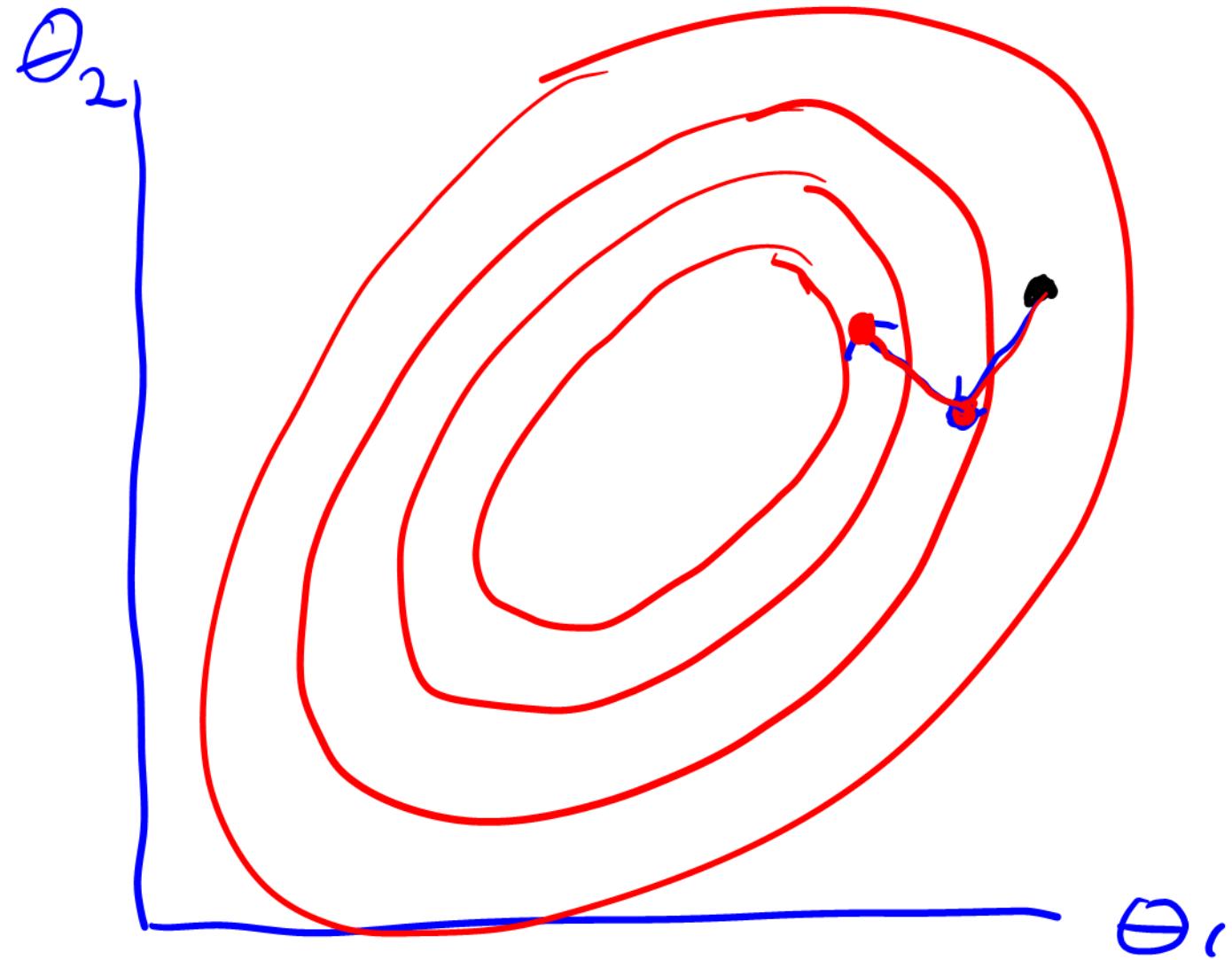
- Otherwise, i.e. if you have moved down in probability, set $\theta_{new} = \theta_{proposal}$ with prob $\frac{p(\theta_{prop}, x)}{p(\theta_{last}, x)}$ else $\theta_{new} = \theta_{last}$

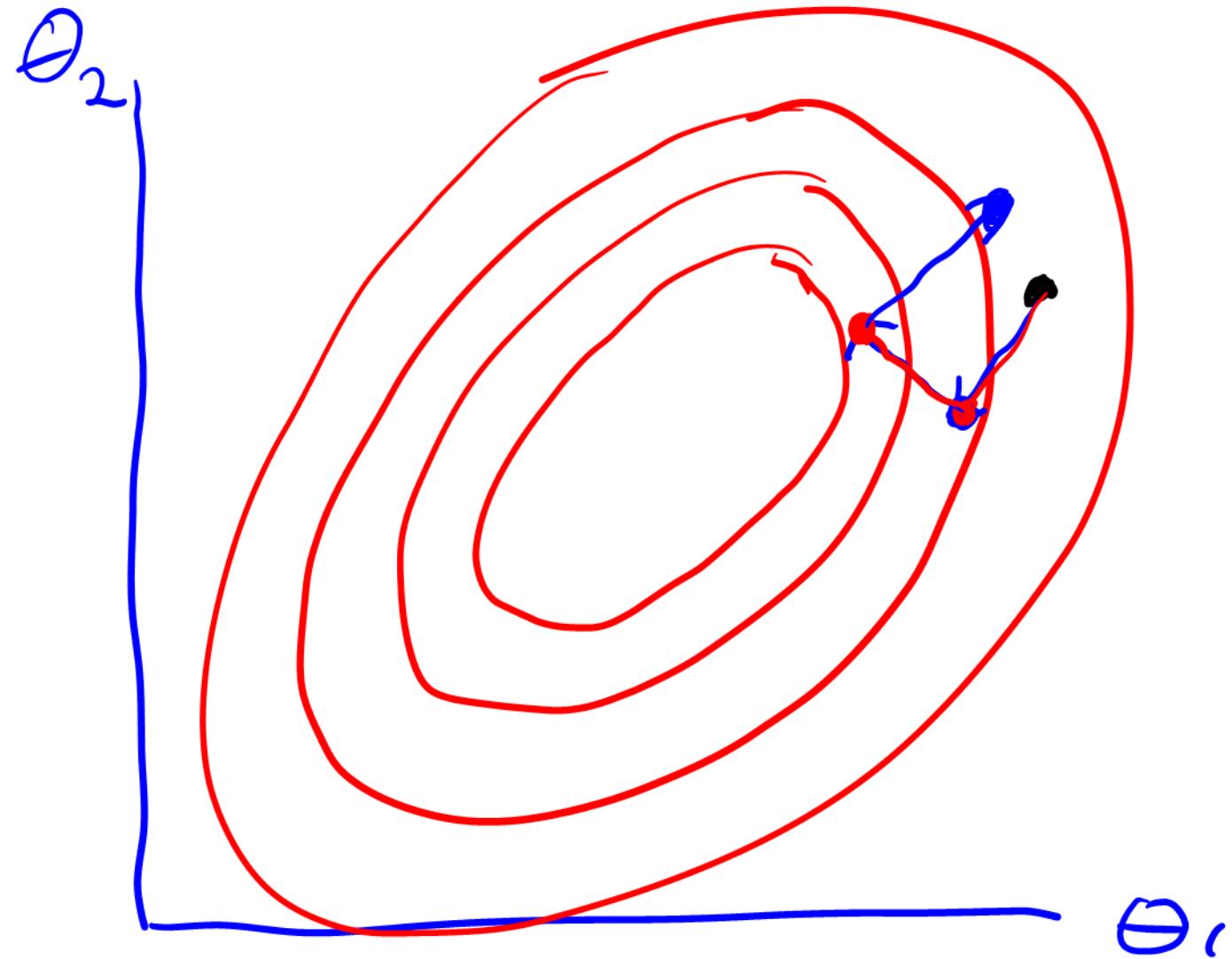




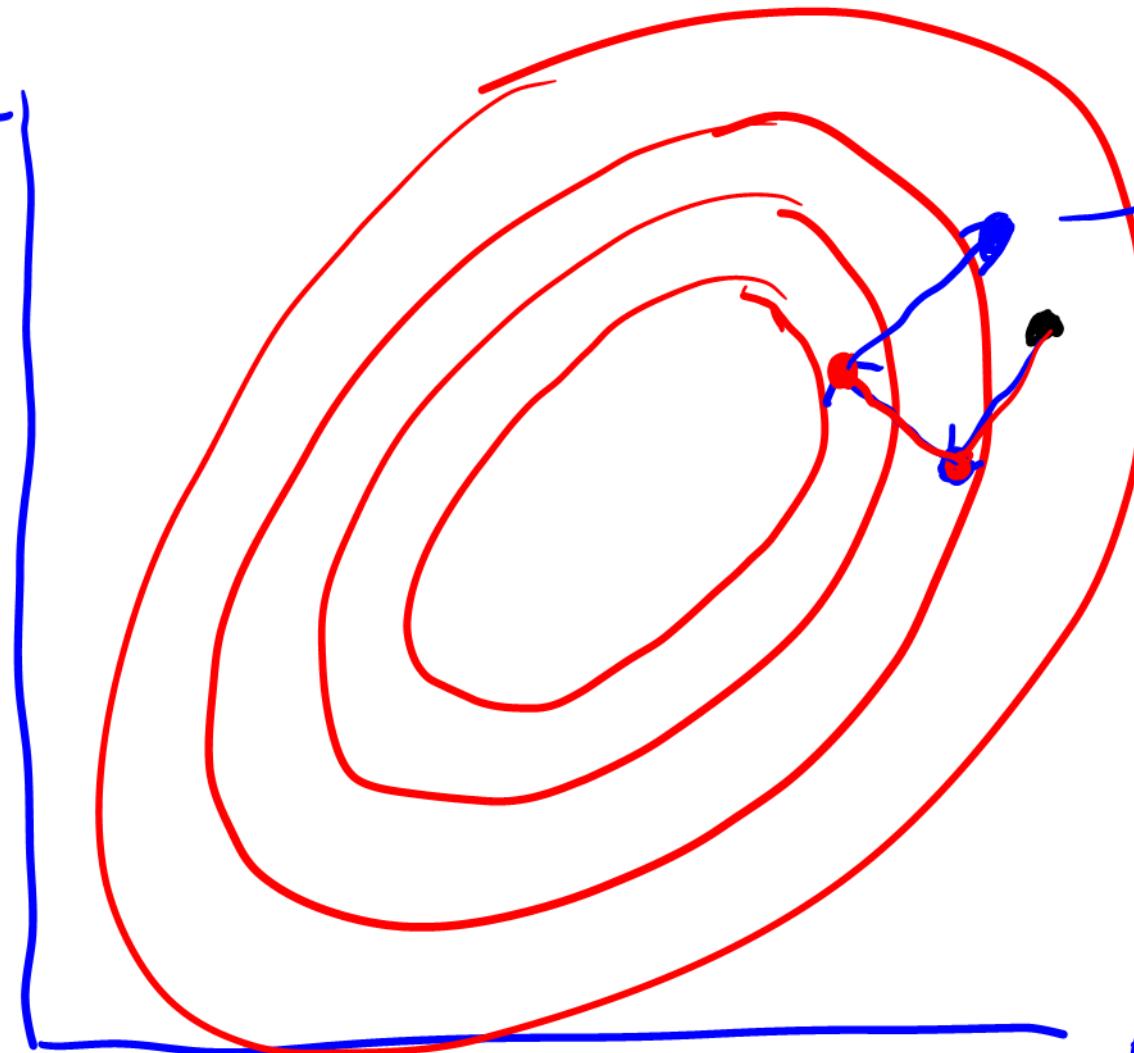








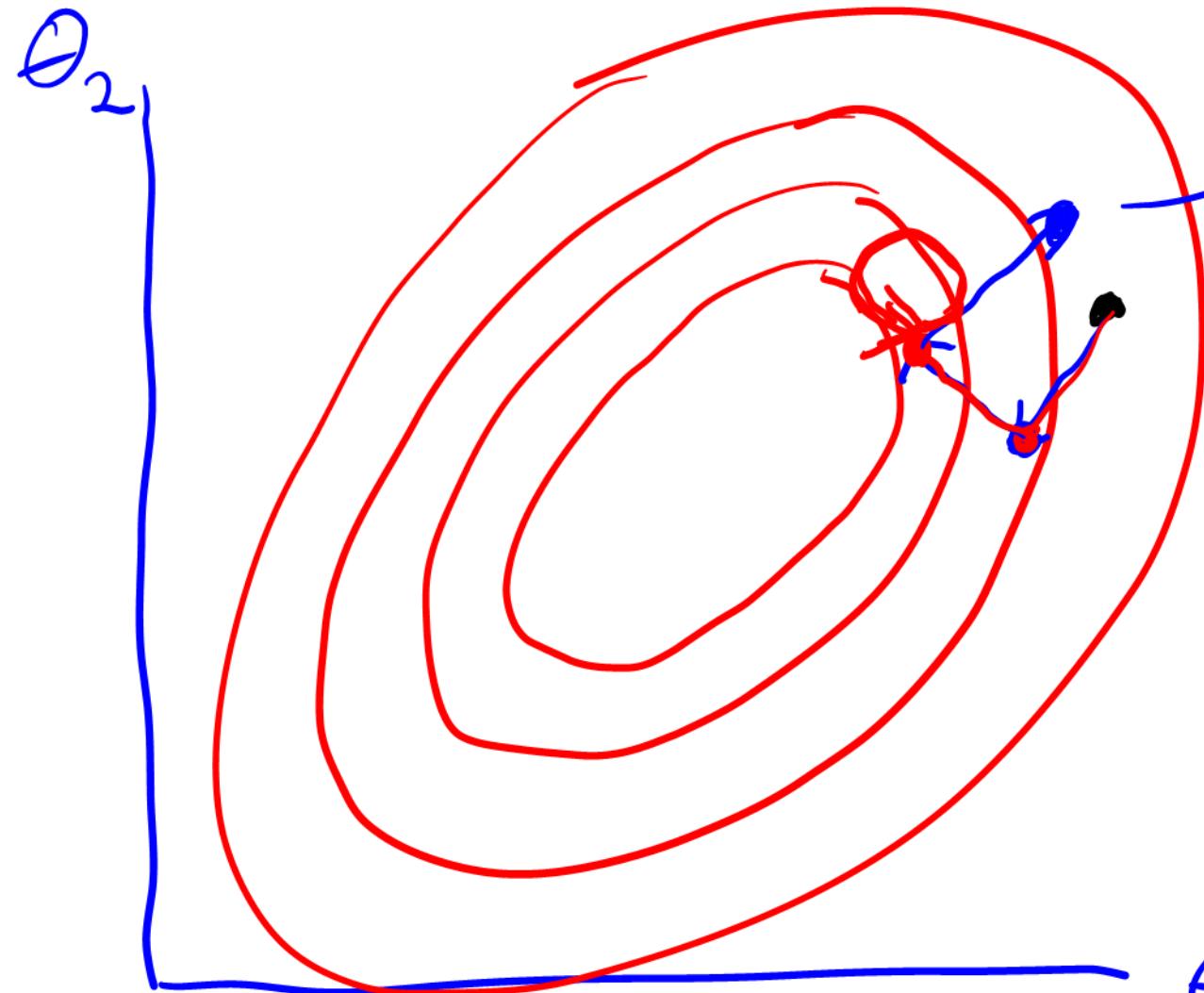
θ_2



toss a biased coin with

$$p(H) = \frac{P(\cdot)}{P(\cdot)}$$

If Heads move
if Tails stay

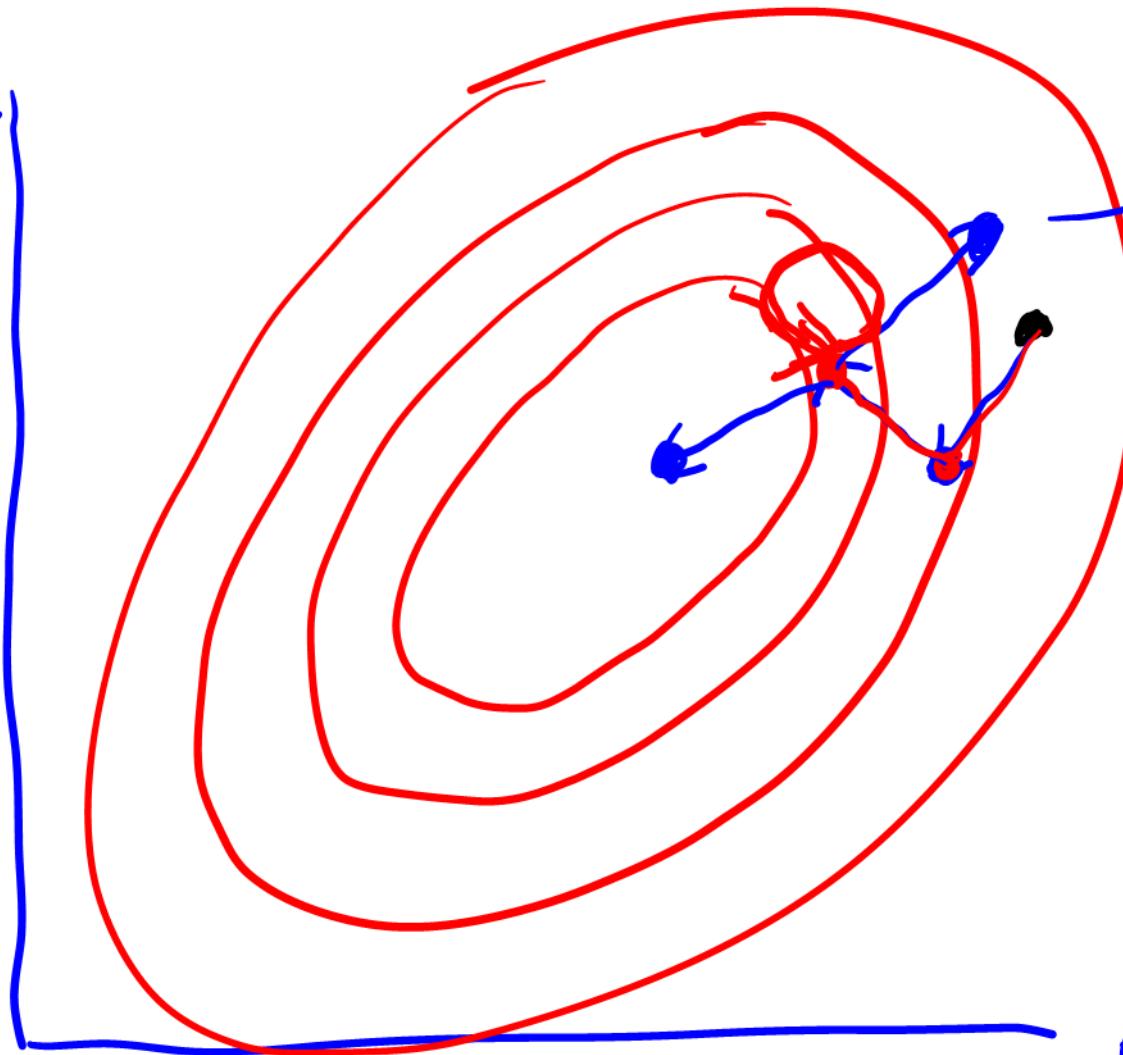


toss a biased coin with

$$p(H) = \frac{P(\cdot)}{P(\cdot)}$$

If Heads move
if Tails stay

θ_2

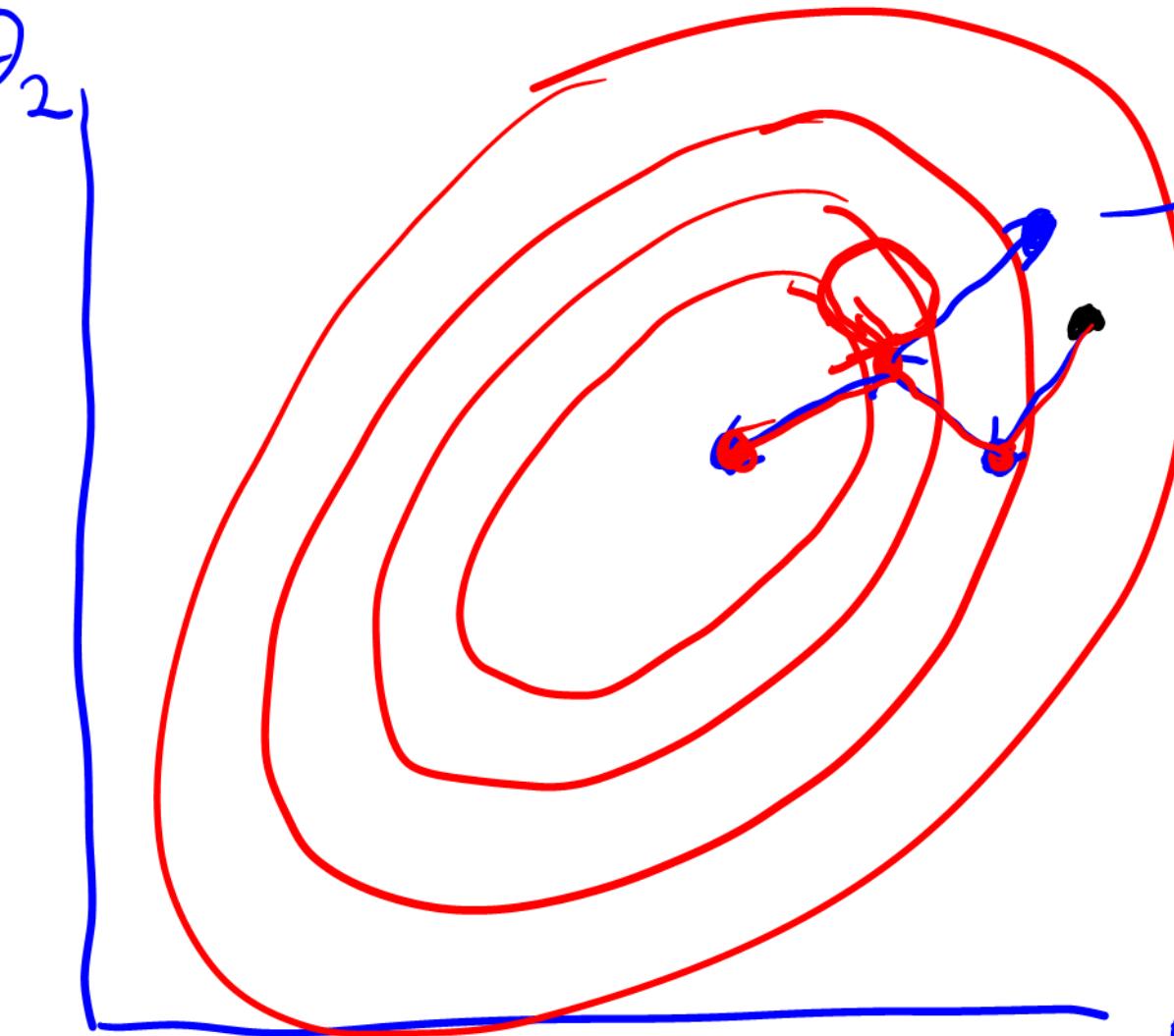


toss a
biased
coin
with

$$p(H) = \frac{P(\cdot)}{P(\cdot)}$$

If Heads
move
if Tails
stay

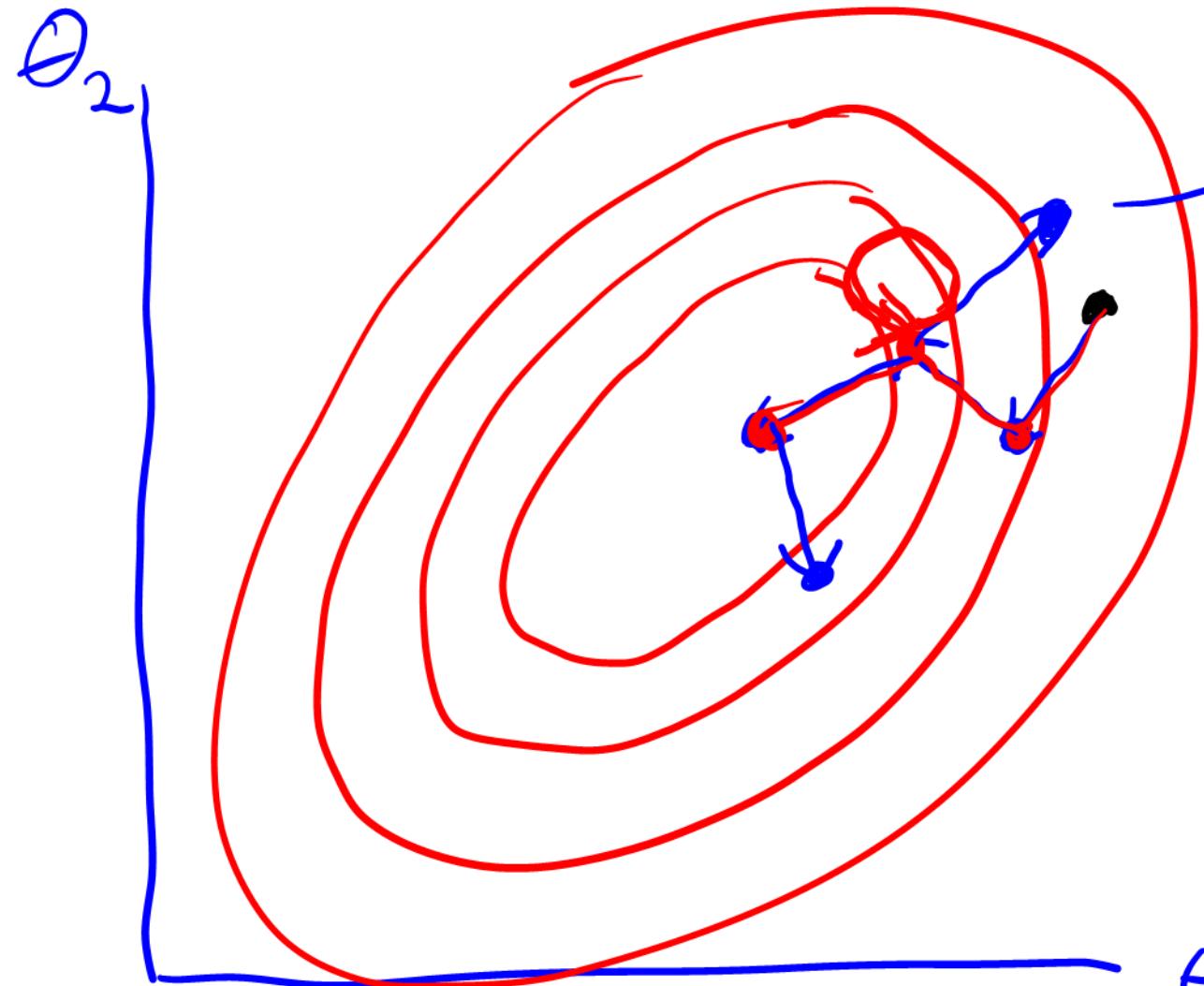
θ_2



toss a
biased
coin
with

$$p(H) = \frac{P(\cdot)}{P(\cdot)}$$

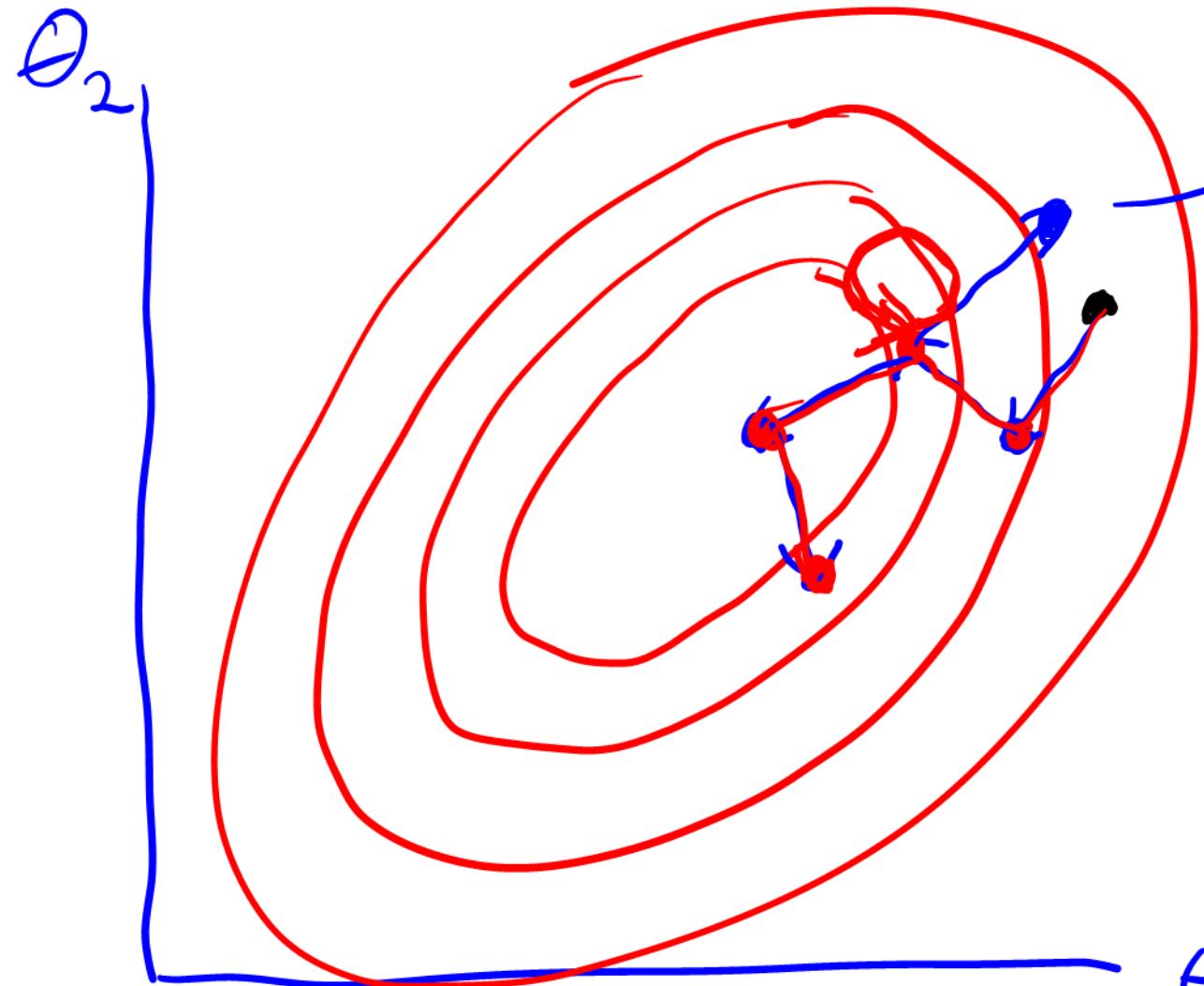
If Heads
move
if Tails
stay



toss a
biased
coin
with

$$p(H) = \frac{P(\cdot)}{P(\cdot)}$$

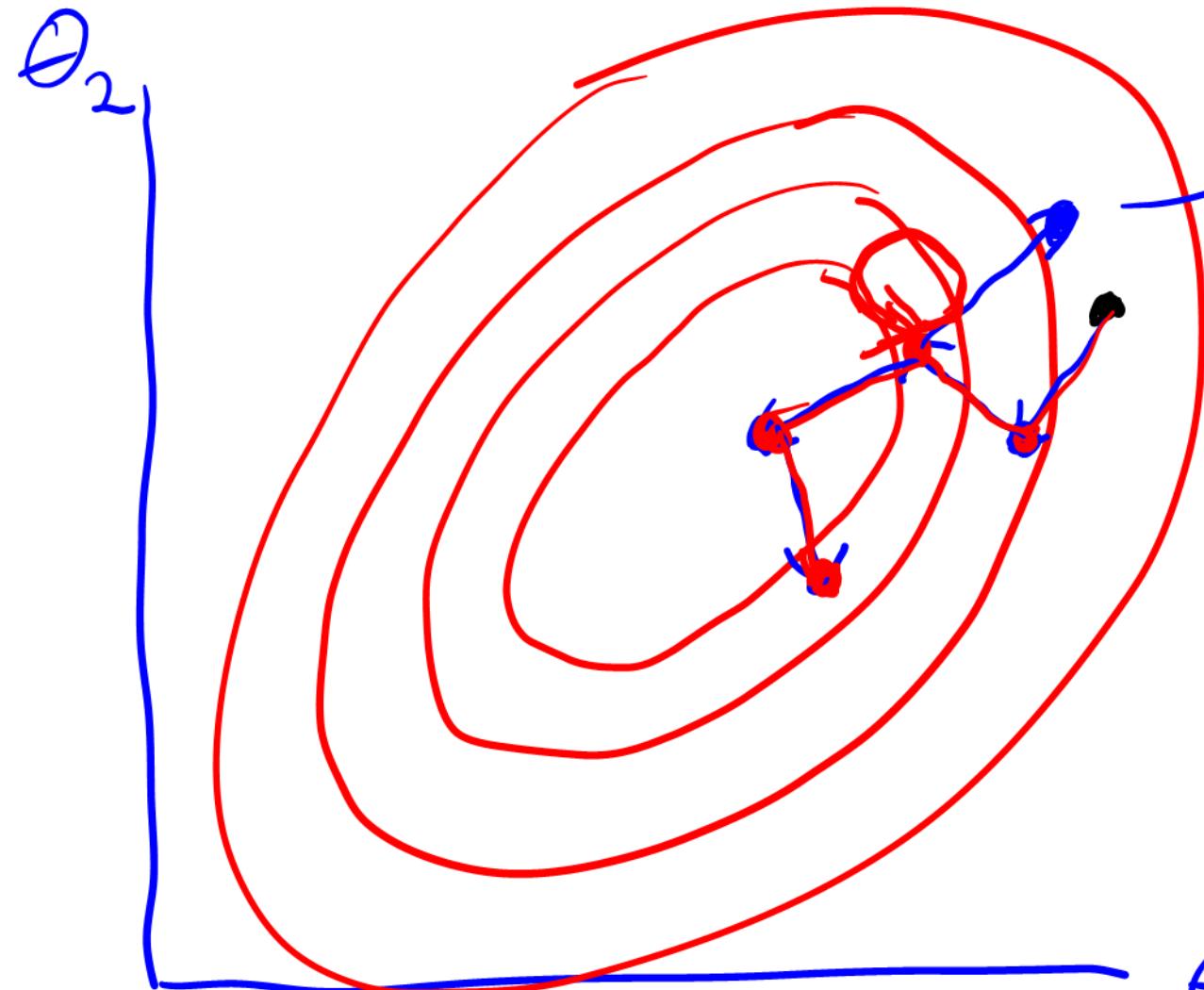
If Heads
move
if Tails
stay



toss a biased coin with

$$p(H) = \frac{P(\cdot)}{P(\cdot)}$$

If Heads move
if Tails stay

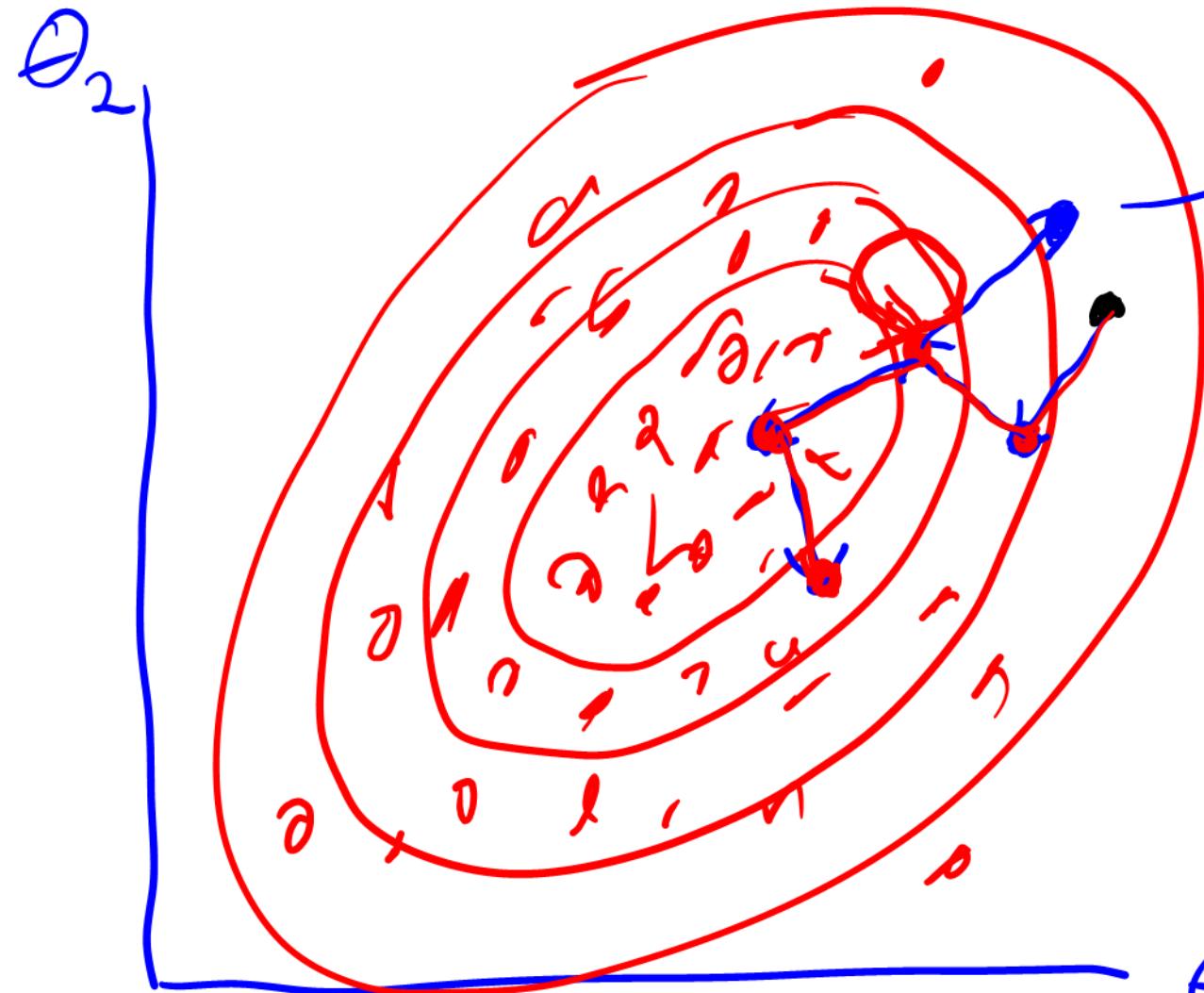


toss a biased coin with

$$p(H) = \frac{P(\cdot)}{P(\cdot)}$$

If Heads move
if Tails stay

Keep doing this for a very long time

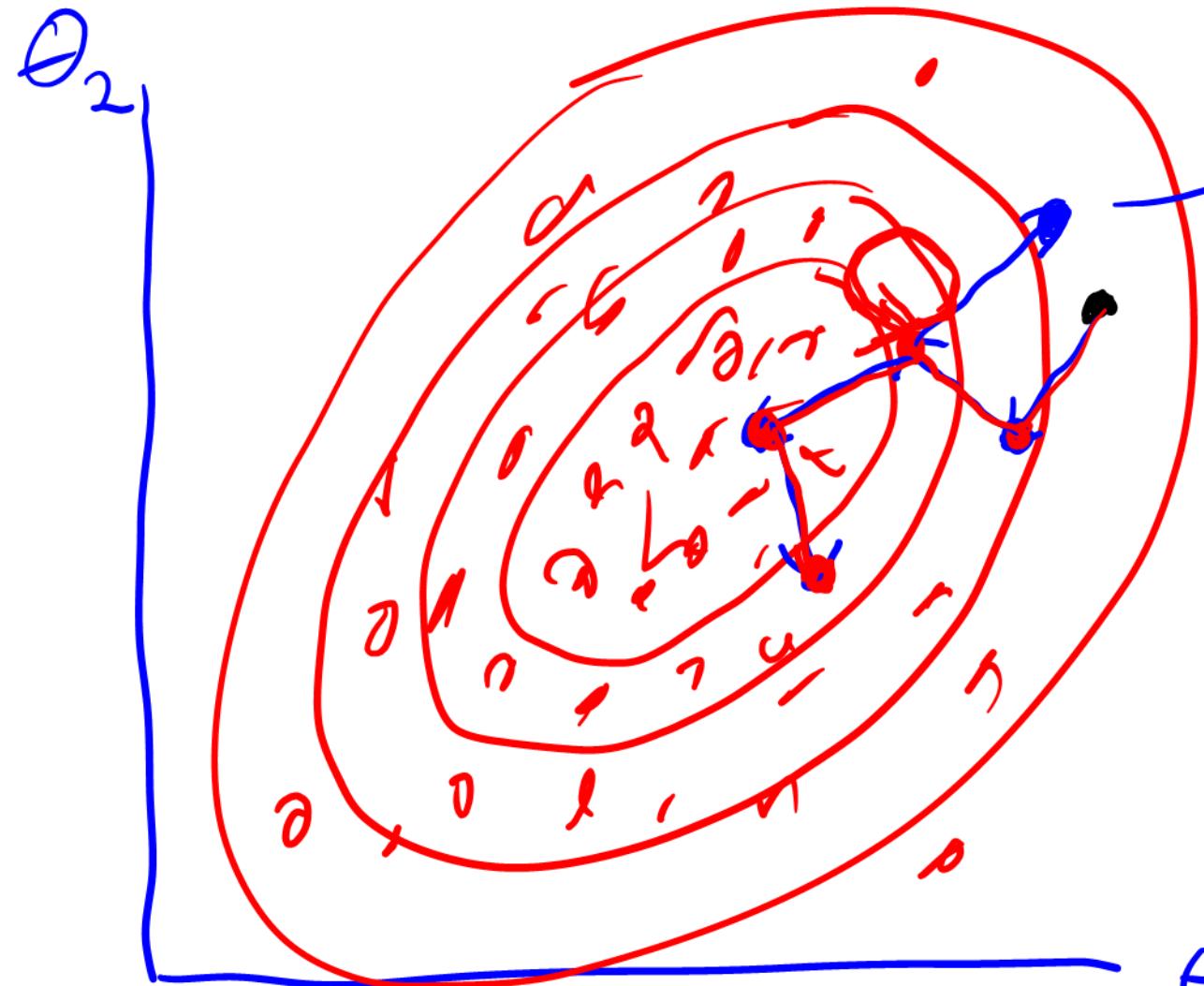


toss a biased coin with

$$p(H) = \frac{P(\cdot)}{P(\cdot)}$$

If Heads move
if Tails stay

Keep doing this for a very long time



toss a biased coin with

$$p(H) = \frac{P(\cdot)}{P(\cdot)}$$

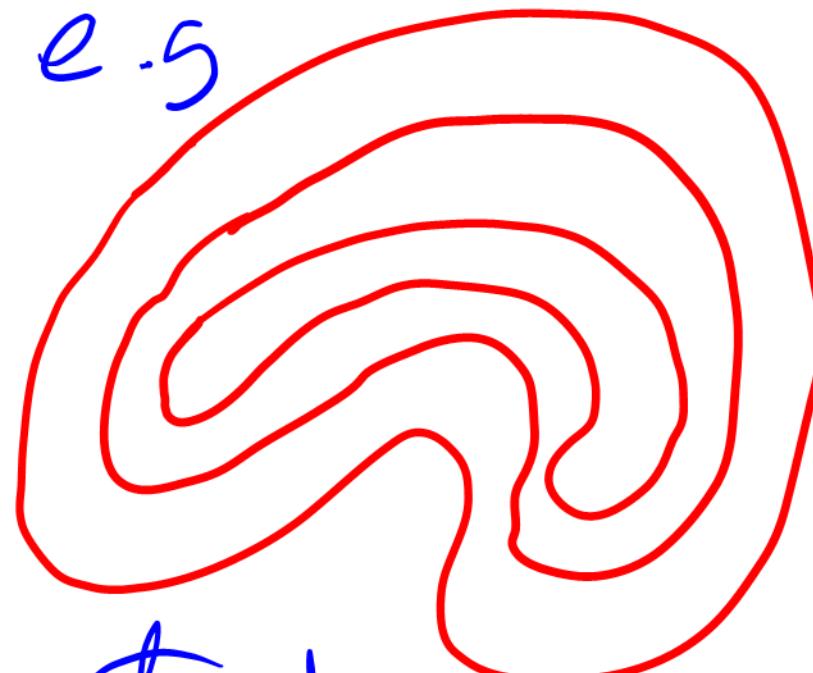
If Heads move
if Tails stay

Generates a sample from $P(\theta|X)$

That's the XX-H algorithm

That's the K-M algorithm

Can be very slow in high dimensions with non-elliptical contours, e.g.



Can get stuck in corners for a long time!

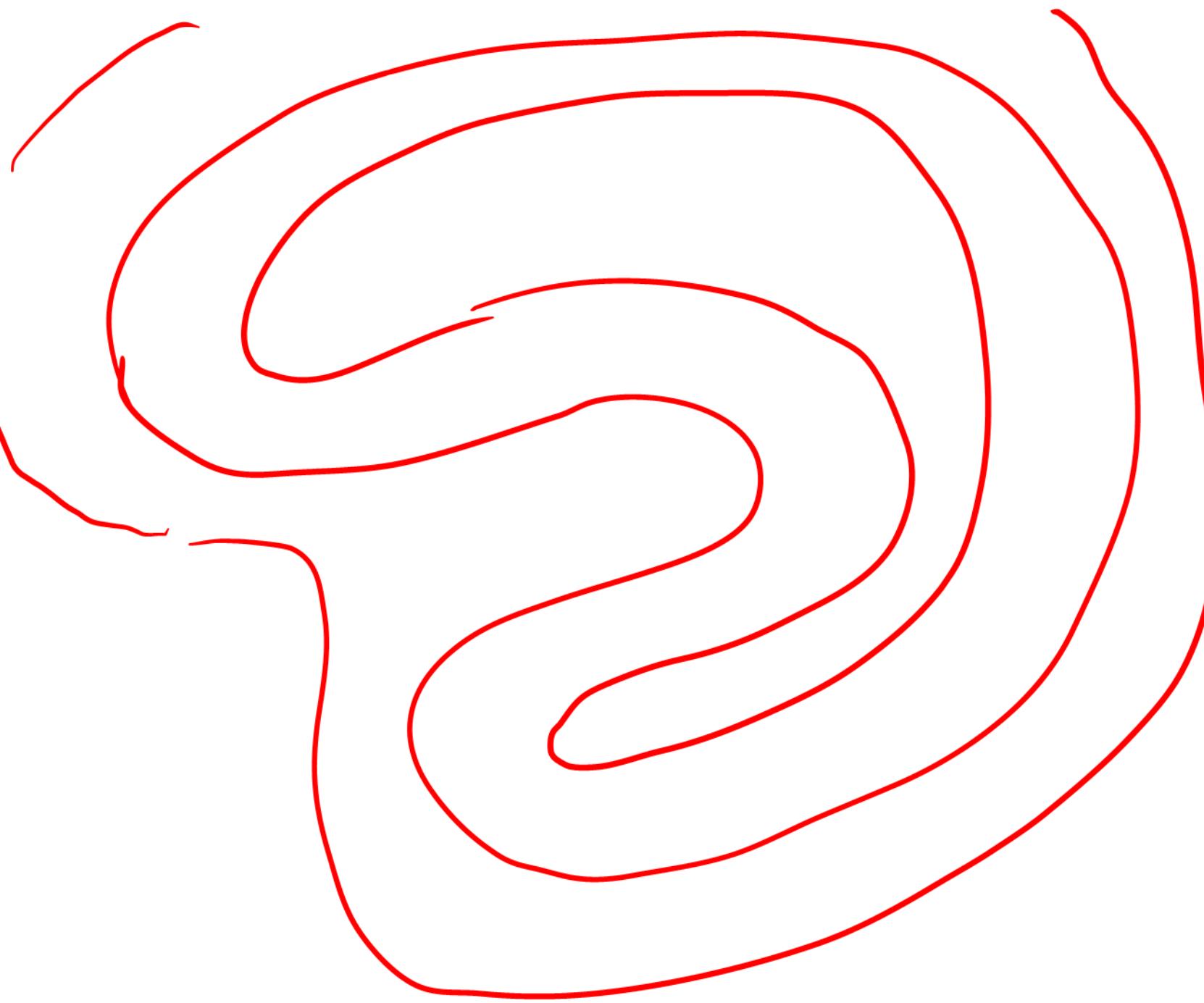
H. Hamiltonian Monte Carlo

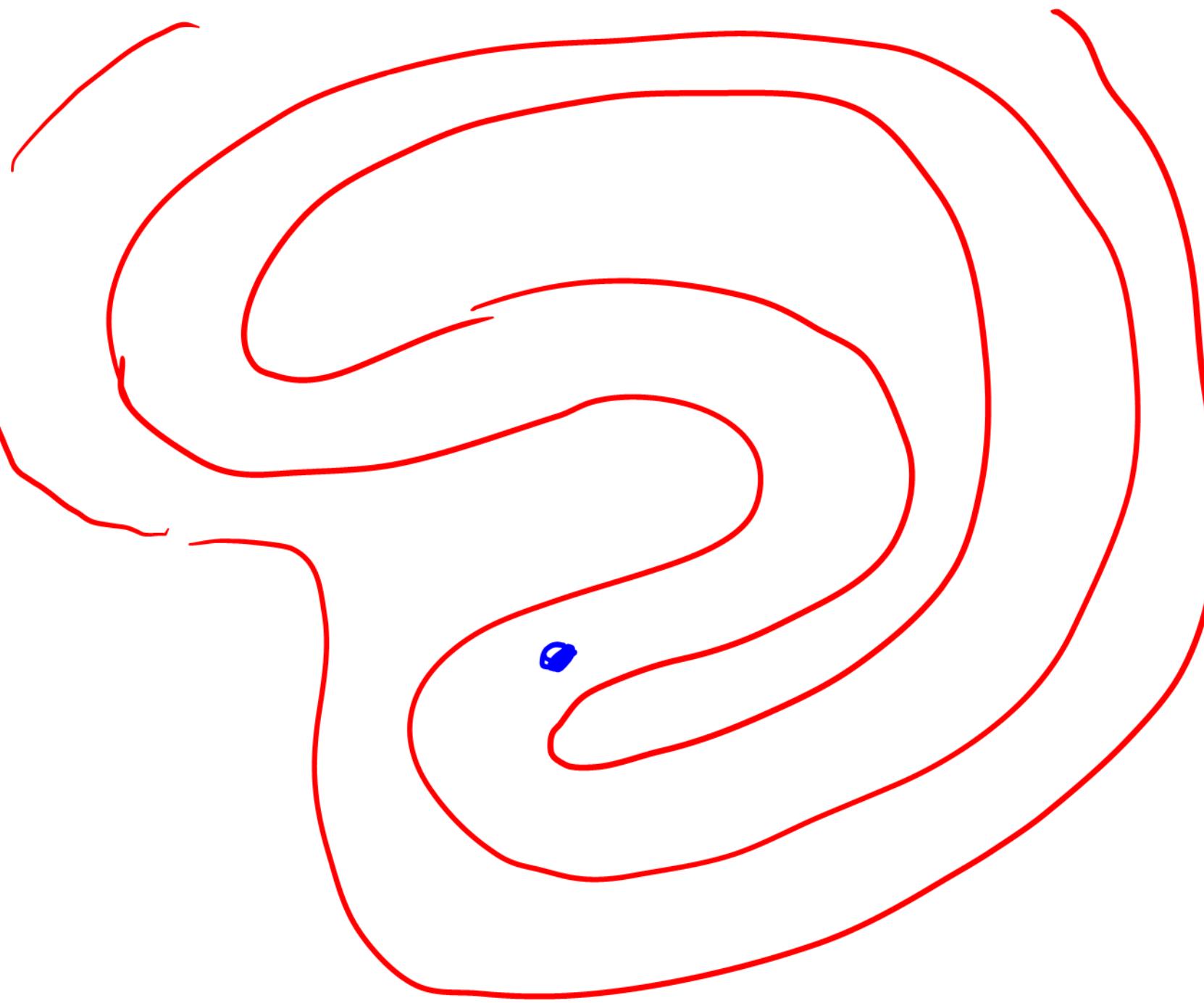
- Turn the mountain into a bowl by using $-\log P(\theta, x)$
- Instead of random steps)
go for a ride on a frictionless skateboard with swivel wheels - starting with a random push.

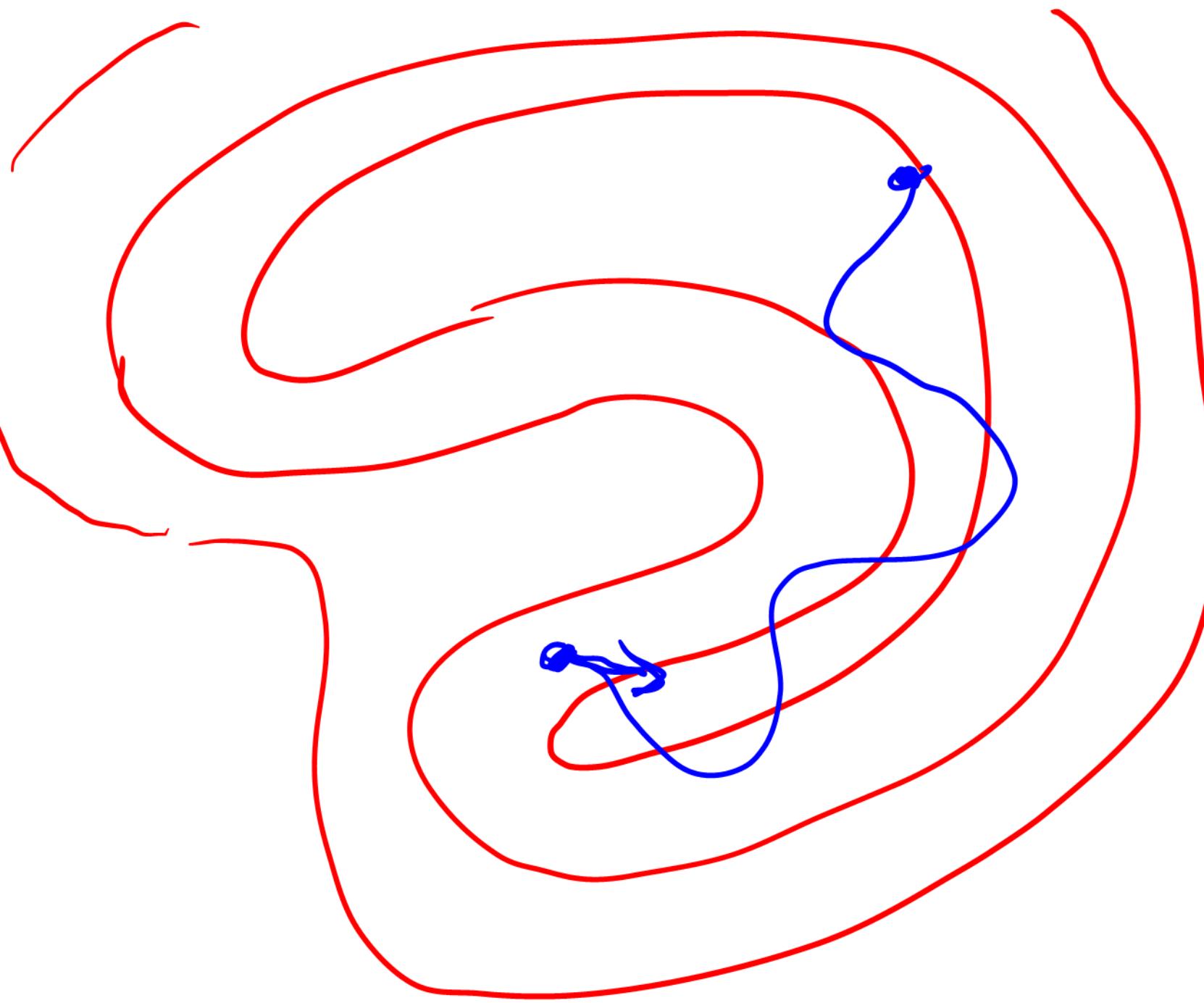
$-\log P(\theta, x)$



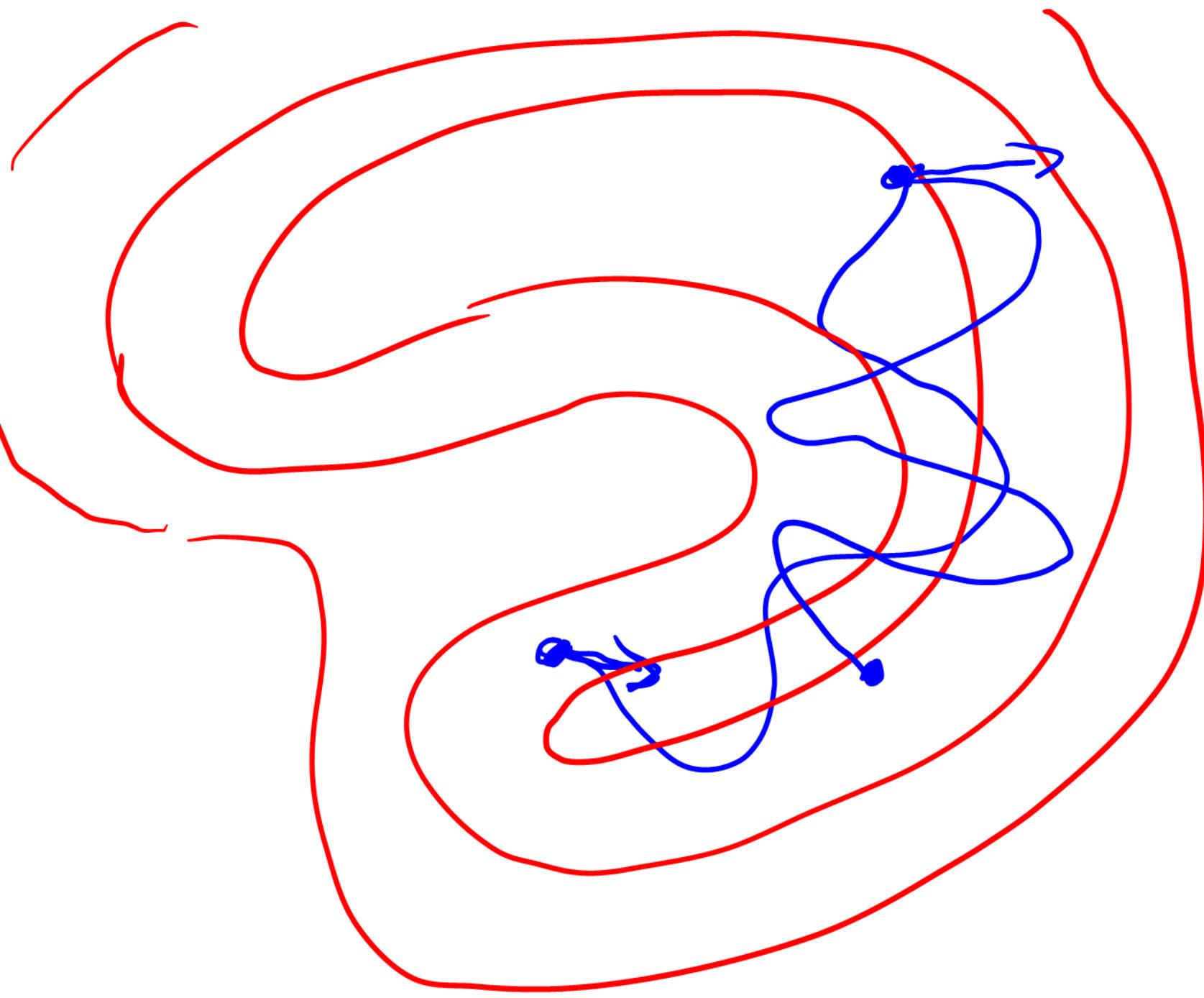
~

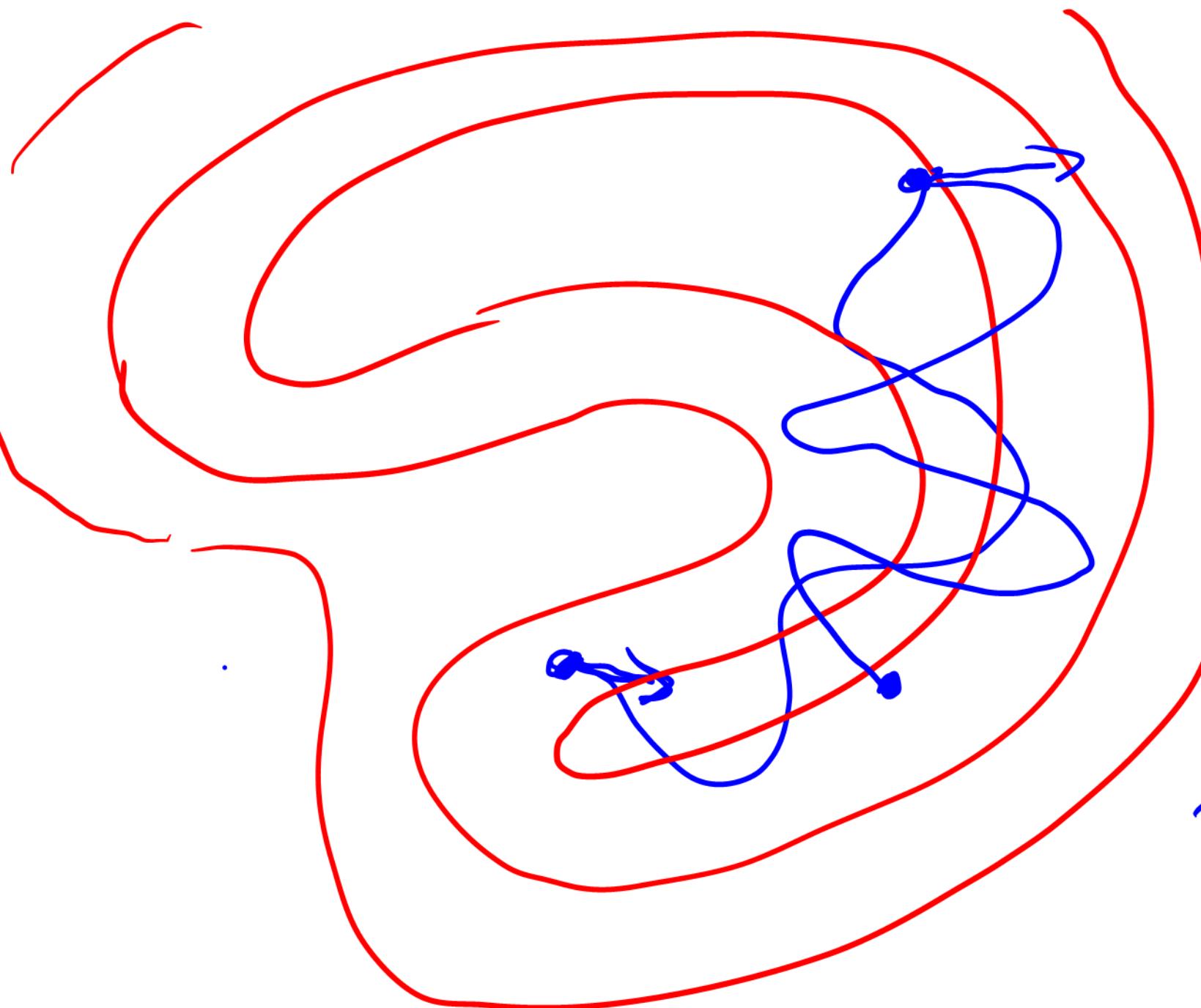




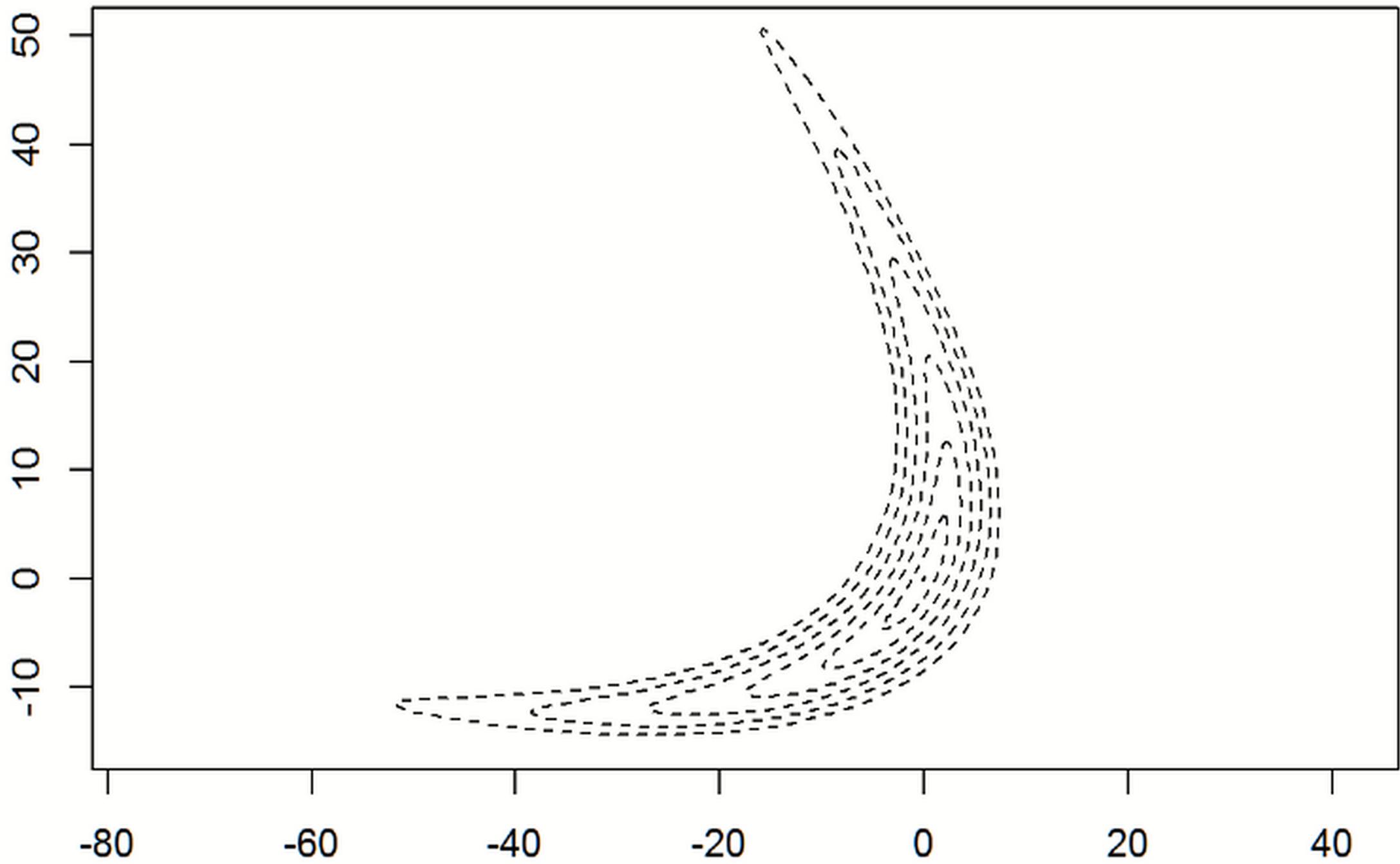


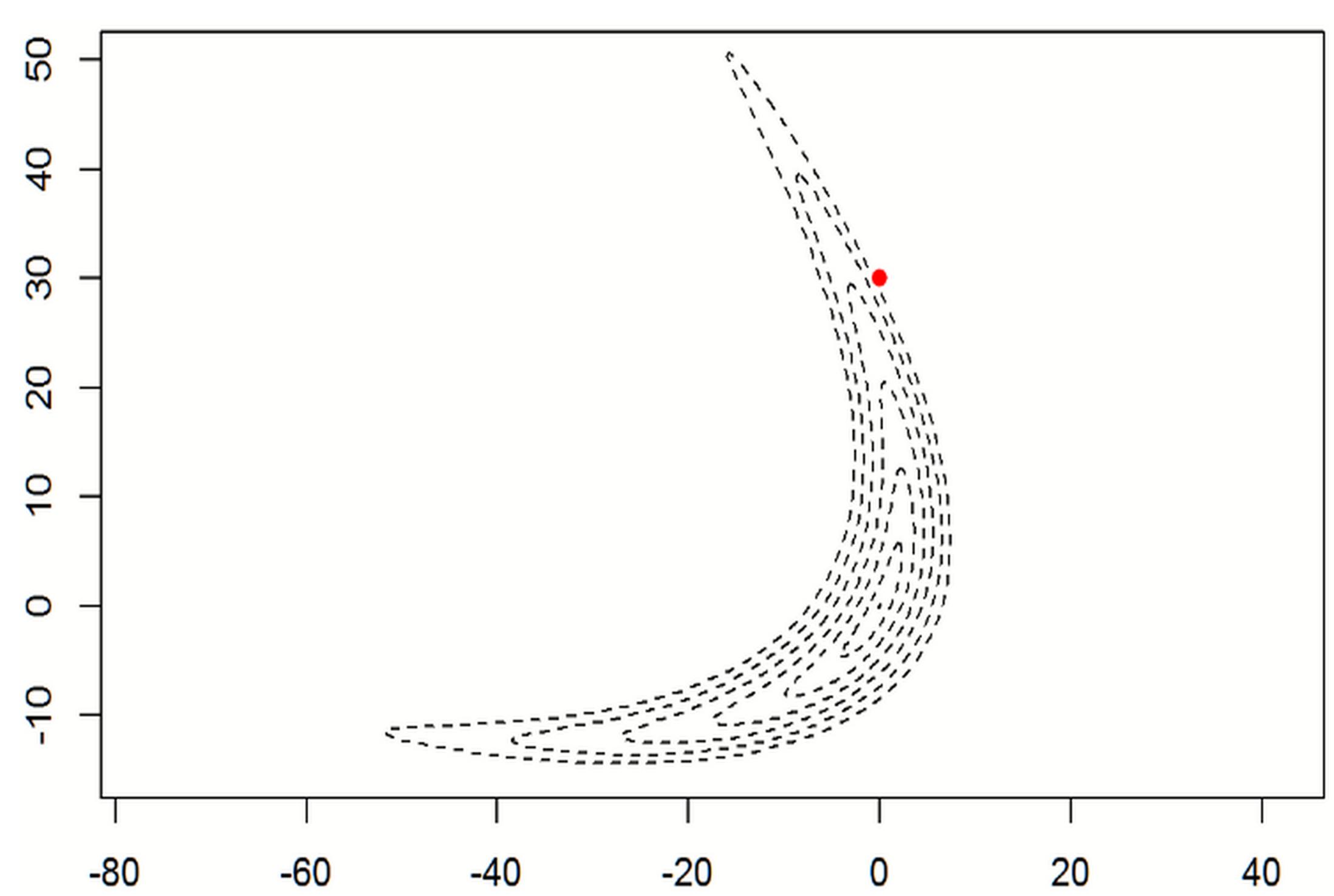
Stop
after
a set
time.

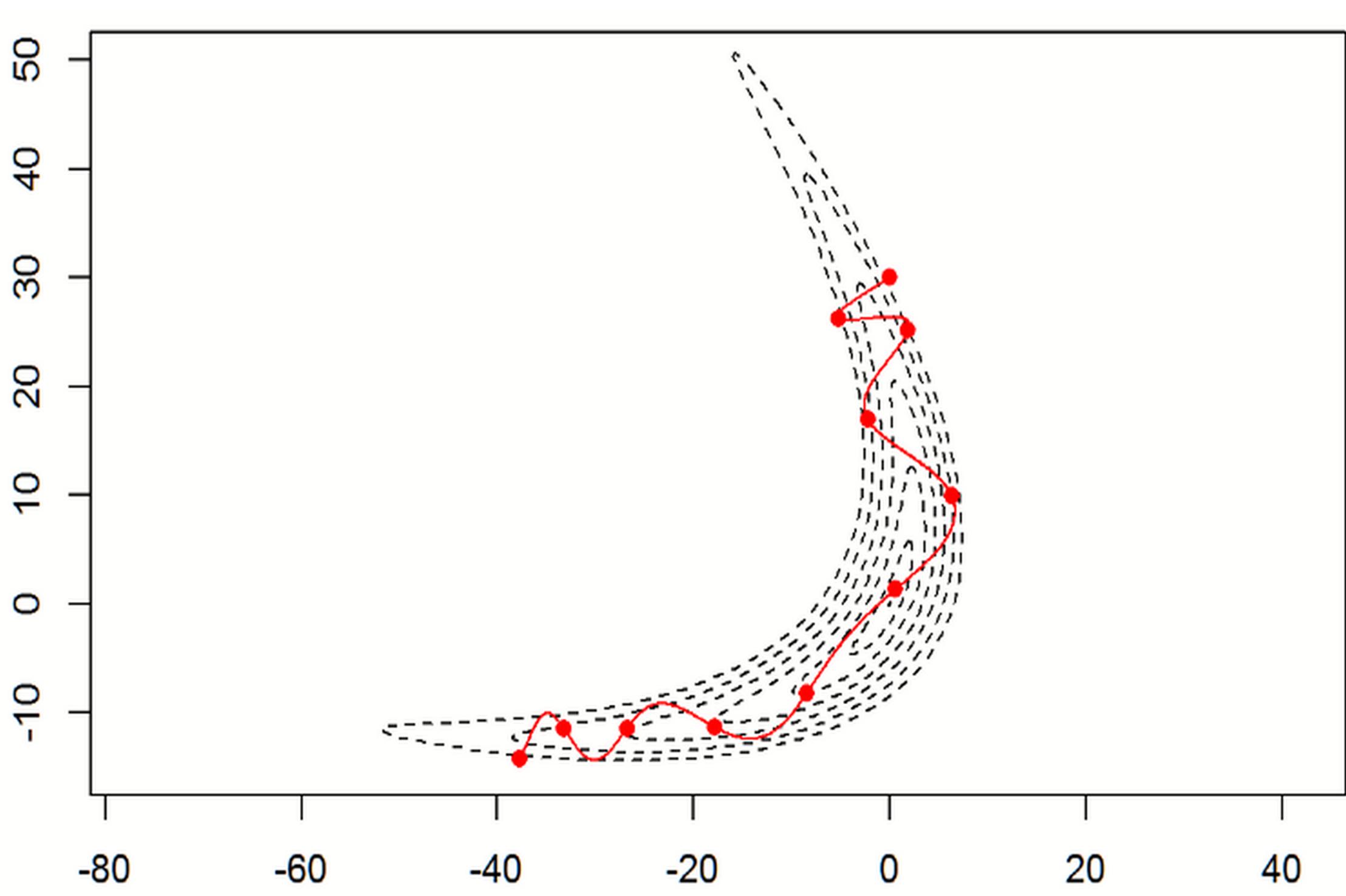


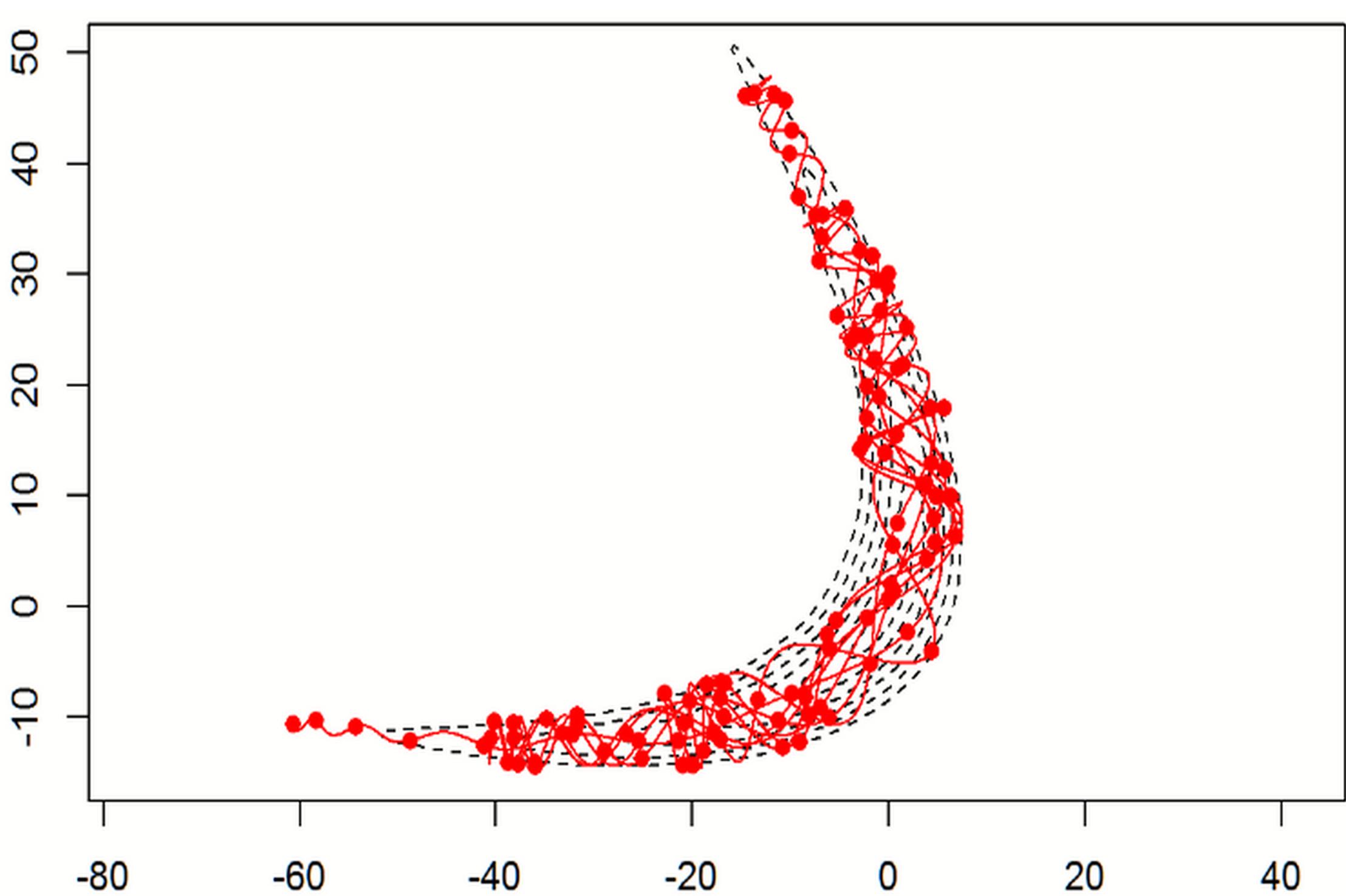


This
can do
a much
better
job of
covering
 $p(\theta|x)$,









L : # of leaps (Leapfrog) steps

ϵ : size of each step

$L\epsilon$: "distance" travelled in Θ -space

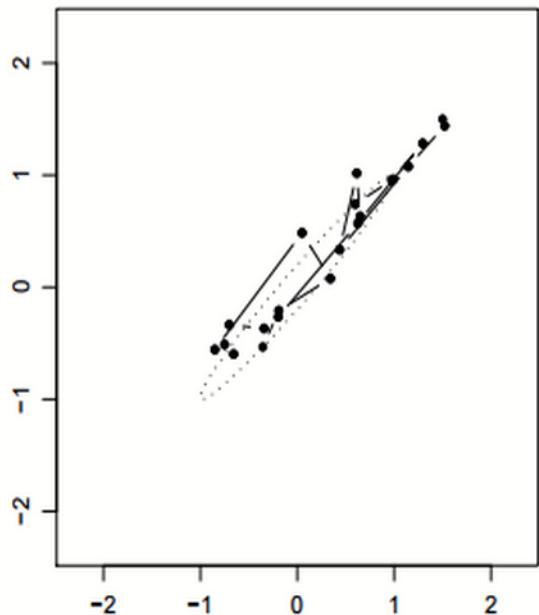
If ϵ too small - too much like random walk

If L too large - too much computing time

If ϵ too large - bad numerical approximation
to continuous path and too
high likelihood of rejecting proposal.

From Neal (2011)

Random-walk Metropolis



Hamiltonian Monte Carlo

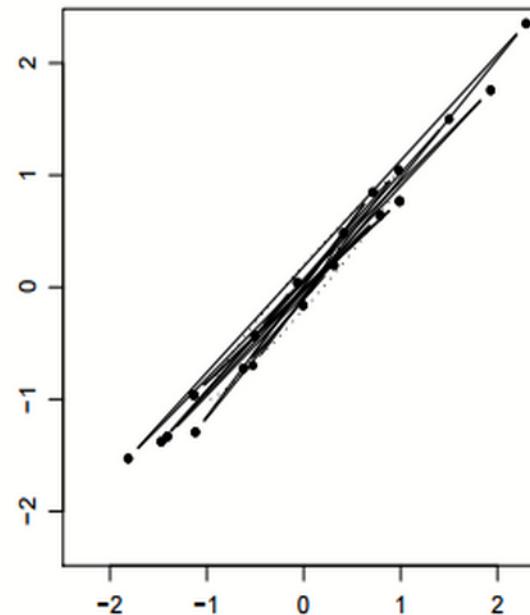
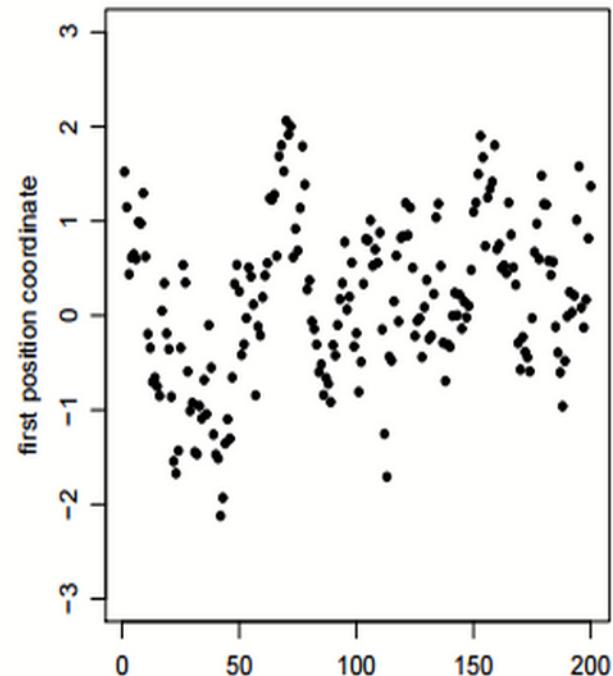


Figure 4: Twenty iterations of the random-walk Metropolis method (with 20 updates per iteration) and of the Hamiltonian Monte Carlo method (with 20 leapfrog steps per trajectory) for a 2D Gaussian distribution with marginal standard deviations of one and correlation 0.98. Only the two position coordinates are plotted, with ellipses drawn one standard deviation away from the mean.

Random-walk Metropolis



Hamiltonian Monte Carlo

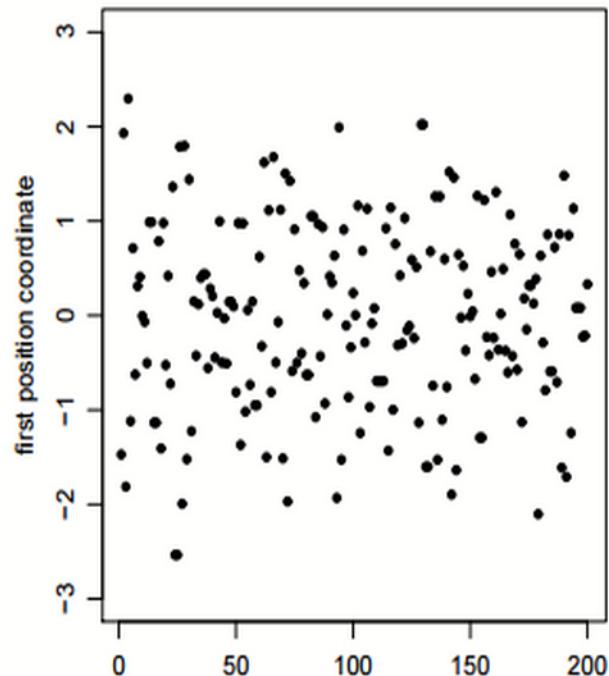


Figure 5: Two hundred iterations, starting with the twenty iterations shown above, with only the first position coordinate plotted.

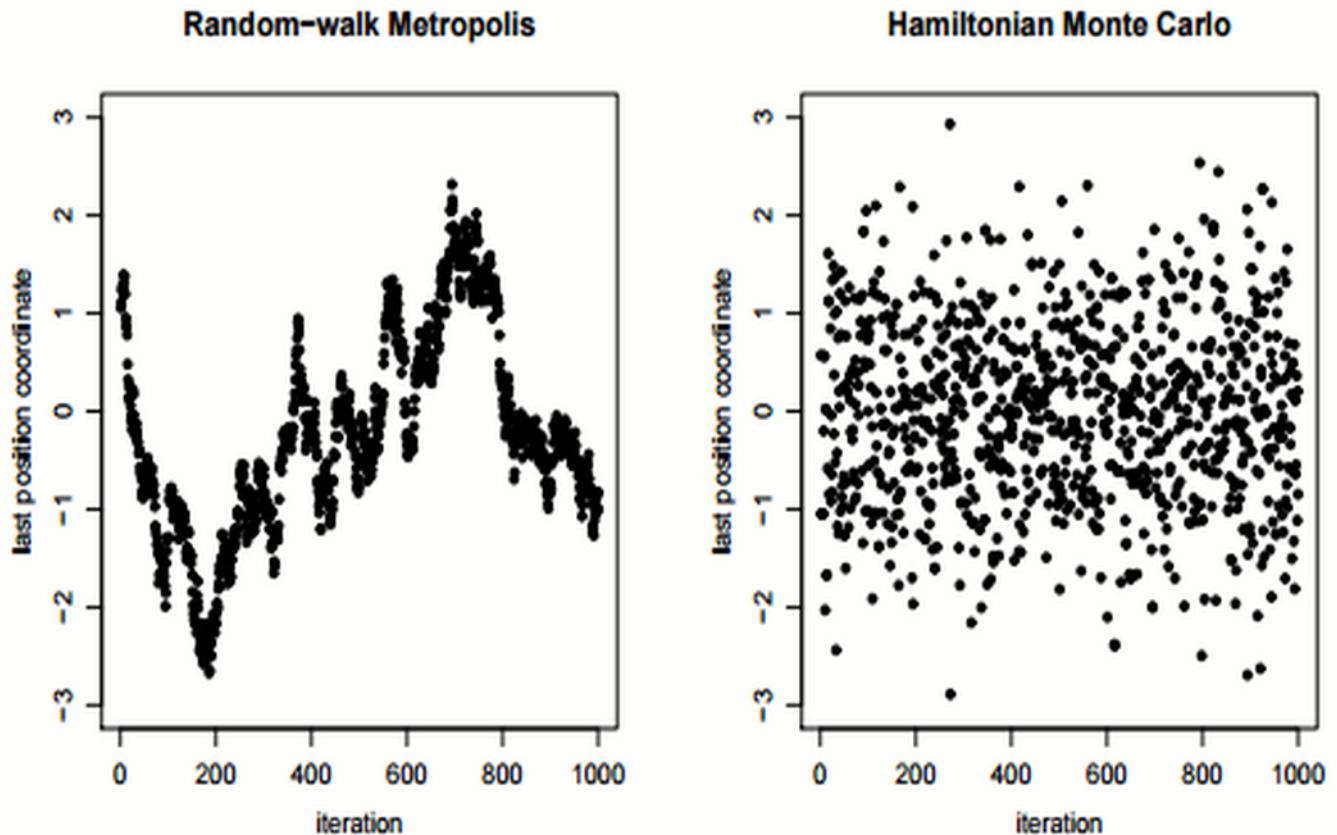


Figure 6: Values for the variable with largest standard deviation for the 100-dimensional example, from a random-walk Metropolis run and an HMC run with $L = 150$. To match computation time, 150 updates were counted as one iteration for random-walk Metropolis.

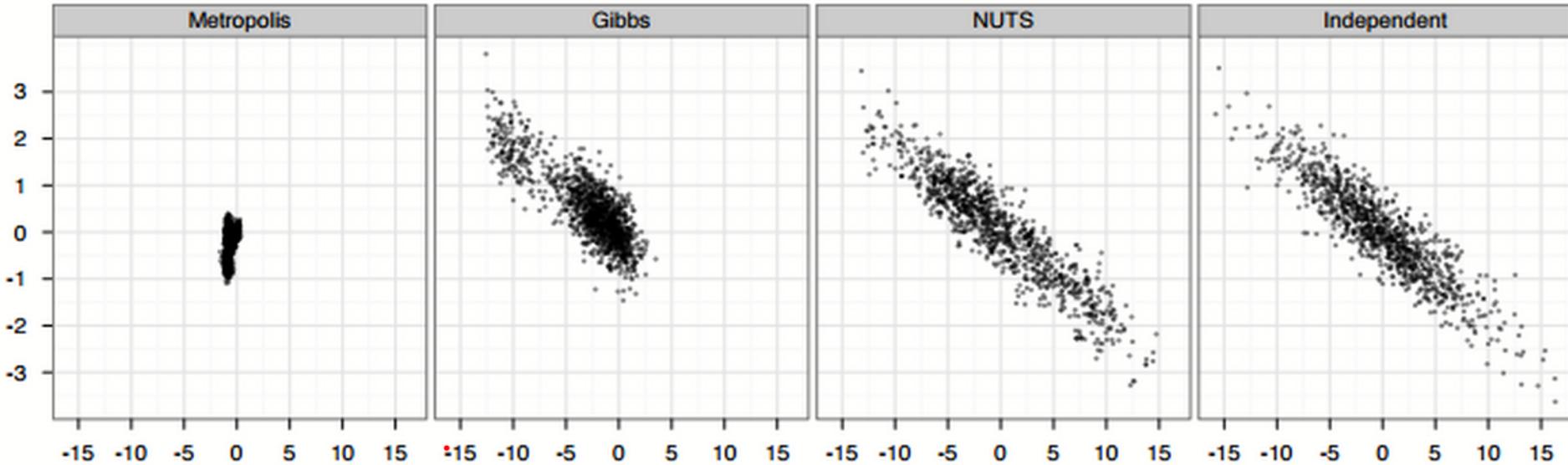


Figure 7: Samples generated by random-walk Metropolis, Gibbs sampling, and NUTS. The plots compare 1,000 independent draws from a highly correlated 250-dimensional distribution (right) with 1,000,000 samples (thinned to 1,000 samples for display) generated by random-walk Metropolis (left), 1,000,000 samples (thinned to 1,000 samples for display) generated by Gibbs sampling (second from left), and 1,000 samples generated by NUTS (second from right). Only the first two dimensions are shown here.

from Hoffman & Gelman (2014)

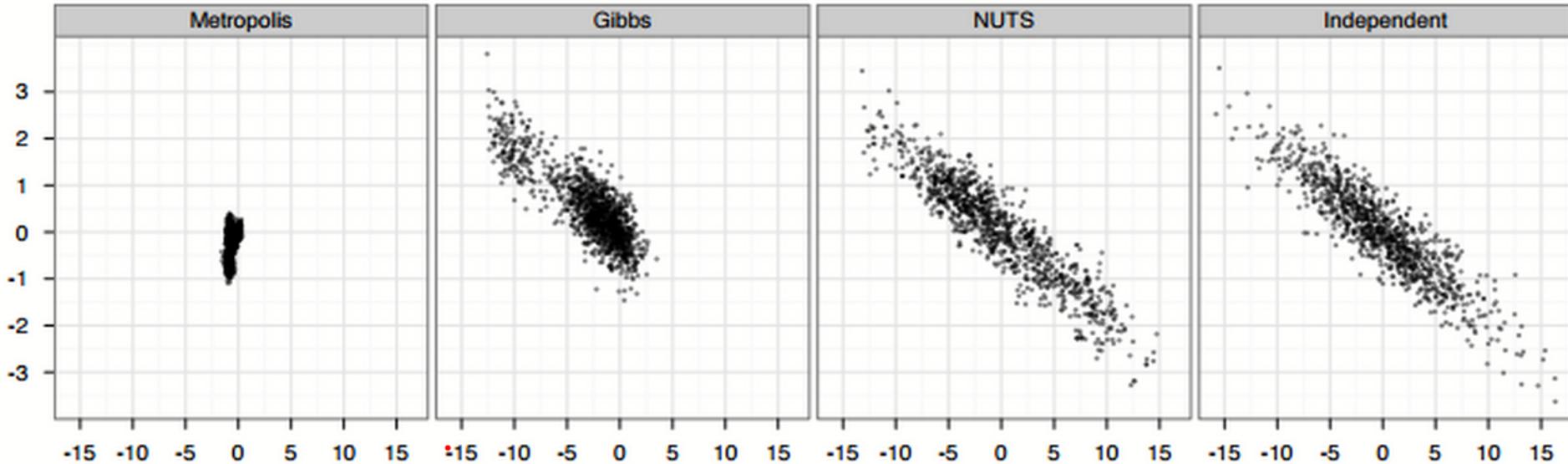
$\mathcal{O}(P^2)$ $\mathcal{O}(P^2)$ $\mathcal{O}(P^{5/4})$ 

Figure 7: Samples generated by random-walk Metropolis, Gibbs sampling, and NUTS. The plots compare 1,000 independent draws from a highly correlated 250-dimensional distribution (right) with 1,000,000 samples (thinned to 1,000 samples for display) generated by random-walk Metropolis (left), 1,000,000 samples (thinned to 1,000 samples for display) generated by Gibbs sampling (second from left), and 1,000 samples generated by NUTS (second from right). Only the first two dimensions are shown here.

from Hoffman & Gelman (2014)