SCS 2017: Longitudinal and Nested Data

Statistics on Trial(s)

Frequentist vs Bayesian Inference

– Objective p-values versus subjective posteriors

Georges Monette, March 2017

P-values on trial: Sir Roy Meadow and Sally Clark

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

Y: 2 children die for no apparent cause

H_0: Sally is innocent

Y: 2 children die for no apparent cause

$$\text{p-value} = \Pr(\underline{Y}^{+} \mid H_0)$$

H_0: Sally is innocent

Y: 2 children die for no apparent cause

$$\text{p-value} = \Pr(Y^+ \mid H_0)$$

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

$$\text{p-value} = \Pr(Y^+ \mid H_0)$$

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-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-value $= \Pr(Y^+ \mid H_0)$

Meadow's calculation

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

Criticism: 1) assumes independence

2) $\frac{1}{8{,}500}$ too small

H_0 : Sally is innocent

y : 2 children die for no apparent cause

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

Meadow's calculation

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

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 $P(y^+ \mid H_0)$?

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 $P(H_0 | Y)$?

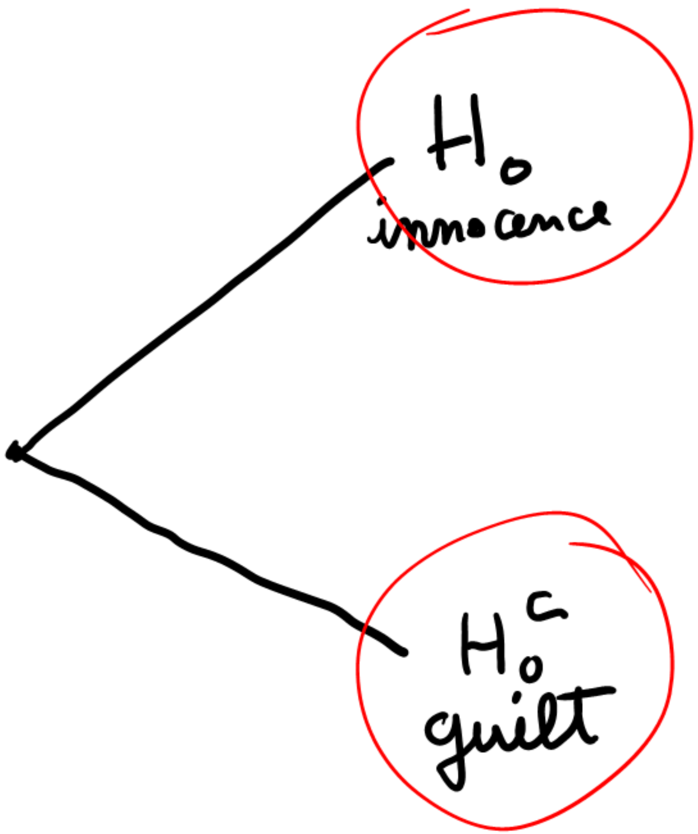
- Must be close!?
- Is $P(Y^+ | H_0)$ a good proxy for $P(H_0 | Y)$?

Let's find out:

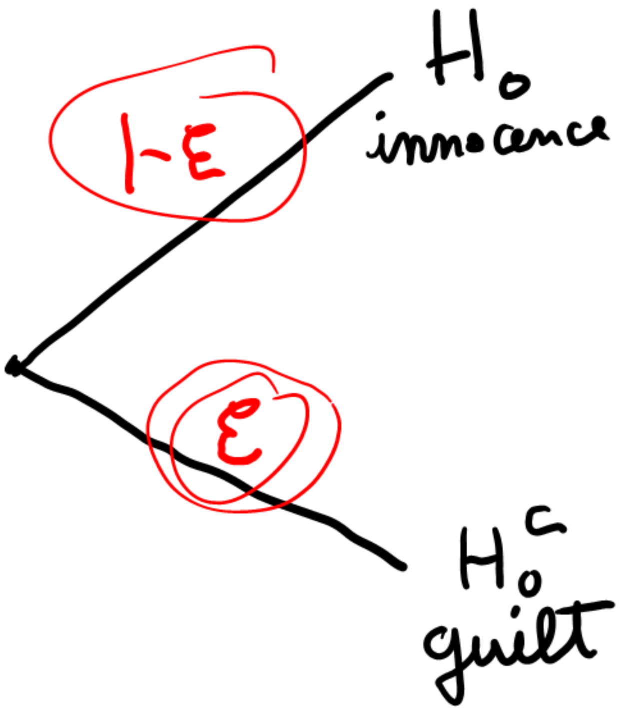
$$\underline{P(H_0 | Y)} = \frac{P(H_0, Y)}{P(Y)} = \frac{P(Y | H_0)\,\underline{P(H_0)}}{P(Y)}$$

Bayesian tree:

Bayesian tree:

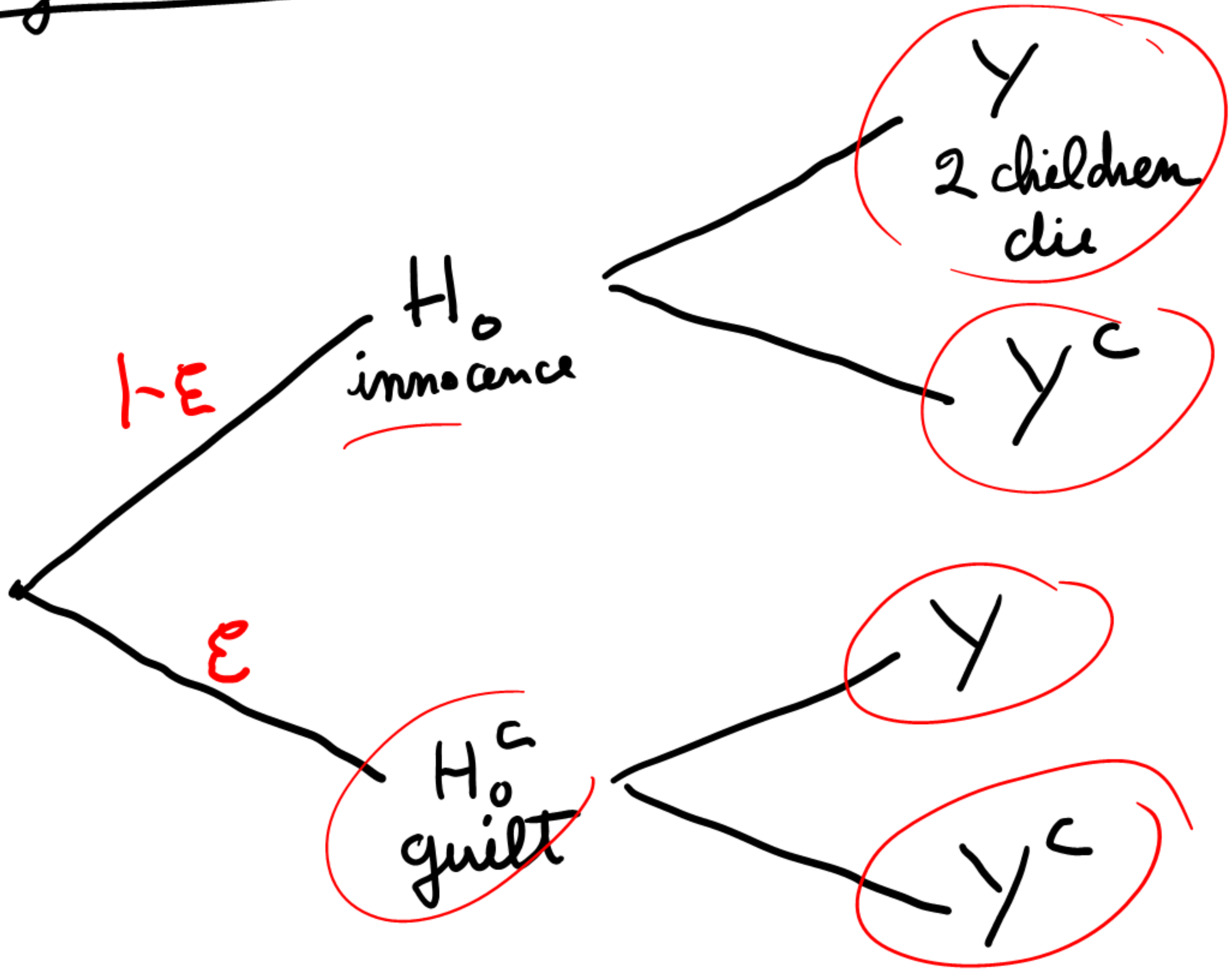


Bayesian tree:

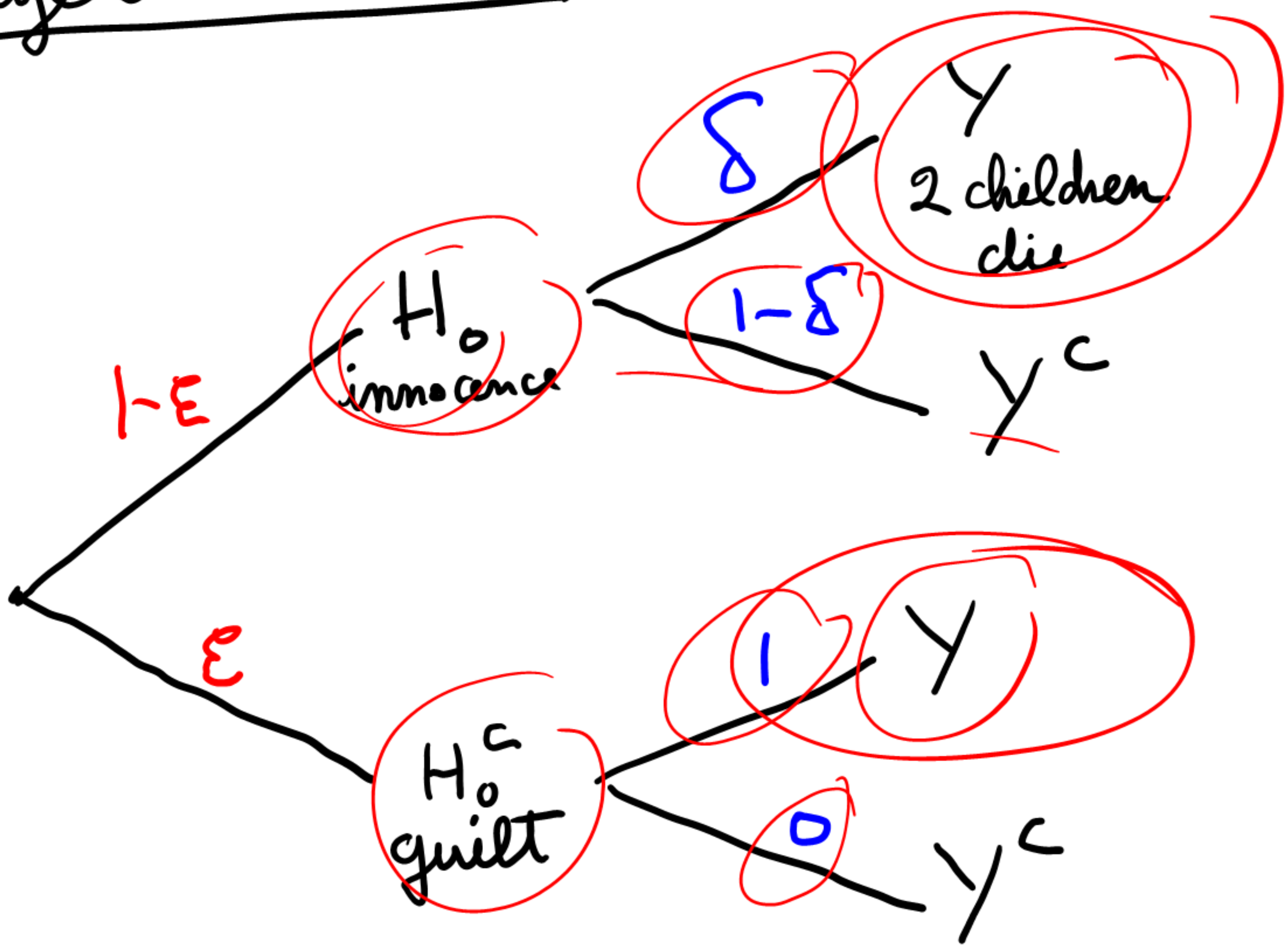


ε = very small number

Bayesian tree:

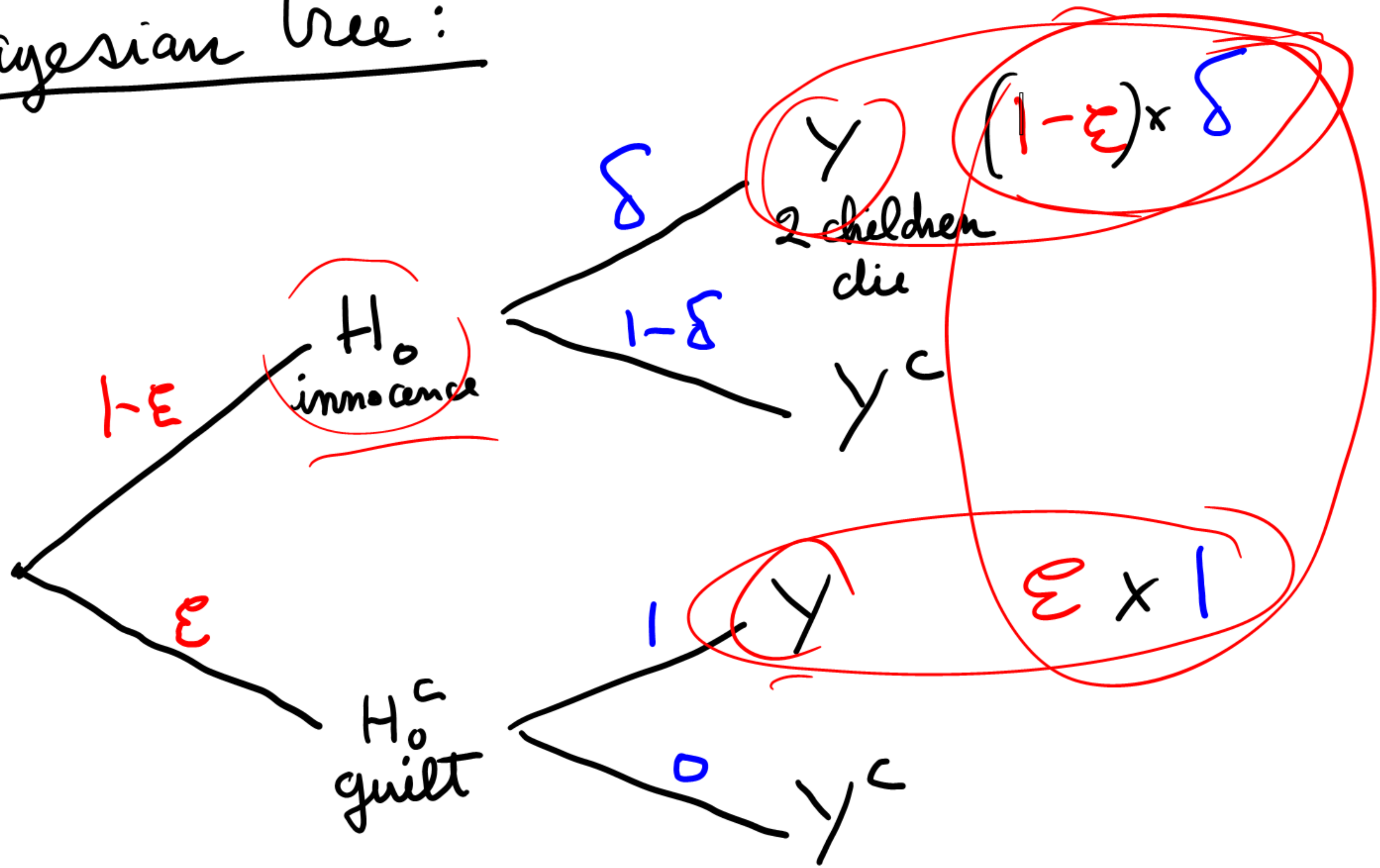


Bayesian tree:



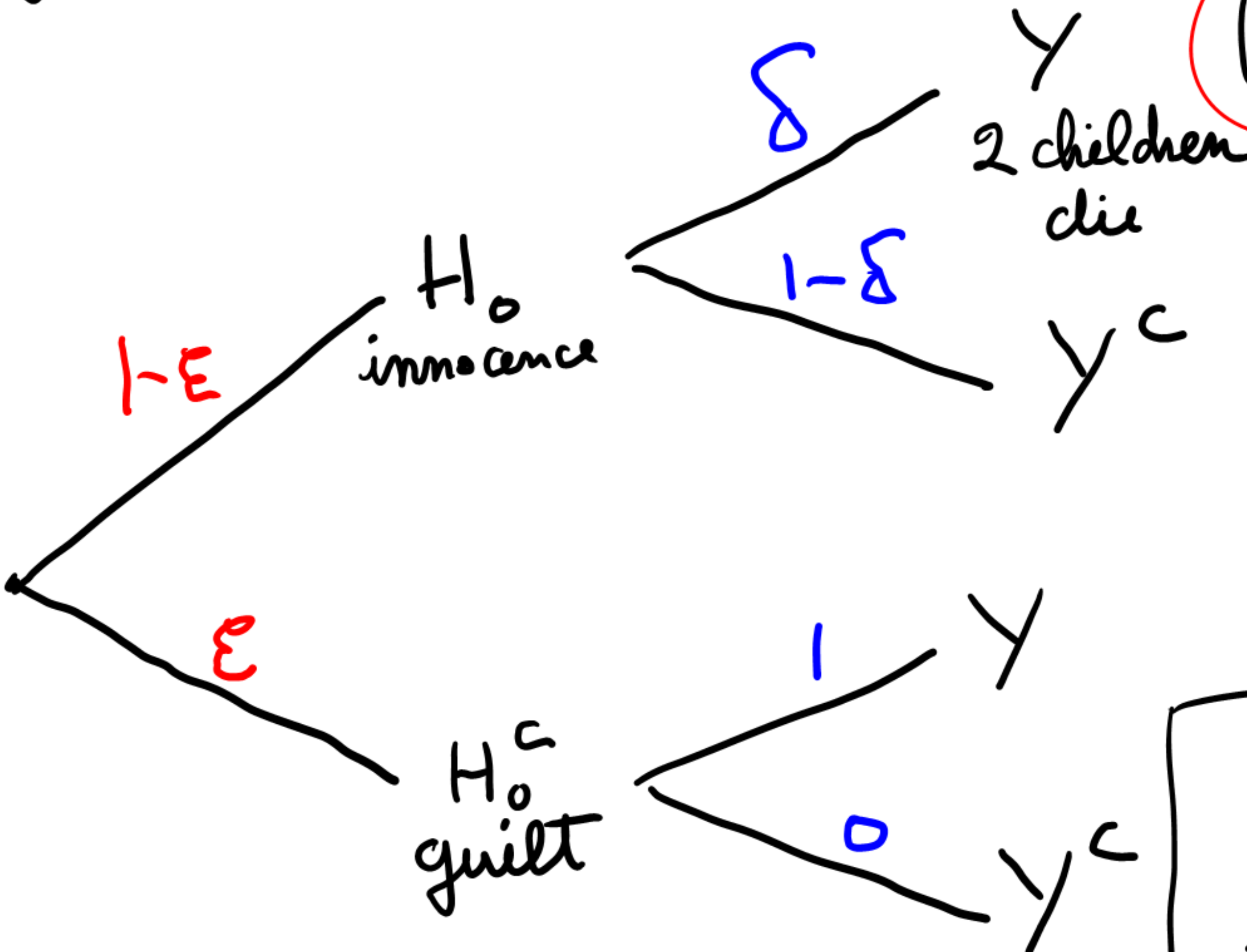
δ = very small number

Bayesian tree:



δ = very small number

Bayesian tree:

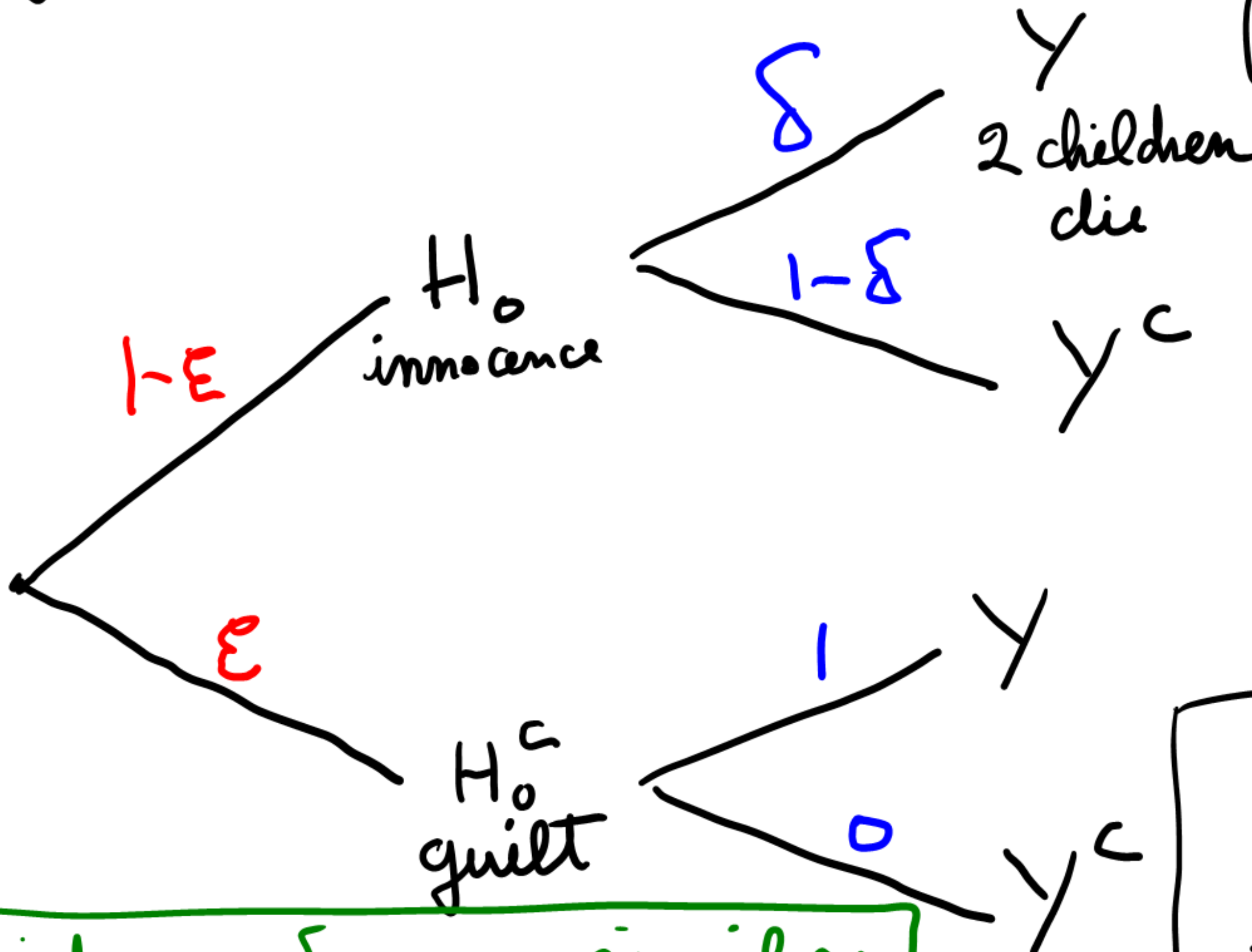


$$(1-\varepsilon)\times\delta$$

$$\varepsilon\times 1$$

$$P(H_0 \mid Y) = \frac{(1-\varepsilon)\,\delta}{(1-\varepsilon)\,\delta + \varepsilon} \approx \frac{\delta}{\delta+\varepsilon}$$

Bayesian tree:



$(1-\varepsilon)\times\delta$

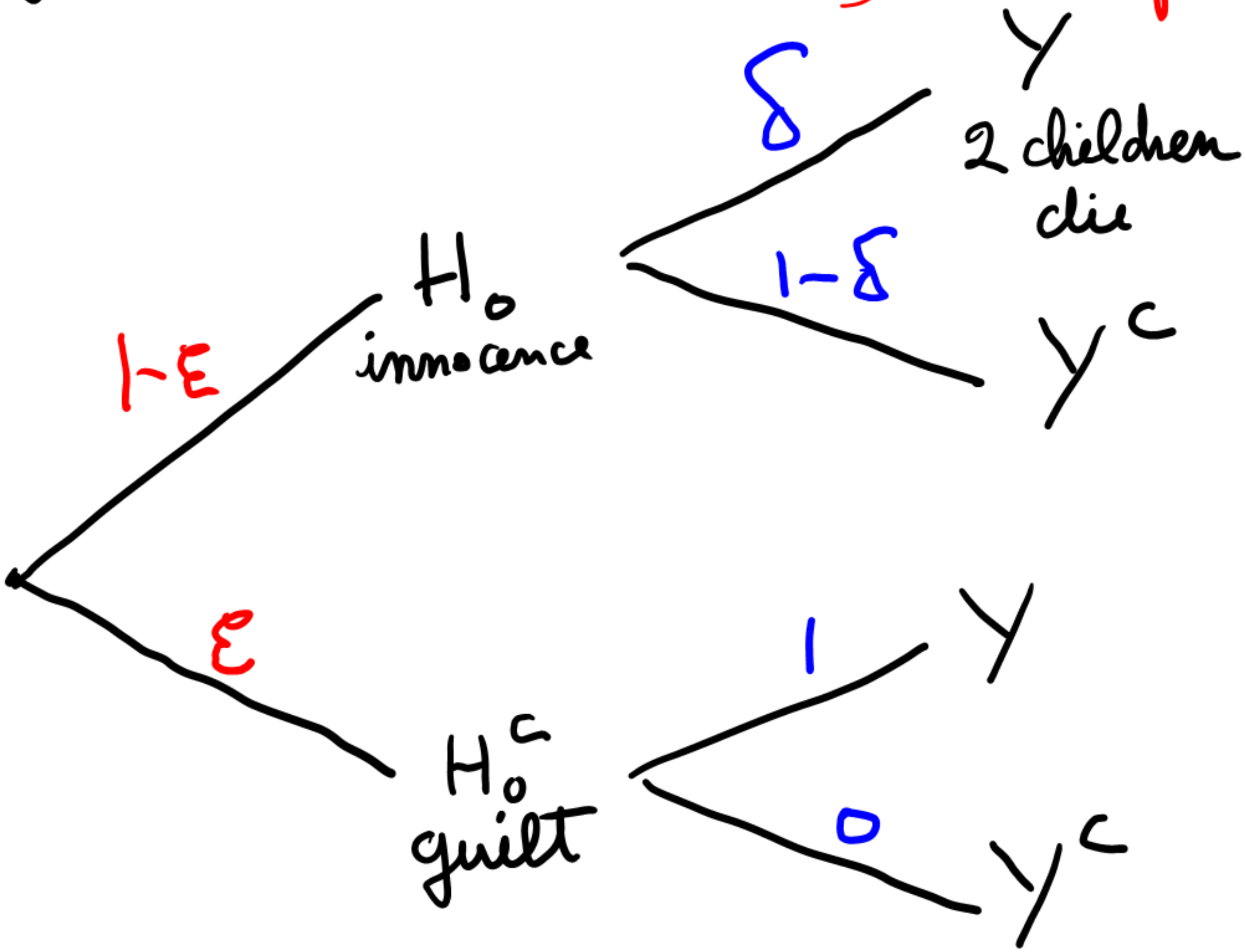
$\varepsilon\times 1$

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

So if ε & δ are similar then $P(H_0|Y) \approx \frac{1}{2}$

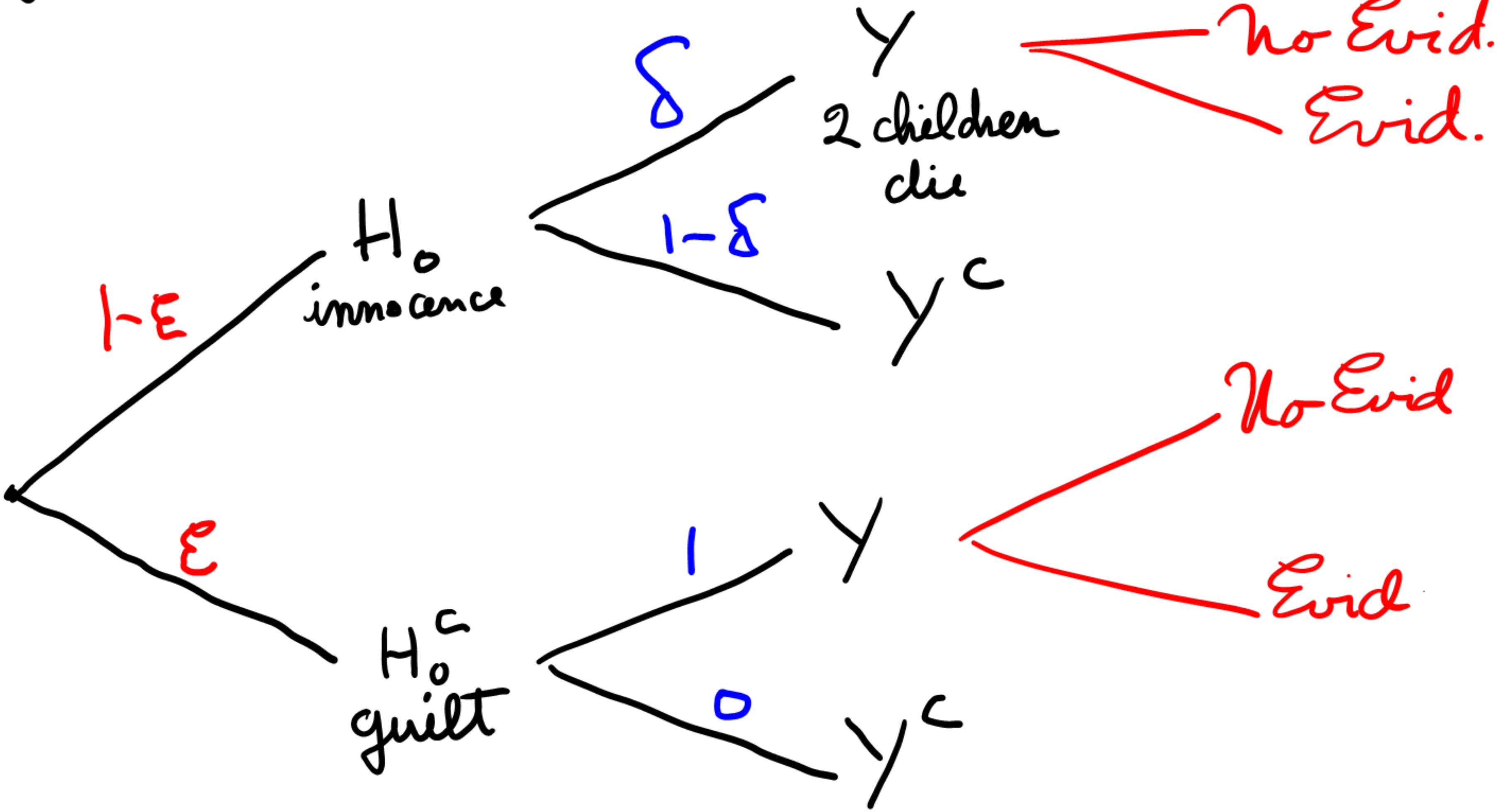
Bayesian tree:

But this ignores lack of evidence

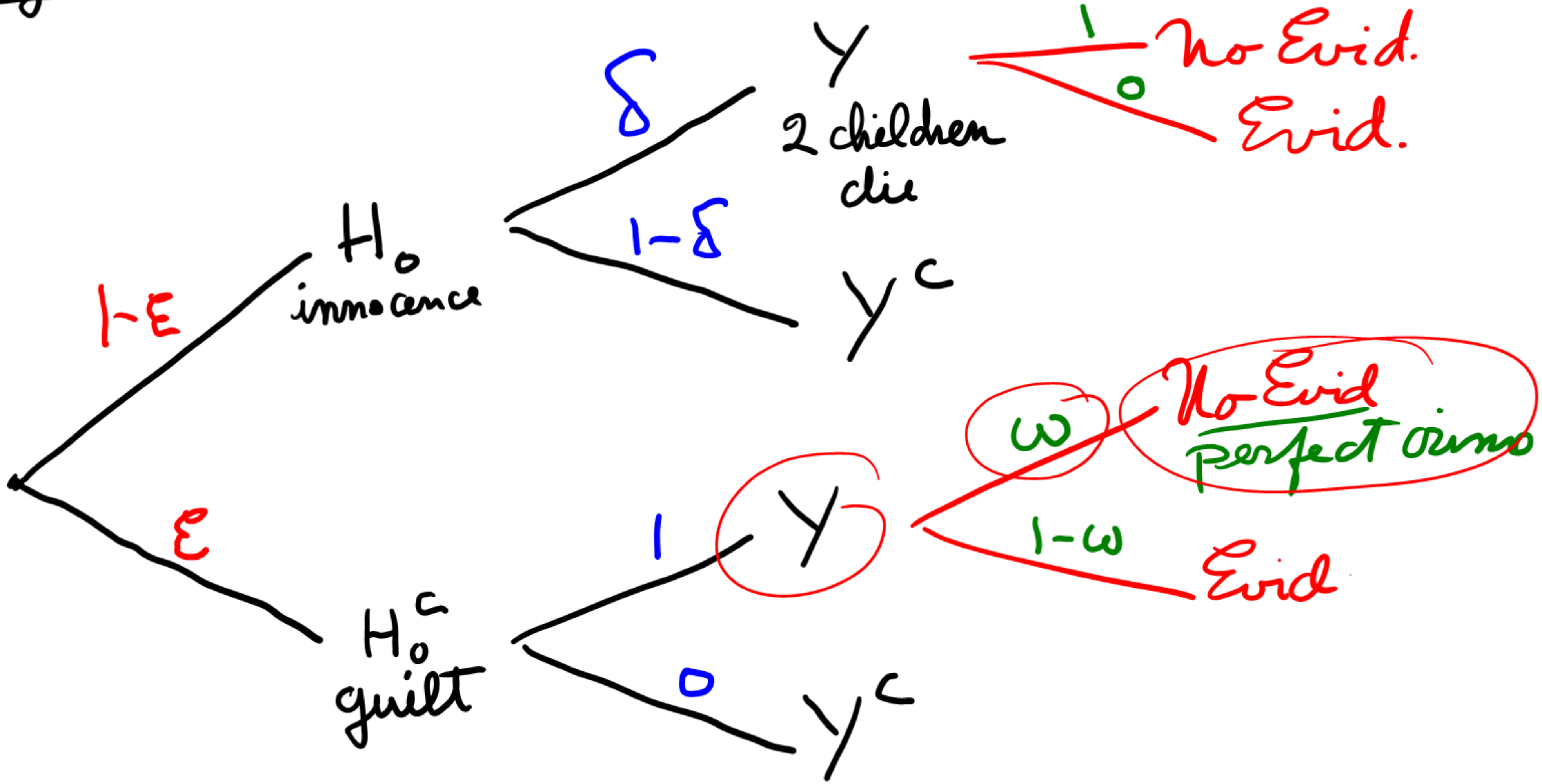


If it's a crime, it's a "perfect" crime – and, presumably, only a small proportion of crimes are perfect crimes.

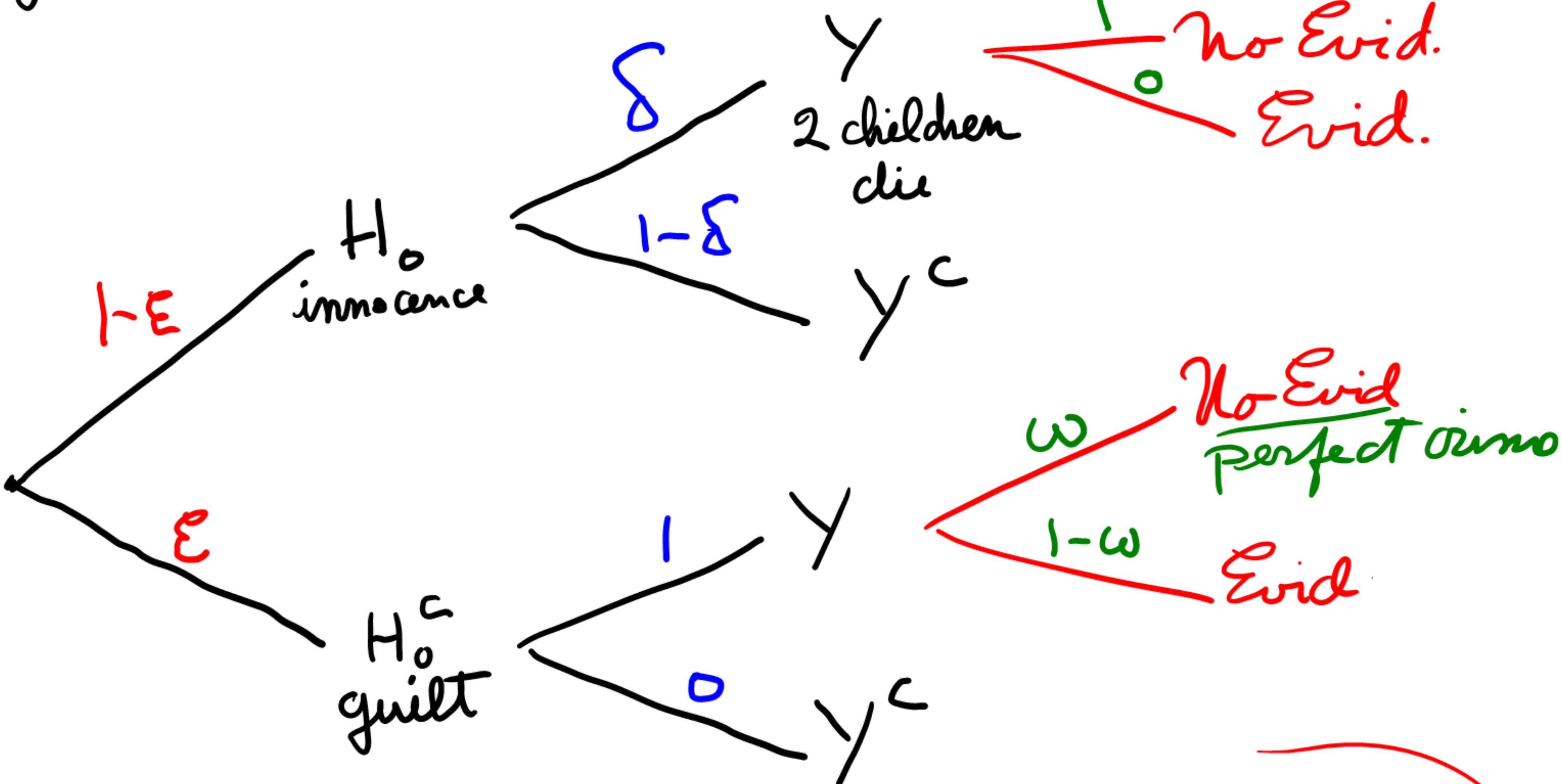
Bayesian tree:



Bayesian tree:



Bayesian tree:



$$\Pr(H_0 \mid Y \,\&\, \text{No Evid}) \approx \frac{\delta}{\delta + \cancel{\varepsilon\omega}} \approx \text{close to } 1 \,!!$$

Sally Clark was innocent beyond a reasonable doubt.

$P(Y^+ \mid H_0)$

was a very poor proxy for

$P(H_0 \mid Y)$

$P(Y^+ | H_0)$

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 $P(H_0 \mid Y)$ has come to be known as the "Prosecutor's Fallacy"

- Using the p. value as a proxy for $P(H_0 \mid Y)$ 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 $P(H_0 \mid y)$ 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

Θ : unknown state of nature

Dilemma

Θ : unknown state of nature

Y : what's observed

Dilemma

Θ : unknown state of nature

Y : what's observed

What does Y say about Θ?

Dilemma

Θ : unknown state of nature

Y : what's observed

What does Y say about Θ?

Model : $P(Y \mid \Theta)$

Dilemma

Θ : unknown state of nature

Y : what's observed

What does Y say about Θ?

Model : $P(Y \mid \Theta)$

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 : $p(Y \mid \Theta)$

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

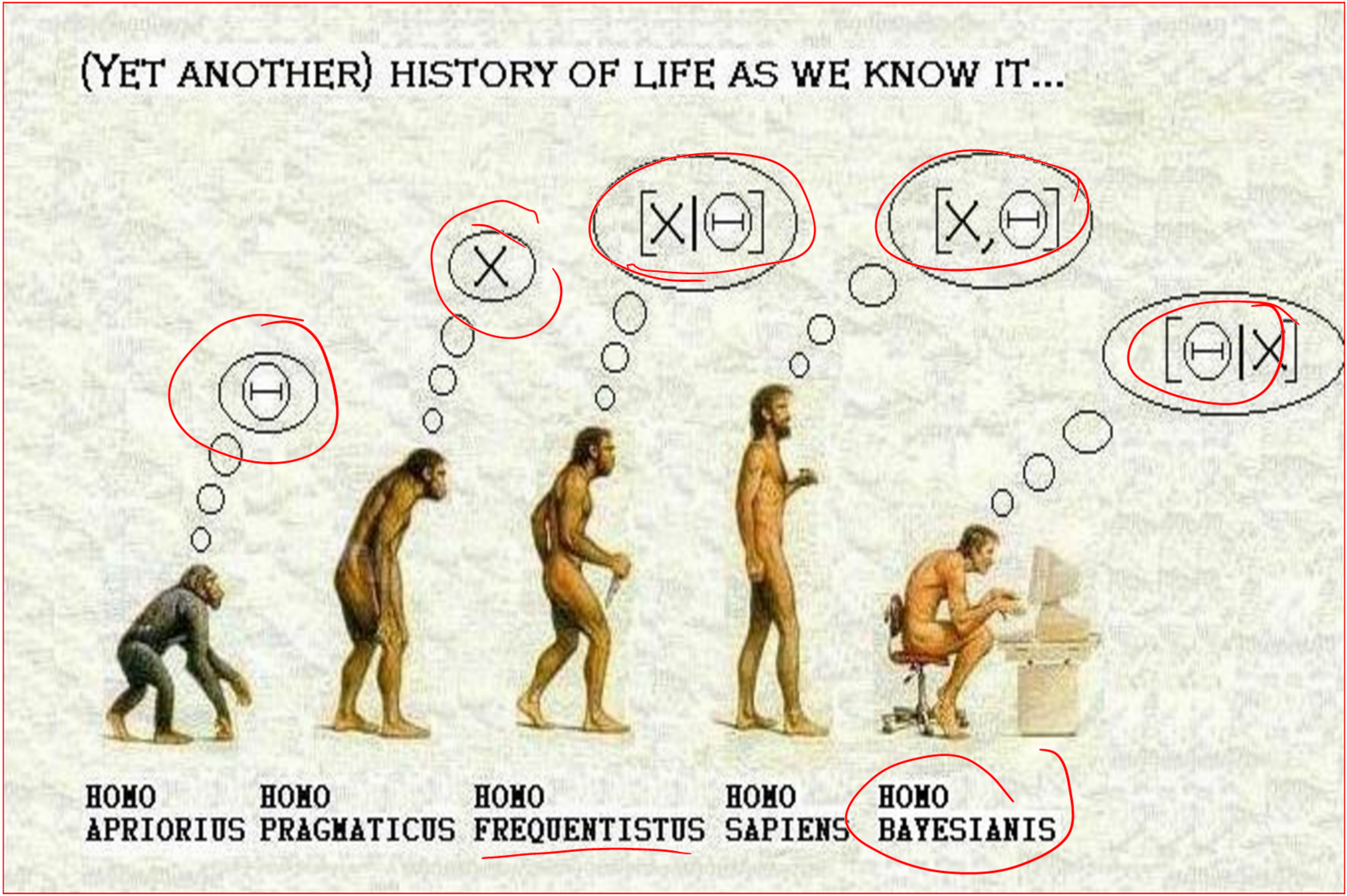
But not $p(\Theta \mid Y)$ posterior

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

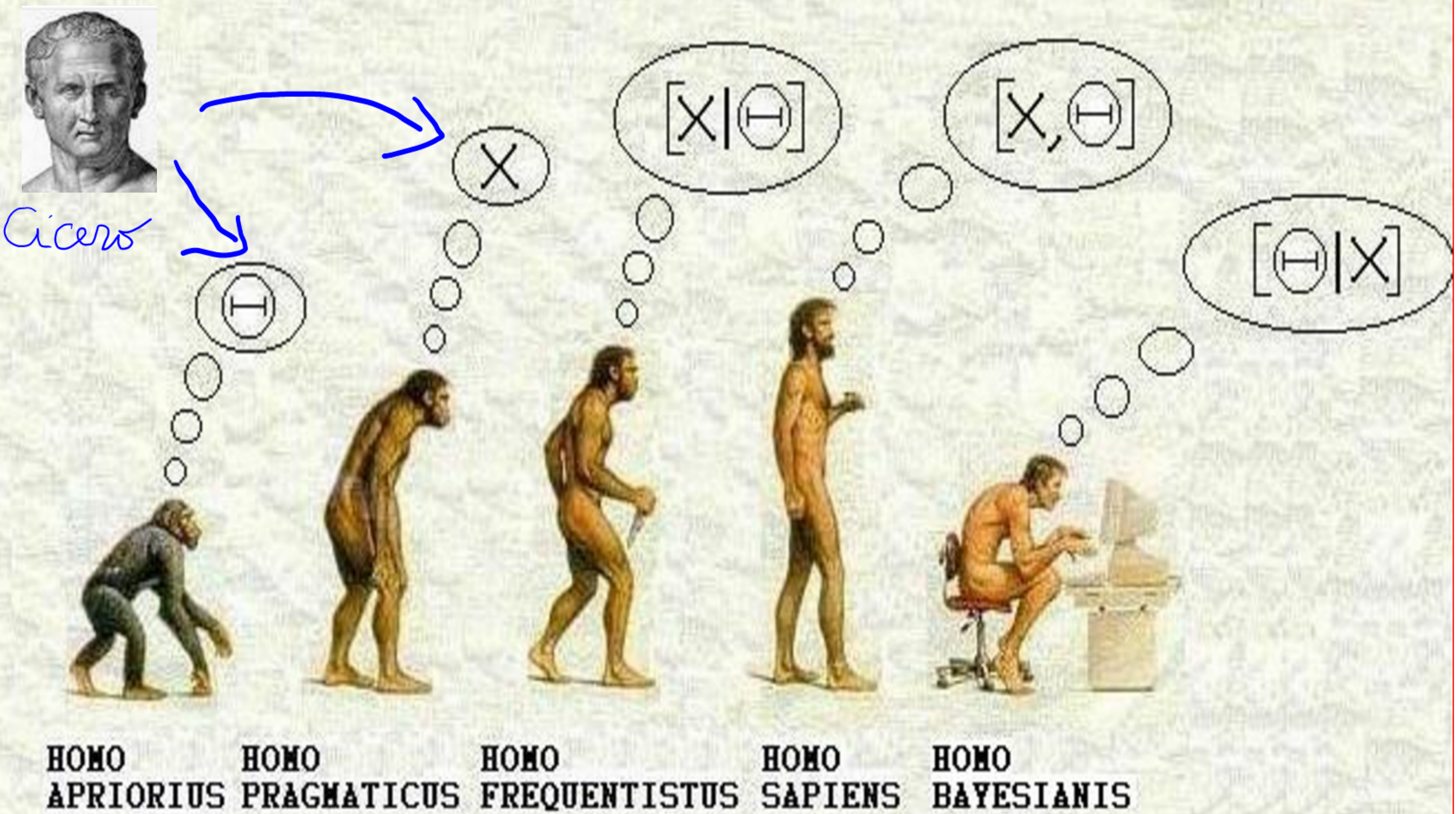
To get what you really want, you need to pay a price many are reluctant to pay.

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

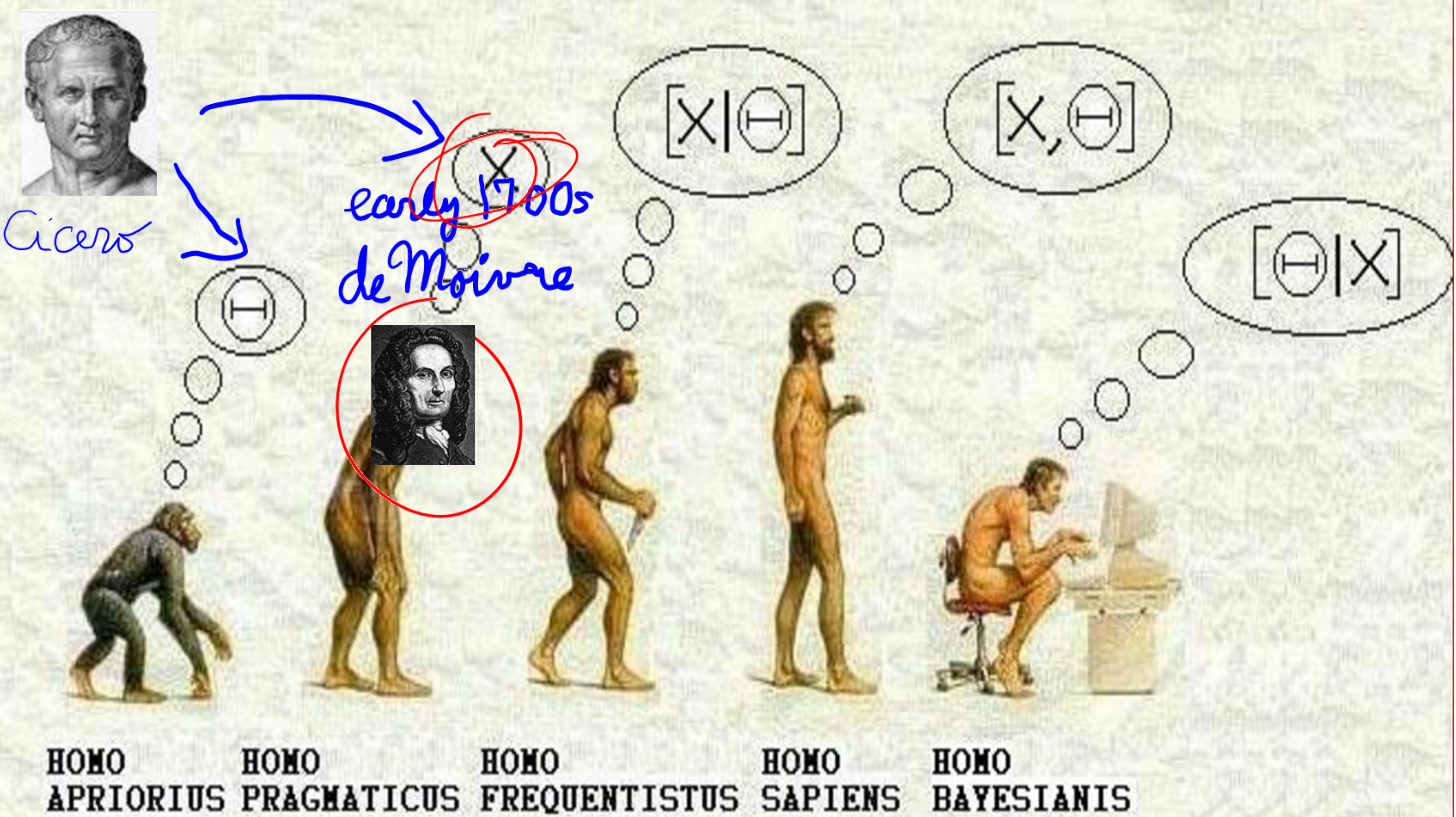
A historical perspective

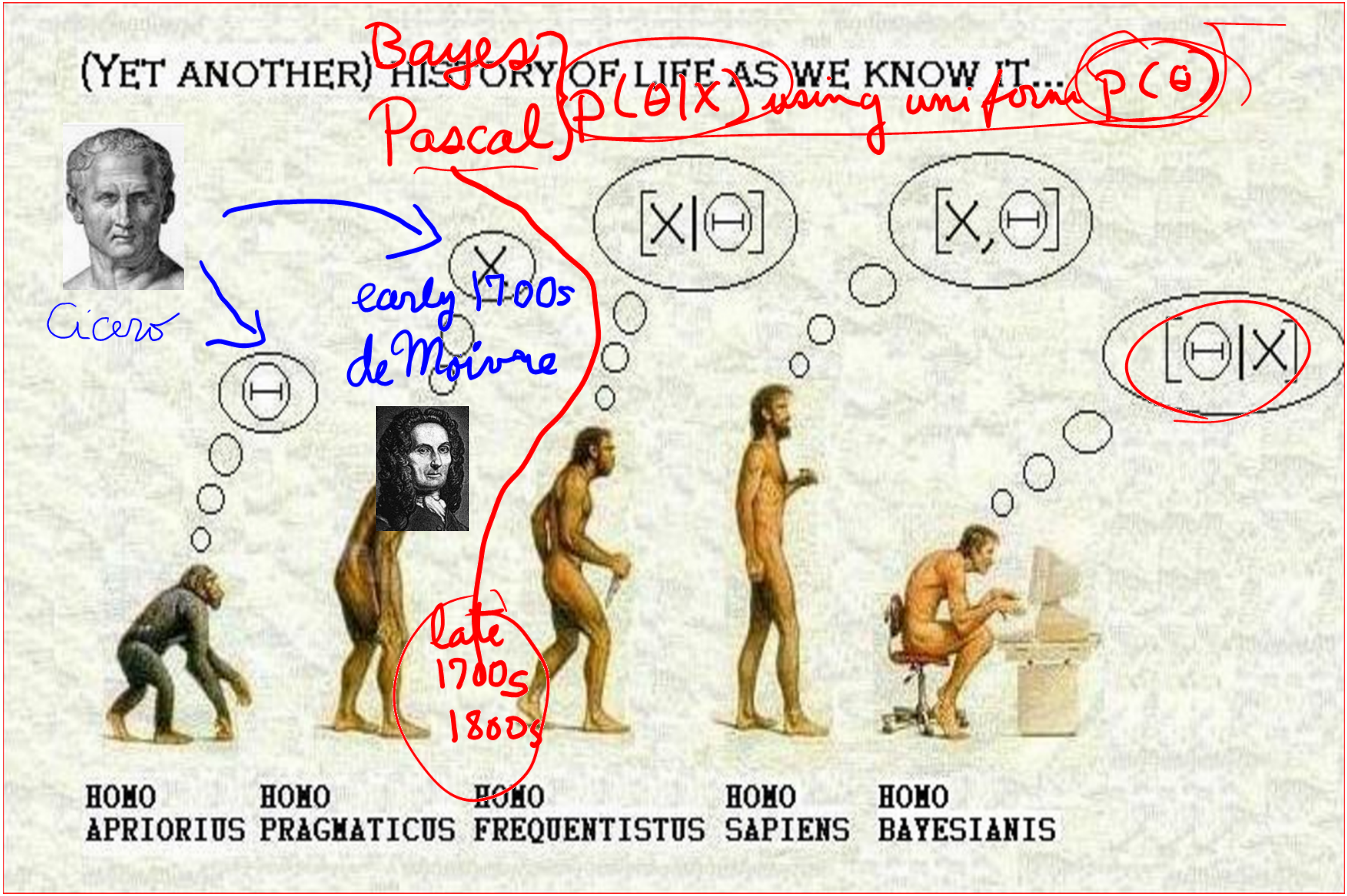

(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...
Θ
X
[X|Θ]
[X,Θ]
[Θ|X]
HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS

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



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

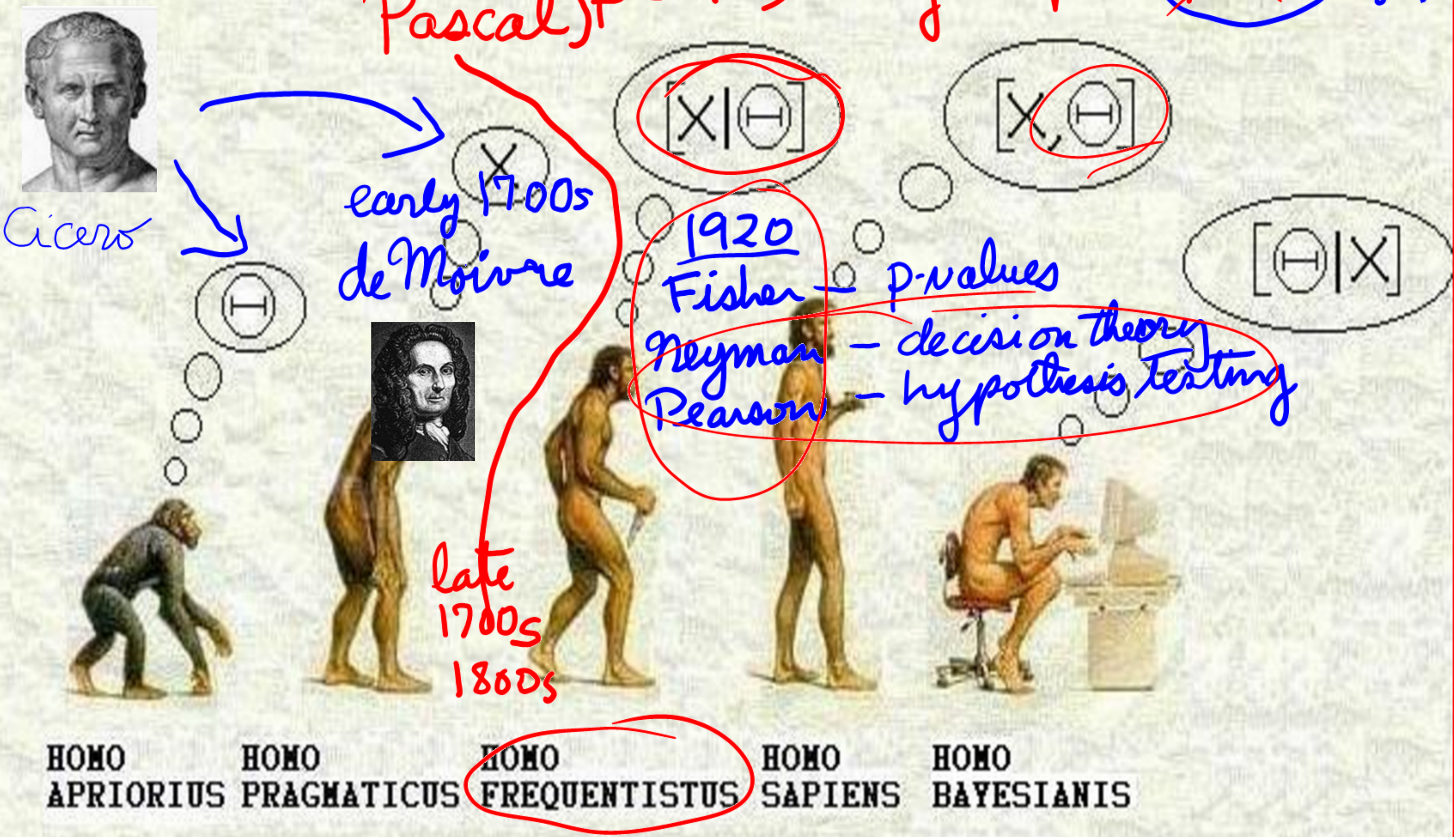


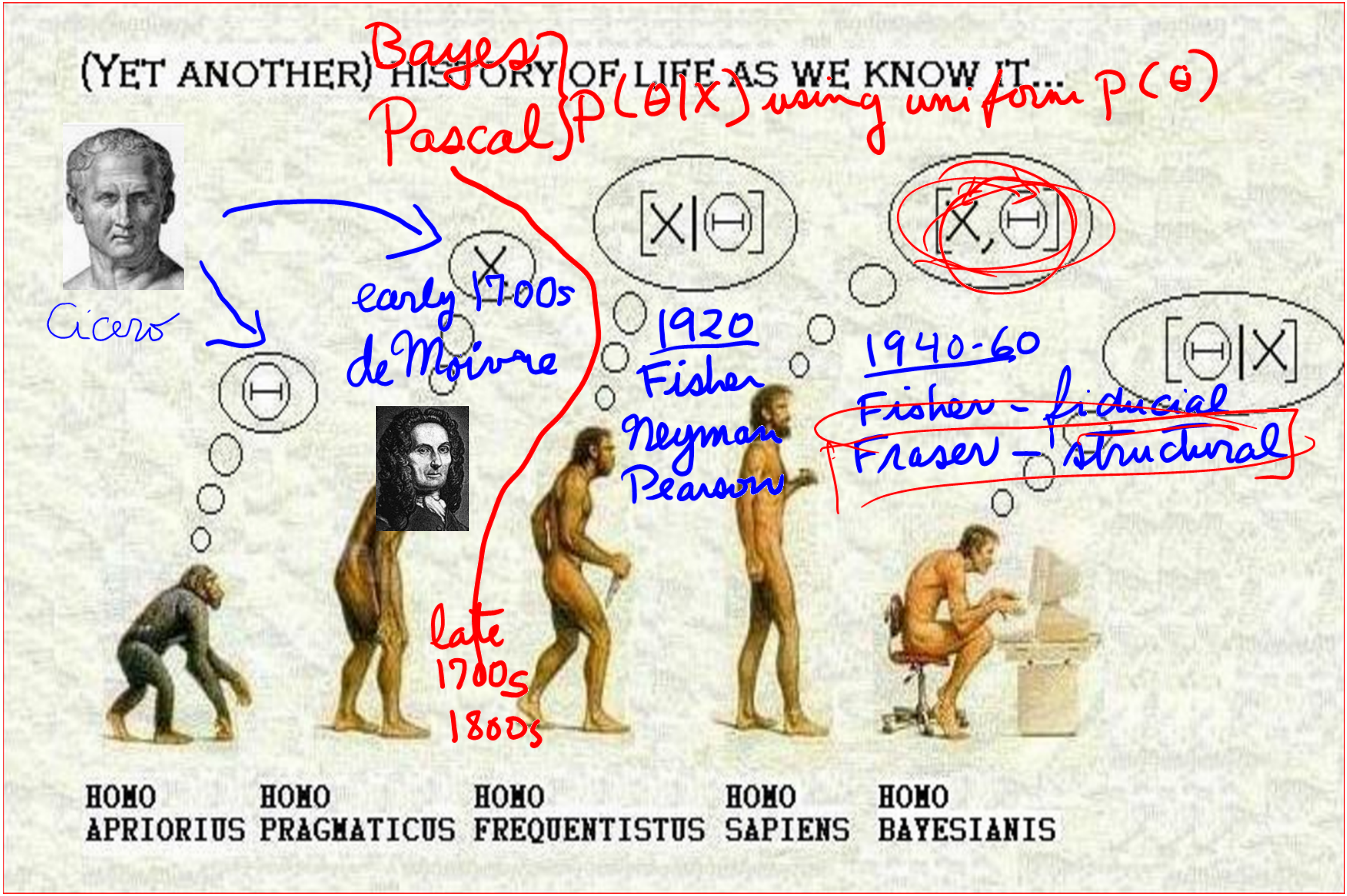


(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...
Bayes
Pascal
P(θ|X) using uniform P(θ)
Cicero
[Θ]
early 1700s
de Moivre
[X]
[X|Θ]
[X,Θ]
[Θ|X]
late 1700s 1800s
HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS

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

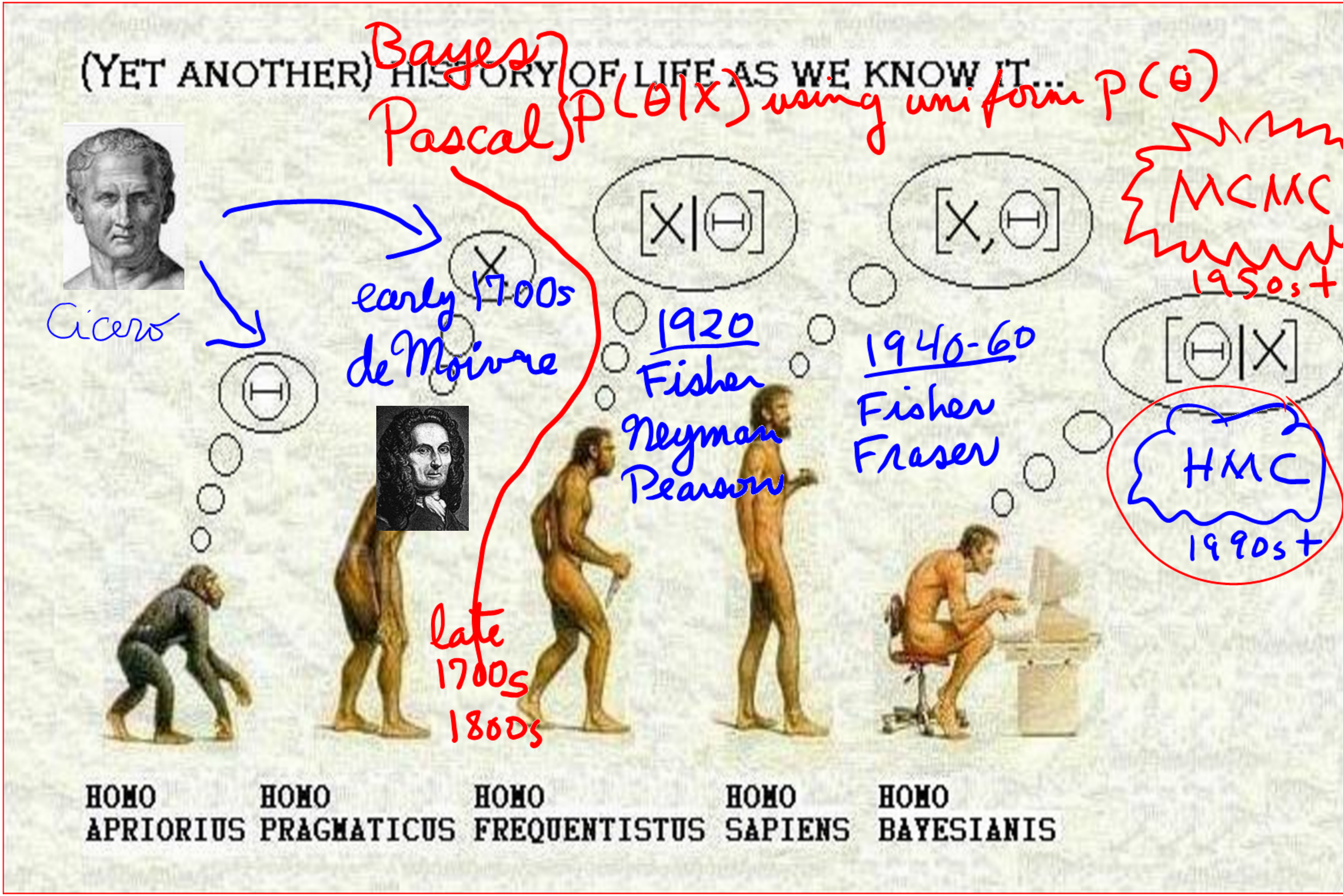
Bayes
Pascal } $P(\Theta|X)$ using uniform $\cancel{P(\Theta)}$??

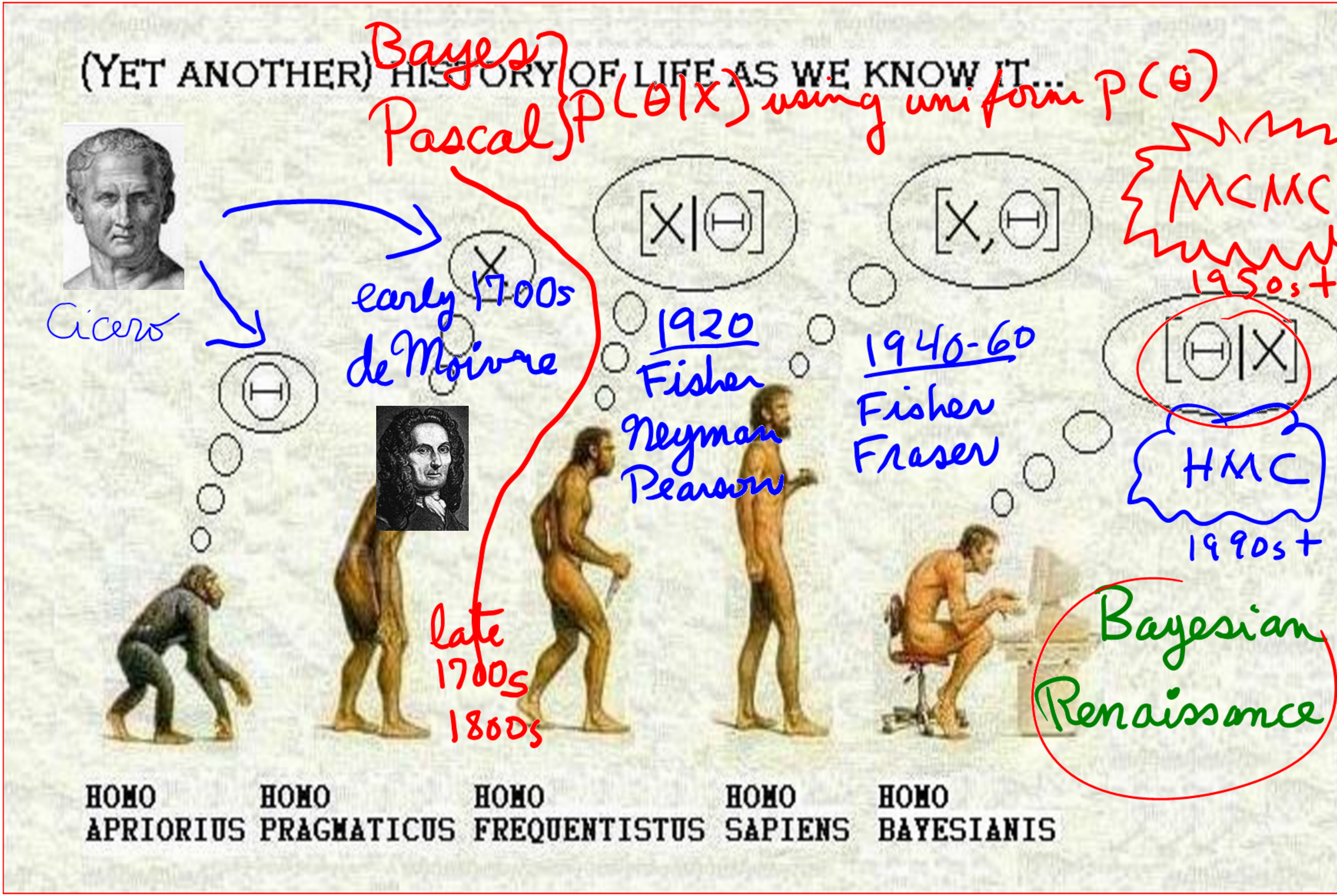



(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...
Bayes
Pascal
P(θ|X) using uniform P(θ)
Cicero
[Θ]
[X]
early 1700s
de Moivre
late 1700s
1800s
[X|Θ]
1920
Fisher
Neyman
Pearson
[X,Θ]
1940-60
Fisher – fiducial
Fraser – structural
[Θ|X]
HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS

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

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


(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...
Bayes
Pascal
$P(\theta|X)$ using uniform $P(\theta)$
Cicero
Θ
X
early 1700s
de Moivre
late 1700s
1800s
$[X|\Theta]$
1920
Fisher
Neyman
Pearson
$[X,\Theta]$
1940-60
Fisher
Fraser
MCMC
1950s+
$[\Theta|X]$
HMC
1990s+
HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS


(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...
Bayes
Pascal
P(θ|X) using uniform P(θ)
Cicero
[Θ]
[X]
early 1700s
de Moivre
late 1700s
1800s
[X|Θ]
1920
Fisher
Neyman
Pearson
[X,Θ]
1940-60
Fisher
Fraser
MCMC
1950s+
[Θ|X]
HMC
1990s+
Bayesian Renaissance
HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS

~~Basic~~ Philosophical 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 $P(\theta)$

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

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

$P(A|B) = \dfrac{P(A,B)}{P(B)}$

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

Posterior

~~Basic~~ Philosophical problem

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

To get $P(\theta|X)$

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

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

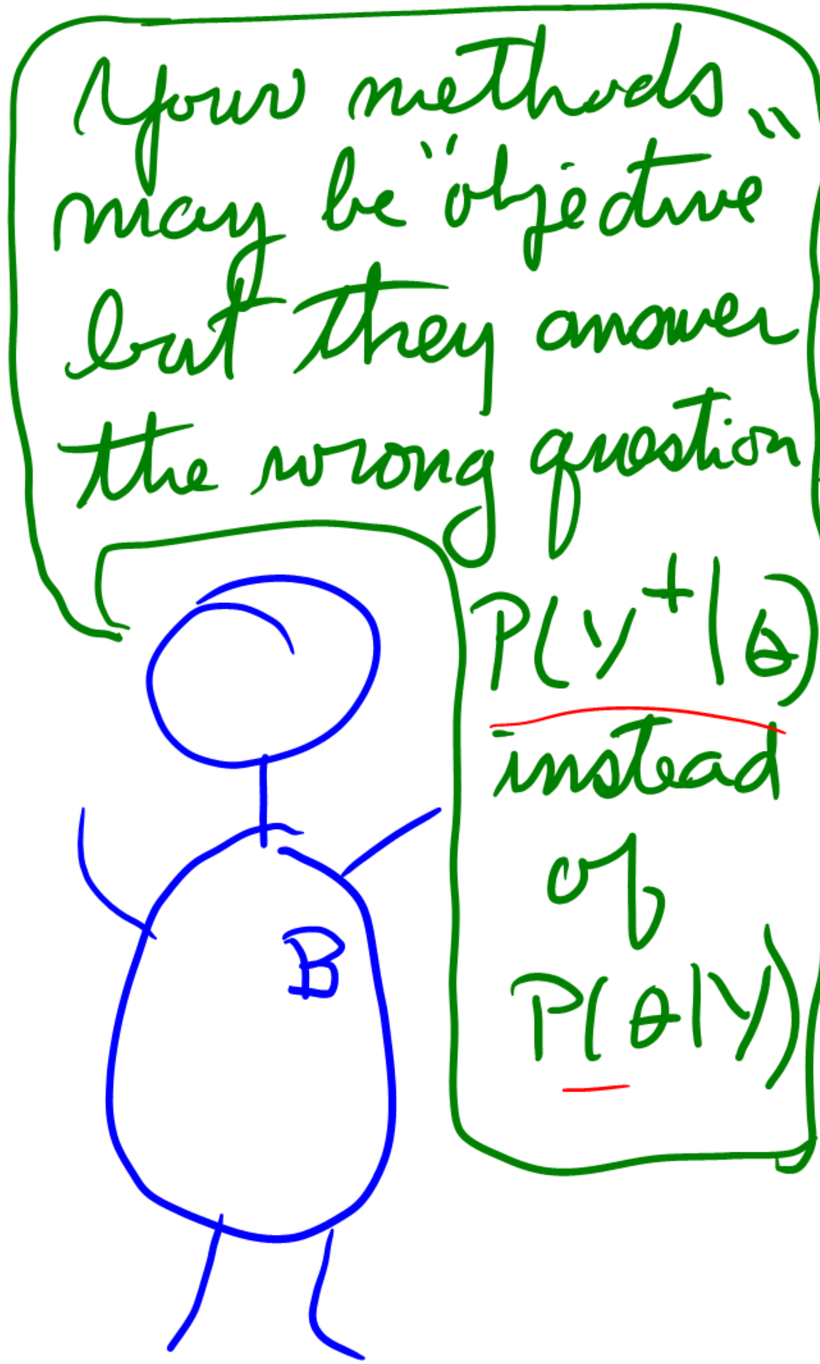
and $\underbrace{P(\theta|X)}_{\text{Posterior}} = \dfrac{P(X,\theta)}{P(X)}$

- 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 $P(X \mid \theta)$ 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

Fisher

Your methods may be "objective" but they answer the wrong question

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

Practical problem:

$$p(X, \theta) = p(X \mid \theta)\, p(\theta)$$

Practical problem:

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

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

Practical problem:

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

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

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

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:

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

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

MCMC (mid 20th C.)

comes to the rescue:

It's possible to sample from $P(\theta \mid X)$ 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

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

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

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

Baking the Bayesian omelette without breaking the Bayesian egg.

But don't take it for granted
Remember Sally Clark

Emerging practice:
Using weakly informative priors.

Markov Chain Monte Carlo

Use

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

Markov Chain Monte Carlo

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

joint model × 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 $P(\theta, X) = P(X|\theta)\, P(\theta)$

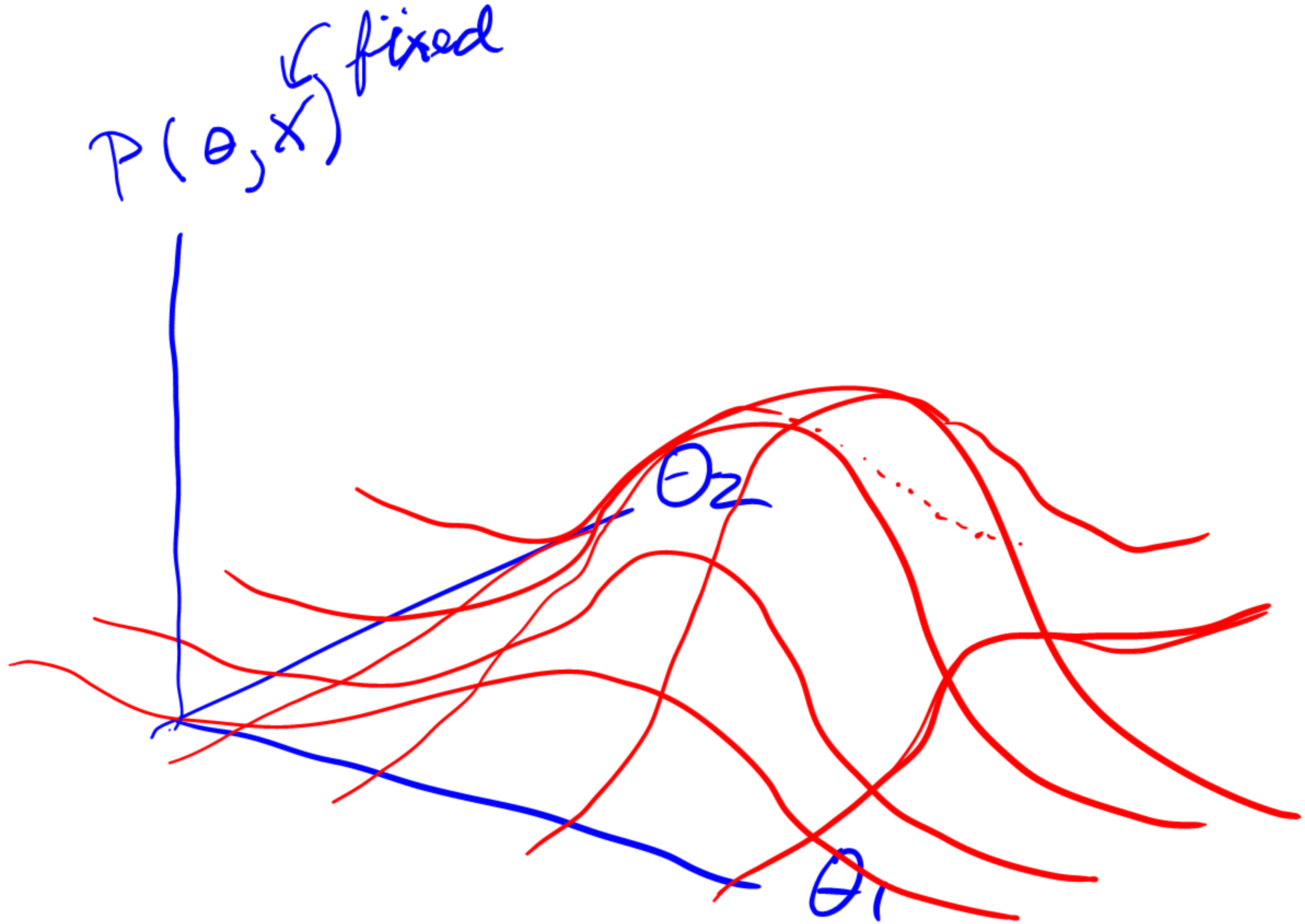
joint model × prior

Samples from $P(\theta|X)$ using only $P(\theta, X)$

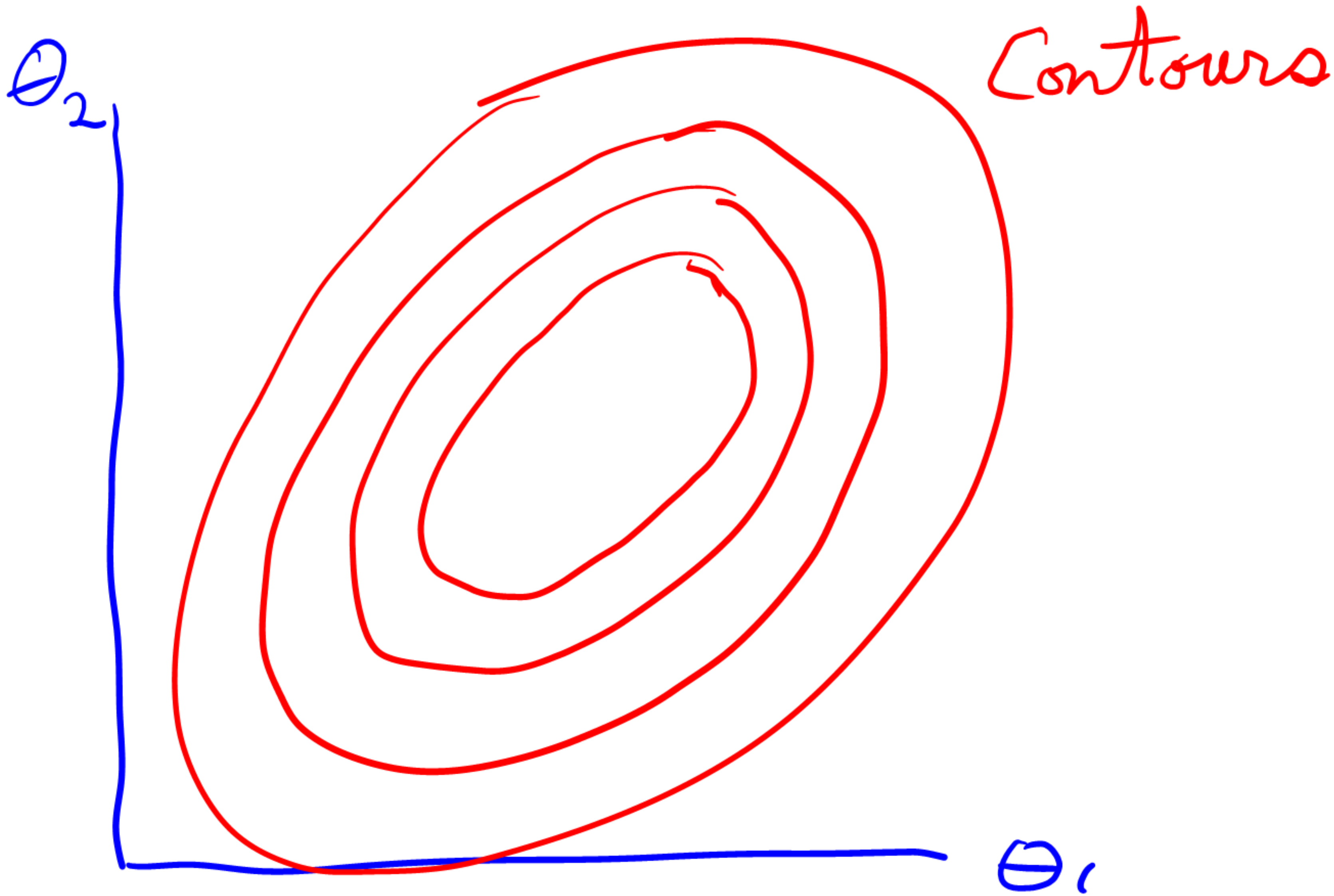
i.e. no need to find elusive $P(X)$

With X fixed, think of $P(\theta, X)$

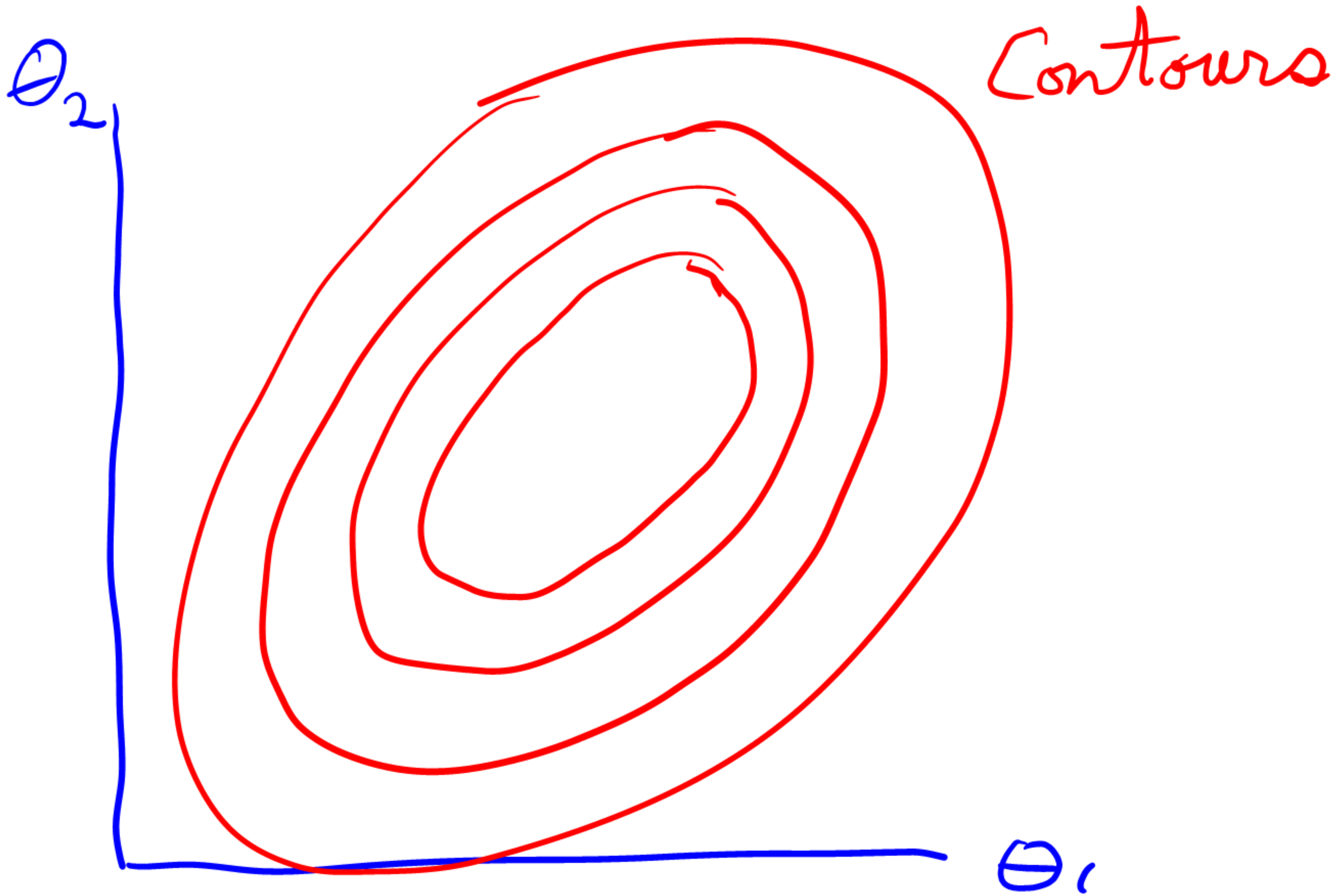
as defining a mountain over θ space



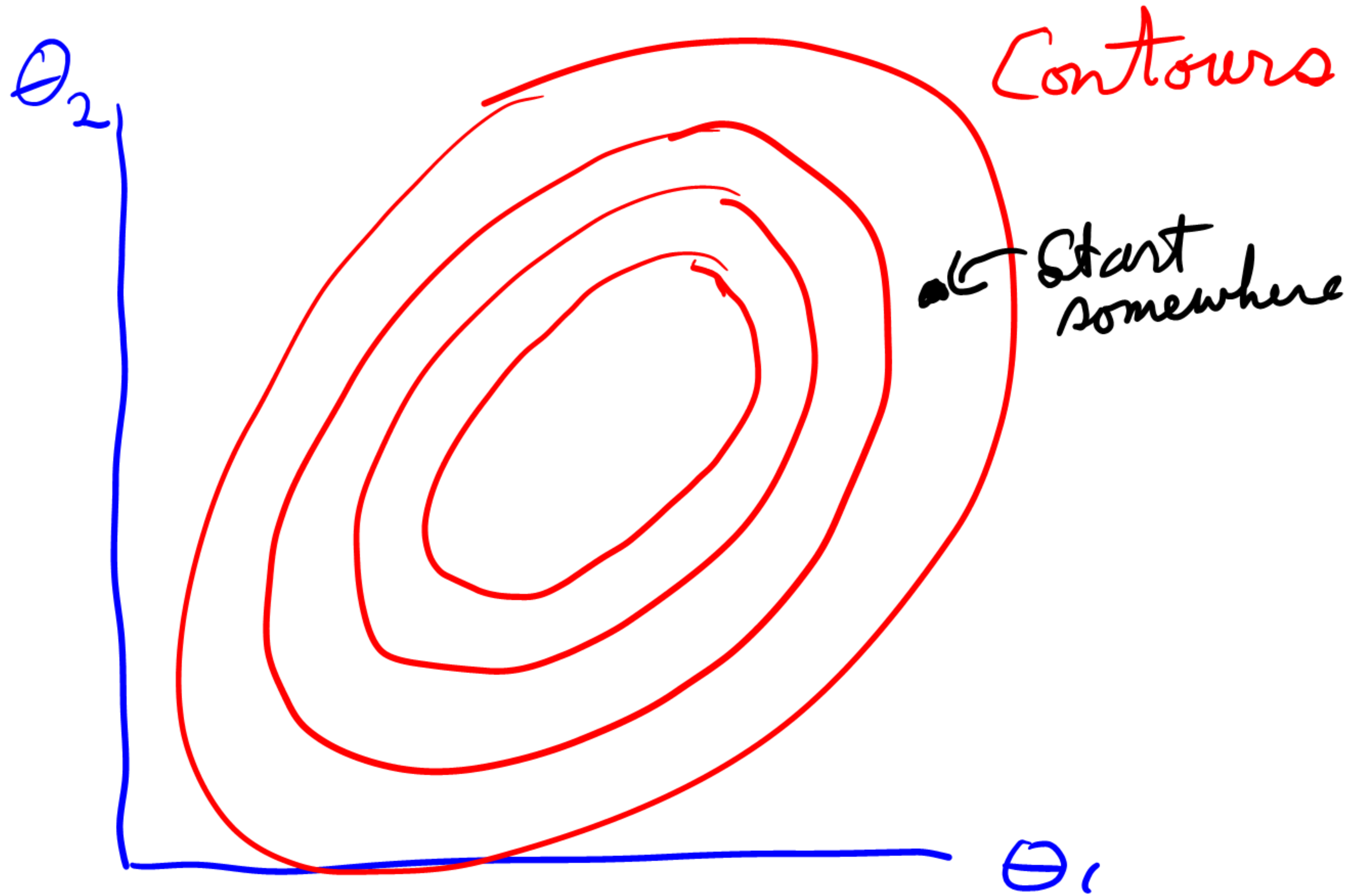
$P(\theta, x)$
fixed
θ_2
θ_1



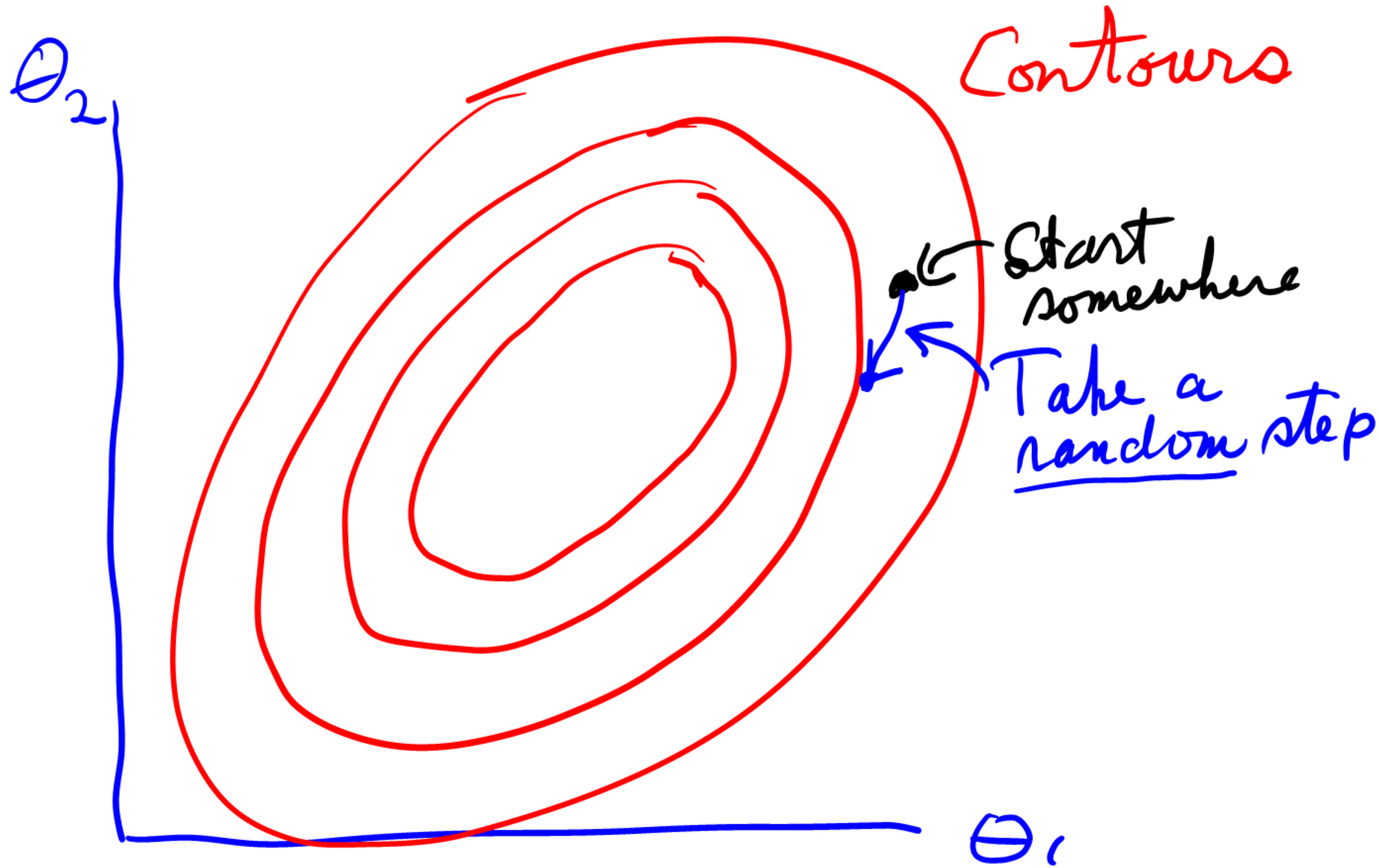
Contours
θ_2
θ_1



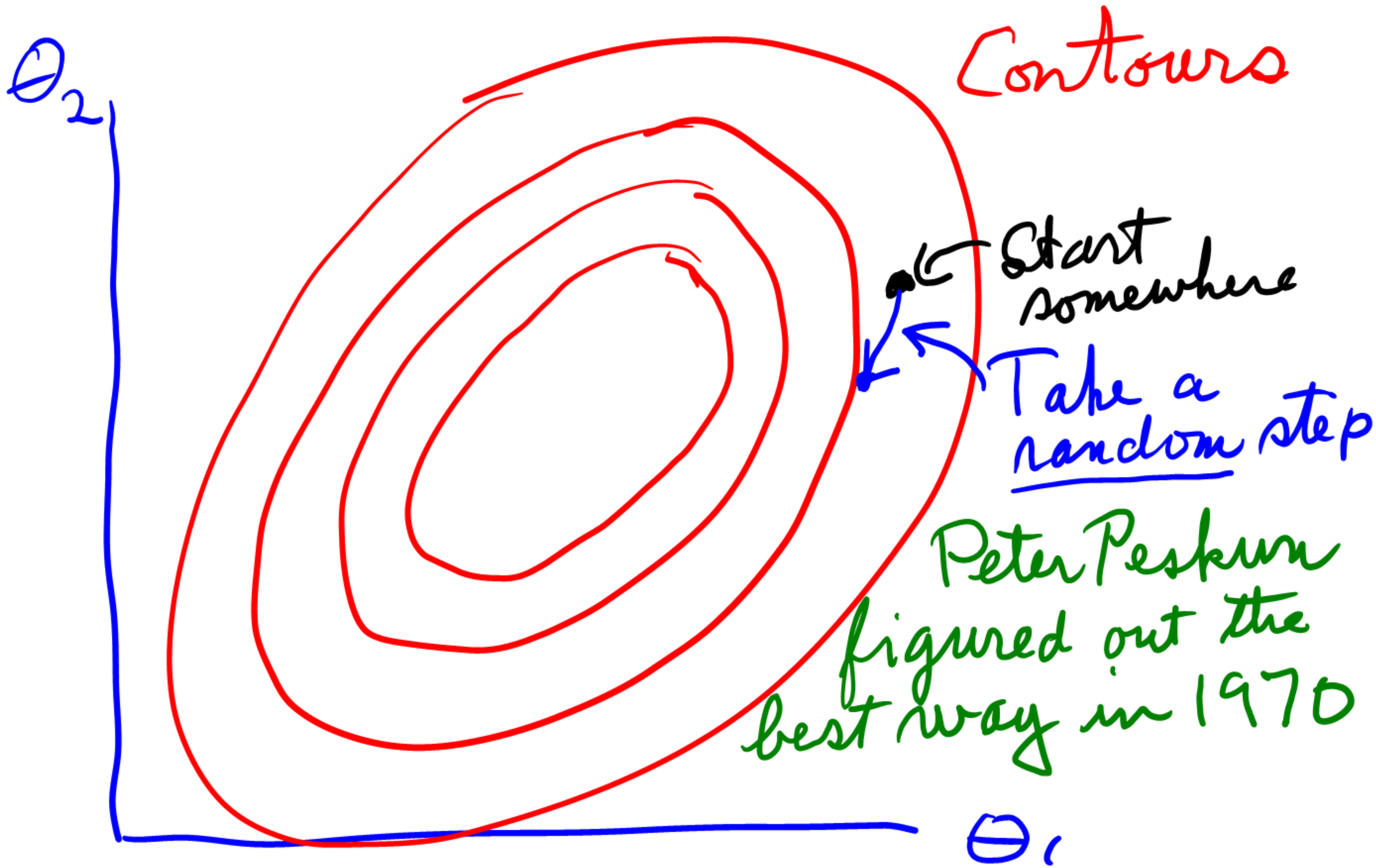
Metropolis-Hastings algorithm



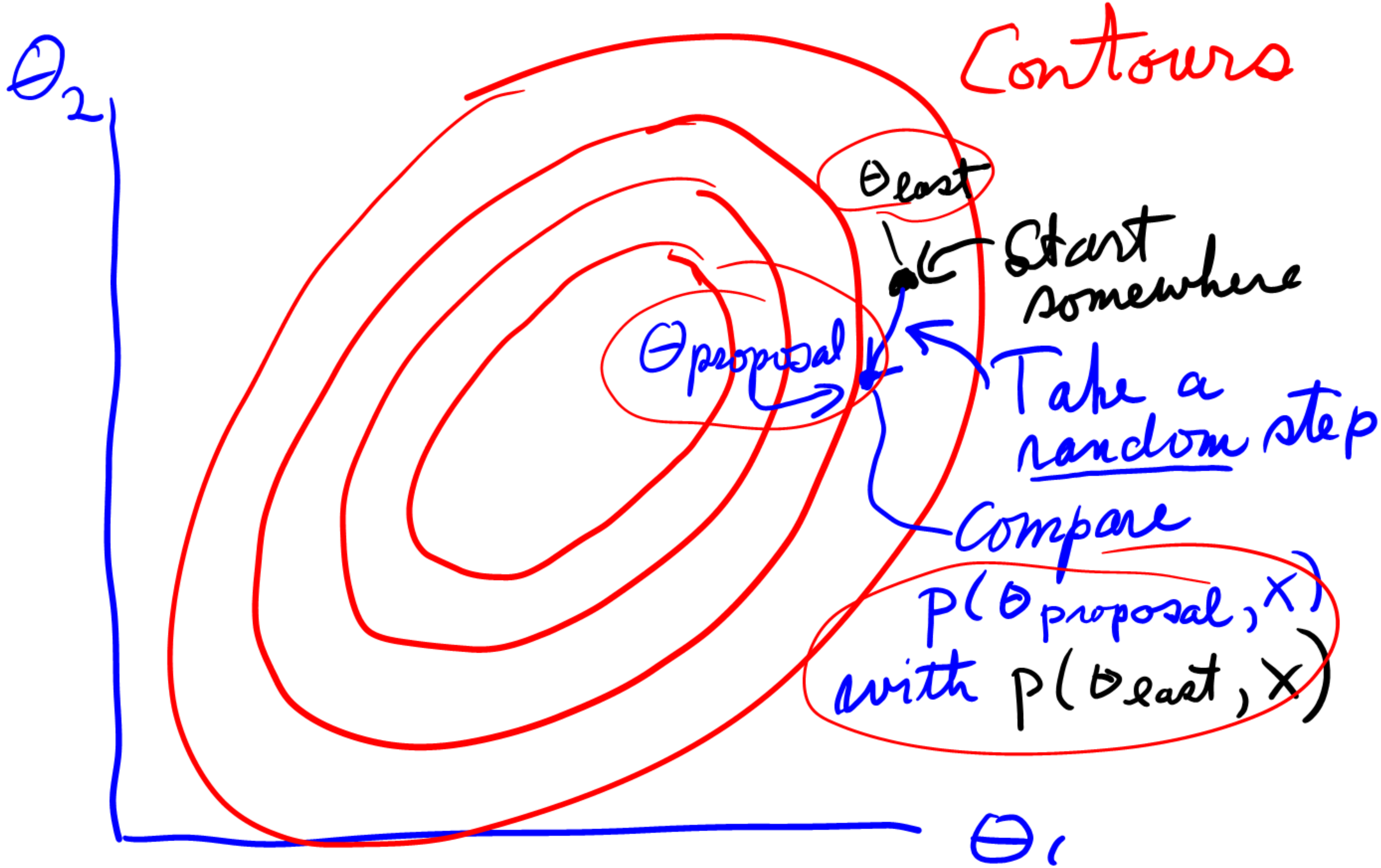
Metropolis-Hastings algorithm



Metropolis-Hastings algorithm

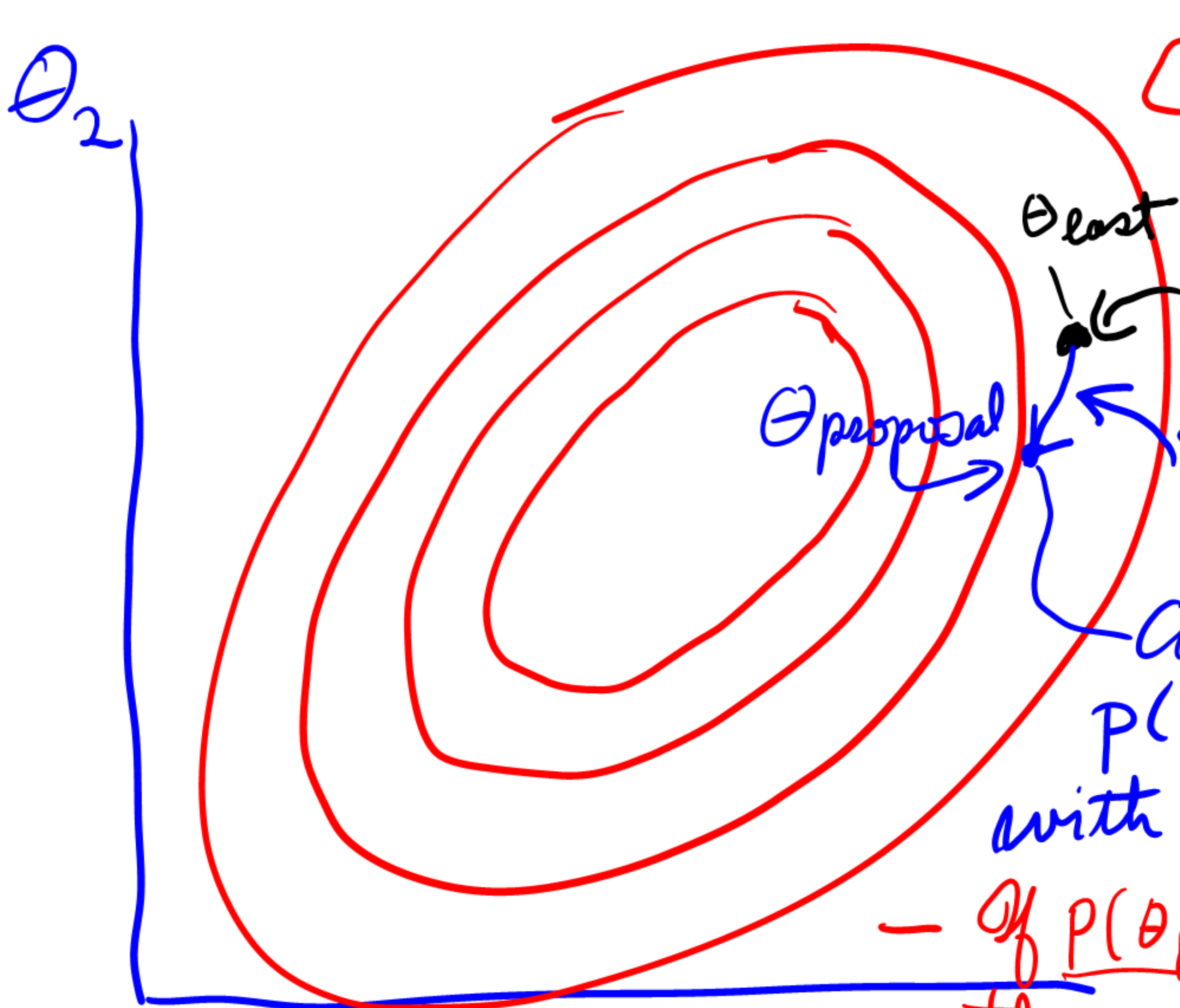


Metropolis-Hastings algorithm



Metropolis-Hastings algorithm

Contours

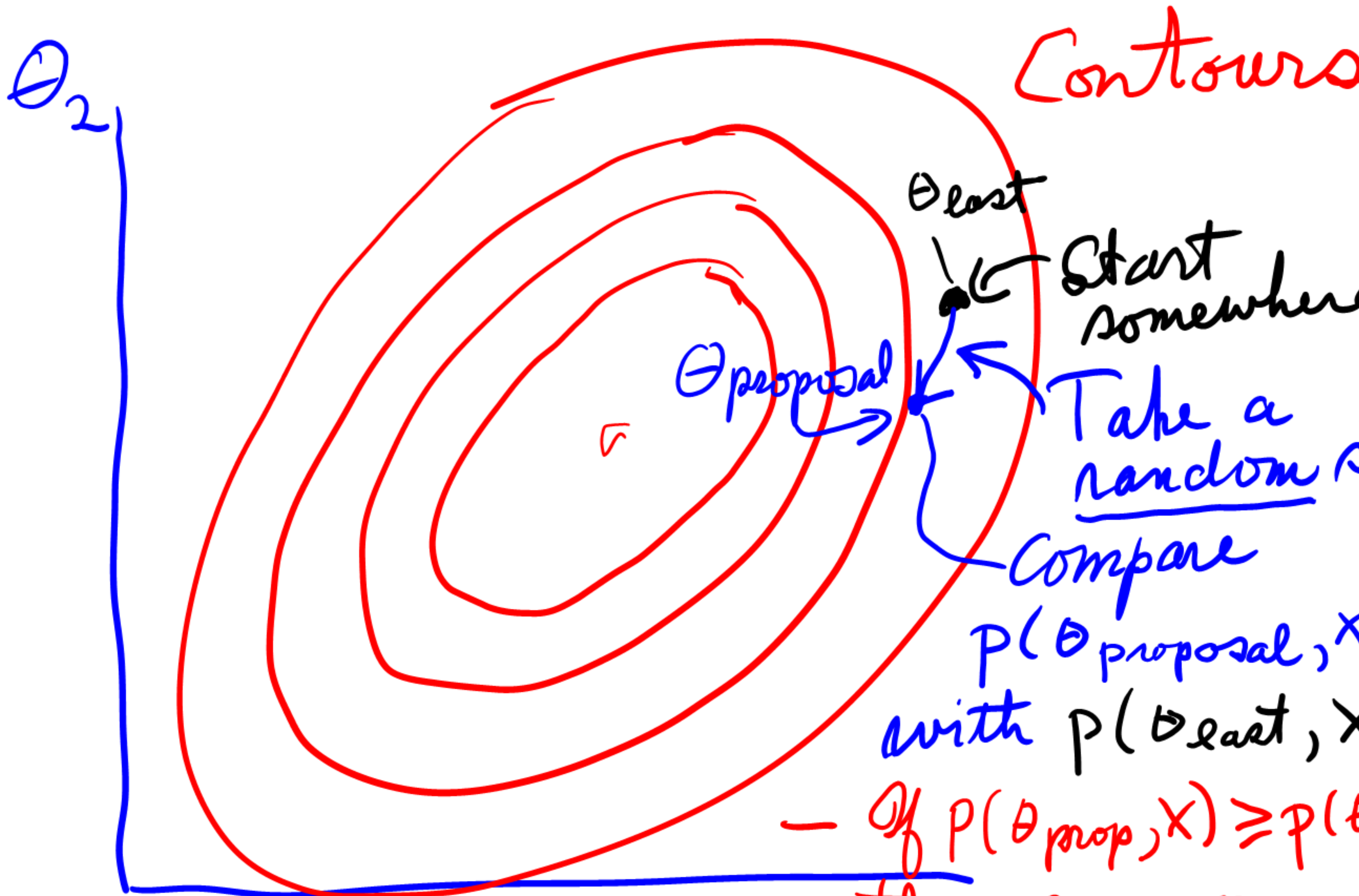


Start somewhere

Take a random step

Compare $p(\theta_{proposal}, X)$
with $p(\theta_{last}, X)$

– If $p(\theta_{prop}, X) \geq p(\theta_{last}, X)$
then $\theta_{new} = \theta_{proposal}$

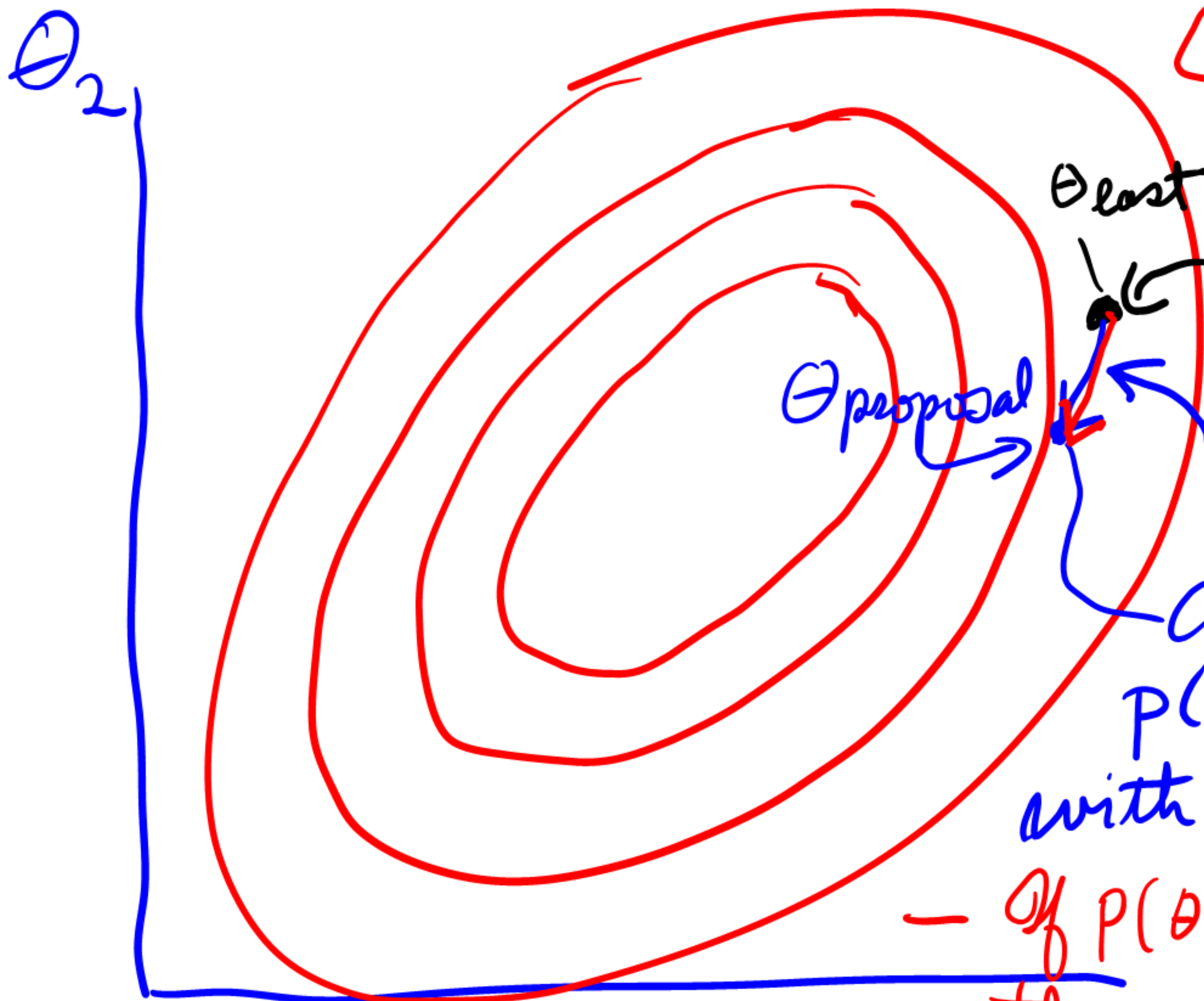


Start somewhere

Take a random step

Compare $p(\theta_{proposal}, X)$ with $p(\theta_{last}, X)$

- If $p(\theta_{prop}, X) \geq p(\theta_{last}, X)$ then $\theta_{new} = \theta_{proposal}$
- Otherwise, i.e. if you have moved down in probability, let $\theta_{new} = \theta_{proposal}$ with prob $\dfrac{p(\theta_{prop}, X)}{p(\theta_{last}, X)}$ else $\theta_{new} = \theta_{last}$



Contours

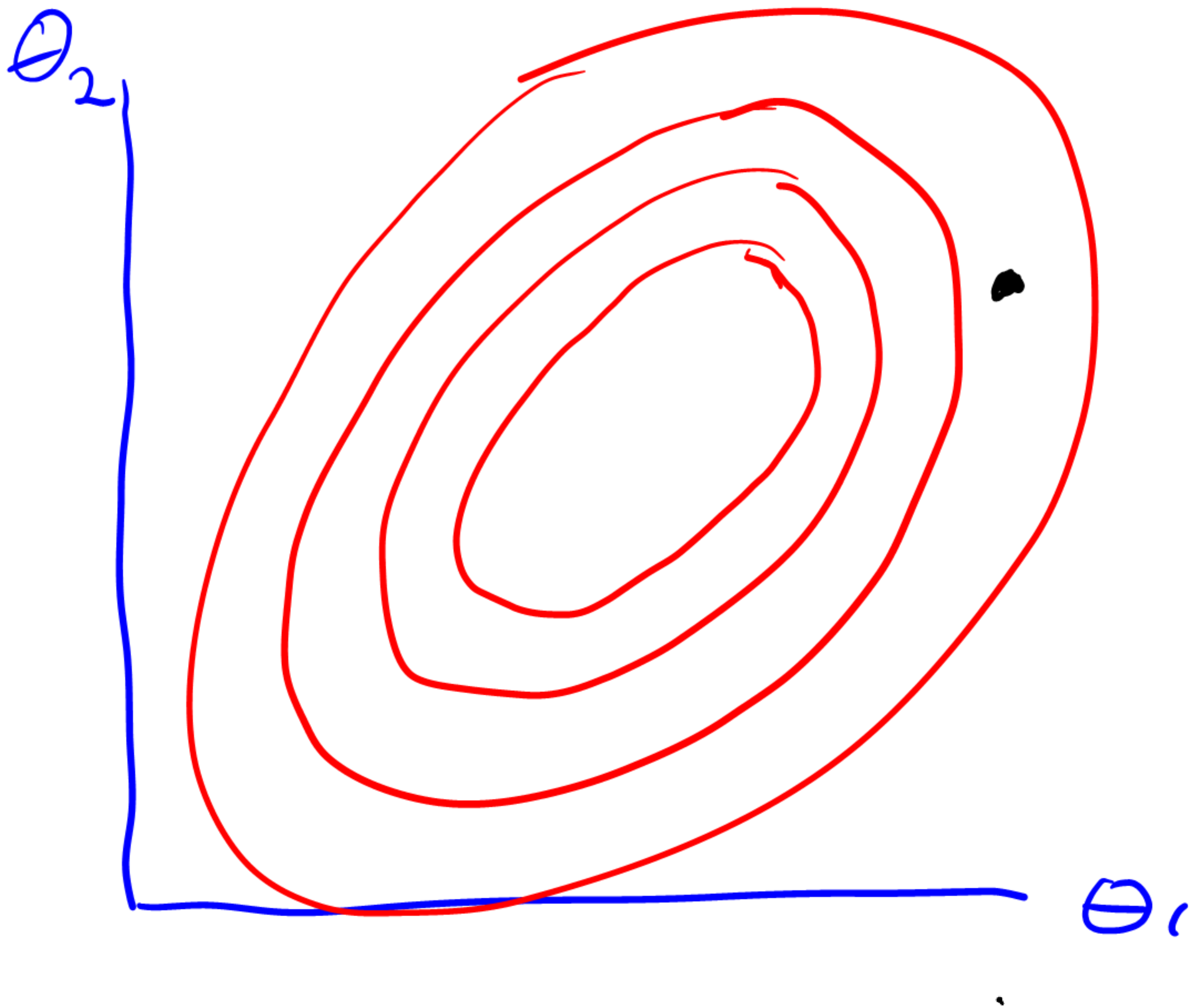
Start somewhere

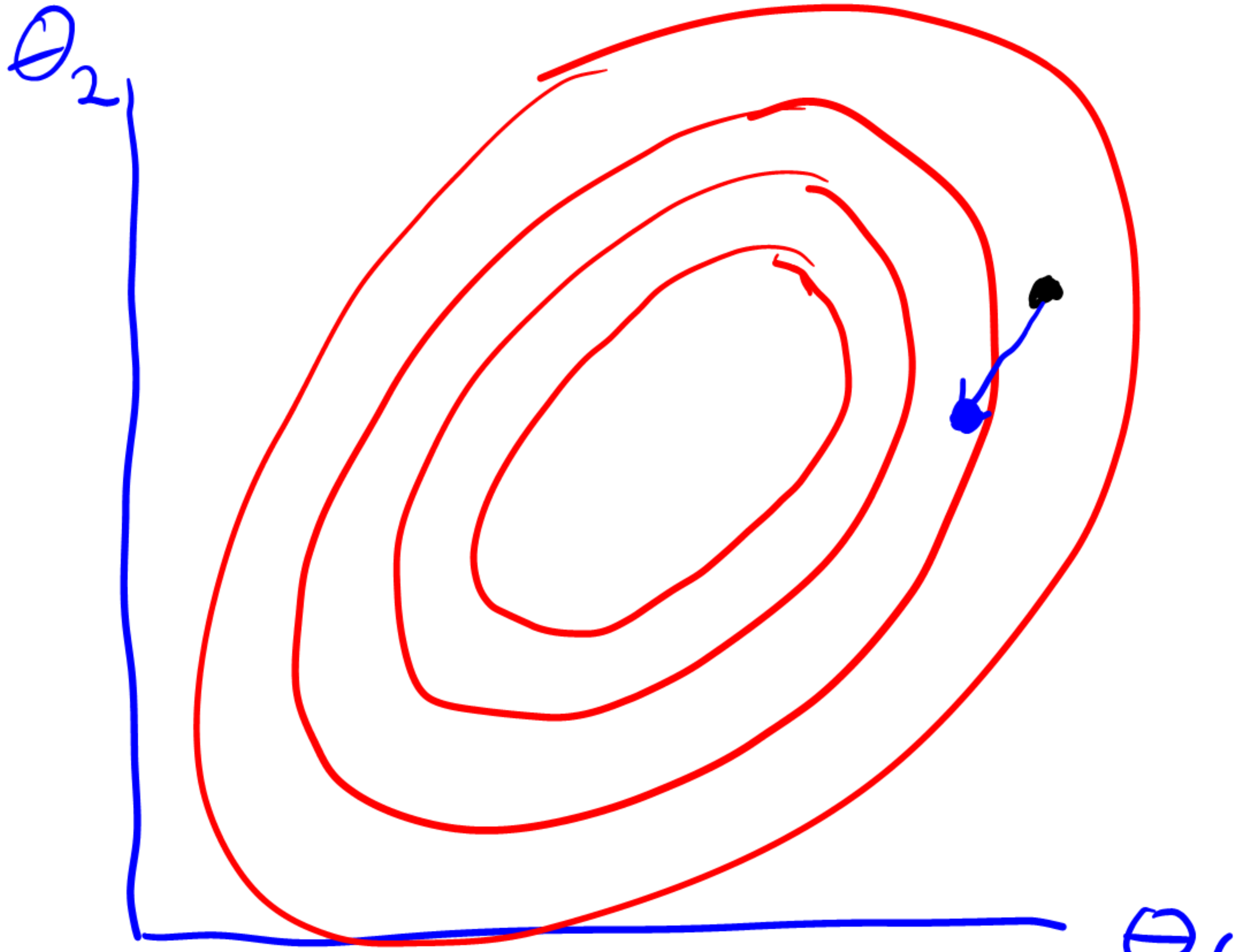
Take a random step

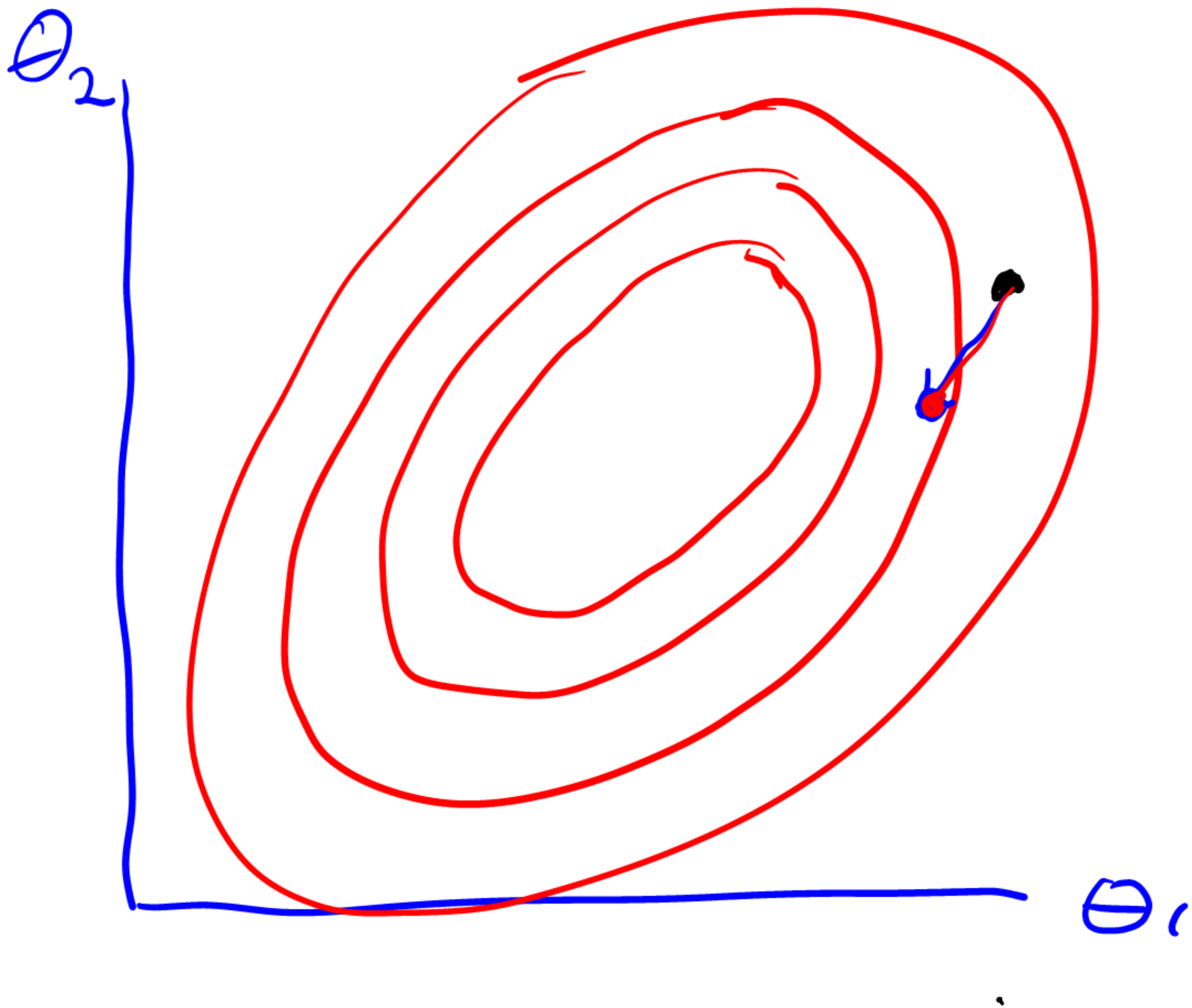
Compare $p(\theta_{proposal}, X)$ with $p(\theta_{last}, X)$

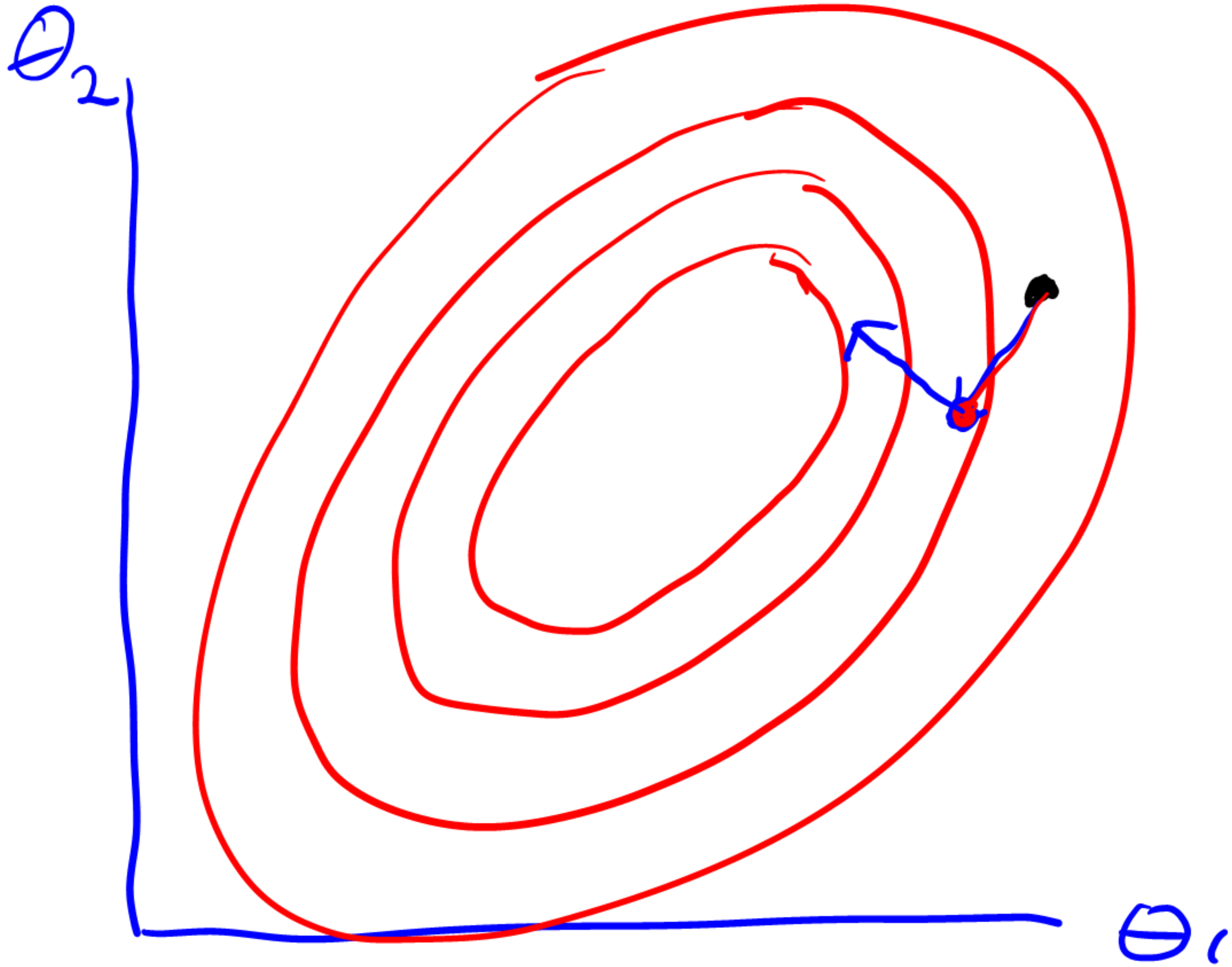
– If $p(\theta_{prop}, X) \geq p(\theta_{last}, X)$ then $\theta_{new} = \theta_{proposal}$

– Otherwise, i.e. if you have moved down in probability, let $\theta_{new} = \theta_{proposal}$ with prob $\dfrac{p(\theta_{prop}, X)}{p(\theta_{last}, X)}$ else $\theta_{new} = \theta_{last}$

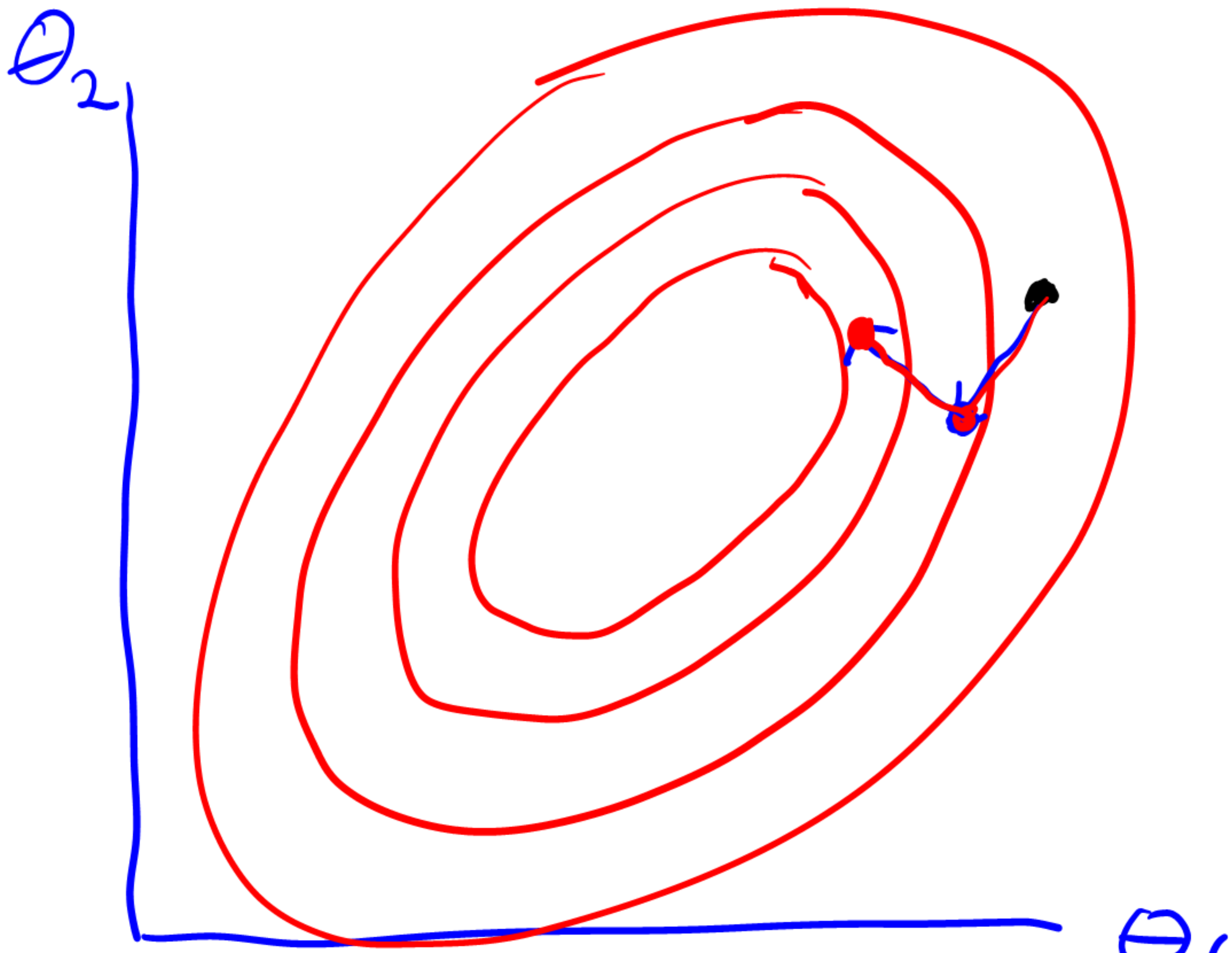

θ_2
θ_1


θ_2
θ_1

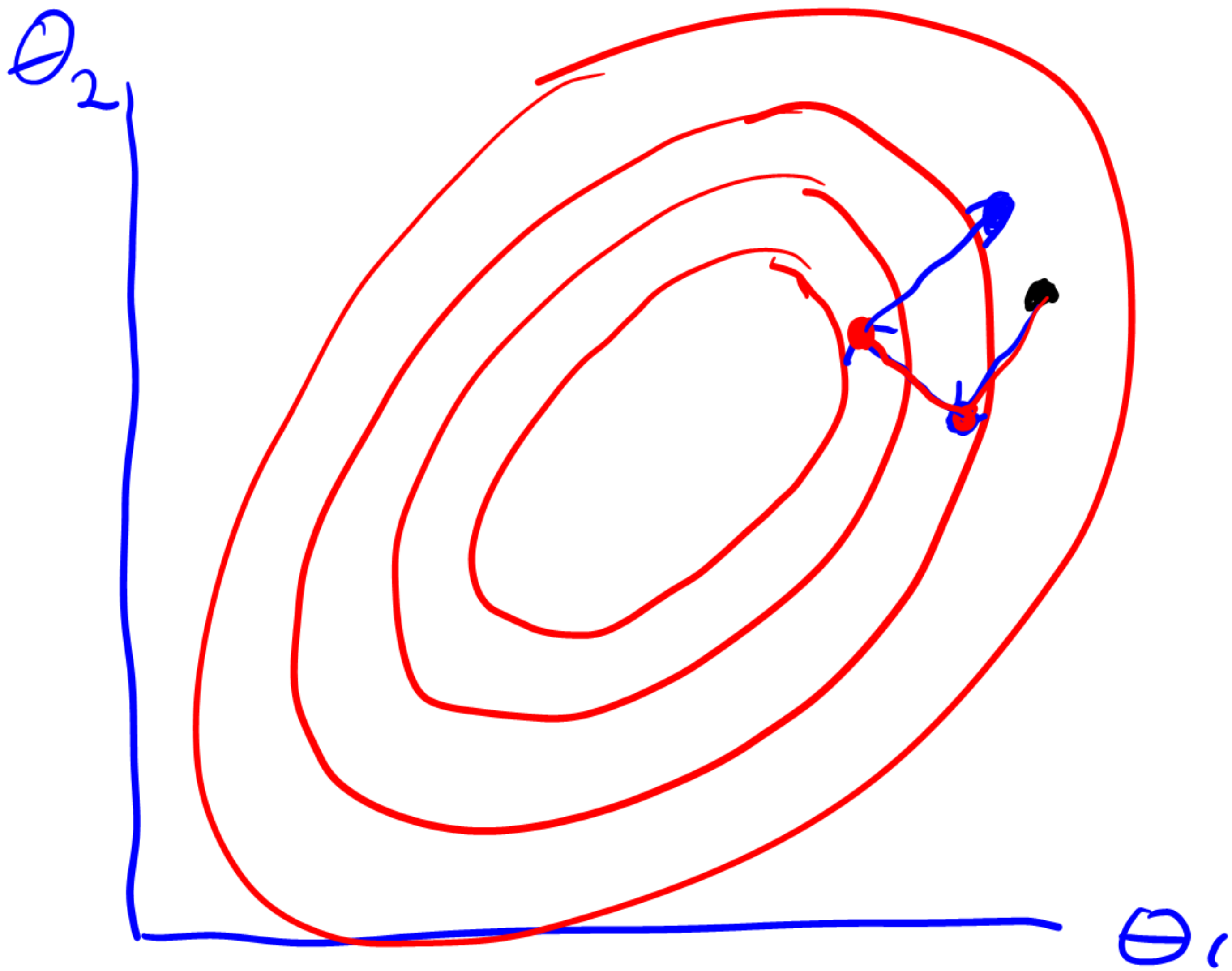

θ_2
θ_1



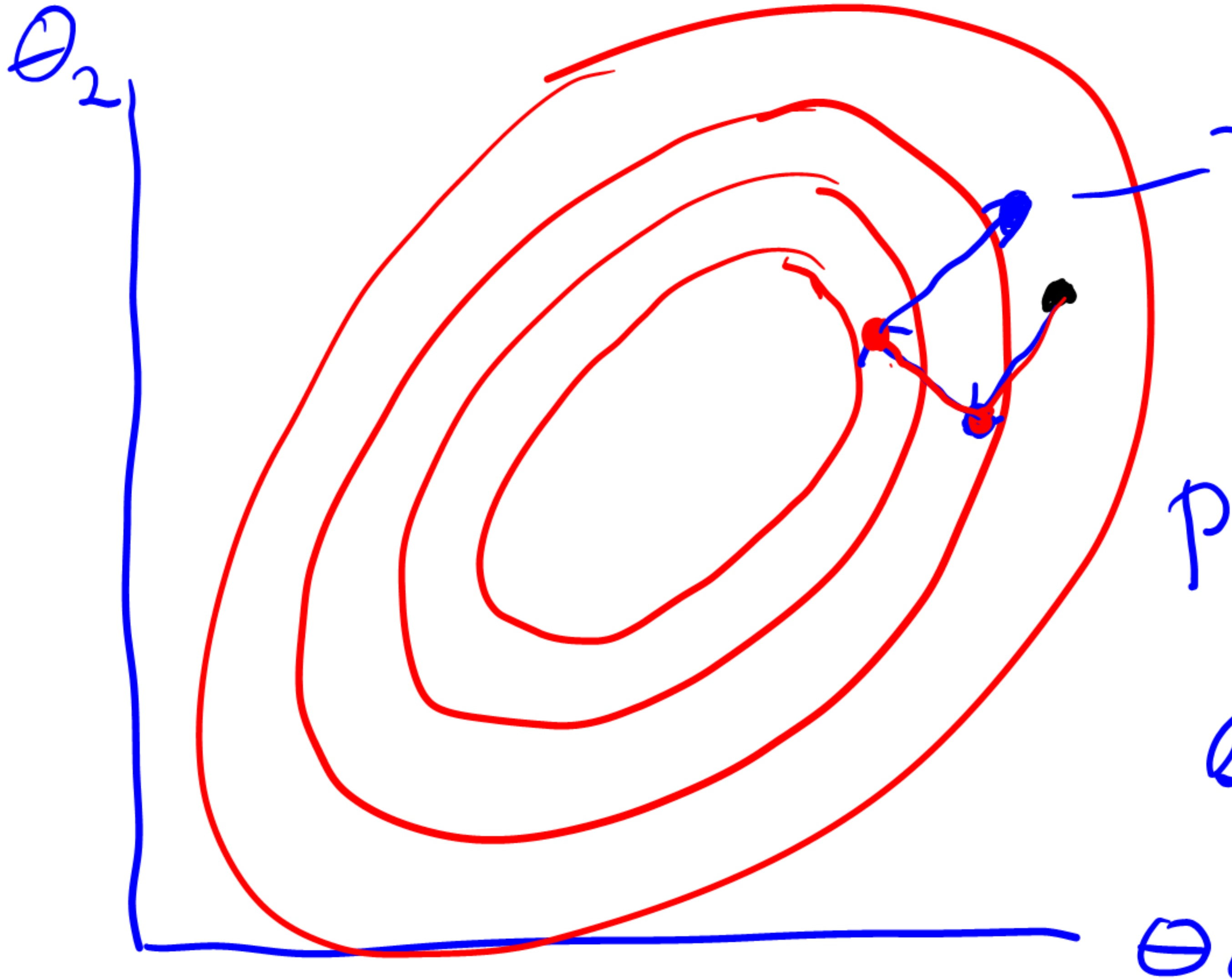
θ2
θ1



θ_2
θ_1



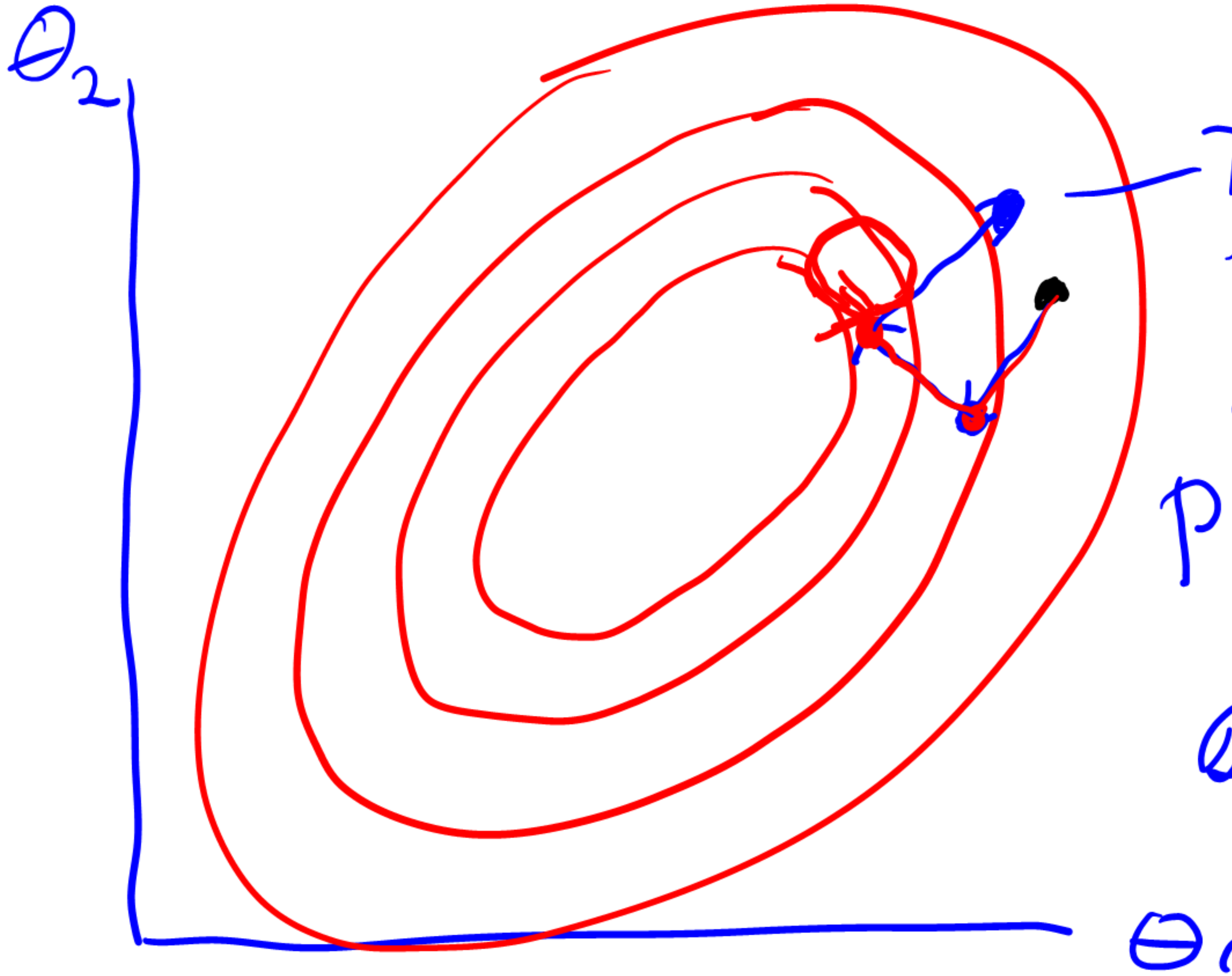
θ2
θ1



toss a biased coin with

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

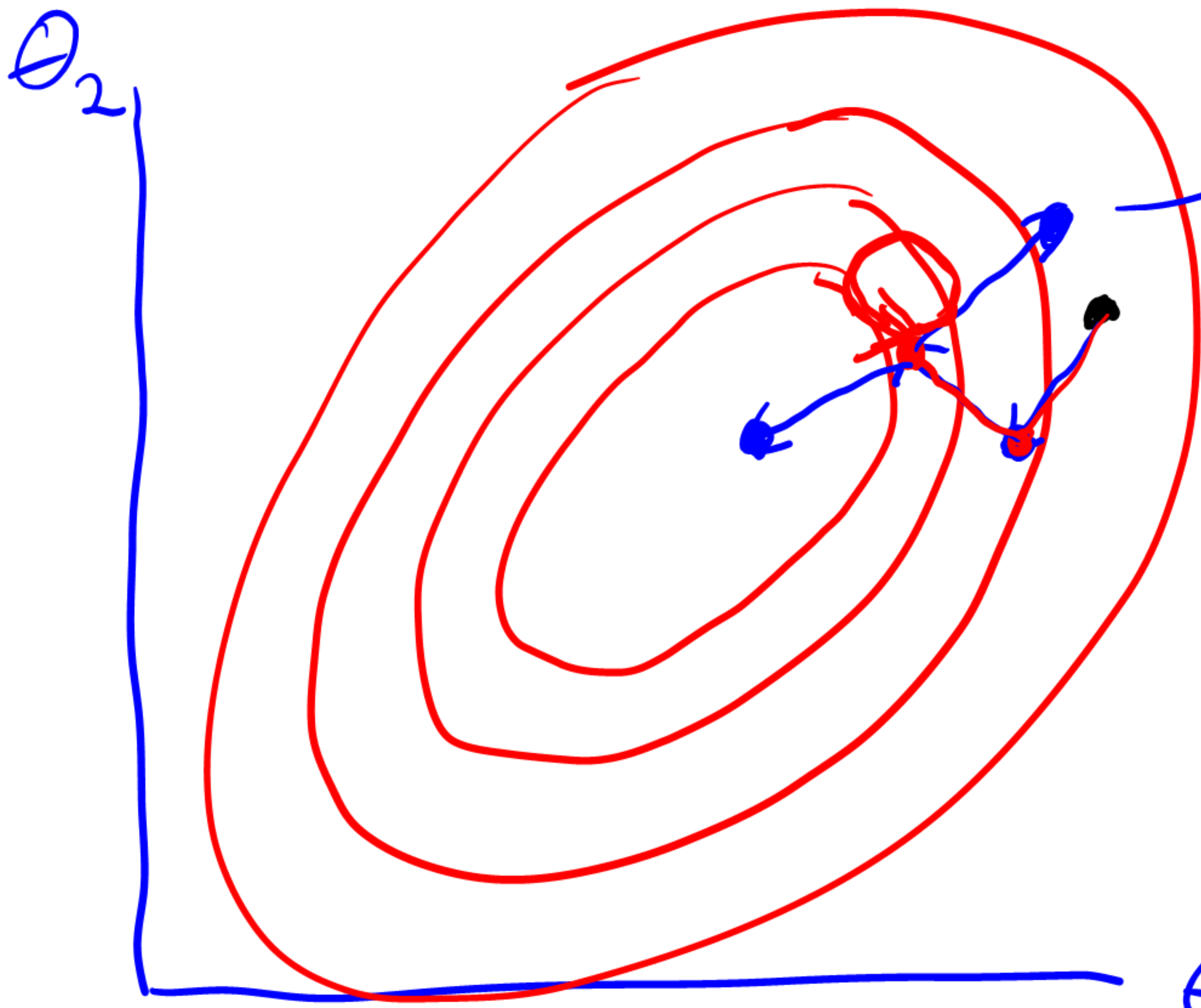
If Heads move
if Tails stay



toss a biased coin with

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

If Heads move
if Tails stay

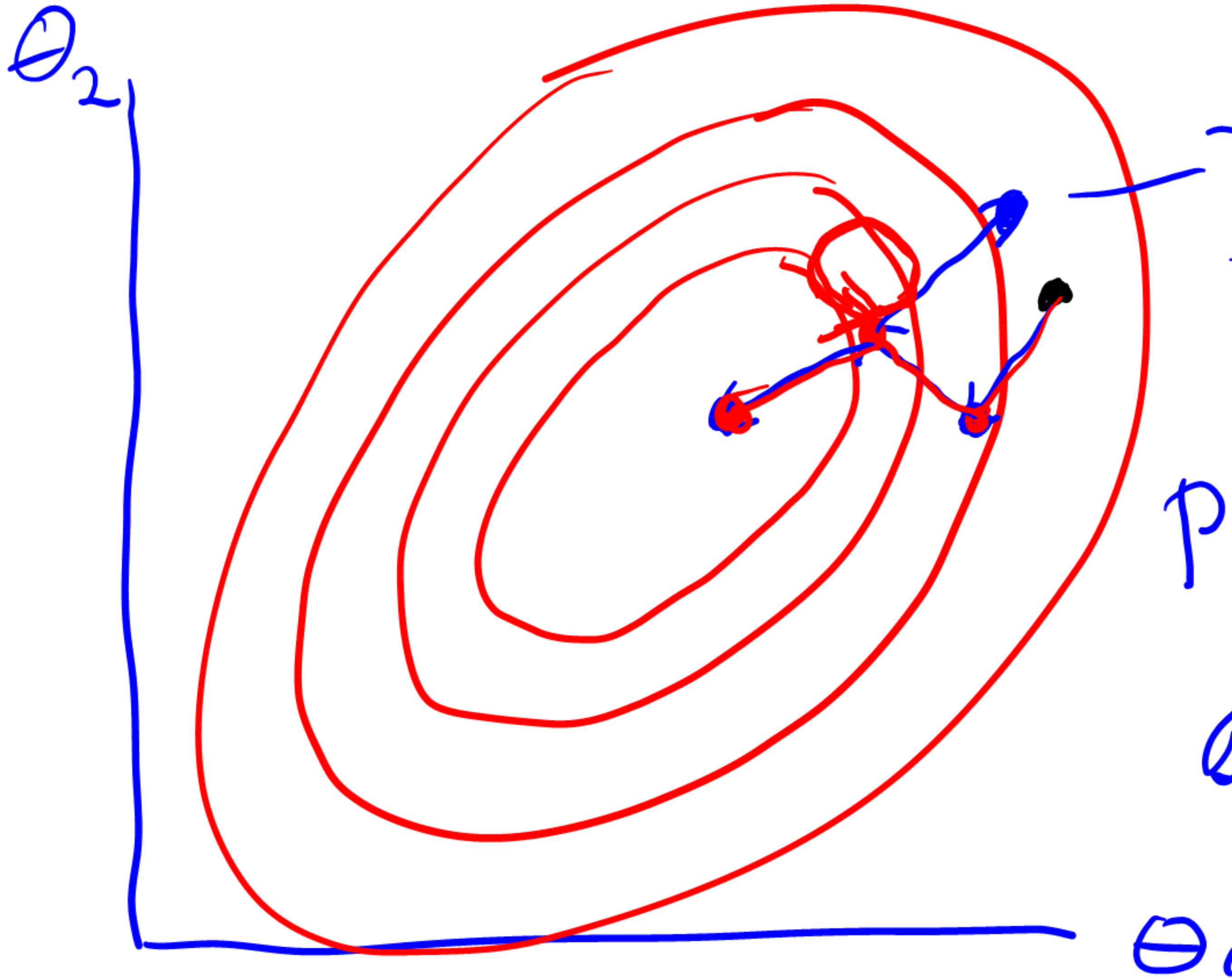


toss a biased coin with

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

If Heads move

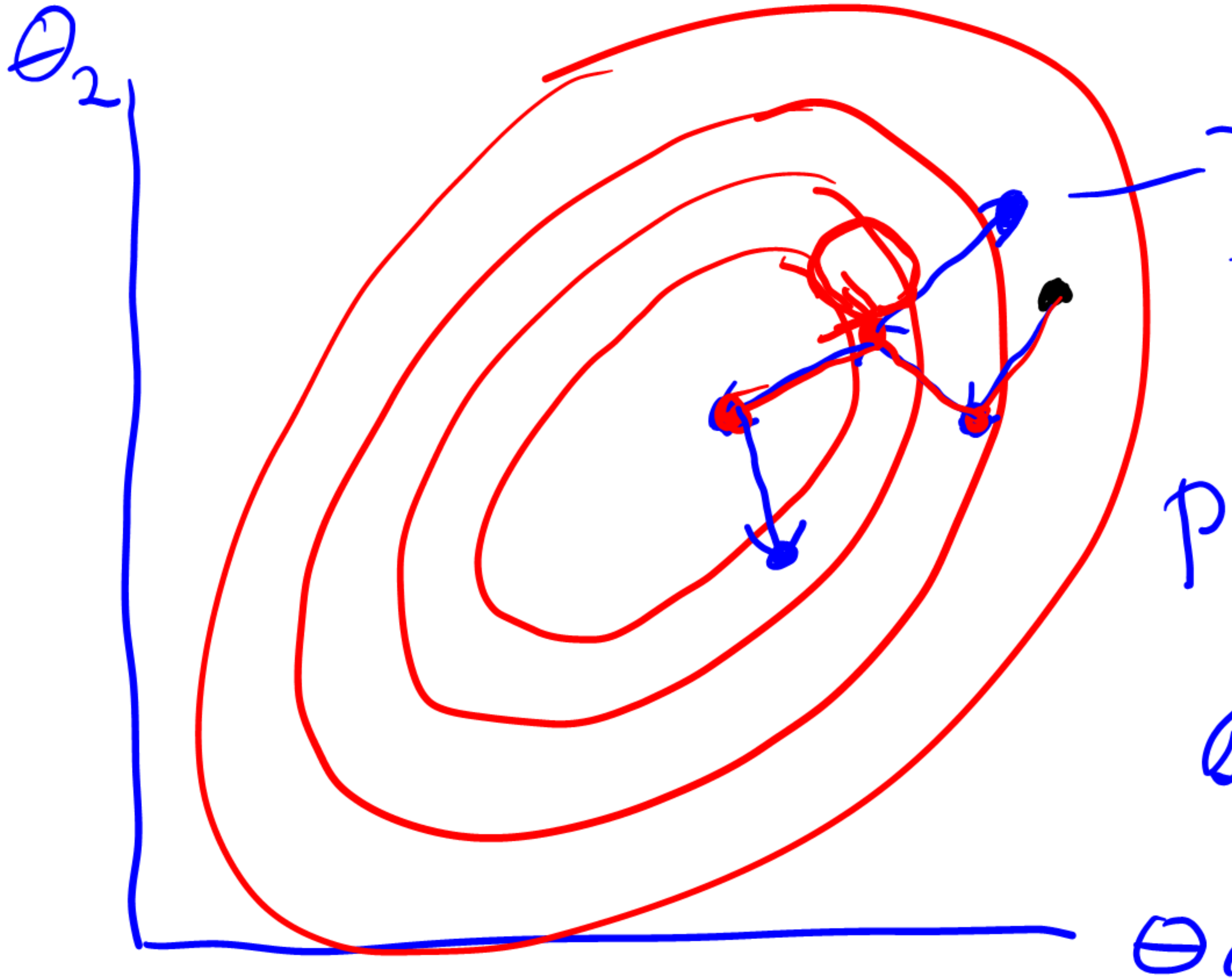
if Tails stay



toss a biased coin with

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

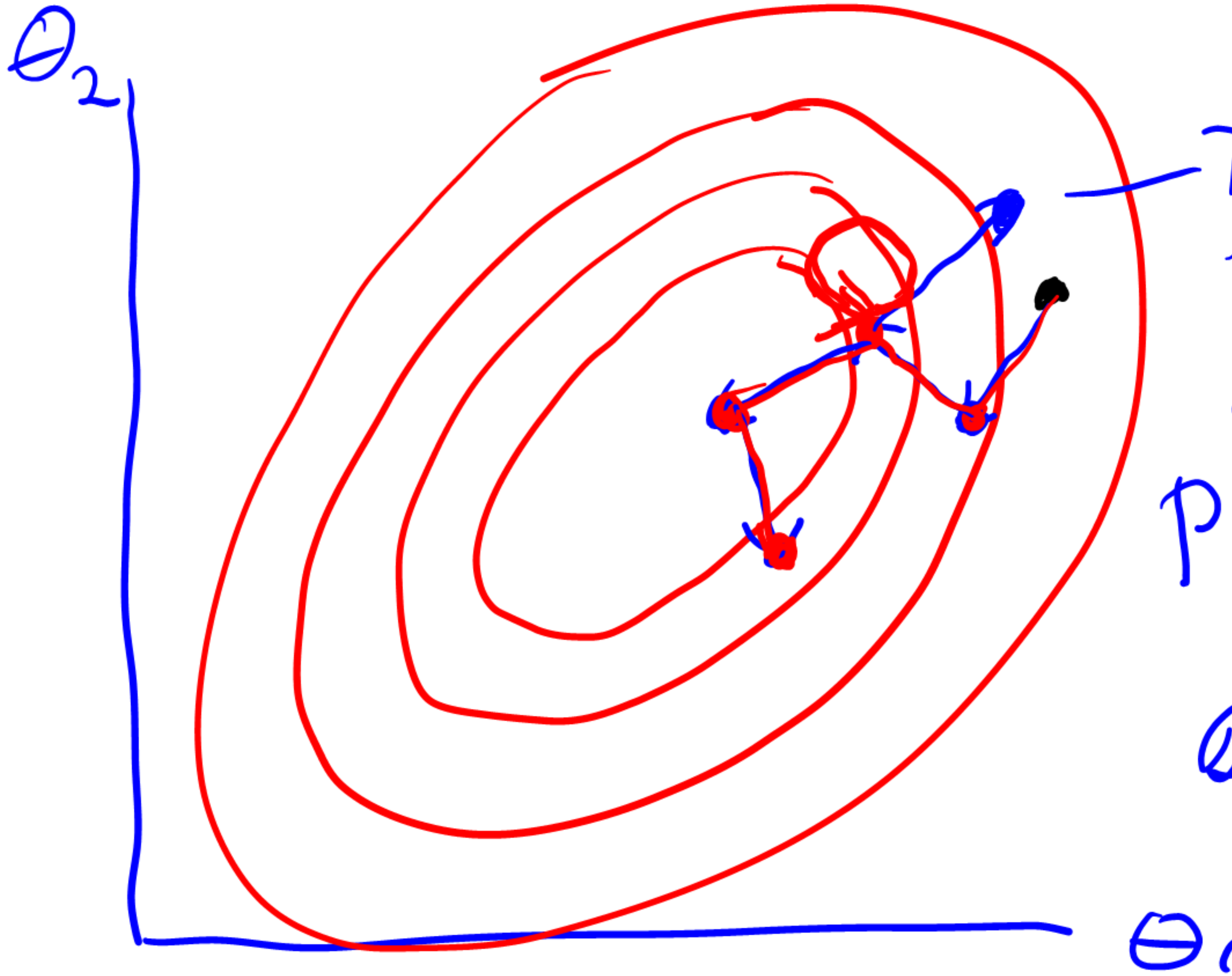
If Heads move
if Tails stay



toss a biased coin with

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

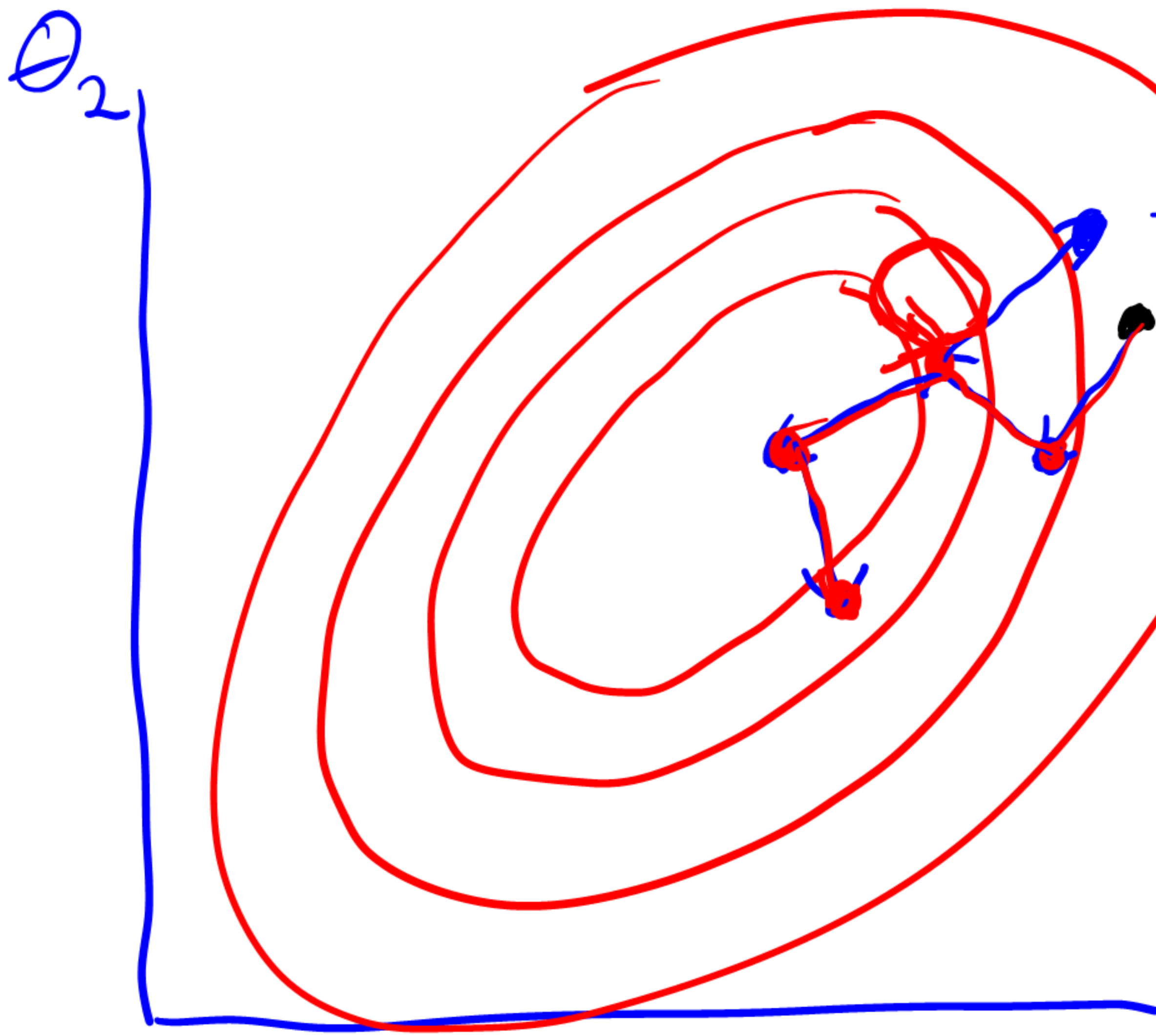
If Heads move
if Tails stay



toss a biased coin with

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

If Heads move
if Tails stay

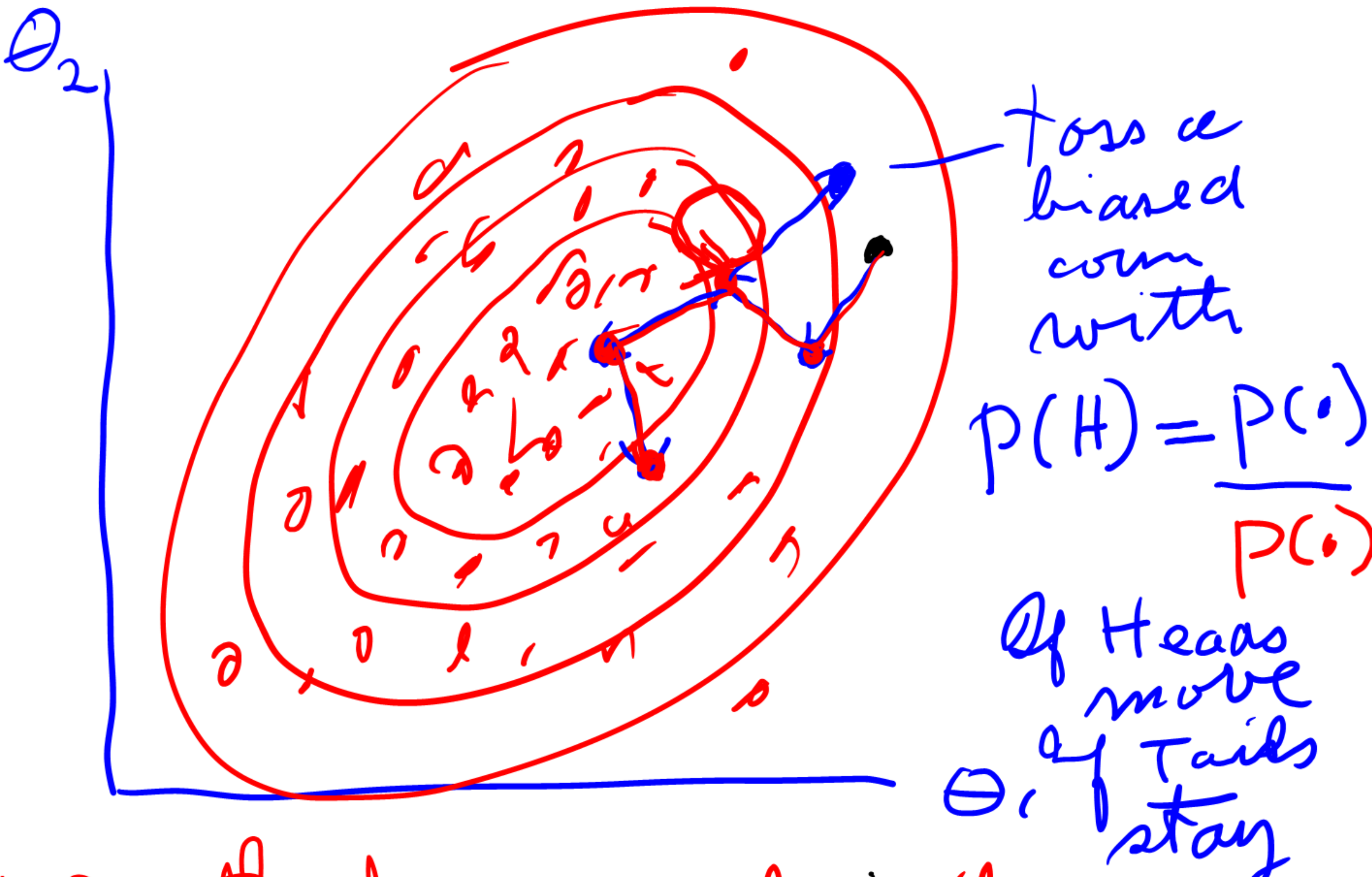


toss a biased coin with

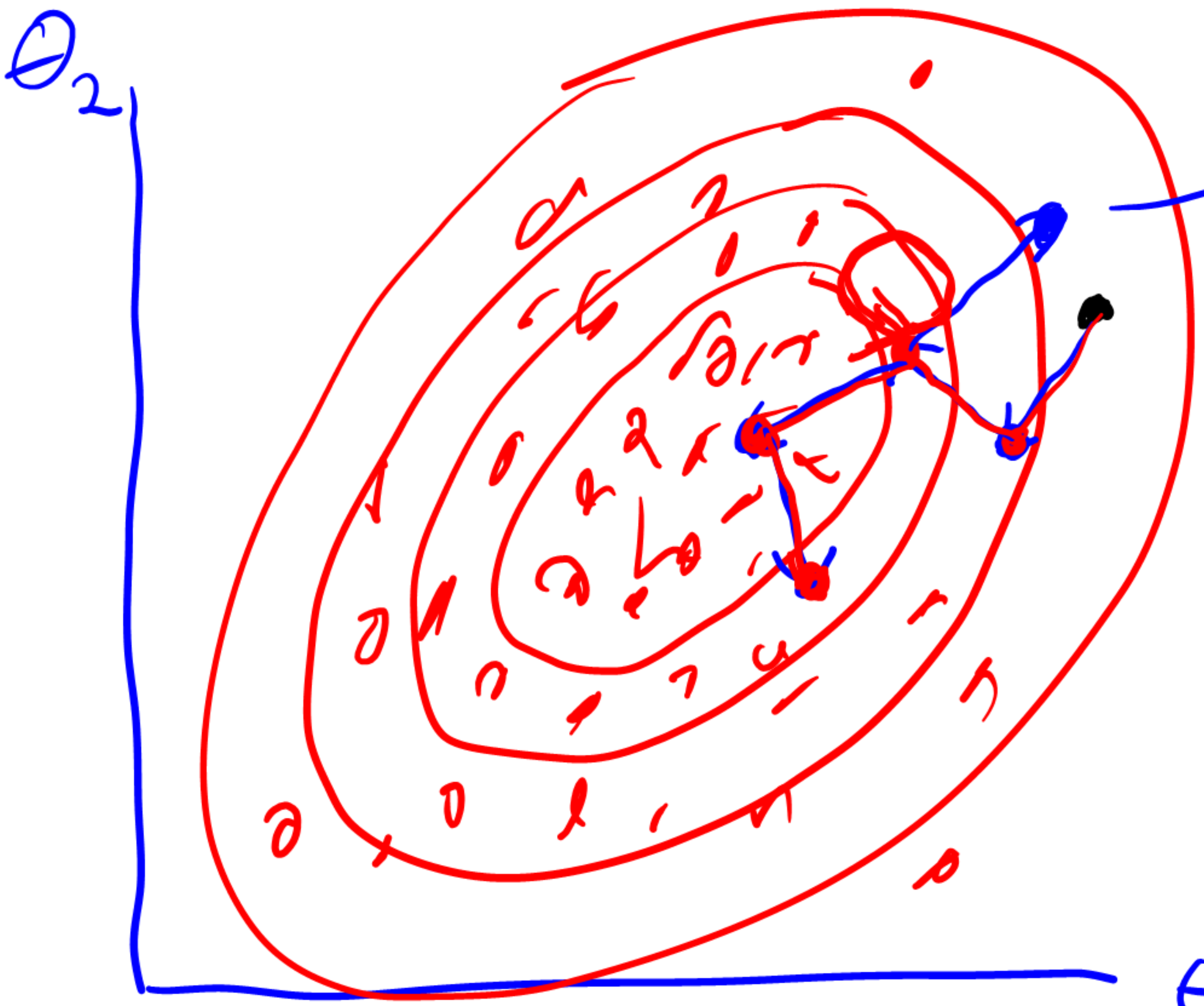
$$P(H) = \frac{P(\cdot)}{P(\cdot)}$$

If Heads move
if Tails stay

Keep doing this for a very long time



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

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

That's the XX-H algorithm

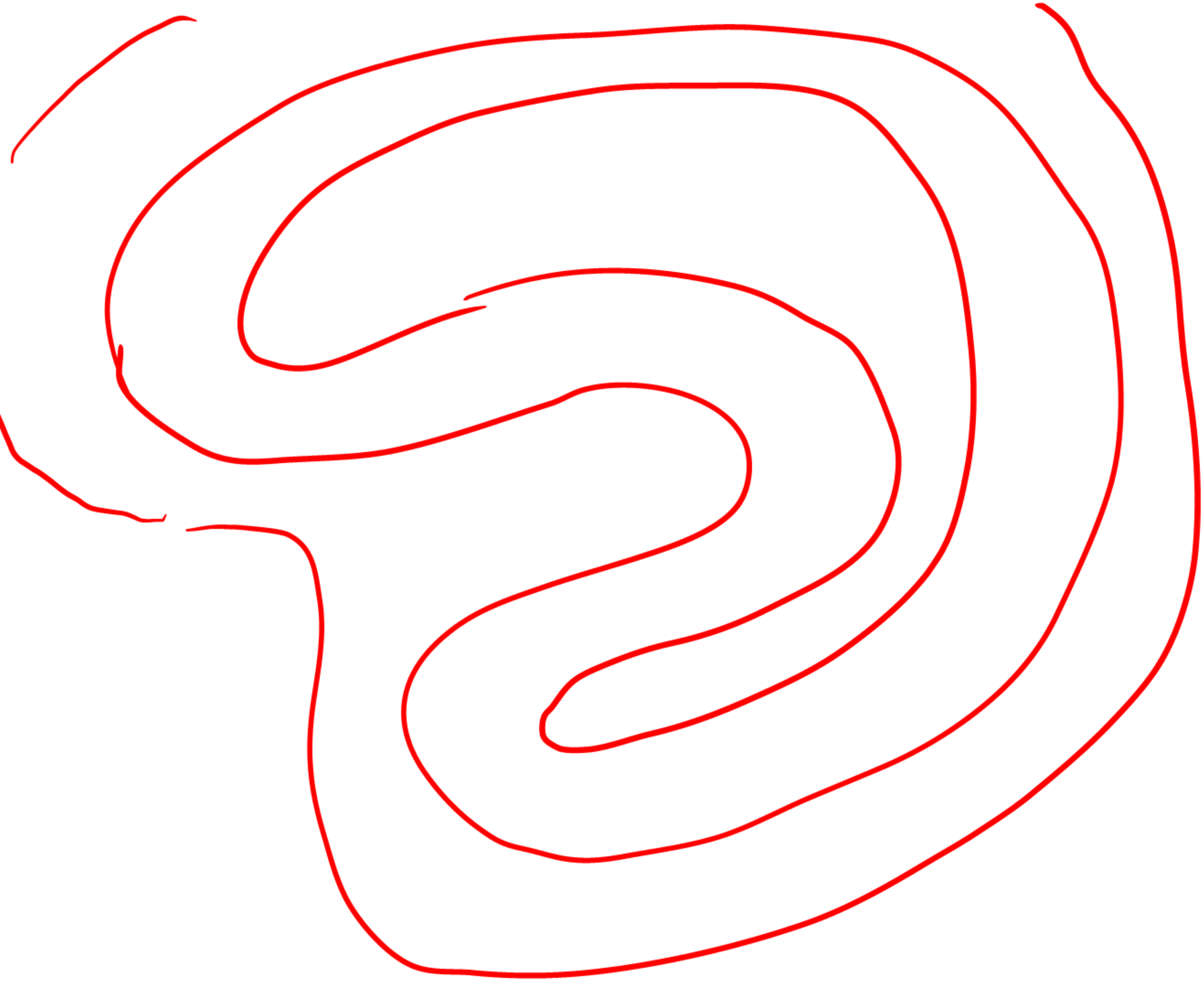
That's the M-H algorithm

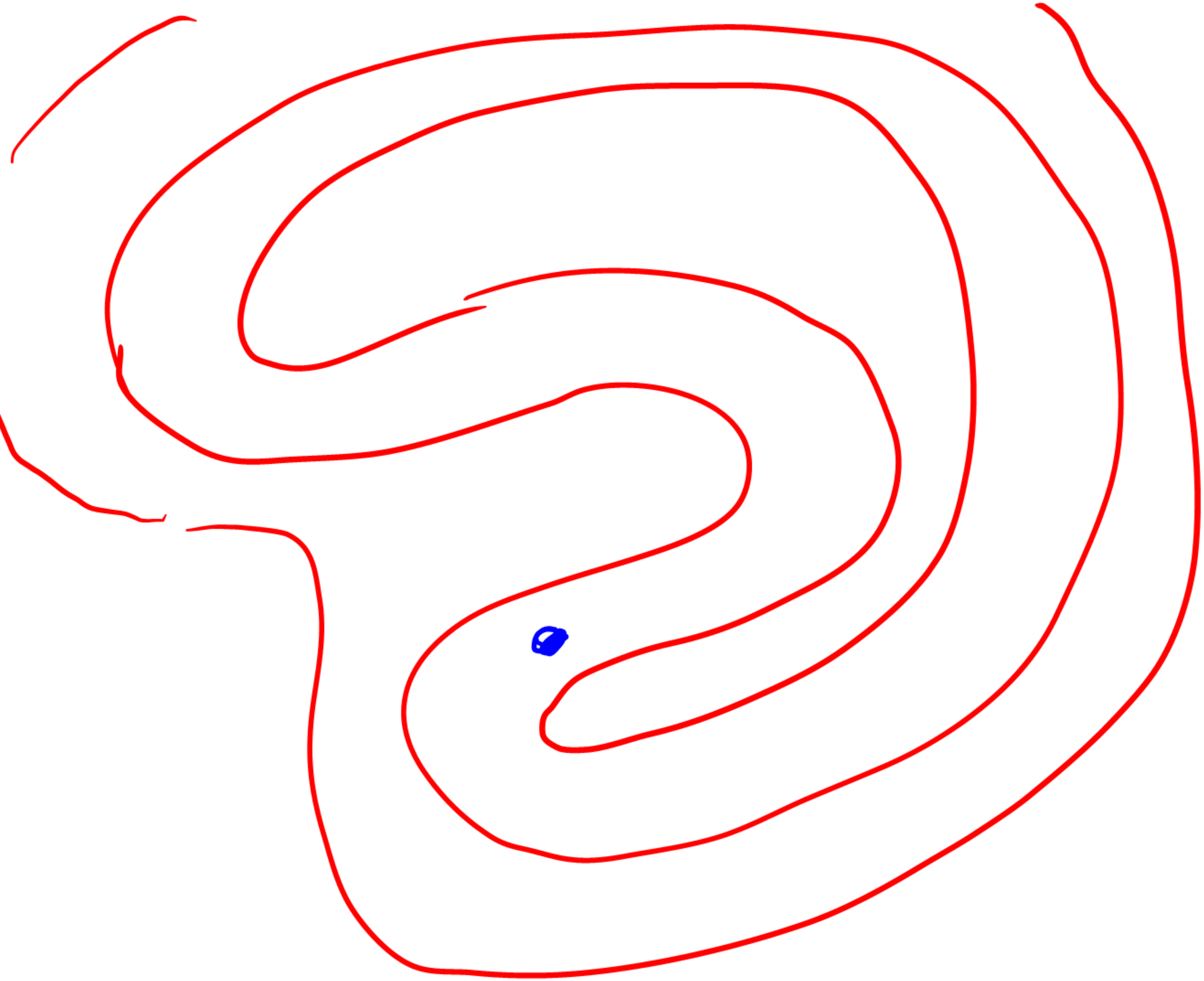
Can be very slow in high dimensions with non-elliptical contours, e.g.

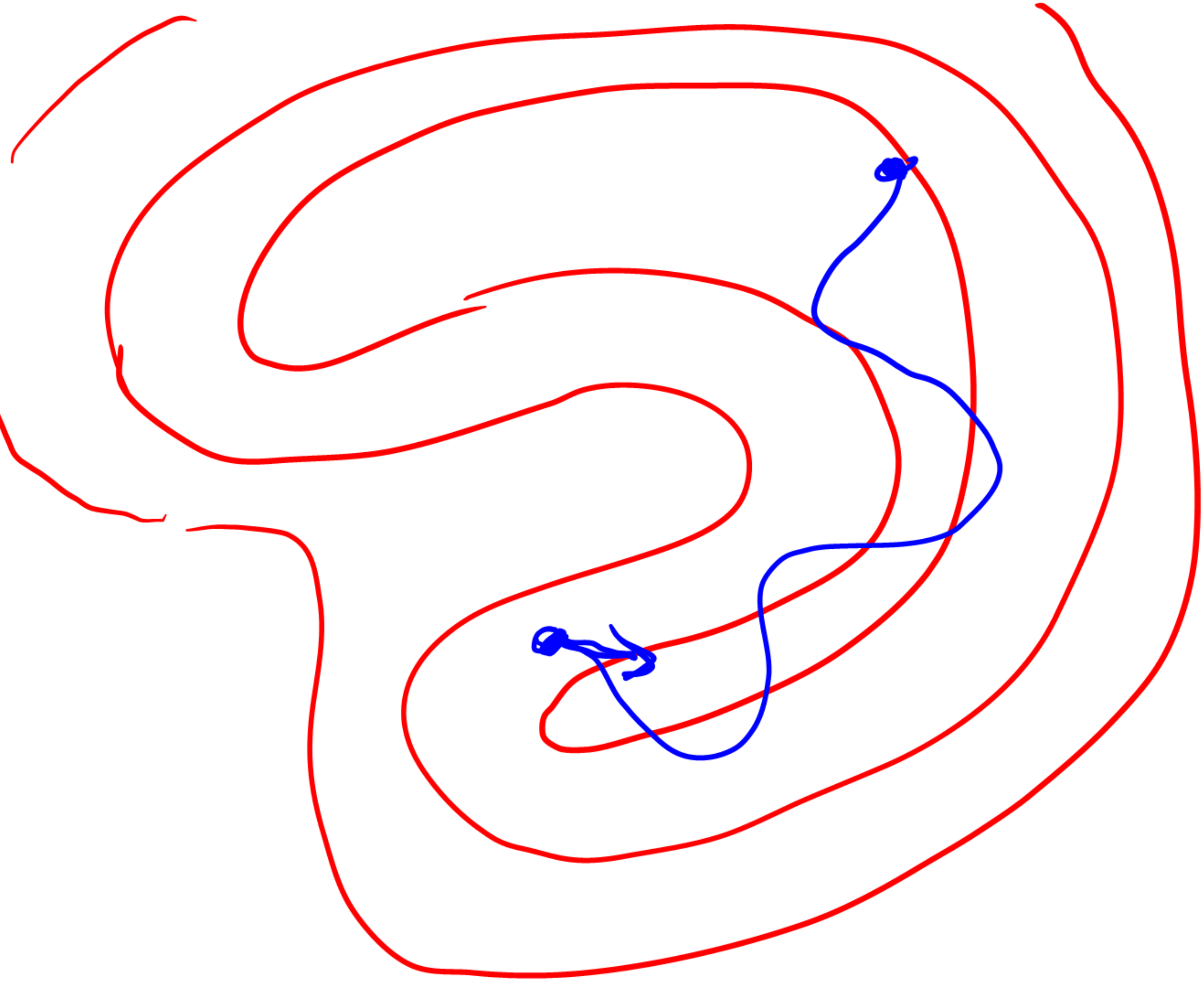
Can get stuck in corners for a long time!

Hamiltonian Monte Carlo

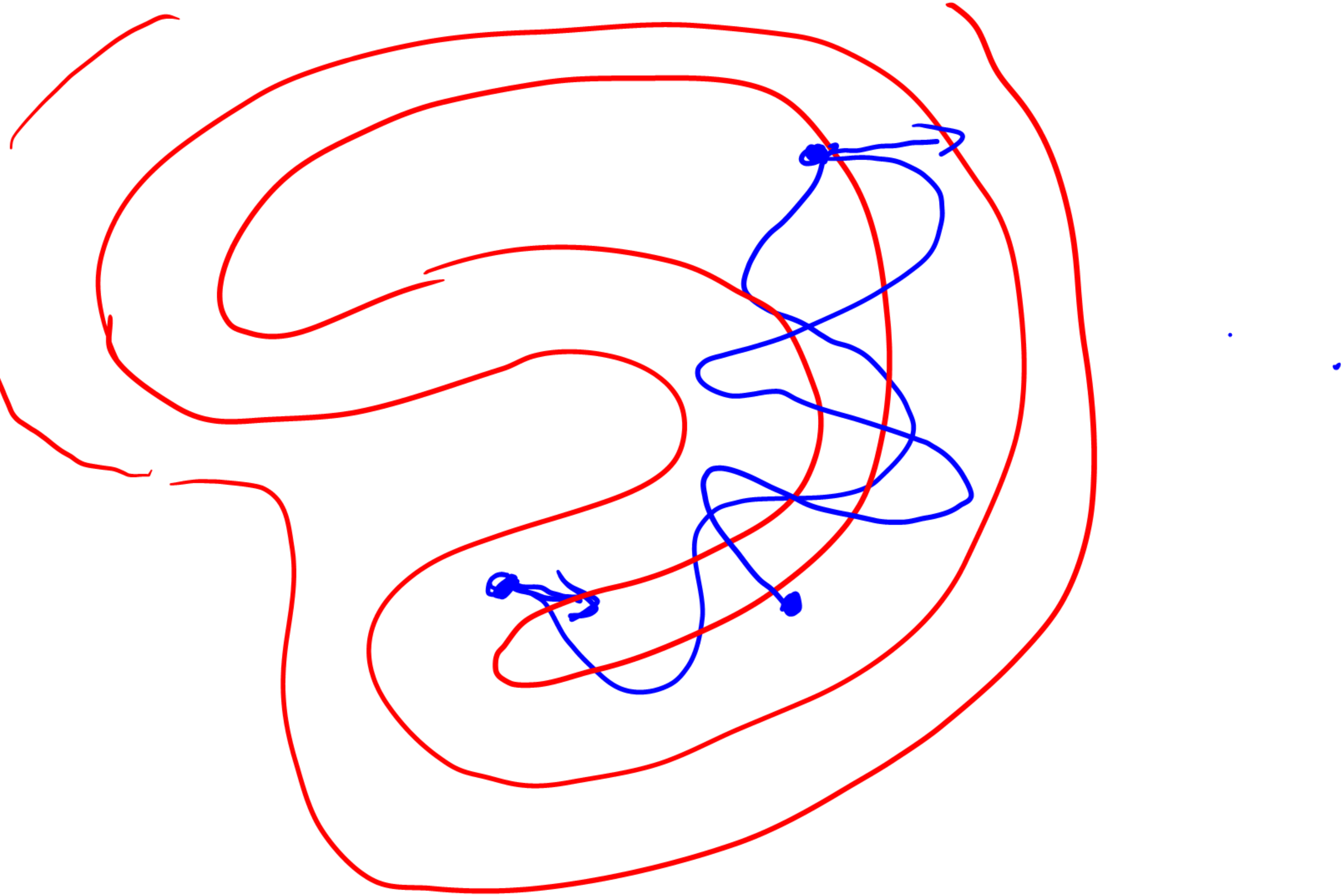
- Turn the mountain into a bowl by using $-\log p(\theta, x)$
- Instead of random steps, go for a ride on a frictionless skateboard with swivel wheels — starting with a random push.

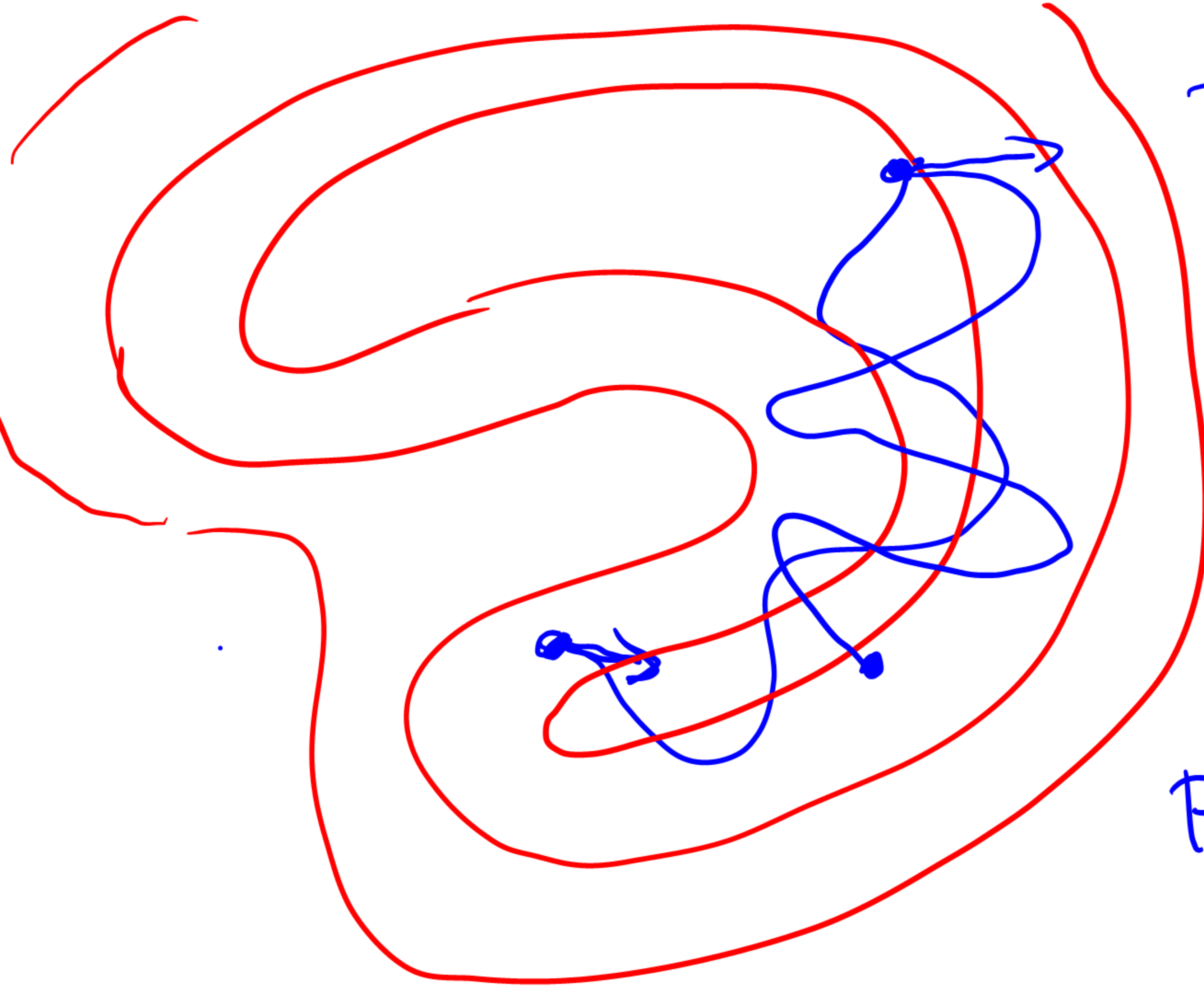




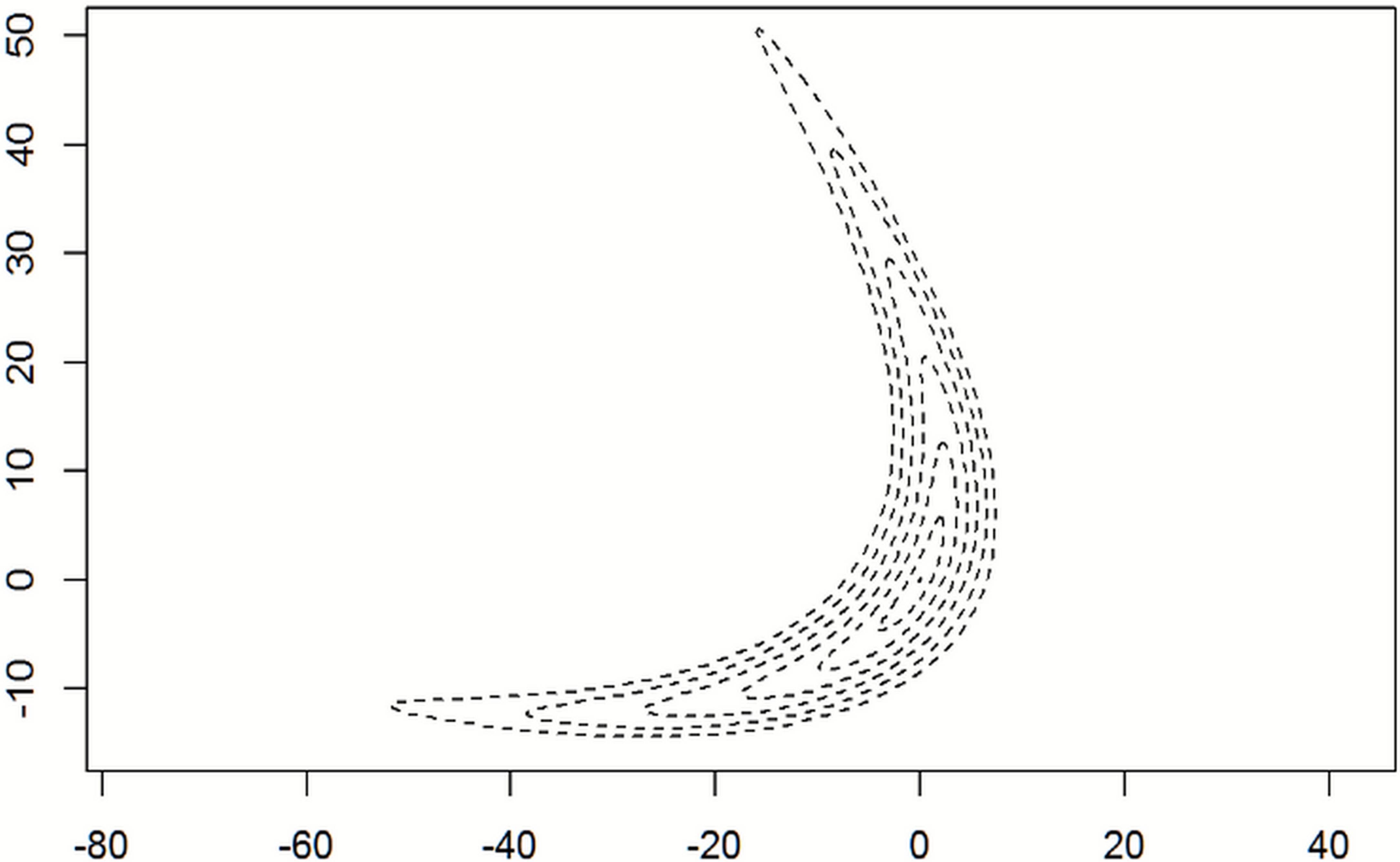


Stop after a set time.

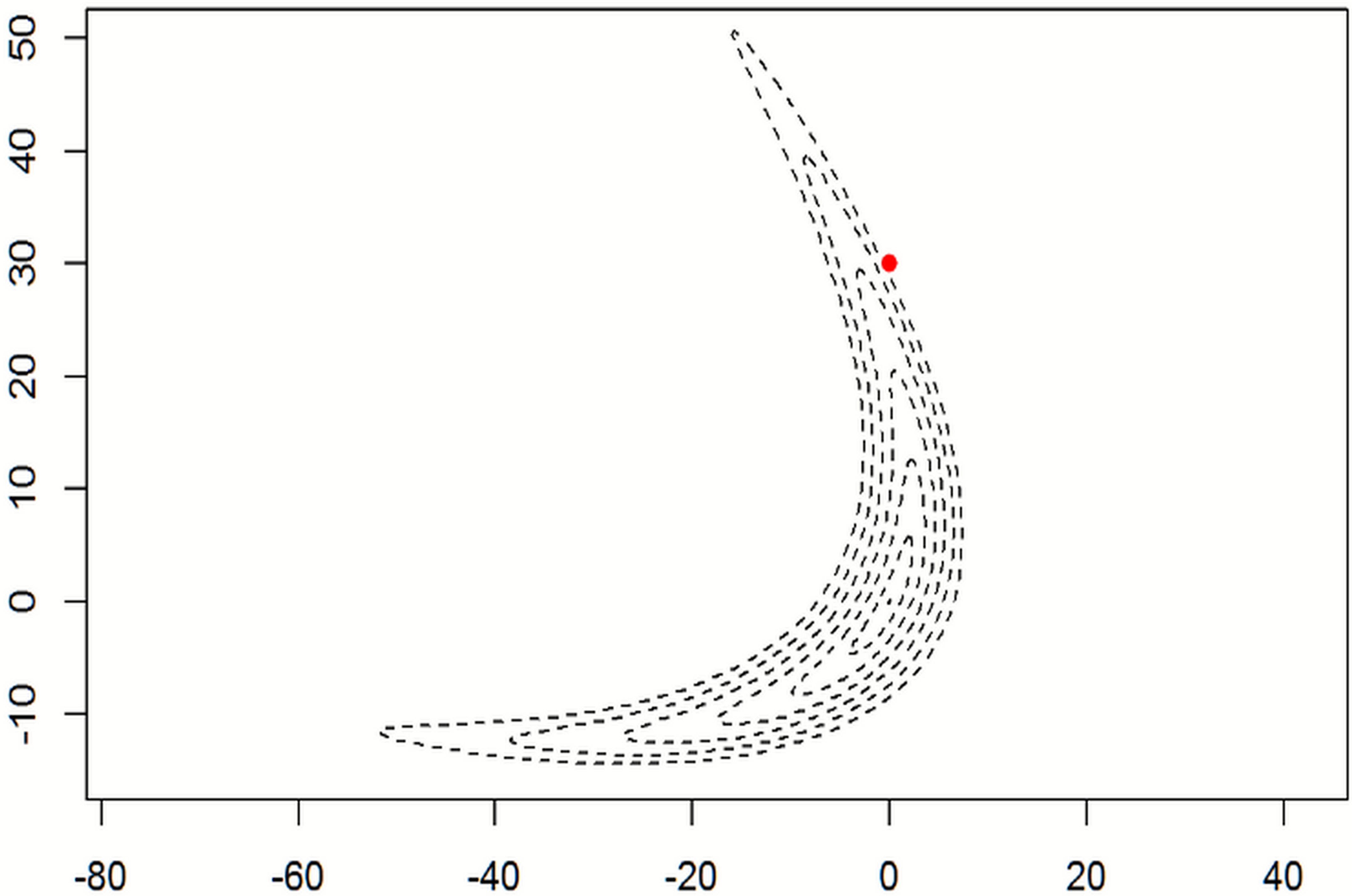




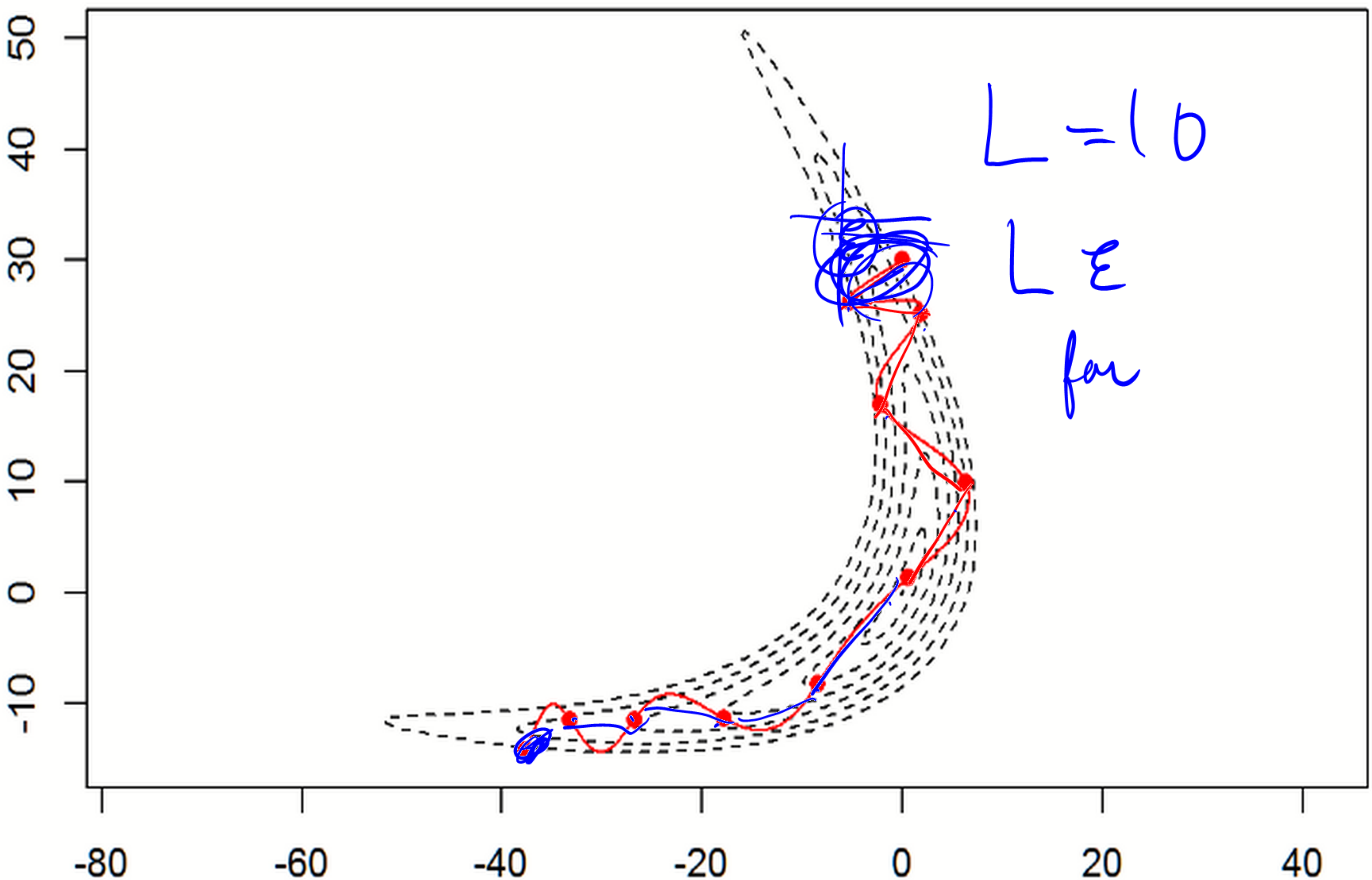
This can do a much better job of covering $p(\theta|X)$.



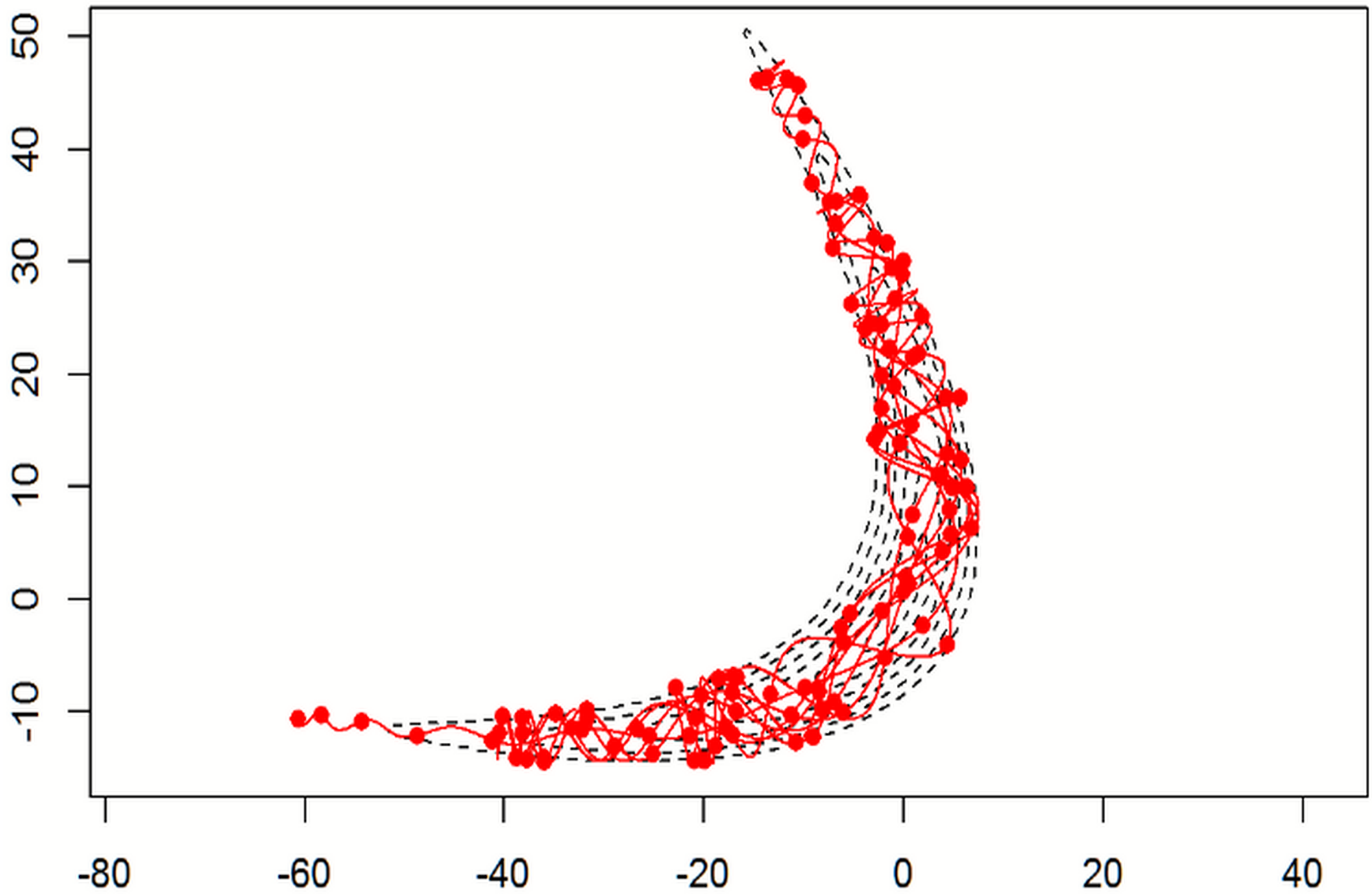
50
40
30
20
10
0
-10
-80
-60
-40
-20
0
20
40



50
40
30
20
10
0
-10
-80
-60
-40
-20
0
20
40



L=10
LE
far
-80
-60
-40
-20
0
20
40
50
40
30
20
10
0
-10


50
40
30
20
10
0
-10
-80
-60
-40
-20
0
20
40

L : # of leaps (leapfrog) steps

ε : size of each step

$L\varepsilon$: "distance" travelled in θ-space

If 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)

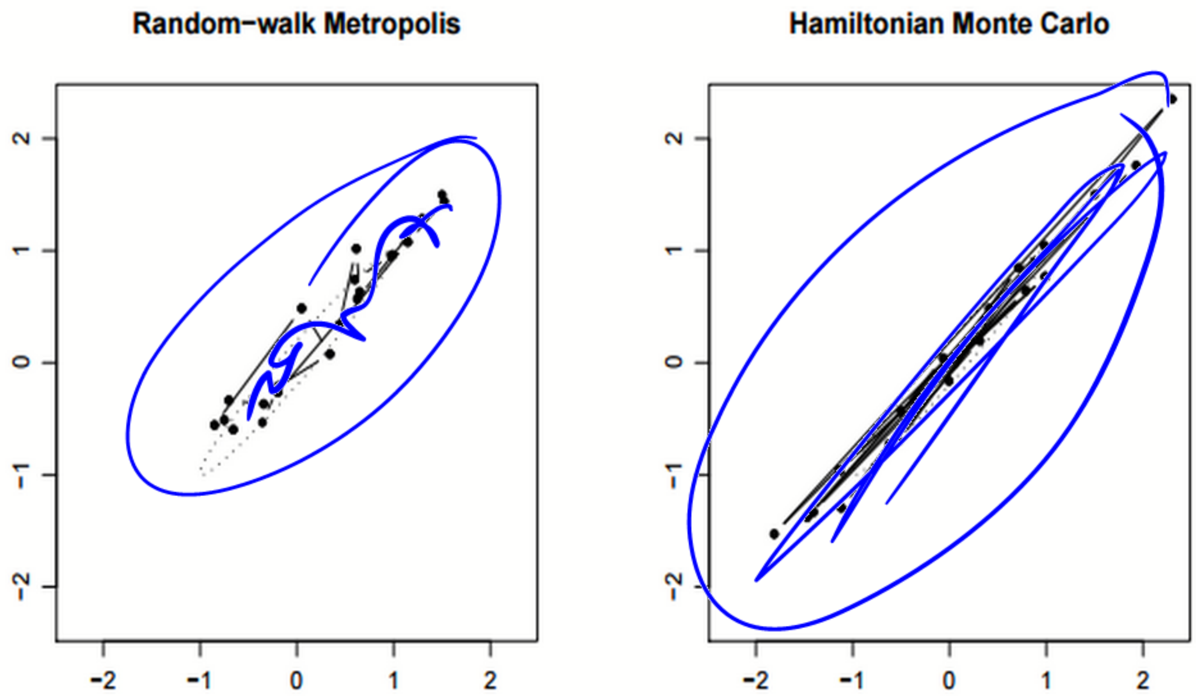


Figure 4: Twenty iterations of the random-walk Metropolis method (with 20 updates per iteration) and of the Hamiltonian Monte Carlo method (with 20 leapfrog steps per trajectory) for a 2D Gaussian distribution with marginal standard deviations of one and correlation 0.98. Only the two position coordinates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

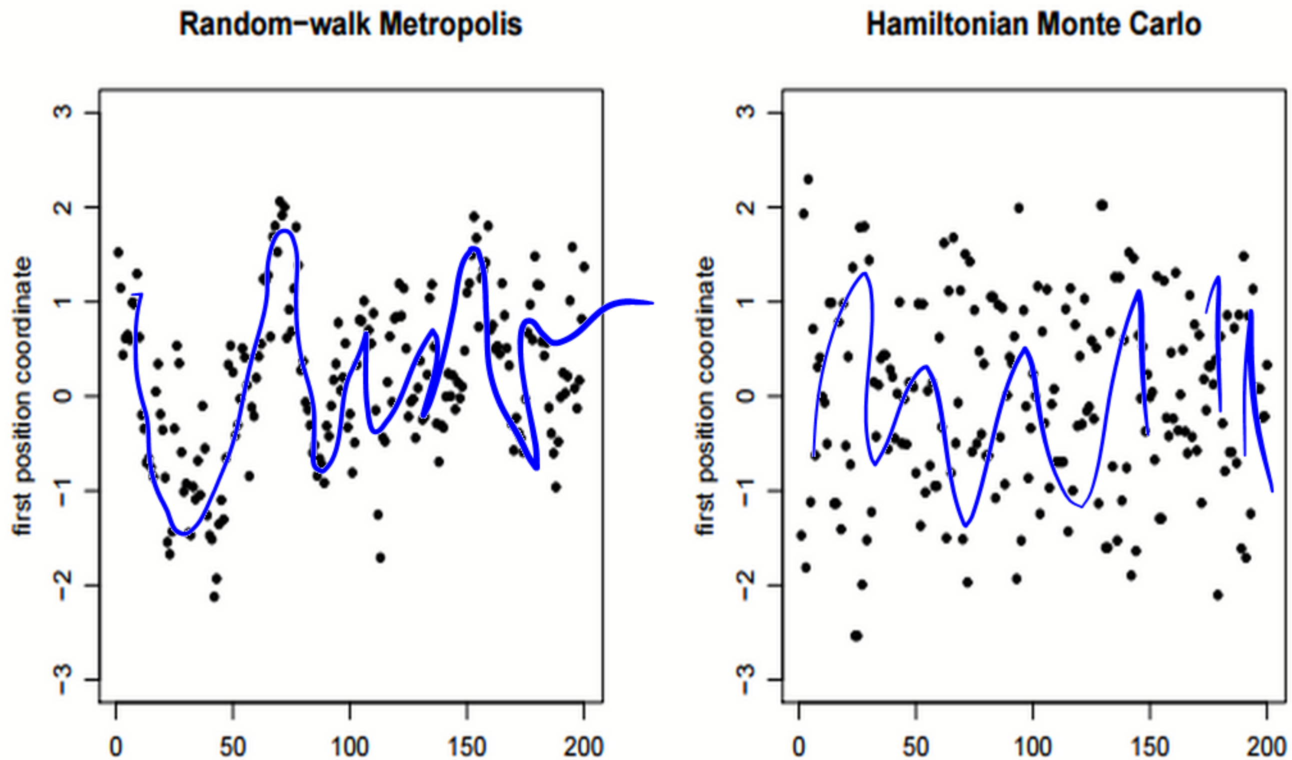


Figure 5: Two hundred iterations, starting with the twenty iterations shown above, with only the first position coordinate plotted.

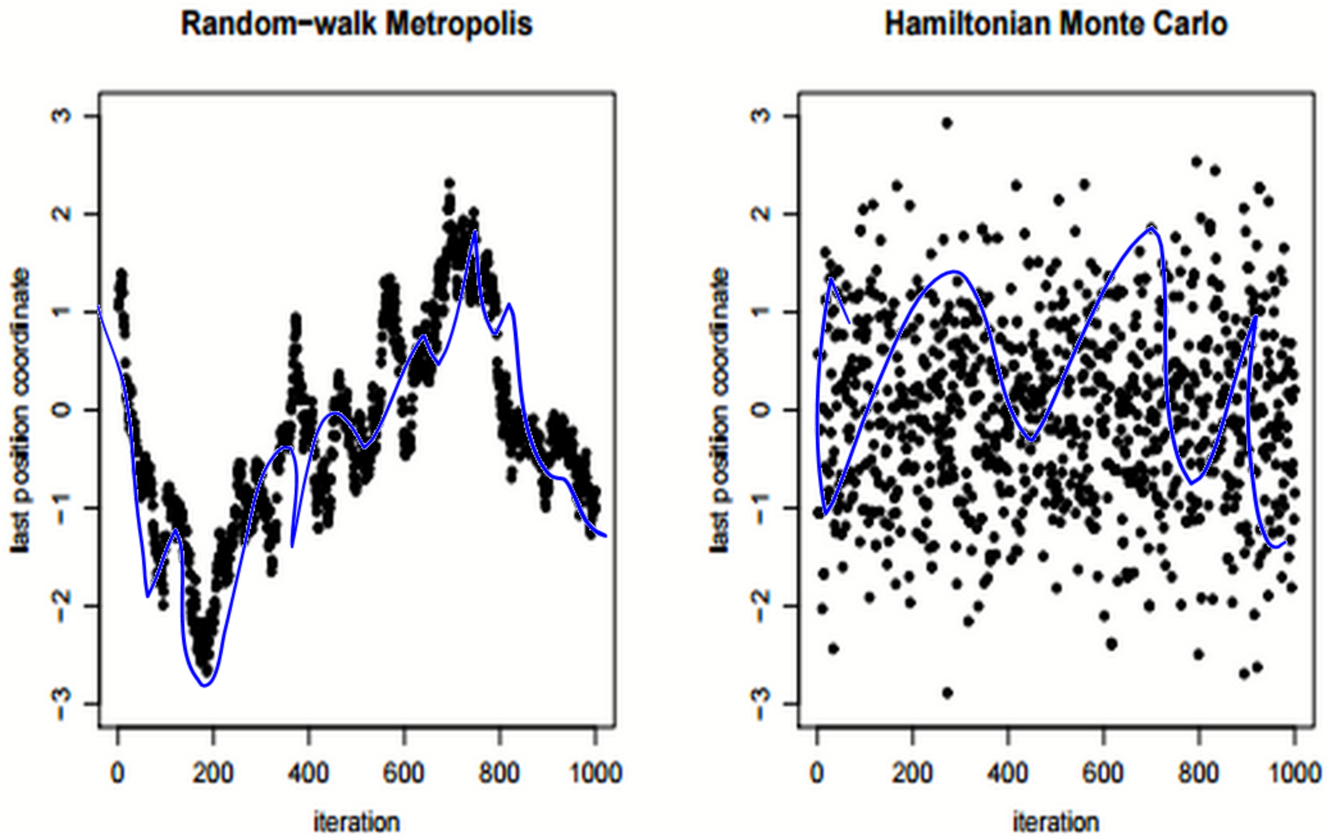


Figure 6: Values for the variable with largest standard deviation for the 100-dimensional example, from a random-walk Metropolis run and an HMC run with $L = 150$. To match computation time, 150 updates were counted as one iteration for random-walk Metropolis.