

Bayesian Ideas and Modern Bayesian Methods – An Introduction

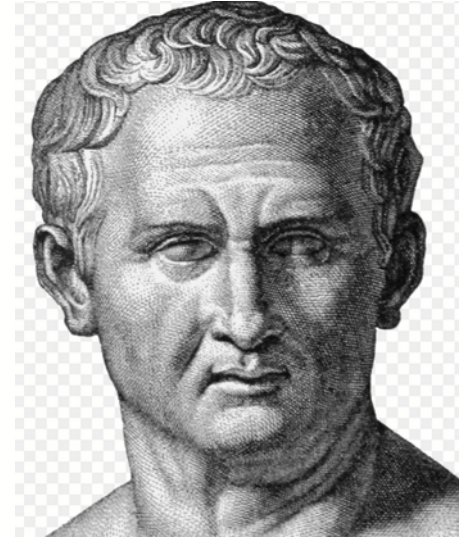
Georges Monette

georges@yorku.ca

Cicero (106 BCE – 43 BCE)

gave two definitions for *probabile*:

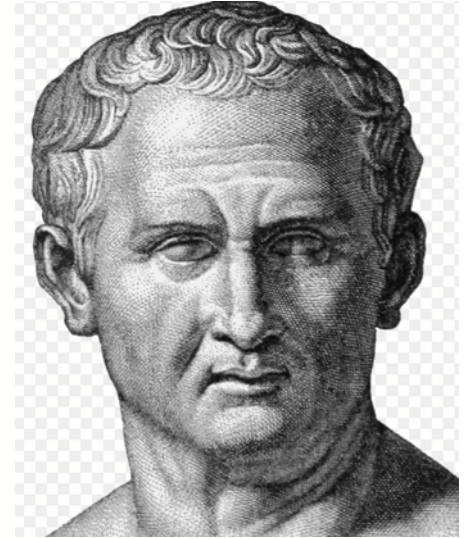
- That which usually happens
- That which is commonly believed



Cicero (106 BCE – 43 BCE)

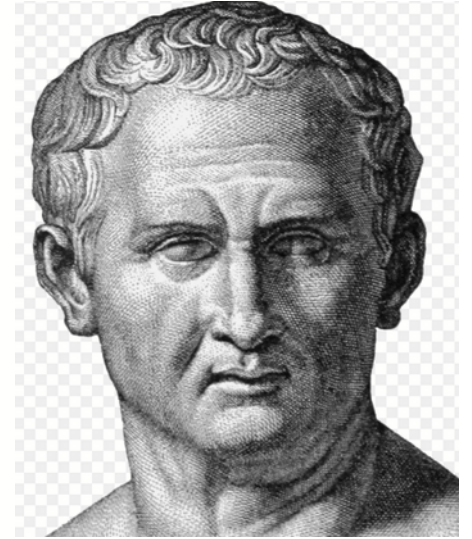
gave two definitions for *probabile*:

- That which usually happens
 - relative frequency
 - frequentist objective interpretation
- That which is commonly believed



Cicero (106 BCE – 43 BCE)

gave two definitions for *probabile*:



- That which usually happens
 - relative frequency
 - frequentist objective interpretation
- That which is commonly believed
 - degree of belief in a hypothesis
 - Bayesian subjective interpretation

What's the difference between Bayesians and Frequentists? An oversimplification:

- ▶ 'Proper subjective Bayesians' believe that the role of empirical evidence is to *transform* beliefs in a coherent way using Bayes theorem by conditioning on what has been observed

What's the difference between Bayesians and Frequentists? An oversimplification:

- ▶ 'Proper subjective Bayesians' believe that the role of empirical evidence is to *transform* beliefs in a coherent way using Bayes theorem by conditioning on what has been observed
- ▶ Frequentists believe (hope?) that it is possible to find an objective basis for the interpretation of empirical evidence, one that does not require a reference to personal beliefs

What's the difference between Bayesians and Frequentists? An oversimplification:

- ▶ 'Proper subjective Bayesians' believe that the role of empirical evidence is to *transform* beliefs in a coherent way using Bayes theorem by conditioning on what has been observed
- ▶ Frequentists believe (hope?) that it is possible to find an objective basis for the interpretation of empirical evidence, one that does not require a reference to personal beliefs
- ▶ Both attempt to incorporate uncertainty – in orthogonal directions – but nonetheless often produce similar results because of symmetries in common statistical models and asymptotically

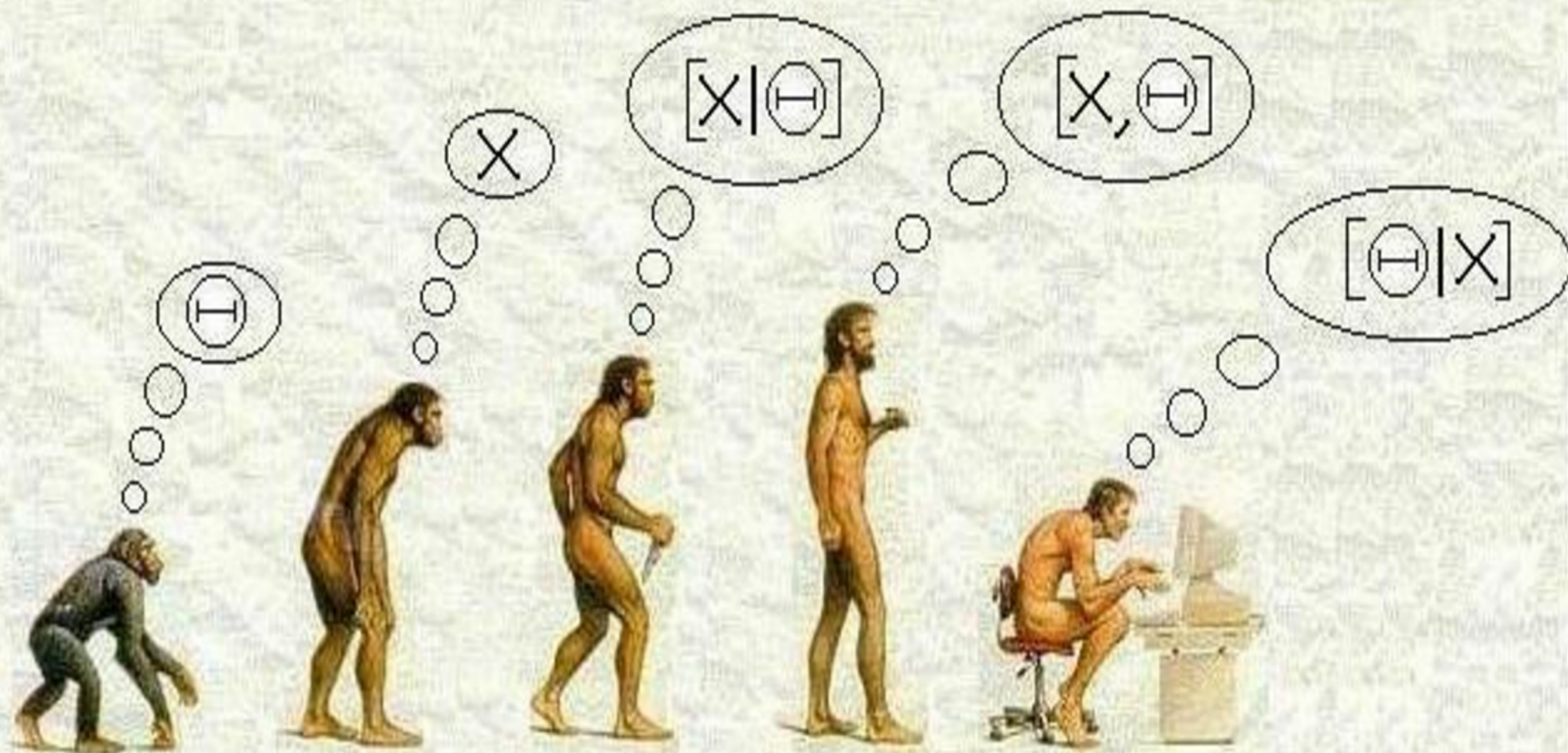
What's the difference between Bayesians and Frequentists? An oversimplification:

- ▶ 'Proper subjective Bayesians' believe that the role of empirical evidence is to *transform* beliefs in a coherent way using Bayes theorem by conditioning on what has been observed
- ▶ Frequentists believe (hope?) that it is possible to find an objective basis for the interpretation of empirical evidence, one that does not require a reference to personal beliefs
- ▶ Both attempt to incorporate uncertainty – in orthogonal directions – but nonetheless often produce similar results because of symmetries in common statistical models and asymptotically

From ideology to utility

- ▶ Until recently the debate was mainly philosophical. F methods were much easier
- ▶ With improvements in MCMC, B methods have become more feasible and surpass F methods for many complex problems

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



HOMO
APRIORIUS

HOMO
PRAGMATICUS

HOMO
FREQUENTISTUS

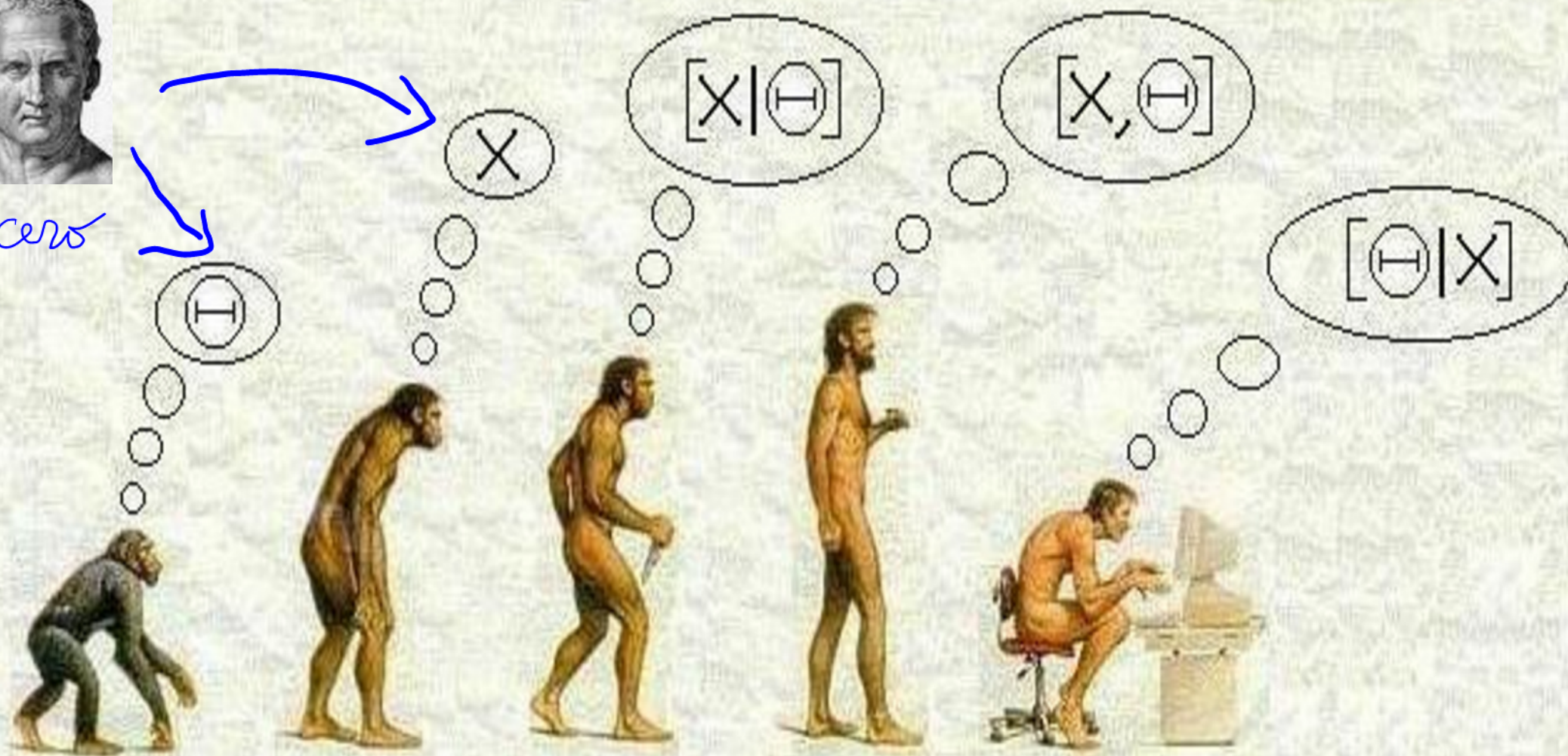
HOMO
SAPIENS

HOMO
BAYESIANIS

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



Cicero



**HOMO
APRIORIUS**

**HOMO
PRAGMATICUS**

**HOMO
FREQUENTISTUS**

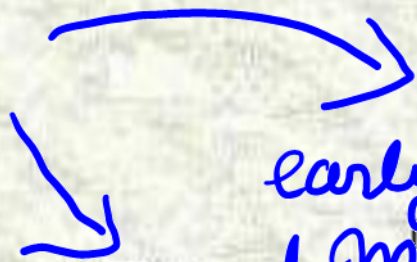
**HOMO
SAPIENS**

**HOMO
BAYESIANIS**

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



Cicero



early 1700s
de Moivre



HOMO
APRIORIUS



HOMO
PRAGMATICUS



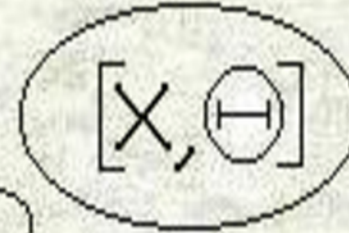
HOMO
FREQUENTISTUS



HOMO
SAPIENS

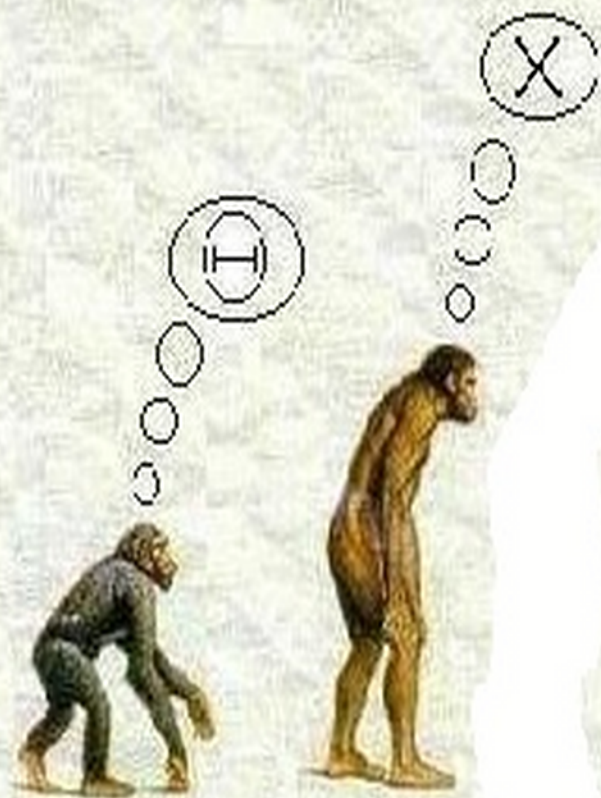


HOMO
BAYESIANIS



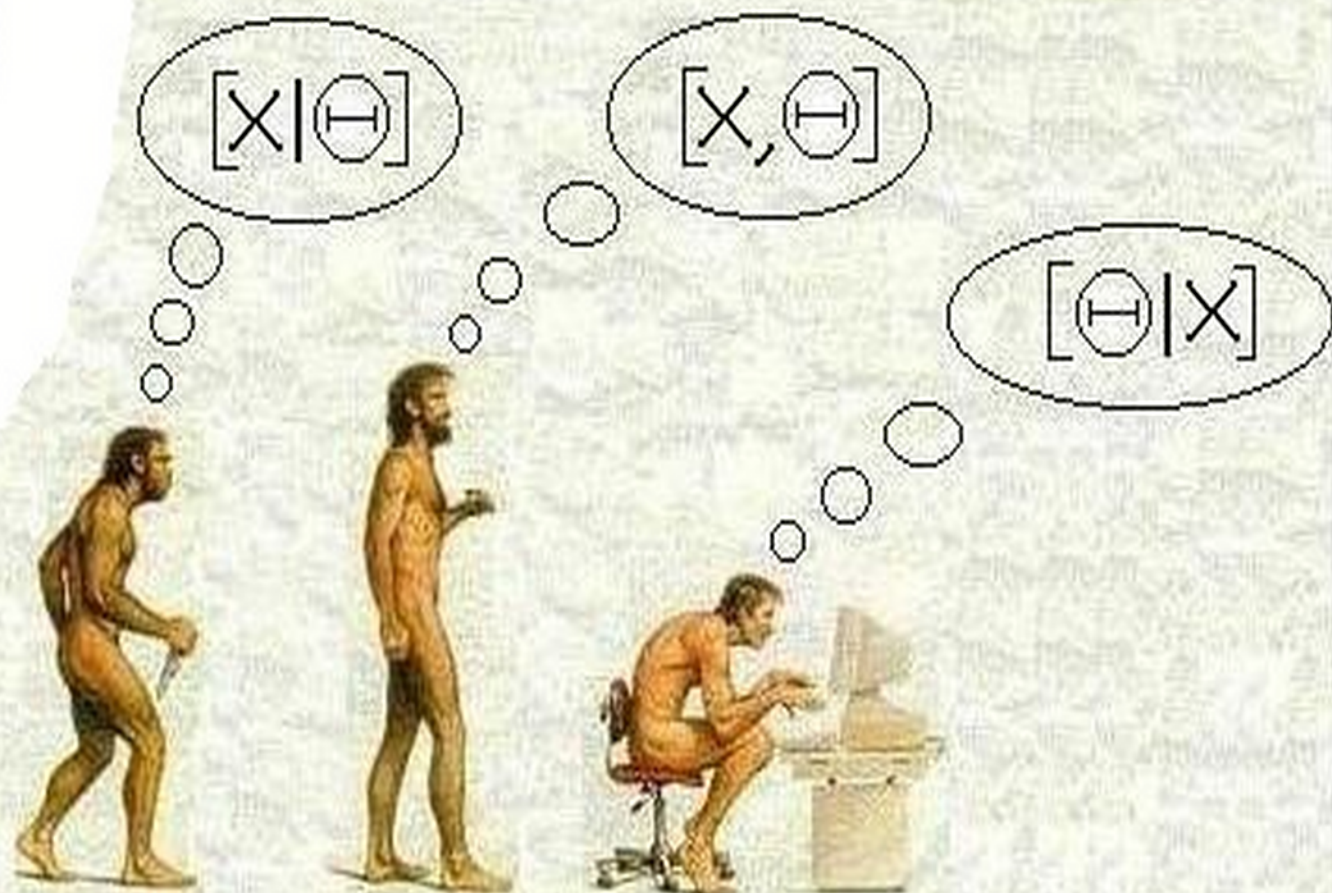
(YET ANOTHER) HISTORY OF

OF LIFE AS WE KNOW IT...



HOMO
APRIORIUS

HOMO
PRAGMATICUS



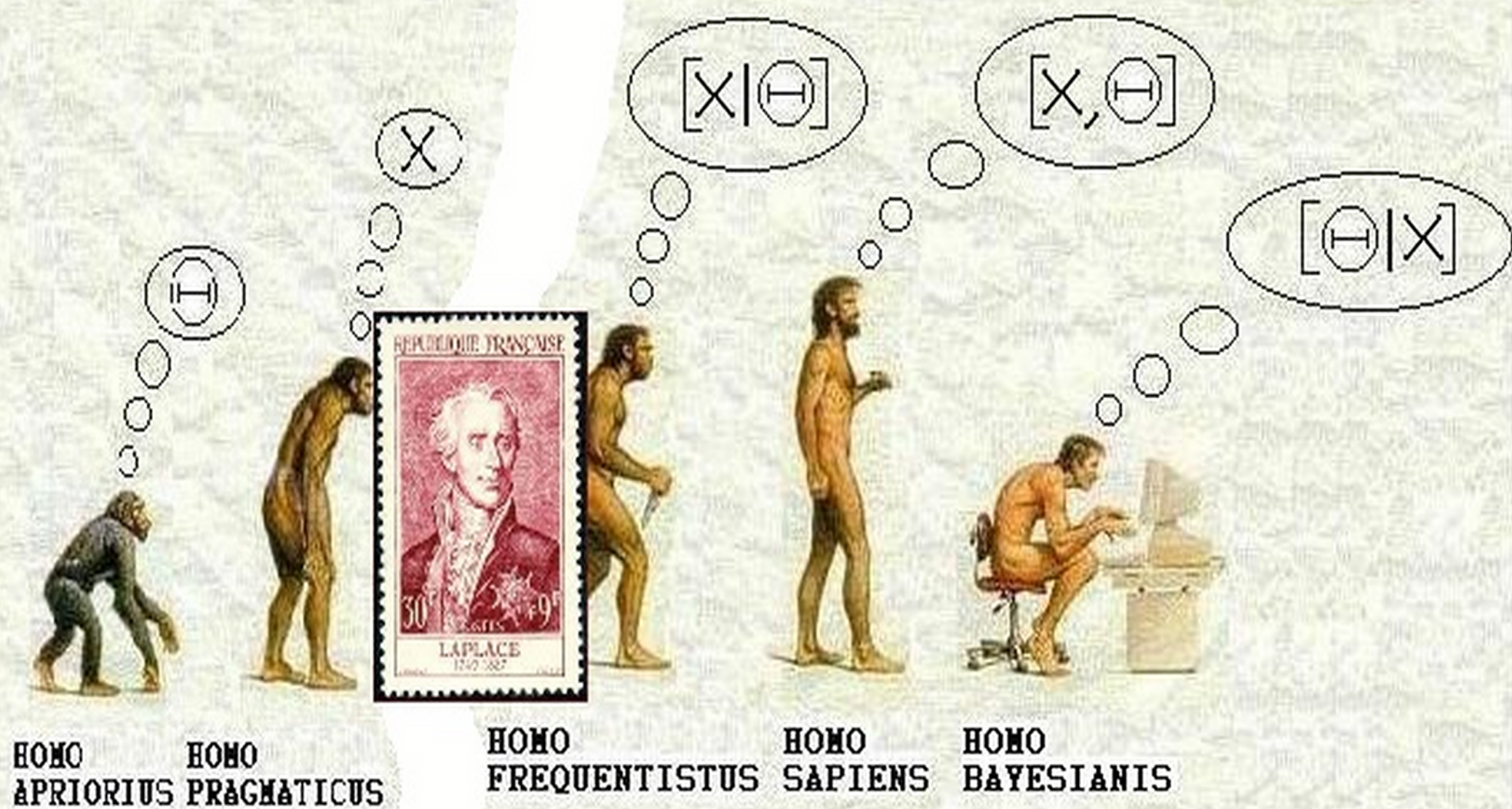
HOMO
FREQUENTISTUS

HOMO
SAPIENS

HOMO
BAYESIANIS

(YET ANOTHER) HISTORY OF

OF LIFE AS WE KNOW IT...



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

$[H|X]$ - 1800s



HOMO
APRIORIUS



HOMO
PRAGMATICUS



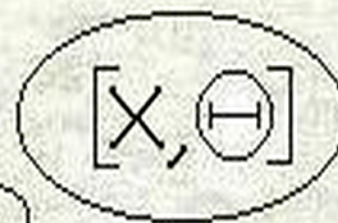
HOMO
FREQUENTISTUS



HOMO
SAPIENS



HOMO
BAYESIANIS



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

$[\Theta | X]$ - 1800s

$[X | \Theta]$

$[X, \Theta]$

$[\Theta | X]$

Fisher
1920



HOMO
APRIORIUS



HOMO
PRAGMATICUS



HOMO
FREQUENTISTUS



HOMO
SAPIENS



HOMO
BAYESIANIS

R. A. Fisher's clever idea

The Lady Tasting Tea and the p-value: a frequentist basis for inference

In 1919, Dr. Muriel Bristol at Rothampsted Experimental Station claimed she could tell whether the milk was poured in first or the tea first.

- ▶ Imagine that she was offered 12 cups of tea in random order – 6 prepared milk first and 6 tea first
- ▶ She got 10 of the 12 right

What does this tell us about her ability to tell the difference?

Rationale behind the p-value

How can we quantify the evidence that she can tell the difference?

- ▶ Pretend that she can't tell the difference: 'null hypothesis' H_0
- ▶ The probability of getting 10 out of 12 right is $p(y|H_0) = 0.038961$
- ▶ But the probability of any single outcome, even one consistent with H_0 , might be very small and might say nothing against H_0
- ▶ Fisher's idea: use the *tail probability*, the probability of y as or more extreme than the observed value of y

p-value:

$$\begin{aligned}\Pr(y^+|H_0) &= p(y = 10|H_0) + p(y = 12|H_0) \\ &= 0.038961 + 0.001082 \\ &= 0.040043\end{aligned}$$

Proof by contradiction/implausibility

Contradiction

A implies not B

B true

Therefore A is false

Implausibility

A implies B is improbable

B is observed

Therefore A is unlikely

Courtroom analogy: presumption of innocence

H_0 : Innocence

Consider probability of data (evidence) | innocence

If evidence inconsistent with innocence, then reject innocence
and find guilt



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

$$p\text{-value} = P_2(Y^+ | H_0)$$

H_0 : Sally is innocent

Y : 2 children die for no apparent cause

$$p\text{-value} = P_2(Y^+ | H_0)$$

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

$$p\text{-value} = P_2(Y^+ | H_0)$$

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

$$p\text{-value} = P_2(Y^+ | H_0)$$

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

$$p\text{-value} = P_2(Y^+ | H_0)$$

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

$$p\text{-value} = P_2(Y^+ | H_0)$$

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 $\frac{1}{10,000}$!

So anyways:

$$P < 0.0001$$

Therefore guilty beyond
a reasonable doubt.

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!?
- So $P(Y^+ | H_0)$ 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!?
- So $P(Y^+ | H_0)$ a good proxy for $P(H_0 | Y)$?

Bayes Theorem:

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

$$= \frac{P(Y|H_0)P(H_0)}{P(Y|H_0)P(H_0) + P(Y|H_0^c)P(H_0^c)}$$

Bayes Theorem:

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

$$= \frac{P(Y|H_0)P(H_0)}{P(Y|H_0)P(H_0) + P(Y|H_0^c)P(H_0^c)}$$

NOT VERY INTUITIVE!

Bayes Theorem:

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

$$= \frac{P(Y|H_0)P(H_0)}{P(Y|H_0)P(H_0) + P(Y|H_0^c)P(H_0^c)}$$

OK - from model

Bayes Theorem:

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

$$= \frac{P(Y|H_0)P(H_0)}{P(Y|H_0)P(H_0) + P(Y|H_0^c)P(H_0^c)}$$

OK - from model

? - not given in model
- "prior"

Bayes Theorem:

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

Posterior

$$= \frac{P(Y|H_0)P(H_0)}{P(Y|H_0)P(H_0) + P(Y|H_0^c)P(H_0^c)}$$

model

prior

Nicer form used by gamblers

odds ratio

$$\frac{P(H_0 | Y)}{P(H_0^c | Y)} = \frac{P(Y | H_0)}{P(Y | H_0^c)} \times \frac{P(H_0)}{P(H_0^c)}$$

Nicer form used by gamblers

odds ratio

$$\frac{P(H_0 | Y)}{P(H_0^c | Y)} = \frac{P(Y | H_0)}{P(Y | H_0^c)} \times \frac{P(H_0)}{P(H_0^c)}$$

posterior
odds

Bayes
Factor
 $\approx$ Likelihood
ratio

prior
odds

Nicer form used by gamblers

odds ratio

$$\frac{P(H_0 | Y)}{P(H_0^c | Y)} = \frac{P(Y | H_0)}{P(Y | H_0^c)} \times \frac{P(H_0)}{P(H_0^c)}$$

posterior
odds

Bayes
Factor

prior
odds

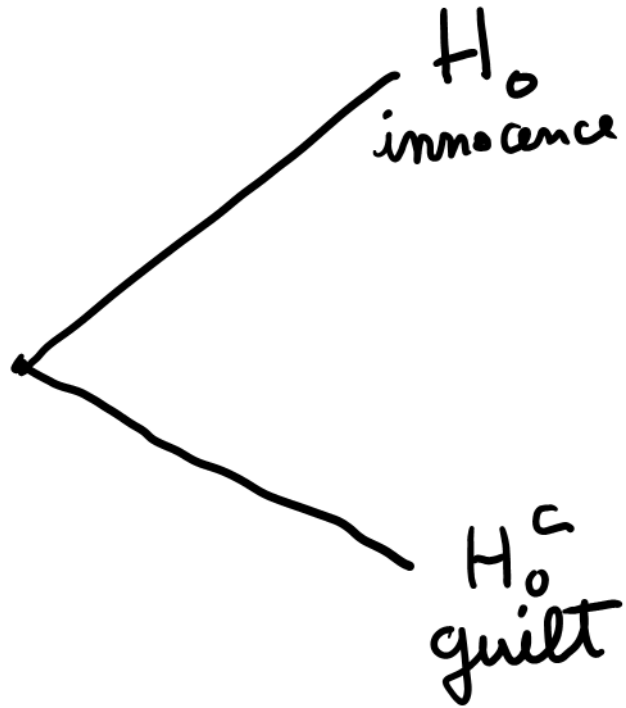
if $BF > 1$ then odds go up

if $BF < 1$ " " " down

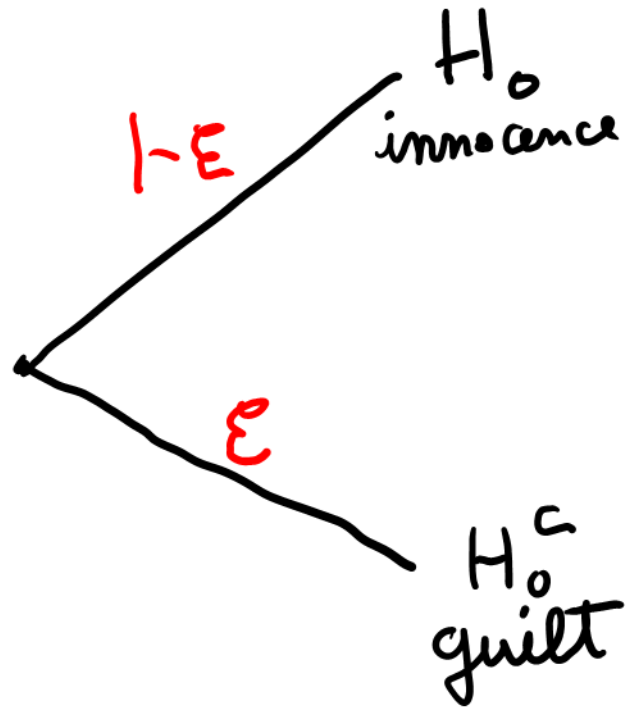
if $BF \approx 1$ then no information

Bayesian tree:

Bayesian tree:

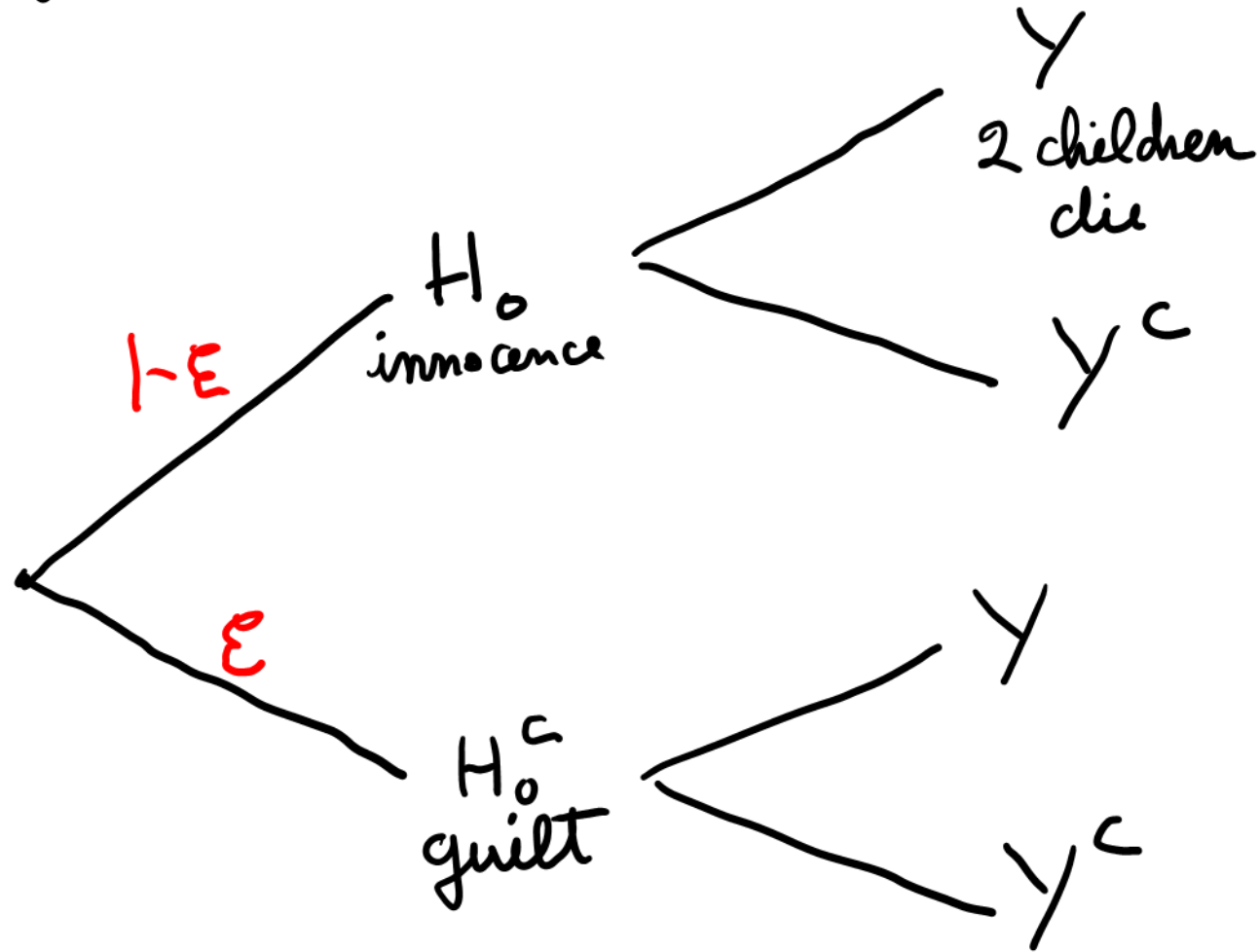


Bayesian tree:

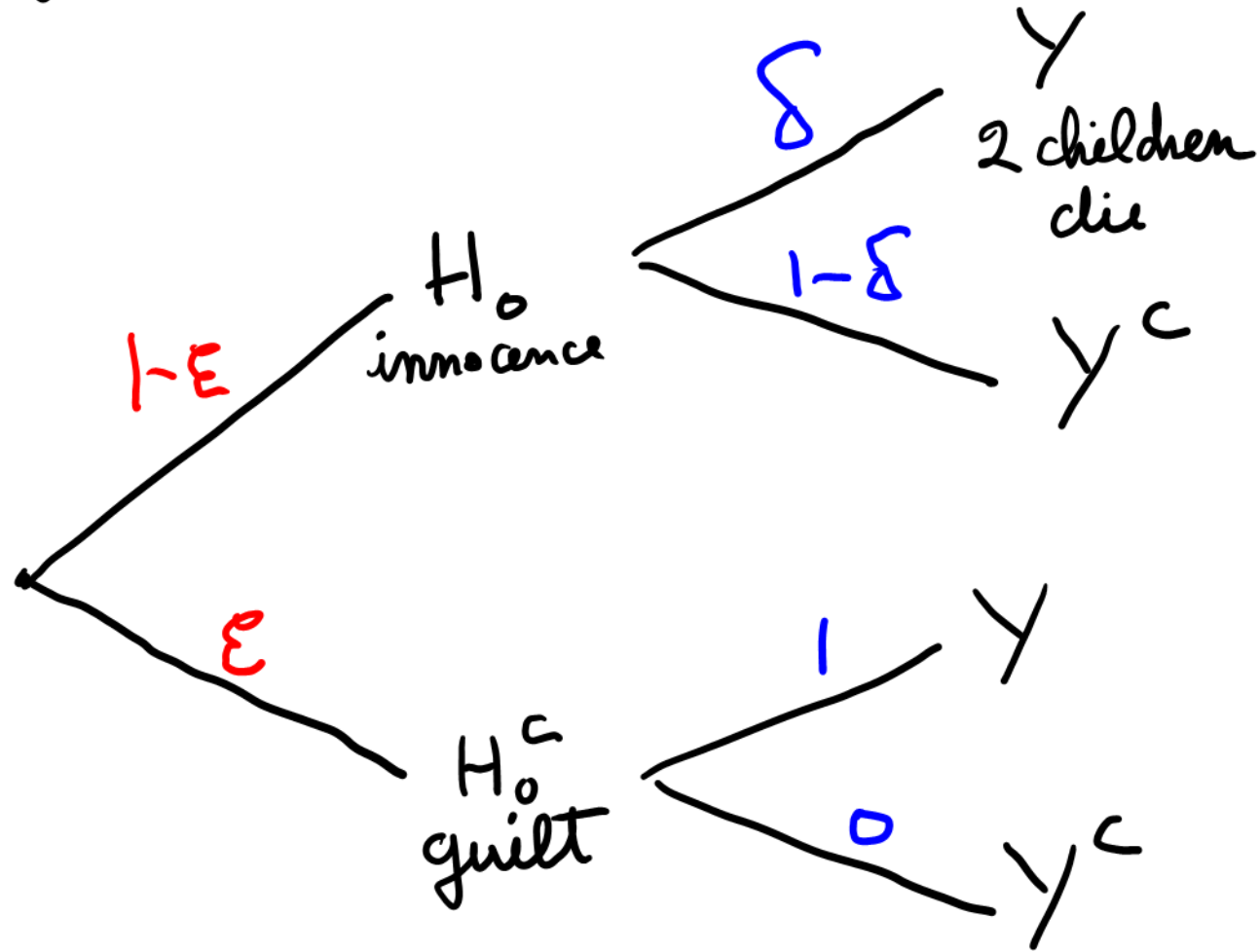


E = very small number

Bayesian tree:

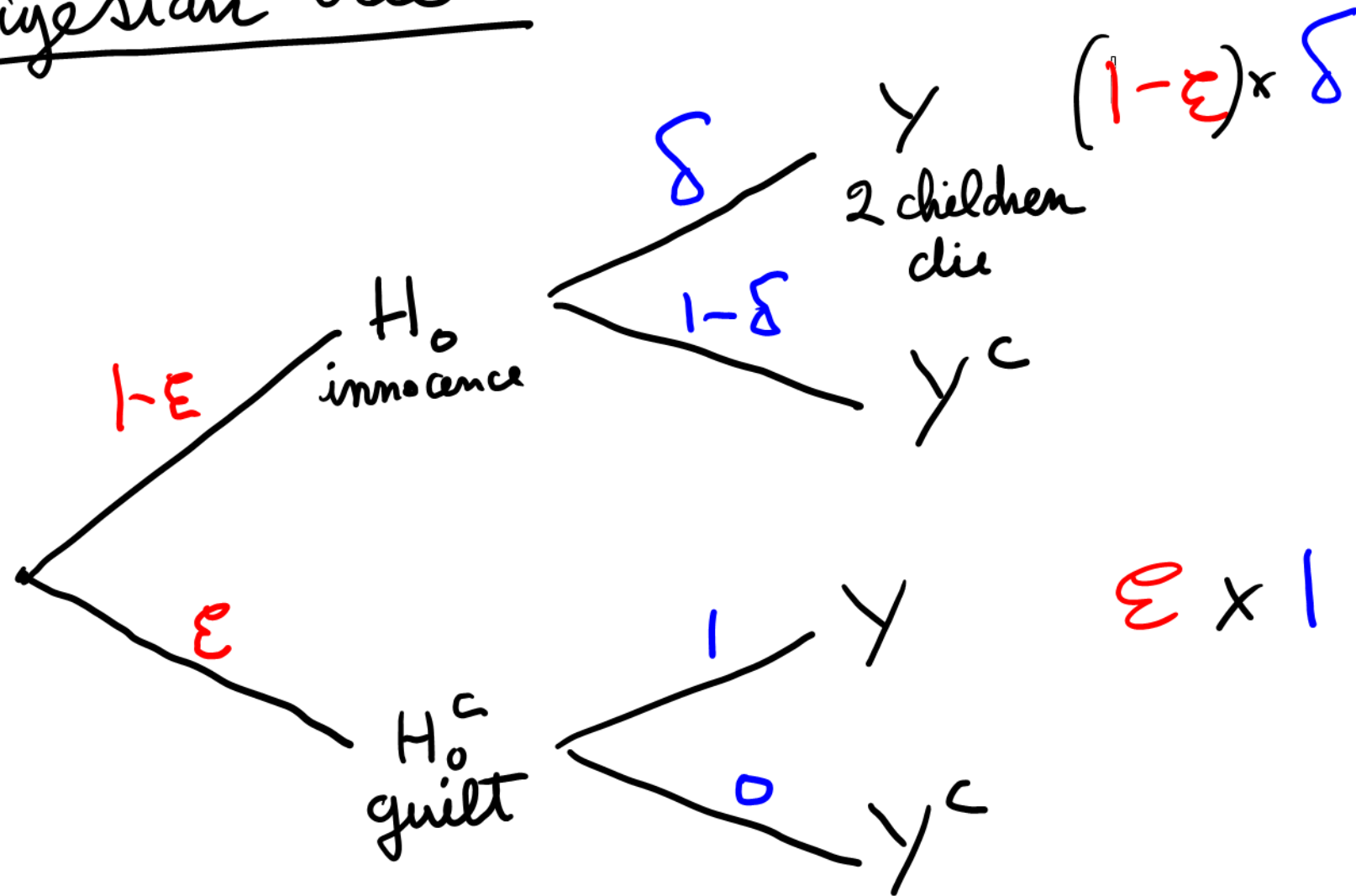


Bayesian tree:



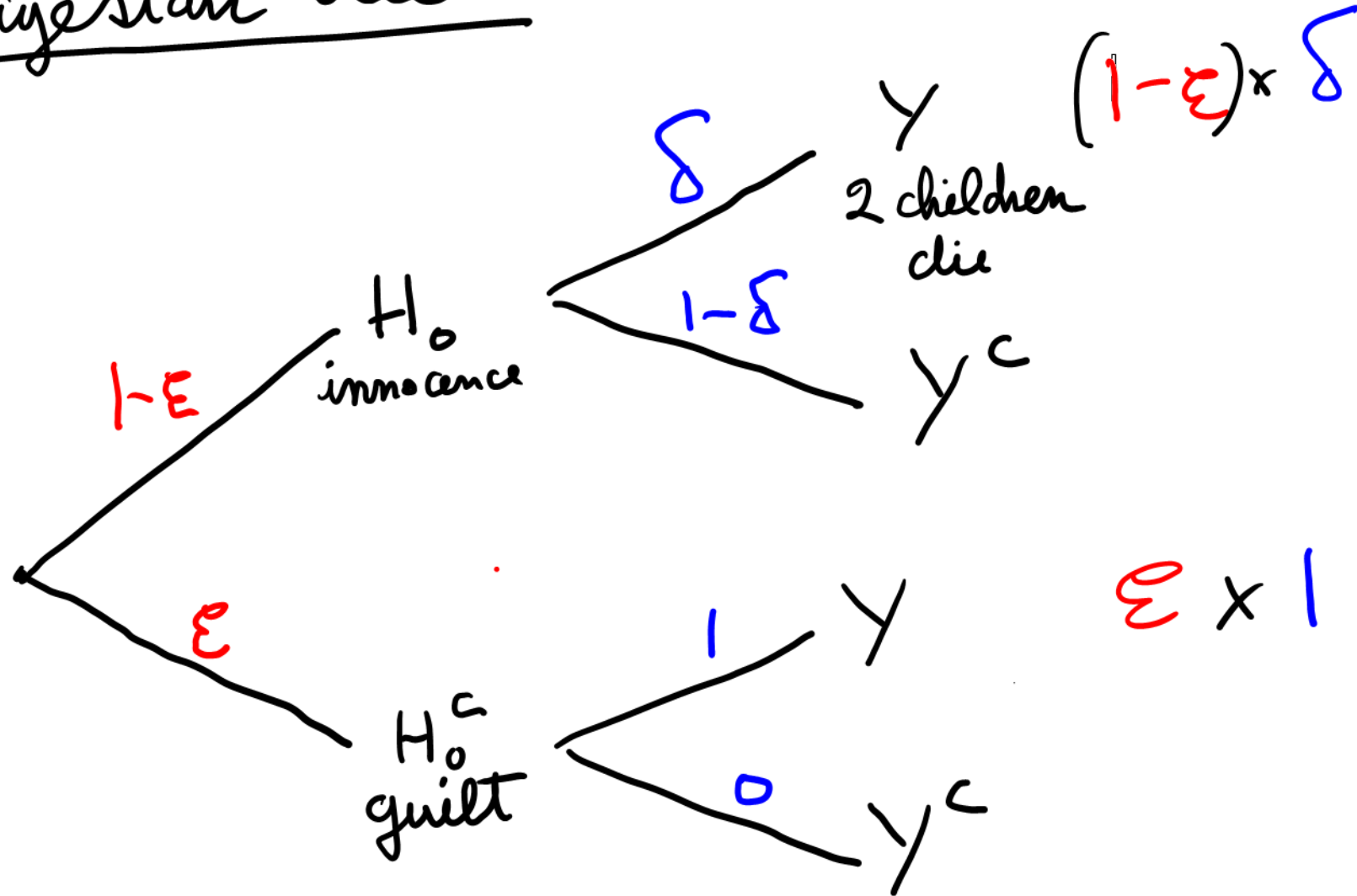
δ = very small number

Bayesian tree:



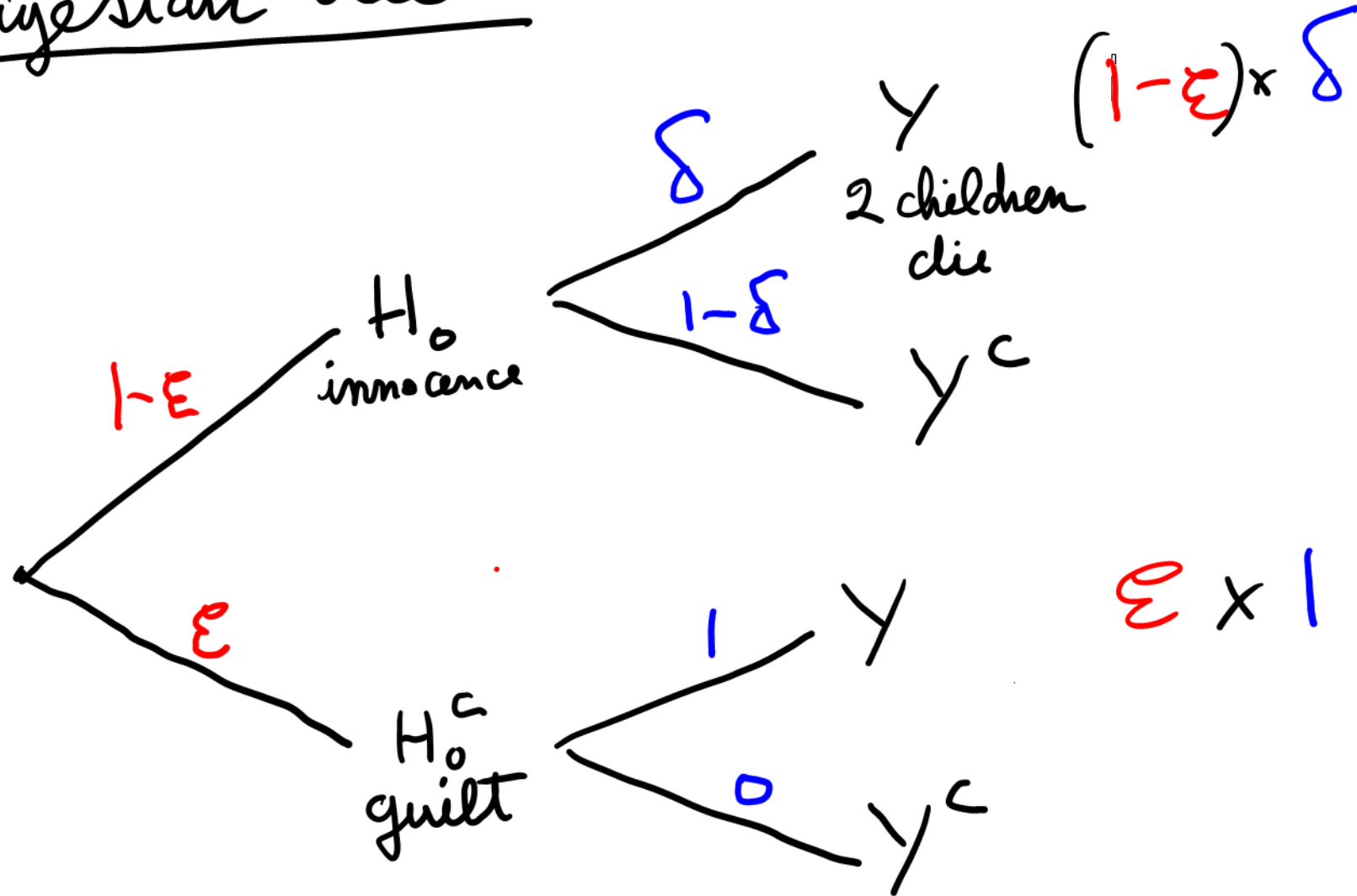
δ = very small number

Bayesian tree:



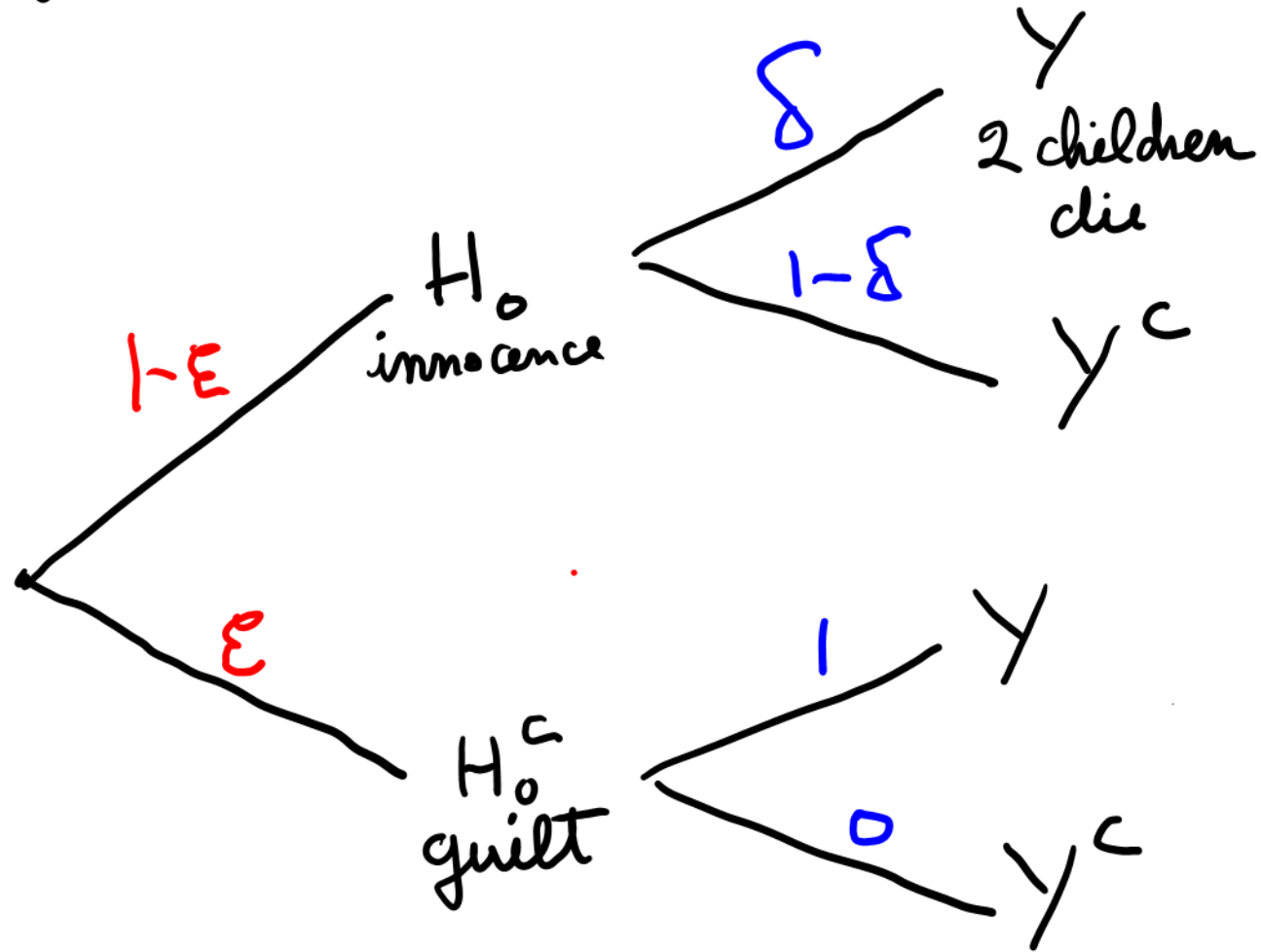
$$P(H_0|Y) = \frac{(1-\epsilon)\delta}{(1-\epsilon)\delta + \epsilon} \approx \frac{\delta}{\delta + \epsilon}$$

Bayesian tree:

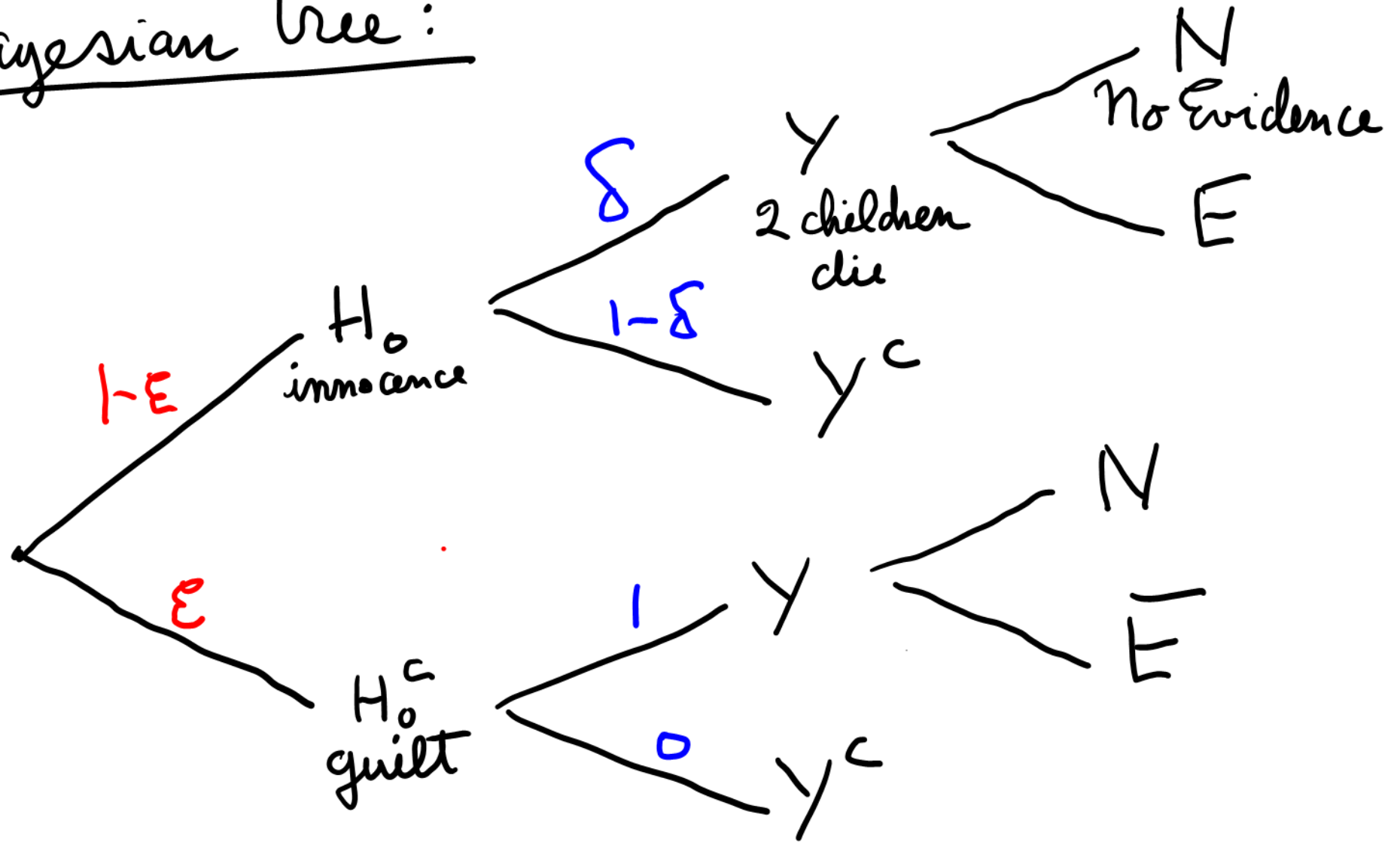


$$P(H_0 | Y) = \frac{(1 - \epsilon) \delta}{(1 - \epsilon) \delta + \epsilon} \approx \frac{\delta}{\delta + \epsilon} \approx \frac{1}{2} ?$$

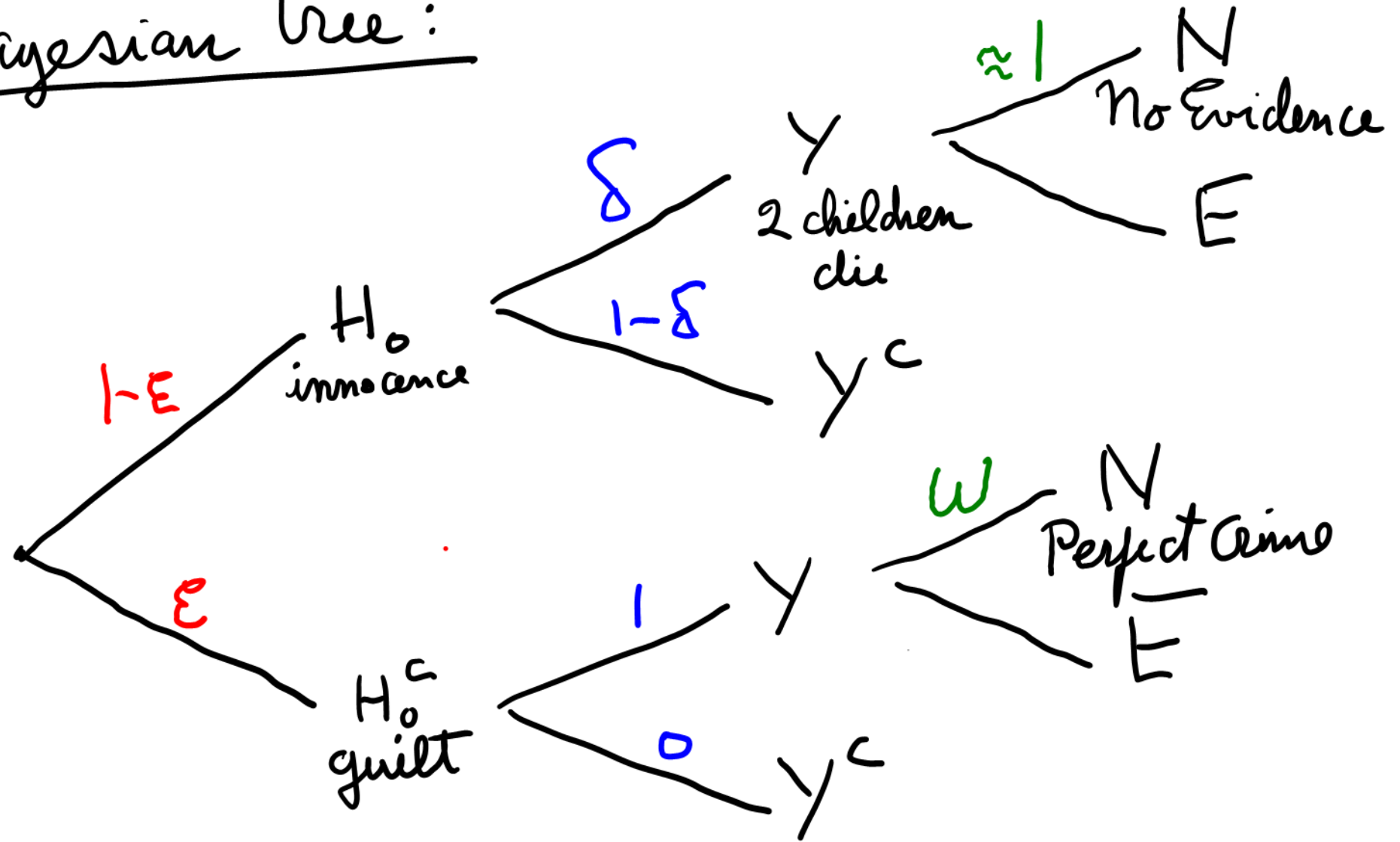
Bayesian tree:



Bayesian tree:

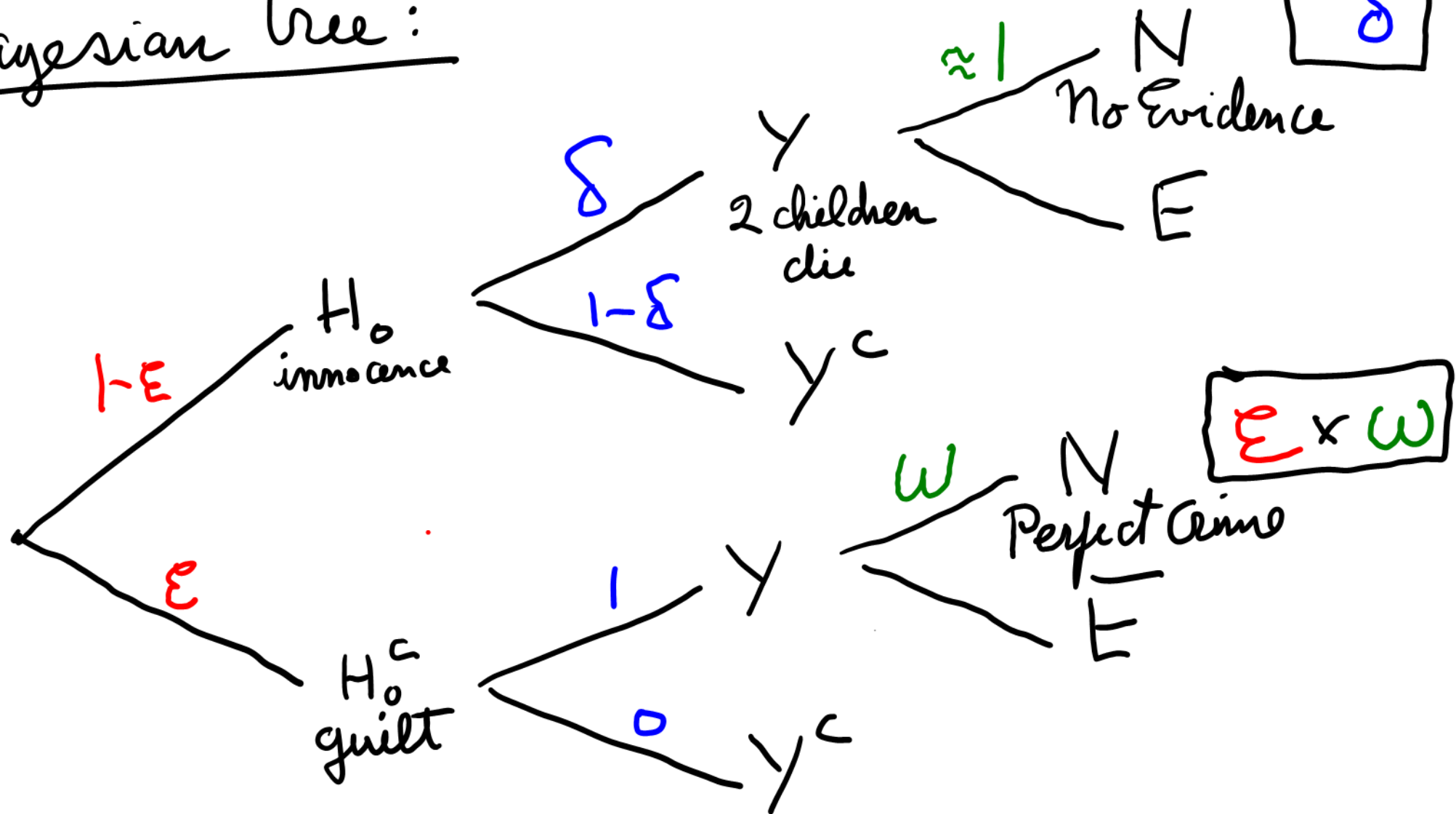


Bayesian tree:

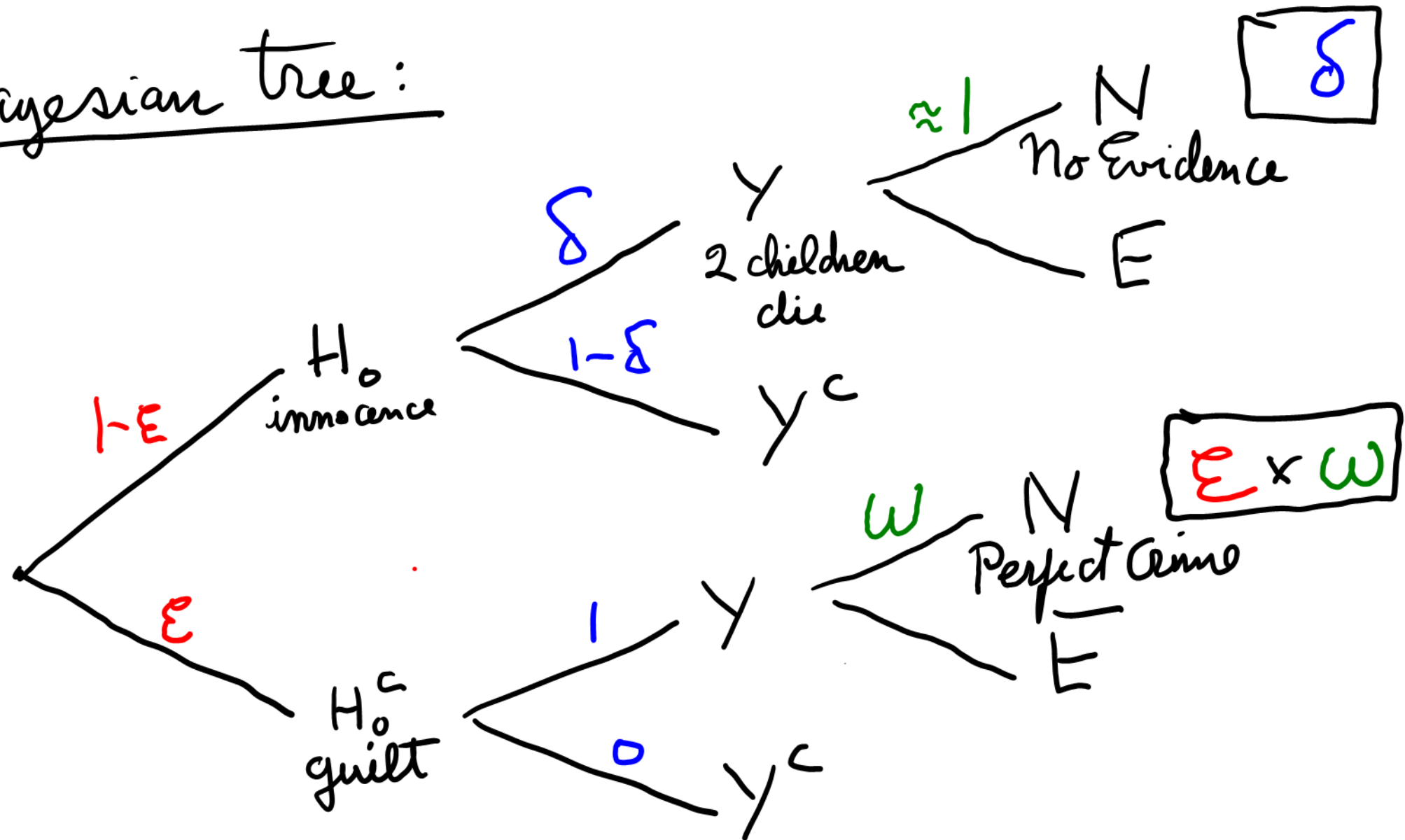


w = another small number

Bayesian tree:

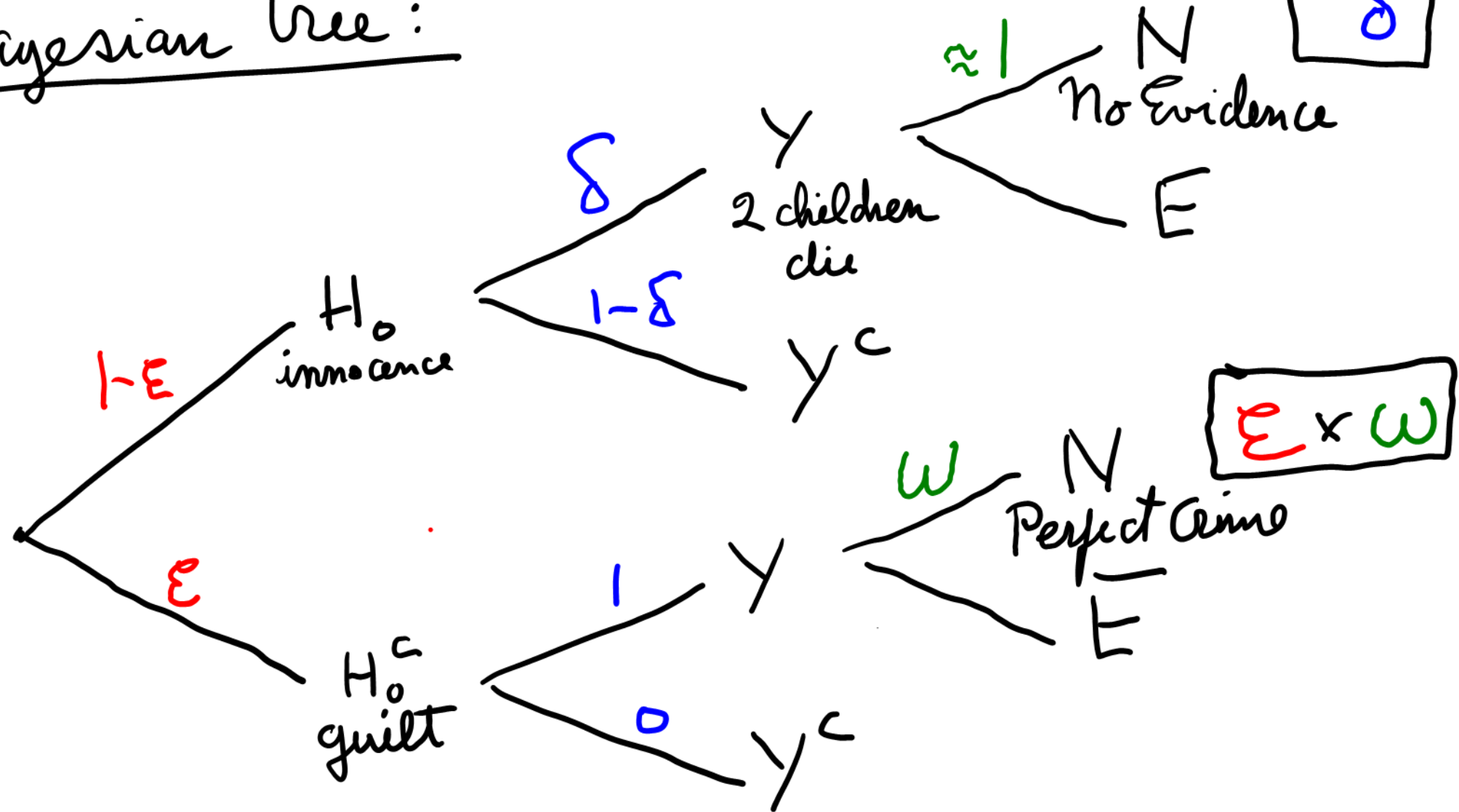


Bayesian tree:



$$P_r(H_0 / Y, N) \approx \frac{\delta}{\delta + E \times \omega}$$

Bayesian tree:



$$P_r(H_0 / Y, N) \approx \frac{\delta}{\delta + E \times \omega} = \text{close to } 1$$

Sally Clark is

innocent .

beyond a reasonable doubt.





GUILTY





GUILTY

INNOCENTE



Proof beyond a reasonable doubt?

Proof beyond a reasonable doubt?

A very small value of the probability of innocence ‘given’ the evidence?

Proof beyond a reasonable doubt?

A very small value of the probability of innocence ‘given’ the evidence?

Probability(Innocence | Evidence)?

Proof beyond a reasonable doubt?

A very small value of the probability of innocence ‘given’ the evidence?

Probability(Innocence | Evidence)?

What did Roy Meadow learn from stats?

How to calculate:

Probability(Evidence⁺ | Innocence)

the *p-value*, the probability of obtaining evidence *as or more contradictory* assuming innocence.

The fundamental neurosis of statistics

- ▶ We really want $p(\theta|y)$ but we'd have to accept $p(\theta)$
- ▶ So we give the world $p(y^+|\theta)$
 - ▶ Most people quietly think it's a proxy for $p(\theta|y)$
 - ▶ if not, what in the world could it be?
- ▶ Gigerenzer:
 - ▶ the confusion created by this unresolved conflict among statisticians, which is both suppressed and inherent in statistics textbooks, leads to a systemic neurosis in science for which the ritual of NHST is a form of conflict resolution – like compulsive hand washing – which makes it resistant to logical arguments
- ▶ One is most strongly committed to the beliefs one does not understand

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

was a poor proxy for

$$P(H_0 | Y)$$

.

If p-values are so bad,
why do we still use
them?

.

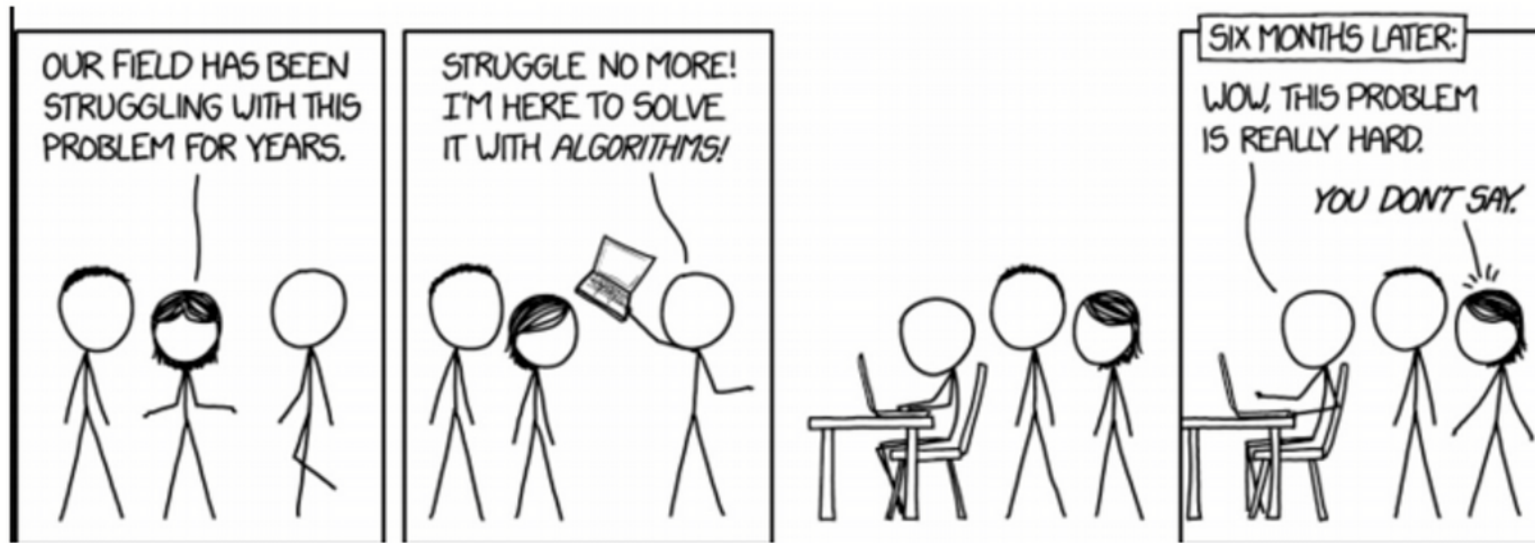
- For many common problems they are consistent with Bayesian answers
 - see Fiducial and Structural Inference

- For many common problems they are consistent with Bayesian answers
 - see Fiducial and Structural Inference
- Bayesian Inference except for very simple problems can be very difficult

- For many common problems they are consistent with Bayesian answers
 - see Fiducial and Structural Inference
- Bayesian Inference except for very simple problems can be very difficult
 - This is changing thanks to MCMC
- You don't need to justify a choice of priors

- For many common problems they are consistent with Bayesian answers
 - see Fiducial and Structural Inference
- Bayesian Inference except for very simple problems can be very difficult
 - This is changing thanks to MCMC
- You don't need to justify a choice of priors
- Fisher finally cautioned to use p-value only if there is little other information on H_0

If you feel puzzled, you are not alone: (Reid, 2017)

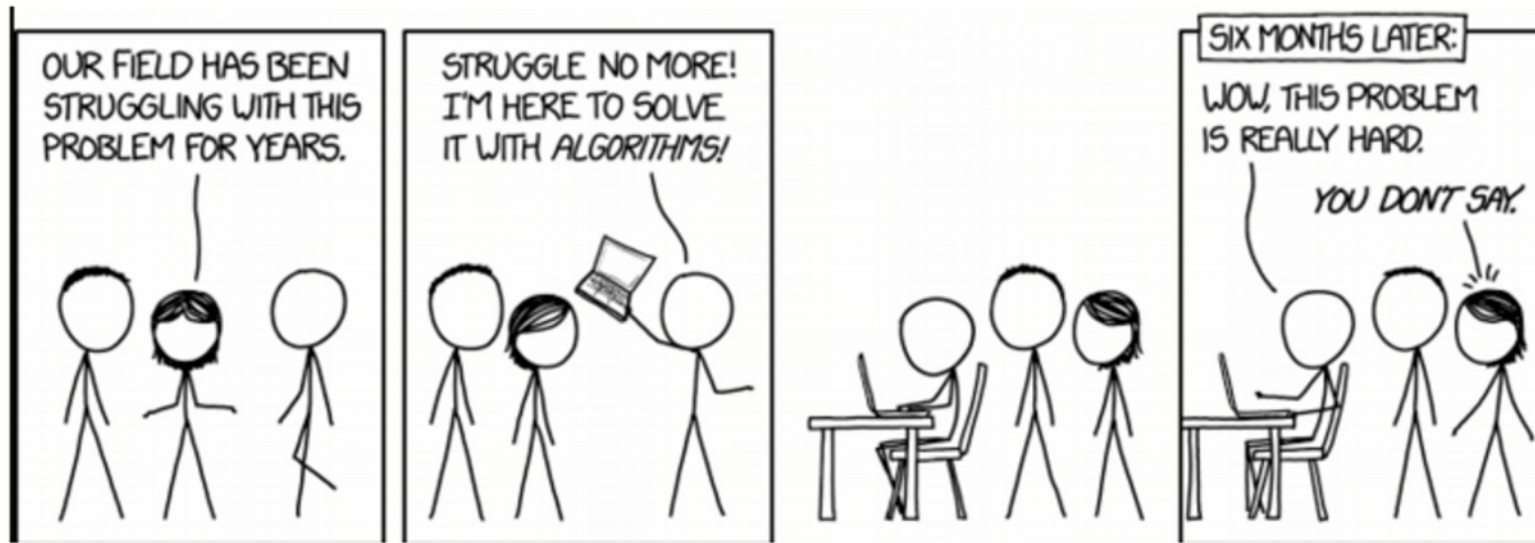


From a 1996 interview:

Nancy Reid: Why is conditional inference so hard?

Sir David Cox: I expect we're all missing something but I don't know what it is.

If you feel puzzled, you are not alone: (Reid, 2017)



From a 1996 interview:

Nancy Reid: Why is conditional inference so hard?

Sir David Cox: I expect we're all missing something but I don't know what it is.

— cited in 2017

Frequentist vs. Bayesian Workflow

Frequentist workflow

Step 1:

Formulate hypotheses and plan comparisons and estimates

Step 3:

1. Observe y
2. Do something clever with $p(y|\theta)$
3. Estimate θ in a way that works well on average – i.e. if you repeat the process and get more y 's

Step 2:

Model: $p(y|\theta)$

	y_1	y_2	sum
θ_1	$p(y_1 \theta_1)$	$p(y_2 \theta_1)$	1
θ_2	$p(y_1 \theta_2)$	$p(y_2 \theta_2)$	1
θ_3	$p(y_1 \theta_3)$	$p(y_2 \theta_3)$	1

Step 4:

Insist how important it is that your results not be confused with $p(\theta|y)$ because that would require a subjective prior and you believe that science should be objective.

Bayesian workflow

Step 1: Prior: $p(\theta)$

Formulate a prior on some basis

θ_1	$p(\theta_1)$
θ_2	$p(\theta_2)$
θ_3	$p(\theta_3)$
sum	1(or ∞ !)

Step 3: Observed joint:

Observe y_{obs}

$$p(y_{obs}, \theta) = p(\theta) \times p(y_{obs}|\theta)$$

θ_1	$p(y_{obs}, \theta_1)$
θ_2	$p(y_{obs}, \theta_2)$
θ_3	$p(y_{obs}, \theta_3)$
sum	$p(y_{obs})$ or c or ∞

Step 2: Model: $p(y|\theta)$

	y_1	y_2	sum
θ_1	$p(y_1 \theta_1)$	$p(y_2 \theta_1)$	1
θ_2	$p(y_1 \theta_2)$	$p(y_2 \theta_2)$	1
θ_3	$p(y_1 \theta_3)$	$p(y_1 \theta_3)$	1

Step 4: Posterior:

$$p(\theta|y_{obs}) = p(y_{obs}, \theta)/p(y_{obs})$$

θ_1	$p(\theta_1 y_{obs})$
θ_2	$p(\theta_2 y_{obs})$
θ_3	$p(\theta_3 y_{obs})$
sum	1 or ?

2 major problems with BI

- 1) Philosophical (or psychological?)
- 2) Practical

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 $P(\theta)$

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

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

Philosophical Basic problem

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

To get $P(\theta|X)$

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

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

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

Philosophical Basic problem

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

To get $P(\theta|X)$

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

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

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

- 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 $P(\theta)$

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



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.

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

$[\Theta | X]$ - 1800s

$[X | \Theta]$

$[X, \Theta]$

$[\Theta | X]$

Fisher
1920



HOMO
APRIORIUS



HOMO
PRAGMATICUS



HOMO
FREQUENTISTUS



HOMO
SAPIENS



HOMO
BAYESIANIS

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

$[H|X]$ - 1800s

$[X|H]$

$[X, H]$

$[H|X]$

Fisher
1920

1950+

MCMC

- Metropolis
- Hastings
- Ulam



HOMO
APRIORIUS



HOMO
PRAGMATICUS



HOMO
FREQUENTISTUS



HOMO
SAPIENS



HOMO
BAYESIANIS

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

$[\Theta | X]$ - 1800s

$[X | \Theta]$

$[X, \Theta]$

$[\Theta | X]$

Fisher
1920

1950+

MCMC

- Metropolis

- Hastings

- Ulam

1987+

HMC



HOMO
APRIORIUS



HOMO
PRAGMATICUS



HOMO
FREQUENTISTUS



HOMO
SAPIENS



HOMO
BAYESIANUS

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

Use proper weakly informative
priors

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

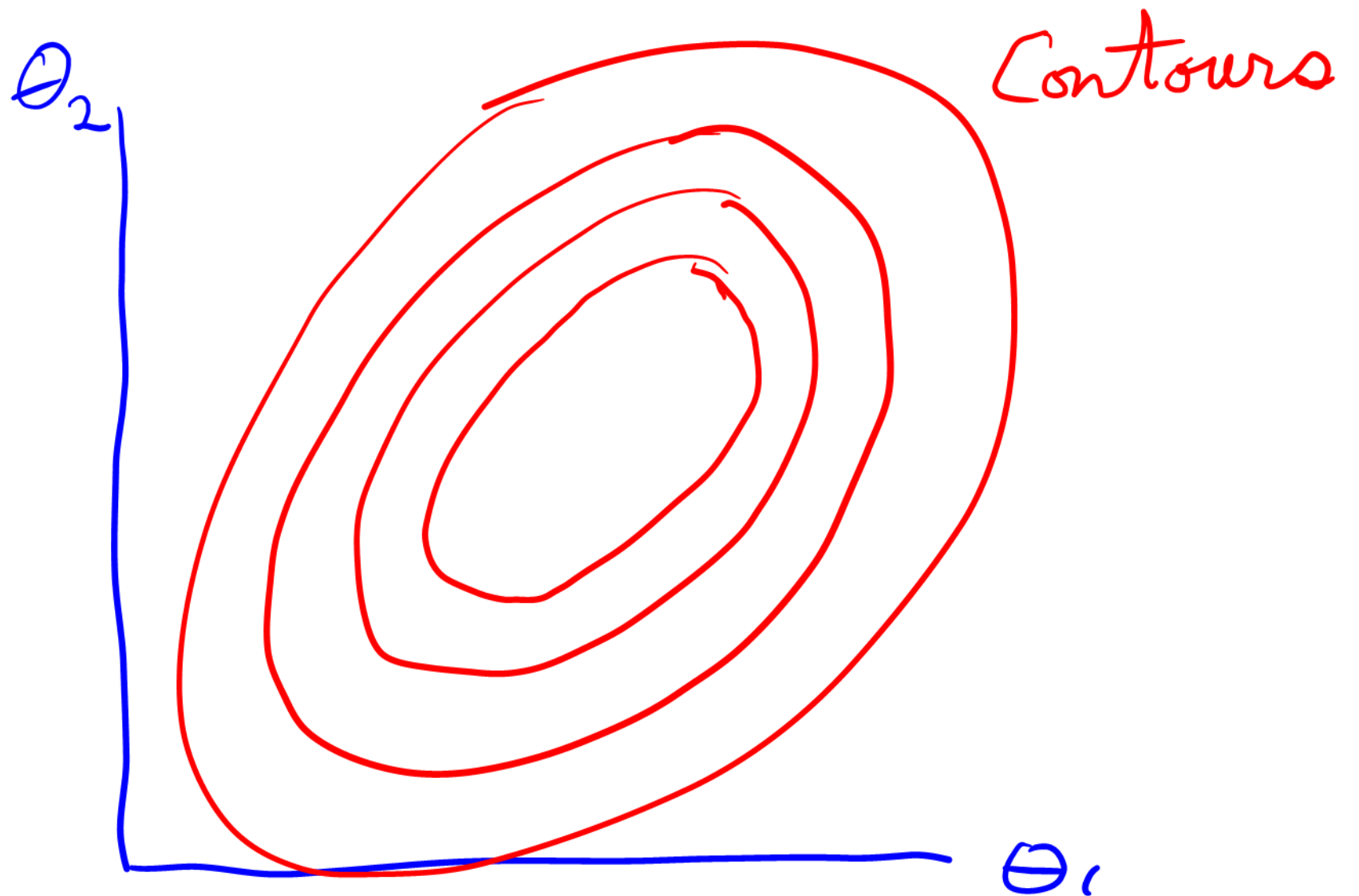
joint model $\times$ prior

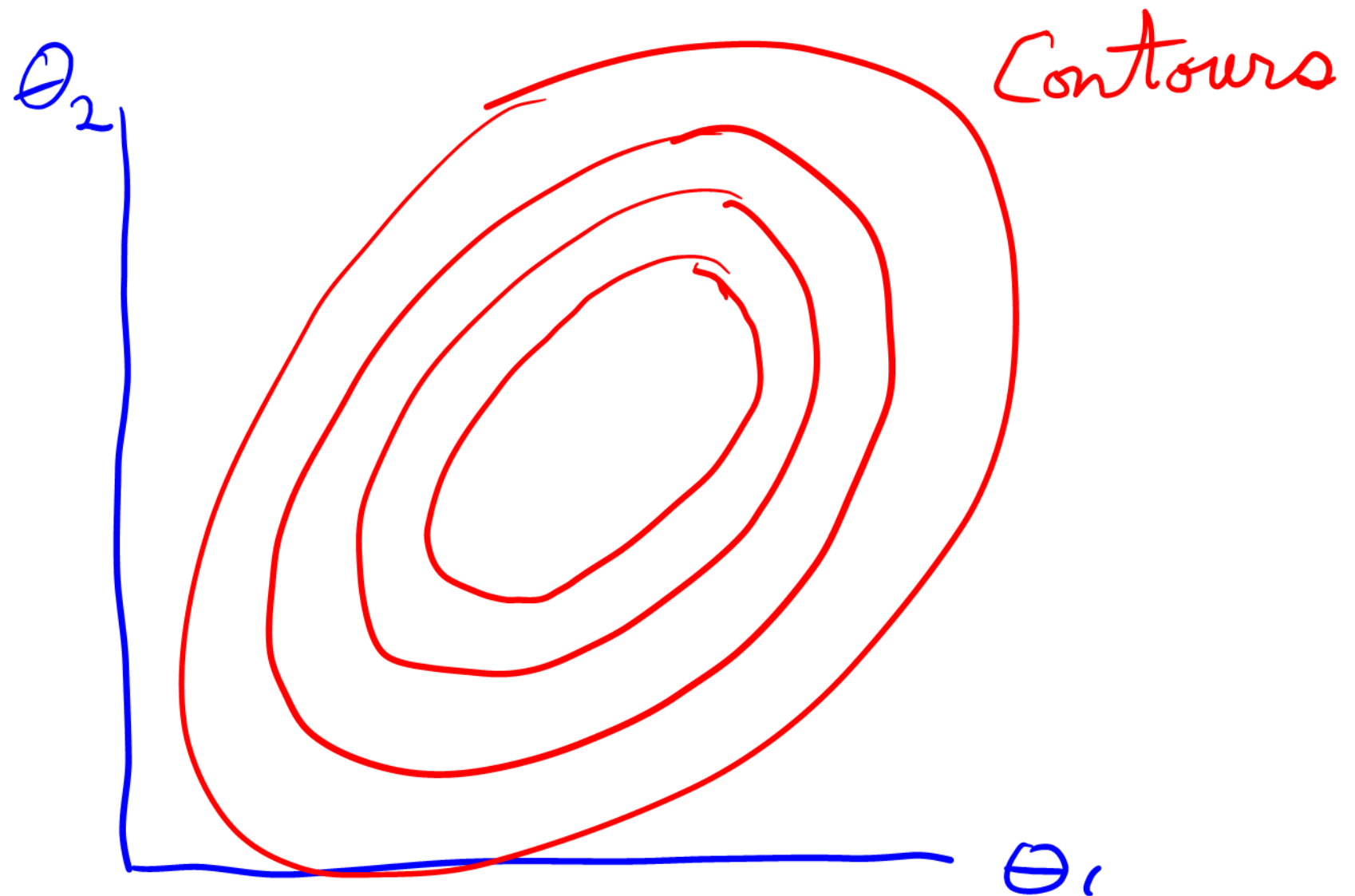
Samples from $P(\theta|x)$ using only $P(\theta, x)$
i.e. no need to find elusive $P(x)$

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

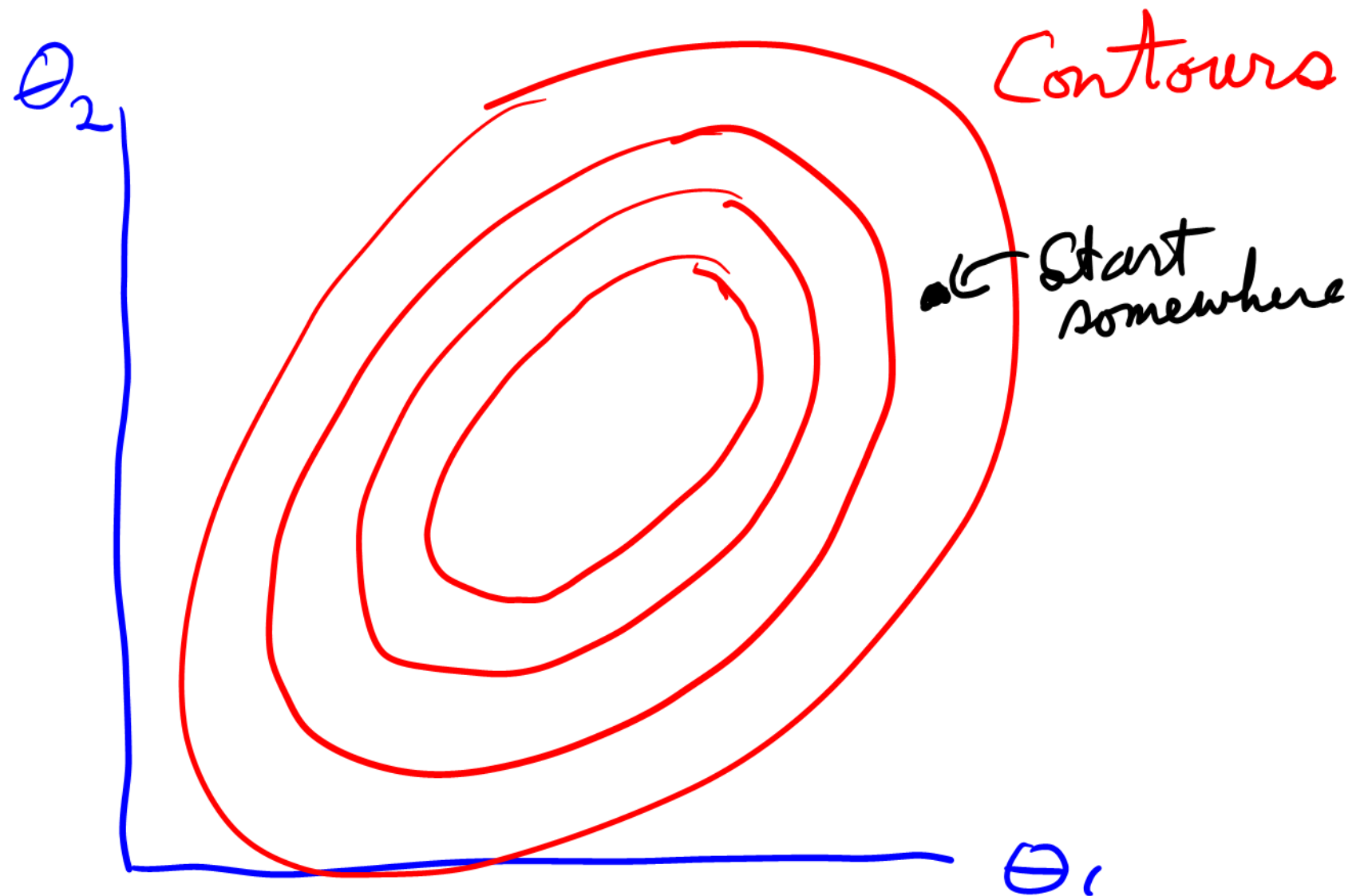
$P(\theta, x)$ $\uparrow$ fixed



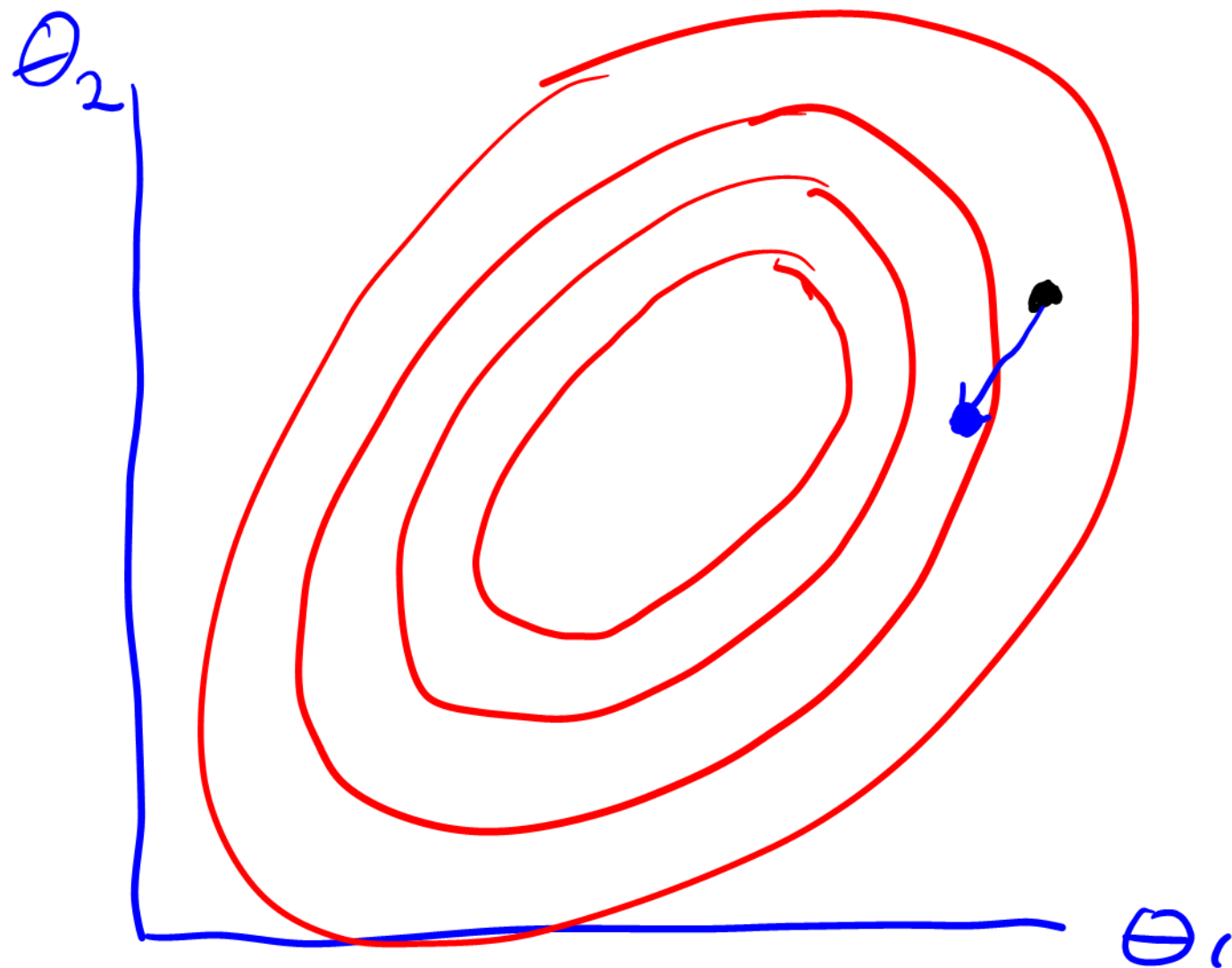


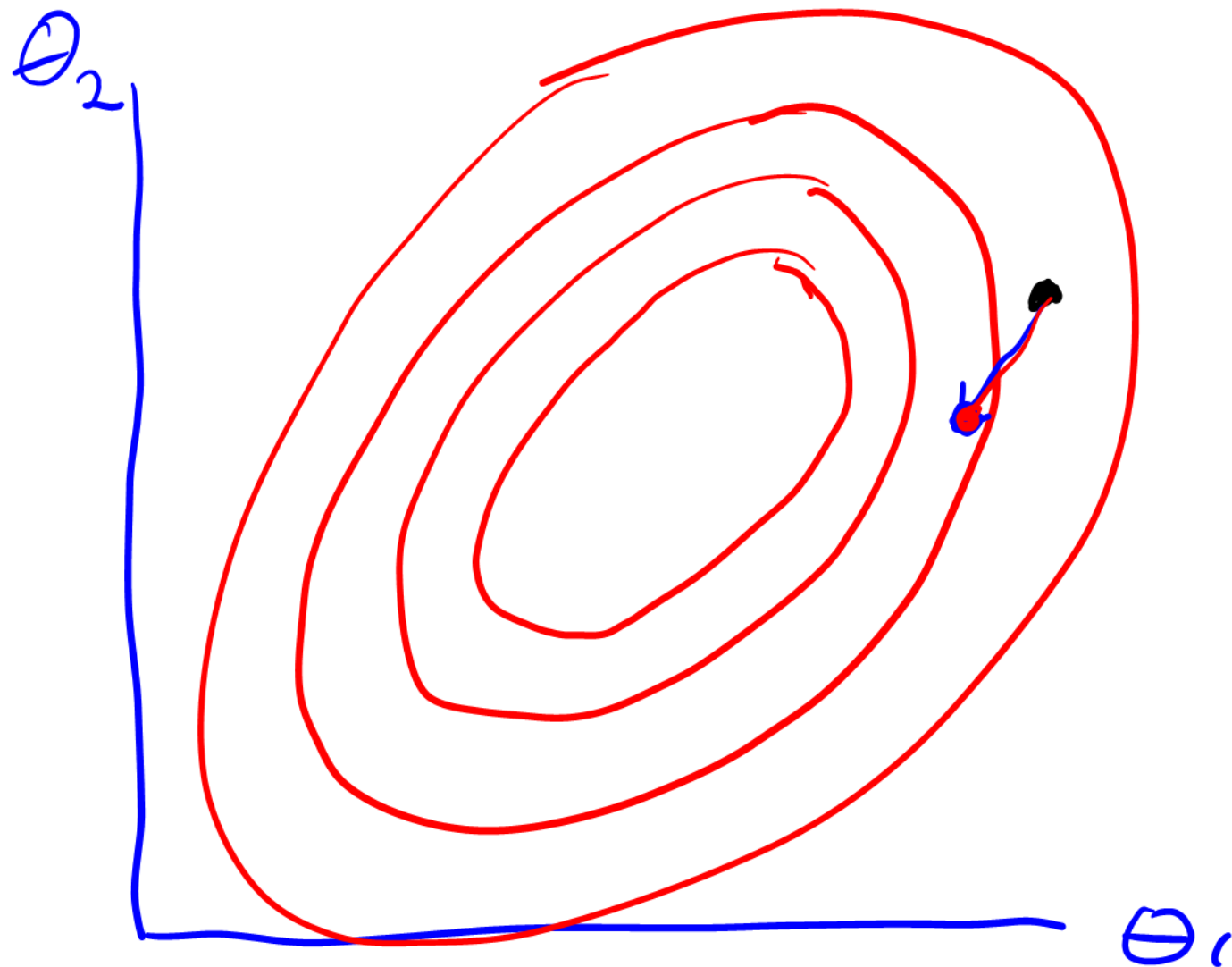


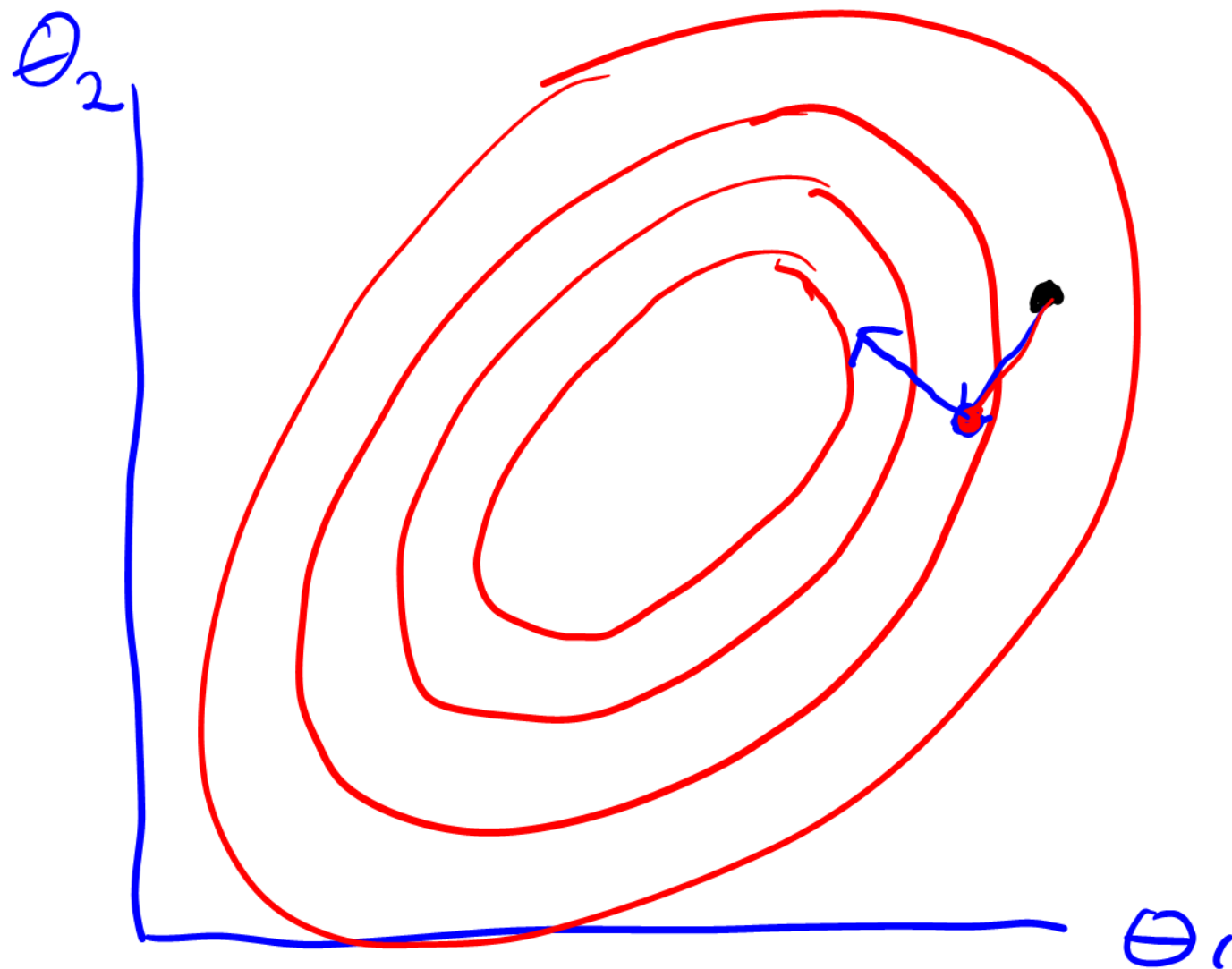
Metropolis-Hastings algorithm

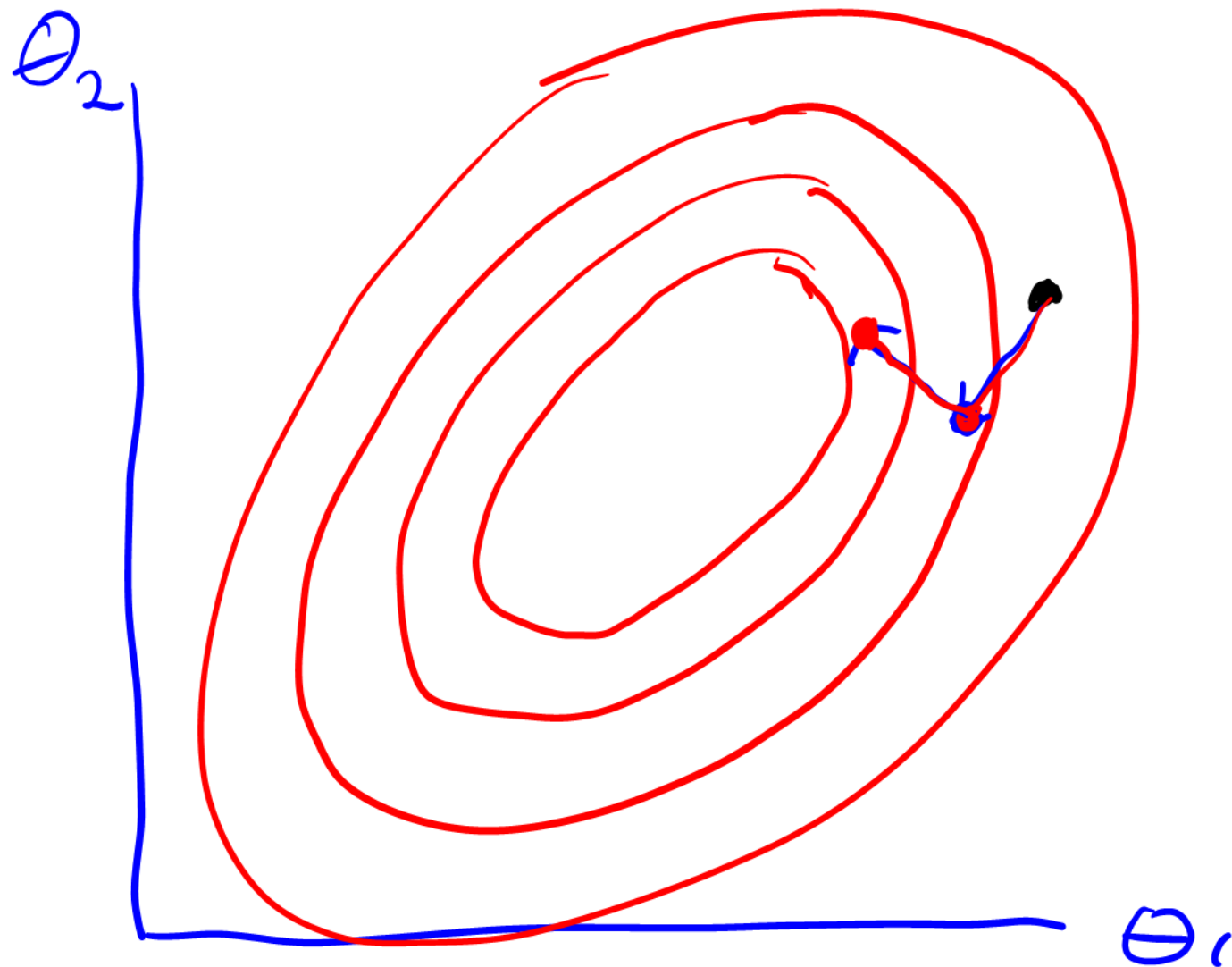


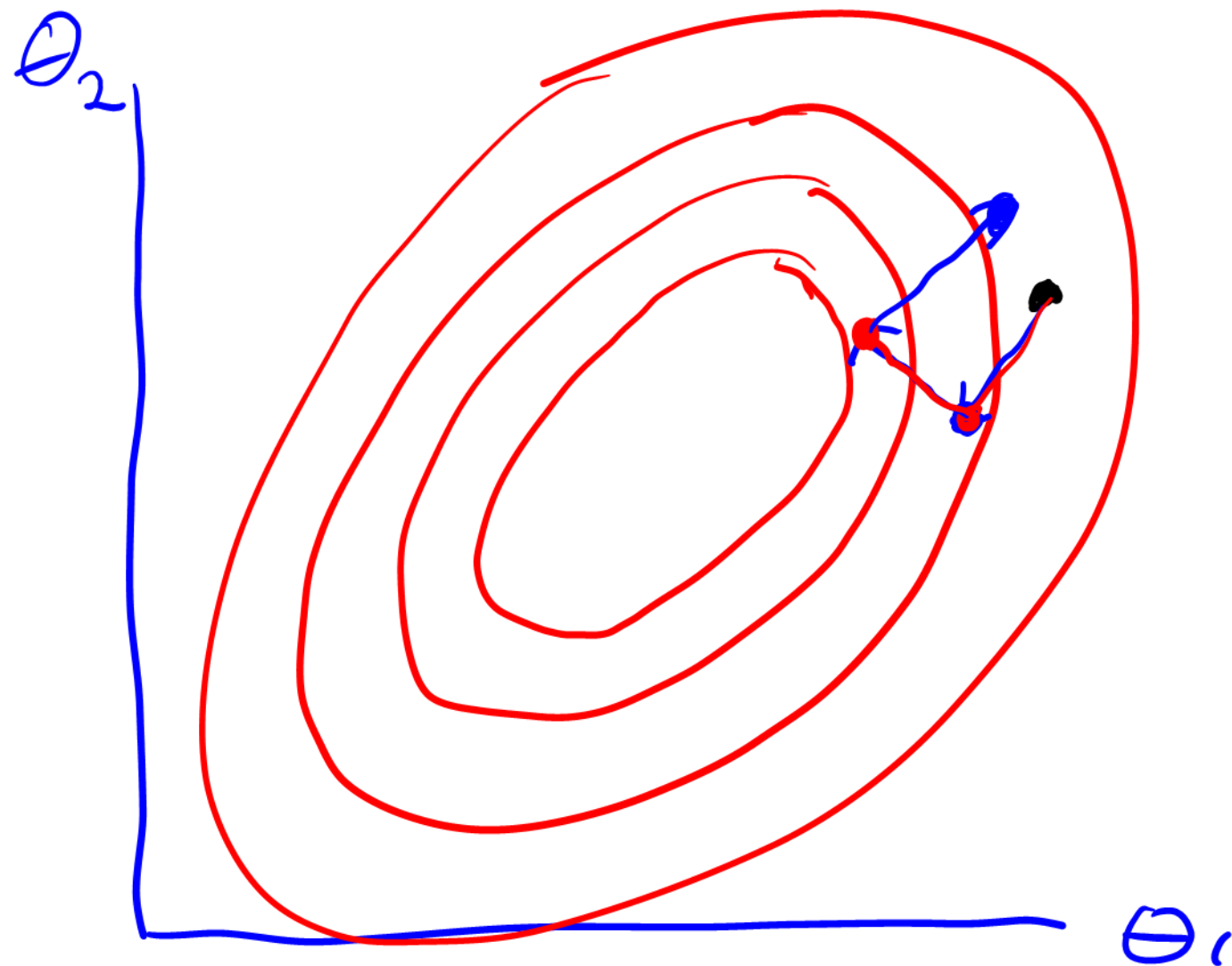
Metropolis-Hastings algorithm











If you've walked downhill, you need to toss a biased coin with:

$$Pr(Heads) = \frac{p(\theta_{new}|Y)}{p(\theta_{last}|Y)}$$

Usually, it's very hard to compute the **numerator** and the **denominator** of this **ratio**.

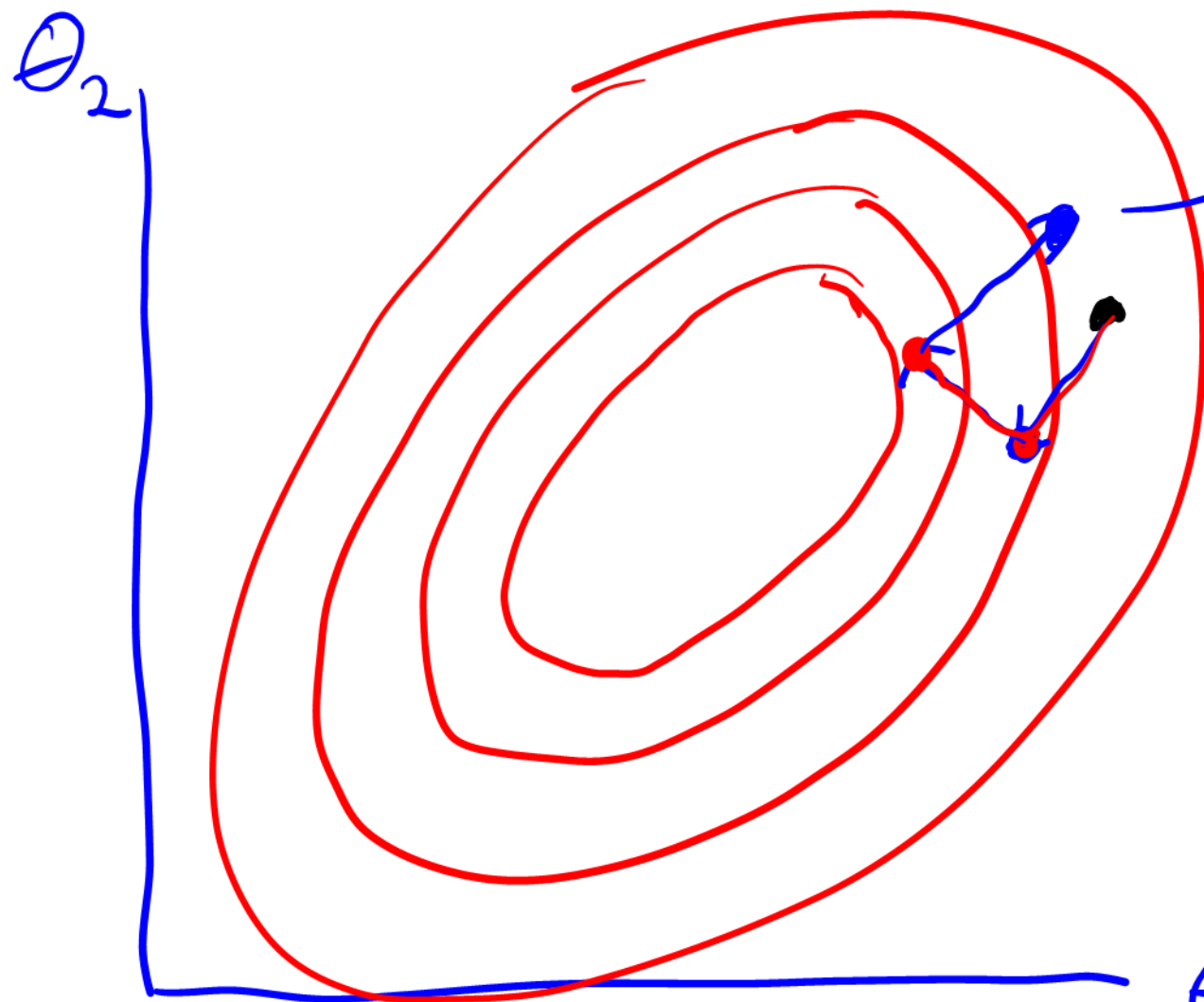
However, the **ratio** itself is, for many models, a relative cinch:

$$Pr(Heads) = \frac{p(\theta_{new}|Y)}{p(\theta_{last}|Y)} = \frac{p(\theta_{new}|Y)/p(Y)}{p(\theta_{last}|Y)/p(Y)} = \frac{p(Y, \theta_{new})}{p(Y, \theta_{last})} \times \frac{p(\theta_{new})}{p(\theta_{last})}$$

Which is just the **likelihood ratio** times the **prior ratio**.

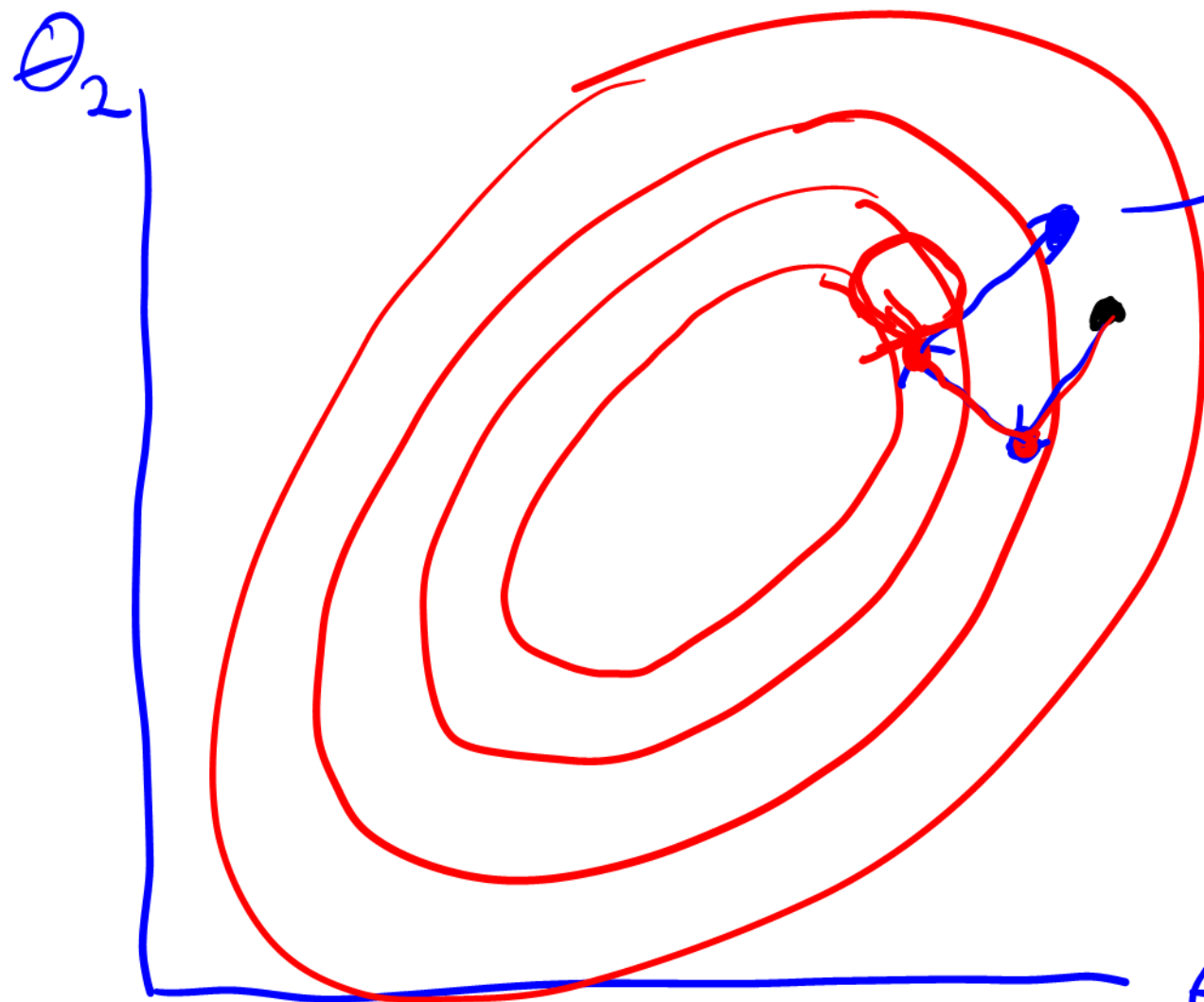
The more you've gone down, the lower the probability of a Head.

- If you get a Head, plant a stake at your new position.
- If you get a Tail, step back to the last position and plant a second stake there.



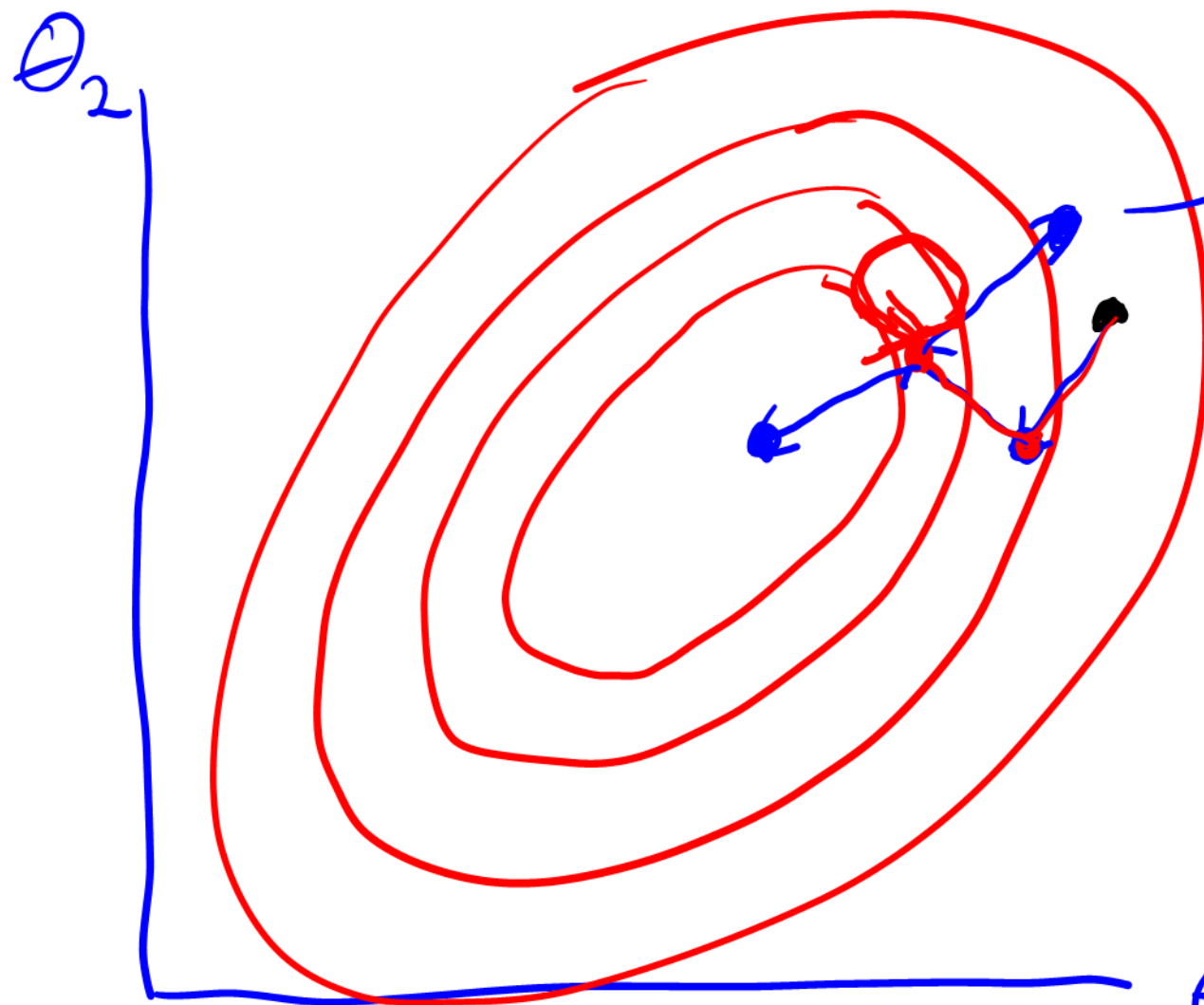
Toss a
biased
coin
with
 $p(H) = \frac{p(\bullet)}{p(\bullet)}$

If Heads
move
 θ_1 If Tails
stay



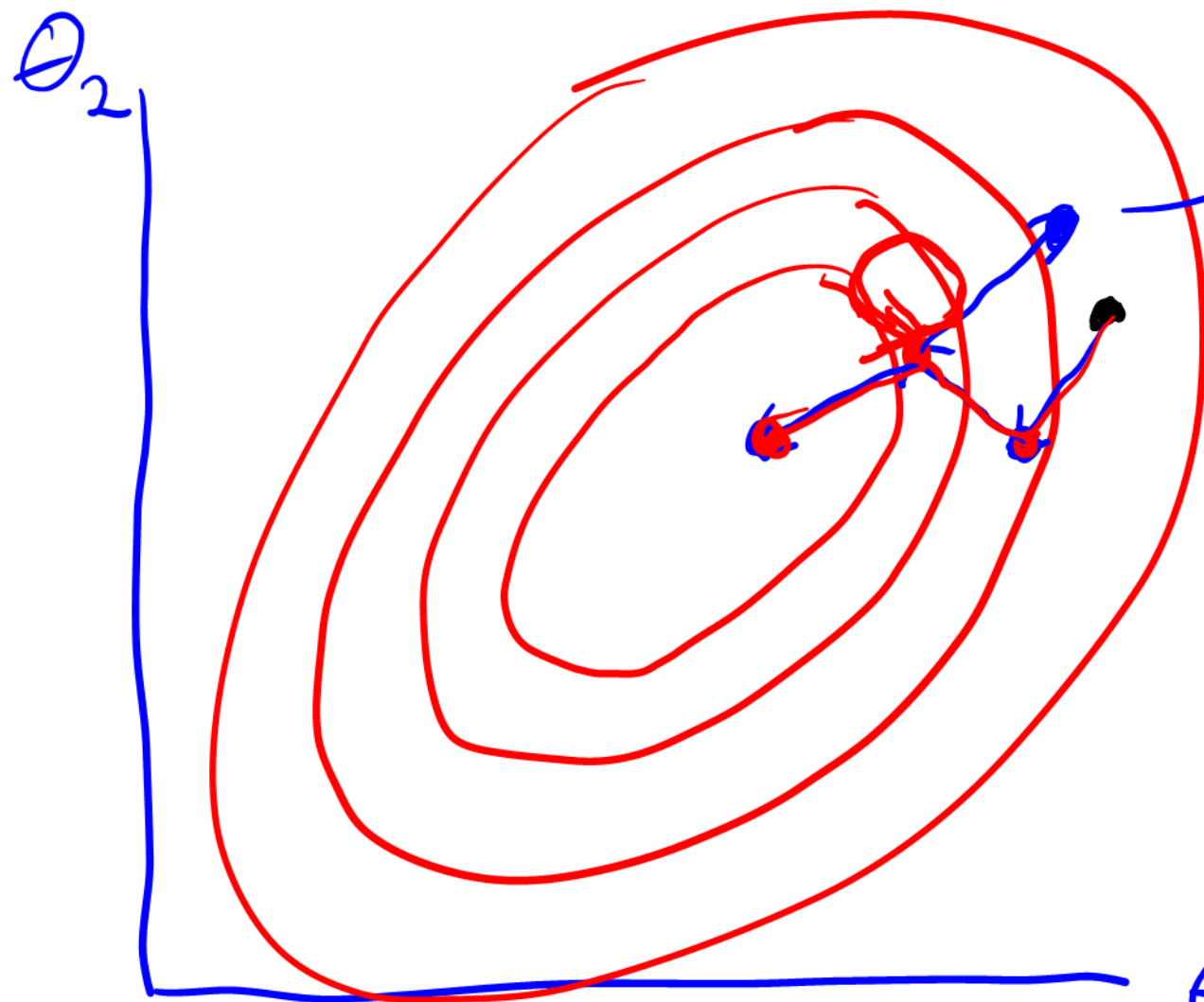
Toss a biased coin with
 $p(H) = \frac{p(\cdot)}{p(\cdot)}$

If Heads move θ_1 , if Tails stay



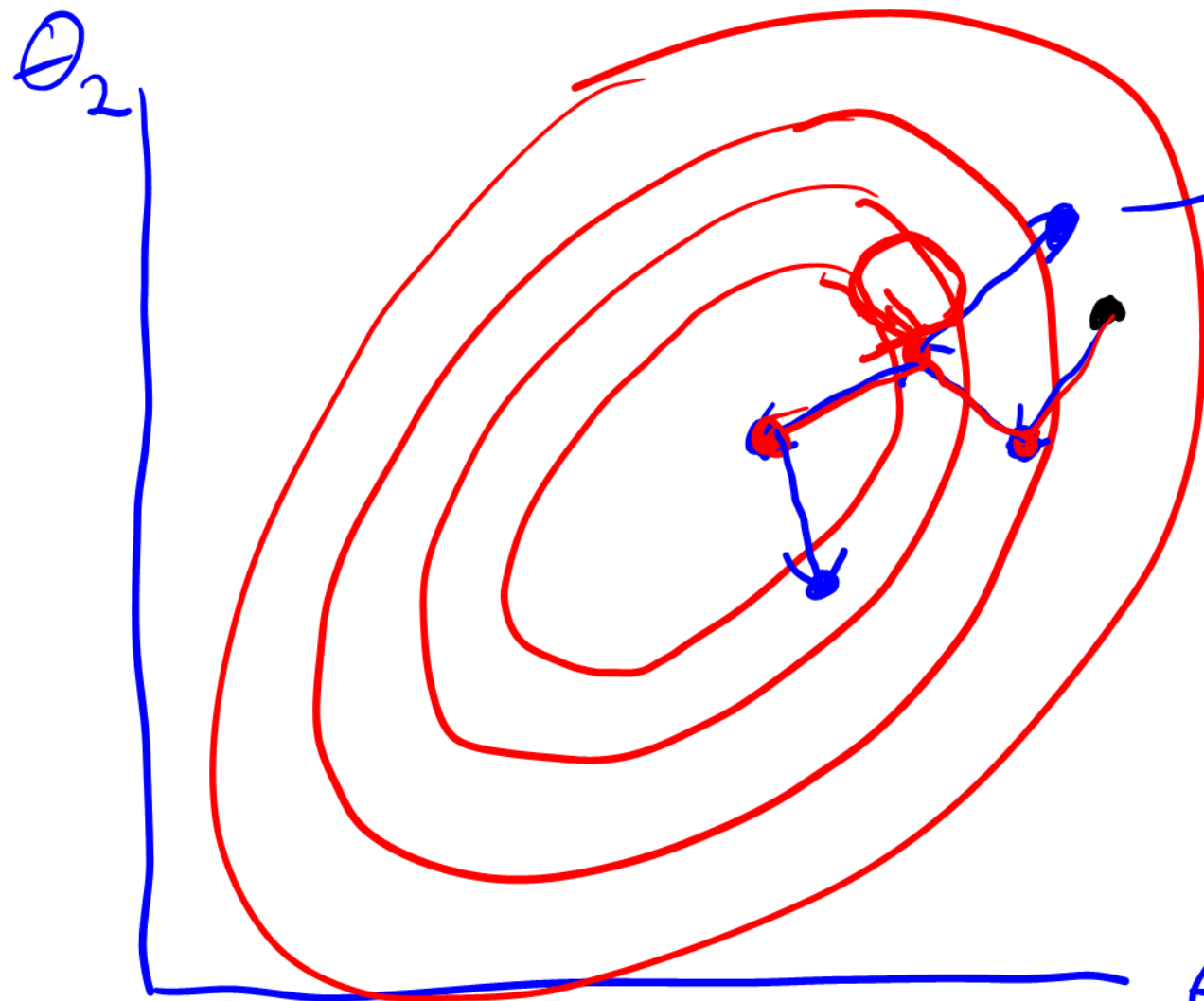
Toss a biased coin with $p(H) = \frac{p(\cdot)}{p(\cdot)}$

If Heads move θ_1
If Tails stay



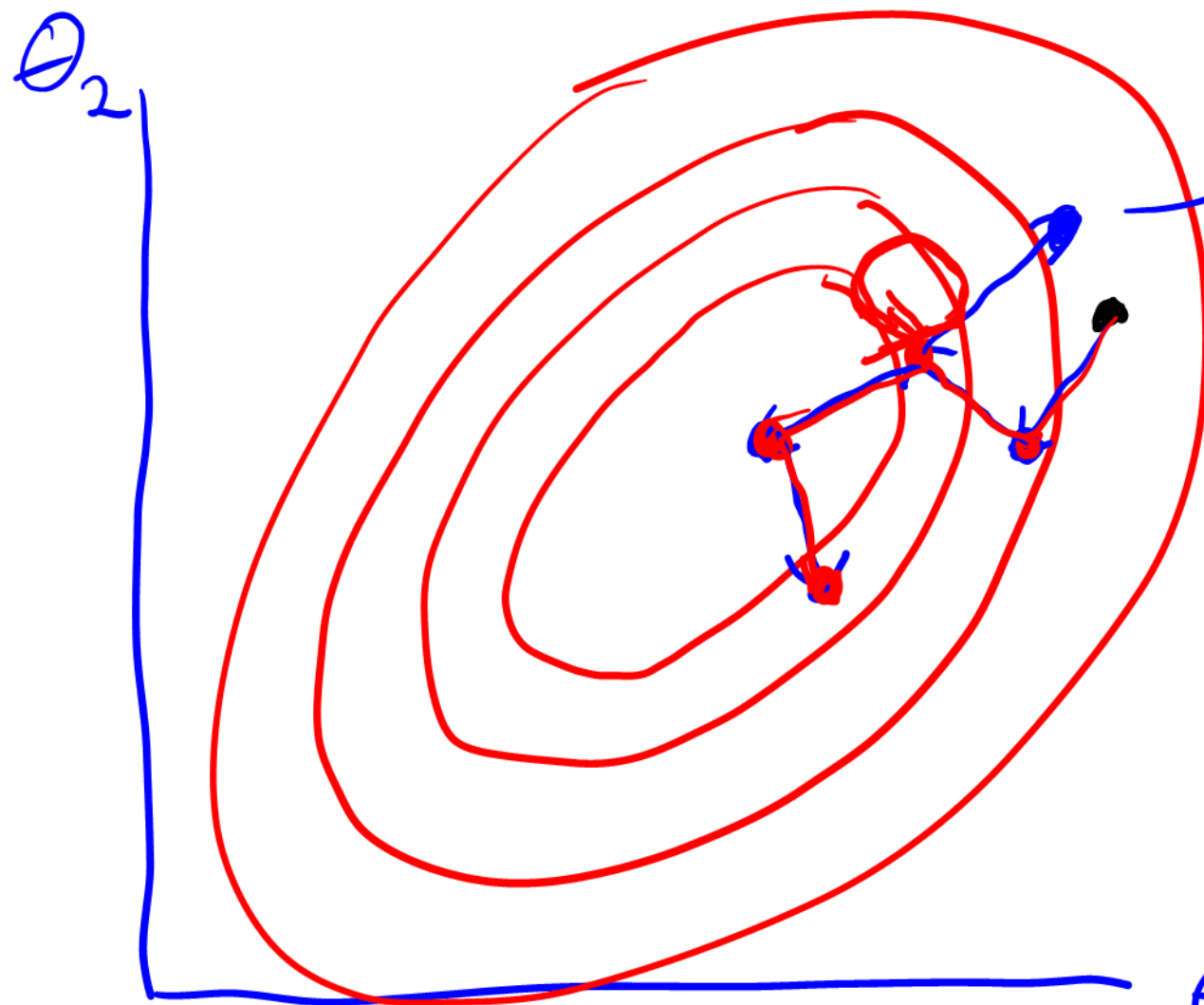
Toss a
biased
coin
with
 $p(H) = \frac{p(\cdot)}{p(\cdot)}$

If Heads
move
 θ_1 If Tails
stay



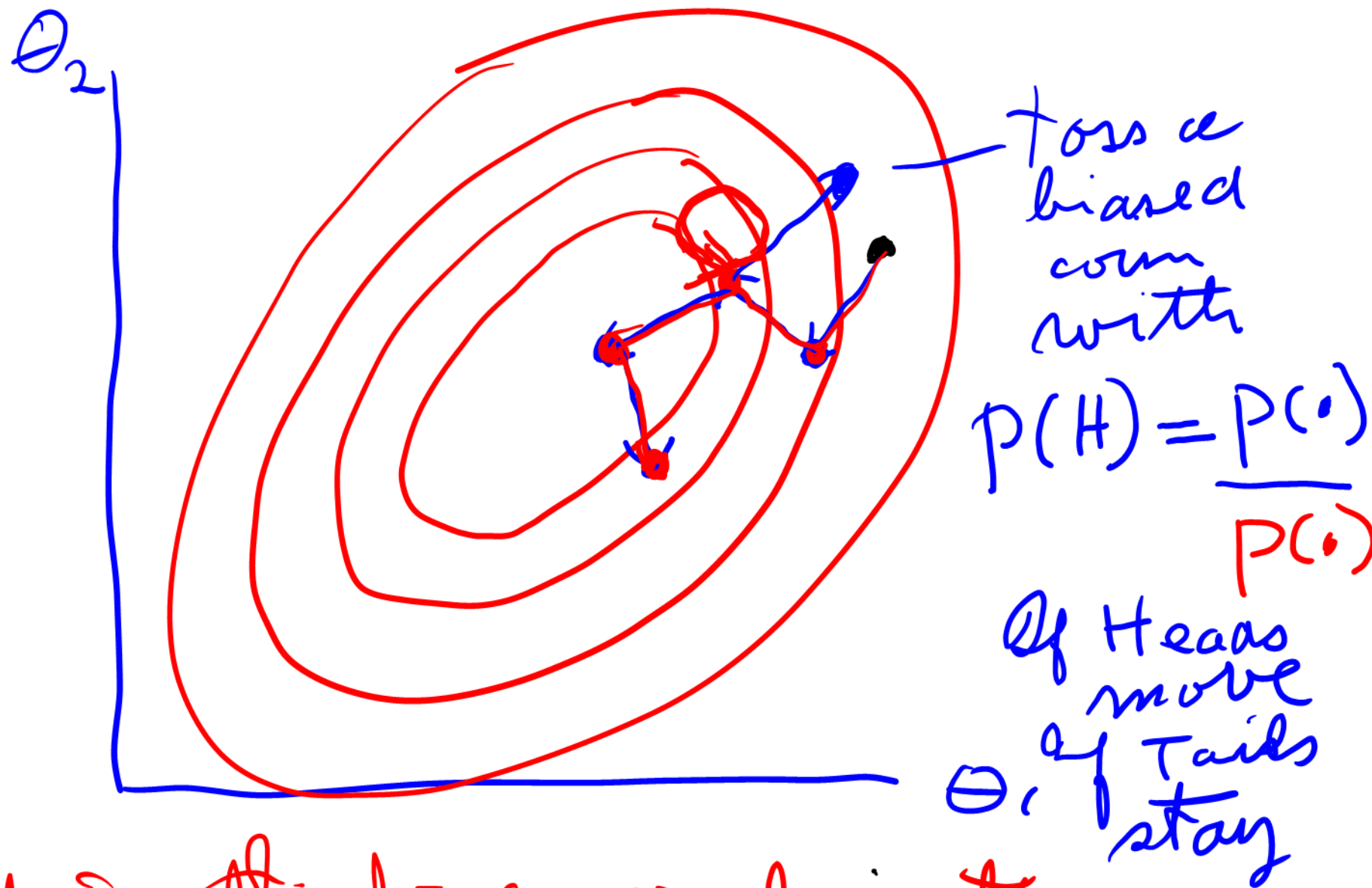
Toss a biased coin with $p(H) = \frac{p(\bullet)}{p(\bullet)}$

If Heads move θ_1
If Tails stay

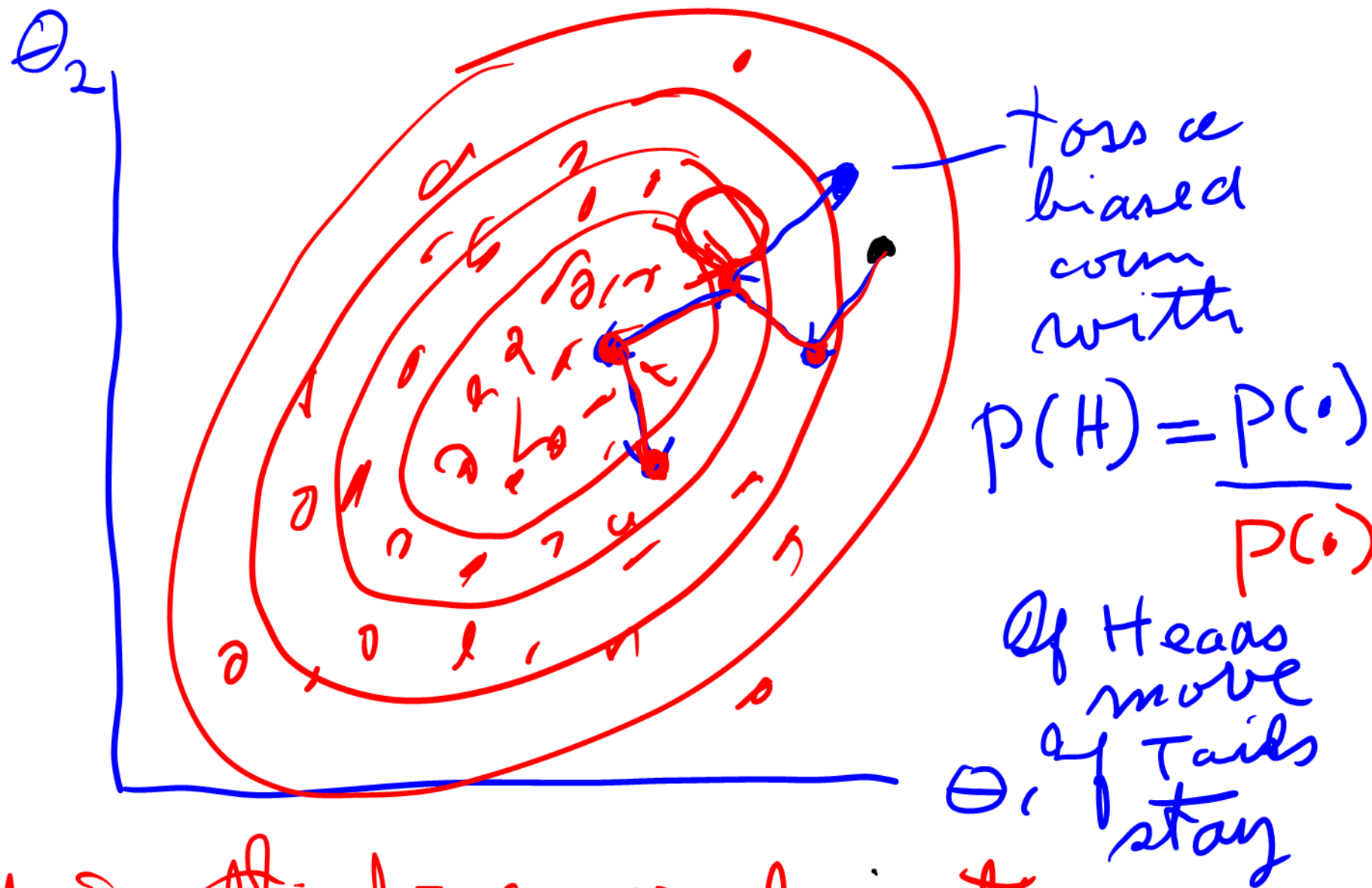


Toss a
biased
coin
with
 $p(H) = \frac{p(\bullet)}{p(\bullet)}$

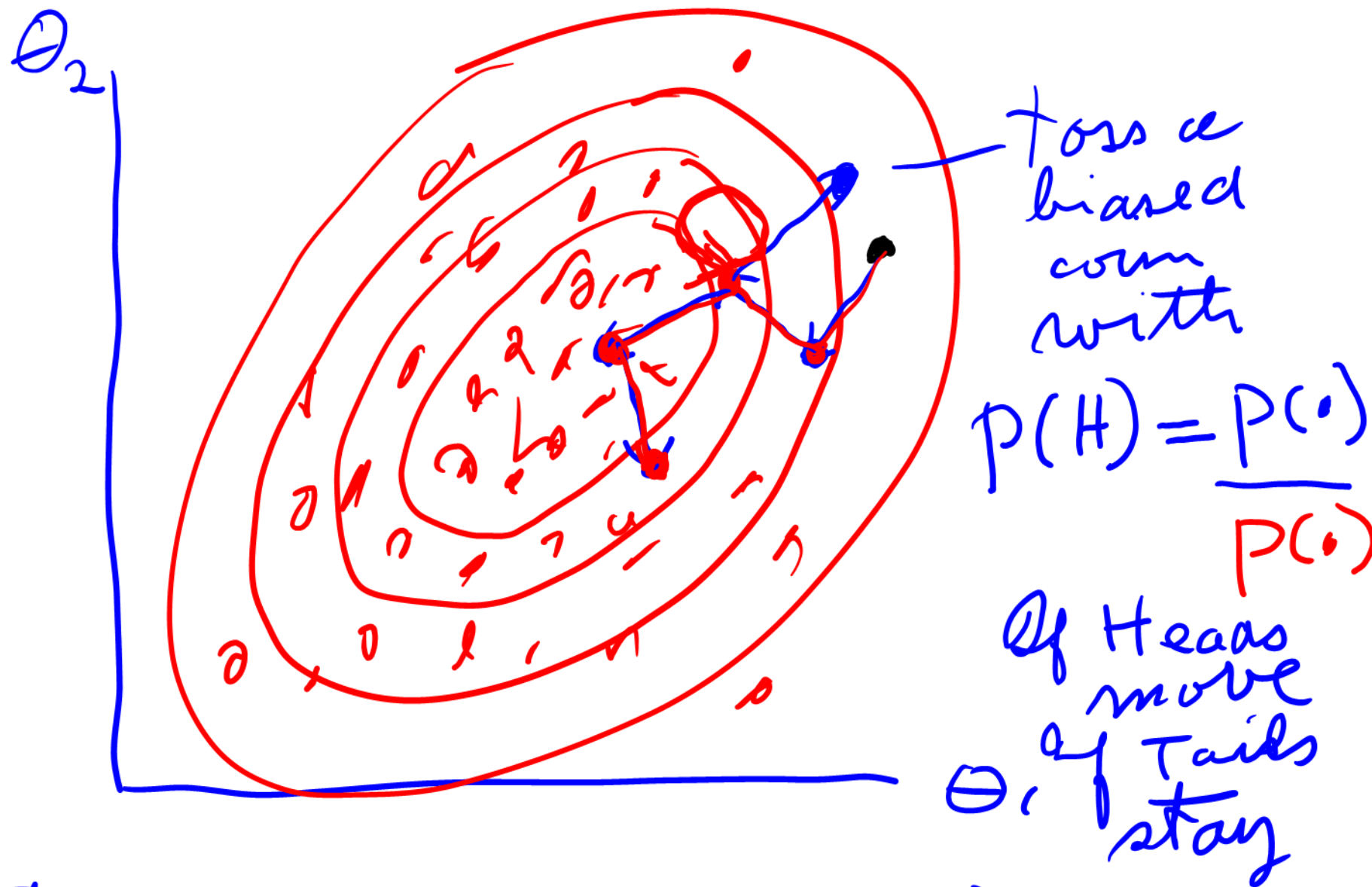
If Heads
move
 θ_1 If Tails
stay



Keep doing this for a very long time



Keep doing this for a very long time



— Generates a sample from $P(\theta | x)$

That's the M-H algorithm

- points can be highly correlated
- very slow to cover distribution
 - especially with large # of parameters, e.g. multilevel models



- can get stuck in corners of distribution

$$O(p^2)$$

$$O(p^2)$$

$$O(p^{5/4})$$

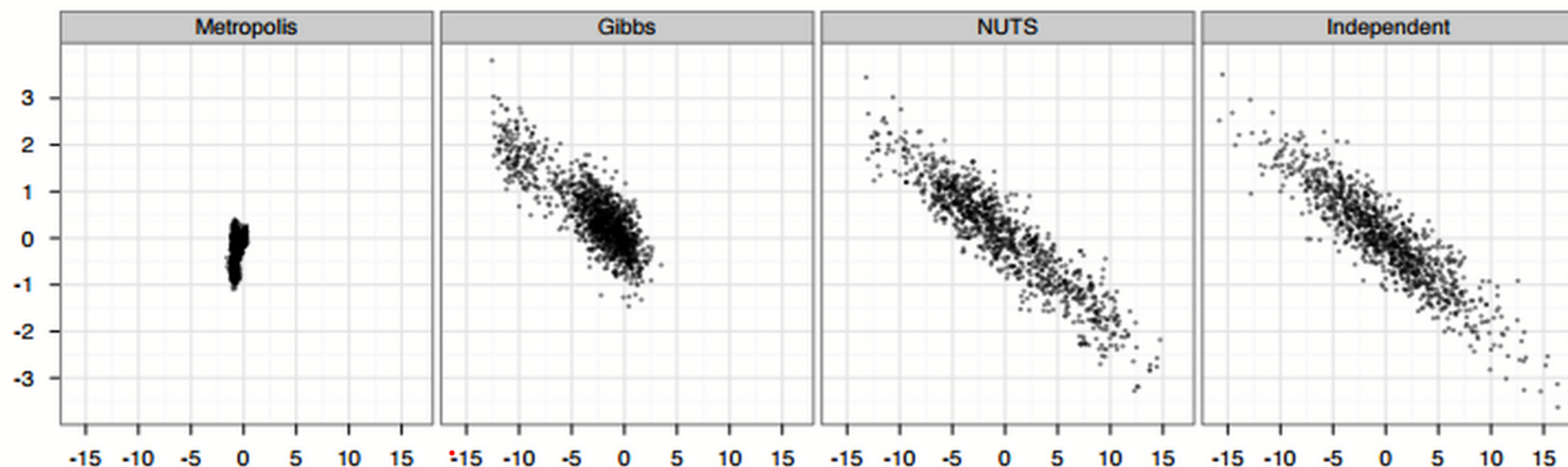


Figure 7: Samples generated by random-walk Metropolis, Gibbs sampling, and NUTS. The plots compare 1,000 independent draws from a highly correlated 250-dimensional distribution (right) with 1,000,000 samples (thinned to 1,000 samples for display) generated by random-walk Metropolis (left), 1,000,000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

$$O(p^2)$$

$$O(p^2)$$

$$O(p^{5/4})$$

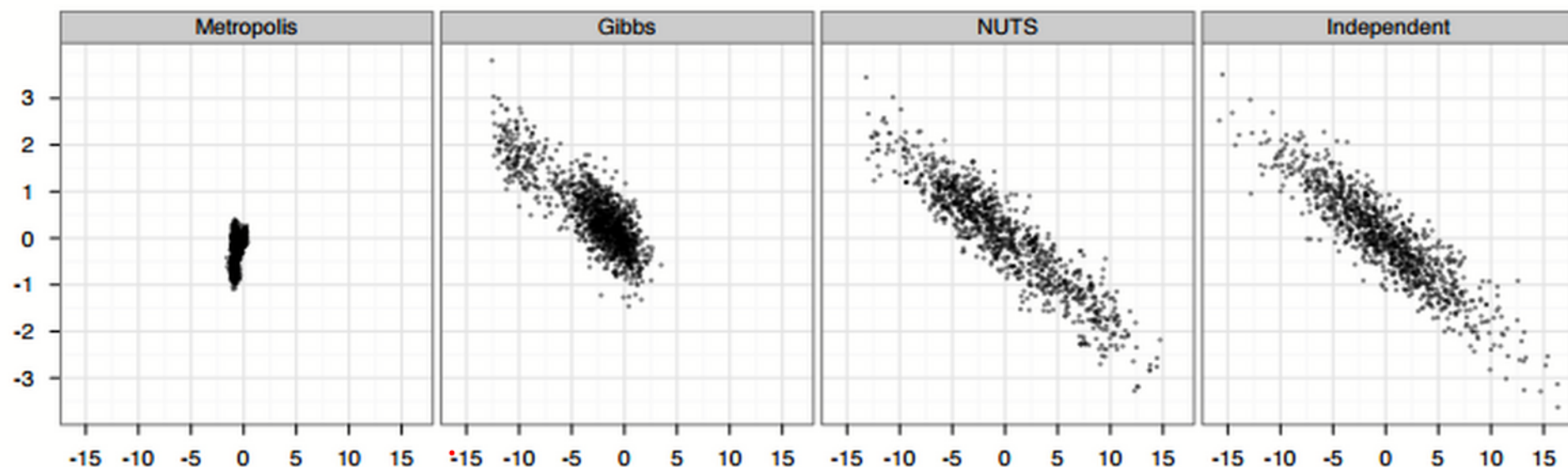


Figure 7: Samples generated by random-walk Metropolis, Gibbs sampling, and NUTS. The plots compare 1,000 independent draws from a highly correlated 250-dimensional distribution (right) with 1,000,000 samples (thinned to 1,000 samples for display) generated by random-walk Metropolis (left), 1,000,000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

$$O(p^2)$$

$$O(p^2)$$

$$O(p^{5/4})$$

Target

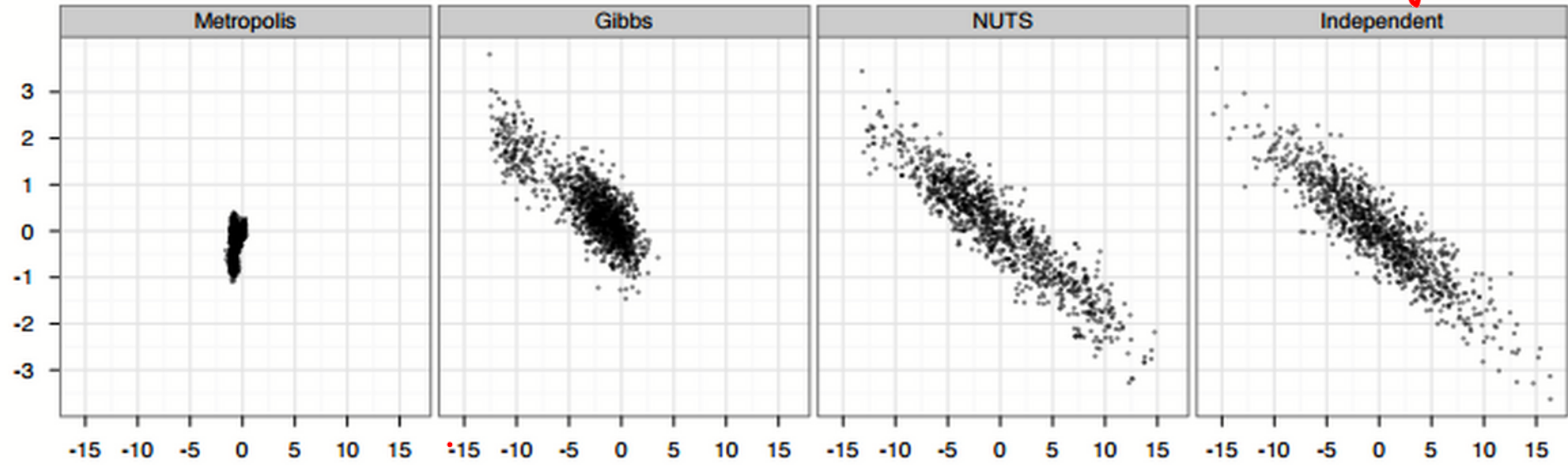


Figure 7: Samples generated by random-walk Metropolis, Gibbs sampling, and NUTS. The plots compare 1,000 independent draws from a highly correlated 250-dimensional distribution (right) with 1,000,000 samples (thinned to 1,000 samples for display) generated by random-walk Metropolis (left), 1,000,000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

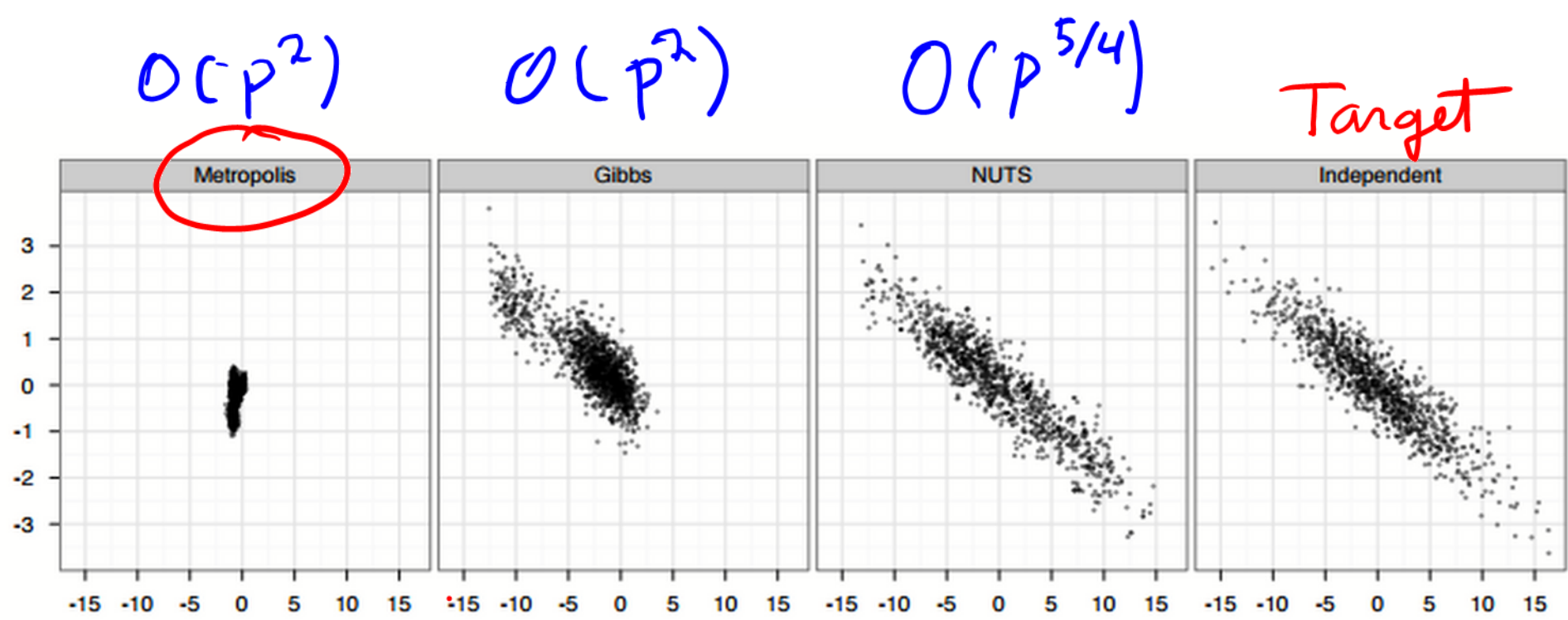


Figure 7: Samples generated by random-walk Metropolis, Gibbs sampling, and NUTS. The plots compare 1,000 independent draws from a highly correlated 250-dimensional distribution (right) with 1,000,000 samples (thinned to 1,000 samples for display) generated by random-walk Metropolis (left), 1,000,000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

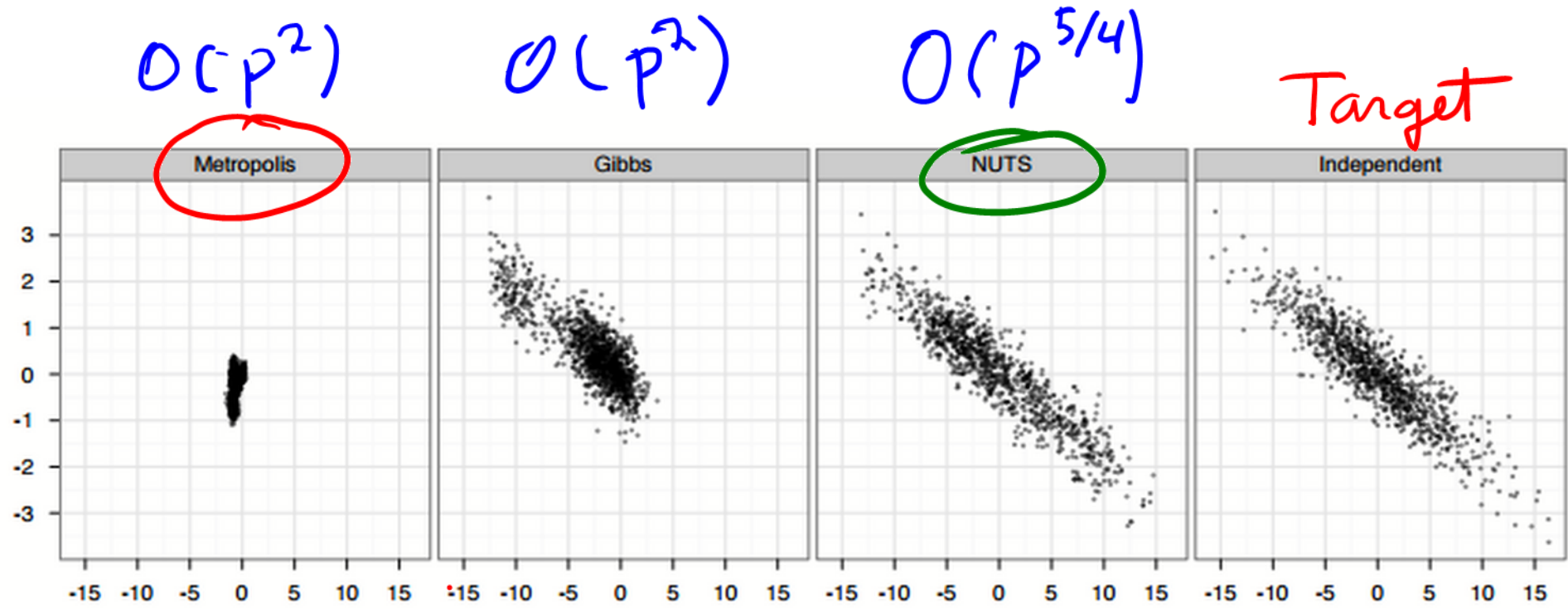


Figure 7: Samples generated by random-walk Metropolis, Gibbs sampling, and NUTS. The plots compare 1,000 independent draws from a highly correlated 250-dimensional distribution (right) with 1,000,000 samples (thinned to 1,000 samples for display) generated by random-walk Metropolis (left), 1,000,000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

$\hookrightarrow$ HMC

from Hoffman & Gelman (2014)

Hamiltonian Monte Carlo

Hamiltonian Monte Carlo

→ Turn the mountain



into a bowl



by using

$$P(X, \underline{\theta})$$

fixed

$$-\log P(X, \underline{\theta})$$

Hamiltonian Monte Carlo

— Turn the mountain



into a bowl



by using

$$P(X, \underline{\theta})$$

fixed

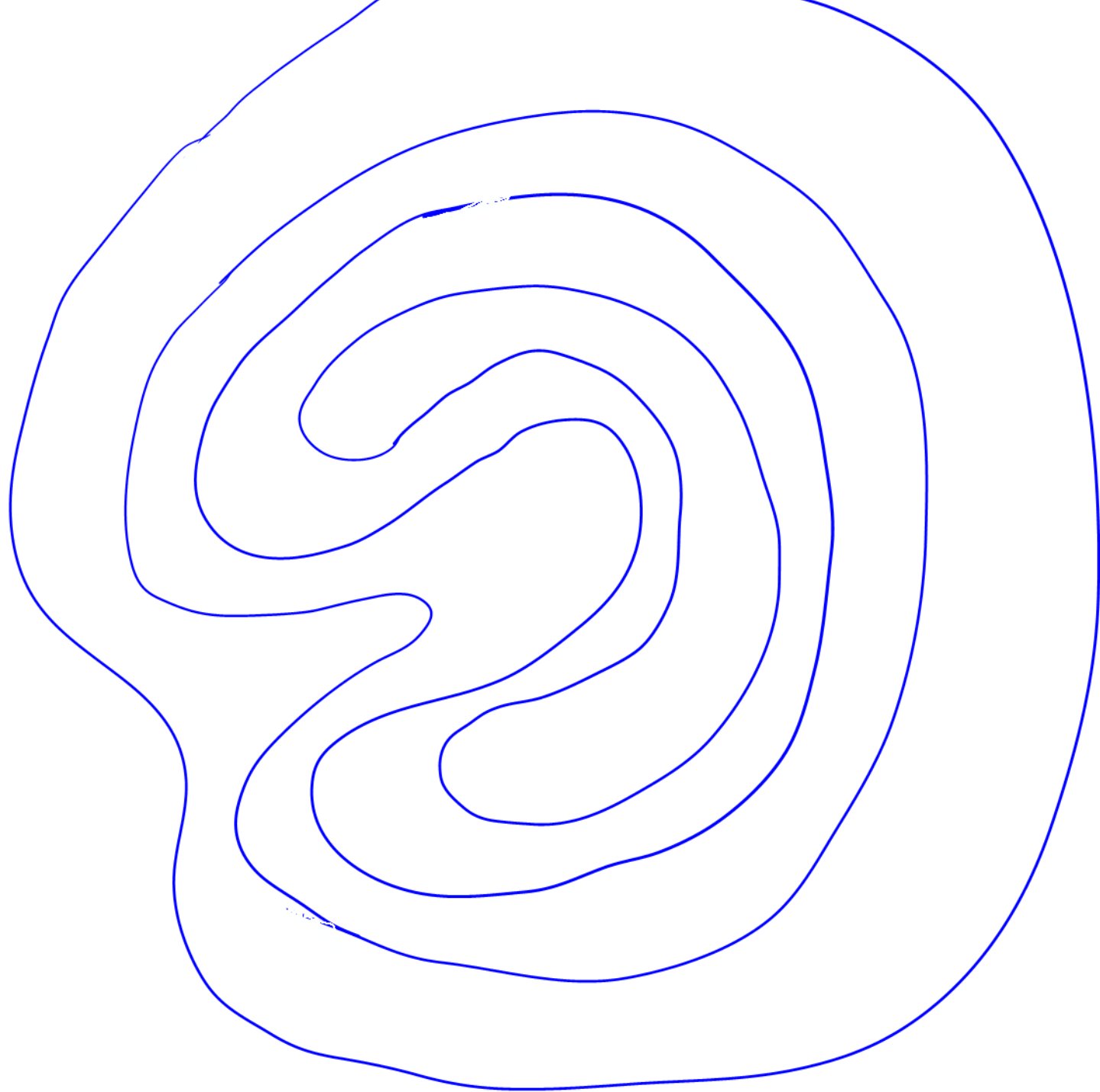
$$-\log P(X, \underline{\theta})$$

— Instead of taking random steps,
go for a ride on a frictionless skateboard
with swivel wheels — starting with a
random push.

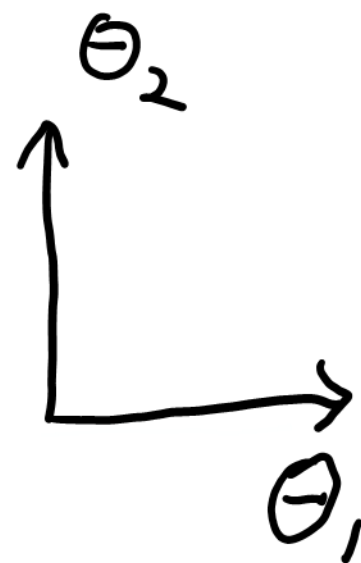
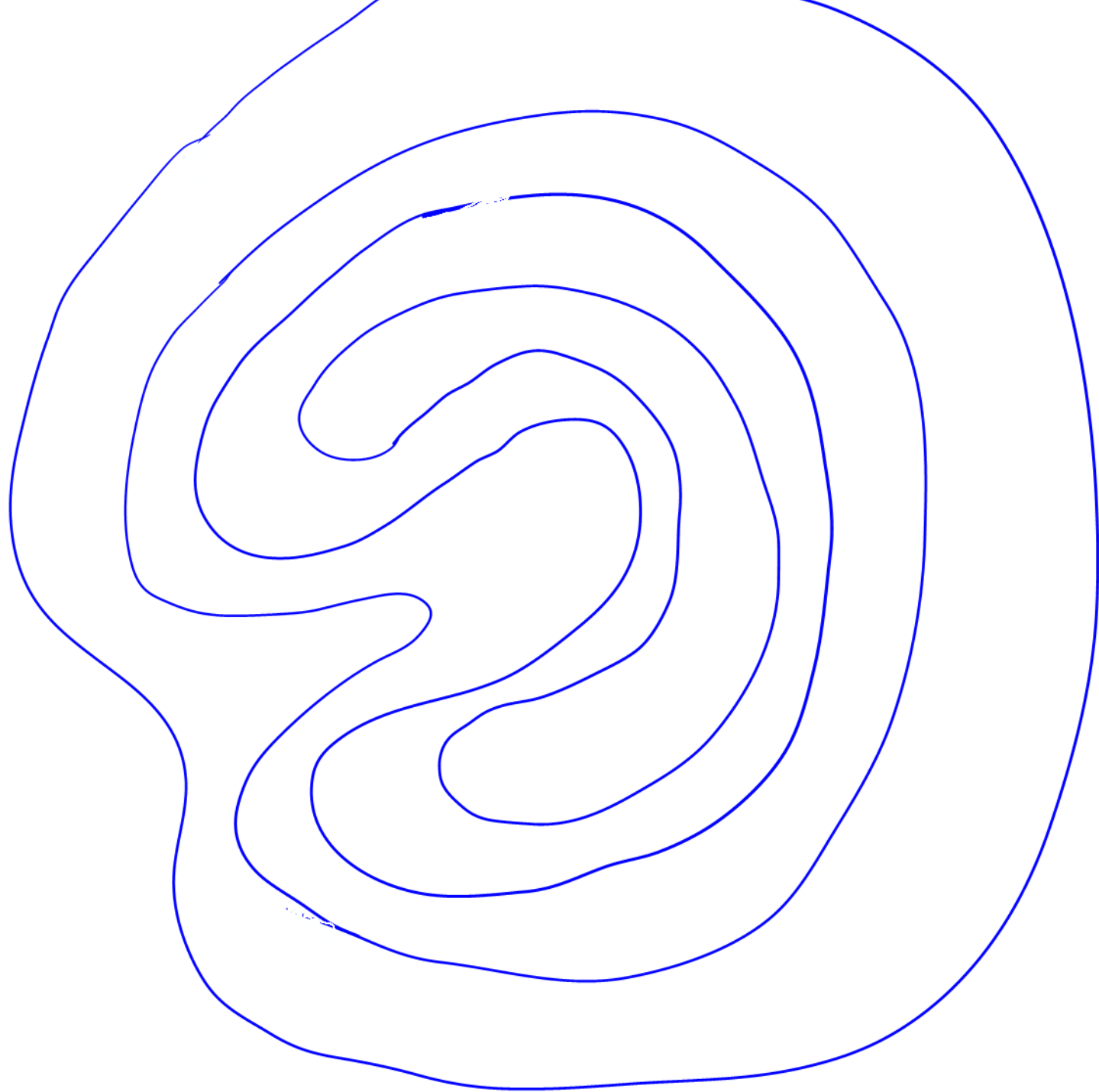




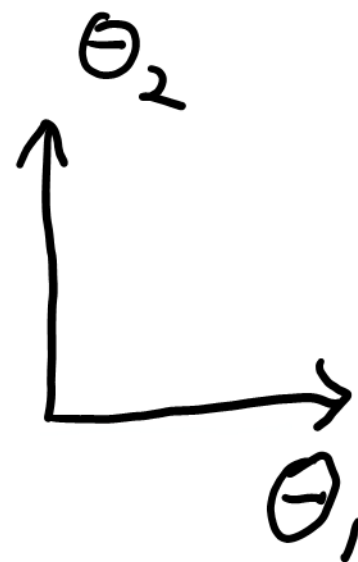
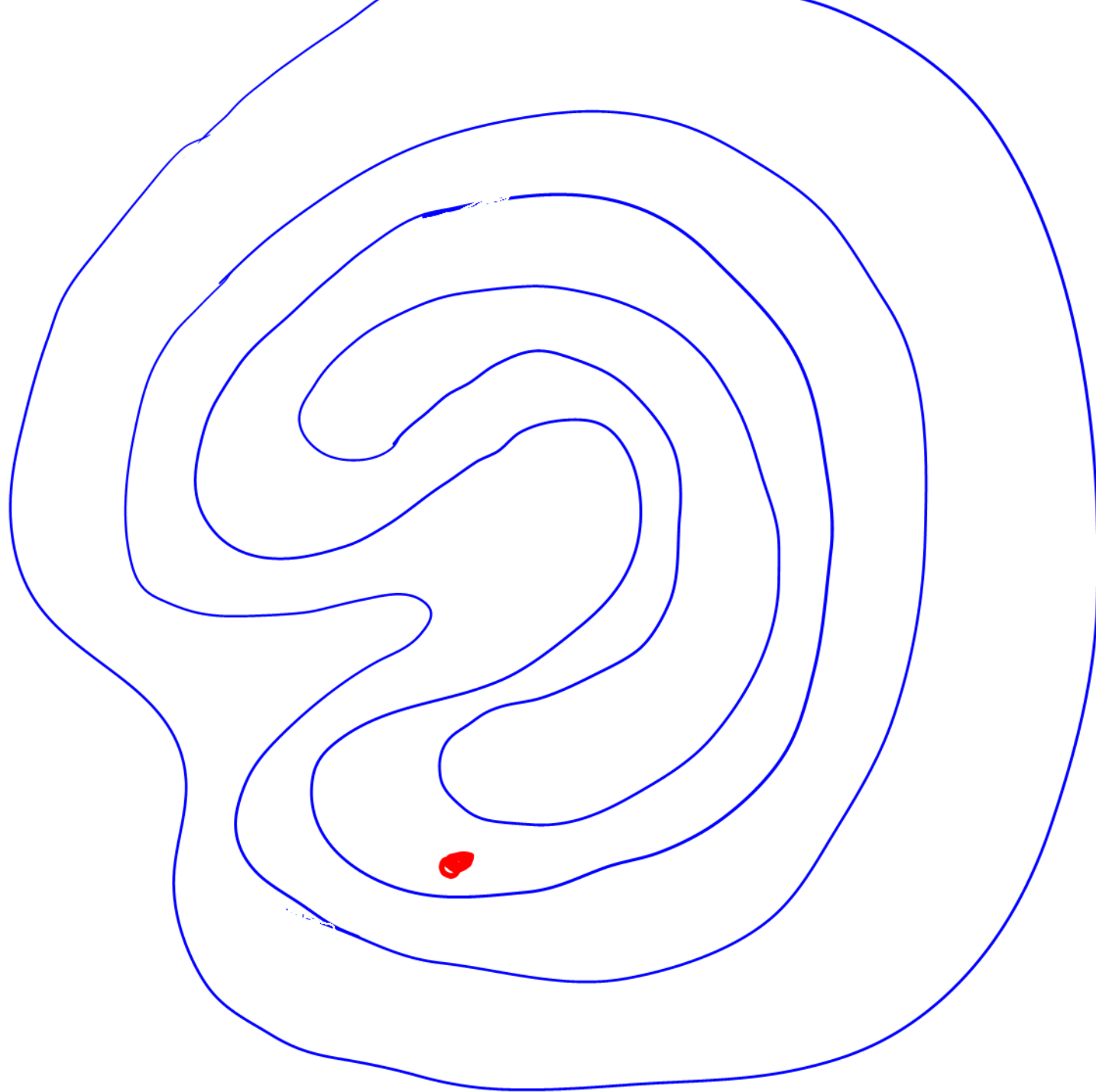


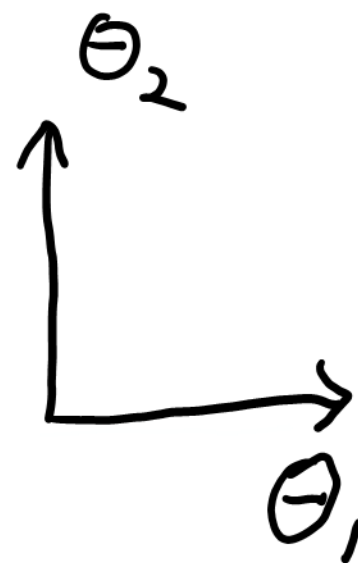
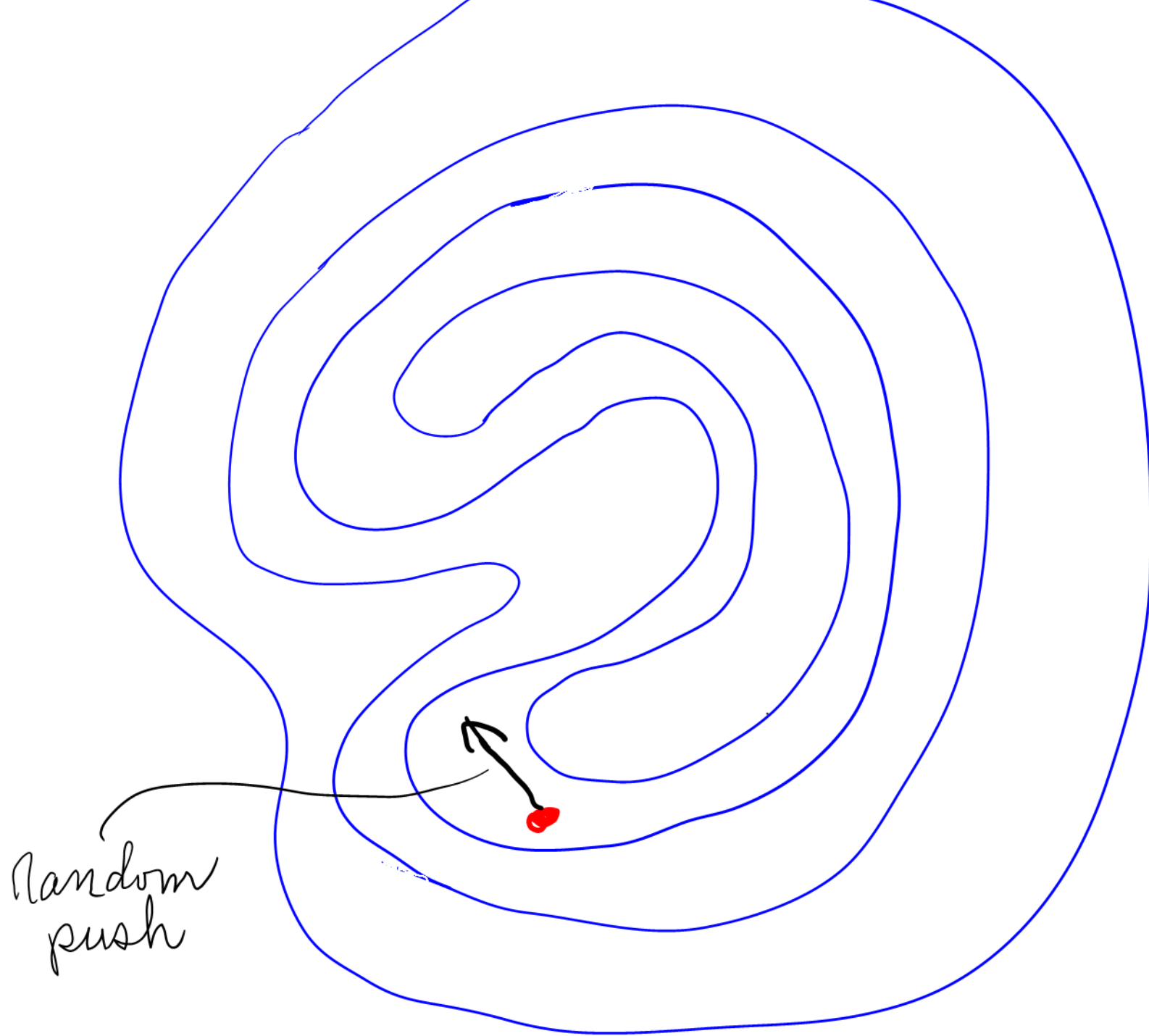


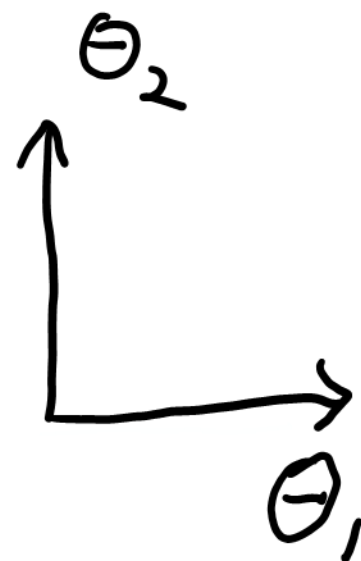
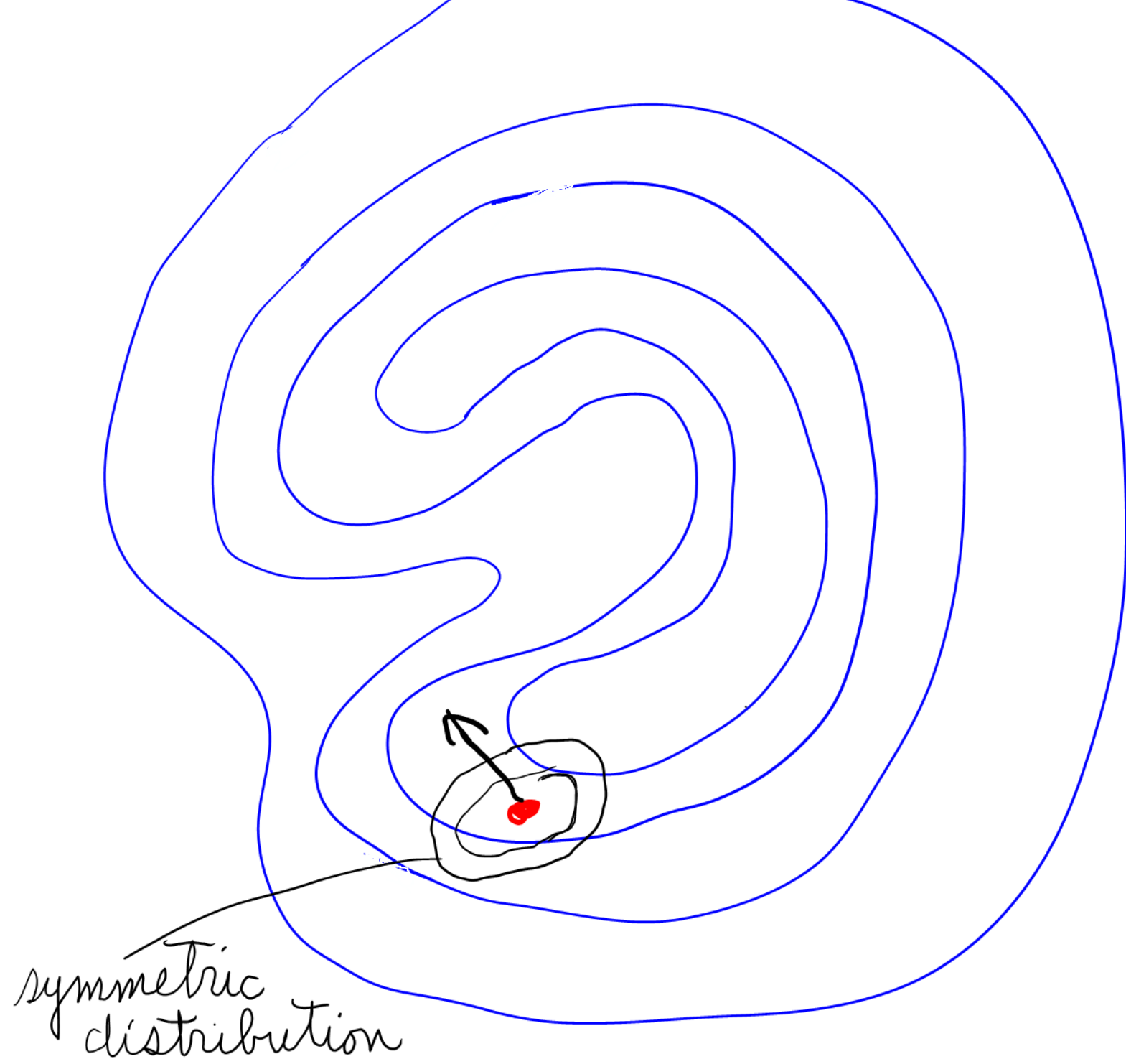


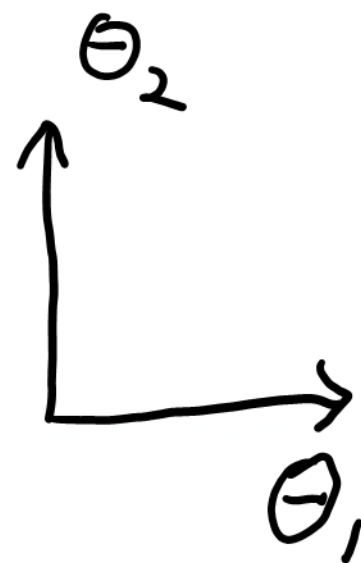
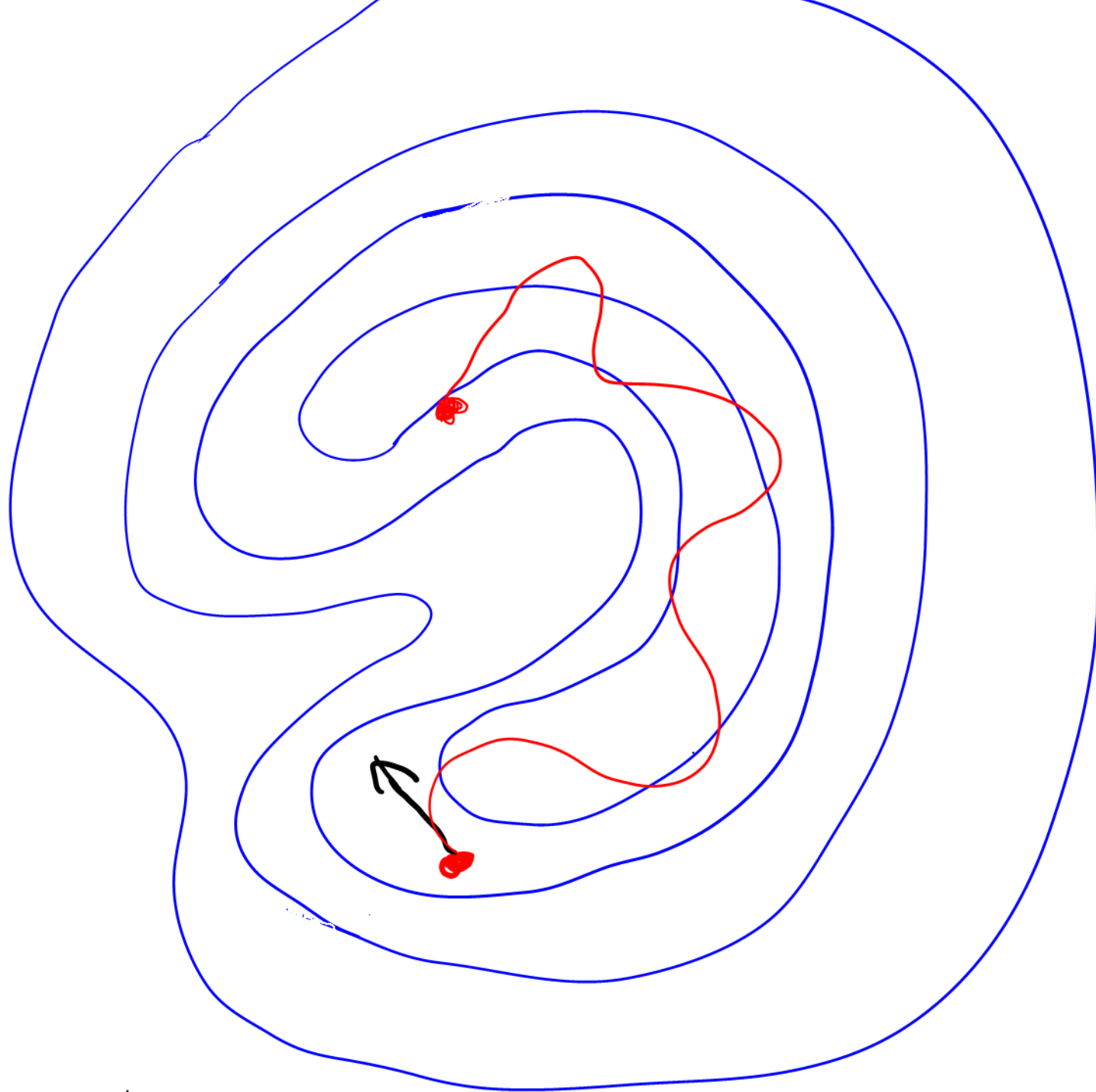


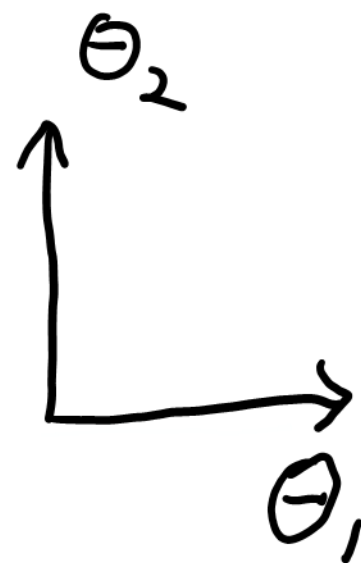
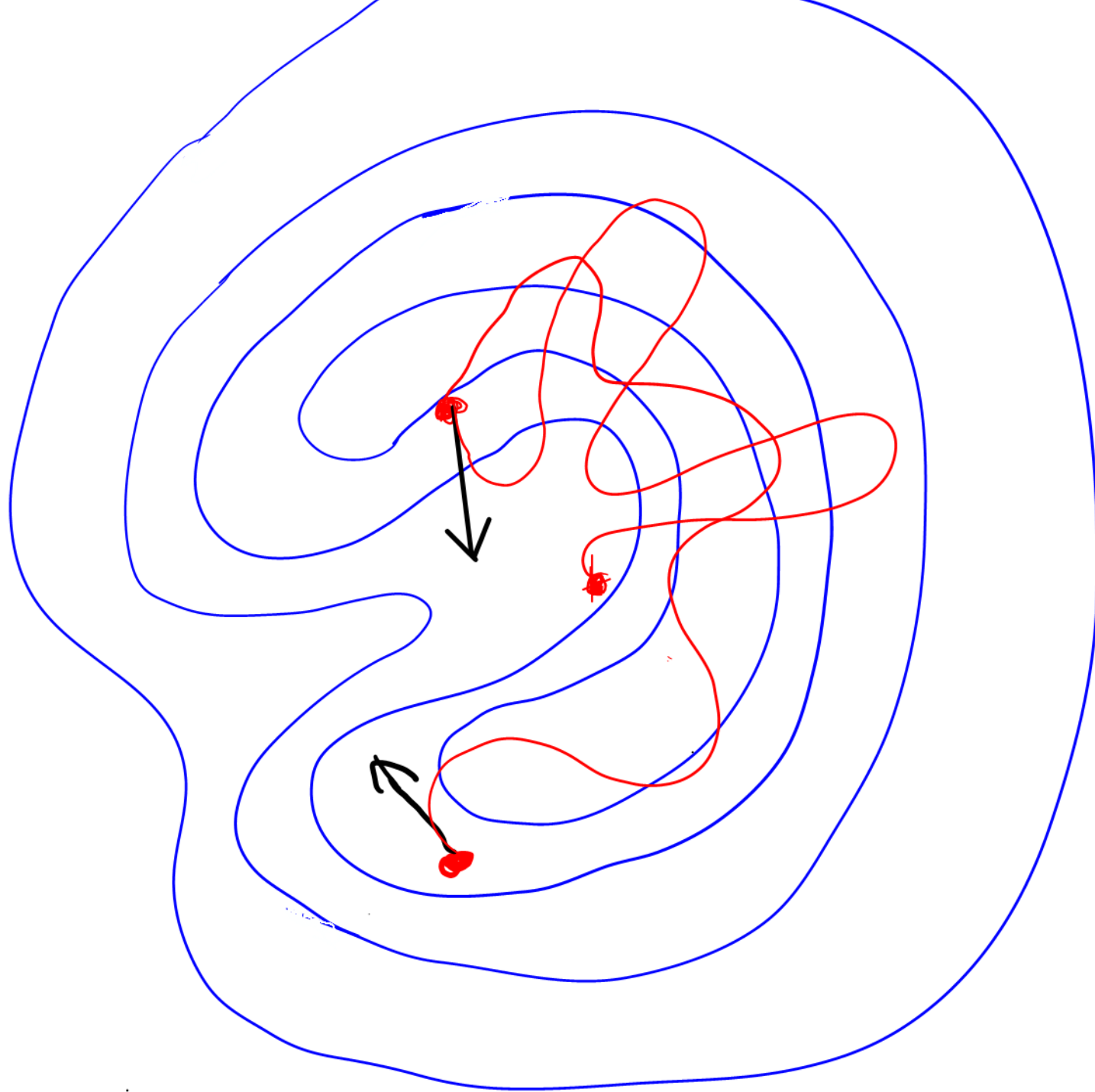
Hamiltonian
monte
carlo

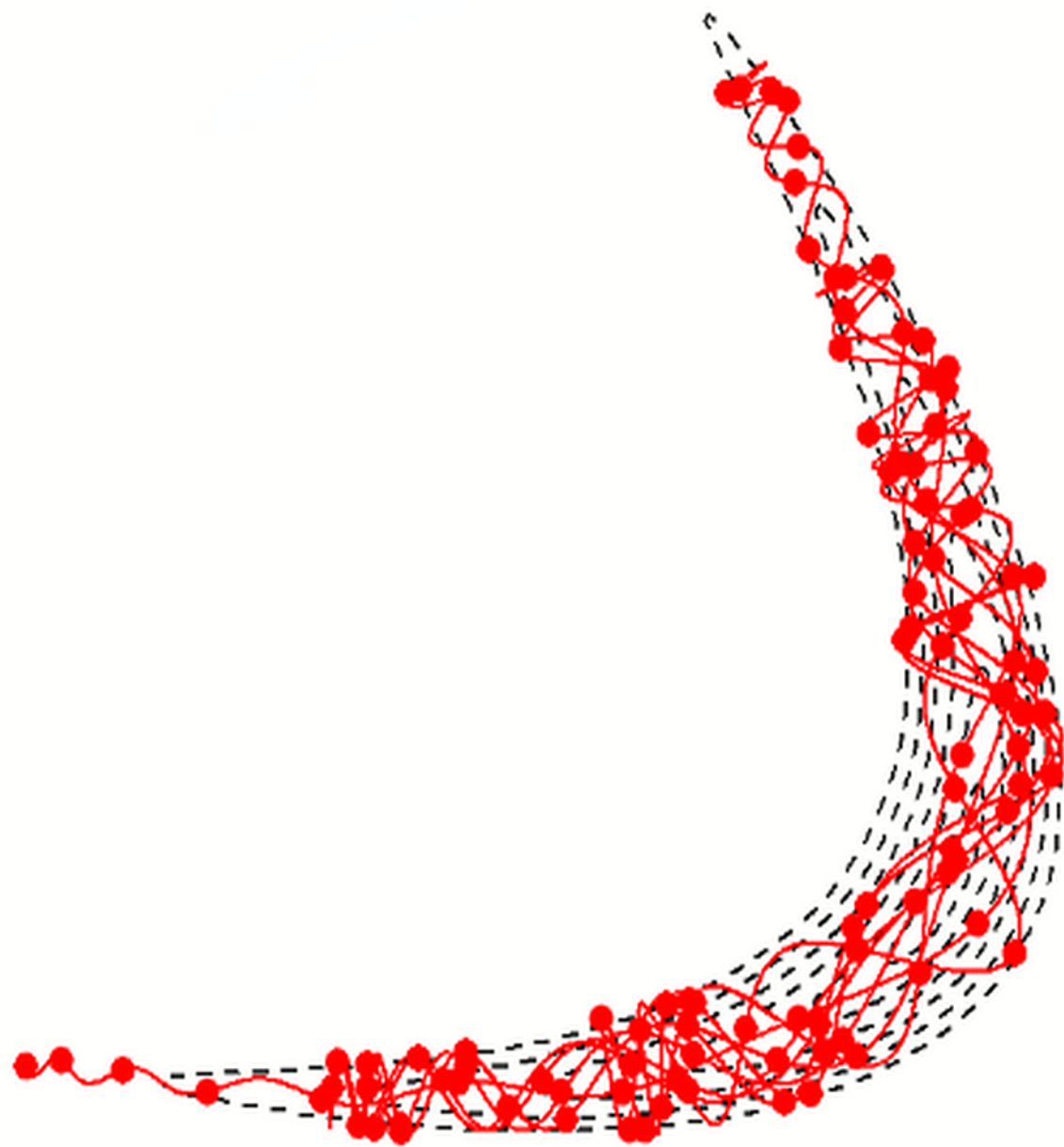










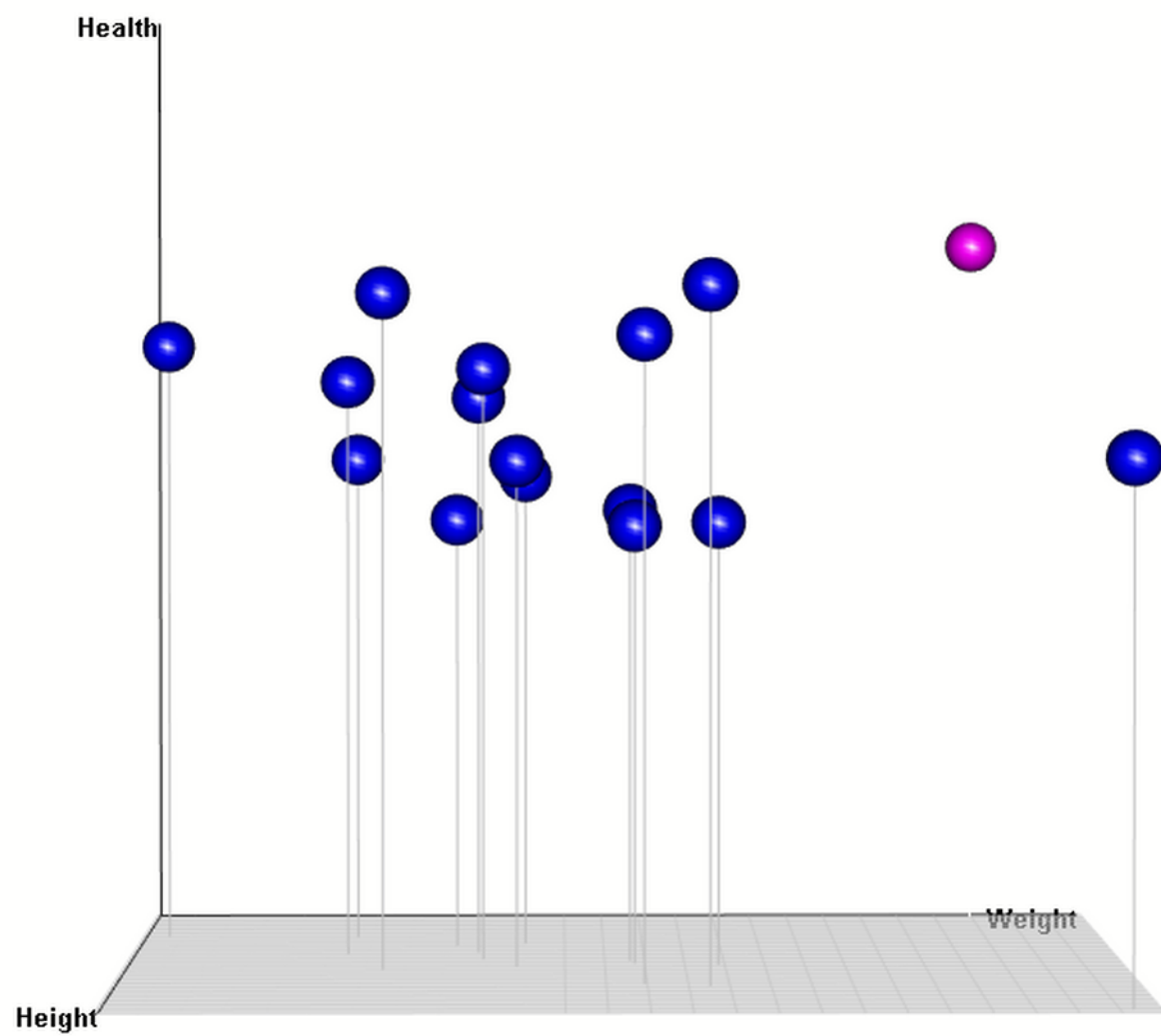


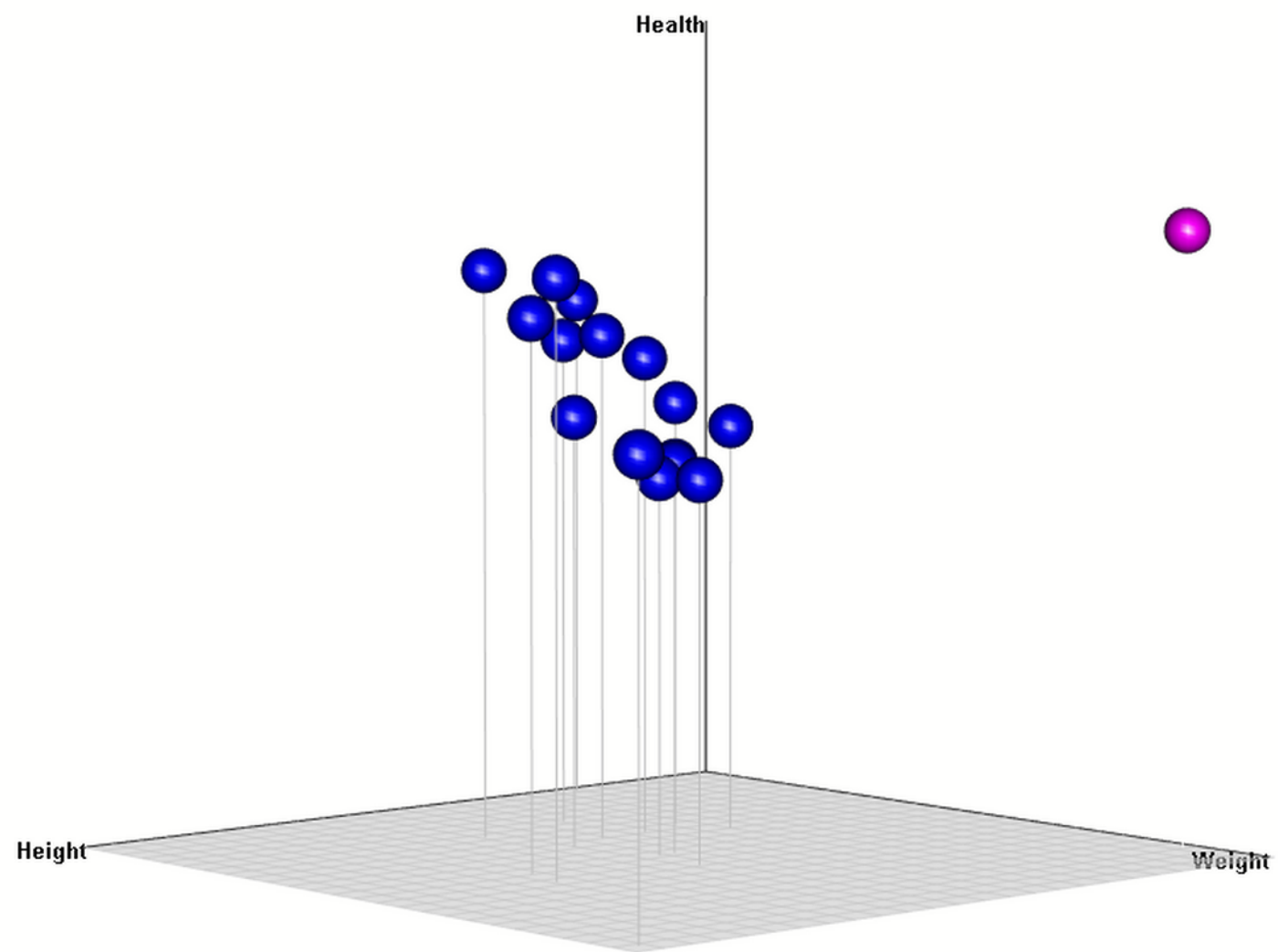
- Much less dependency between successive points in comparison with other MCMC methods
- However, each point is more expensive to generate
- Much faster with large # of parameters

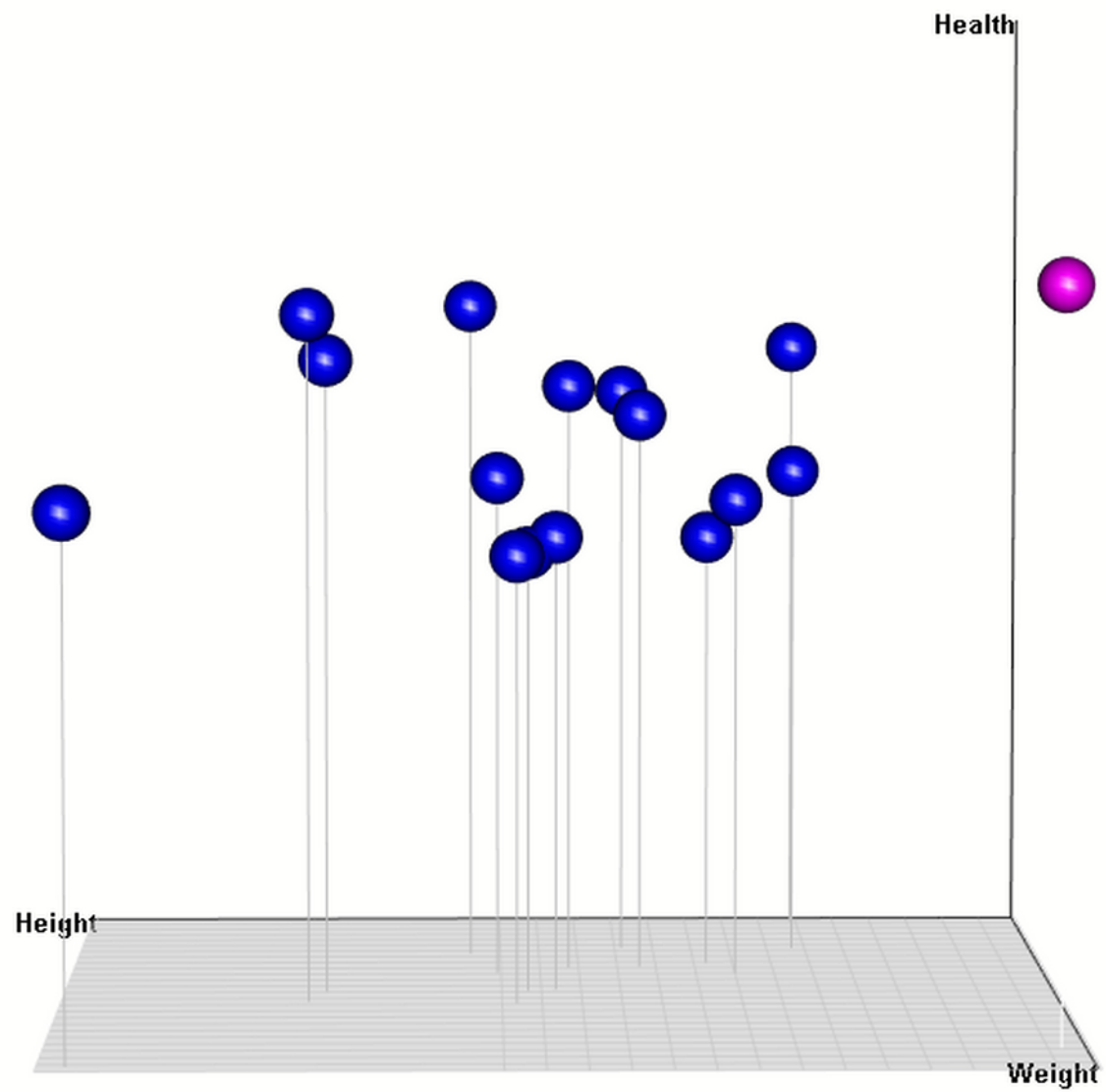


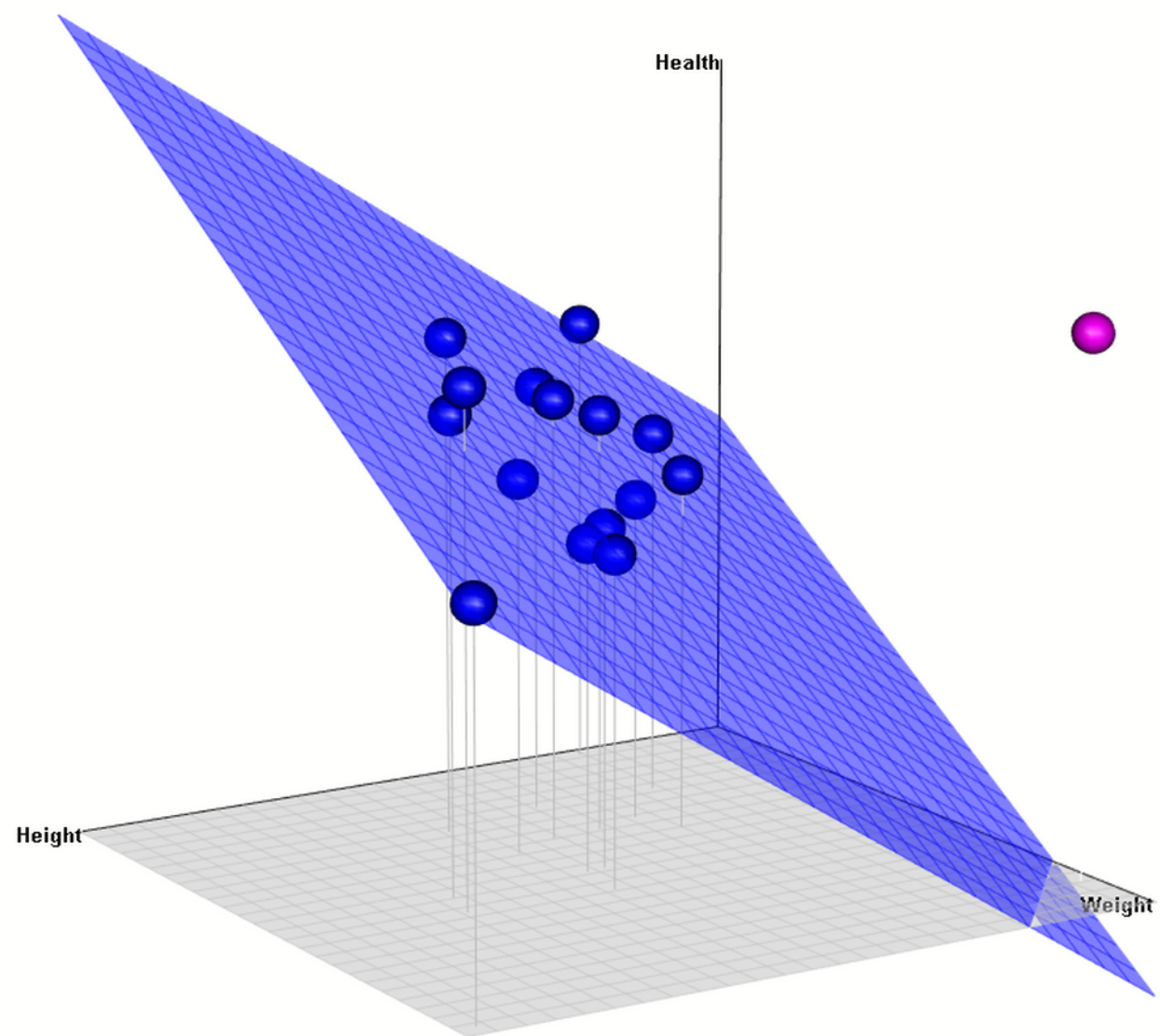
Stan

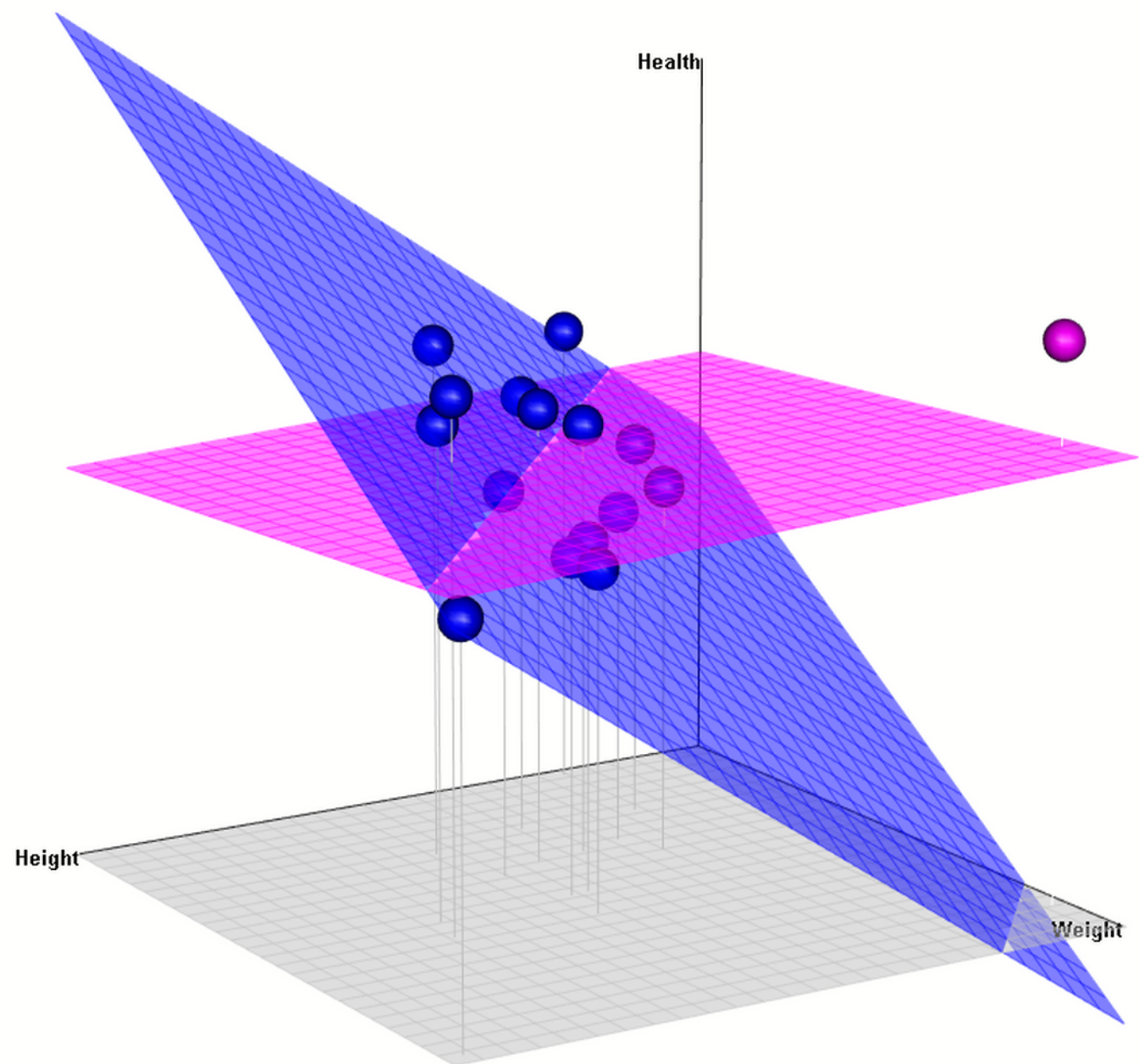
Thousands of users rely on Stan for statistical modeling, data analysis, and prediction in the social, biological, and physical sciences, engineering, and business.

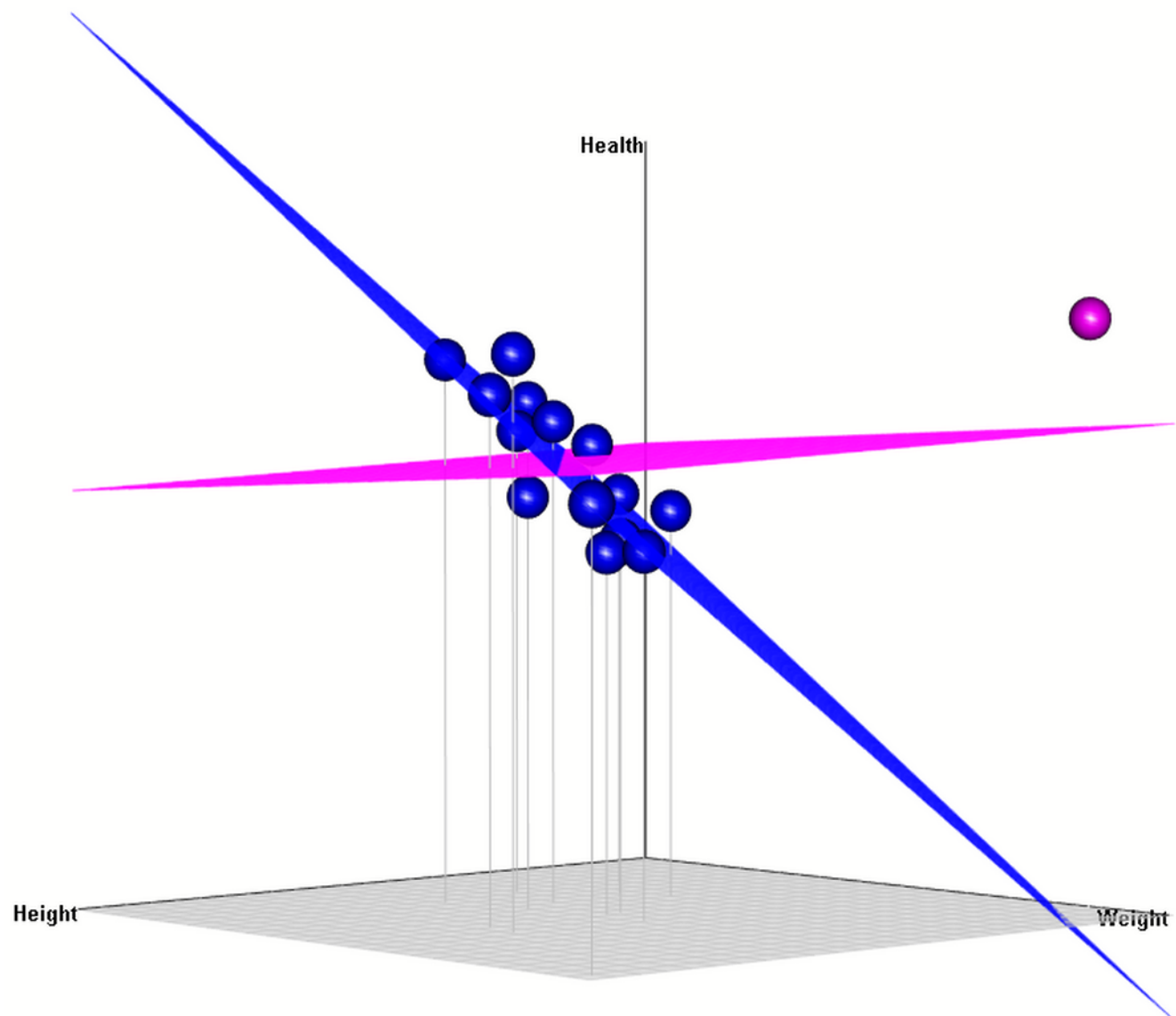












Defining a model in Stan

```
data {  
  int N;    // number of observations  
  int P;    // number of columns of X matrix (including intercept)  
  matrix[N,P] X;  // X matrix including intercept  
  vector[N] y;  // response  
}  
  
parameters {  
  vector[P] beta;      // default uniform prior if nothing specied in model  
  real <lower=0> sigma; // uniform on positive reals  
}  
  
model {  
  y ~ normal( X * beta, sigma ); // note that * is matrix mult.  
                                   // For elementwise multiplication use .*  
}
```


Data

```
head( Xmat <- model.matrix(Health ~ Weight + Height, dd) )
```

```
      (Intercept) Weight Height
1             1 0.3355 0.6008
2             1 0.6890 0.9440
3             1 0.6980 0.6150
4             1 0.7617 1.2340
5             1 0.8910 0.7870
6             1 0.9330 0.9150
```

```
dat_list <- list(N = nrow(Xmat), P = ncol(Xmat),
                 X = Xmat, y = dd$Health)
```

\$N

[1] 16

\$P

[1] 3

\$X

	(Intercept)	Weight	Height
1	1	0.3355	0.6008
2	1	0.6890	0.9440
3	1	0.6980	0.6150
4	1	0.7617	1.2340
5	1	0.8910	0.7870
6	1	0.9330	0.9150
7	1	0.9430	1.0490
8	1	1.0060	1.1840
9	1	1.0200	0.7370
10	1	1.2150	1.0770
11	1	1.2230	1.1280
12	1	1.2360	1.5000
13	1	1.3530	1.5310
14	1	1.3770	1.1500
15	1	2.0734	1.9340
18	1	1.9000	0.2000

attr(,"assign")

[1] 0 1 2

\$y

[1] 1.280 1.208 1.036 1.395 0.912 1.175 1.237 1.048 1.003 0.943 0.912
[12] 1.311 1.411 0.920 1.073 1.500

```
fit_stan <- sampling(reg_model_dso, dat_list)
```

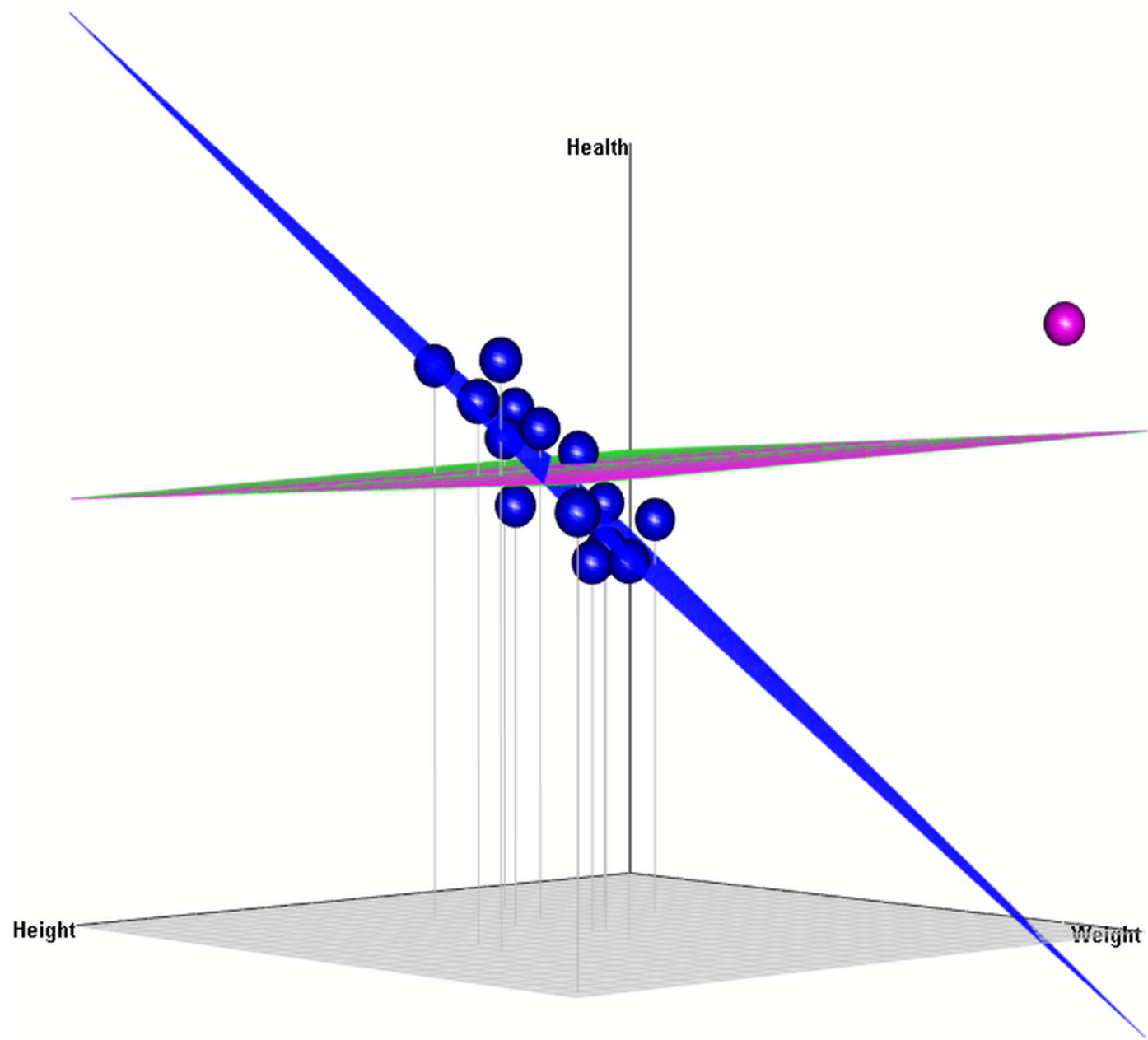
fit_stan

w Snip

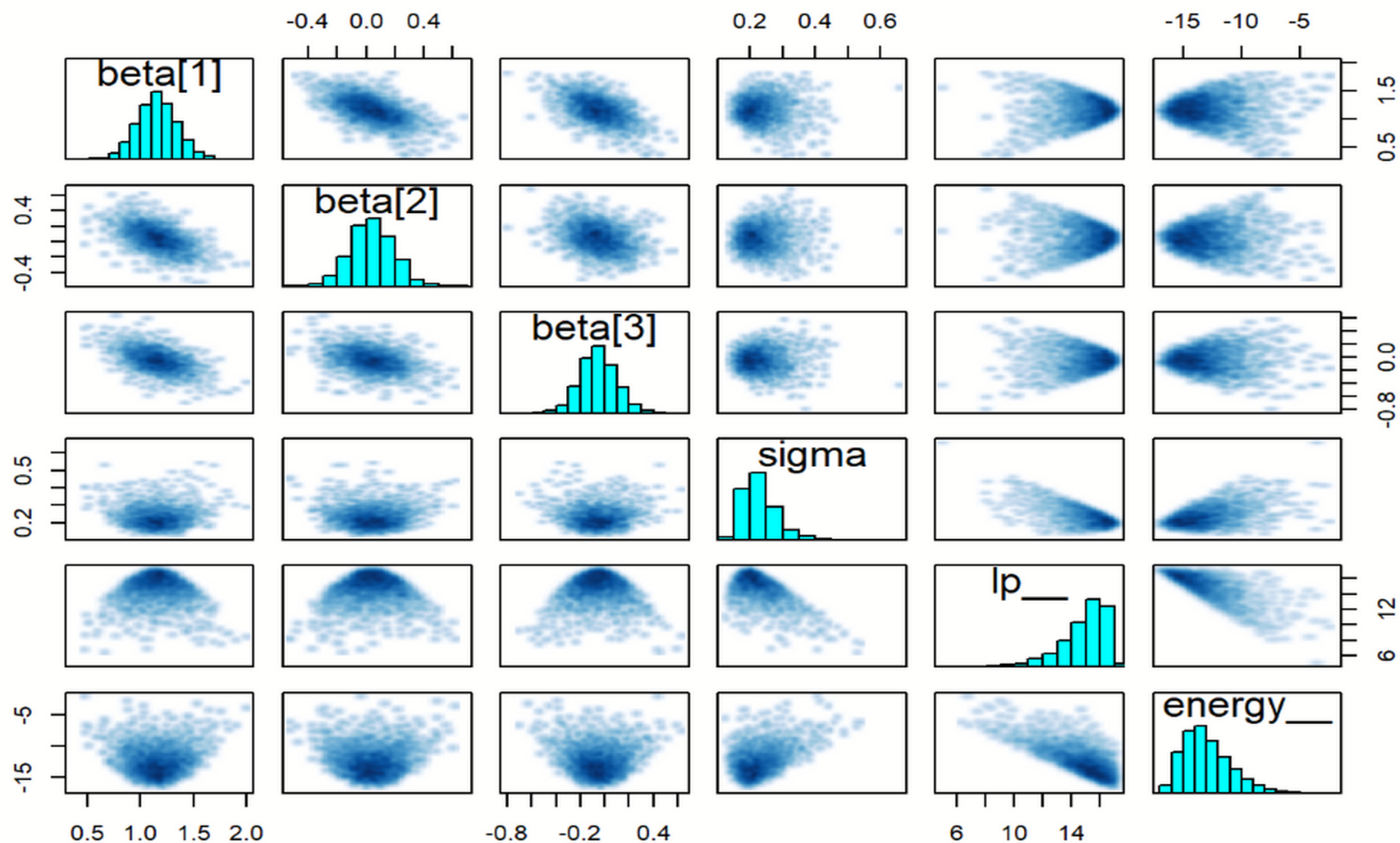
Inference for Stan model: 5a6361673e39acd797b5518d36e20615.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
beta[1]	1.16	0.00	0.20	0.77	1.03	1.16	1.29	1.55	2035	1
beta[2]	0.04	0.00	0.15	-0.26	-0.06	0.04	0.14	0.34	1980	1
beta[3]	-0.05	0.00	0.16	-0.36	-0.15	-0.06	0.04	0.26	2135	1
sigma	0.23	0.00	0.05	0.15	0.19	0.22	0.26	0.36	1713	1
lp__	14.88	0.05	1.64	10.52	14.08	15.25	16.08	16.93	1084	1

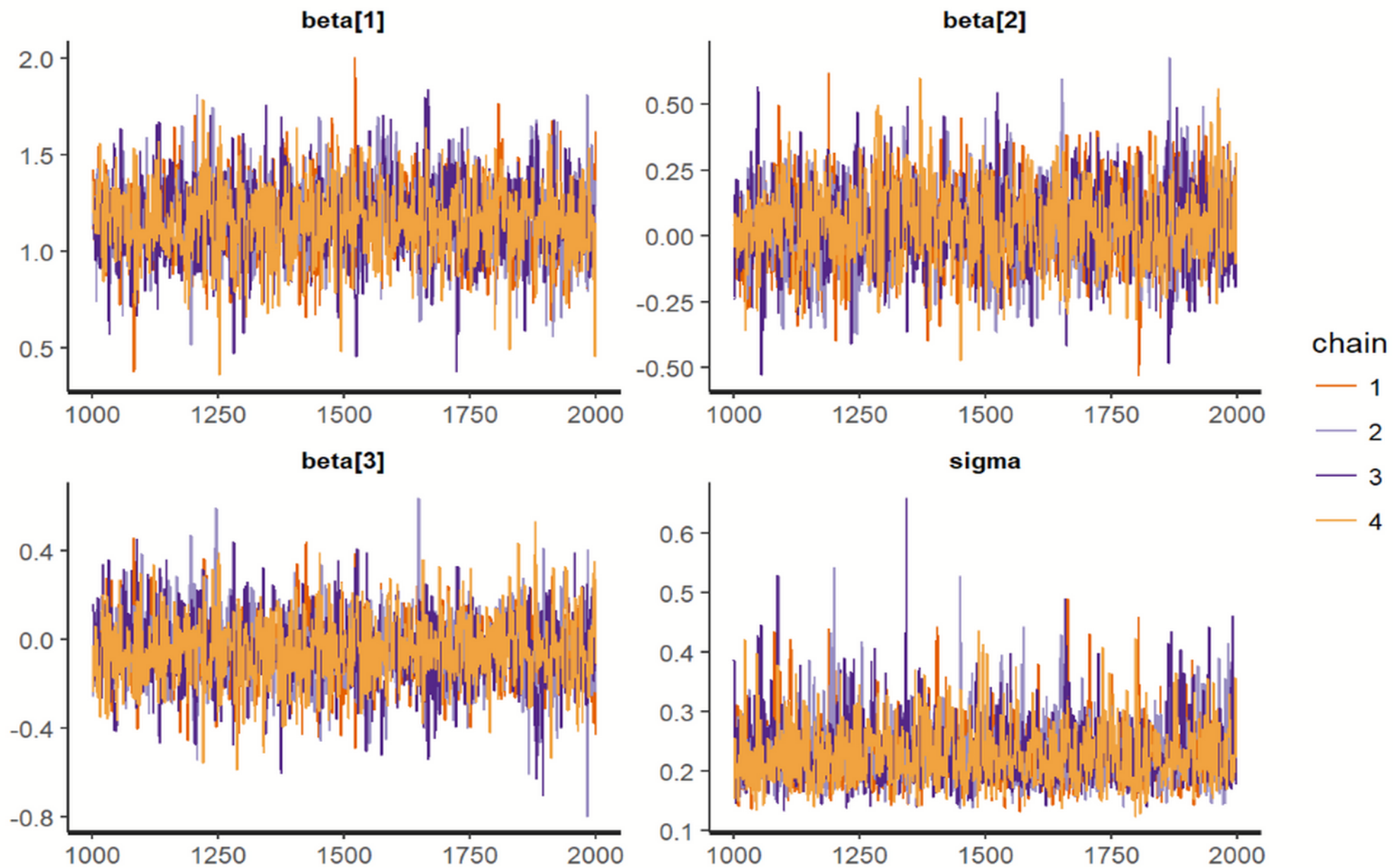
Samples were drawn using NUTS(diag_e) at Mon Jun 05 22:52:12 2017.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).



```
pairs(fit_stan)
```



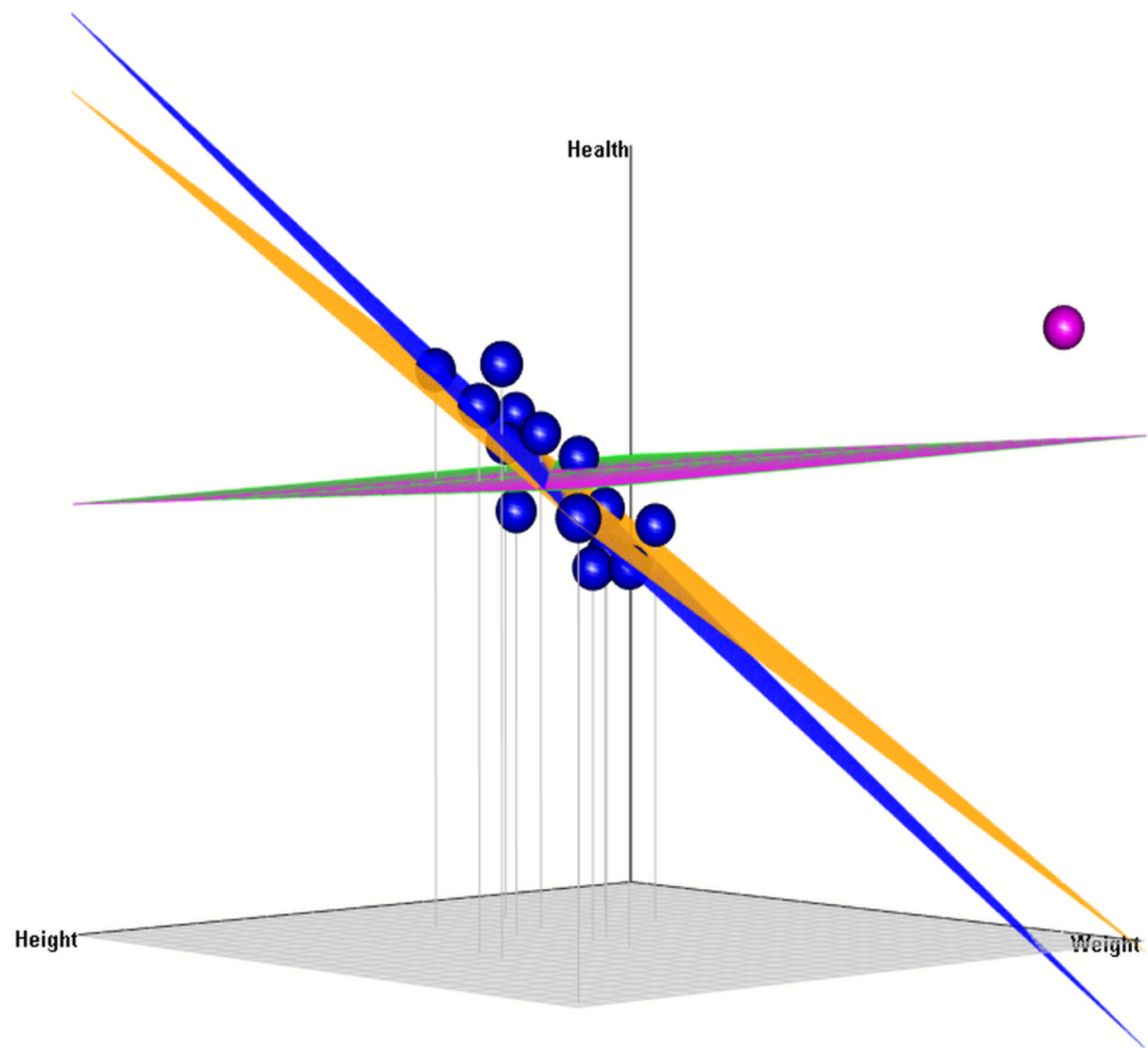
```
traceplot(fit_stan)
```



Robust model: Just change the distribution of Y

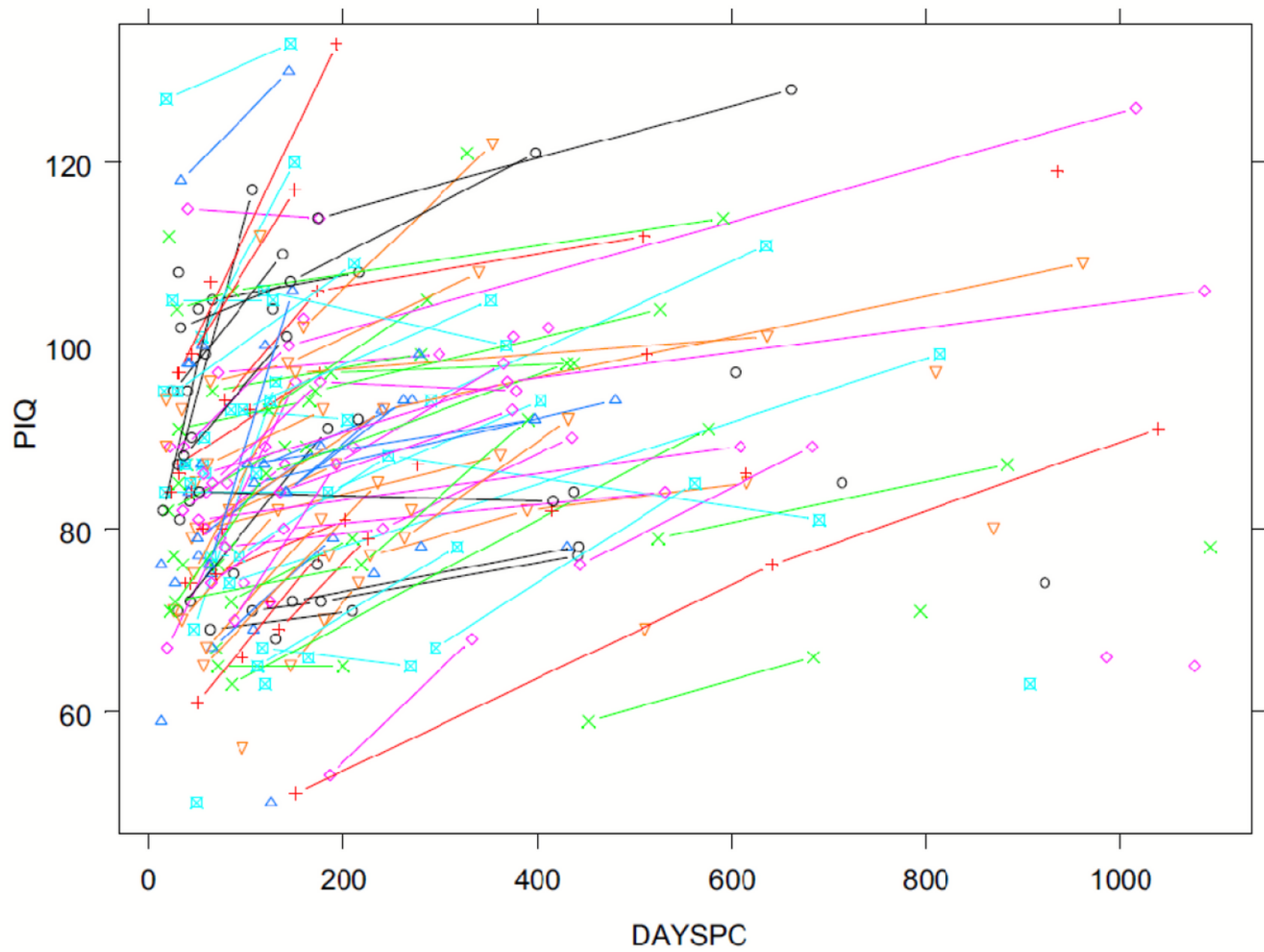
```
data {  
  int N;    // number of observations  
  int P;    // number of columns of X matrix (including intercept)  
  matrix[N,P] X;  // X matrix including intercept  
  vector[N] y;  // response  
  int nu;    // degrees for freedom for student_t  
}  
  
parameters {  
  vector[P] beta;  // default uniform prior if nothing specied in model  
  real <lower=0> sigma;  
}  
  
model {  
  y ~ student_t(nu, X * beta, sigma );  
}  
..
```

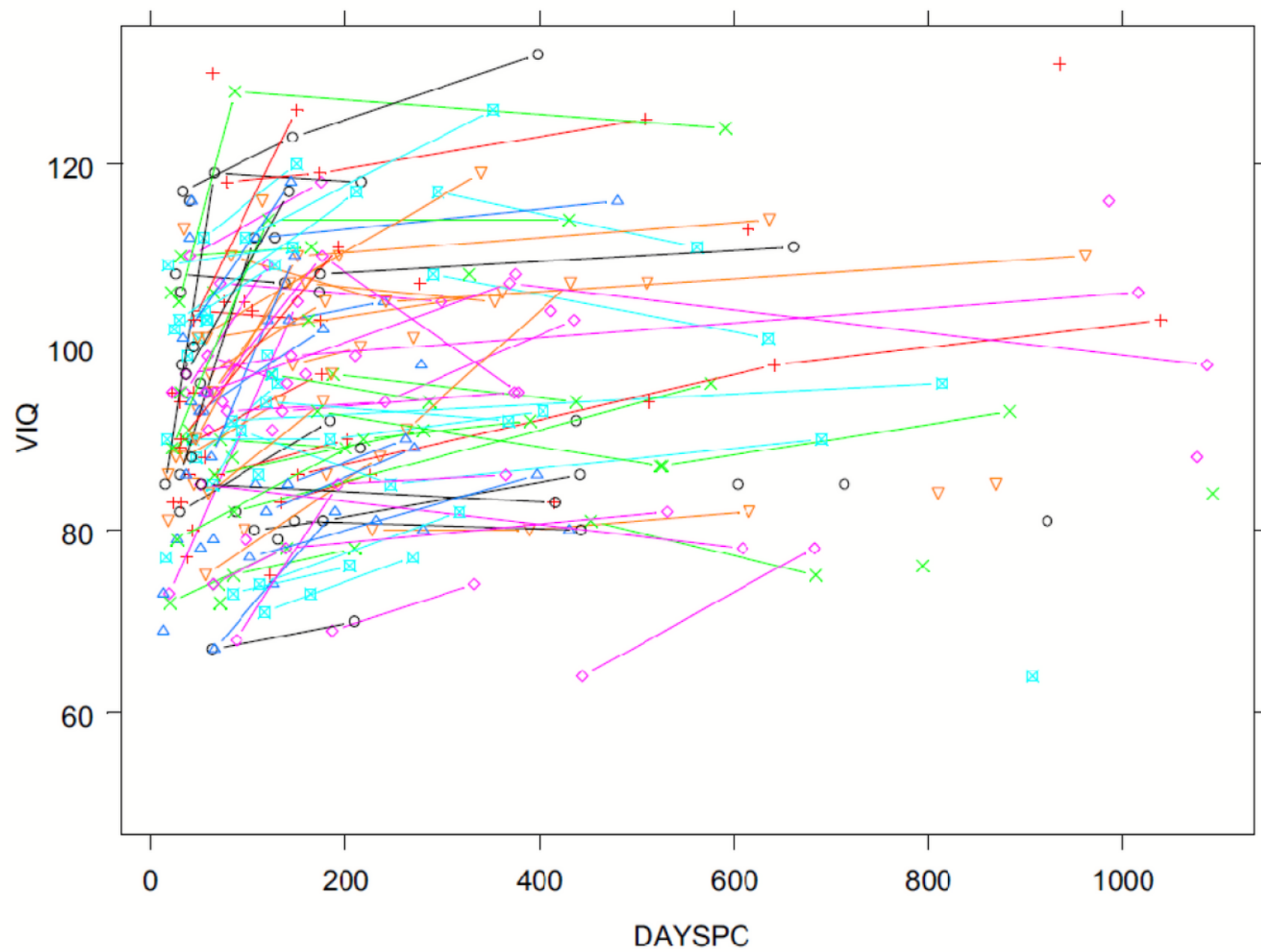
```
fit3_stan_2 <- sampling(robust_model_dso, c(dat_list, nu = 2))
```

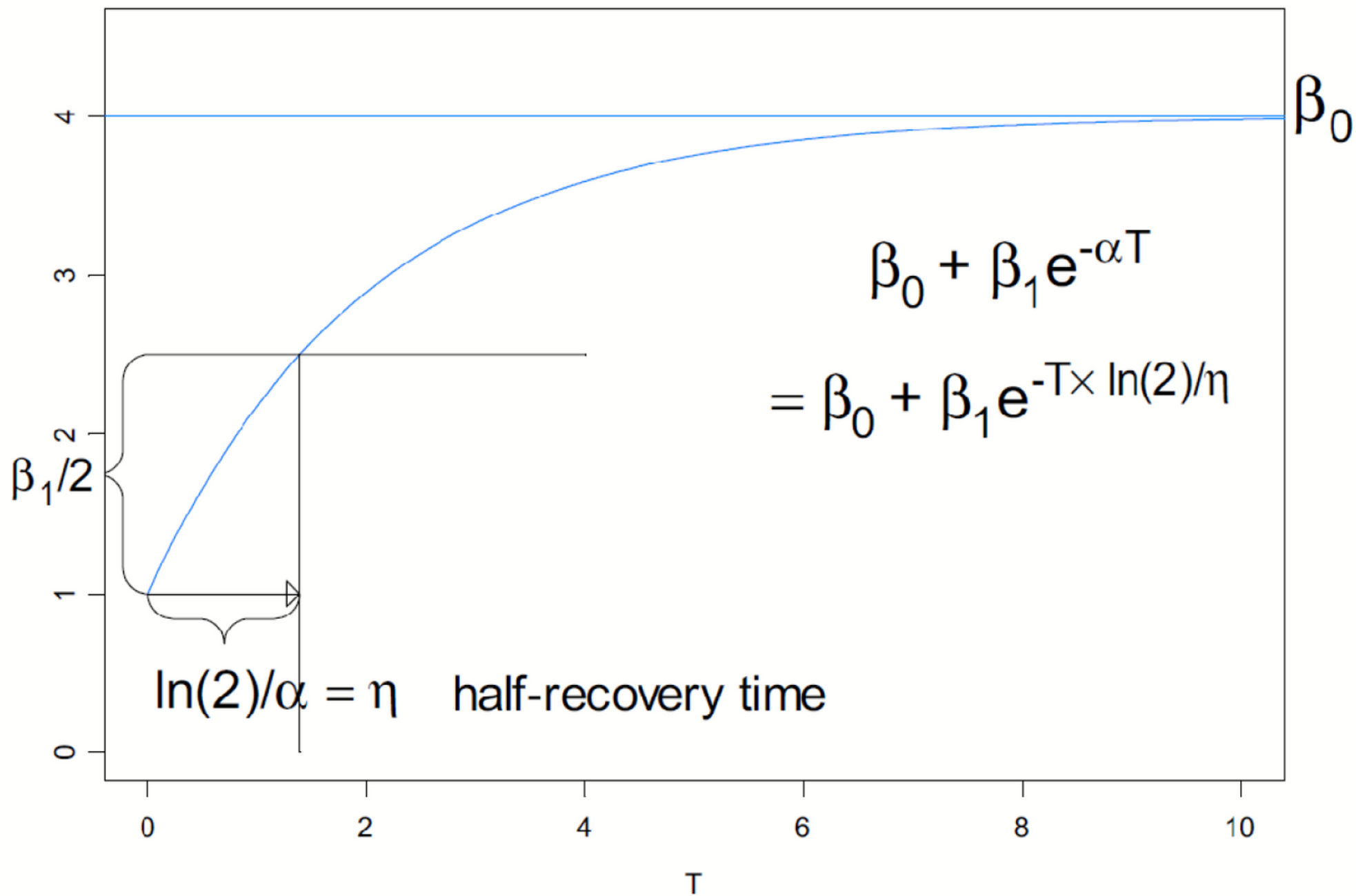


Traumatic Brain Injury

- Recovery after coma
- Non-linear asymptotic recovery curves








```
//  
// Multivariate model for VIQ and PIQ  
//
```

```
data {  
  int N;  
  int J;  
  matrix[N,2] iq;  
  vector[N] time;  
  vector[N] coma;  
  int id[N];  
}
```

```

parameters {
  vector <lower=1,upper=10000>[2] hrt;
  vector <lower=0,upper=200>[2] asymp;
  vector <lower=-100,upper=100>[2] init_def;
  vector [2] bcoma;
  vector[2] u[J];
  cov_matrix[2] Sigma;
  cov_matrix[2] Sigma_u;
}
transformed parameters {
  real hrt_diff;
  real bcoma_diff;
  hrt_diff = hrt[2] - hrt[1];
  bcoma_diff = bcoma[2] - bcoma[1];
}

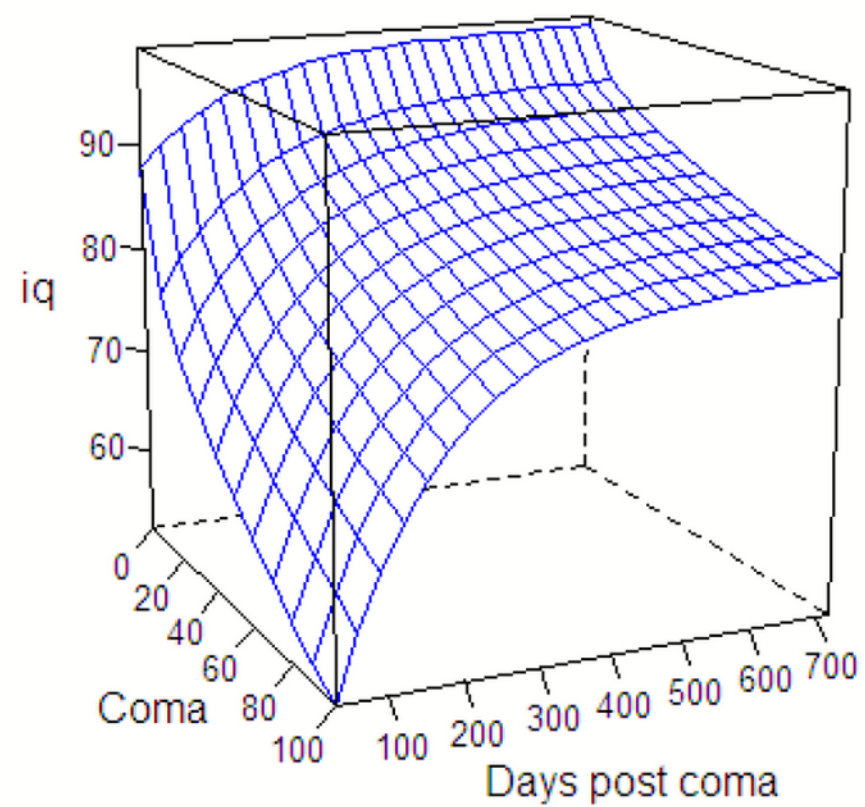
```

```

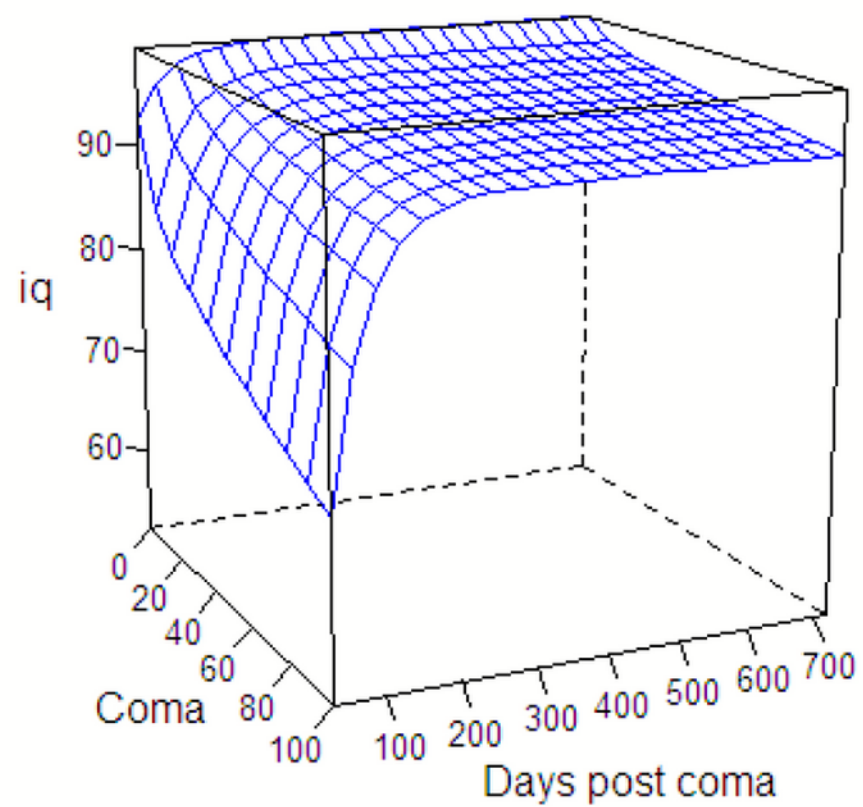
model {
  vector[2] eta;
  // for the multinormal distribution we need to loop over observations
  for(j in 1:J) u[j] ~ multi_normal(zero, Sigma_u);
  for(n in 1:N) {
    eta[1] = asymp[1] + u[id[n],1] + bcoma[1] * coma[n] +
              init_def[1] * exp(-time[n]/(hrt[1]*ln2));
    eta[2] = asymp[2] + u[id[n],2] + bcoma[2] * coma[n] +
              init_def[2] * exp(-time[n]/(hrt[2]*ln2));
    iq[n,] ~ multi_normal(eta, Sigma);
  }
}

```

PIQ



VIQ



Inference for Stan model: asymp_model_4.

4 chains, each with iter=2000; warmup=1000; thin=1;

post-warmup draws per chain=1000, total post-warmup draws=4000.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
hrt[1]	65.18	0.42	18.23	36.67	52.36	62.78	75.12	108.37	1867	1.00
hrt[2]	249.32	1.02	53.51	160.66	211.31	244.72	279.99	373.35	2741	1.00
asymp[1]	99.79	0.06	1.61	96.57	98.73	99.80	100.87	102.86	639	1.01
asymp[2]	100.53	0.07	1.97	96.62	99.23	100.52	101.86	104.38	818	1.00
init_def[1]	-23.34	0.12	4.88	-34.91	-25.94	-22.77	-20.00	-15.66	1633	1.00
init_def[2]	-19.46	0.04	1.95	-23.17	-20.77	-19.52	-18.15	-15.57	1916	1.00
bcoma[1]	-0.72	0.02	0.40	-1.50	-0.99	-0.72	-0.44	0.05	565	1.00
bcoma[2]	-1.93	0.02	0.42	-2.78	-2.22	-1.94	-1.64	-1.11	629	1.00
Sigma[1,1]	33.16	0.10	4.19	25.82	30.24	32.82	35.79	42.37	1632	1.00
Sigma[2,1]	20.38	0.12	4.22	13.02	17.47	20.00	22.96	29.64	1298	1.00
Sigma[1,2]	20.38	0.12	4.22	13.02	17.47	20.00	22.96	29.64	1298	1.00
Sigma[2,2]	49.72	0.17	6.55	38.23	45.07	49.34	53.68	63.89	1446	1.00
Sigma_u[1,1]	162.62	0.31	19.63	128.29	148.57	160.99	175.41	202.97	4000	1.00
Sigma_u[2,1]	119.42	0.34	18.19	86.67	106.59	118.38	131.11	158.27	2916	1.00
Sigma_u[1,2]	119.42	0.34	18.19	86.67	106.59	118.38	131.11	158.27	2916	1.00
Sigma_u[2,2]	176.27	0.43	22.67	135.45	160.26	174.76	191.17	224.59	2766	1.00
hrt_diff	184.14	0.93	52.12	99.58	147.66	178.34	213.67	303.20	3125	1.00
bcoma_diff	-1.22	0.01	0.33	-1.88	-1.44	-1.22	-1.01	-0.56	3120	1.00
lp__	-2604.98	0.66	20.41	-2644.44	-2619.18	-2605.04	-2590.67	-2565.84	945	1.00

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. “Stan: A probabilistic programming language.” *Journal of Statistical Software* 76 (1): 1–32. doi:[10.18637/jss.v076.i01](https://doi.org/10.18637/jss.v076.i01).

Cicero, Marcus Tullius, and H. M. & Hubbell. 1949. *De Inventione. de Optimo Genere Oratorum. Topica*. Cambridge: Harvard University Press.

Hoffman, Md, and Andrew Gelman. 2014. “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo.” *Journal of Machine Learning Research* 15: 30.

Neal, Radford M. 2011. “MCMC using Hamiltonian dynamics.” *Handbook of Markov Chain Monte Carlo*, 113–62. doi:[doi:10.1201/b10905-6](https://doi.org/10.1201/b10905-6).

Reid, Nancy. 2017. “BFF Four: Are We Converging?” In.
<http://www.utstat.utoronto.ca/reid/research/reid-bff.pdf>.

Stan Development Team. 2016. “Stan Modeling Language Users Guide and Reference Manual, Version 2.15.0.” <http://mc-stan.org>.

Wasserstein, Ronald L, and Nicole A Lazar. 2016. “The Asa’s Statement on P-Values: Context, Process, and Purpose.” *Am Stat* 70 (2): 129–33.

Wong, Pauline P, Georges Monette, and Neil I Weiner. 2001. “Mathematical models of cognitive recovery.” *Brain Injury* 15 (6). Taylor & Francis: 519–30.
doi:[10.1080/02699050010005995](https://doi.org/10.1080/02699050010005995).