Statisztikai tanulás az idegrendszerben
Közelítő inferencia, mintavételezés

Orbán Gergő
golab.wigner.mta.hu

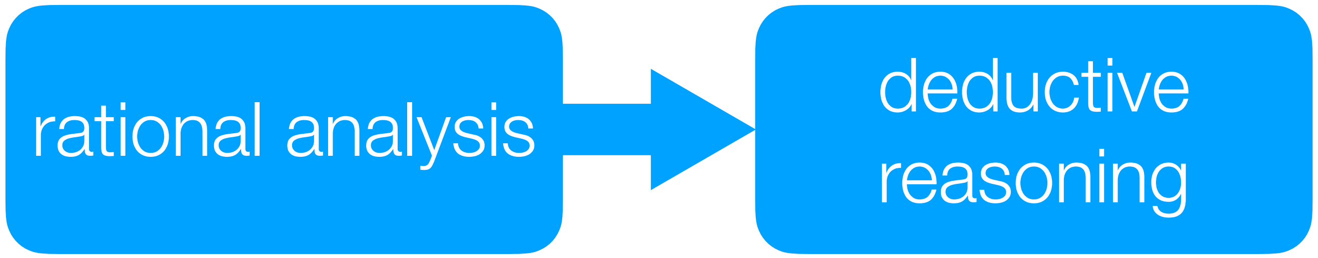
WIGNER RESEARCH CENTRE FOR PHYSICS

csnl

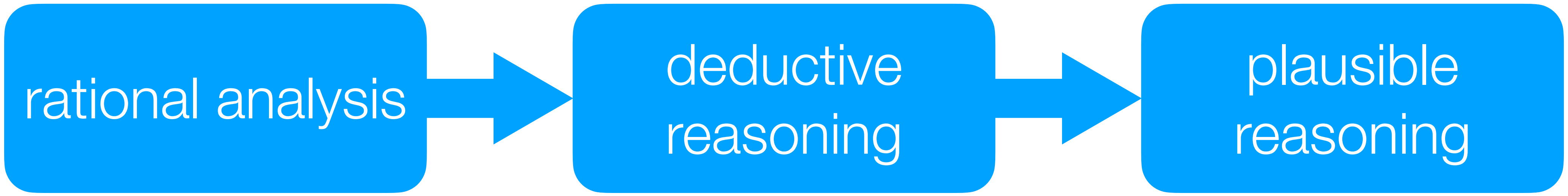
Recap: graphical models

rational analysis

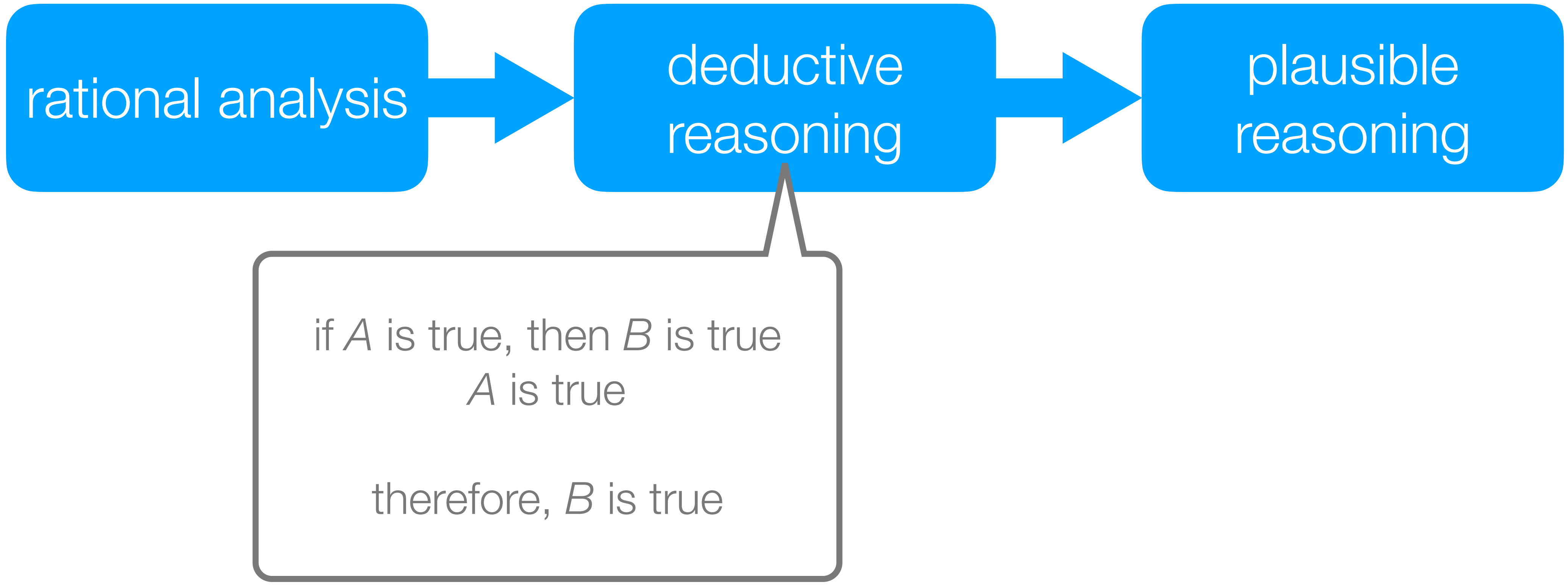
Recap: graphical models



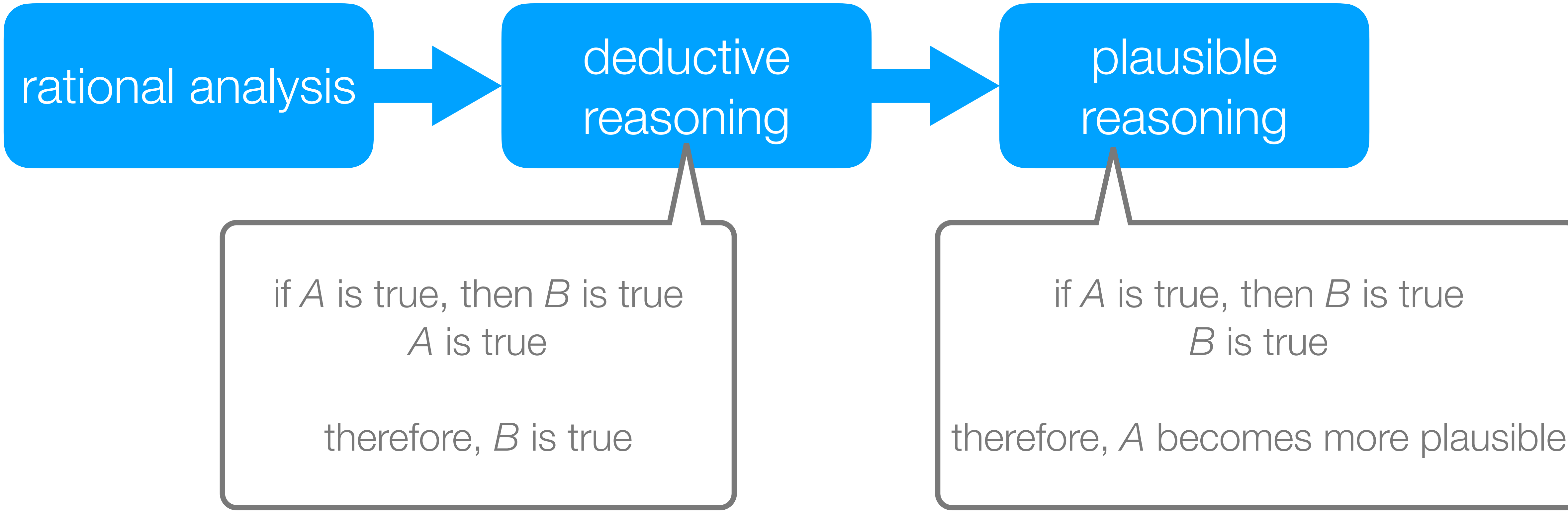
Recap: graphical models



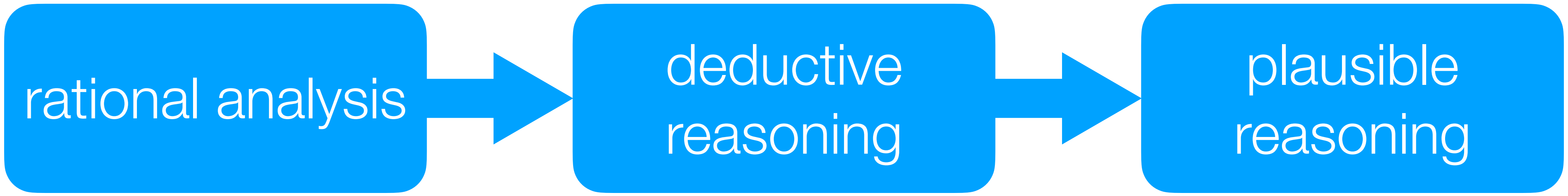
Recap: graphical models



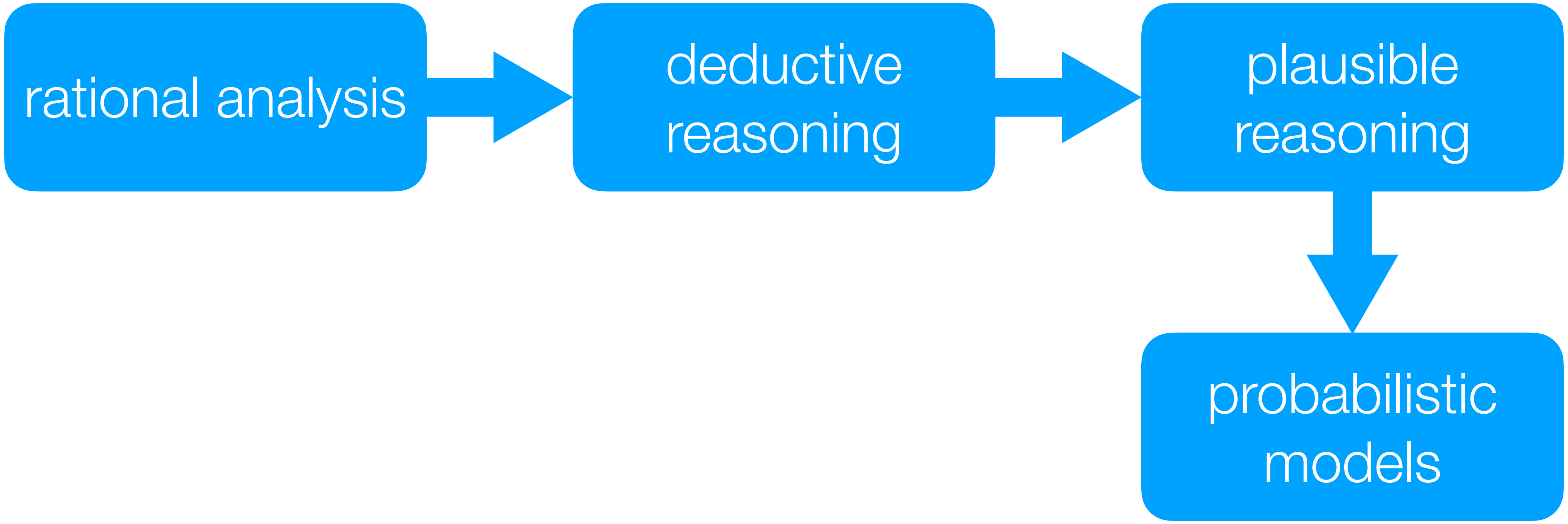
Recap: graphical models



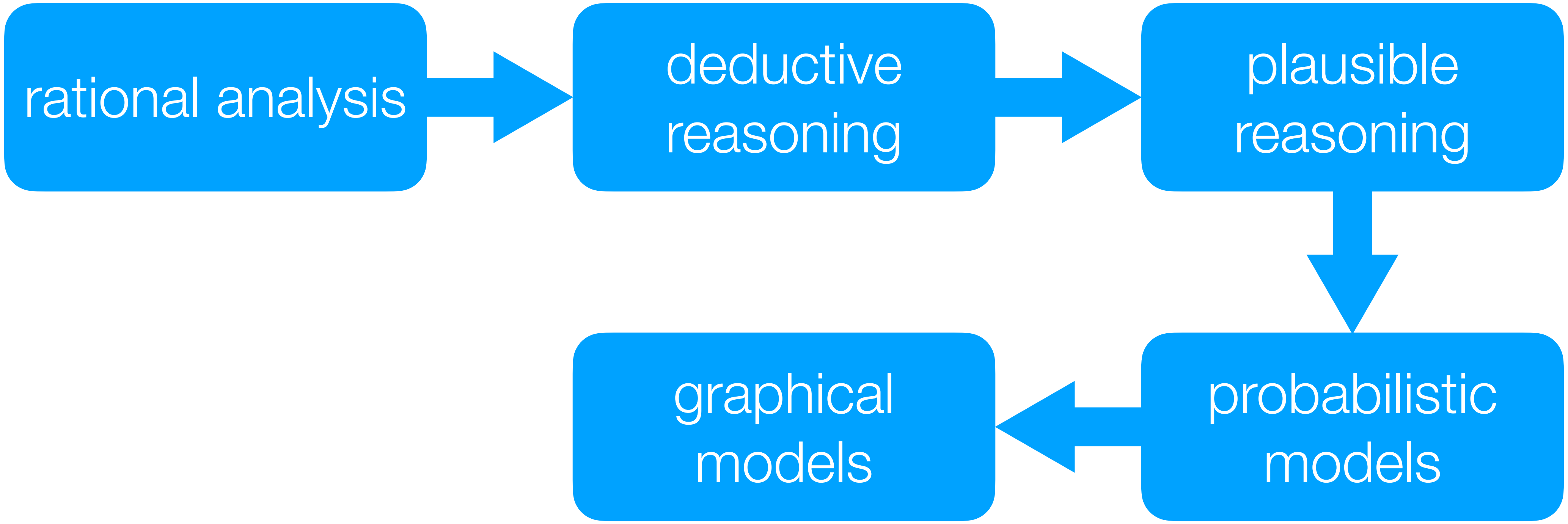
Recap: graphical models



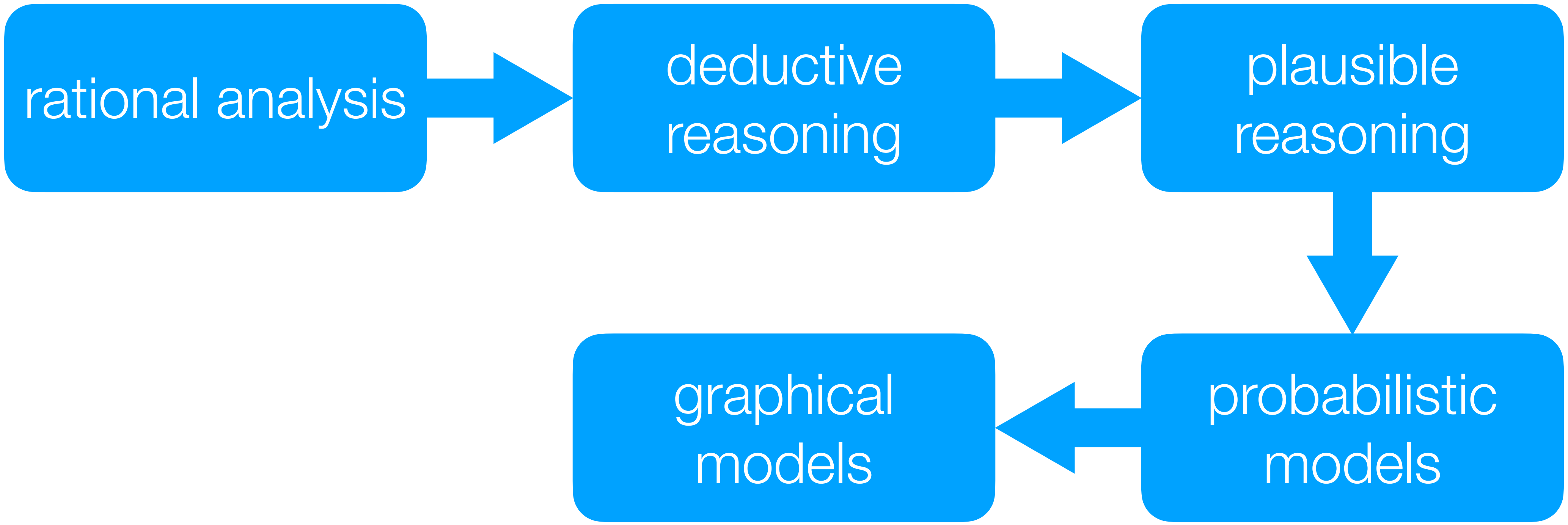
Recap: graphical models



Recap: graphical models

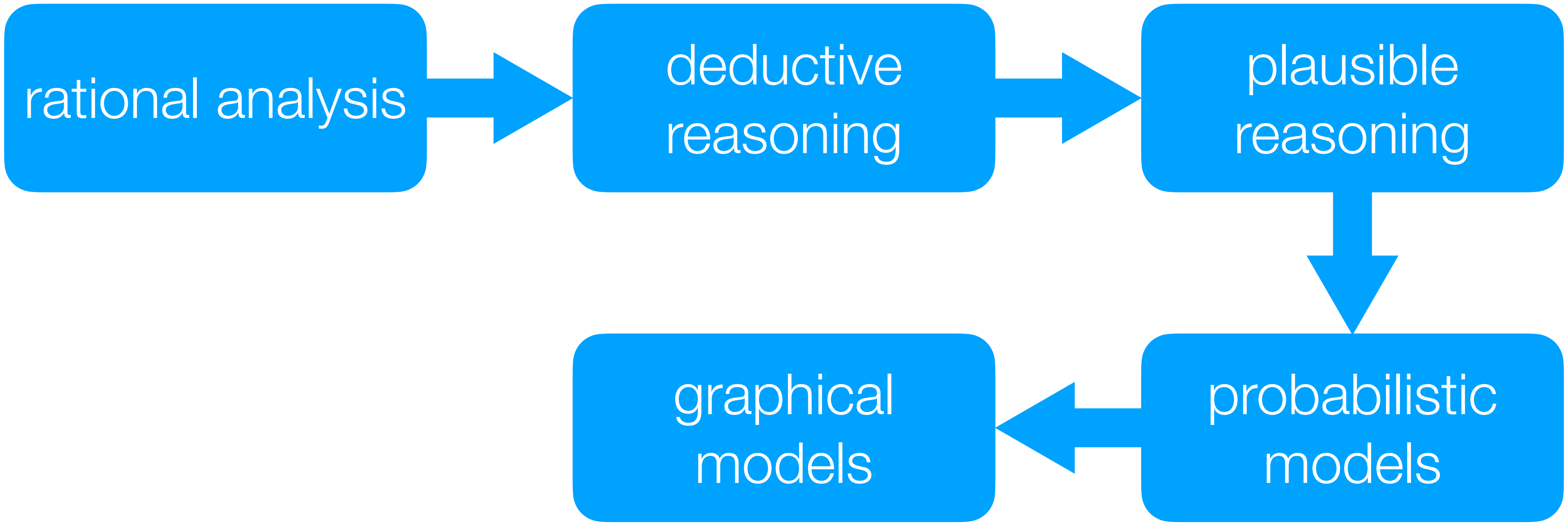


Recap: graphical models



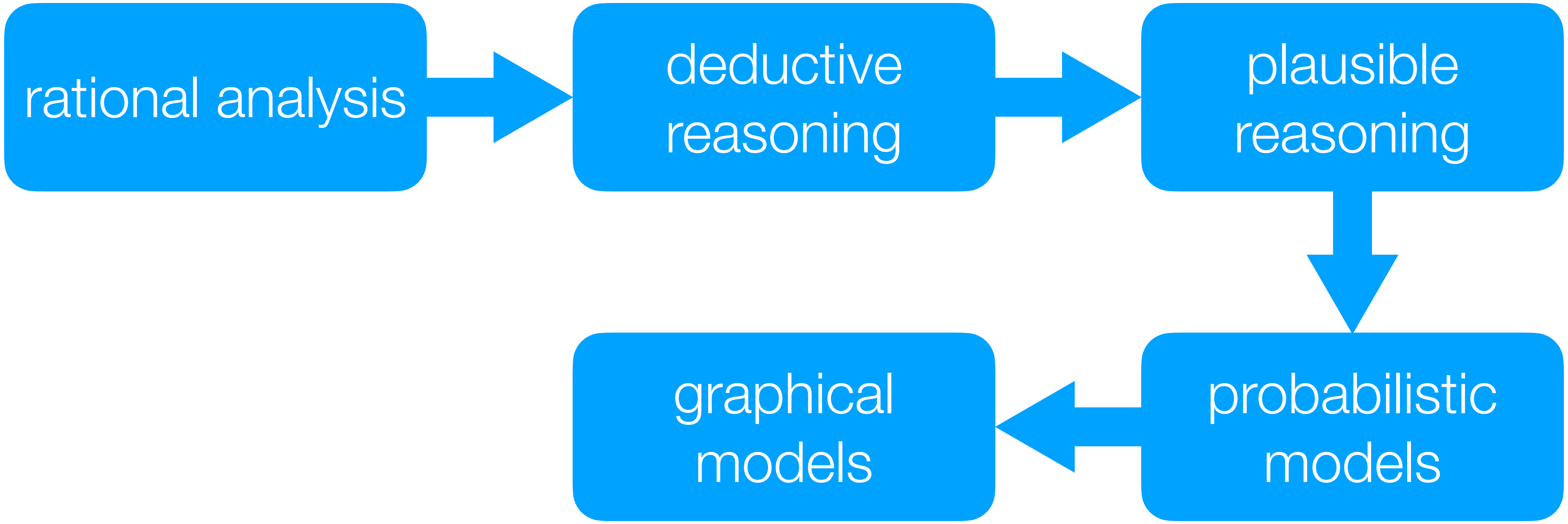
- intuitive parametrization of the probabilistic model (probability table)

Recap: graphical models



- intuitive parametrization of the probabilistic model (probability table)
- independence or conditional independence of variables can be conveniently identified

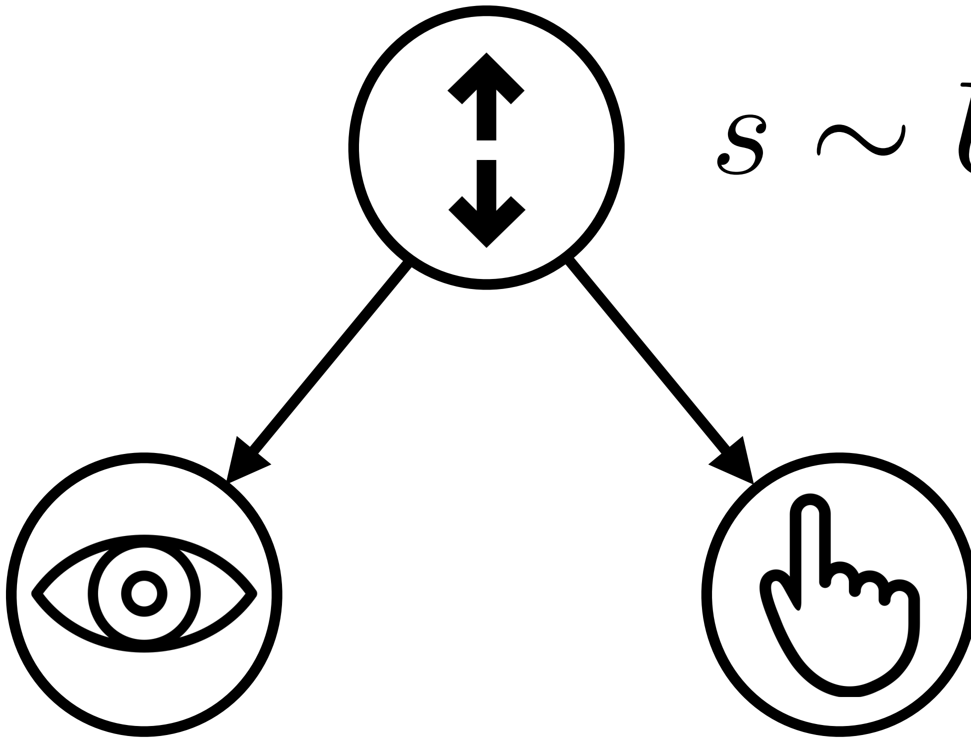
Recap: graphical models



- intuitive parametrization of the probabilistic model (probability table)
- independence or conditional independence of variables can be conveniently identified
- breaks down the joint distribution into simpler conditional

Recap: cue integration

CUE INTEGRATION



$$s \sim Unif(0, M)$$

$$X_v | s \sim N(s, \sigma_v) \quad X_h | s \sim N(s, \sigma_h)$$

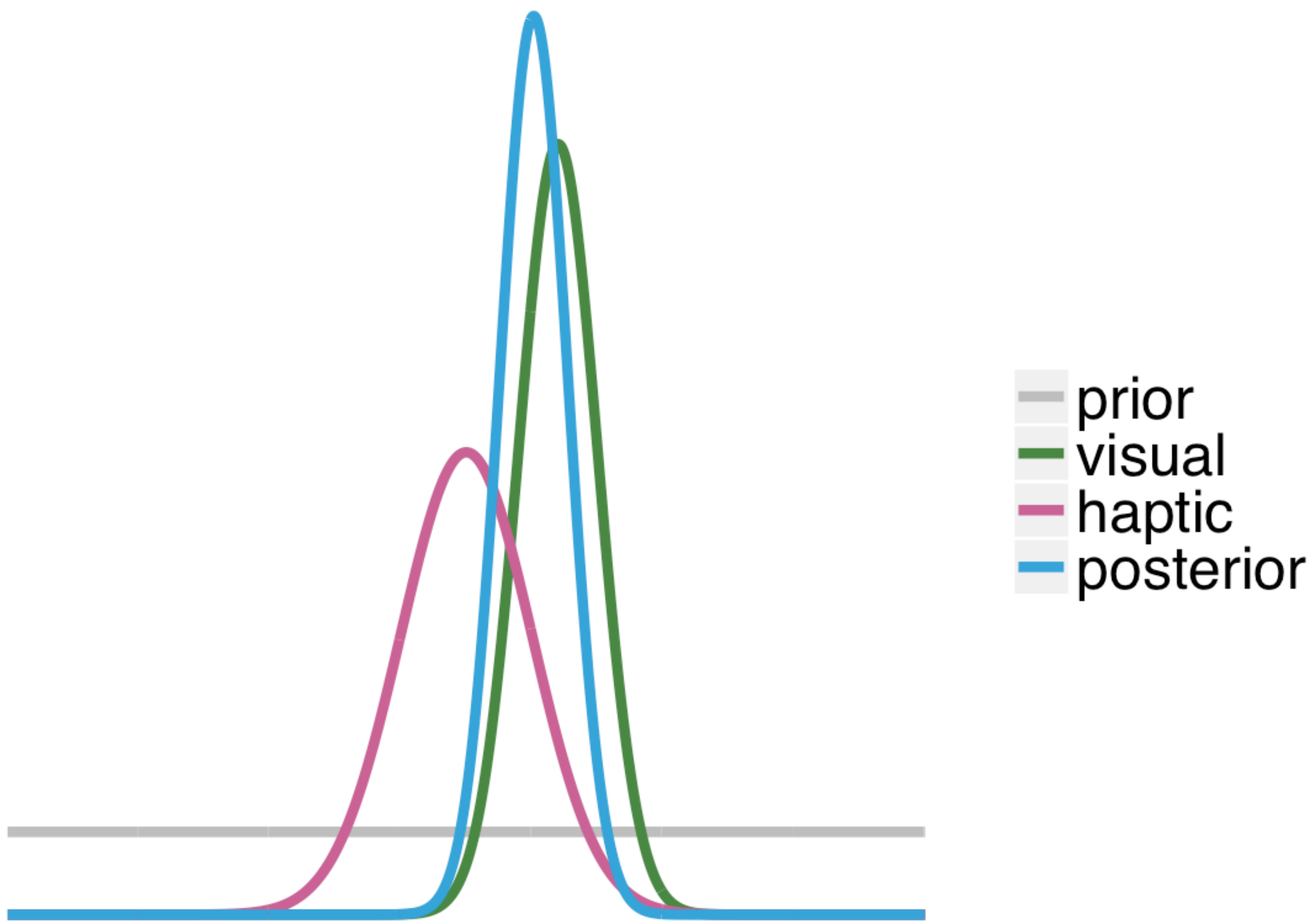
Recap: cue integration

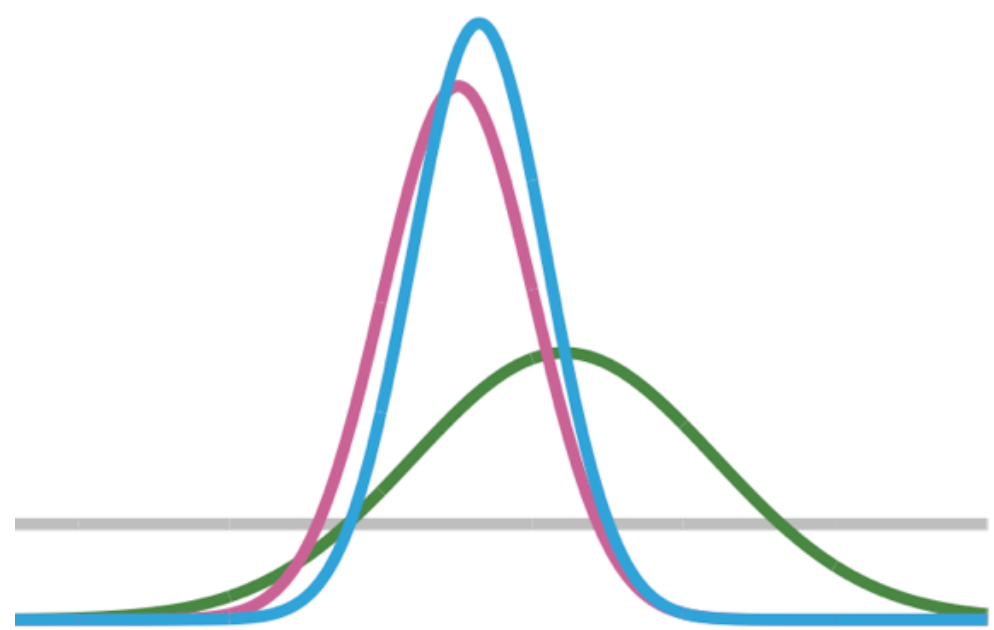
CUE INTEGRATION

$$\hat{s} \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$$

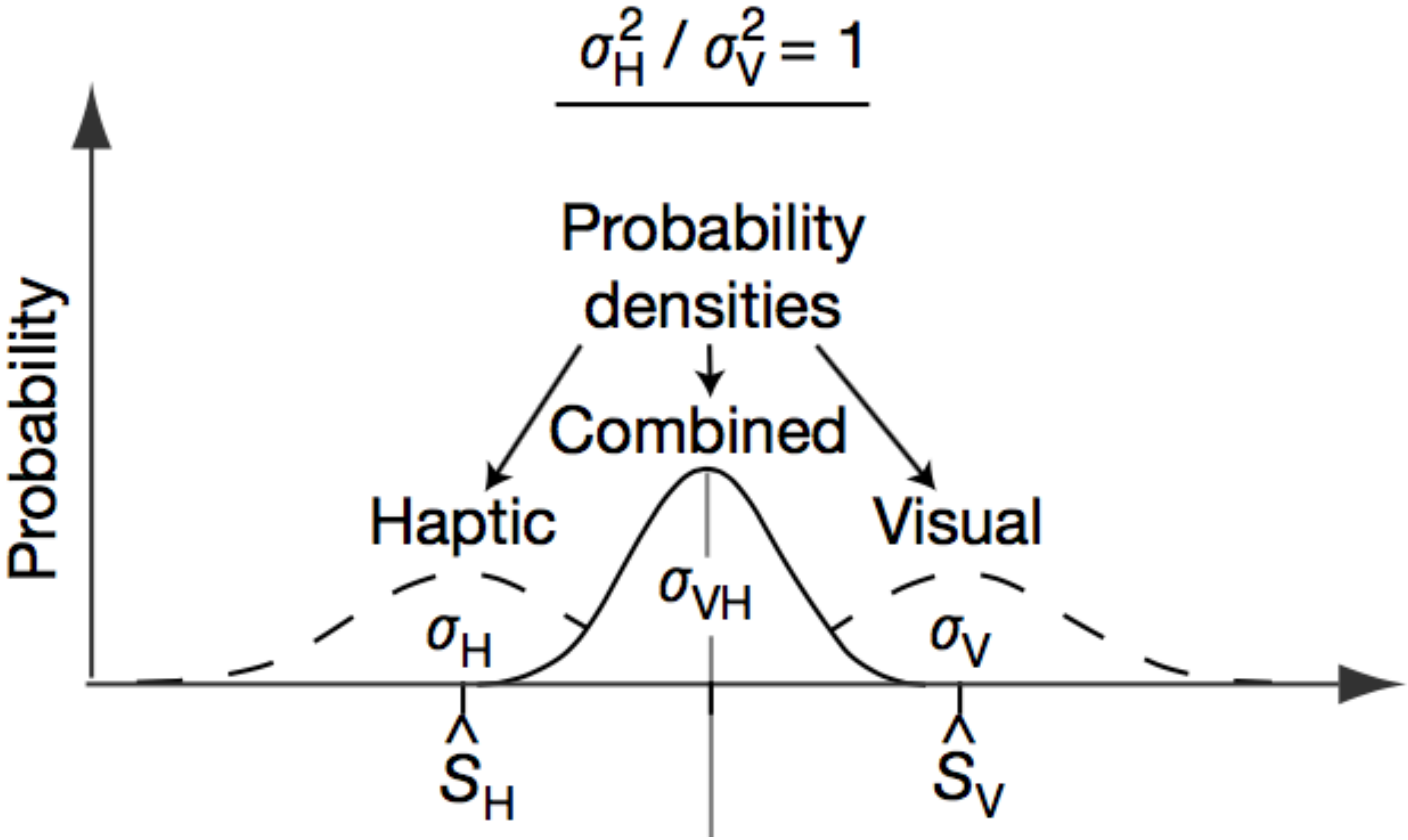
$$\hat{\mu} = \frac{\frac{X_v}{\sigma_v^2} + \frac{X_h}{\sigma_h^2}}{\frac{1}{\sigma_h^2} + \frac{1}{\sigma_v^2}}$$

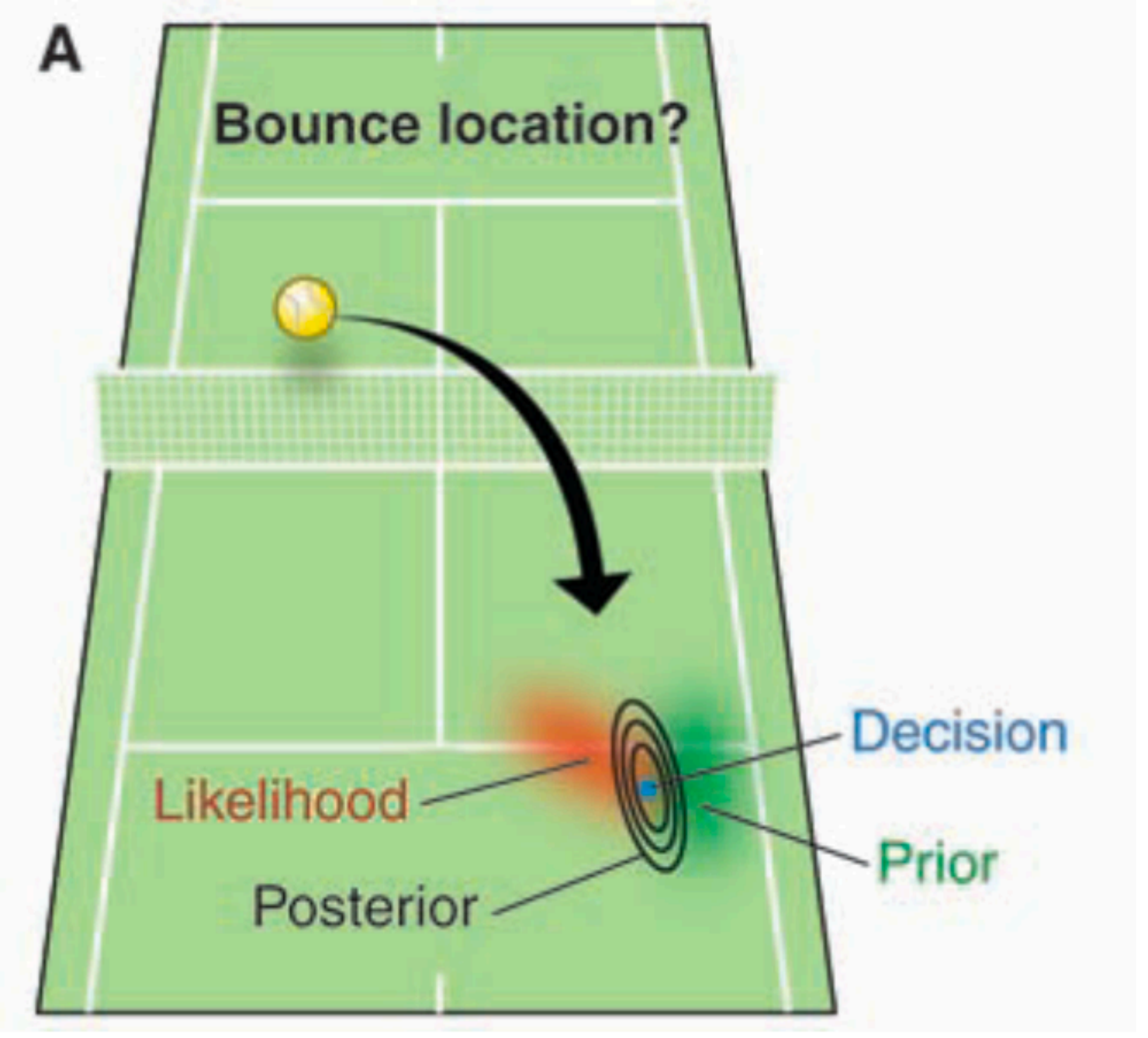
$$\frac{1}{\hat{\sigma}^2} = \frac{1}{\sigma_h^2} + \frac{1}{\sigma_v^2}$$





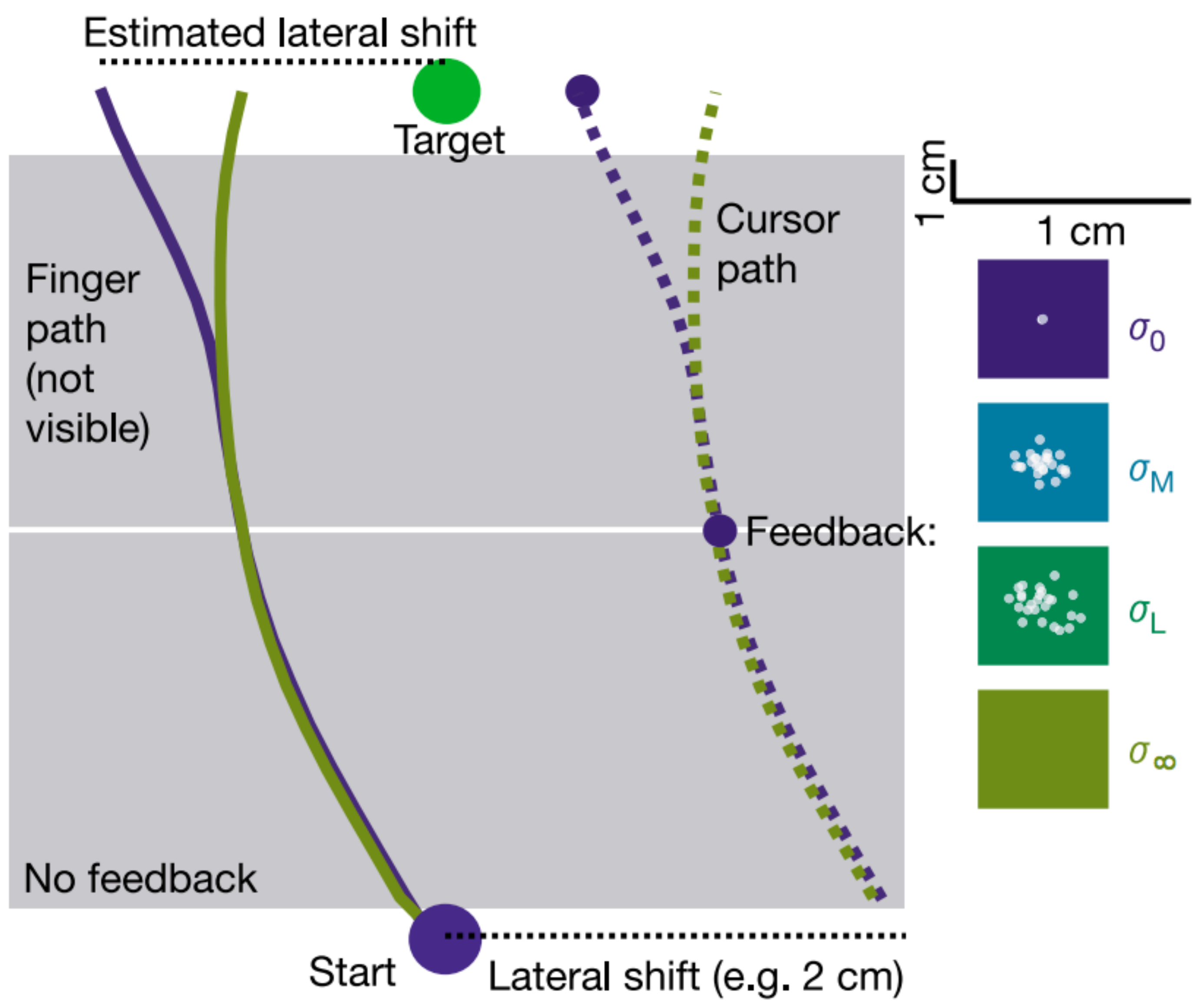
Recap: cue integration




A
Bounce location?
Decision
Likelihood
Prior
Posterior

BAYESIAN INTEGRATION IN SENSORIMOTOR LEARNING

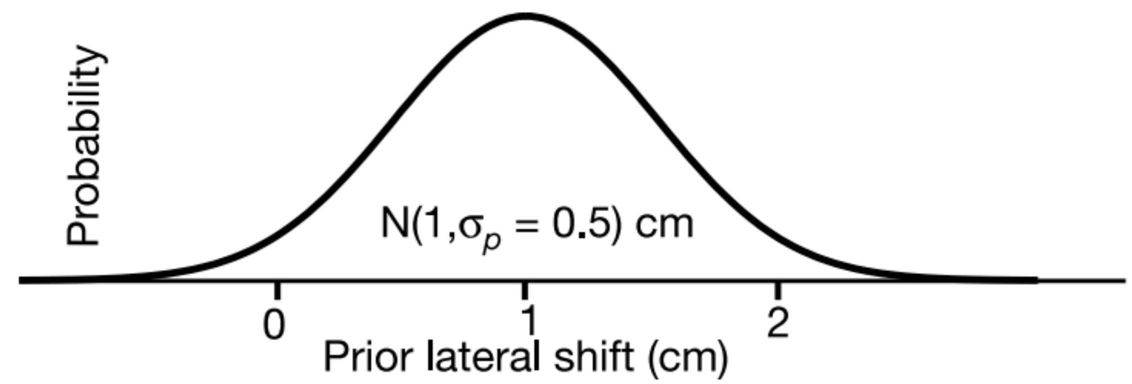
- Mutatóujj két pont közötti mozgatása
- VR setup: nem valódi helyzet, ráadásul csaknéha látható
  - tréning: félúton és út végén
  - többi kondíció: csak félúton



KÖRDING & WOLPERT, 2004

TASK VARIABILITY

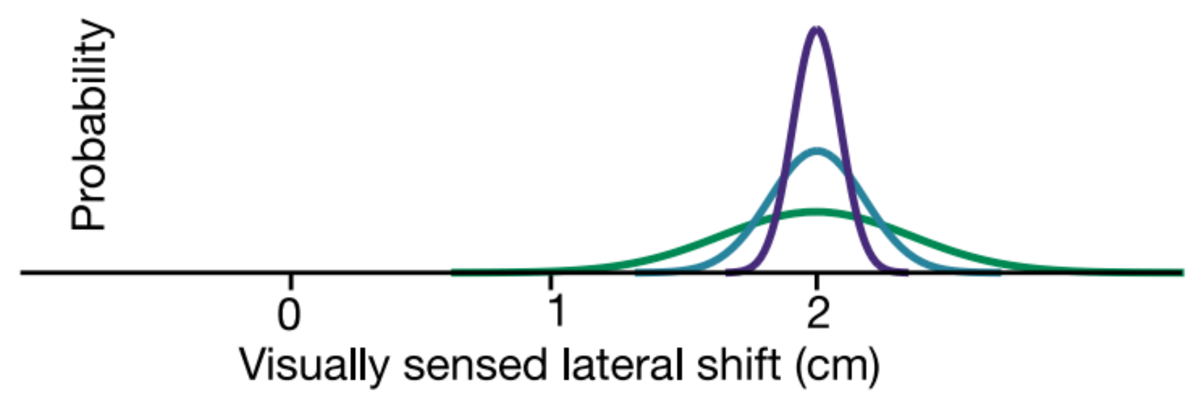
- a megjelenő kurzor eltolása a megadott priorból lett generálva véletlenszerűen
- ezt 1000 tréning próba alatt lehetett kikövetkeztetni



$$p(x) \sim \mathcal{N}(x|\mu = 1cm, \sigma_0 = 0.5cm)$$

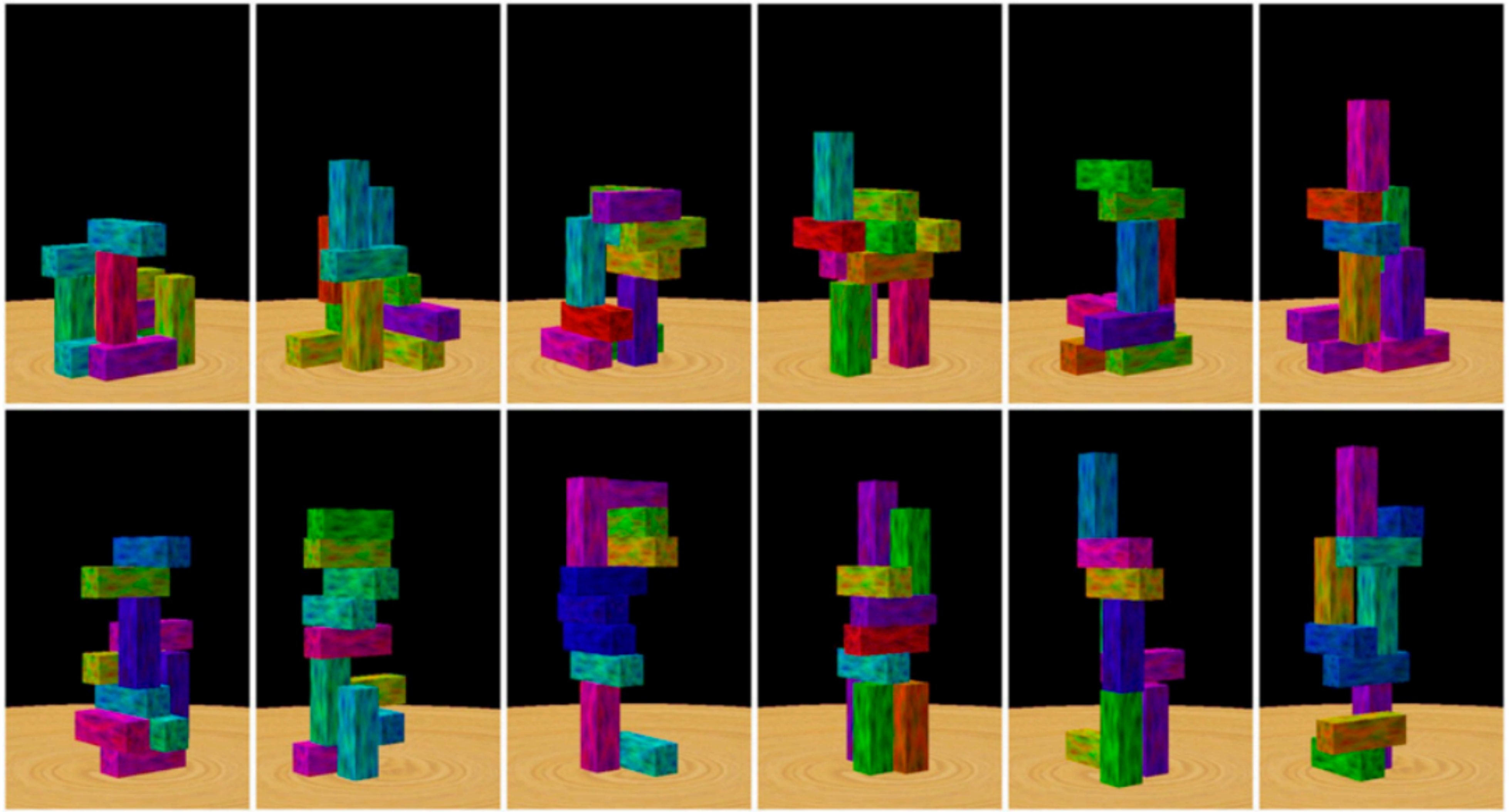
SENSORY VARIABILITY

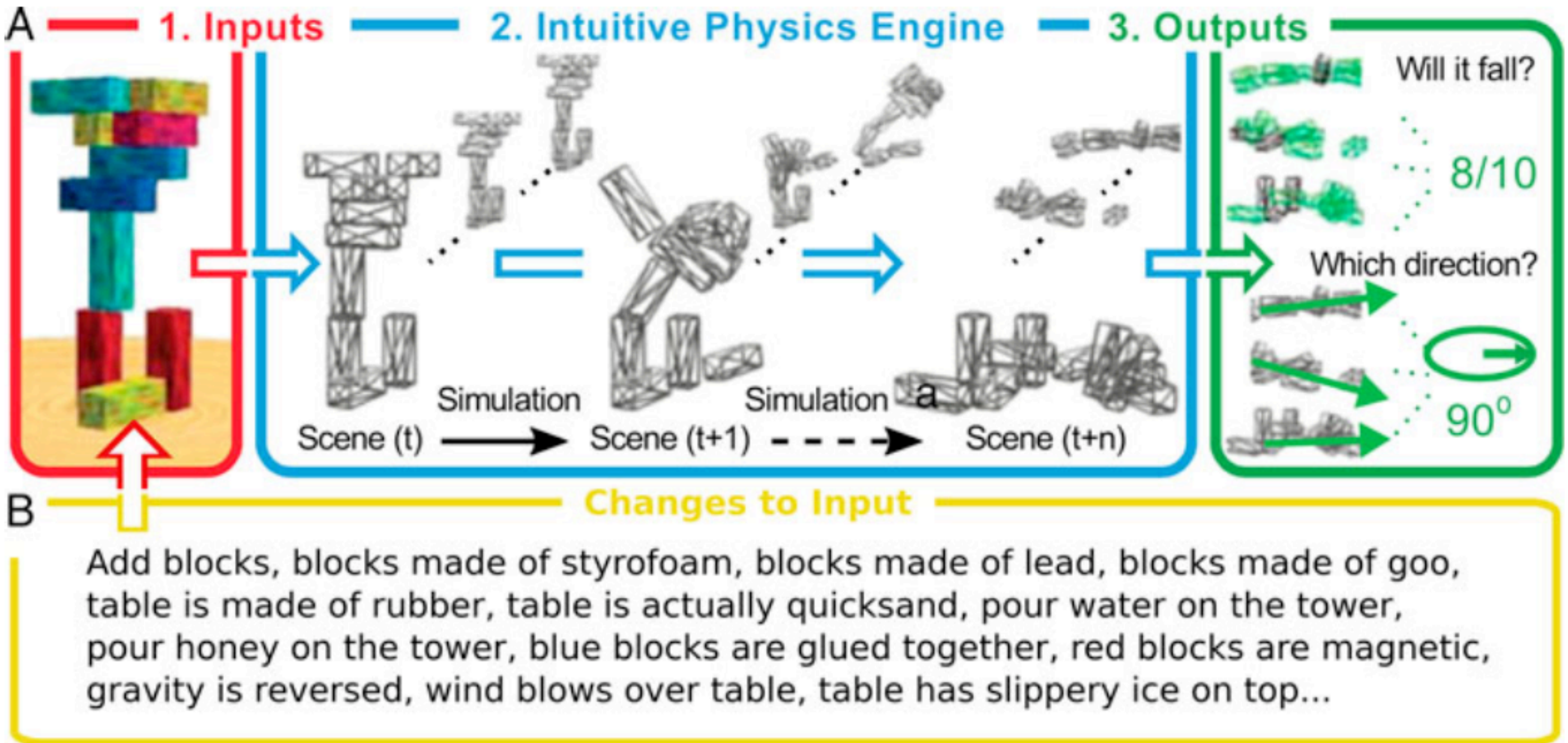
- A különböző kondíciókban a kurzor különböző mértékben volt elmosódva:

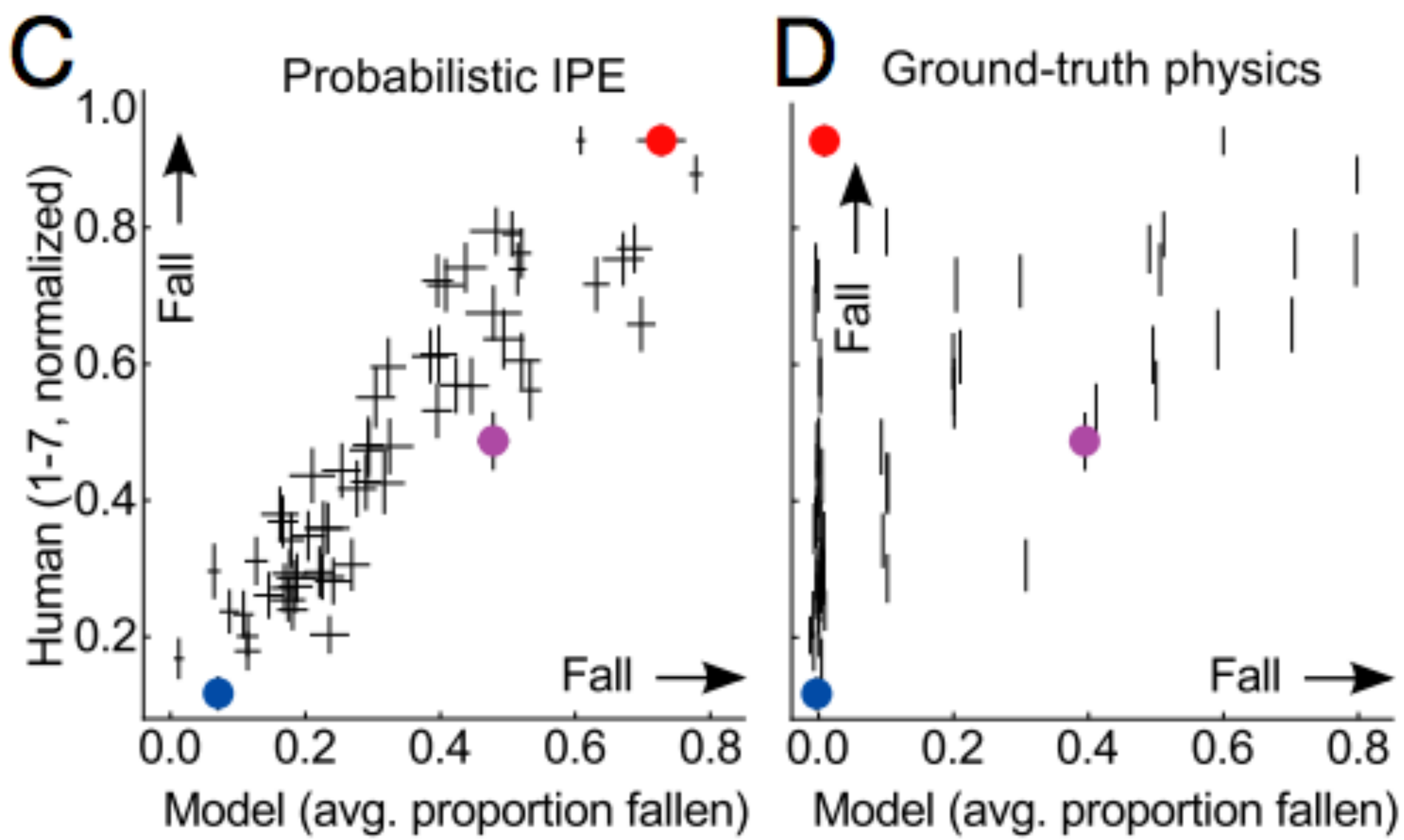


INTUITIVE PHYSICS

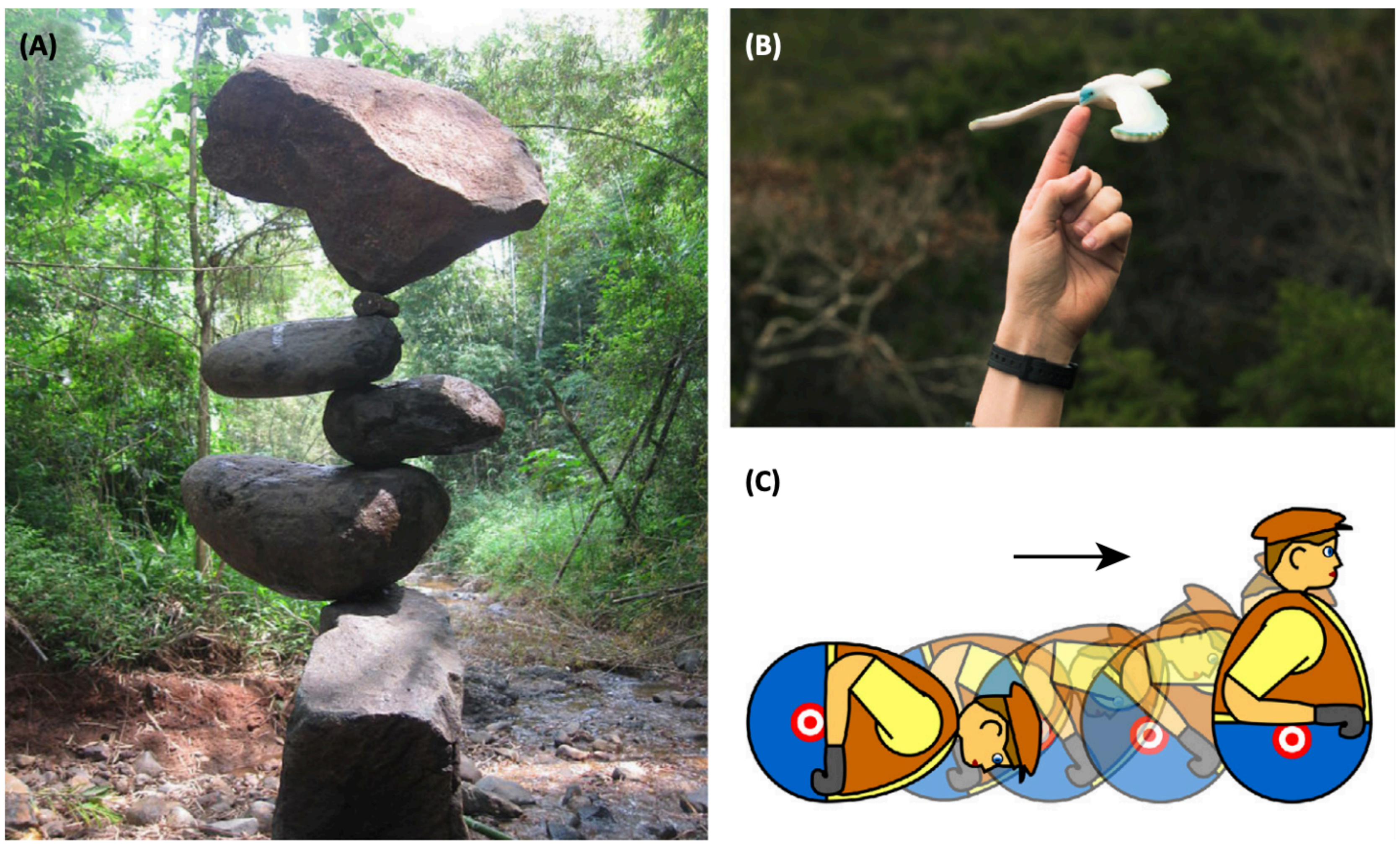
INTUITIVE PHYSICS

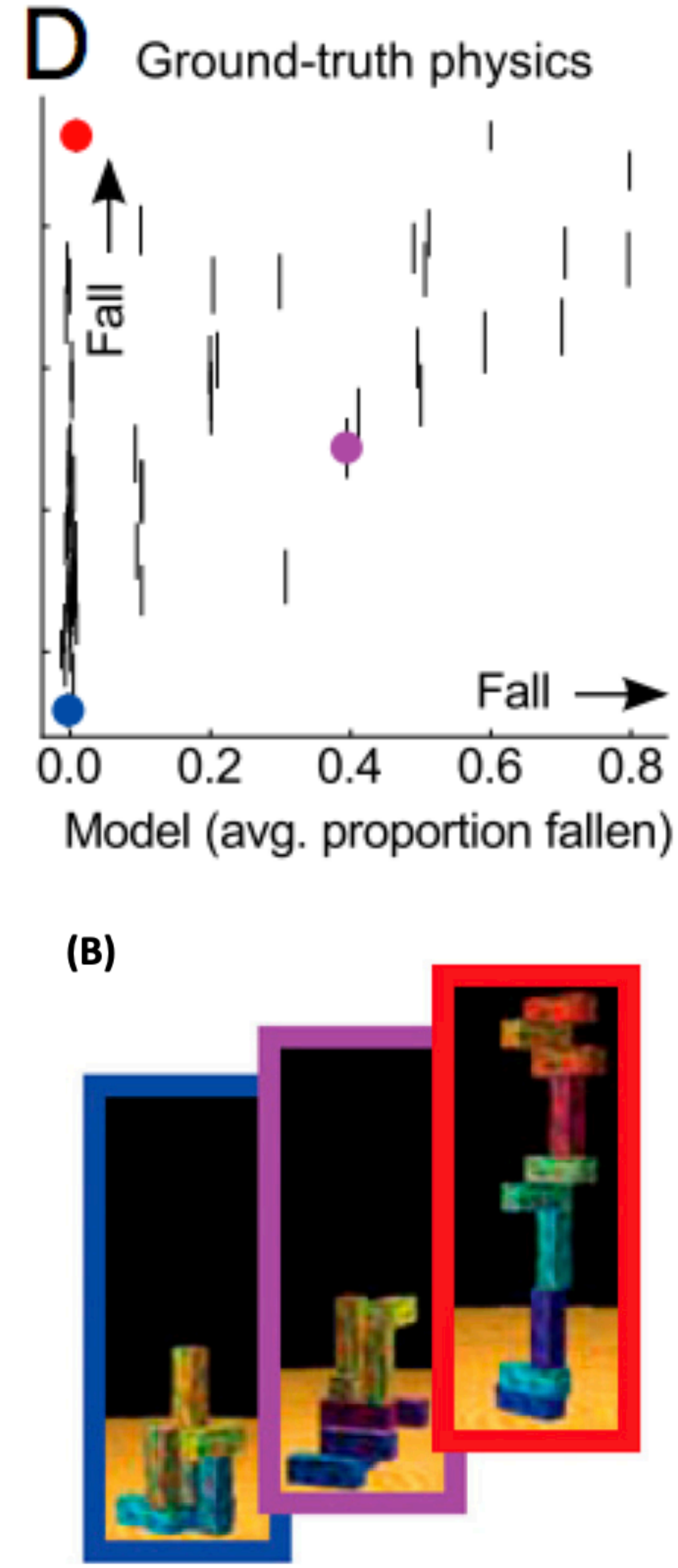



A
1. Inputs
2. Intuitive Physics Engine
3. Outputs
Will it fall?
8/10
Which direction?
90°
Simulation
Simulation
Scene (t)
Scene (t+1)
Scene (t+n)
B
Changes to Input
Add blocks, blocks made of styrofoam, blocks made of lead, blocks made of goo, table is made of rubber, table is actually quicksand, pour water on the tower, pour honey on the tower, blue blocks are glued together, red blocks are magnetic, gravity is reversed, wind blows over table, table has slippery ice on top...

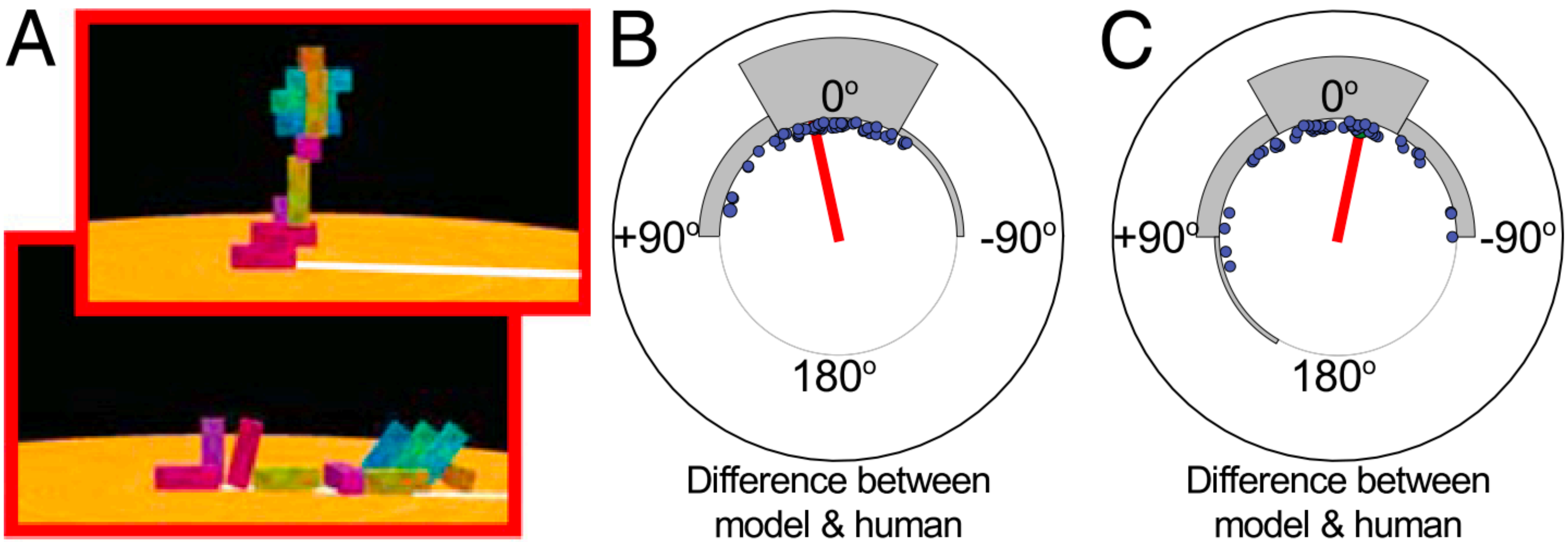

C
Probabilistic IPE
Human (1-7, normalized)
1.0
0.8
0.6
0.4
0.2
Fall
Fall
0.0
0.2
0.4
0.6
0.8
Model (avg. proportion fallen)
D
Ground-truth physics
Fall
Fall
0.0
0.2
0.4
0.6
0.8
Model (avg. proportion fallen)

PHYSICAL ILLUSIONS





WHICH DIRECTION?



Graphical model example

Graphical model example

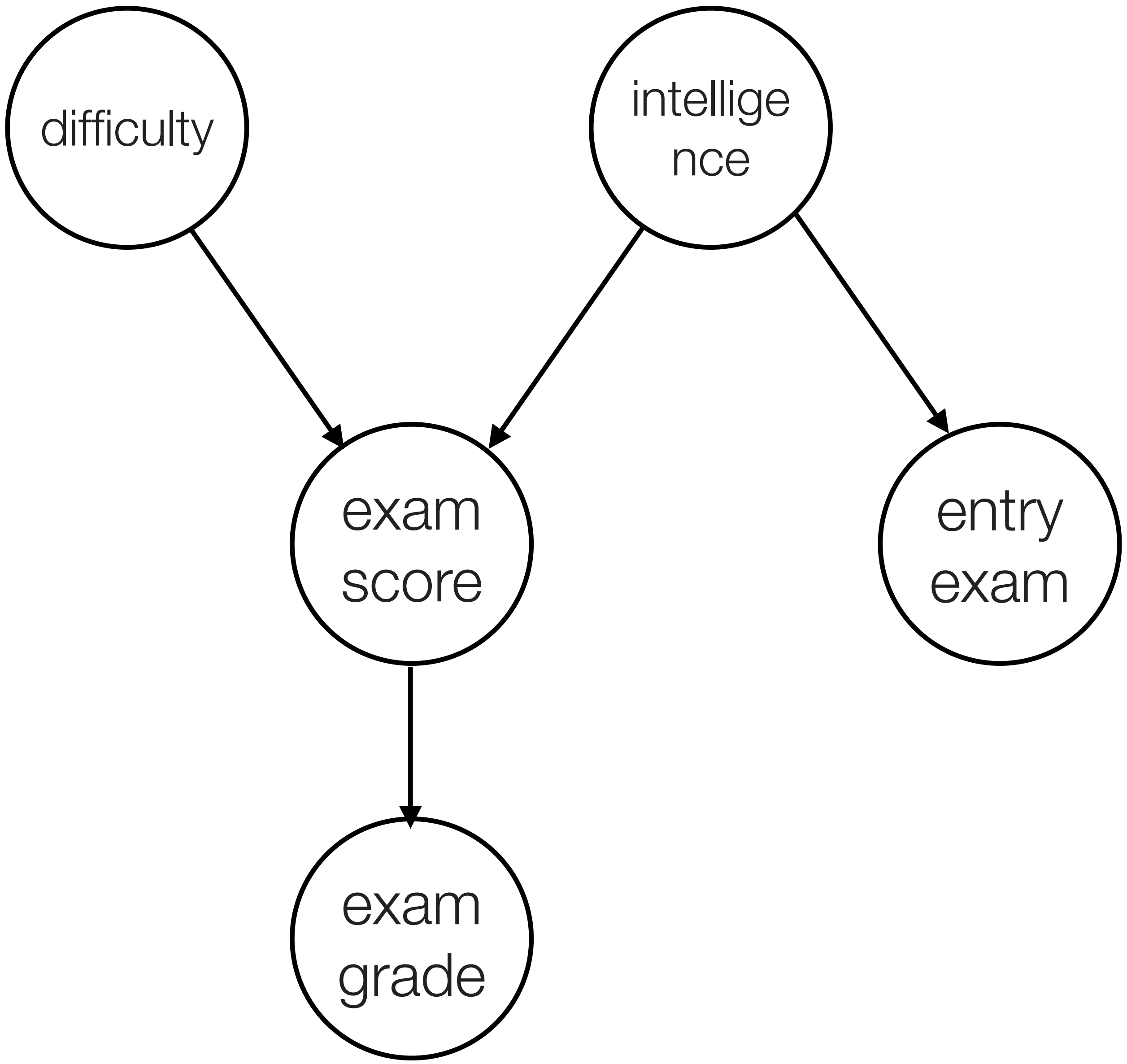
- learn about a five dimensional model
  - exam difficulty (d)
  - intellectual capacity (i)
  - entry exam quality (y)
  - exam score obtained (s)
  - exam grade obtained (g)

Graphical model example

- learn about a five dimensional model
  - exam difficulty (d)
  - intellectual capacity (i)
  - entry exam quality (y)
  - exam score obtained (s)
  - exam grade obtained (g)
- the joint distribution: $P(g, s, y, d, i)$

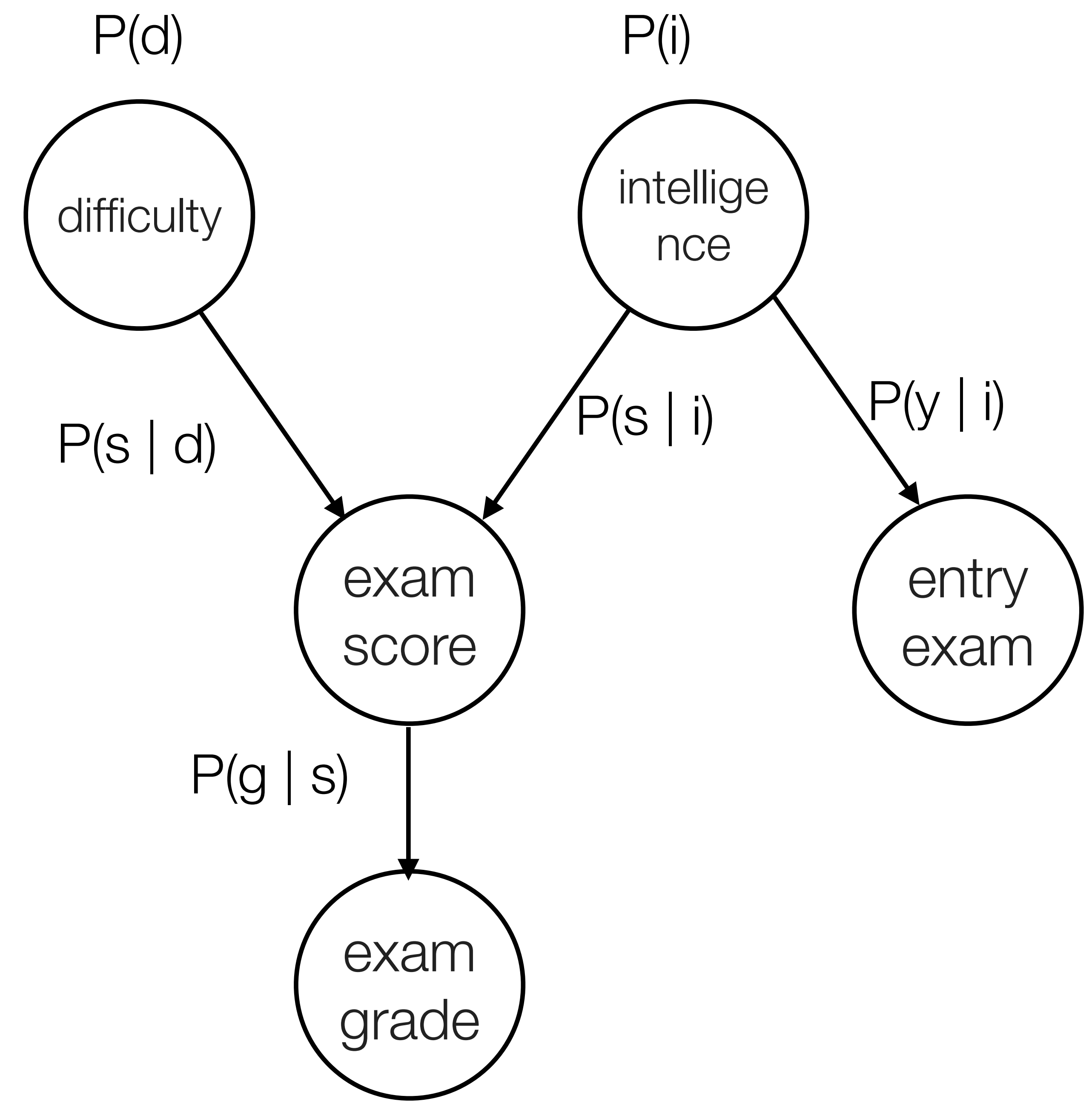
Graphical model example

- learn about a five dimensional model
  - exam difficulty (d)
  - intellectual capacity (i)
  - entry exam quality (y)
  - exam score obtained (s)
  - exam grade obtained (g)
- the joint distribution: P(g, s, y, d, i)



Graphical model example

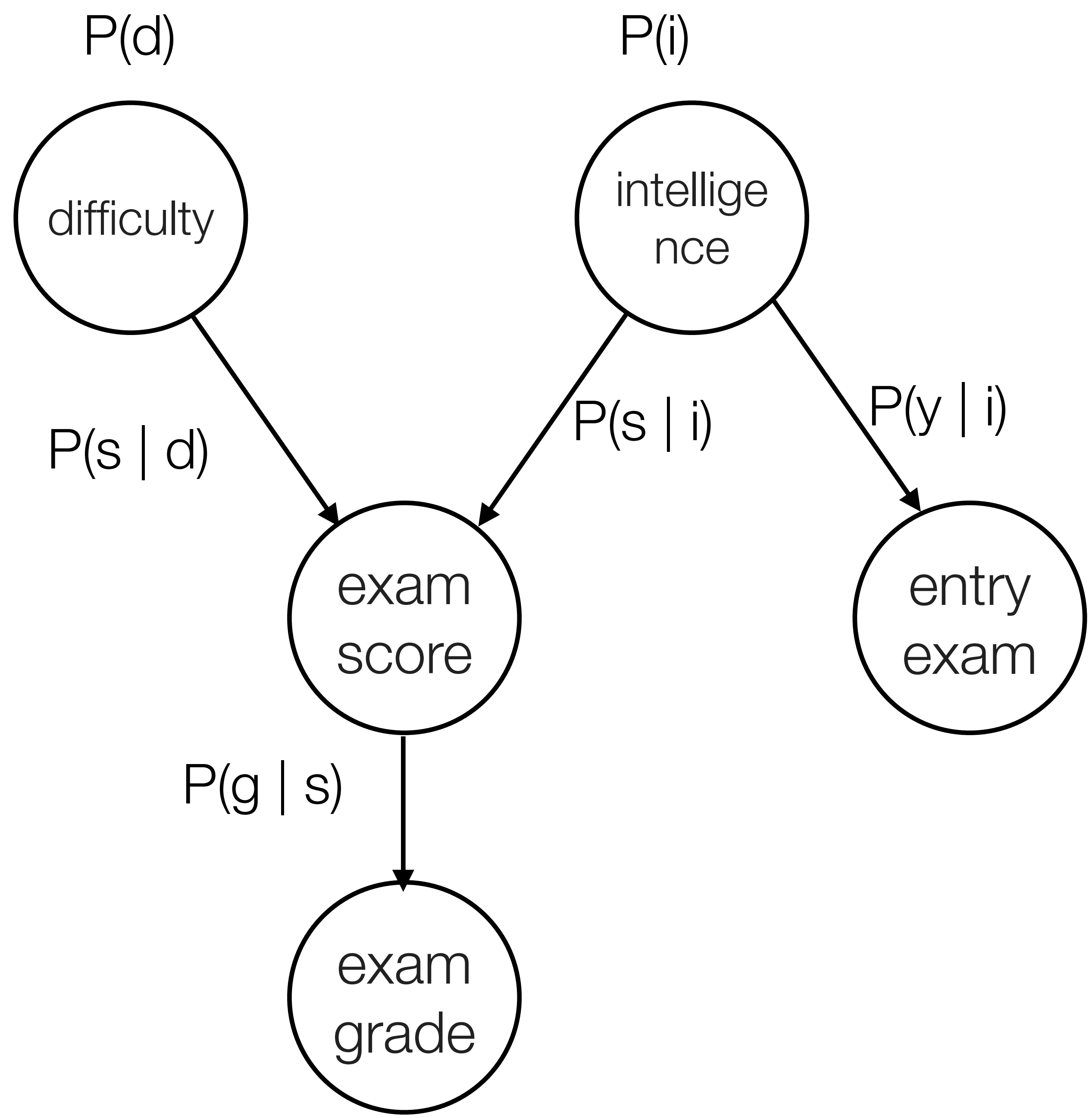
- learn about a five dimensional model
  - exam difficulty (d)
  - intellectual capacity (i)
  - entry exam quality (y)
  - exam score obtained (s)
  - exam grade obtained (g)
- the joint distribution: P(g, s, y, d, i)



Graphical model example

- learn about a five dimensional model
  - exam difficulty (d)
  - intellectual capacity (i)
  - entry exam quality (y)
  - exam score obtained (s)
  - exam grade obtained (g)
- the joint distribution: $P(g, s, y, d, i)$

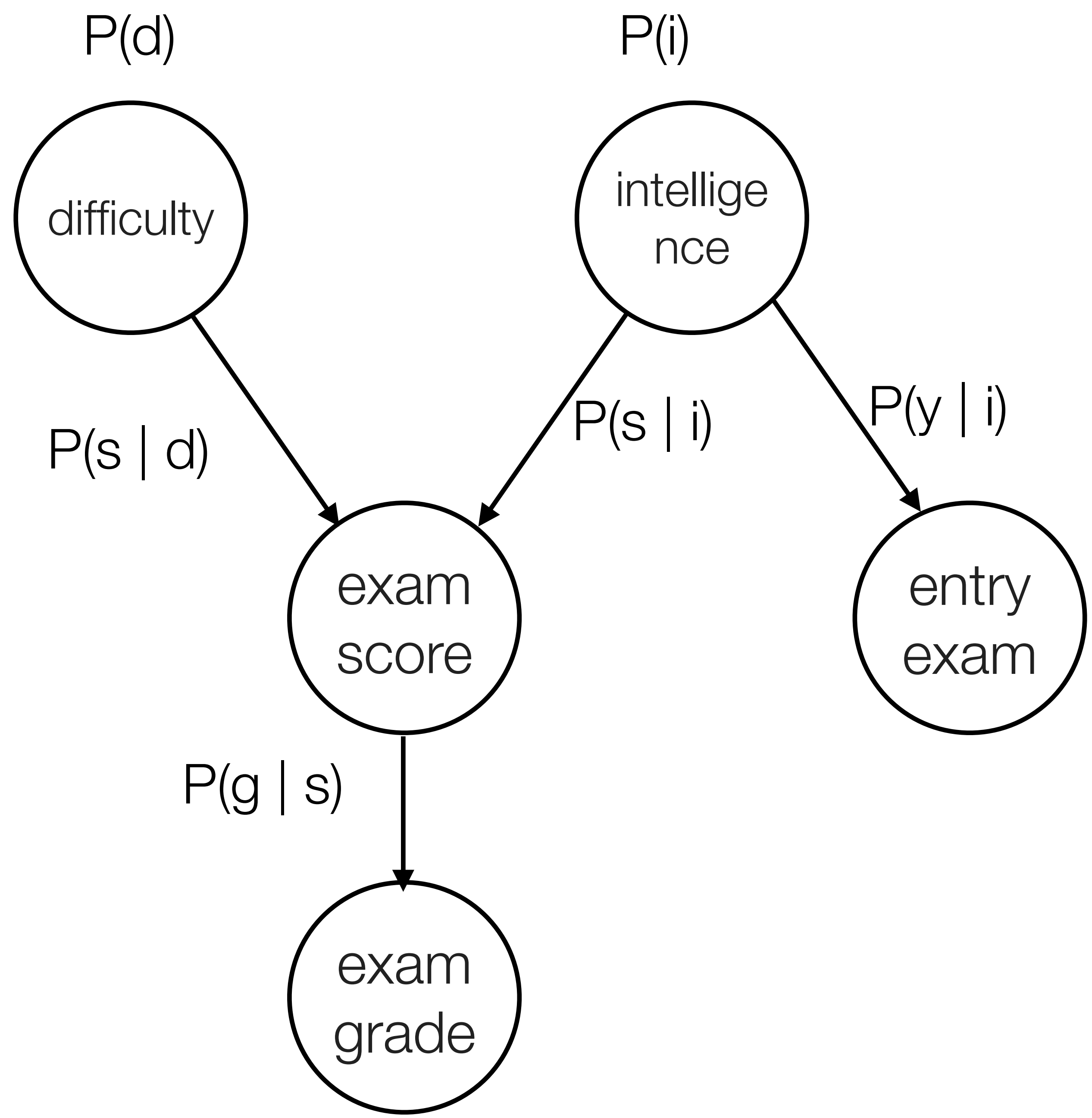
$$= P(g \mid s, y, d, i)\, P(s, y, d, i)$$



Graphical model example

- learn about a five dimensional model
  - exam difficulty (d)
  - intellectual capacity (i)
  - entry exam quality (y)
  - exam score obtained (s)
  - exam grade obtained (g)
- the joint distribution: $P(g, s, y, d, i)$

$$= P(g \mid s, y, d, i)\, P(s, y, d, i)$$

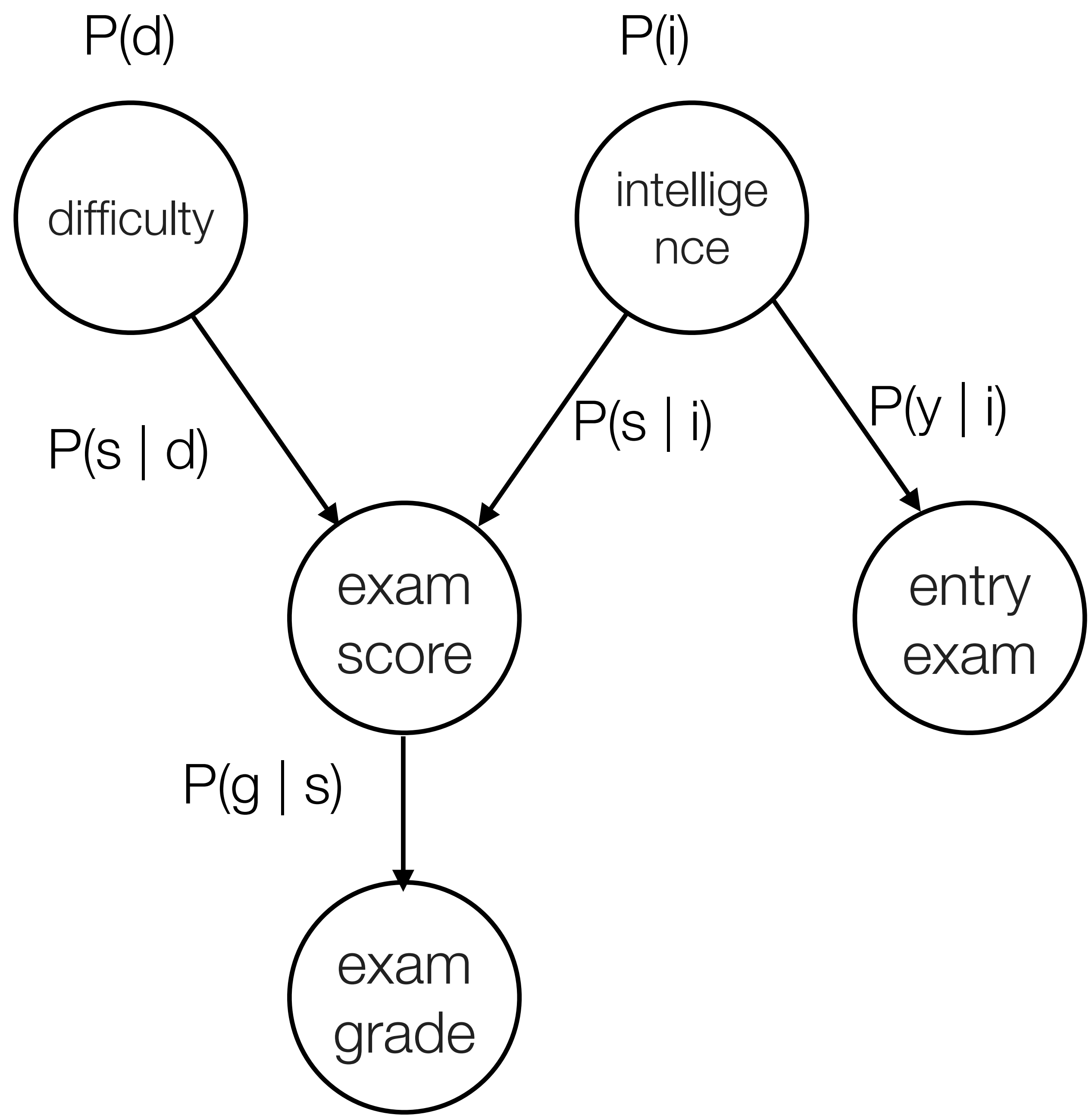


Graphical model example

- learn about a five dimensional model
  - exam difficulty (d)
  - intellectual capacity (i)
  - entry exam quality (y)
  - exam score obtained (s)
  - exam grade obtained (g)
- the joint distribution: $P(g, s, y, d, i)$

$$= P(g \mid s, y, d, i)\, P(s, y, d, i)$$

$$= P(g \mid s)\, P(s \mid y, d, i)\, P(y, d, i)$$

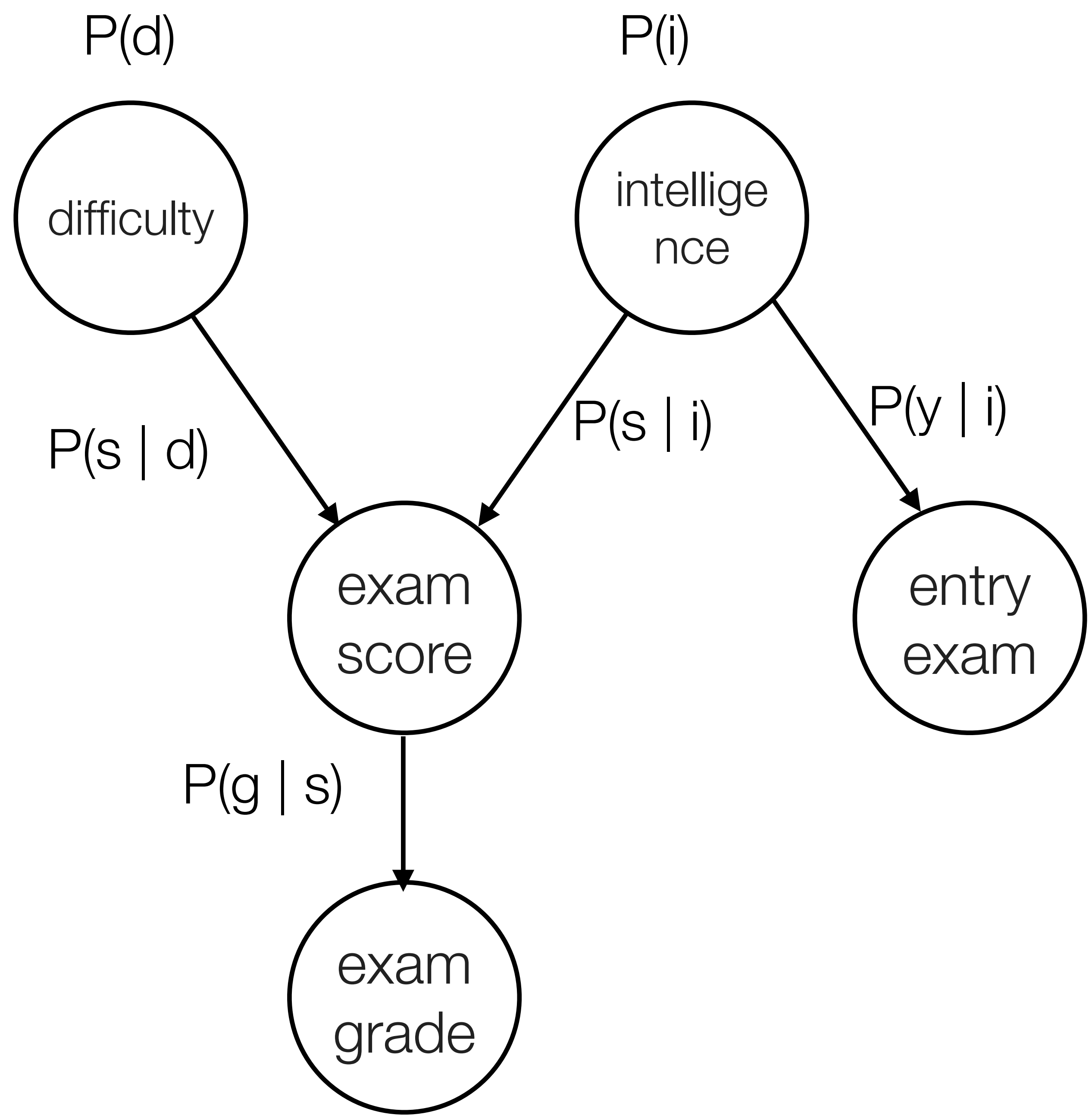


Graphical model example

- learn about a five dimensional model
  - exam difficulty (d)
  - intellectual capacity (i)
  - entry exam quality (y)
  - exam score obtained (s)
  - exam grade obtained (g)
- the joint distribution: P(g, s, y, d, i)

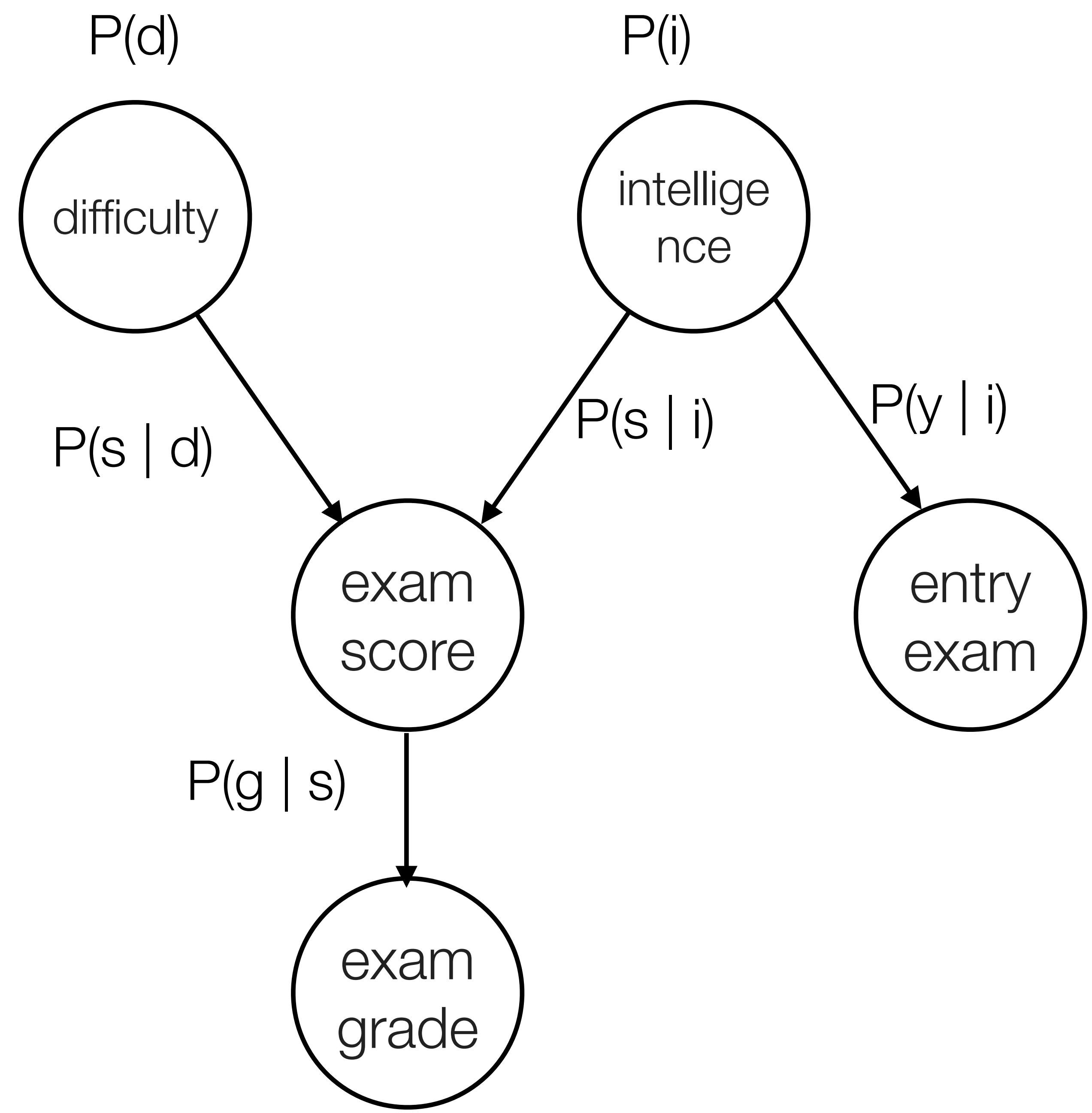
$$= P(g \mid s, \cancel{y}, \cancel{d}, \cancel{i})\, P(s, y, d, i)$$

$$= P(g \mid s)\, P(s \mid \cancel{y}, d, i)\, P(y, d, i)$$



Graphical model example

- learn about a five dimensional model
  - exam difficulty (d)
  - intellectual capacity (i)
  - entry exam quality (y)
  - exam score obtained (s)
  - exam grade obtained (g)
- the joint distribution: $P(g, s, y, d, i)$

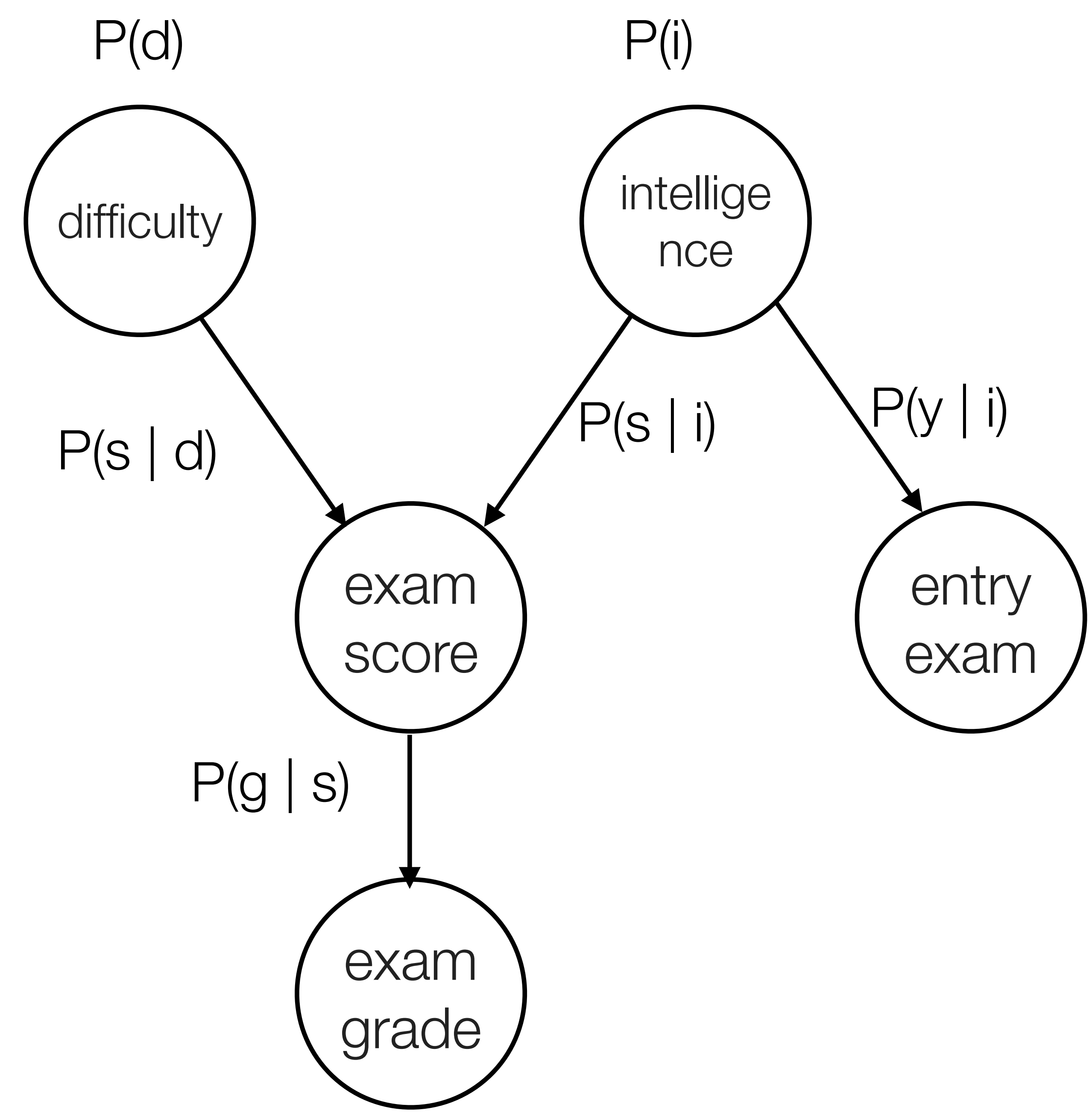


$$= P(g \mid s, y, d, i)\, P(s, y, d, i)$$

$$= P(g \mid s)\, P(s \mid y, d, i)\, P(y, d, i) \quad = P(g \mid s)\, P(s \mid d, i)\, P(y \mid d, i)\, P(d, i)$$

Graphical model example

- learn about a five dimensional model
  - exam difficulty (d)
  - intellectual capacity (i)
  - entry exam quality (y)
  - exam score obtained (s)
  - exam grade obtained (g)
- the joint distribution: $P(g, s, y, d, i)$

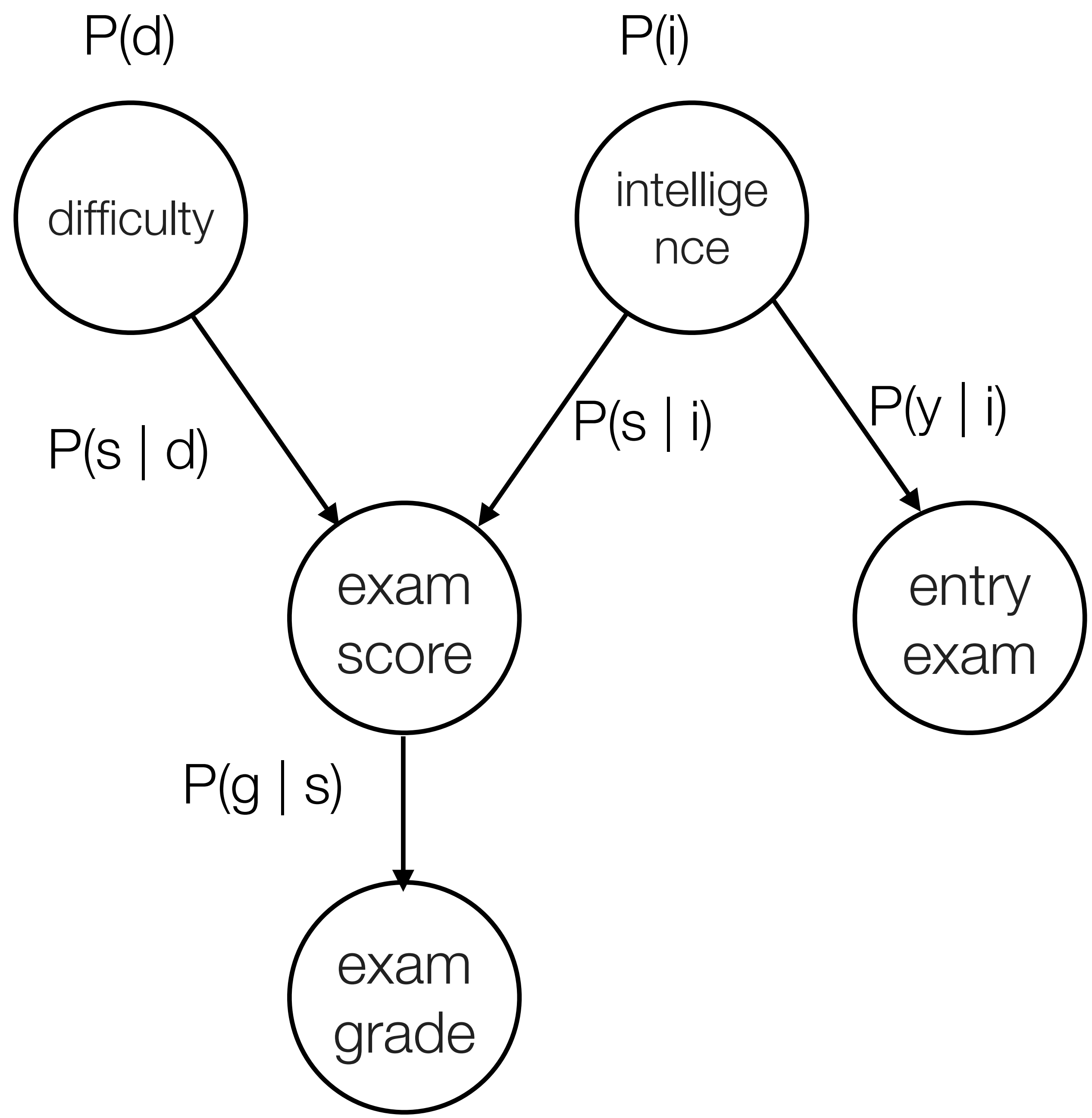


$$= P(g \mid s, y, d, i)\, P(s, y, d, i)$$

$$= P(g \mid s)\, P(s \mid y, d, i)\, P(y, d, i) \quad = P(g \mid s)\, P(s \mid d, i)\, P(y \mid d, i)\, P(d, i)$$

Graphical model example

- learn about a five dimensional model
  - exam difficulty (d)
  - intellectual capacity (i)
  - entry exam quality (y)
  - exam score obtained (s)
  - exam grade obtained (g)
- the joint distribution: $P(g, s, y, d, i)$



$$= P(g \mid s, \cancel{y}, \cancel{d}, \cancel{i})\, P(s, y, d, i)$$

$$= P(g \mid s)\, P(s \mid \cancel{y}, d, i)\, P(y, d, i) \quad = P(g \mid s)\, P(s \mid d, i)\, P(y \mid \cancel{d}, i)\, P(d, i)$$

$$= P(g \mid s)\, P(s \mid d, i)\, P(y \mid i)\, P(d)\, P(i)$$

Graphical models vs probability tables

Assignment The following probabilistic model is learned in Japan about occurrence of earthquakes, public radio announcements on tsunami alerts, and car alarms setting off

earthquake	radio announcement	alarm	P(E,R,A)
1	1	1	0.036
1	1	0	0.054
1	0	1	0.004
1	0	0	0.006
0	1	1	0.00045
0	1	0	0.00855
0	0	1	0.04455
0	0	0	0.84645

Graphical models vs probability tables

Assignment The following probabilistic model is learned in Japan about occurrence of earthquakes, public radio announcements on tsunami alerts, and car alarms setting off

earthquake	radio announcement	alarm	P(E,R,A)
1	1	1	0.036
1	1	0	0.054
1	0	1	0.004
1	0	0	0.006
0	1	1	0.00045
0	1	0	0.00855
0	0	1	0.04455
0	0	0	0.84645

We would like to answer the following questions:

* What is the probability of earthquakes to happen?
* Is the radio announcement independent of the alarm turning on?
* If not, can a conditional independence relationship be established?
* Can you write up a graphical model of this data?
* What are the parameters of the graphical model?

Application: Causal learning in infants

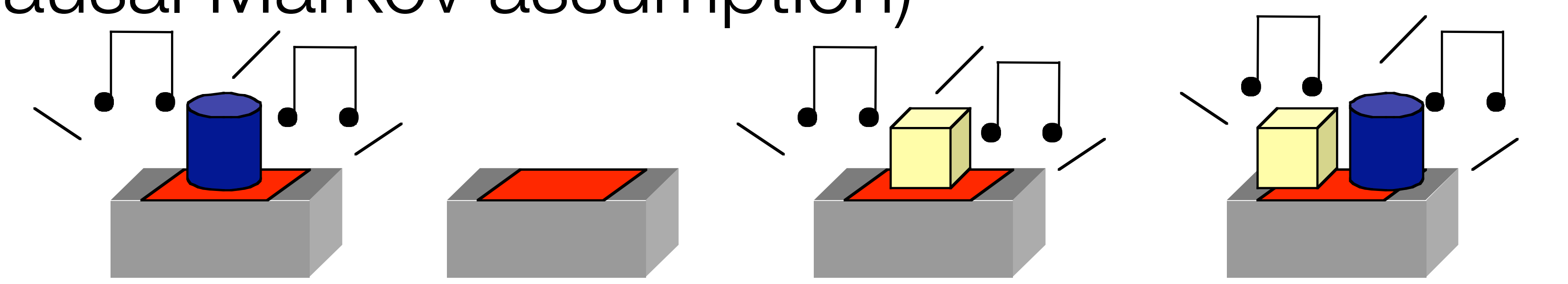
Gopnik et al (2004) Cog Sci

- Representation of causal structure is through graphical models (directed acyclic graphs, DAGs)
- Causal structure imply conditional independencies (causal Markov assumption)

Application: Causal learning in infants

Gopnik et al (2004) Cog Sci

- Representation of causal structure is through graphical models (directed acyclic graphs, DAGs)
- Causal structure imply conditional independencies (causal Markov assumption)



Object B is placed on the detector and the detector activates

Object B is removed. The detector stops activating

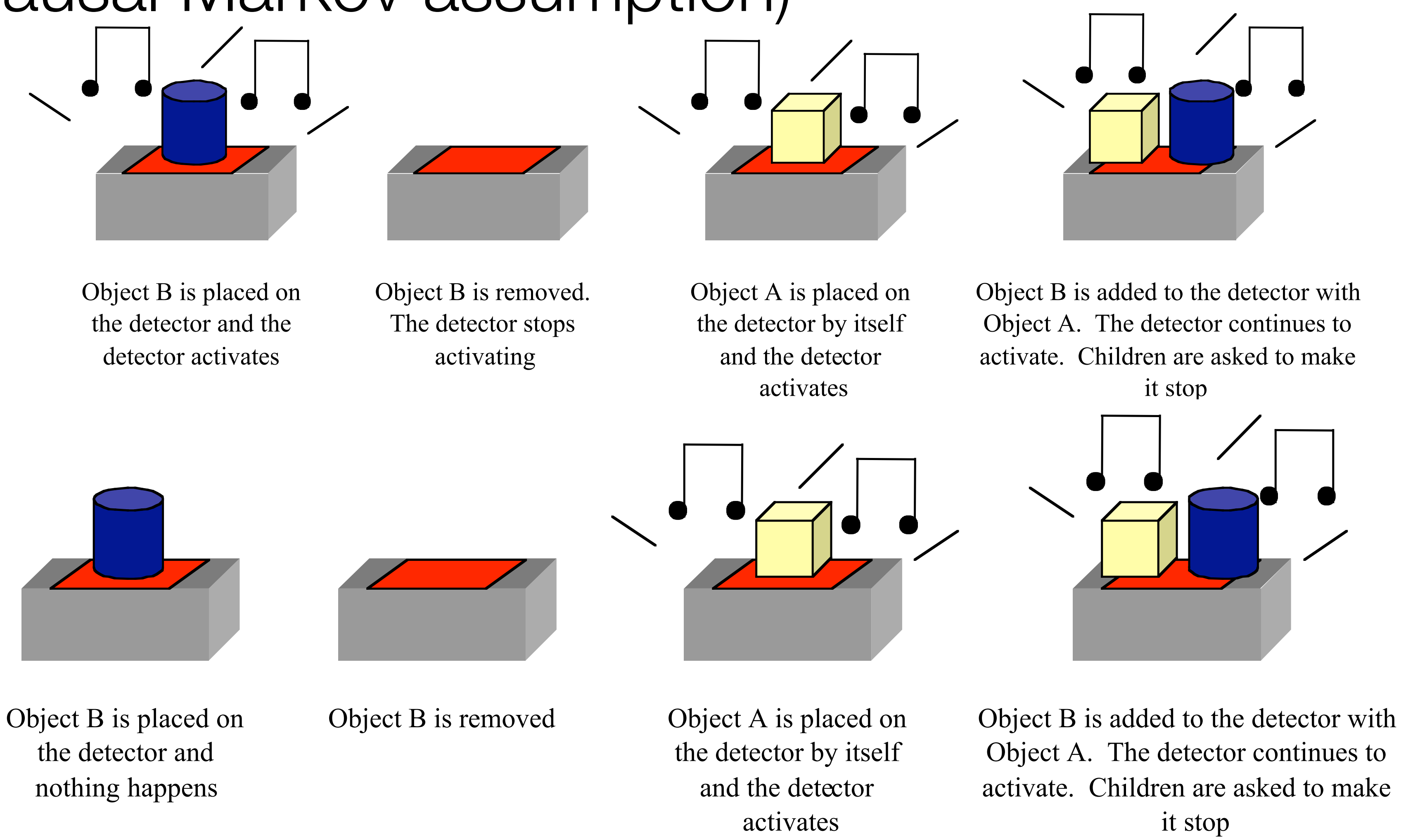
Object A is placed on the detector by itself and the detector activates

Object B is added to the detector with Object A. The detector continues to activate. Children are asked to make it stop

Application: Causal learning in infants

Gopnik et al (2004) Cog Sci

- Representation of causal structure is through graphical models (directed acyclic graphs, DAGs)
- Causal structure imply conditional independencies (causal Markov assumption)



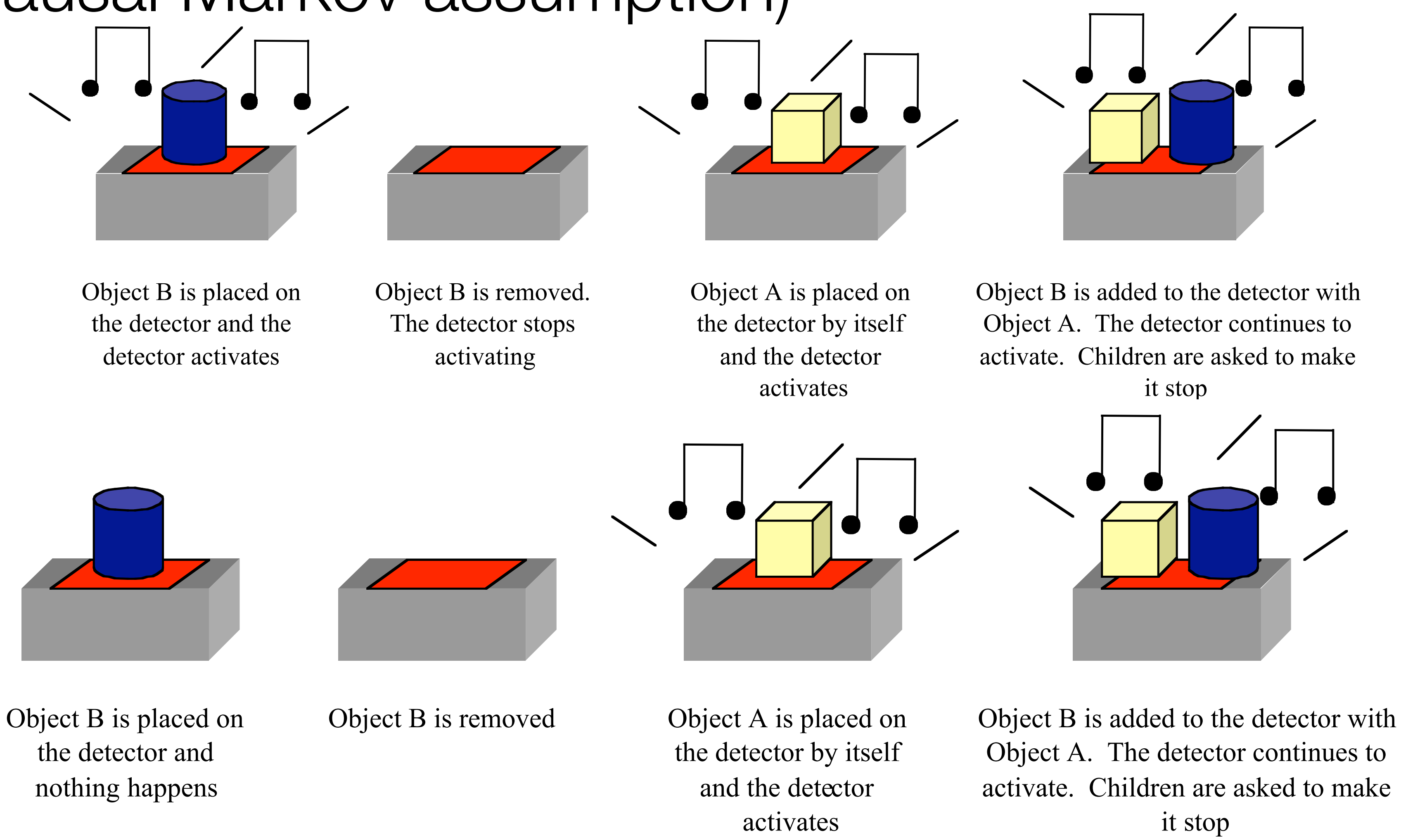
Application: Causal learning in infants

Gopnik et al (2004) Cog Sci

- Representation of causal structure is through graphical models (directed acyclic graphs, DAGs)
- Causal structure imply conditional independencies (causal Markov assumption)

hypothesis #1

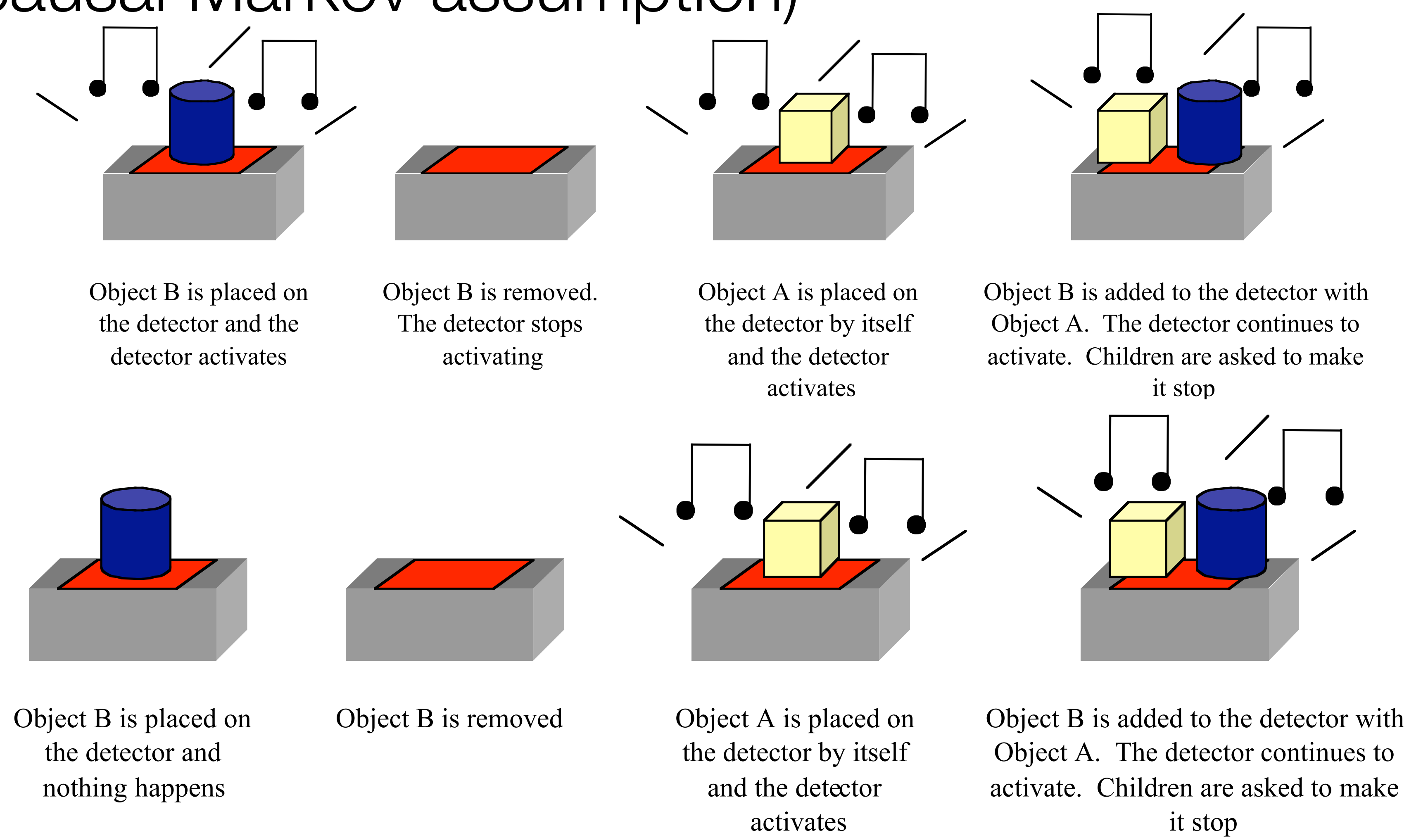
hypothesis #2

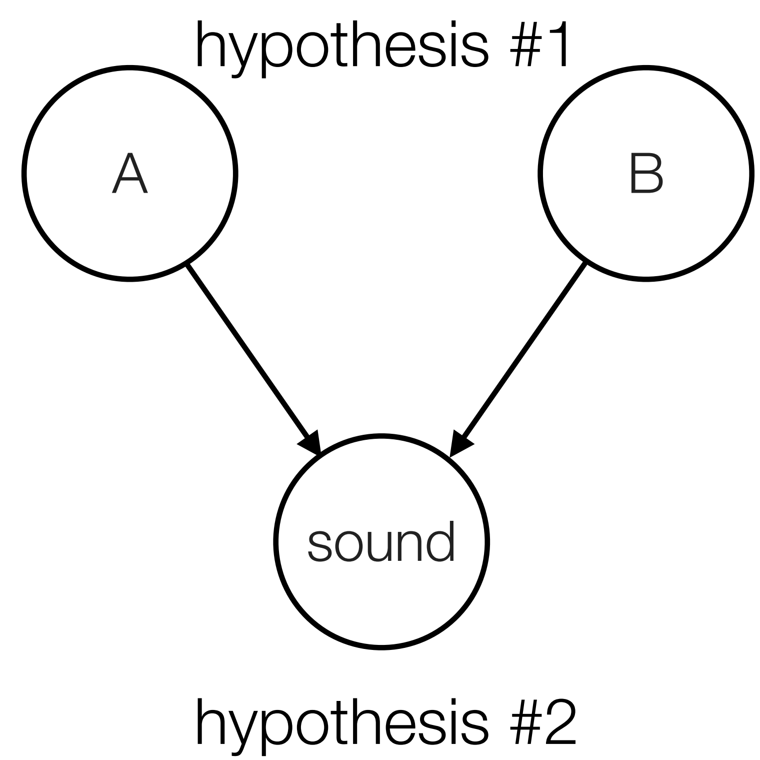


Application: Causal learning in infants

Gopnik et al (2004) Cog Sci

- Representation of causal structure is through graphical models (directed acyclic graphs, DAGs)
- Causal structure imply conditional independencies (causal Markov assumption)

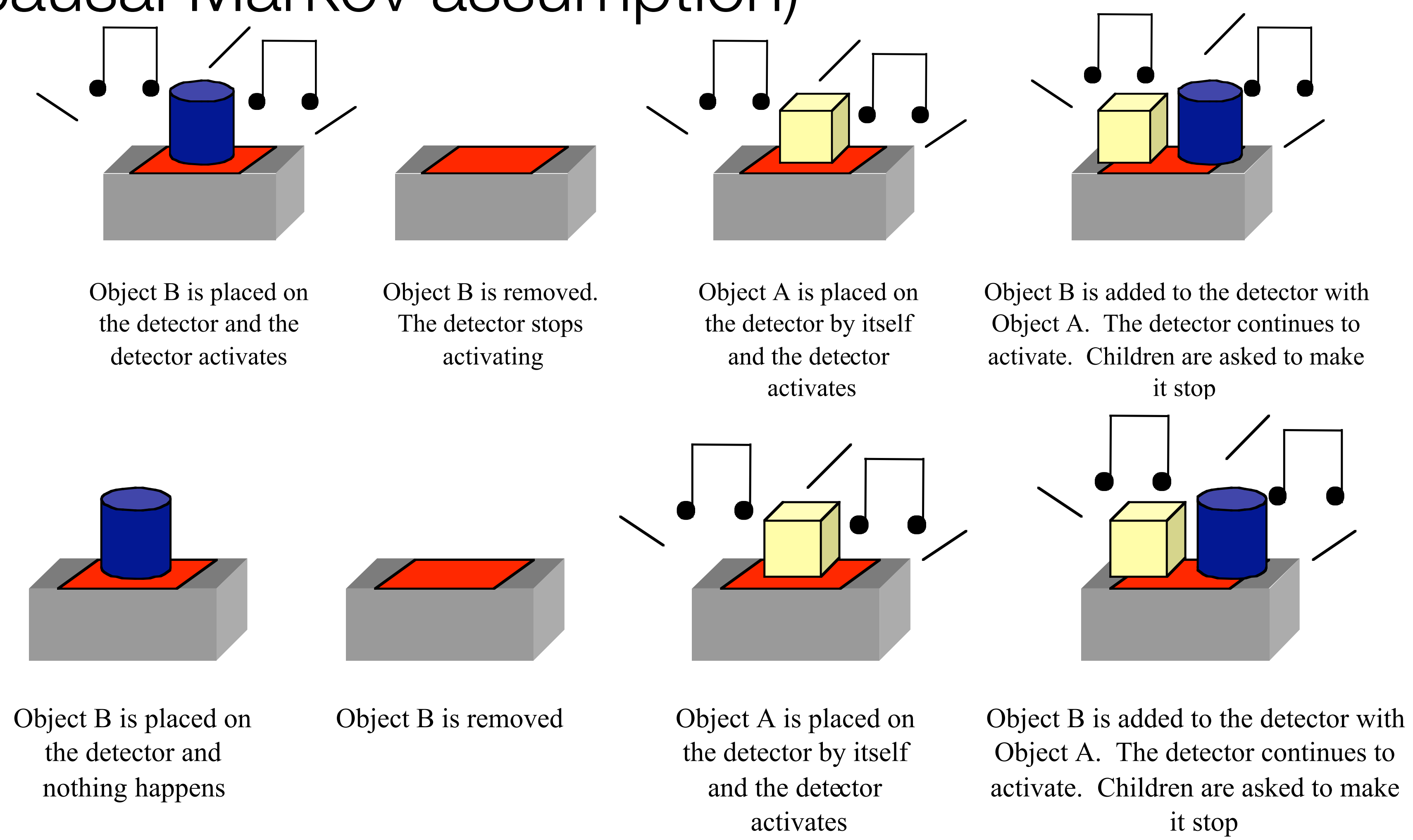


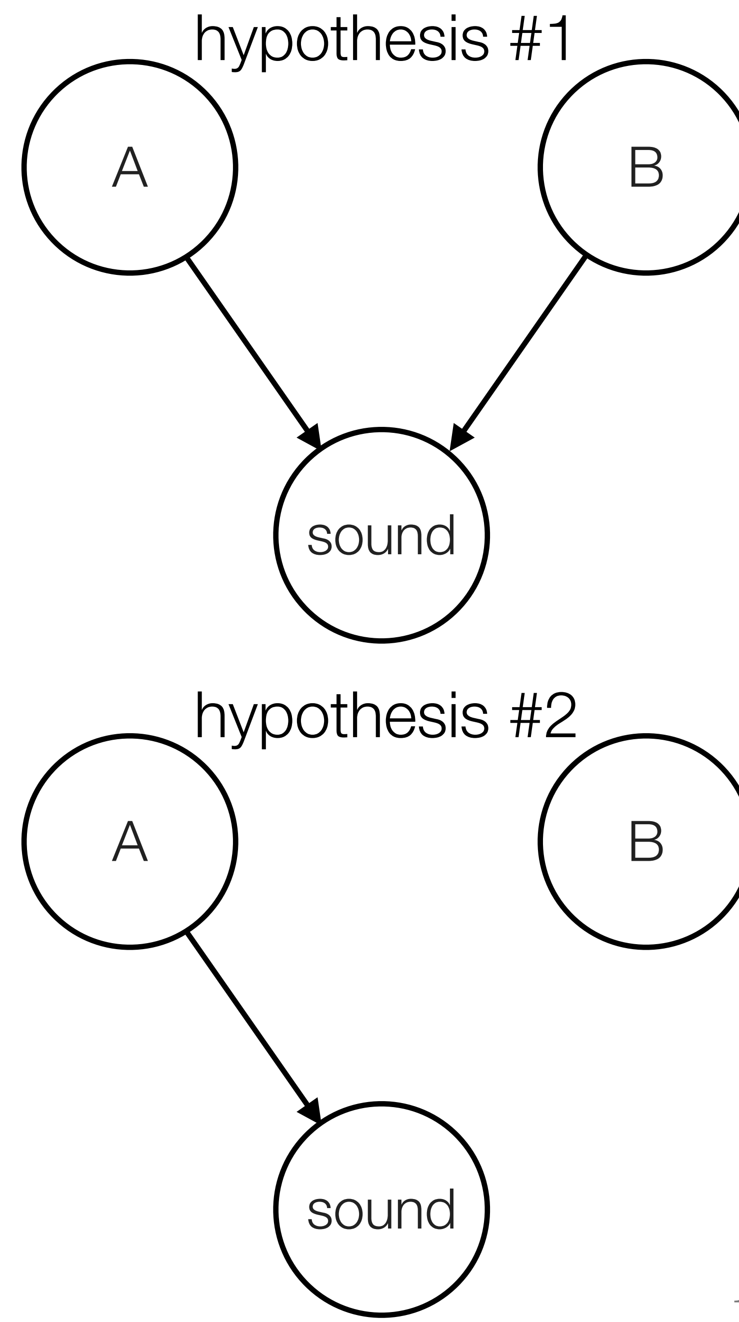


Application: Causal learning in infants

Gopnik et al (2004) Cog Sci

- Representation of causal structure is through graphical models (directed acyclic graphs, DAGs)
- Causal structure imply conditional independencies (causal Markov assumption)

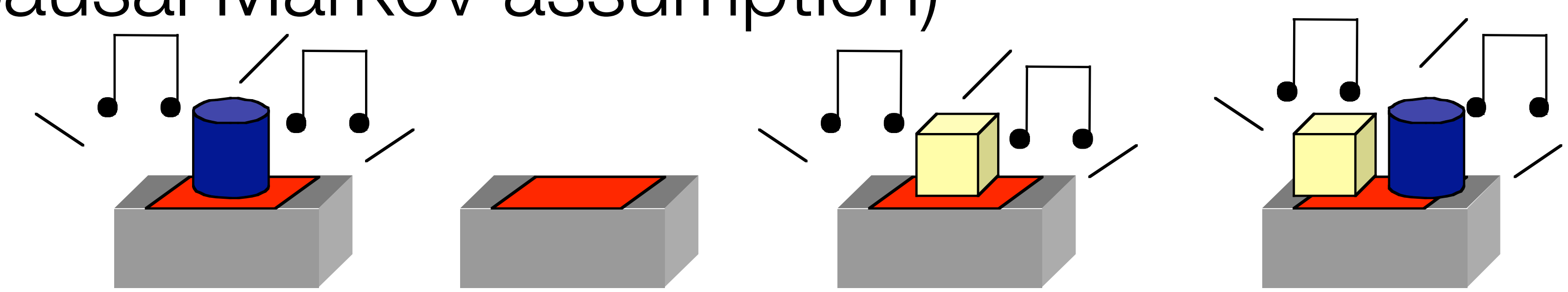




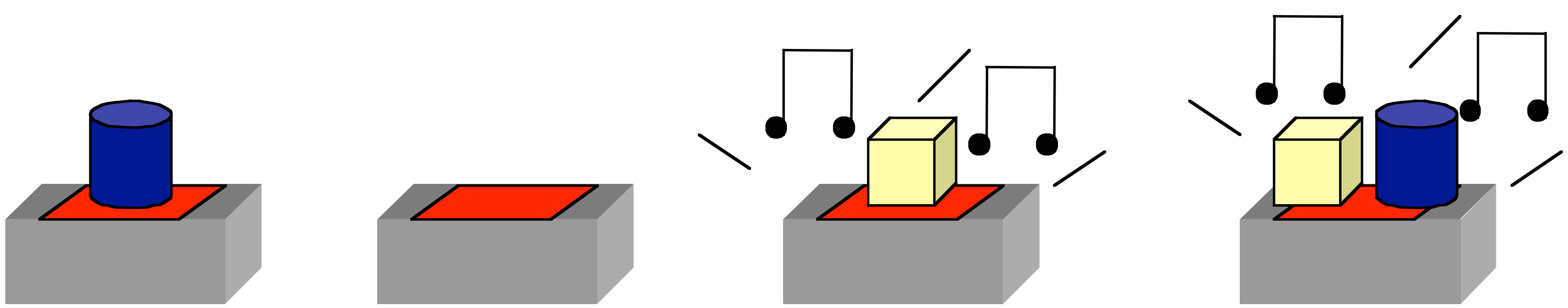
Application: Causal learning in infants

Gopnik et al (2004) Cog Sci

- Representation of causal structure is through graphical models (directed acyclic graphs, DAGs)
- Causal structure imply conditional independencies (causal Markov assumption)

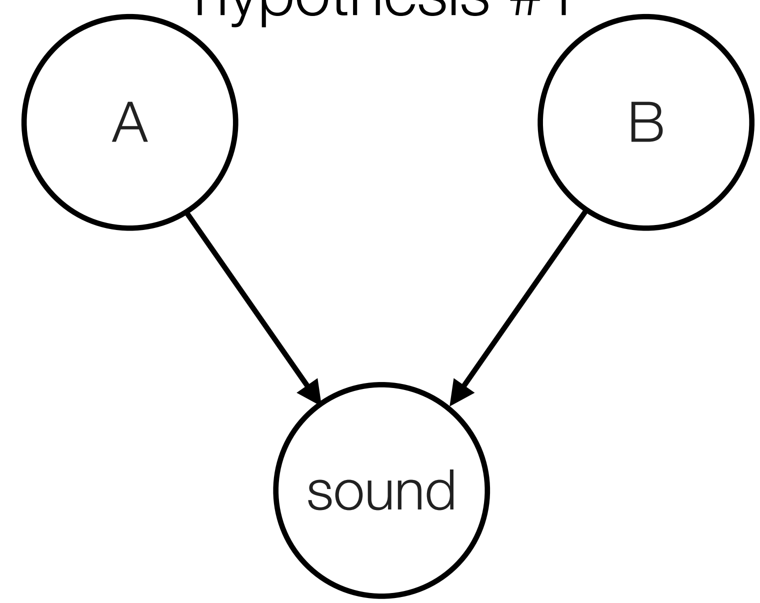


Object B is placed on the detector and the detector activates | Object B is removed. The detector stops activating | Object A is placed on the detector by itself and the detector activates | Object B is added to the detector with Object A. The detector continues to activate. Children are asked to make it stop

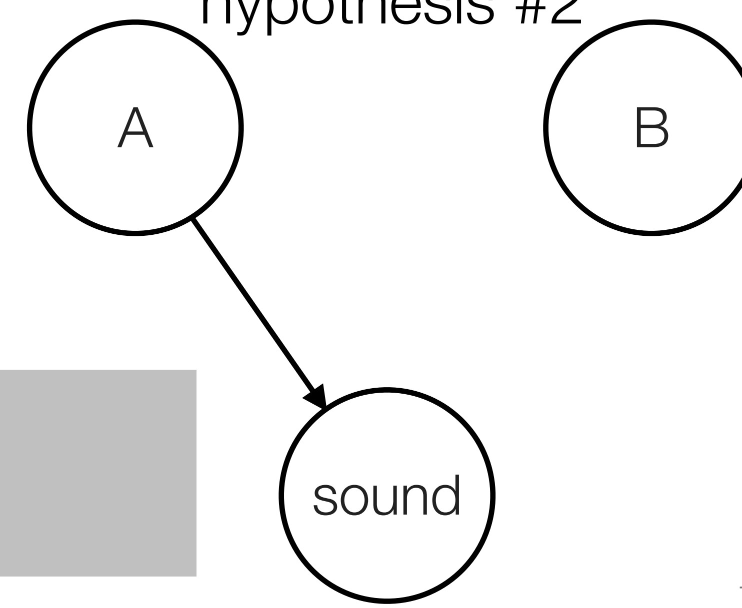


Object B is placed on the detector and nothing happens | Object B is removed | Object A is placed on the detector by itself and the detector activates | Object B is added to the detector with Object A. The detector continues to activate. Children are asked to make it stop

hypothesis #1



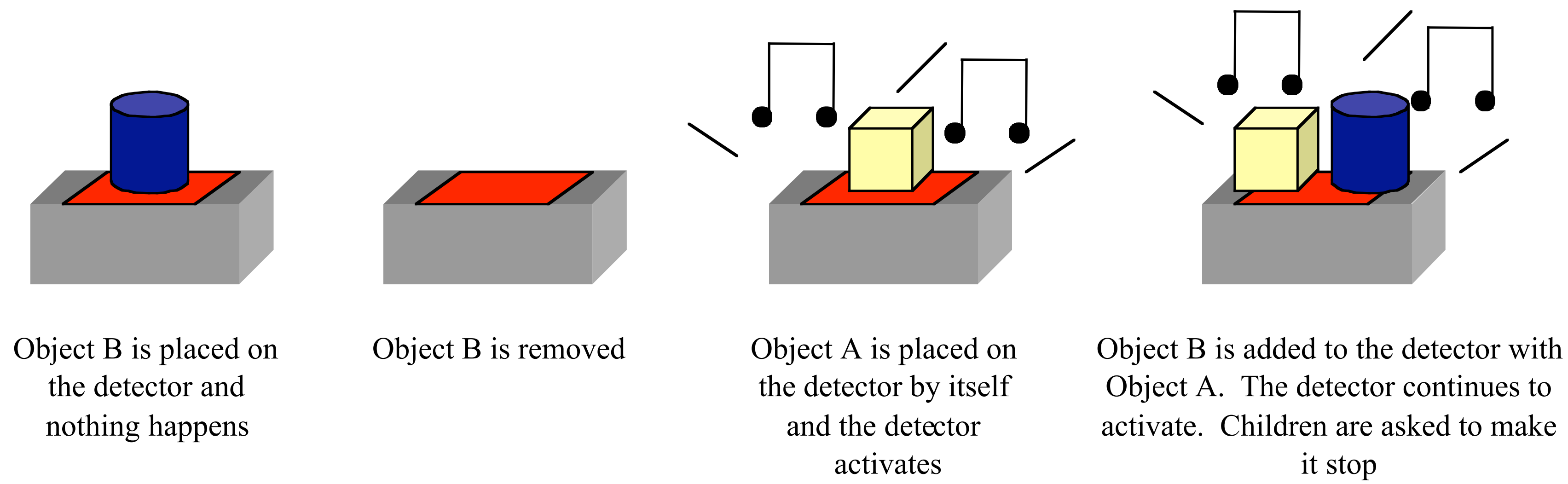
hypothesis #2



discovering causal structure → screening-off:
assessment of conditional probabilities

Application: Causal learning in infants

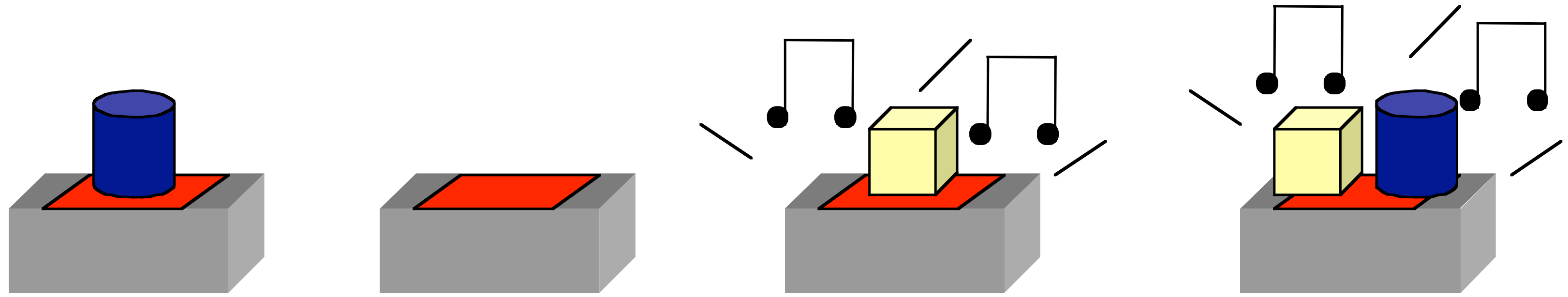
Gopnik et al (2004) Cog Sci



- A, B, S are correlated
- A & B are potential causes of S
- S is independent of B conditional on A
- S is not independent of A conditional on B
- A causes S and B does not

Application: Causal learning in infants

Gopnik et al (2004) Cog Sci



Object B is placed on the detector and nothing happens

Object B is removed

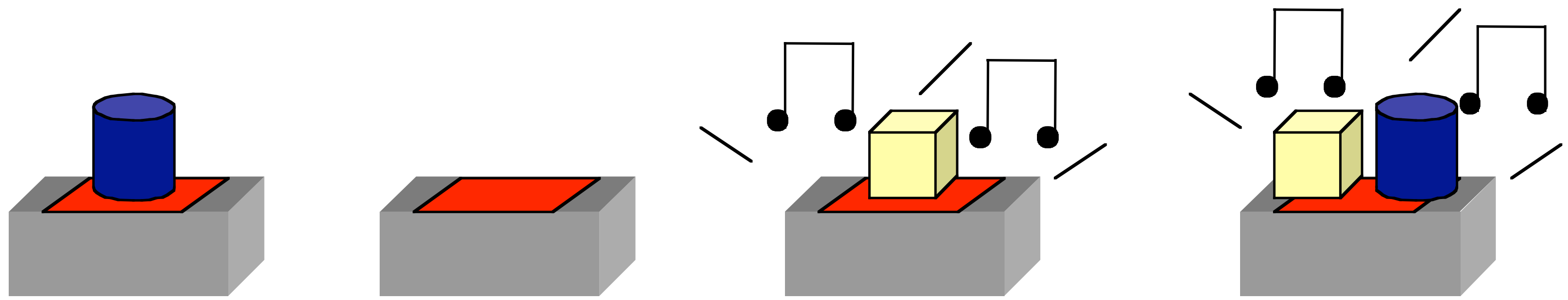
Object A is placed on the detector by itself and the detector activates

Object B is added to the detector with Object A. The detector continues to activate. Children are asked to make it stop

- A, B, S are correlated
- A & B are potential causes of S
- S is independent of B conditional on A
- S is not independent of A conditional on B
- A causes S and B does not

Application: Causal learning in infants

Gopnik et al (2004) Cog Sci



Object B is placed on the detector and nothing happens

Object B is removed

Object A is placed on the detector by itself and the detector activates

Object B is added to the detector with Object A. The detector continues to activate. Children are asked to make it stop

- A, B, S are correlated

associative accounts solely rely on this measurement (e.g. Rescola Wagner)

- A & B are potential causes of S
- S is independent of B conditional on A
- S is not independent of A conditional on B
- A causes S and B does not

Assignment

- Mini-esszé: végezzetek kutatást Shepard Universal Law of Generalization kapcsán: miről szól ez, mennyire univerzális, milyen következményei vannak; s: mi köze a Bayes-i infrenciához? (hint: Tenenbaum & Griffiths)

Recap: Bayes rule

measurement and **inference**

Recap: Bayes rule

measurement and **inference**

measurement:
measuring the probability of coughing when having a flu or not

Recap: Bayes rule

measurement and **inference**

measurement:
measuring the probability of coughing when having a flu or not

$$P(\text{coughing} = 1 \,|\, \text{flu} = 1)$$

Recap: Bayes rule

measurement and **inference**

measurement:
measuring the probability of coughing when having a flu or not

$$P(\text{coughing} = 1 \,|\, \text{flu} = 1)$$

inference:
infer the (posterior) probability of a hypothesis

Recap: Bayes rule

measurement and **inference**

measurement:
measuring the probability of coughing when having a flu or not

$$P(\text{coughing} = 1 \,|\, \text{flu} = 1)$$

inference:
infer the (posterior) probability of a hypothesis

$$P(\text{flu} = 1 \,|\, \text{coughing} = 1)$$

Recap: Bayes rule

measurement and **inference**

measurement:
measuring the probability of coughing when having a flu or not

$$P(\text{coughing} = 1 \,|\, \text{flu} = 1)$$

inference:
infer the (posterior) probability of a hypothesis

$$P(\text{flu} = 1 \,|\, \text{coughing} = 1)$$

what is the connection between the two quantities?

Recap: Bayes rule

measurement and **inference**

measurement:
measuring the probability of coughing when having a flu or not

$$P(\text{coughing} = 1 \,|\, \text{flu} = 1)$$

inference:
infer the (posterior) probability of a hypothesis

$$P(\text{flu} = 1 \,|\, \text{coughing} = 1)$$

what is the connection between the two quantities?

remember: **multiplication rule:** $P(x, y) = P(x \,|\, y)P(y)$

Recap: Bayes rule

measurement and **inference**

measurement:
measuring the probability of coughing when having a flu or not

$$P(\text{coughing} = 1 \,|\, \text{flu} = 1)$$

inference:
infer the (posterior) probability of a hypothesis

$$P(\text{flu} = 1 \,|\, \text{coughing} = 1)$$

what is the connection between the two quantities?

remember: **multiplication rule:** $P(x, y) = P(x \,|\, y)P(y)$

or equivalently:

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

Recap: Bayes rule

measurement and **inference**

measurement:
measuring the probability of coughing when having a flu or not

$$P(\text{coughing} = 1 \,|\, \text{flu} = 1)$$

inference:
infer the (posterior) probability of a hypothesis

$$P(\text{flu} = 1 \,|\, \text{coughing} = 1)$$

what is the connection between the two quantities?

remember: **multiplication rule:** $P(x, y) = P(x \,|\, y)P(y)$

or equivalently:

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

$$P(\text{flu} = 1 \,|\, \text{coughing} = 1) = \frac{P(\text{coughing} = 1 \,|\, \text{flu} = 1)P(\text{flu} = 1)}{P(\text{coughing} = 1)}$$

Recap: Bayes rule

measurement and **inference**

measurement:
measuring the probability of coughing when having a flu or not

$$P(\text{coughing} = 1 \,|\, \text{flu} = 1)$$

inference:
infer the (posterior) probability of a hypothesis

$$P(\text{flu} = 1 \,|\, \text{coughing} = 1)$$

what is the connection between the two quantities?

remember: **multiplication rule:** $P(x, y) = P(x \,|\, y)P(y)$

or equivalently:

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

$$P(\text{flu} = 1 \,|\, \text{coughing} = 1) = \frac{P(\text{coughing} = 1 \,|\, \text{flu} = 1)P(\text{flu} = 1)}{P(\text{coughing} = 1)}$$

Bayes rule: ‘inverts’ a probabilistic relationship

Recap: Bayes rule

measurement and **inference**

measurement:
measuring the probability of coughing when having a flu or not

$$P(\text{coughing} = 1 \,|\, \text{flu} = 1)$$

inference:
infer the (posterior) probability of a hypothesis

$$P(\text{flu} = 1 \,|\, \text{coughing} = 1)$$

what is the connection between the two quantities?

remember: **multiplication rule:** $P(x, y) = P(x \,|\, y)P(y)$

or equivalently:

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

$$P(\text{flu} = 1 \,|\, \text{coughing} = 1) = \frac{P(\text{coughing} = 1 \,|\, \text{flu} = 1)P(\text{flu} = 1)}{P(\text{coughing} = 1)}$$

Bayes rule: ‘inverts’ a probabilistic relationship

if the inferred variable is continuous, then the posterior assigns probabilities to all possible hypotheses

Coin tossing: an example

Head: 0
Tail: 1

Result

Coin tossing: an example

Head:	0
Tail:	1

Result **0**

Coin tossing: an example

Head:	0
Tail:	1

Result **0 0**

Coin tossing: an example

Head: 0
Tail: 1

Result 0 0 0

Coin tossing: an example

Head: 0
Tail: 1

Result 0 0 0 1

Coin tossing: an example

Head:	0
Tail:	1

Result 0 0 0 I I

Coin tossing: an example

Head: 0
Tail: 1

Result 0 0 0 1 1 0

Coin tossing: an example

Head:	0
Tail:	1

Result 0 0 0 1 1 0 1

Coin tossing: an example

Head:	0
Tail:	1

Result 0 0 0 1 1 0 1 ...

Coin tossing: an example

Head: 0 Tail: 1	

Result 0 0 0 1 1 0 1 ...

Estimated bias

Coin tossing: an example

Head:	0
Tail:	1

Result 0 0 0 1 1 0 1 ...

Estimated bias

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	I	I	0	I	...
Estimated bias	0							

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

Coin tossing: an example

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0						

Head: 0
Tail: 1

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	…
Estimated bias	0	0	0					

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25				

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4			

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33		

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias								

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias								

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0						

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\,(\vartheta) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle \right)^2$$

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0					

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25				

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3			

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26		

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

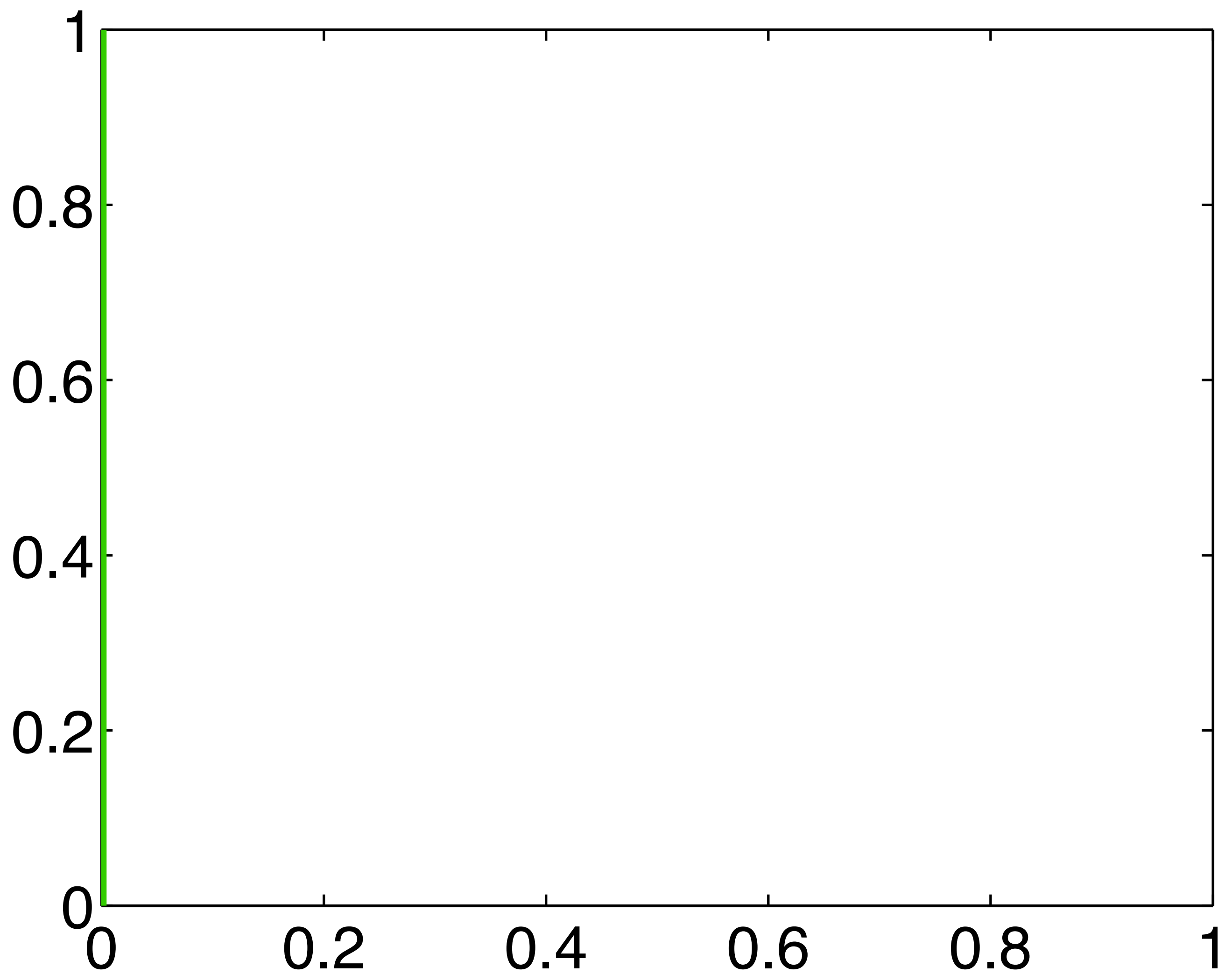
Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle\vartheta\rangle = \frac{1}{N}\sum_i x_i$$

$$\mathrm{Var}\,(\vartheta) = \frac{1}{N-1}\sum_i (x_i - \langle\vartheta\rangle)^2$$



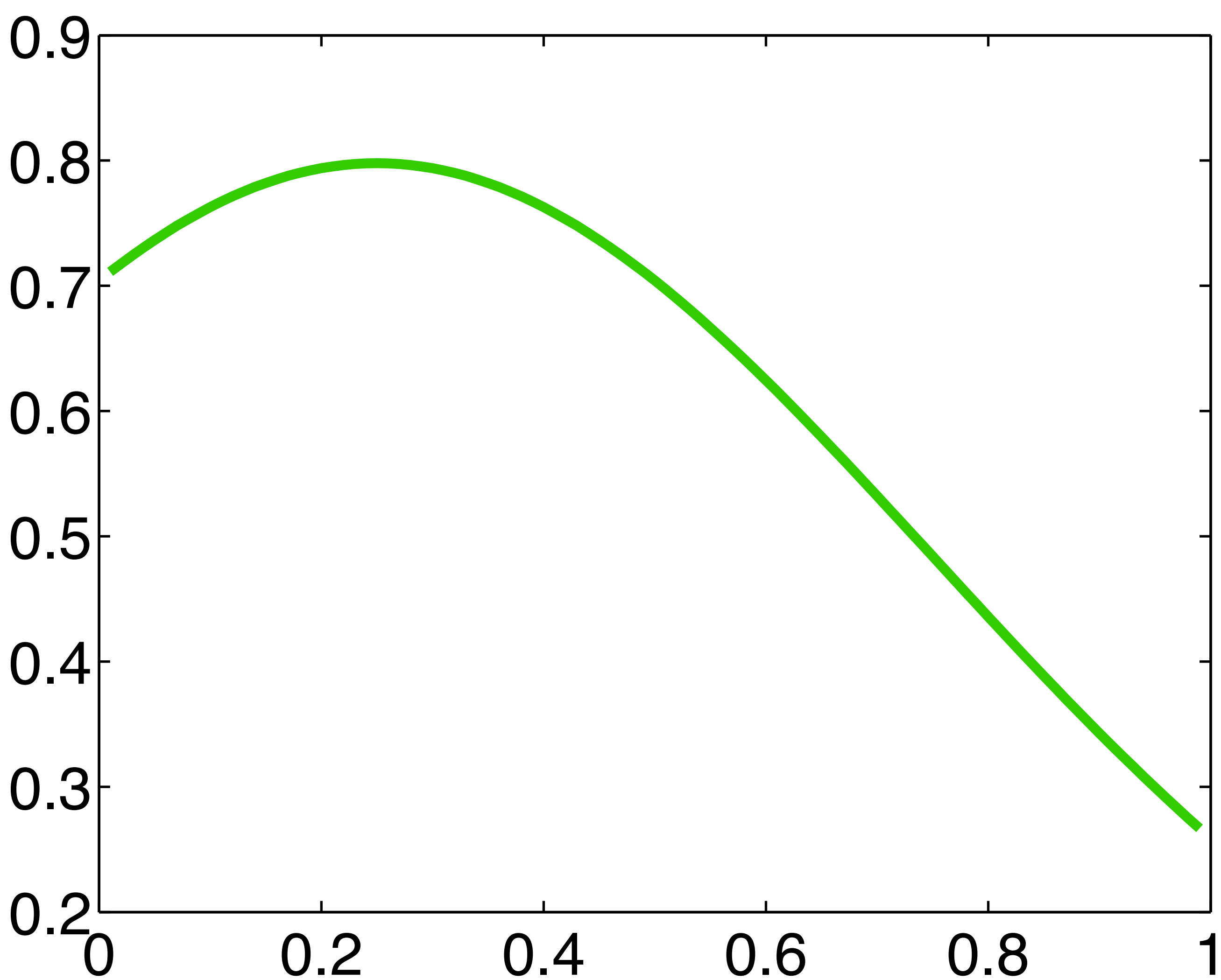
Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle\vartheta\rangle = \frac{1}{N}\sum_i x_i$$

$$\mathrm{Var}\,(\vartheta) = \frac{1}{N-1}\sum_i \left(x_i - \langle\vartheta\rangle\right)^2$$



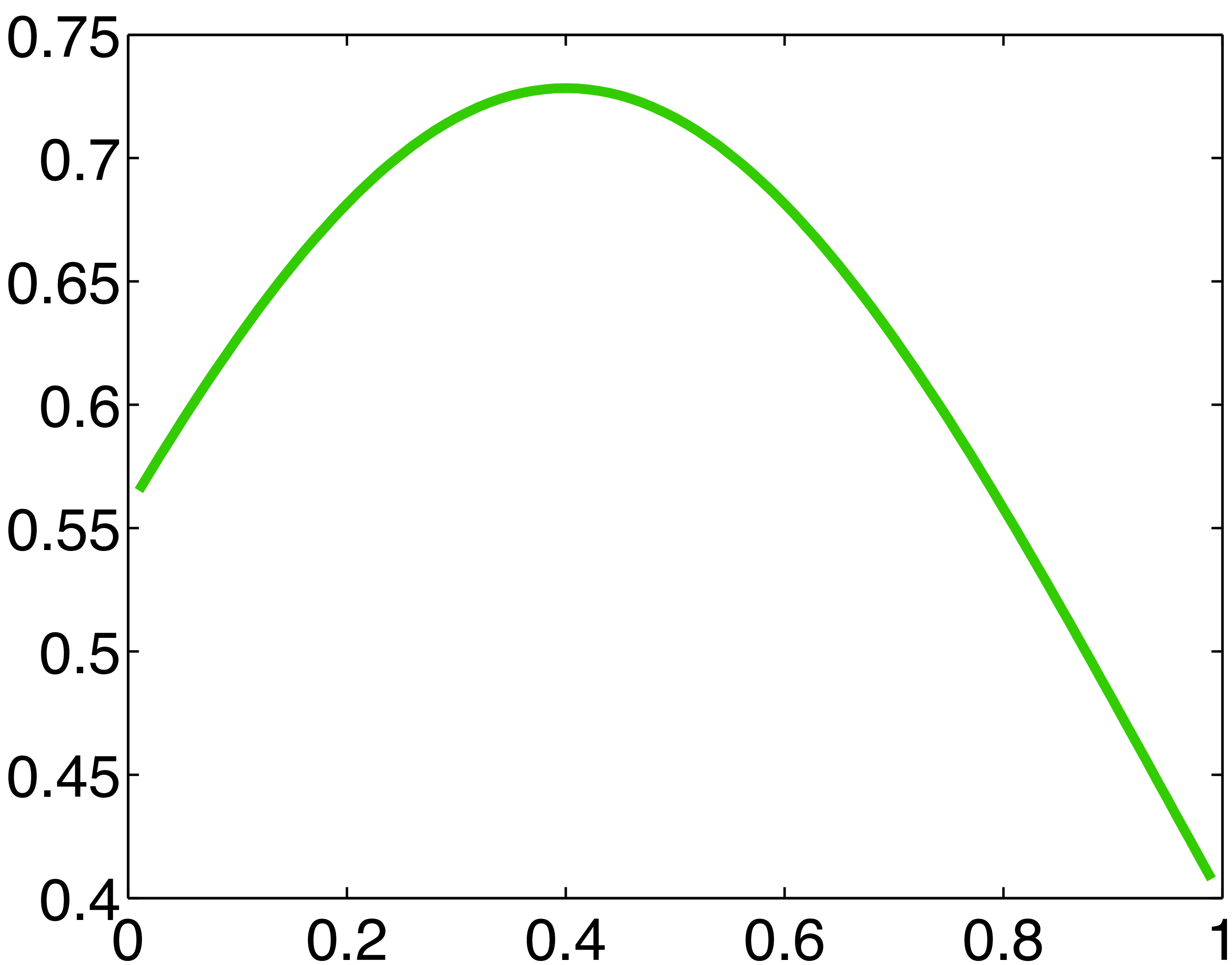
Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\,(\vartheta) = \frac{1}{N-1} \sum_i (x_i - \langle \vartheta \rangle)^2$$



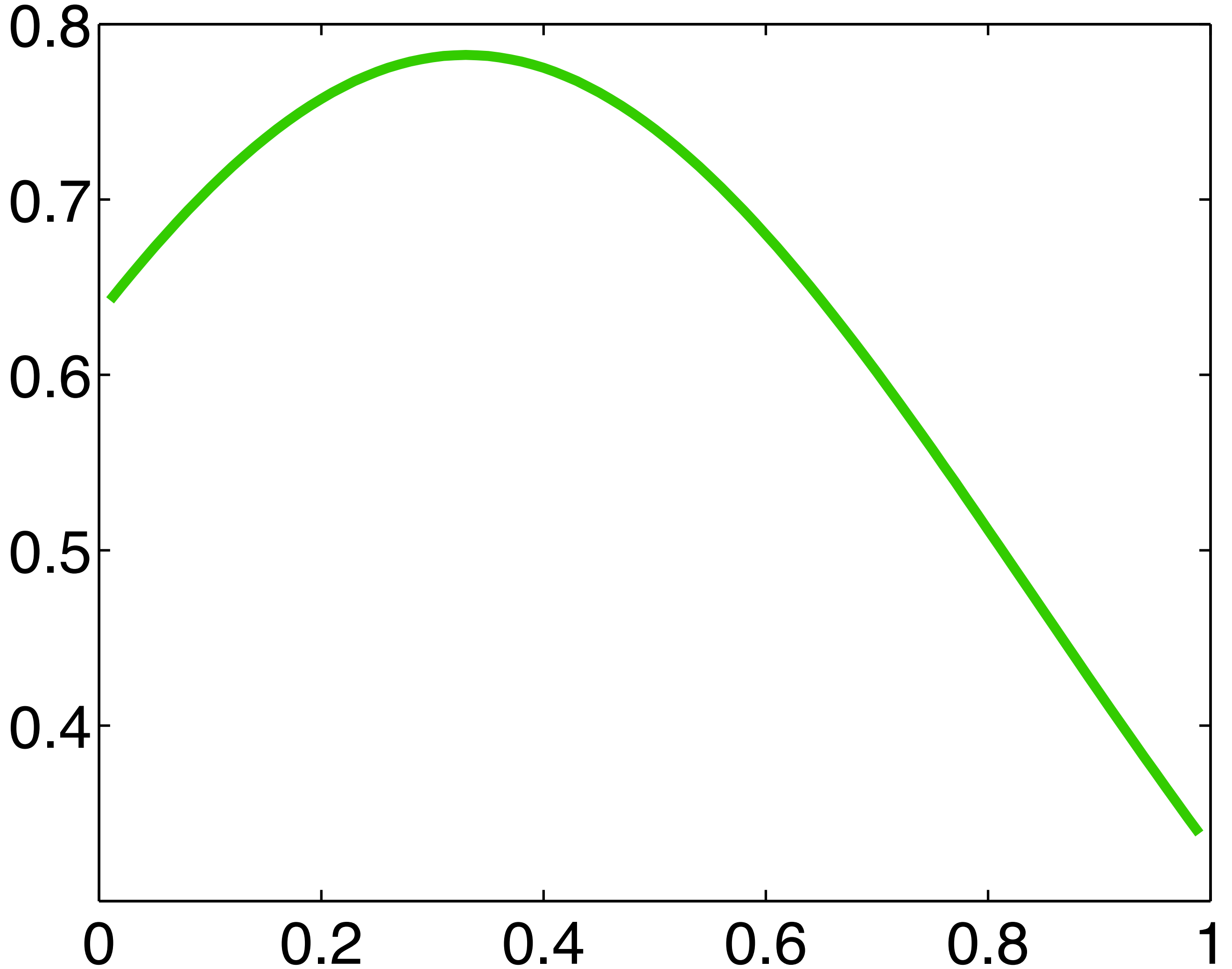
Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$



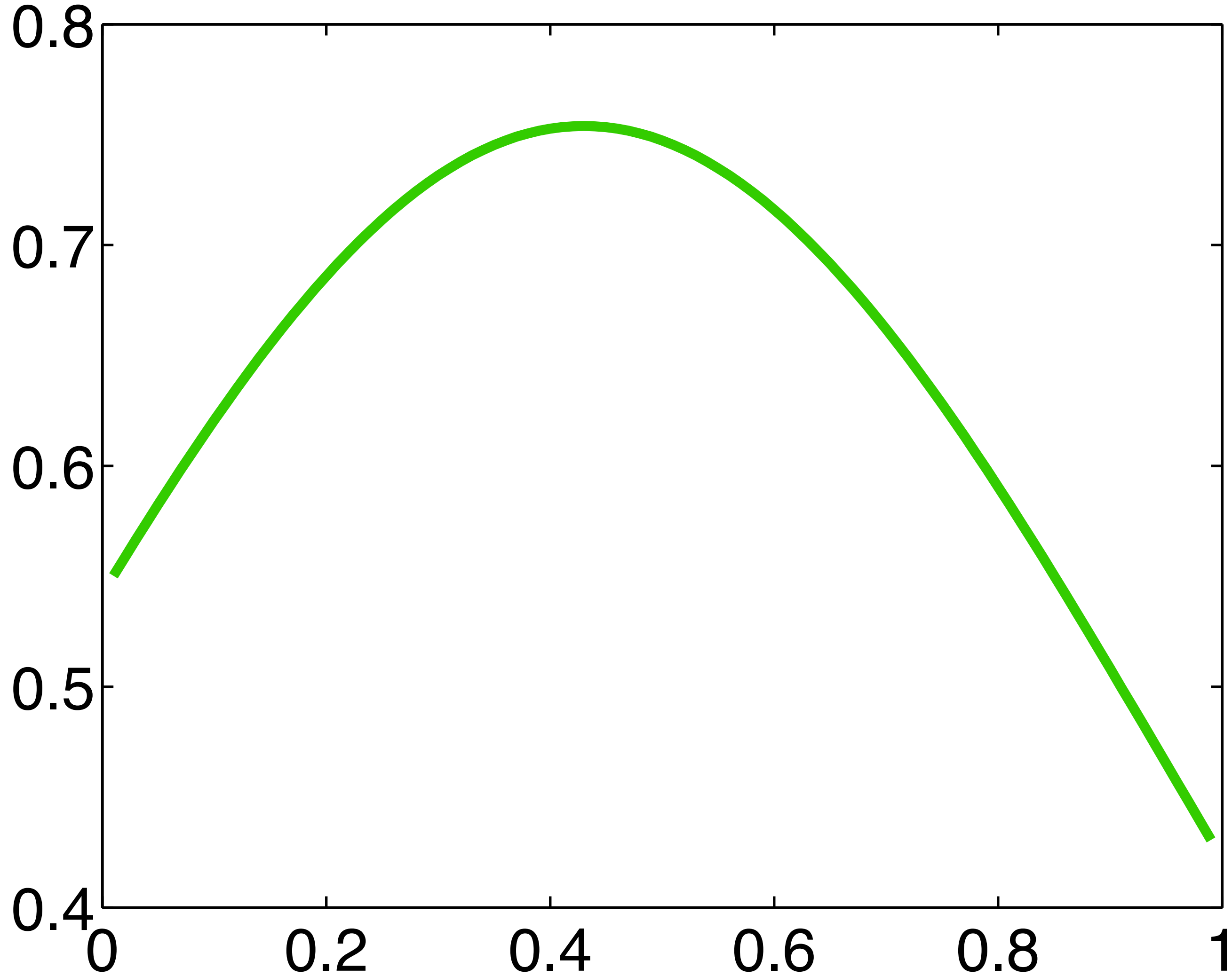
Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1 ...
Estimated bias	0	0	0	.25	.4	.33	.43
Variance of bias		0	0	.25	.3	.26	.28

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}(\vartheta) = \frac{1}{N-1} \sum_i (x_i - \langle \vartheta \rangle)^2$$



Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}(\vartheta) = \frac{1}{N-1} \sum_i (x_i - \langle \vartheta \rangle)^2$$

Trial likelihood $P(x \mid \vartheta) = \mathrm{Bernoulli}(x; \vartheta)$

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\,(\vartheta) = \frac{1}{N-1} \sum_i (x_i - \langle \vartheta \rangle)^2$$

Trial likelihood $P(x \,|\, \vartheta) = \mathrm{Bernoulli}\,(x; \vartheta) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}(\vartheta) = \frac{1}{N-1} \sum_i (x_i - \langle \vartheta \rangle)^2$$

Trial likelihood $P(x \mid \vartheta) = \mathrm{Bernoulli}(x; \vartheta) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

Trial likelihood $P(x \mid \vartheta) = \mathrm{Bernoulli}\left(x; \vartheta\right) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P\left(\mathrm{data}|\vartheta\right) = \prod_t P\left(x_t|\vartheta\right)$

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

Trial likelihood $P(x \mid \vartheta) = \mathrm{Bernoulli}\left(x; \vartheta\right) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P\left(\mathrm{data} | \vartheta\right) = \prod_t P\left(x_t | \vartheta\right)$

Is this the important quantity?

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

Trial likelihood $$P(x \mid \vartheta) = \mathrm{Bernoulli}\left(x; \vartheta\right) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$$

Data likelihood $$P\left(\mathrm{data}|\vartheta\right) = \prod_t P\left(x_t|\vartheta\right)$$

Is this the important quantity? $P\left(\vartheta|\mathrm{data}\right)$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}(\vartheta) = \frac{1}{N-1} \sum_i (x_i - \langle \vartheta \rangle)^2$$

Trial likelihood $P(x \mid \vartheta) = \mathrm{Bernoulli}(x; \vartheta) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P(\mathrm{data} | \vartheta) = \prod_t P(x_t | \vartheta)$

Is this the important quantity? $P(\vartheta | \mathrm{data}) = P(\mathrm{data} | \vartheta) P(\vartheta) / P(\mathrm{data})$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}(\vartheta) = \frac{1}{N-1} \sum_i (x_i - \langle \vartheta \rangle)^2$$

Trial likelihood $P(x \mid \vartheta) = \mathrm{Bernoulli}(x; \vartheta) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P(\mathrm{data}|\vartheta) = \prod_t P(x_t|\vartheta)$

Bayes rule

Is this the important quantity?

$$P(\vartheta|\mathrm{data}) = P(\mathrm{data}|\vartheta)\, P(\vartheta) \,/\, P(\mathrm{data})$$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}(\vartheta) = \frac{1}{N-1} \sum_i (x_i - \langle \vartheta \rangle)^2$$

Trial likelihood $P(x \mid \vartheta) = \mathrm{Bernoulli}(x; \vartheta) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P(\mathrm{data}|\vartheta) = \prod_t P(x_t|\vartheta)$

Is this the important quantity?

Bayes rule

$$P(\vartheta|\mathrm{data}) = P(\mathrm{data}|\vartheta)\, P(\vartheta) / P(\mathrm{data})$$

Posterior

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}(\vartheta) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle \right)^2$$

Trial likelihood $P(x \mid \vartheta) = \mathrm{Bernoulli}(x; \vartheta) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P(\mathrm{data} | \vartheta) = \prod_t P(x_t | \vartheta)$

Bayes rule

Is this the important quantity? $P(\vartheta | \mathrm{data}) = P(\mathrm{data} | \vartheta)\, P(\vartheta) / P(\mathrm{data})$

Posterior

Likelihood

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

Trial likelihood $P(x \,|\, \vartheta) = \mathrm{Bernoulli}\left(x; \vartheta\right) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P\left(\mathrm{data}|\vartheta\right) = \prod_t P\left(x_t|\vartheta\right)$

Bayes rule

Is this the important quantity?

$$P\left(\vartheta|\mathrm{data}\right) = P\left(\mathrm{data}|\vartheta\right) P\left(\vartheta\right) / P\left(\mathrm{data}\right)$$

Posterior

Likelihood

Prior

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1 ...
Estimated bias	0	0	0	.25	.4	.33	.43
Variance of bias		0	0	.25	.3	.26	.28

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

Trial likelihood $P(x \mid \vartheta) = \mathrm{Bernoulli}\left(x; \vartheta\right) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P\left(\mathrm{data}|\vartheta\right) = \prod_t P\left(x_t|\vartheta\right)$

Is this the important quantity?

Bayes rule

$$P\left(\vartheta|\mathrm{data}\right) = P\left(\mathrm{data}|\vartheta\right) P\left(\vartheta\right) / P\left(\mathrm{data}\right)$$

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}(\vartheta) = \frac{1}{N-1} \sum_i (x_i - \langle \vartheta \rangle)^2$$

Trial likelihood $P(x \mid \vartheta) = \mathrm{Bernoulli}(x; \vartheta) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P(\mathrm{data}|\vartheta) = \prod_t P(x_t|\vartheta)$

Bayes rule

Is this the important quantity?

$$P(\vartheta|\mathrm{data}) = P(\mathrm{data}|\vartheta)\, P(\vartheta) \,/\, P(\mathrm{data})$$

$$P(\vartheta) = \mathrm{Beta}(\vartheta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \vartheta^{\alpha-1} (1-\vartheta)^{\beta-1}$$

Coin tossing: an example

Head: 0 Tail: 1								
Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}(\vartheta) = \frac{1}{N-1} \sum_i (x_i - \langle \vartheta \rangle)^2$$

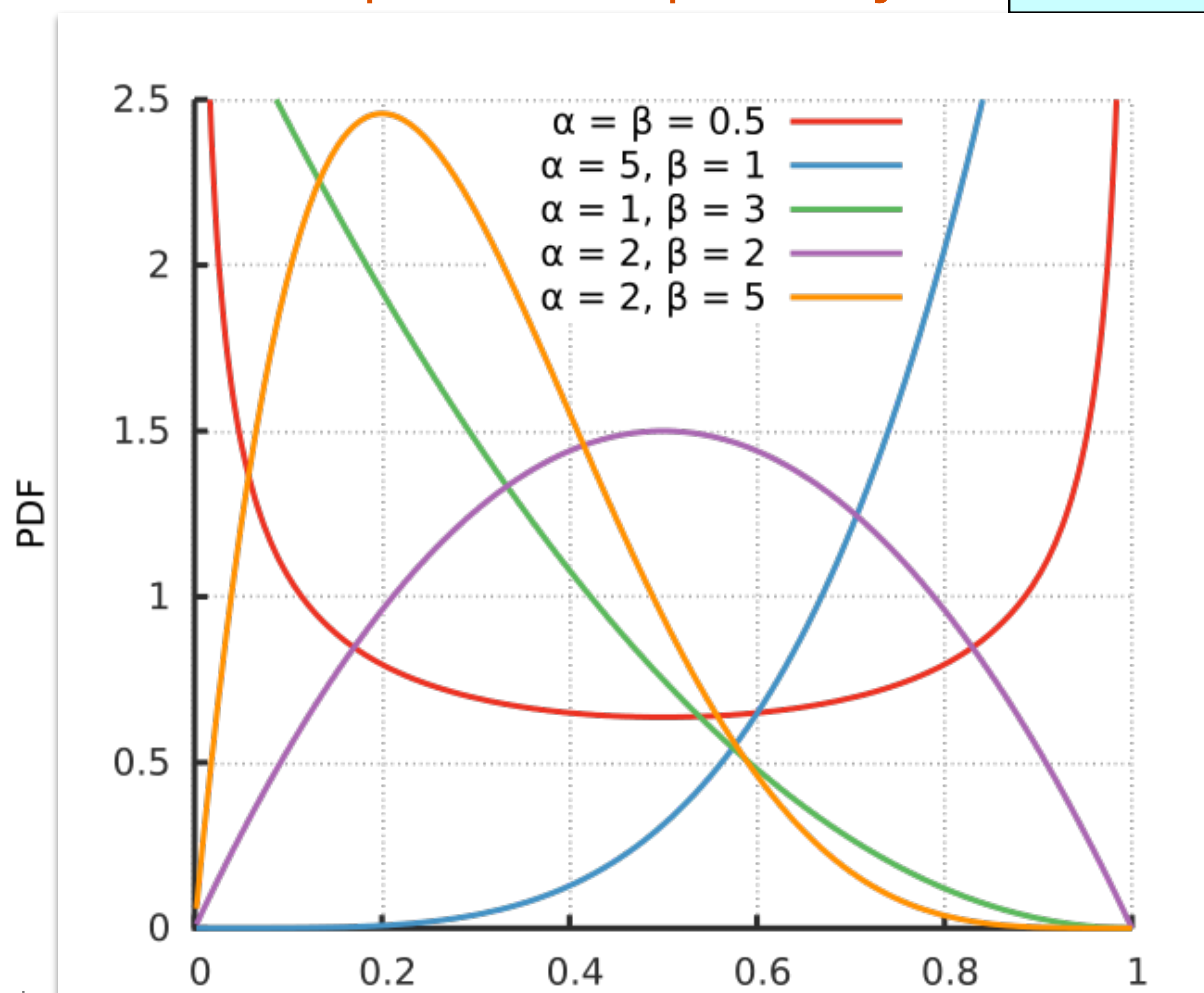
Trial likelihood $P(x \mid \vartheta) = \mathrm{Bernoulli}(x; \vartheta) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P(\mathrm{data}|\vartheta) = \prod_t P(x_t|\vartheta)$

Bayes rule

$$P(\vartheta|\mathrm{data}) = P(\mathrm{data}|\vartheta)\, P(\vartheta) / P(\mathrm{data})$$

Is this the important quantity?



$$P(\vartheta) = \mathrm{Beta}(\vartheta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \vartheta^{\alpha-1} (1-\vartheta)^{\beta-1}$$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

Trial likelihood $P(x \mid \vartheta) = \mathrm{Bernoulli}\left(x; \vartheta\right) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P\left(\mathrm{data}|\vartheta\right) = \prod_t P\left(x_t|\vartheta\right)$

Is this the important quantity?

Bayes rule

$$P\left(\vartheta|\mathrm{data}\right) = P\left(\mathrm{data}|\vartheta\right) P\left(\vartheta\right) / P\left(\mathrm{data}\right)$$

$$P(\vartheta) = \mathrm{Beta}(\vartheta; \alpha, \beta) = \frac{1}{B\left(\alpha, \beta\right)} \vartheta^{\alpha-1} \left(1-\vartheta\right)^{\beta-1}$$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1 ...
Estimated bias	0	0	0	.25	.4	.33	.43
Variance of bias		0	0	.25	.3	.26	.28

$$\langle\vartheta\rangle = \frac{1}{N}\sum_i x_i$$

$$\mathrm{Var}\,(\vartheta) = \frac{1}{N-1}\sum_i \left(x_i - \langle\vartheta\rangle\right)^2$$

Trial likelihood $P(x\,|\,\vartheta) = \mathrm{Bernoulli}\,(x;\vartheta) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P\,(\mathrm{data}|\vartheta) = \prod_t P\,(x_t|\vartheta)$

Bayes rule

Is this the important quantity?

$$P\,(\vartheta|\mathrm{data}) = P\,(\mathrm{data}|\vartheta)\,P\,(\vartheta)\,/P\,(\mathrm{data})$$

$$P(\vartheta) = \mathrm{Beta}(\vartheta;\alpha,\beta) = \frac{1}{B\,(\alpha,\beta)}\vartheta^{\alpha-1}\,(1-\vartheta)^{\beta-1}$$

$$P(\vartheta\,|\,\mathrm{data}) = \mathrm{Beta}\left(\vartheta;\alpha^{(t)},\beta^{(t)}\right)$$

$$\alpha^{(t)} = \alpha + \sum_{i=1}^{t} x_i$$

$$\beta^{(t)} = \beta + t - \sum_{i=1}^{t} x_i$$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}(\vartheta) = \frac{1}{N-1} \sum_i (x_i - \langle \vartheta \rangle)^2$$

Trial likelihoo[...] $(x; \vartheta) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likeliho[...] $P(x_t | \vartheta)$

conjugate prior

Bayes rule

Is this the important quantity?

$$P(\vartheta | \mathrm{data}) = P(\mathrm{data} | \vartheta)\, P(\vartheta) / P(\mathrm{data})$$

$$P(\vartheta) = \mathrm{Beta}(\vartheta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \vartheta^{\alpha-1} (1-\vartheta)^{\beta-1}$$

$$P(\vartheta \,|\, \mathrm{data}) = \mathrm{Beta}\left(\vartheta; \alpha^{(t)}, \beta^{(t)}\right)$$

$$\alpha^{(t)} = \alpha + \sum_{i=1}^{t} x_i$$

$$\beta^{(t)} = \beta + t - \sum_{i=1}^{t} x_i$$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\,(\vartheta) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

Trial likelihood $P(x \mid \vartheta) = \mathrm{Bernoulli}\,(x; \vartheta) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P\,(\mathrm{data}|\vartheta) = \prod_t P\,(x_t|\vartheta)$

Is this the important quantity?

Bayes rule

$$P\,(\vartheta|\mathrm{data}) = P\,(\mathrm{data}|\vartheta)\, P\,(\vartheta) \,/\, P\,(\mathrm{data})$$

$$P(\vartheta) = \mathrm{Beta}(\vartheta; \alpha, \beta) = \frac{1}{B\,(\alpha, \beta)} \vartheta^{\alpha-1} \,(1-\vartheta)^{\beta-1}$$

$$P(\vartheta \mid \mathrm{data}) = \mathrm{Beta}\left(\vartheta; \alpha^{(t)}, \beta^{(t)}\right)$$

$$\alpha^{(t)} = \alpha + \sum_{i=1}^{t} x_i$$

$$\beta^{(t)} = \beta + t - \sum_{i=1}^{t} x_i$$

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}(\vartheta) = \frac{1}{N-1} \sum_i (x_i - \langle \vartheta \rangle)^2$$

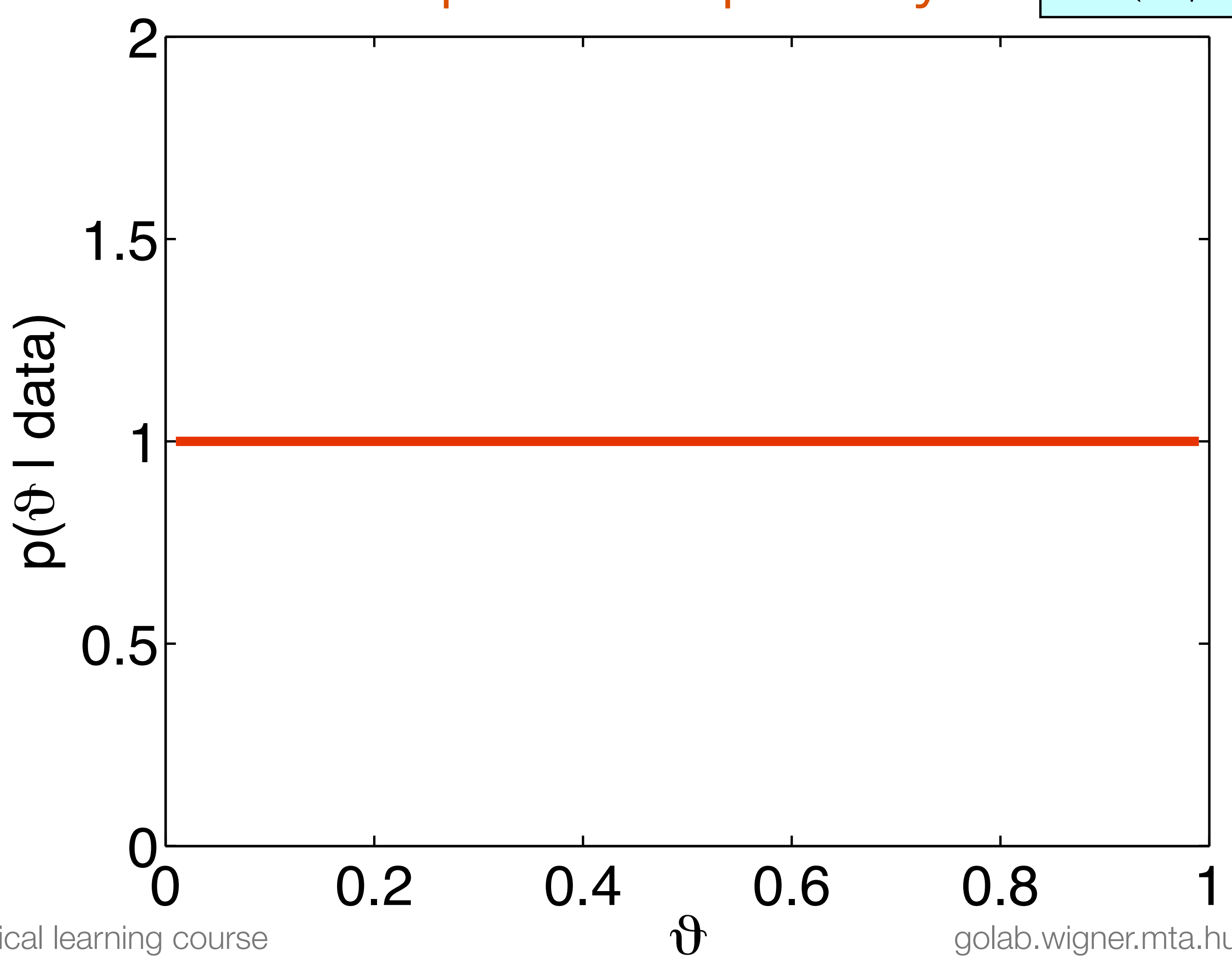
Trial likelihood $P(x \,|\, \vartheta) = \mathrm{Bernoulli}(x; \vartheta) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P(\mathrm{data}|\vartheta) = \prod_t P(x_t|\vartheta)$

Bayes rule

$$P(\vartheta|\mathrm{data}) = P(\mathrm{data}|\vartheta)\, P(\vartheta) \,/\, P(\mathrm{data})$$

Is this the important quantity?



$$P(\vartheta) = \mathrm{Beta}(\vartheta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \vartheta^{\alpha-1} (1-\vartheta)^{\beta-1}$$

$$P(\vartheta \,|\, \mathrm{data}) = \mathrm{Beta}\left(\vartheta; \alpha^{(t)}, \beta^{(t)}\right)$$

$$\alpha^{(t)} = \alpha + \sum_{i=1}^{t} x_i$$

$$\beta^{(t)} = \beta + t - \sum_{i=1}^{t} x_i$$

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

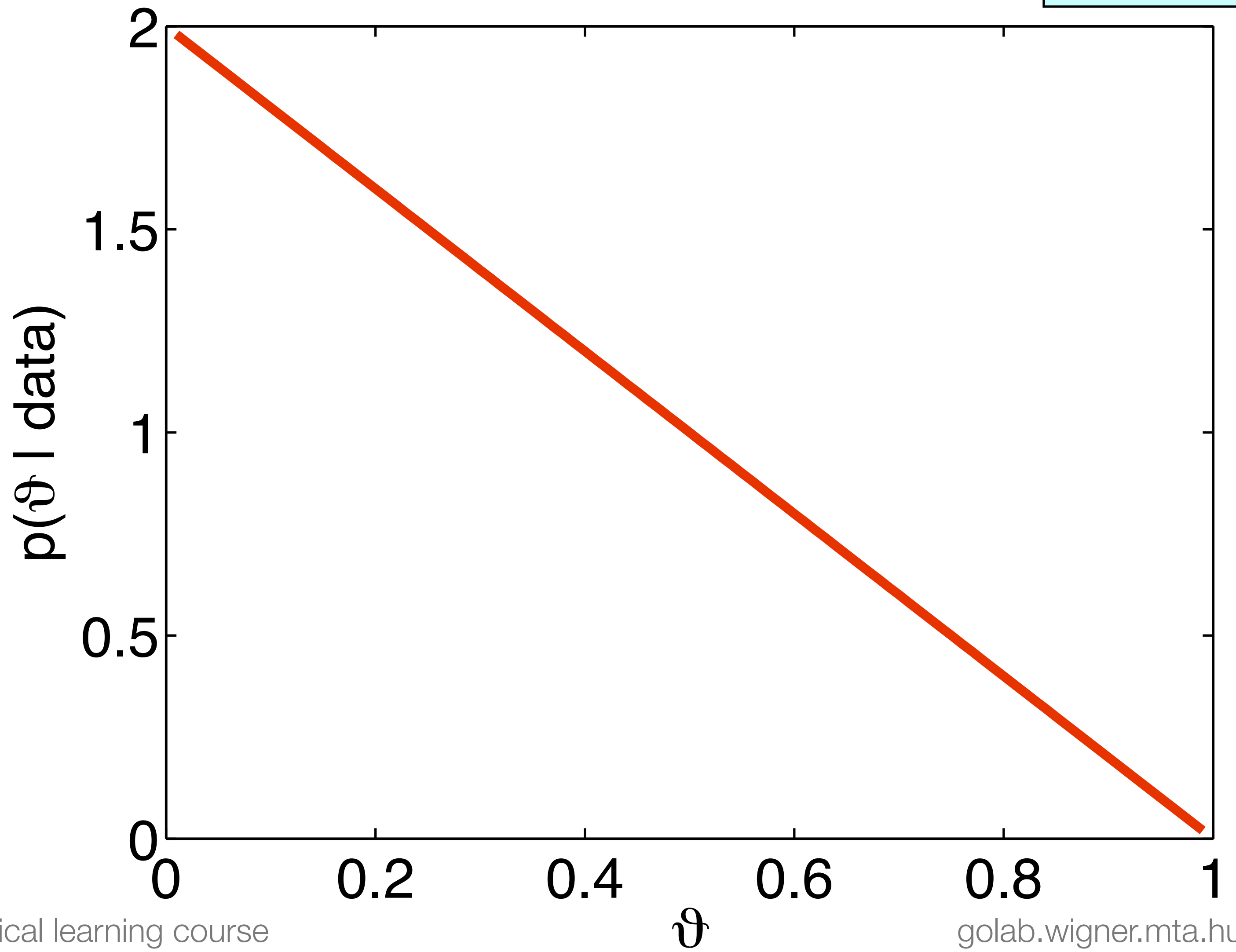
Trial likelihood $P(x \,|\, \vartheta) = \mathrm{Bernoulli}\left(x; \vartheta\right) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P\left(\mathrm{data}|\vartheta\right) = \prod_t P\left(x_t|\vartheta\right)$

Bayes rule

$$P\left(\vartheta|\mathrm{data}\right) = P\left(\mathrm{data}|\vartheta\right) P\left(\vartheta\right) / P\left(\mathrm{data}\right)$$

Is this the important quantity?



$$P(\vartheta) = \mathrm{Beta}(\vartheta; \alpha, \beta) = \frac{1}{B\left(\alpha, \beta\right)} \vartheta^{\alpha - 1} \left(1-\vartheta\right)^{\beta-1}$$

$$P(\vartheta \,|\, \mathrm{data}) = \mathrm{Beta}\Big(\vartheta; \alpha^{(t)}, \beta^{(t)}\Big)$$

$$\alpha^{(t)} = \alpha + \sum_{i=1}^{t} x_i$$

$$\beta^{(t)} = \beta + t - \sum_{i=1}^{t} x_i$$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}(\vartheta) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

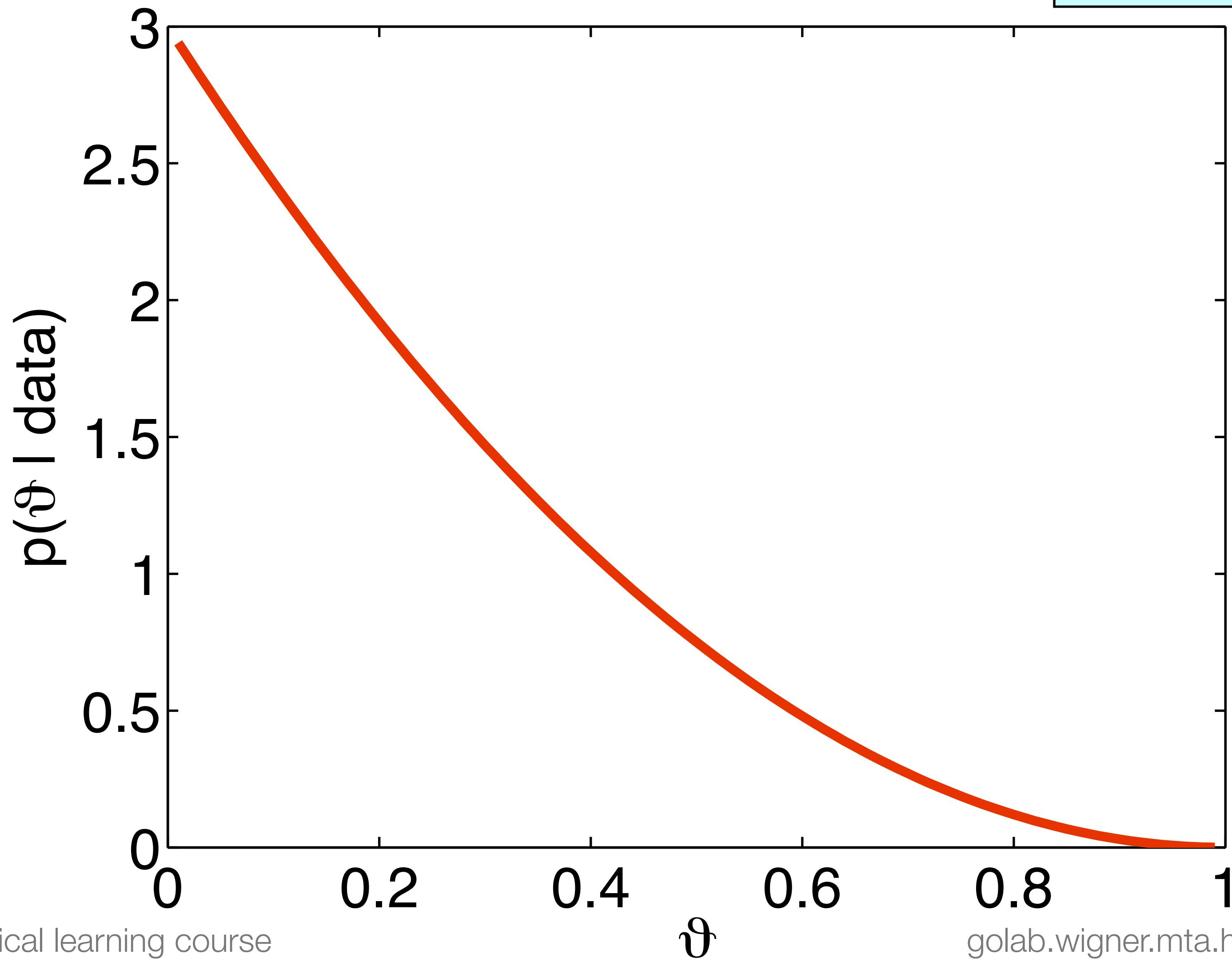
Trial likelihood $P(x \mid \vartheta) = \mathrm{Bernoulli}(x; \vartheta) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P(\mathrm{data}|\vartheta) = \prod_t P(x_t|\vartheta)$

Bayes rule

$$P(\vartheta|\mathrm{data}) = P(\mathrm{data}|\vartheta)\, P(\vartheta) / P(\mathrm{data})$$

Is this the important quantity?



$$P(\vartheta) = \mathrm{Beta}(\vartheta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \vartheta^{\alpha-1} (1-\vartheta)^{\beta-1}$$

$$P(\vartheta \mid \mathrm{data}) = \mathrm{Beta}\left(\vartheta; \alpha^{(t)}, \beta^{(t)}\right)$$

$$\alpha^{(t)} = \alpha + \sum_{i=1}^{t} x_i$$

$$\beta^{(t)} = \beta + t - \sum_{i=1}^{t} x_i$$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

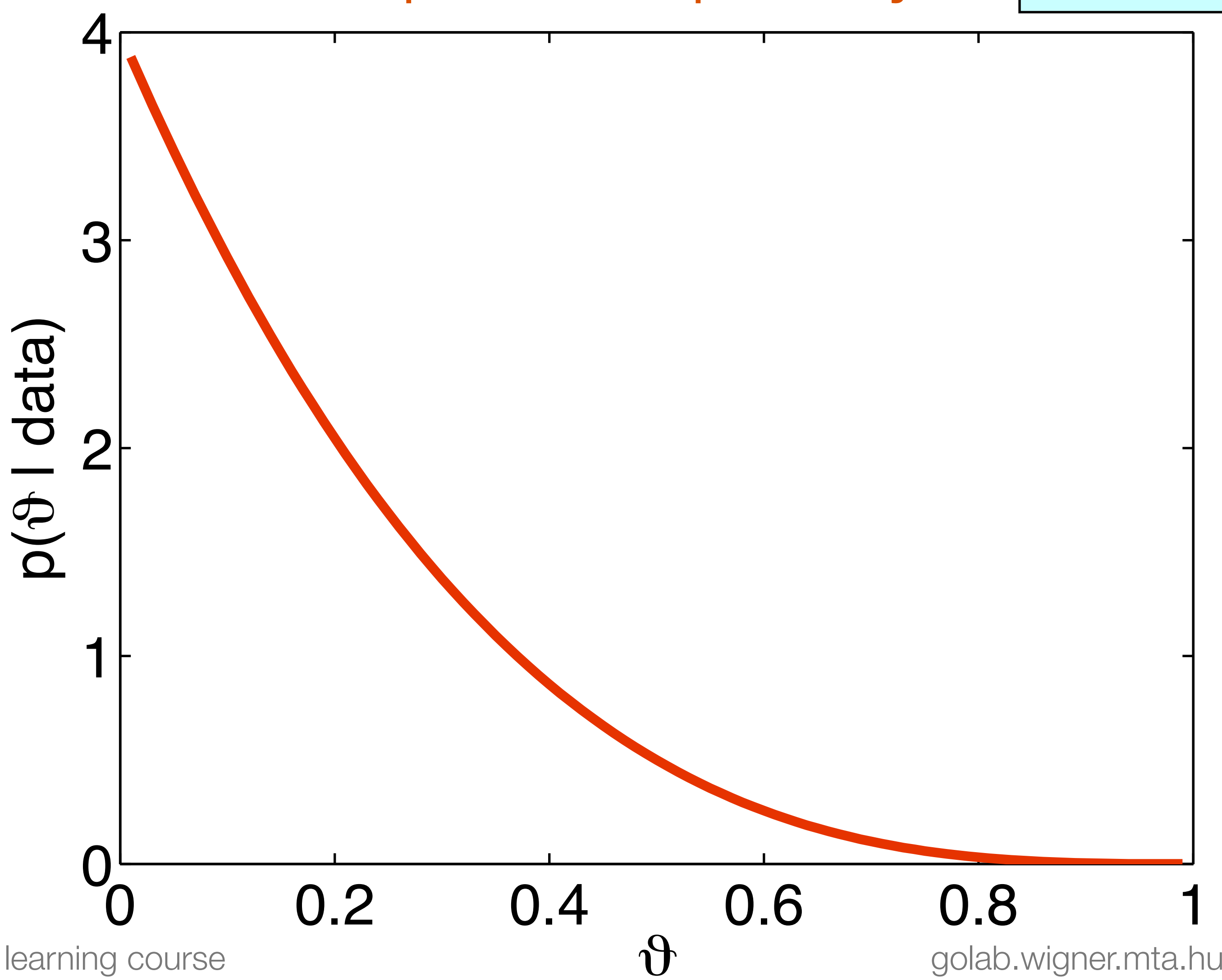
Trial likelihood $P(x \mid \vartheta) = \mathrm{Bernoulli}\left(x; \vartheta\right) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P\left(\mathrm{data}|\vartheta\right) = \prod_t P\left(x_t|\vartheta\right)$

Bayes rule

$$P\left(\vartheta|\mathrm{data}\right) = P\left(\mathrm{data}|\vartheta\right) P\left(\vartheta\right) / P\left(\mathrm{data}\right)$$

Is this the important quantity?



$$P(\vartheta) = \mathrm{Beta}(\vartheta; \alpha, \beta) = \frac{1}{B\left(\alpha, \beta\right)} \vartheta^{\alpha-1} \left(1-\vartheta\right)^{\beta-1}$$

$$P(\vartheta \mid \mathrm{data}) = \mathrm{Beta}\Big(\vartheta; \alpha^{(t)}, \beta^{(t)}\Big)$$

$$\alpha^{(t)} = \alpha + \sum_{i=1}^{t} x_i$$

$$\beta^{(t)} = \beta + t - \sum_{i=1}^{t} x_i$$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}(\vartheta) = \frac{1}{N-1} \sum_i (x_i - \langle \vartheta \rangle)^2$$

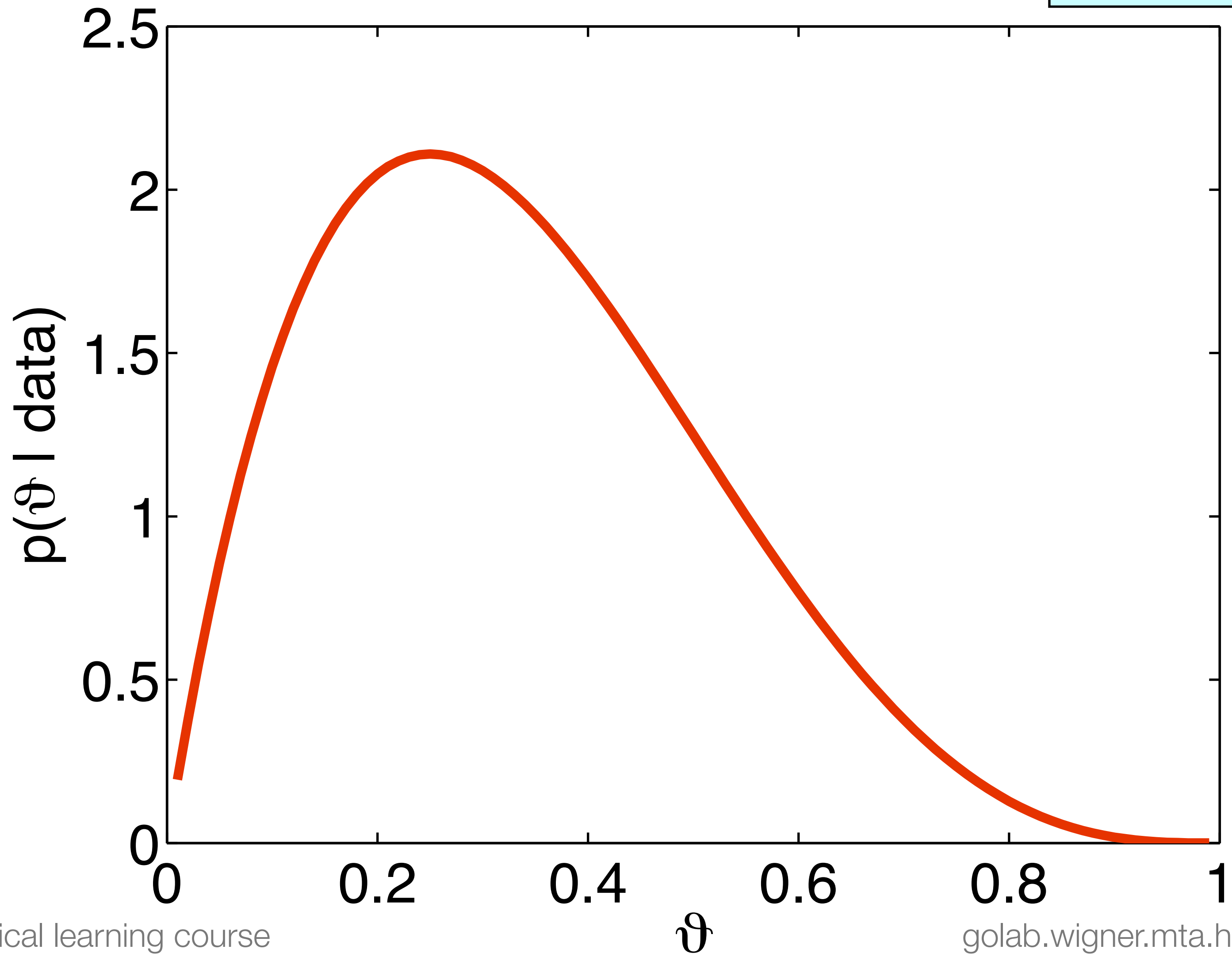
Trial likelihood $P(x \mid \vartheta) = \mathrm{Bernoulli}(x; \vartheta) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P(\mathrm{data}|\vartheta) = \prod_t P(x_t|\vartheta)$

Bayes rule

$$P(\vartheta|\mathrm{data}) = P(\mathrm{data}|\vartheta)\, P(\vartheta) / P(\mathrm{data})$$

Is this the important quantity?



$$P(\vartheta) = \mathrm{Beta}(\vartheta; \alpha, \beta) = \frac{1}{B(\alpha,\beta)} \vartheta^{\alpha-1} (1-\vartheta)^{\beta-1}$$

$$P(\vartheta \mid \mathrm{data}) = \mathrm{Beta}\left(\vartheta; \alpha^{(t)}, \beta^{(t)}\right)$$

$$\alpha^{(t)} = \alpha + \sum_{i=1}^{t} x_i$$

$$\beta^{(t)} = \beta + t - \sum_{i=1}^{t} x_i$$

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	I	I	0	I	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}(\vartheta) = \frac{1}{N-1} \sum_i (x_i - \langle \vartheta \rangle)^2$$

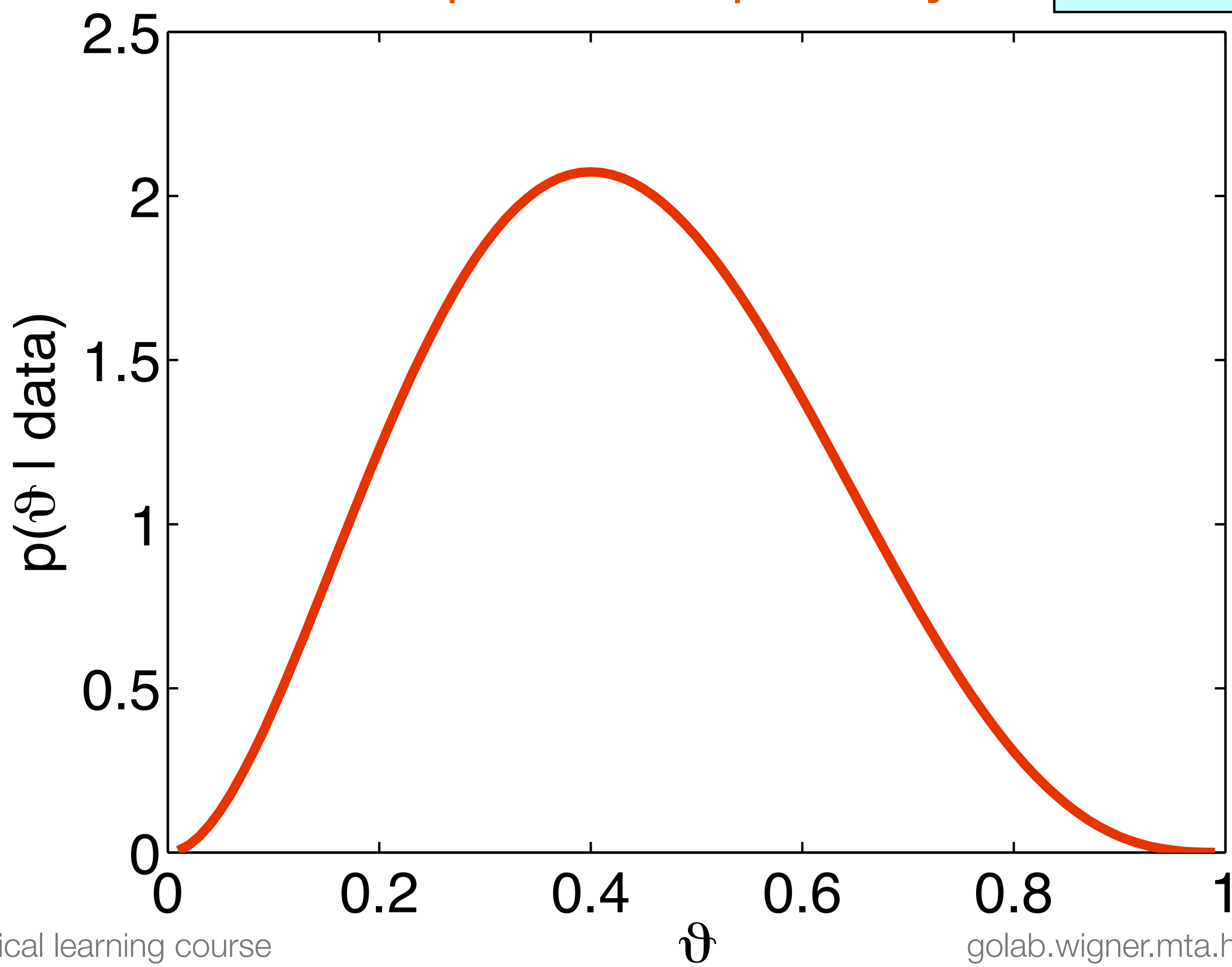
Trial likelihood $P(x \,|\, \vartheta) = \mathrm{Bernoulli}(x; \vartheta) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P(\mathrm{data}|\vartheta) = \prod_t P(x_t|\vartheta)$

Bayes rule

$$P(\vartheta|\mathrm{data}) = P(\mathrm{data}|\vartheta)\, P(\vartheta) / P(\mathrm{data})$$

Is this the important quantity?



$$P(\vartheta) = \mathrm{Beta}(\vartheta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \vartheta^{\alpha-1} (1-\vartheta)^{\beta-1}$$

$$P(\vartheta \,|\, \mathrm{data}) = \mathrm{Beta}\left(\vartheta; \alpha^{(t)}, \beta^{(t)}\right)$$

$$\alpha^{(t)} = \alpha + \sum_{i=1}^{t} x_i$$

$$\beta^{(t)} = \beta + t - \sum_{i=1}^{t} x_i$$

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle\vartheta\rangle = \frac{1}{N}\sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1}\sum_i \left(x_i - \langle\vartheta\rangle\right)^2$$

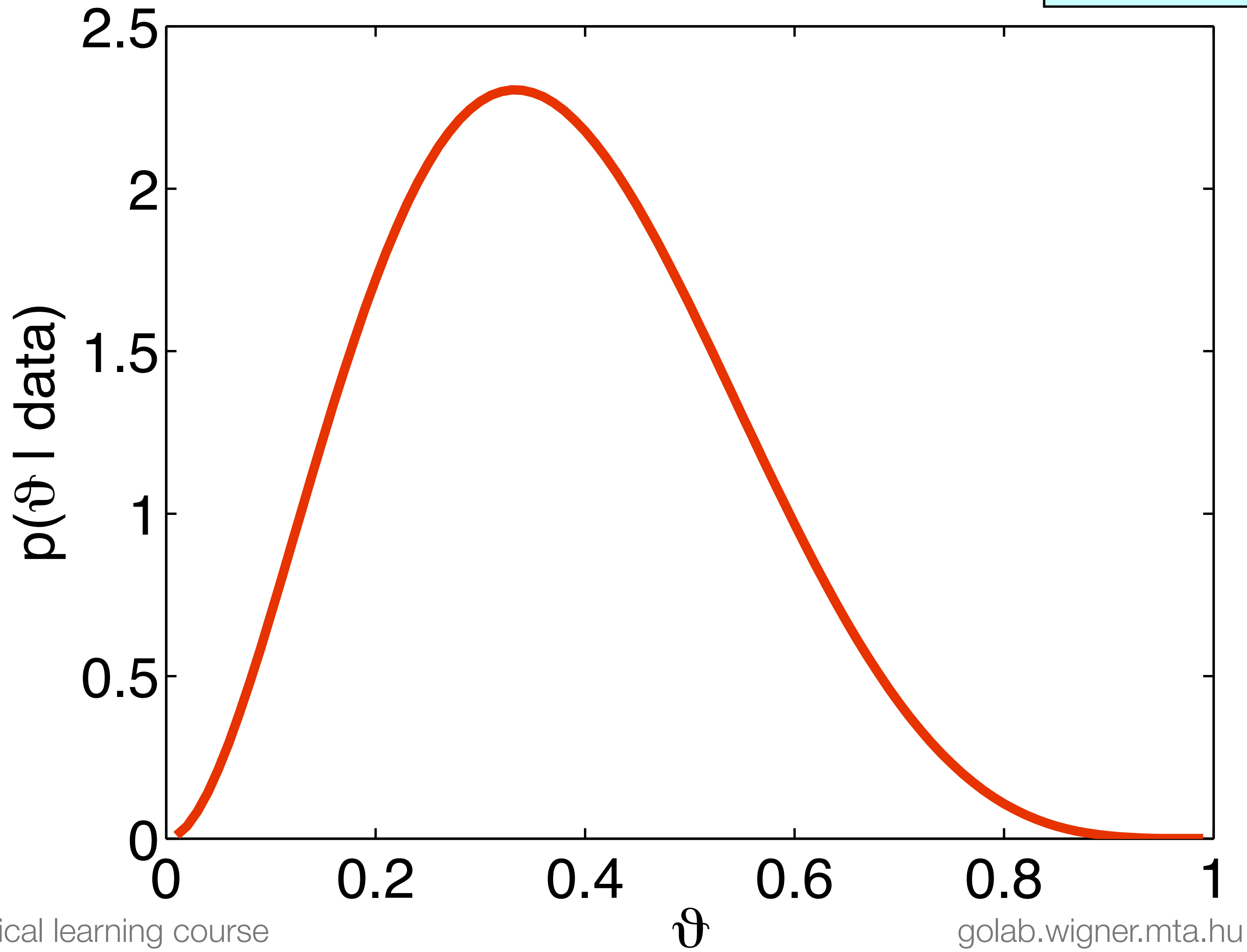
Trial likelihood $P(x\,|\,\vartheta) = \mathrm{Bernoulli}\left(x;\vartheta\right) = \vartheta^x \cdot \left(1-\vartheta\right)^{(1-x)}$

Data likelihood $P\left(\mathrm{data}|\vartheta\right) = \prod_t P\left(x_t|\vartheta\right)$

Bayes rule

$$P\left(\vartheta|\mathrm{data}\right) = P\left(\mathrm{data}|\vartheta\right) P\left(\vartheta\right) / P\left(\mathrm{data}\right)$$

Is this the important quantity?



$$P(\vartheta) = \mathrm{Beta}(\vartheta;\alpha,\beta) = \frac{1}{B\left(\alpha,\beta\right)}\vartheta^{\alpha-1}\left(1-\vartheta\right)^{\beta-1}$$

$$P(\vartheta\,|\,\mathrm{data}) = \mathrm{Beta}\Big(\vartheta;\alpha^{(t)},\beta^{(t)}\Big)$$

$$\alpha^{(t)} = \alpha + \sum_{i=1}^{t} x_i$$

$$\beta^{(t)} = \beta + t - \sum_{i=1}^{t} x_i$$

Coin tossing: an example

Head: 0
Tail: 1

Result	0	0	0	1	1	0	1 ...
Estimated bias	0	0	0	.25	.4	.33	.43
Variance of bias		0	0	.25	.3	.26	.28

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}\left(\vartheta\right) = \frac{1}{N-1} \sum_i \left(x_i - \langle \vartheta \rangle\right)^2$$

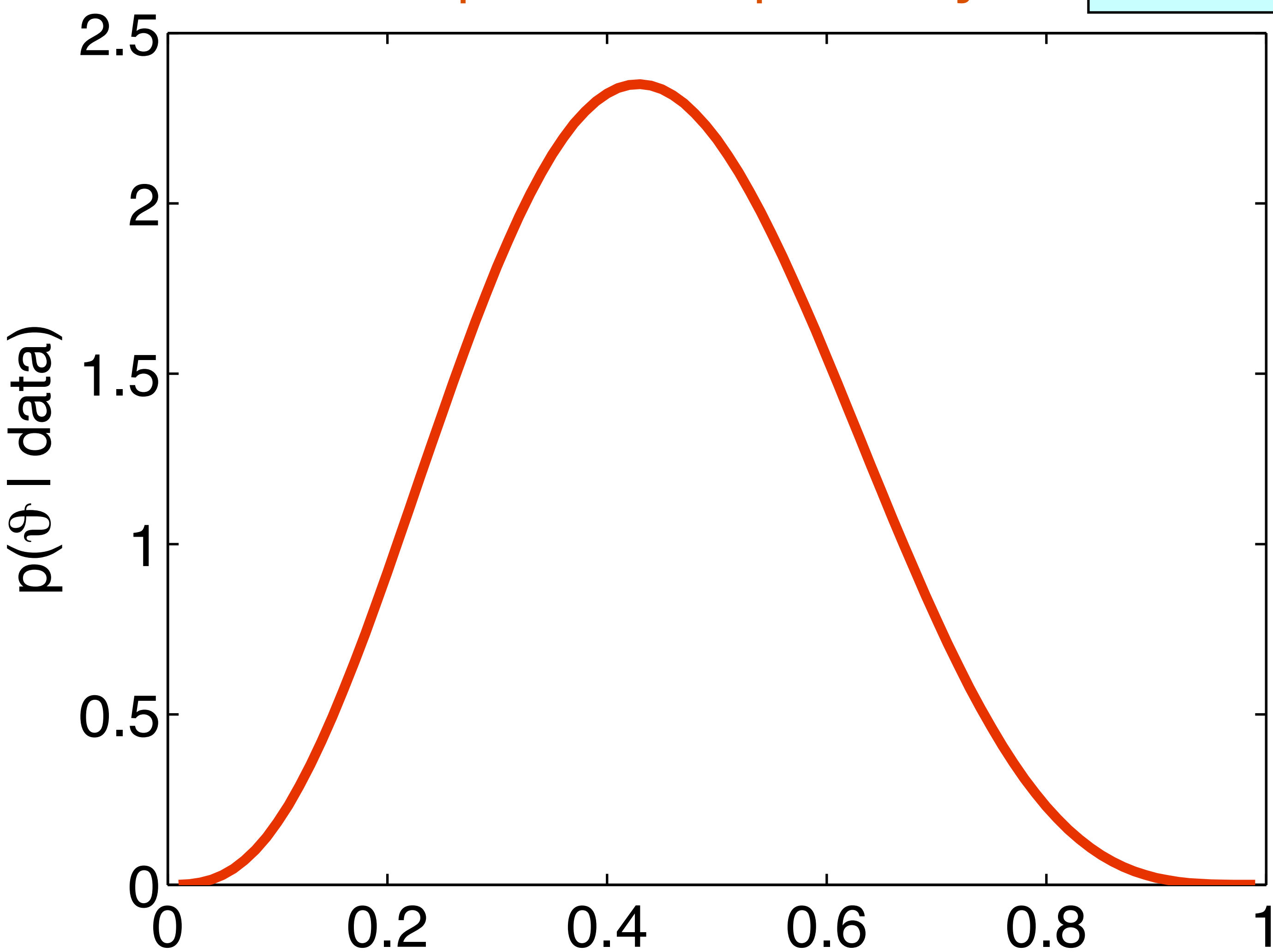
Trial likelihood $P(x \,|\, \vartheta) = \mathrm{Bernoulli}\left(x; \vartheta\right) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P\left(\mathrm{data}|\vartheta\right) = \prod_t P\left(x_t|\vartheta\right)$

Bayes rule

$$P\left(\vartheta|\mathrm{data}\right) = P\left(\mathrm{data}|\vartheta\right) P\left(\vartheta\right) / P\left(\mathrm{data}\right)$$

Is this the important quantity?



$$P(\vartheta) = \mathrm{Beta}(\vartheta; \alpha, \beta) = \frac{1}{B\left(\alpha, \beta\right)} \vartheta^{\alpha-1} \left(1-\vartheta\right)^{\beta-1}$$

$$P(\vartheta \,|\, \mathrm{data}) = \mathrm{Beta}\left(\vartheta; \alpha^{(t)}, \beta^{(t)}\right)$$

$$\alpha^{(t)} = \alpha + \sum_{i=1}^{t} x_i$$

$$\beta^{(t)} = \beta + t - \sum_{i=1}^{t} x_i$$

Coin tossing: an example

Head:	0
Tail:	1

Result	0	0	0	1	1	0	1	...
Estimated bias	0	0	0	.25	.4	.33	.43	
Variance of bias		0	0	.25	.3	.26	.28	

$$\langle \vartheta \rangle = \frac{1}{N} \sum_i x_i$$

$$\mathrm{Var}(\vartheta) = \frac{1}{N-1} \sum_i (x_i - \langle \vartheta \rangle)^2$$

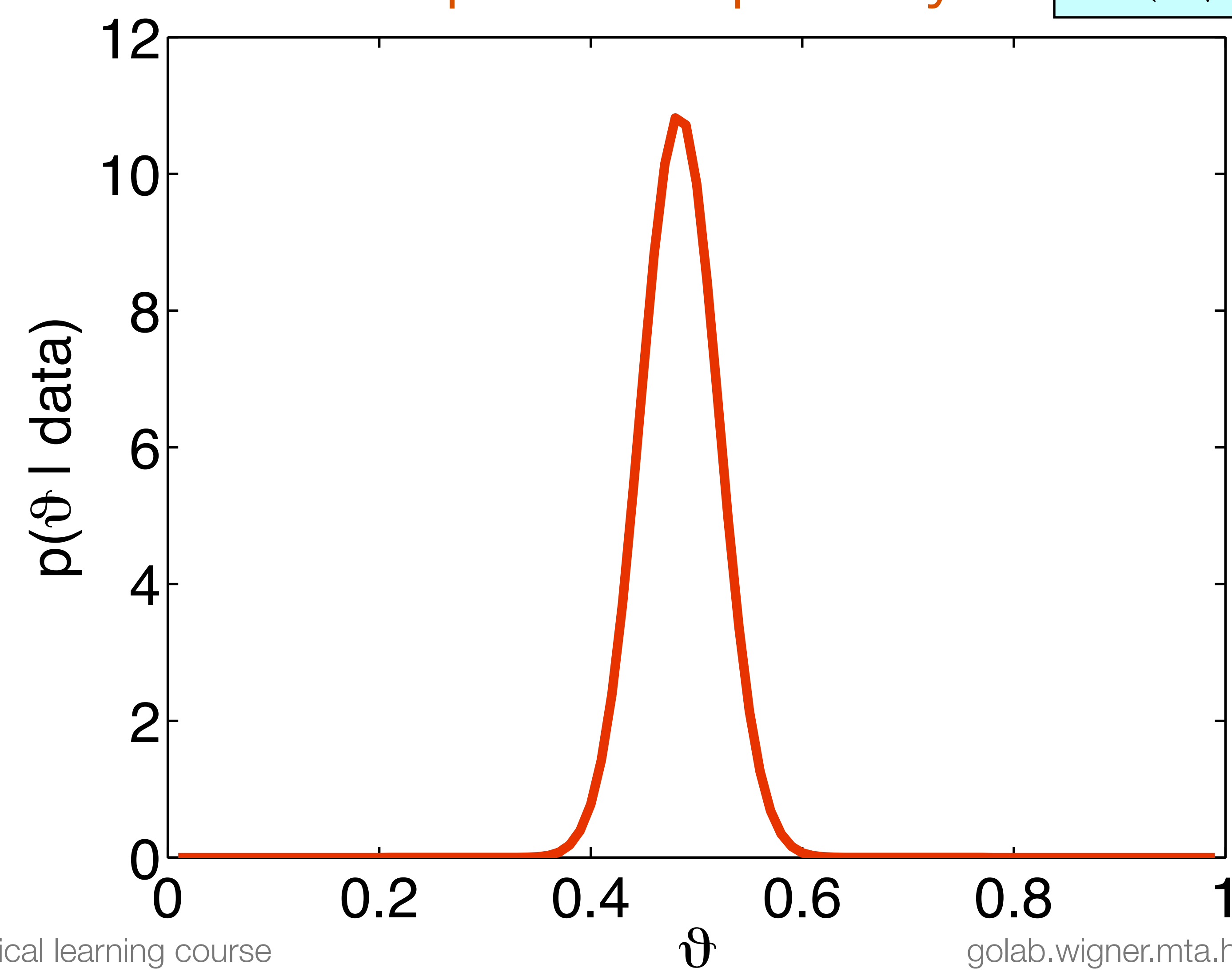
Trial likelihood $P(x \mid \vartheta) = \mathrm{Bernoulli}(x; \vartheta) = \vartheta^x \cdot (1-\vartheta)^{(1-x)}$

Data likelihood $P(\mathrm{data}|\vartheta) = \prod_t P(x_t|\vartheta)$

Bayes rule

$$P(\vartheta|\mathrm{data}) = P(\mathrm{data}|\vartheta)\, P(\vartheta) / P(\mathrm{data})$$

Is this the important quantity?



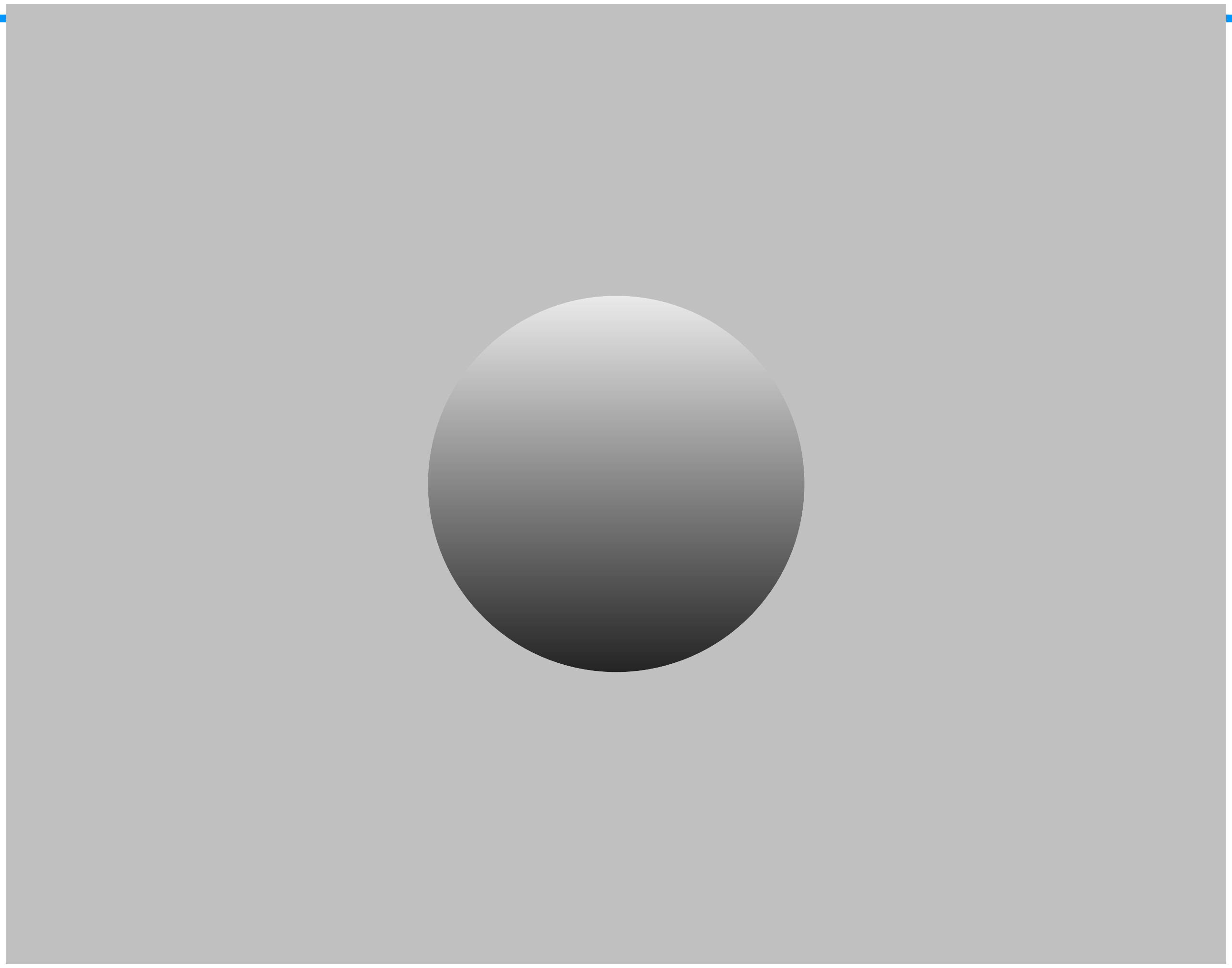
$$P(\vartheta) = \mathrm{Beta}(\vartheta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \vartheta^{\alpha-1} (1-\vartheta)^{\beta-1}$$

$$P(\vartheta \mid \mathrm{data}) = \mathrm{Beta}\left(\vartheta; \alpha^{(t)}, \beta^{(t)}\right)$$

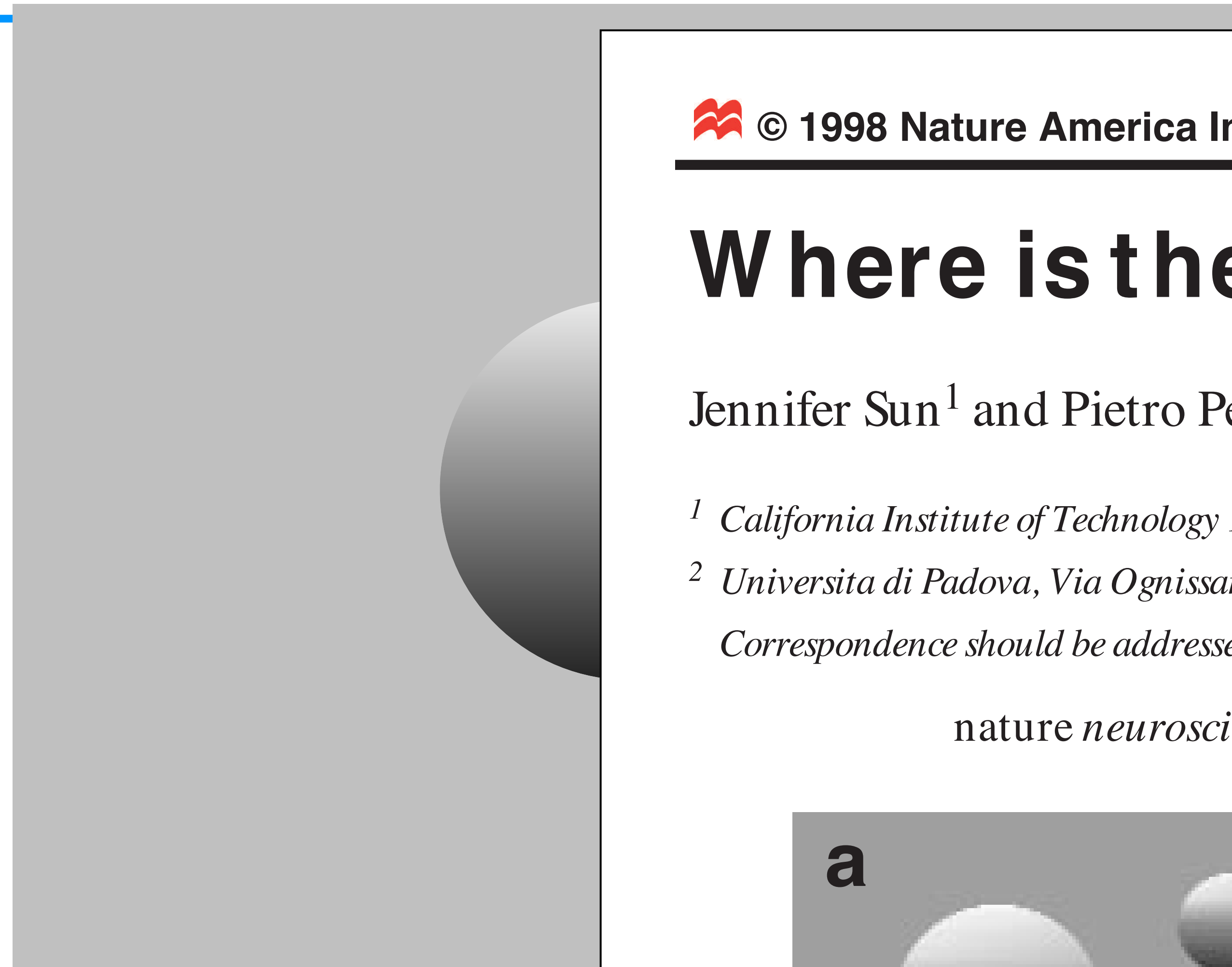
$$\alpha^{(t)} = \alpha + \sum_{i=1}^{t} x_i$$

$$\beta^{(t)} = \beta + t - \sum_{i=1}^{t} x_i$$

Bayesian inference



Bayesian inference



© 1998 Nature America Inc. • http://neurosci.nature.com

Where is the sun?

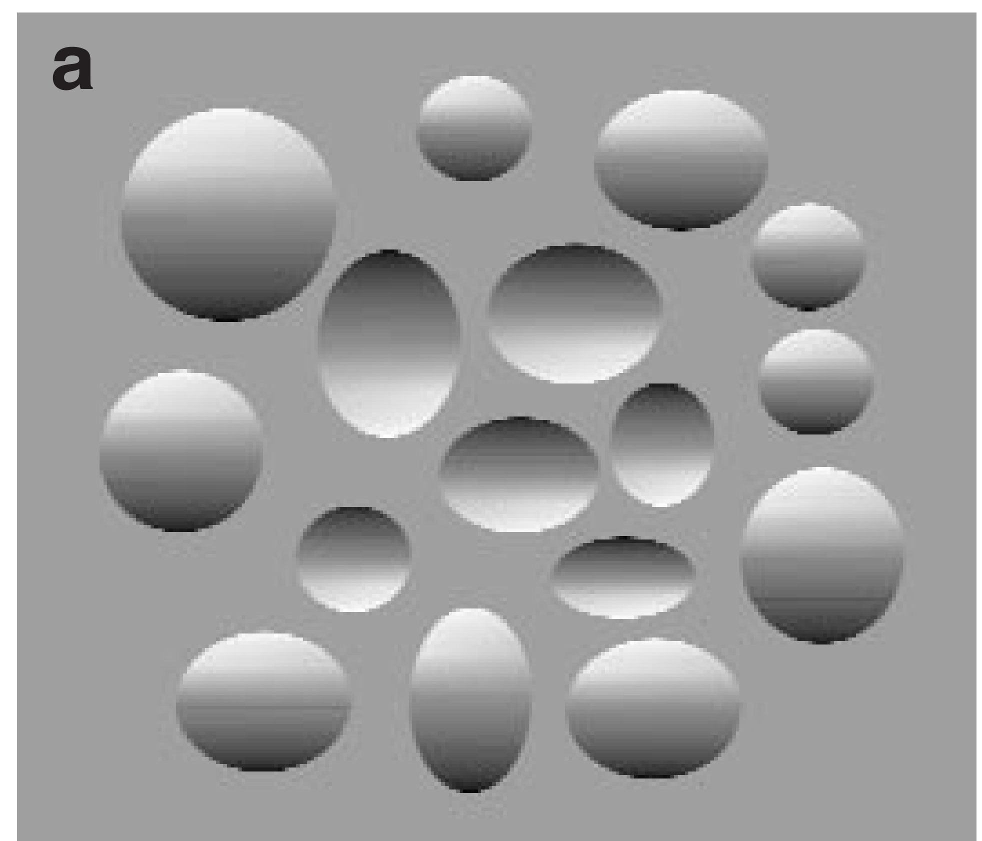
Jennifer Sun[1] and Pietro Perona[1,2]

[1] *California Institute of Technology 136-93, Pasadena, California 91125, USA*

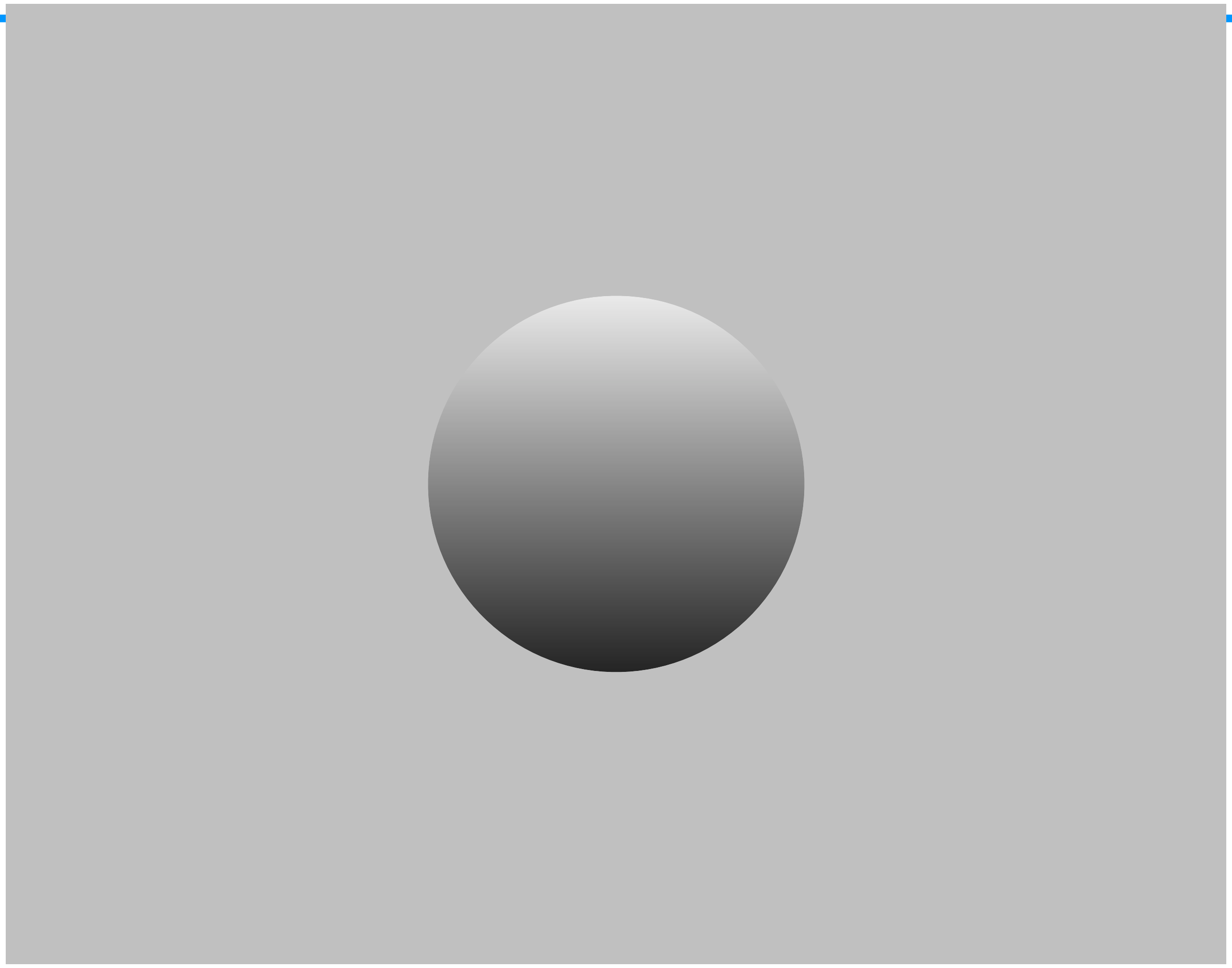
[2] *Universita di Padova, Via Ognissanti 72, 35131 Padova, Italy*

Correspondence should be addressed to P.P. (perona@vision.caltech.edu)

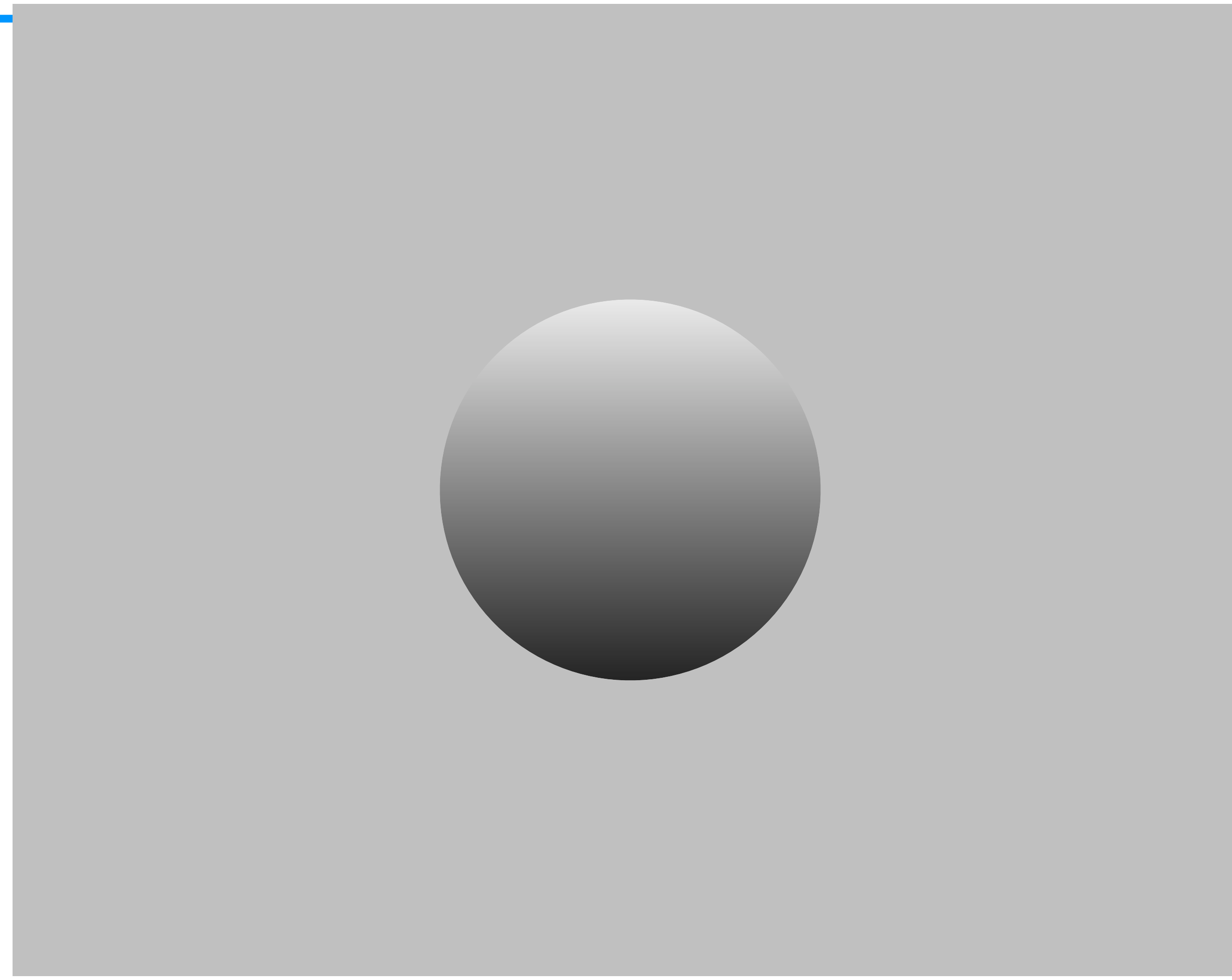
nature *neuroscience* • volume 1 no 3 • july 1998

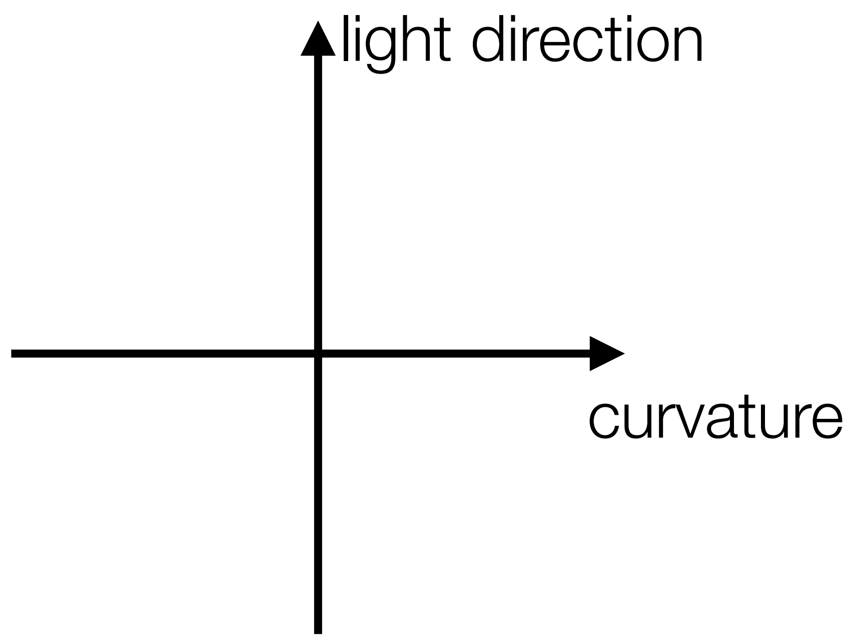


Bayesian inference

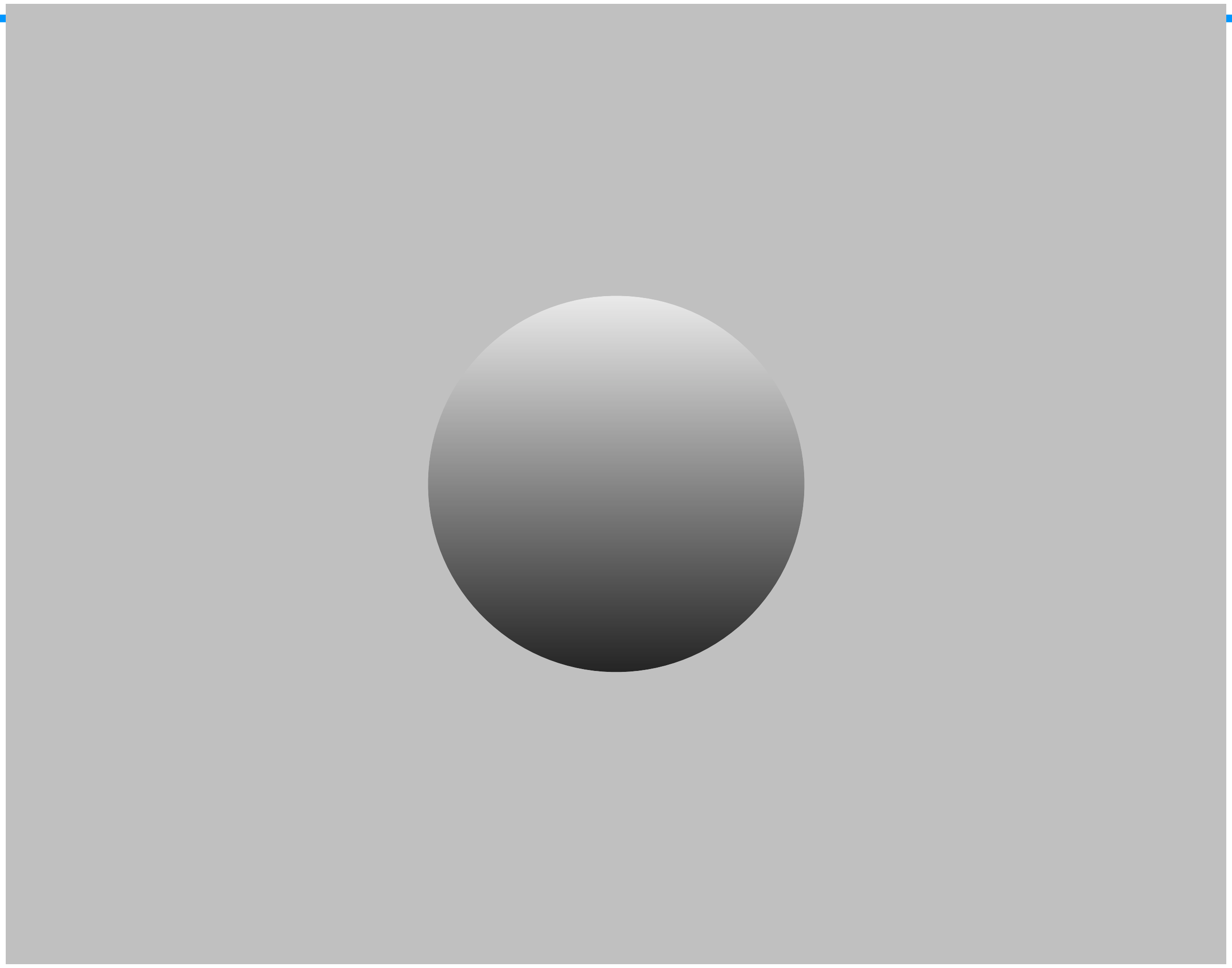


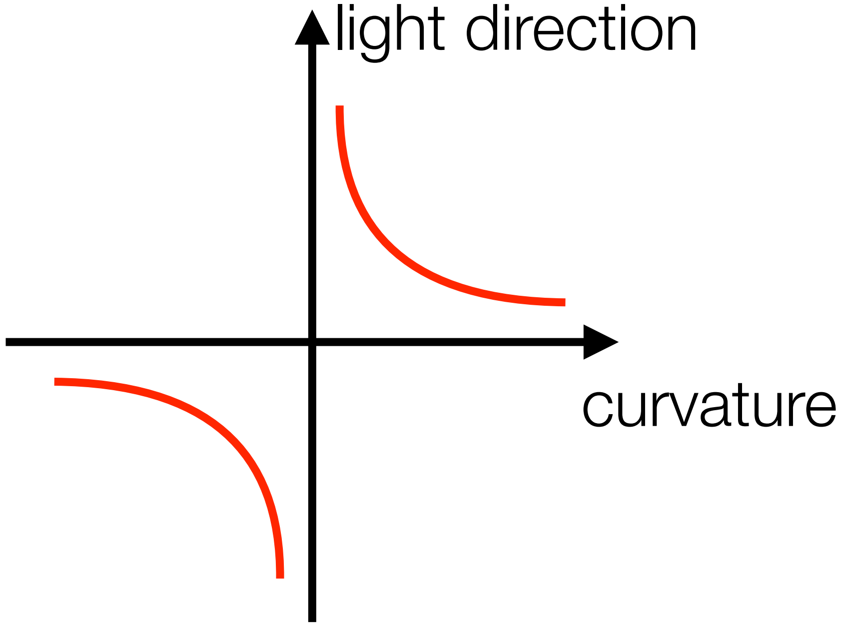
Bayesian inference



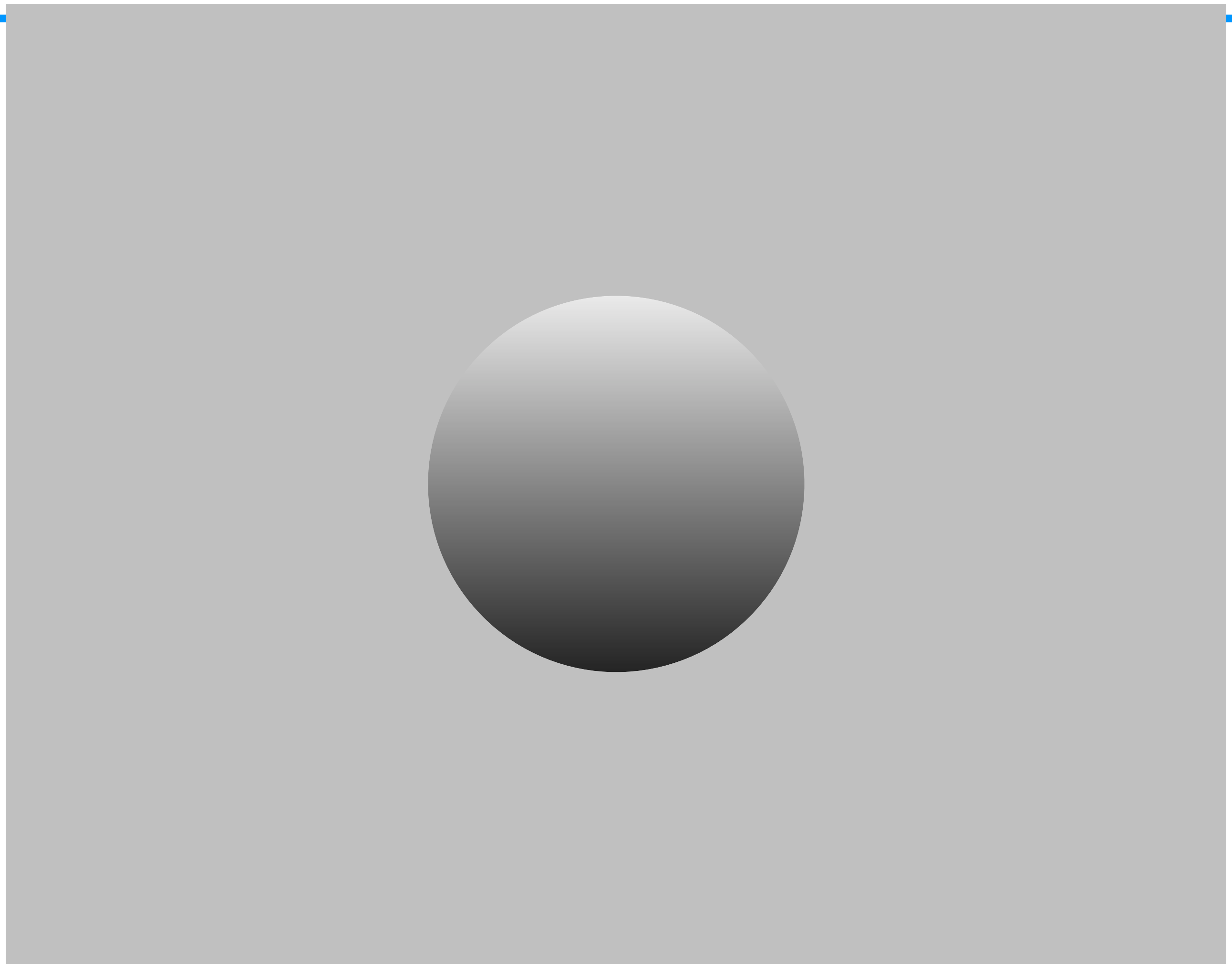


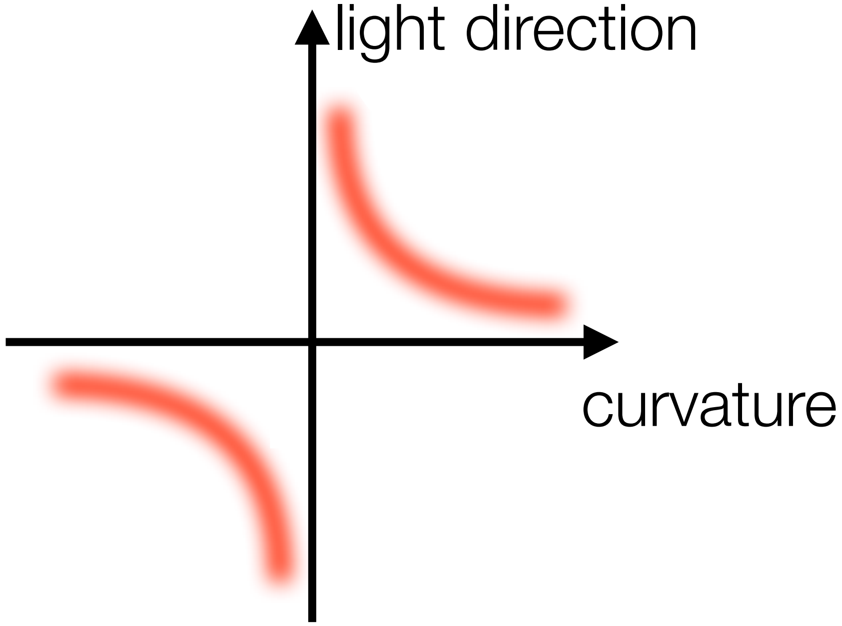
Bayesian inference





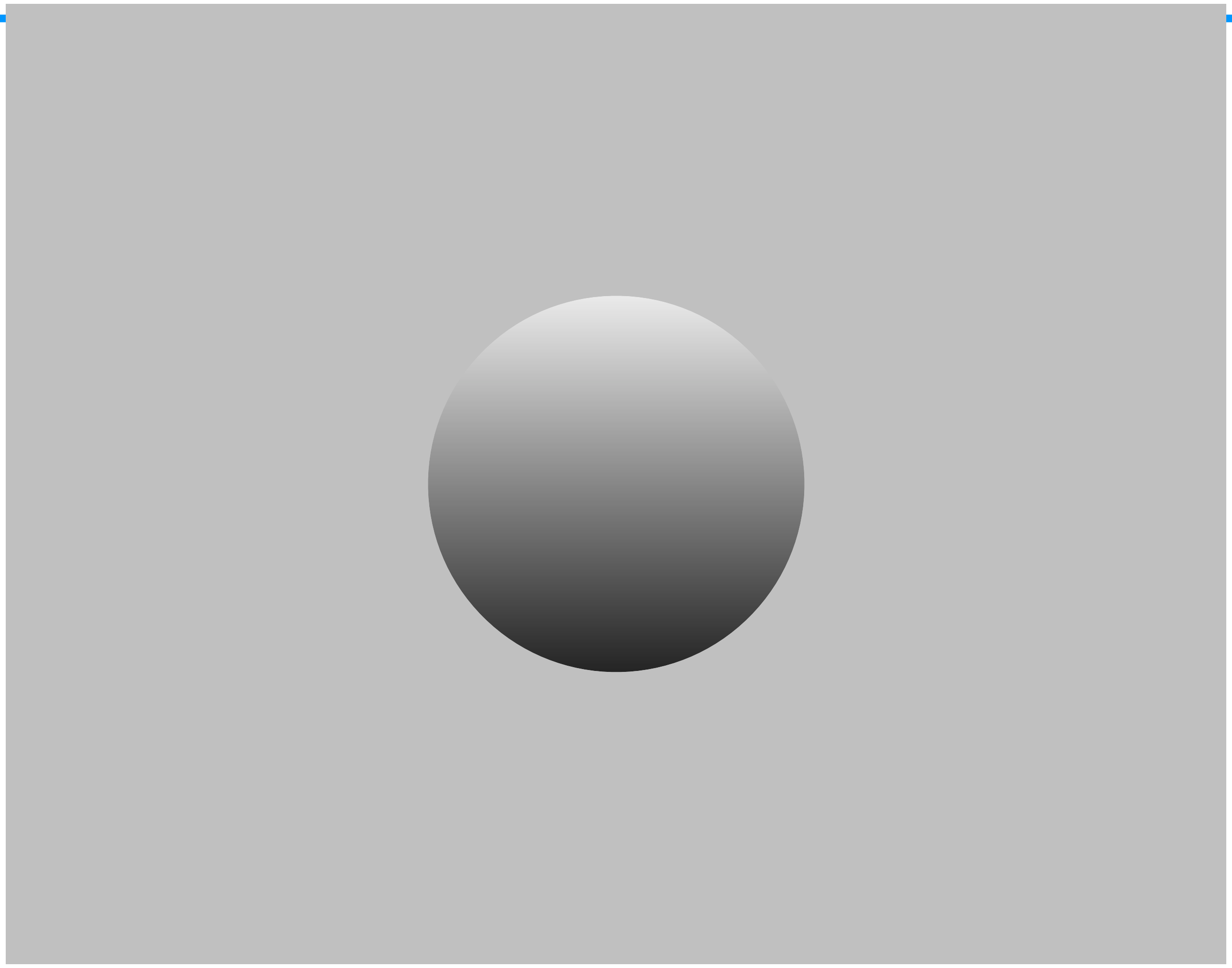
Bayesian inference

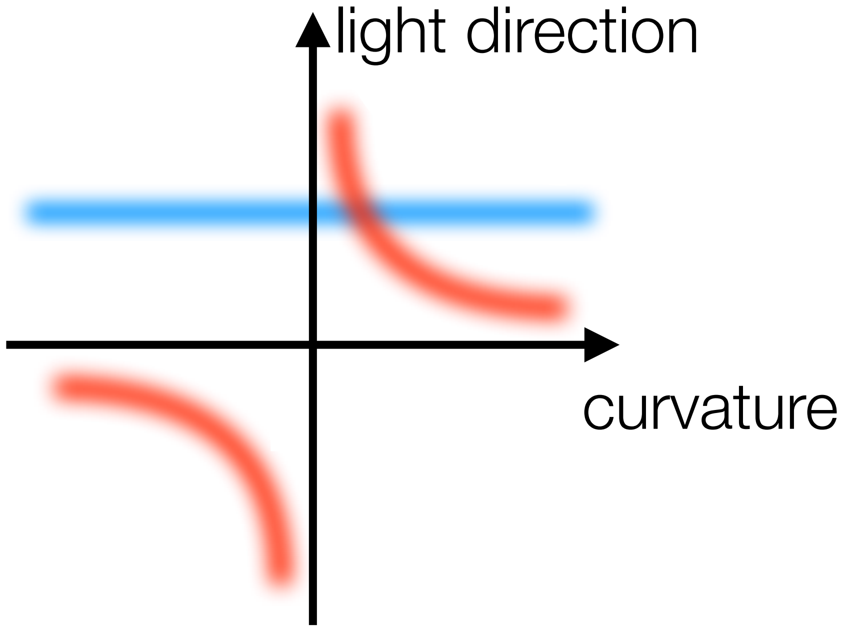




evidence

Bayesian inference

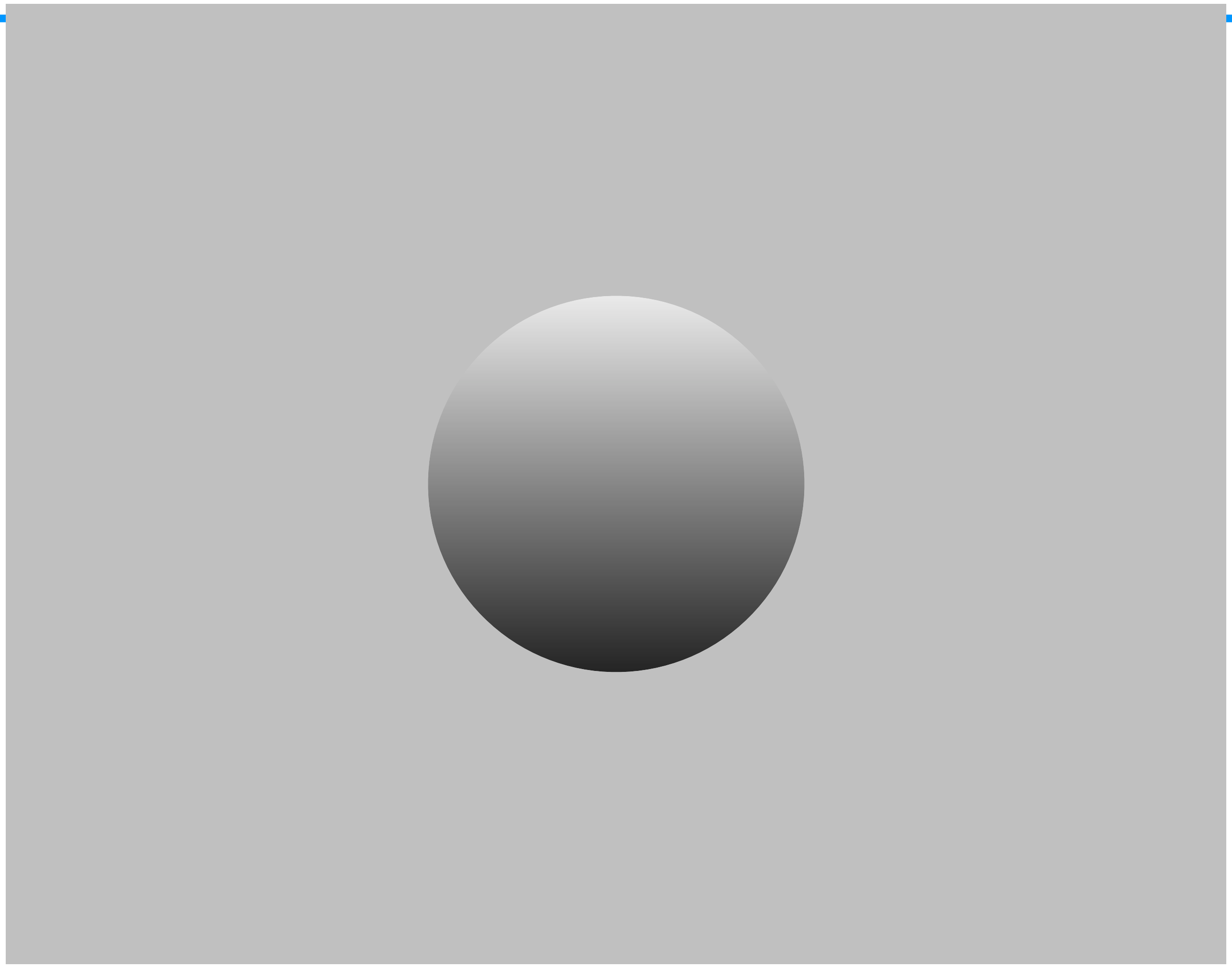


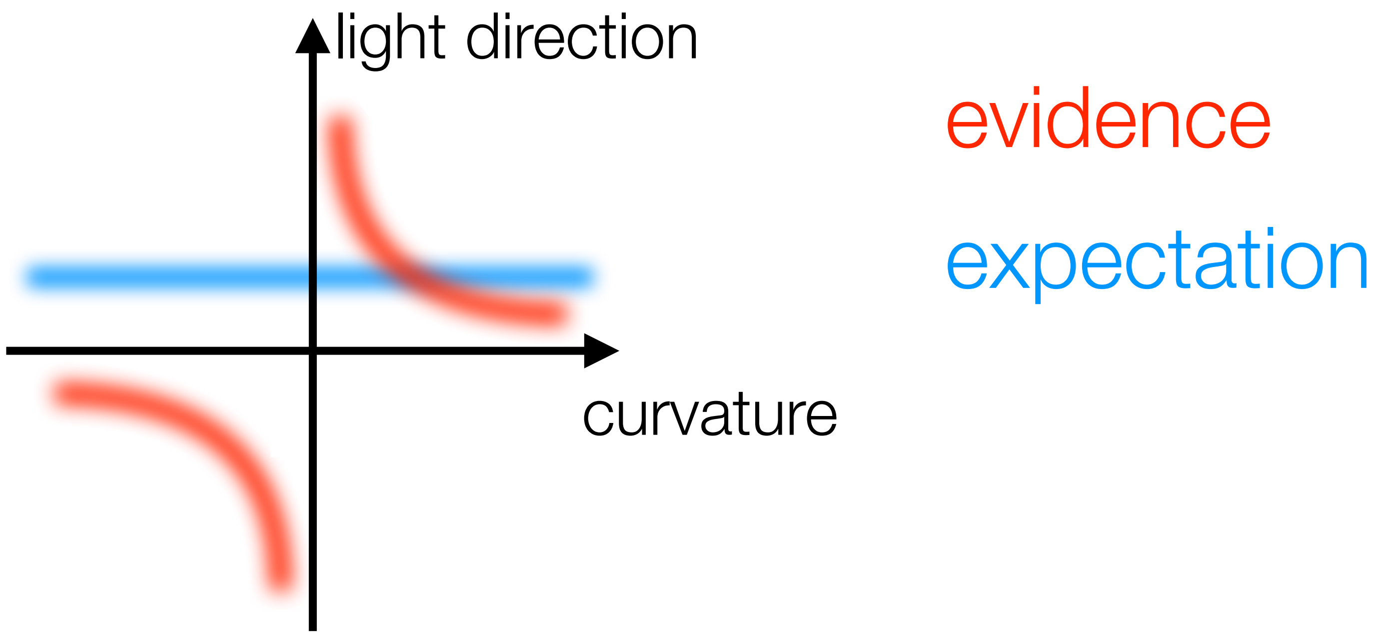


evidence

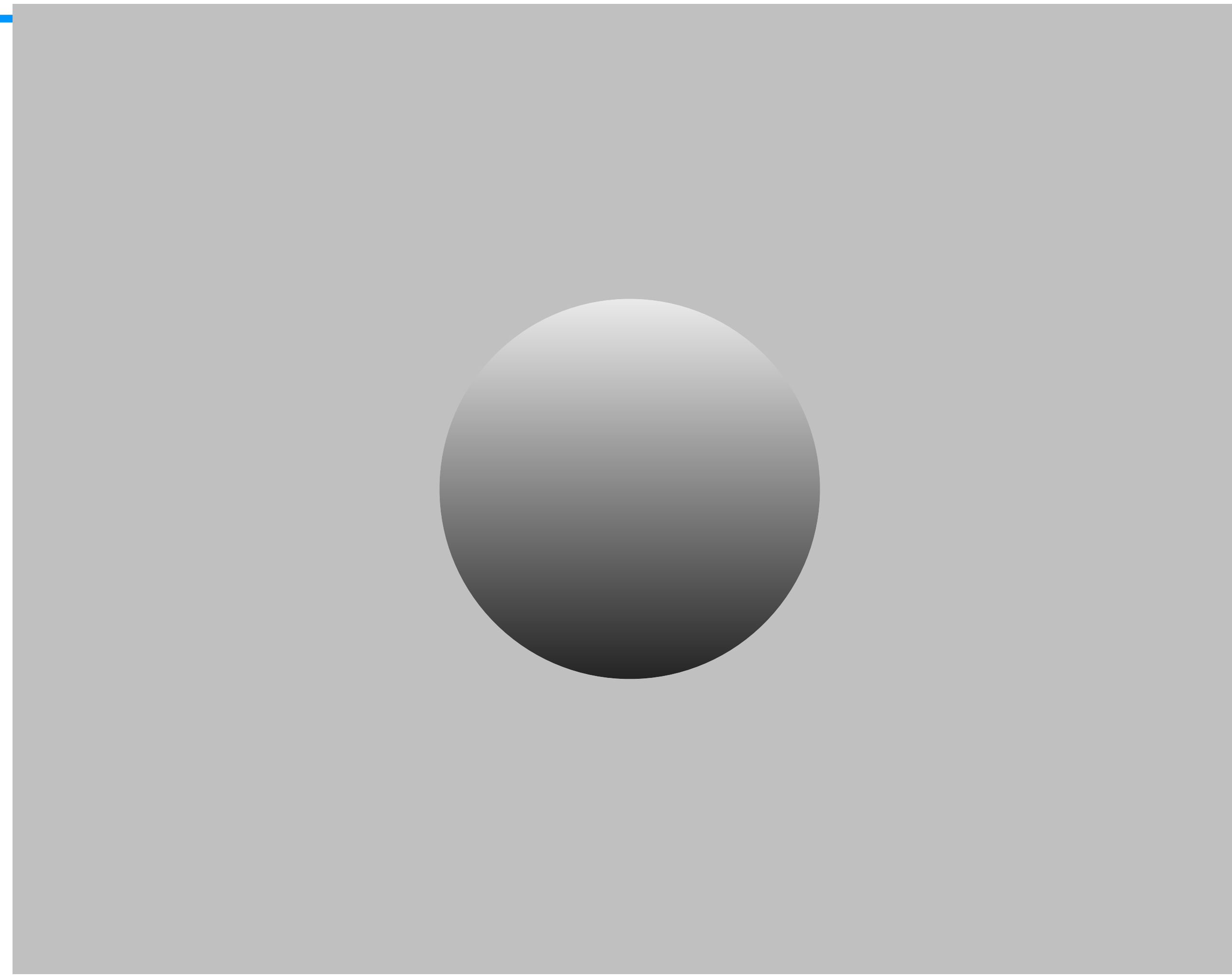
expectation

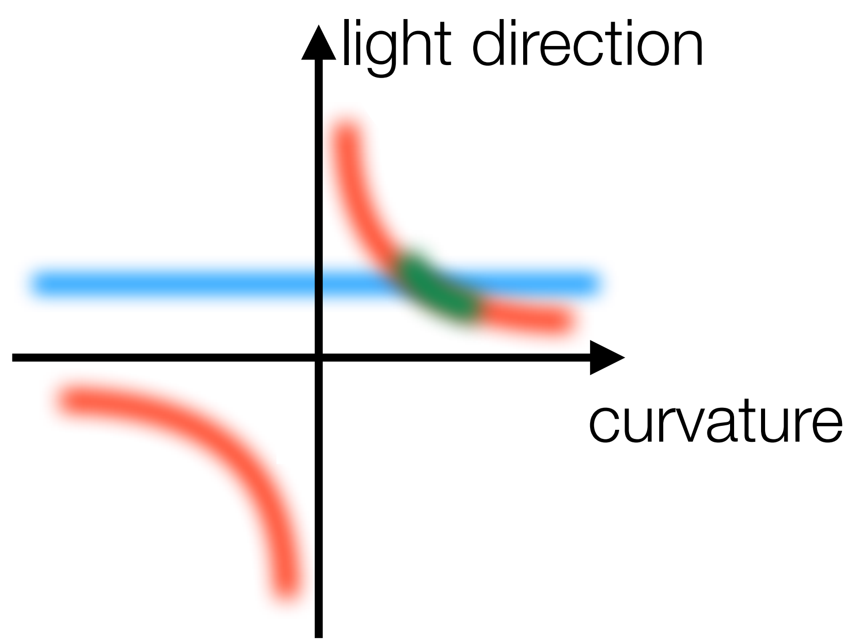
Bayesian inference





Bayesian inference



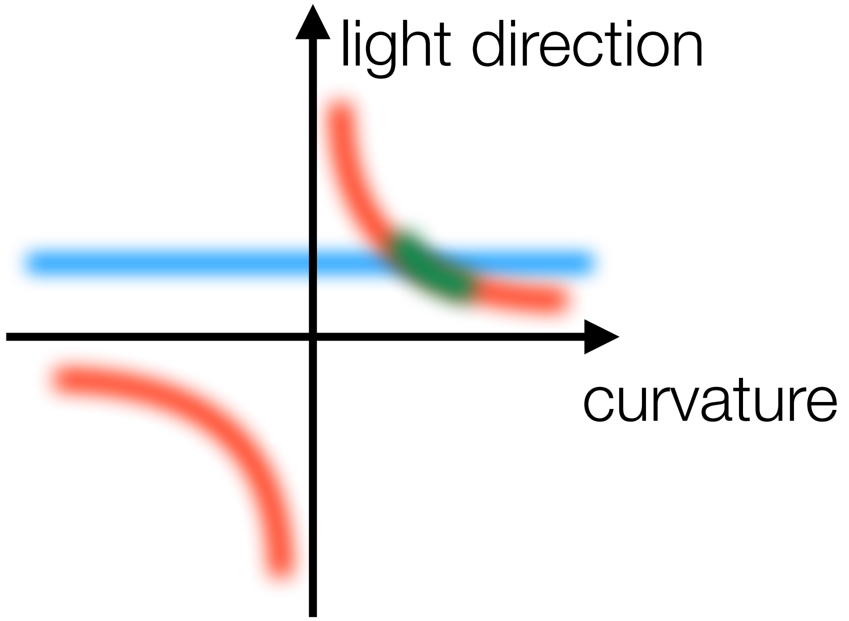


evidence

expectation

inference

Bayesian inference

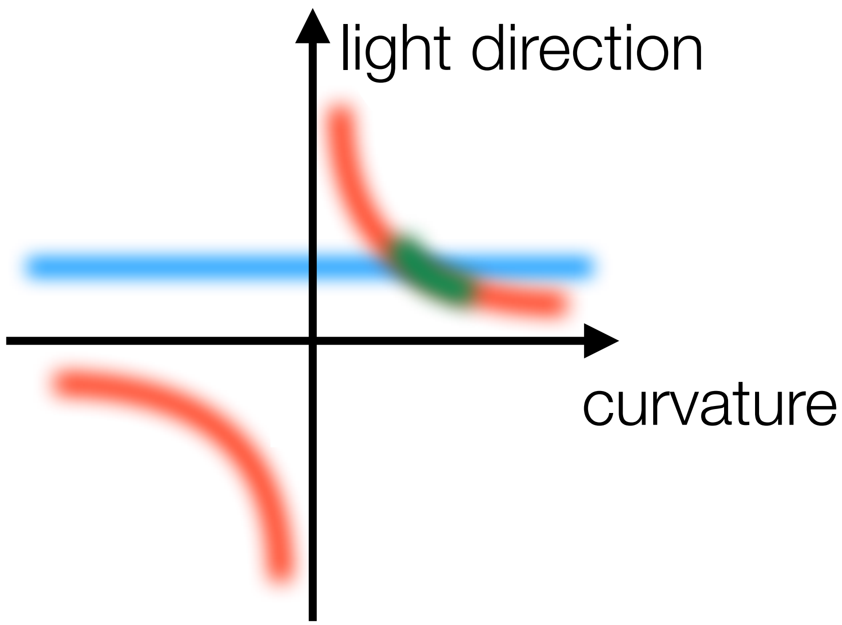


evidence

expectation

inference

Bayesian inference

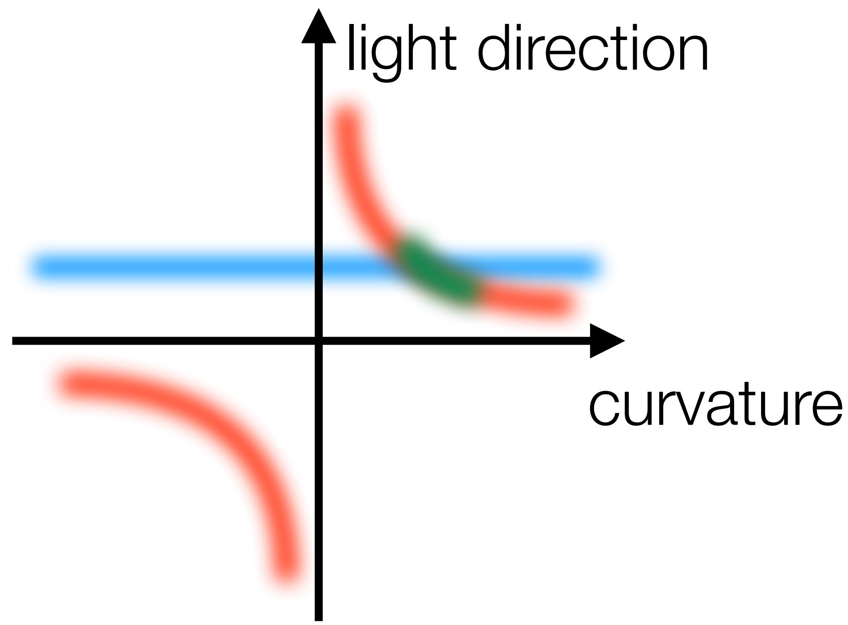


evidence

expectation

inference

Bayesian inference



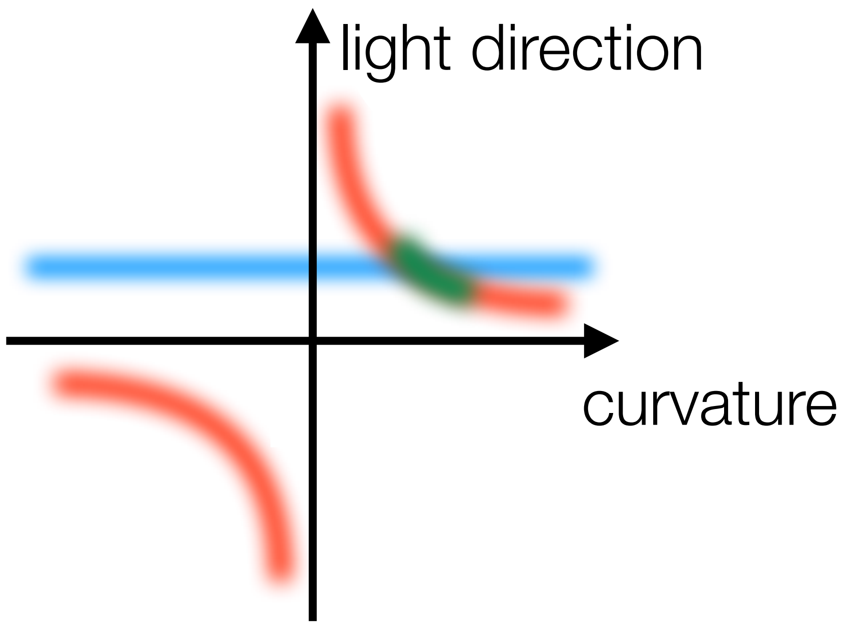
$$\color{green}{P(\text{feature}\,|\,\text{stimulus})} \propto \color{red}{P(\text{stimulus}\,|\,\text{feature})} \times \color{blue}{P(\text{feature})}$$

evidence

expectation

inference

Bayesian inference



$$P(\text{feature} \mid \text{stimulus}) \propto P(\text{stimulus} \mid \text{feature}) \times P(\text{feature})$$

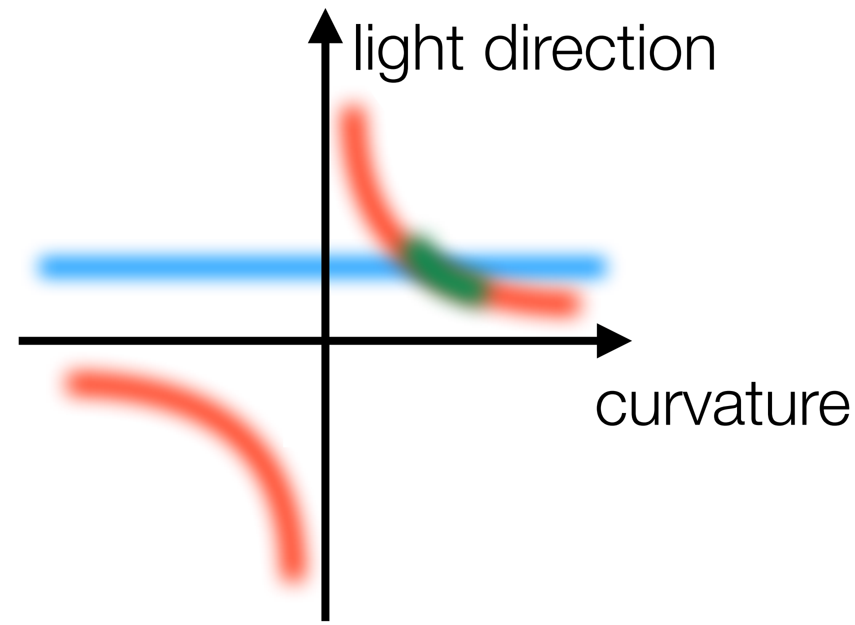
posterior: inference

evidence

expectation

inference

Bayesian inference



$$P(\text{feature} \mid \text{stimulus}) \propto P(\text{stimulus} \mid \text{feature}) \times P(\text{feature})$$

posterior: inference

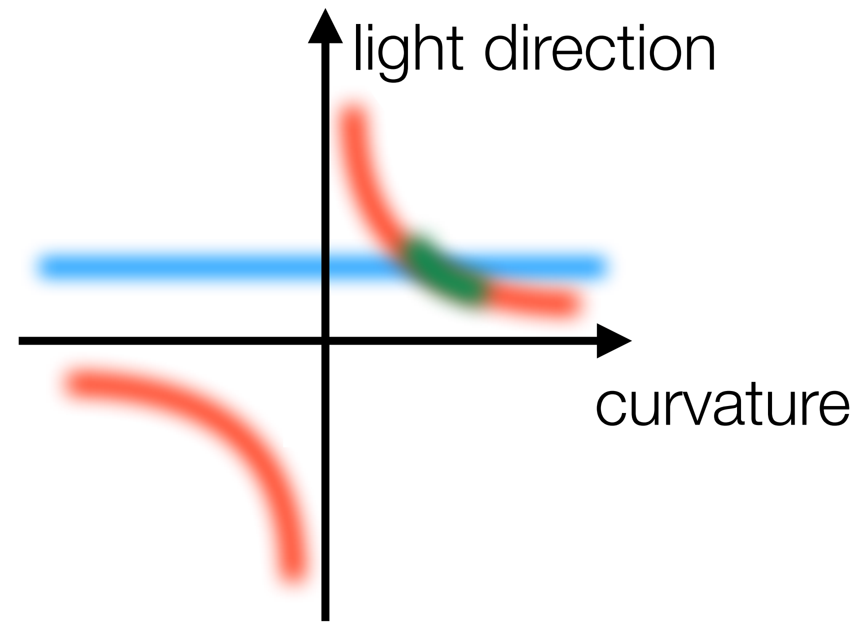
likelihood: evidence

evidence

expectation

inference

Bayesian inference



$$\color{green}{P(\text{feature} \mid \text{stimulus})} \propto \color{red}{P(\text{stimulus} \mid \text{feature})} \times \color{blue}{P(\text{feature})}$$

posterior: inference

likelihood: evidence

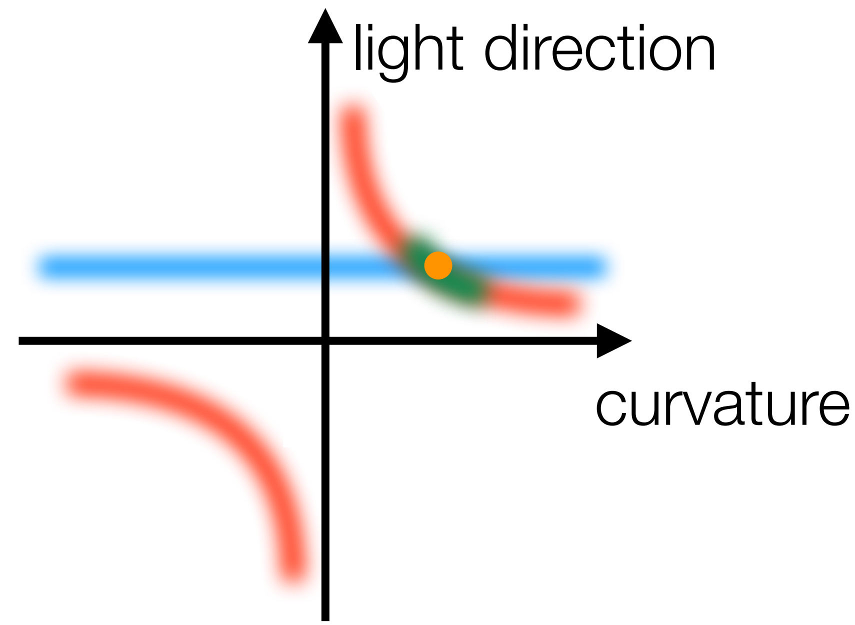
prior : expectations

evidence

expectation

inference

Bayesian inference



$$\color{green}{P(\text{feature} \mid \text{stimulus})} \propto \color{red}{P(\text{stimulus} \mid \text{feature})} \times \color{blue}{P(\text{feature})}$$

posterior: inference

likelihood: evidence

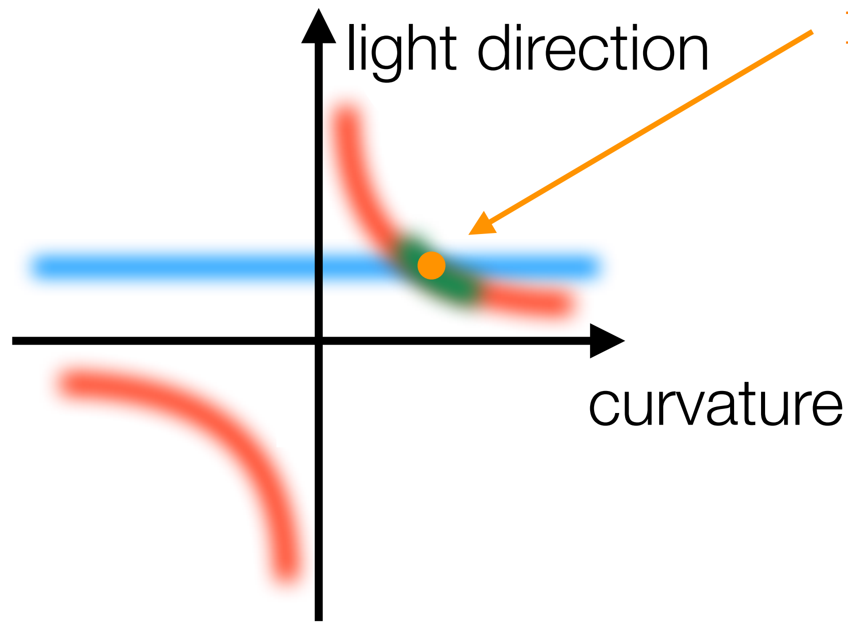
prior : expectations

evidence

expectation

inference

Bayesian inference



$\max[P(\text{feature} \mid \text{stimulus})]$

$$P(\text{feature} \mid \text{stimulus}) \propto P(\text{stimulus} \mid \text{feature}) \times P(\text{feature})$$

posterior: inference

likelihood: evidence

prior : expectations

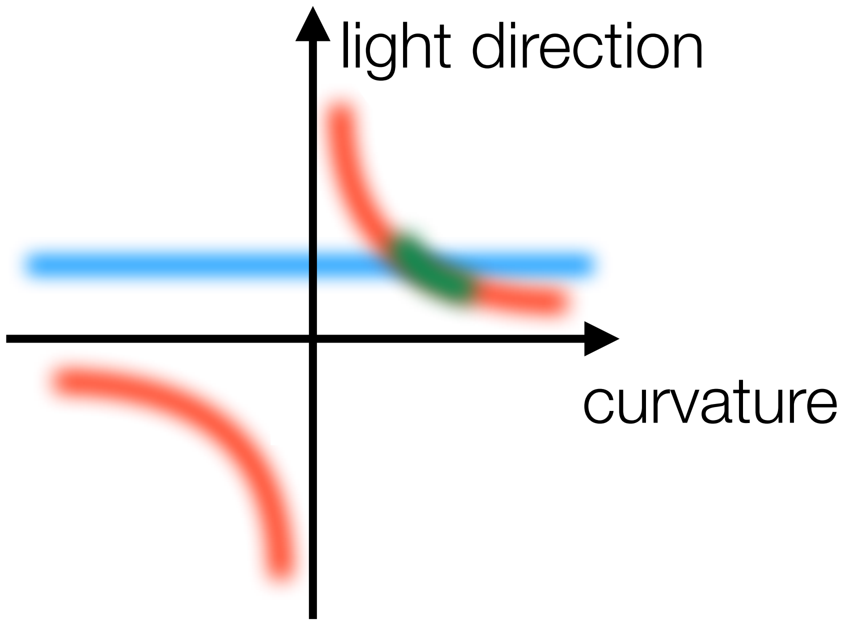
evidence

expectation

inference

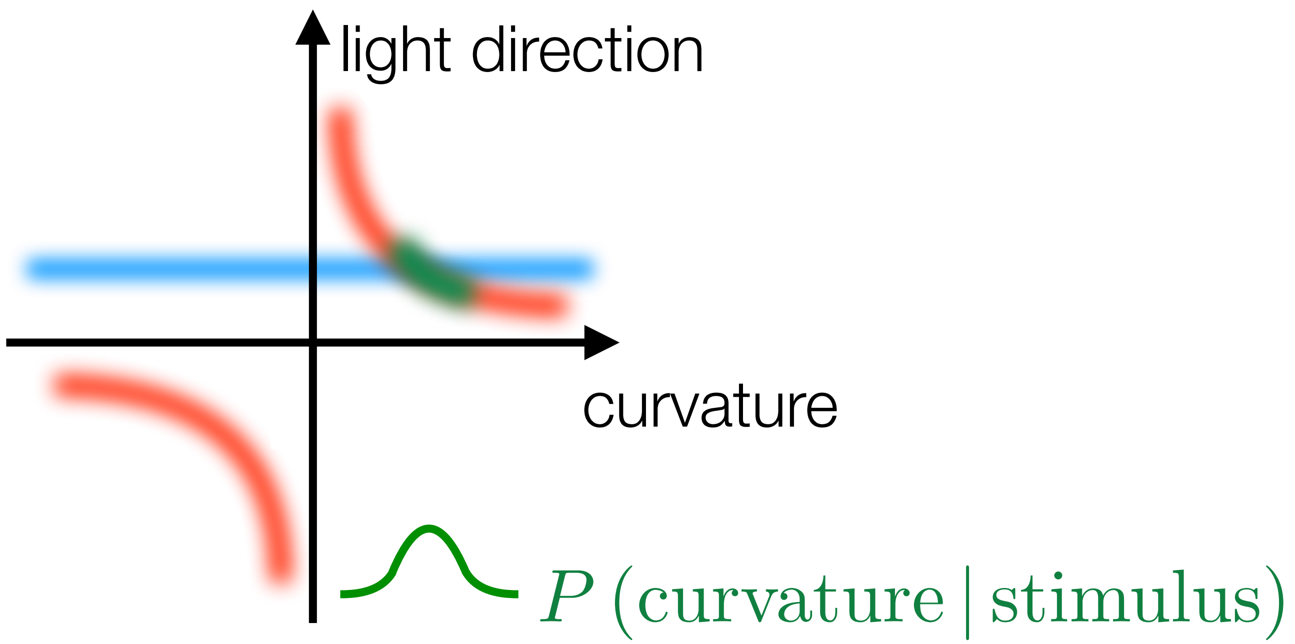
Mathematical challenges

Marginalization



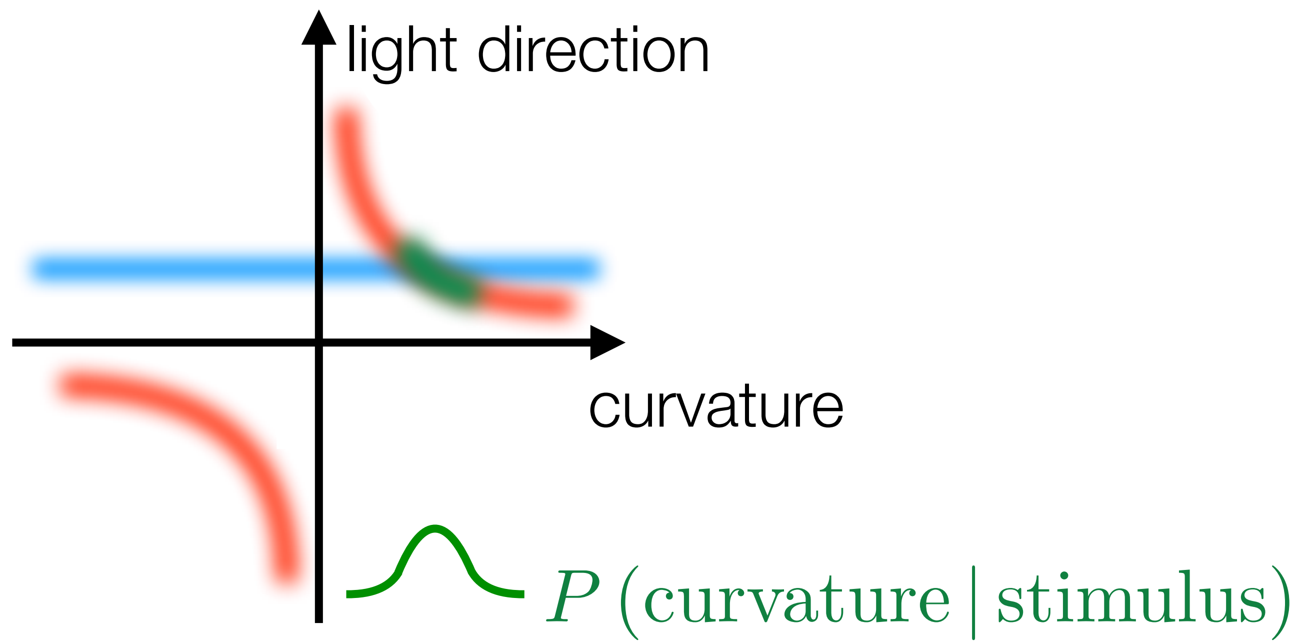
Mathematical challenges

Marginalization



Mathematical challenges

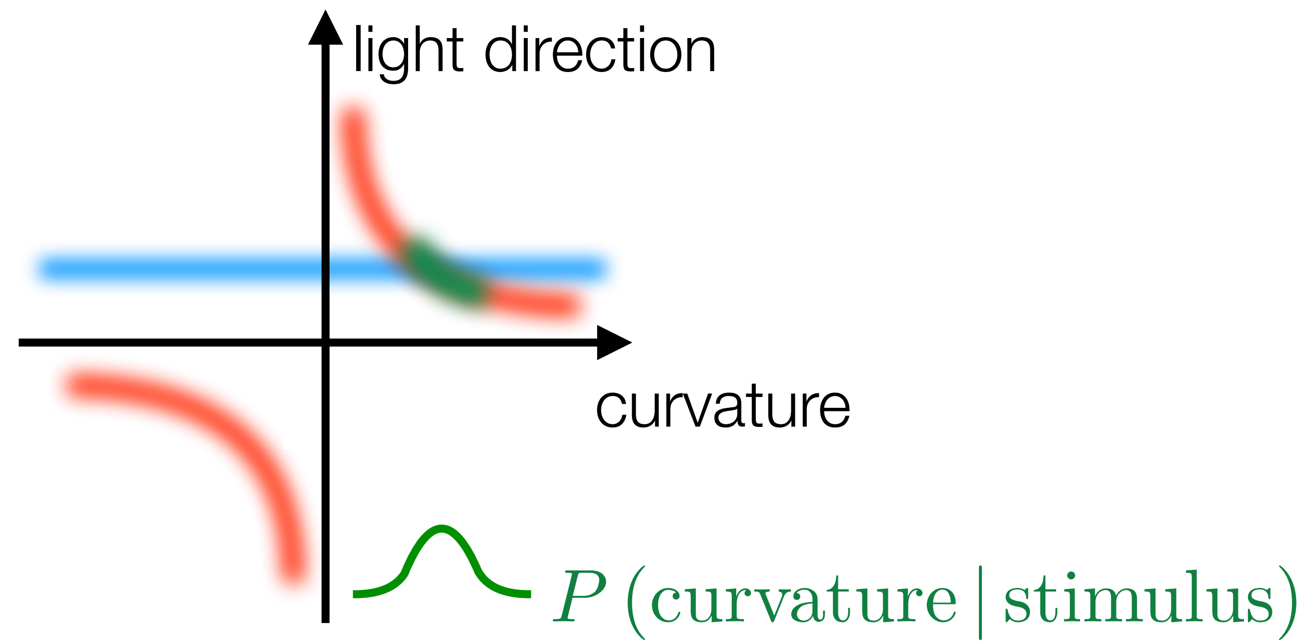
Marginalization



$$\text{P(curvature}\,|\,\text{stimulus)} = \int \text{P(curvature, light direction}\,|\,\text{stimulus)}\; d\text{light direction}$$

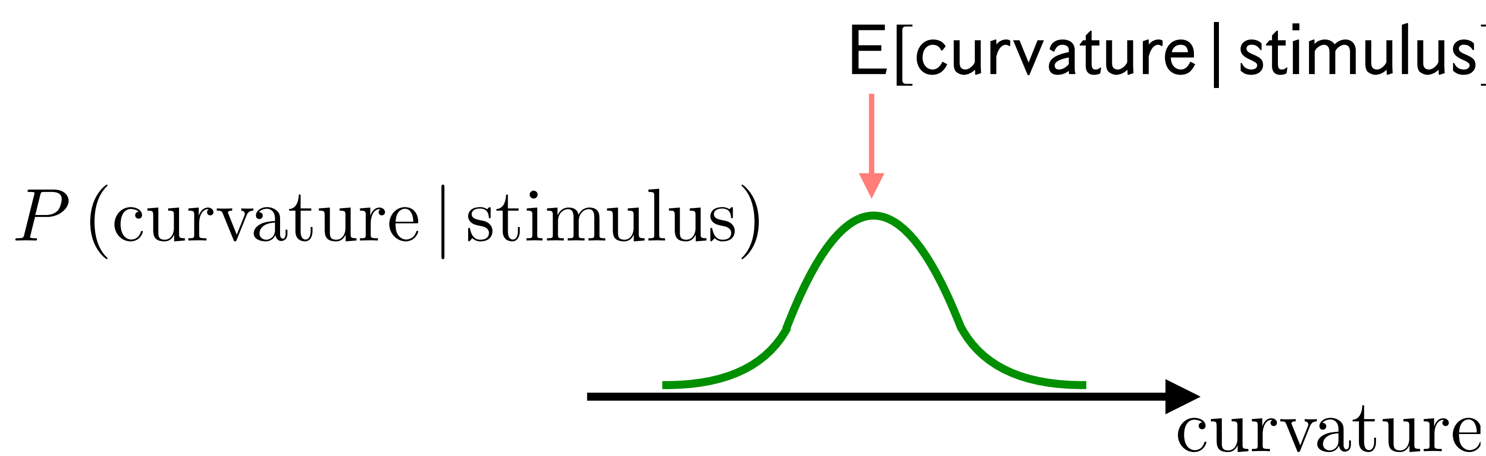
Mathematical challenges

Marginalization



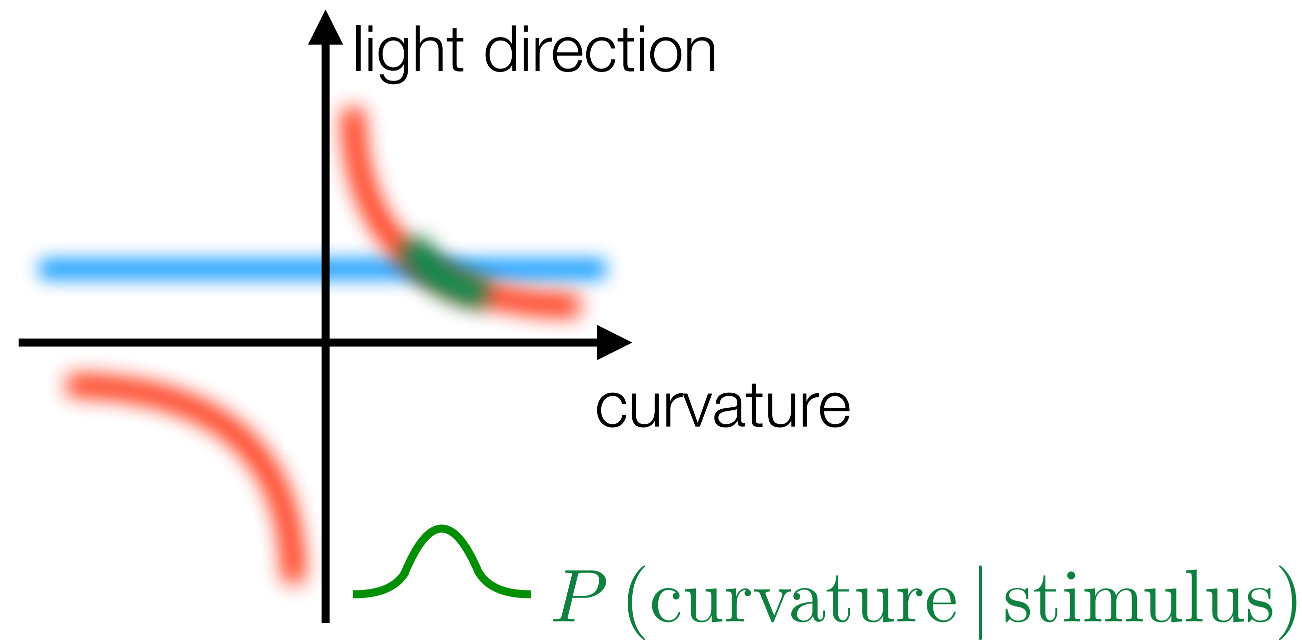
$$\text{P(curvature | stimulus)} = \int \text{P(curvature, light direction | stimulus)}\ d\text{light direction}$$

Expected value



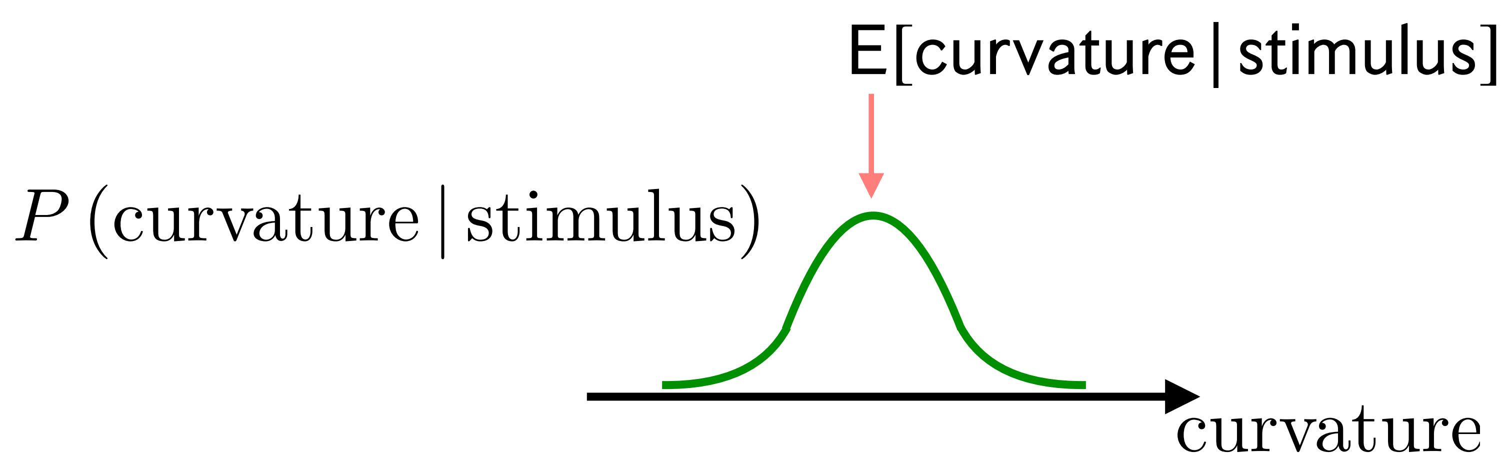
Mathematical challenges

Marginalization



$$\text{P(curvature}\,|\,\text{stimulus)} = \int \text{P(curvature, light direction}\,|\,\text{stimulus)}\;d\text{light direction}$$

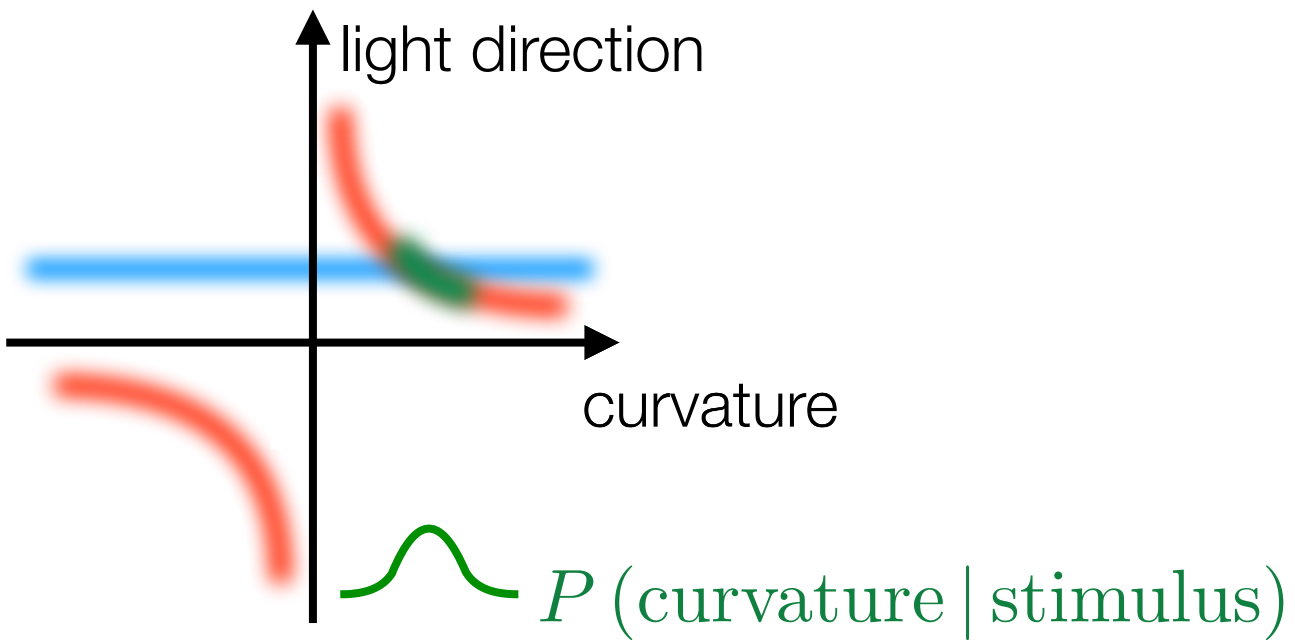
Expected value



$$\text{E[curvature}\,|\,\text{stimulus]} = \int \text{curvature P(curvature}\,|\,\text{stimulus)}\;d\text{curvature}$$

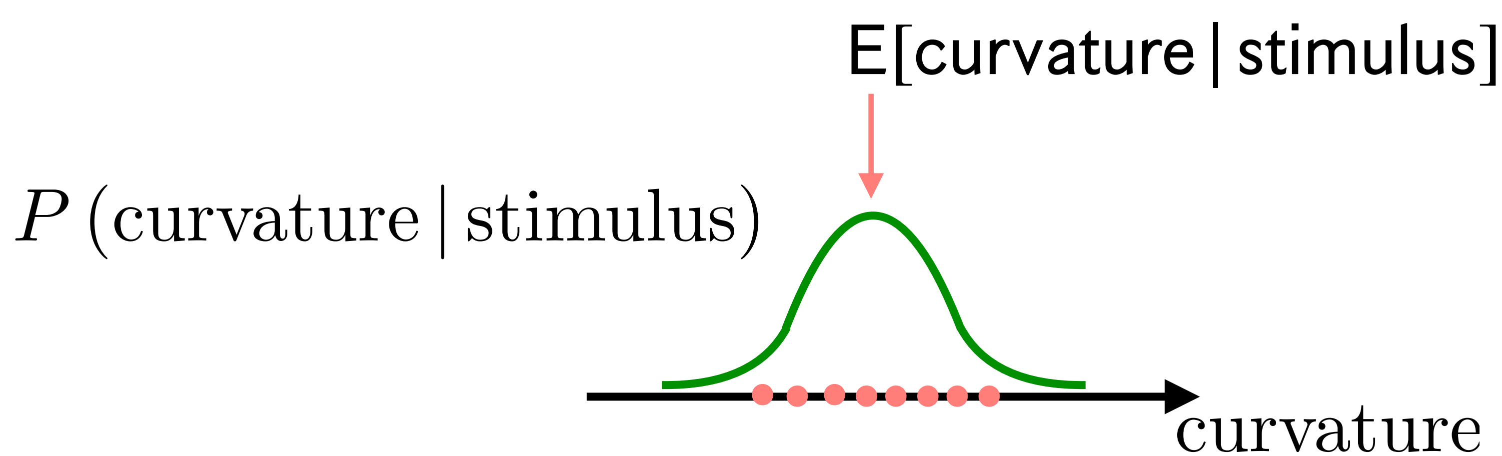
Mathematical challenges

Marginalization



$$\text{P(curvature}\,|\,\text{stimulus)} = \int \text{P(curvature, light direction}\,|\,\text{stimulus)}\; d\text{light direction}$$

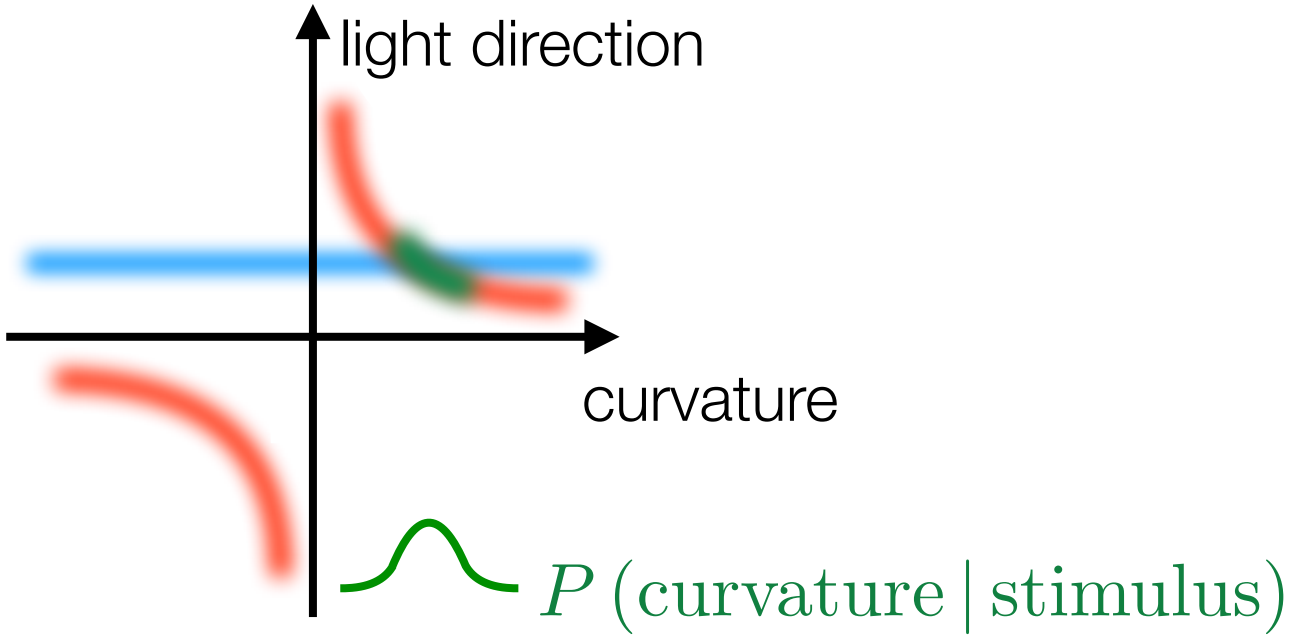
Expected value



$$\text{E[curvature}\,|\,\text{stimulus]} = \int \text{curvature P(curvature}\,|\,\text{stimulus)}\; d\text{curvature}$$

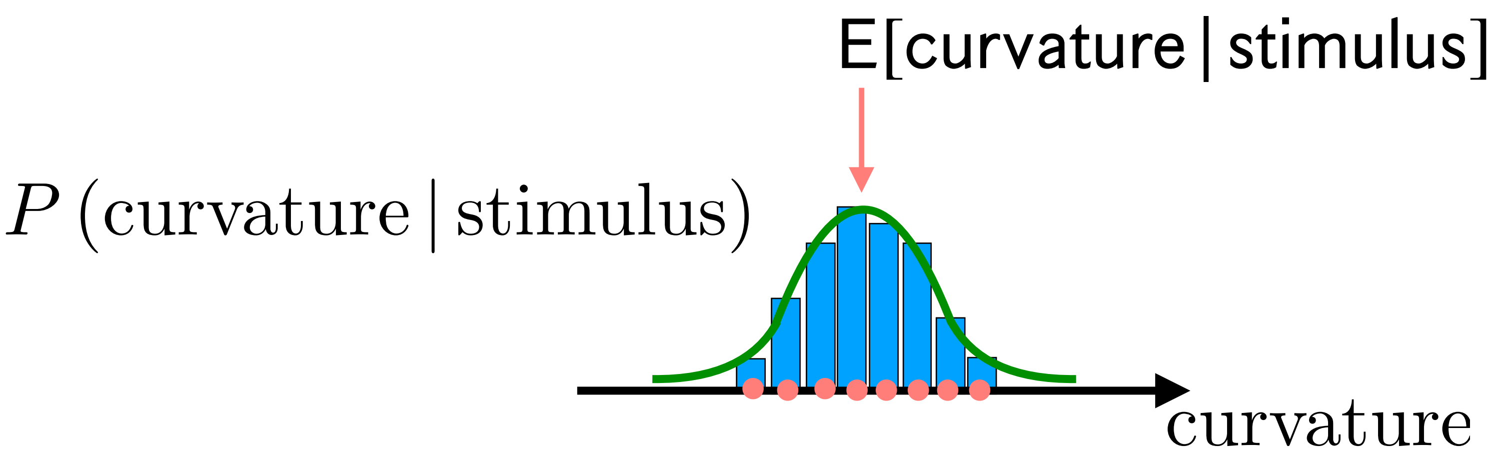
Mathematical challenges

Marginalization



$$\text{P(curvature}\,|\,\text{stimulus}) = \int \text{P(curvature, light direction}\,|\,\text{stimulus})\ d\text{light direction}$$

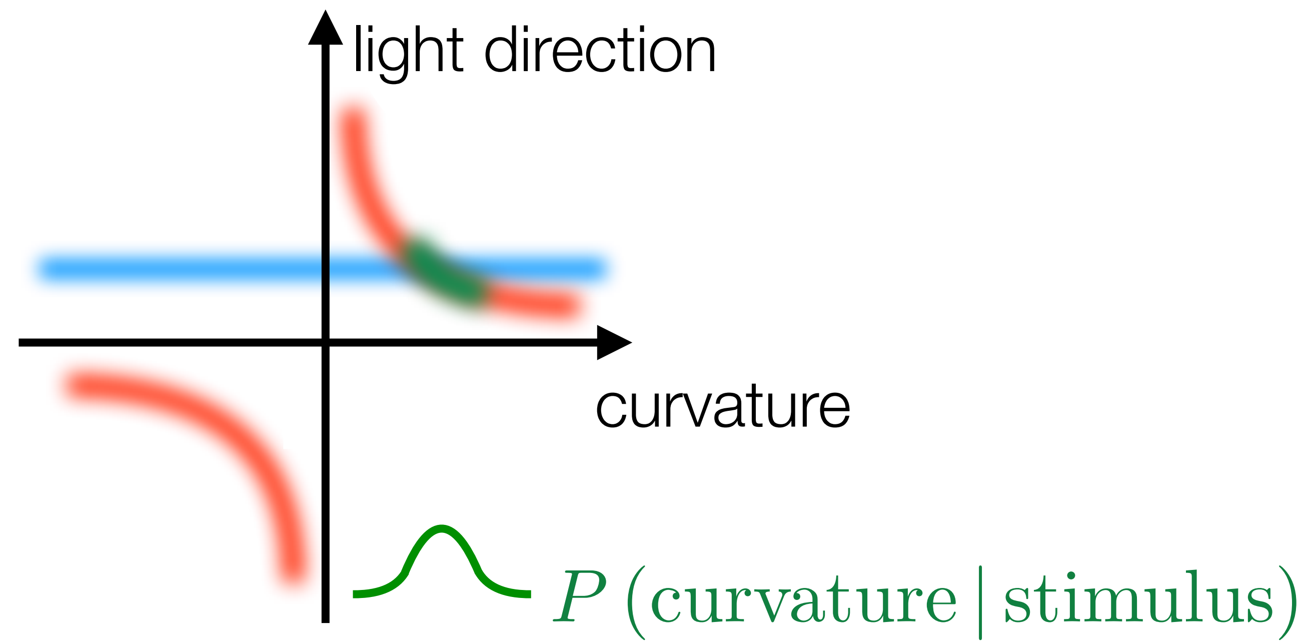
Expected value



$$\text{E[curvature}\,|\,\text{stimulus]} = \int \text{curvature P(curvature}\,|\,\text{stimulus})\ d\text{curvature}$$

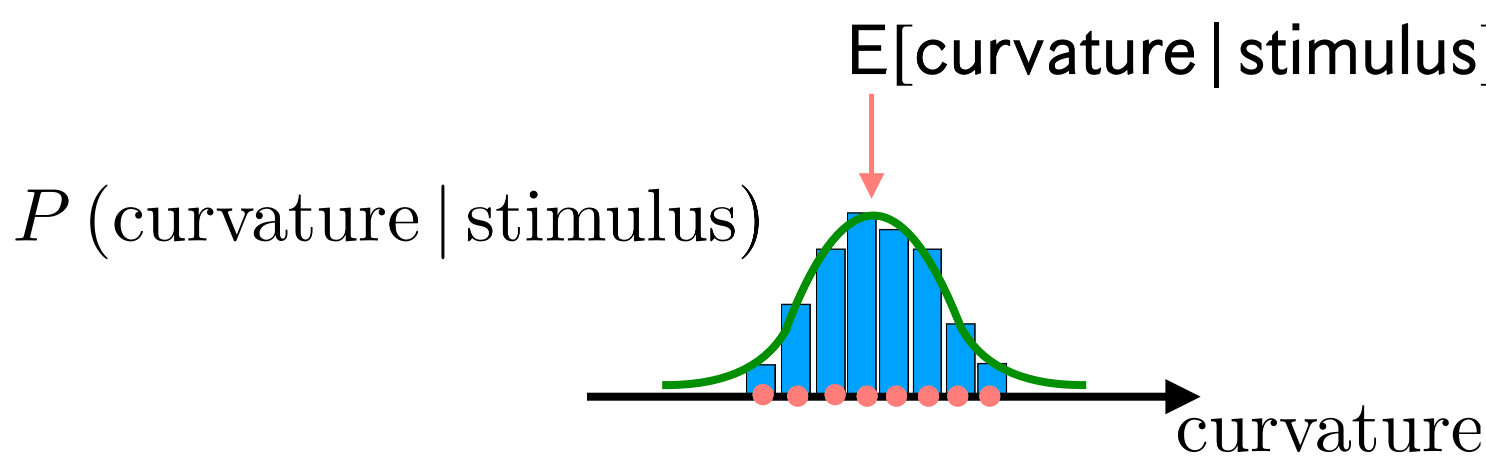
Mathematical challenges

Marginalization



$$\text{P(curvature} \,|\, \text{stimulus)} = \int \text{P(curvature, light direction} \,|\, \text{stimulus)} \; d\text{light direction}$$

Expected value

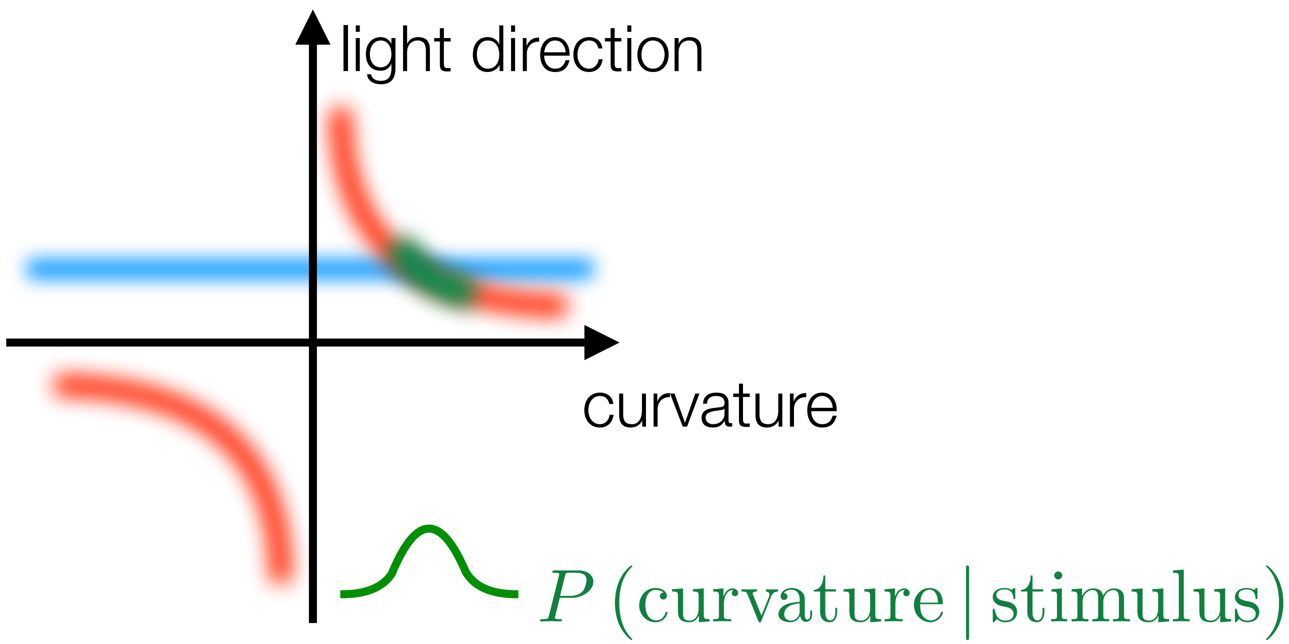


$$\text{E[curvature} \,|\, \text{stimulus]} = \int \text{curvature P(curvature} \,|\, \text{stimulus)} \; d\text{curvature}$$

OK, if P( ) is a Normal distribution (Gaussian)

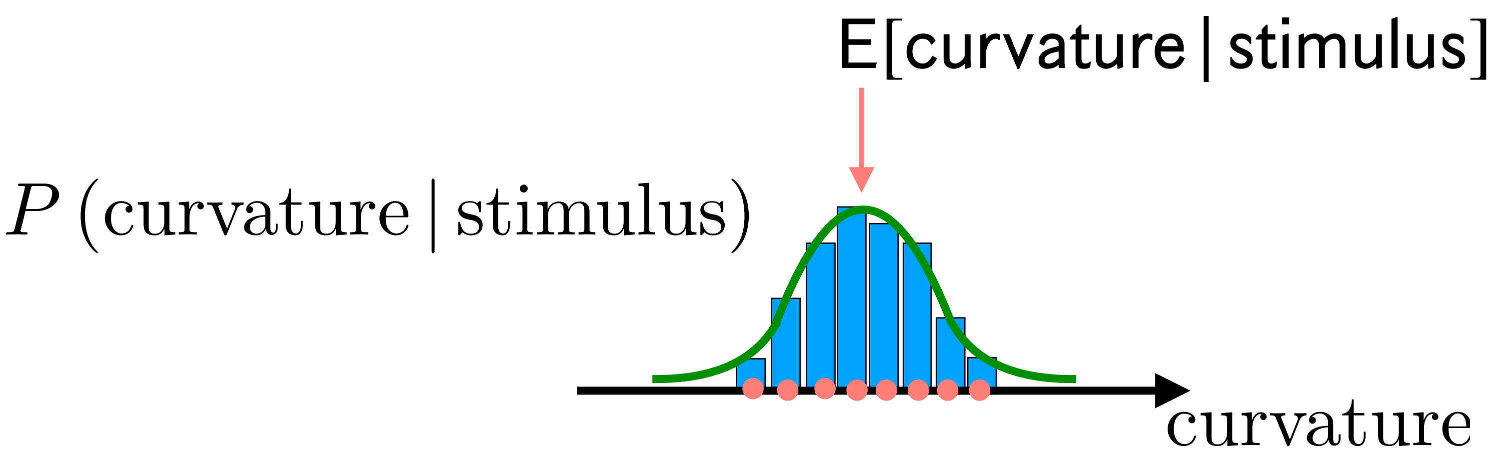
Mathematical challenges

Marginalization



$$\text{P(curvature | stimulus)} = \int \text{P(curvature, light direction | stimulus)}\; d\text{light direction}$$

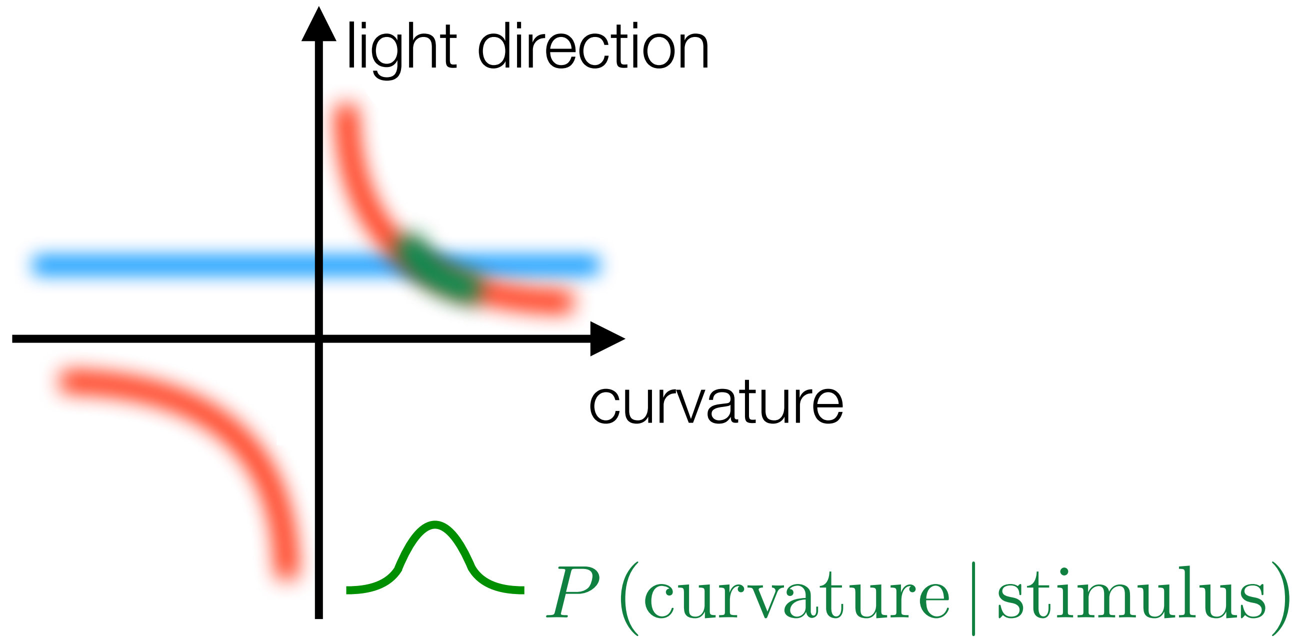
Expected value



$$\text{E[curvature | stimulus]} = \int \text{curvature P(curvature | stimulus)}\; d\text{curvature}$$

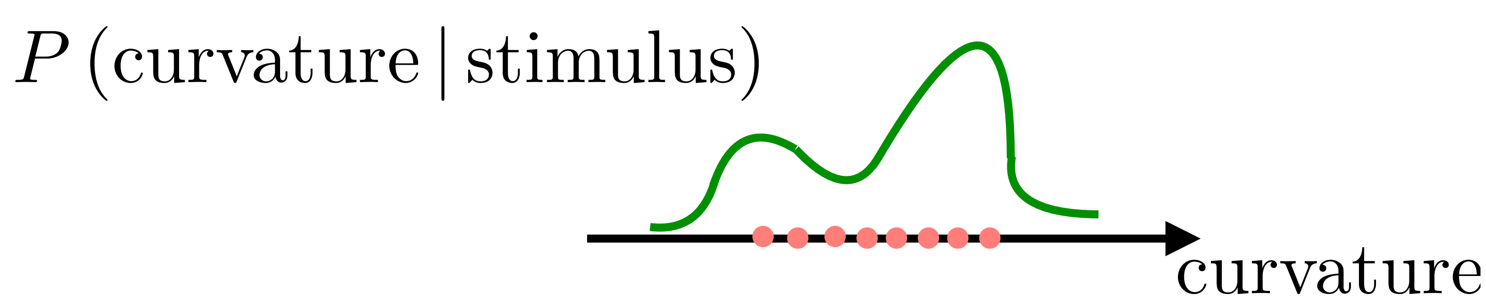
Mathematical challenges

Marginalization



$$\mathrm{P(curvature\,|\,stimulus)} = \int \mathrm{P(curvature, light\ direction\,|\,stimulus)}\ d\mathrm{light\ direction}$$

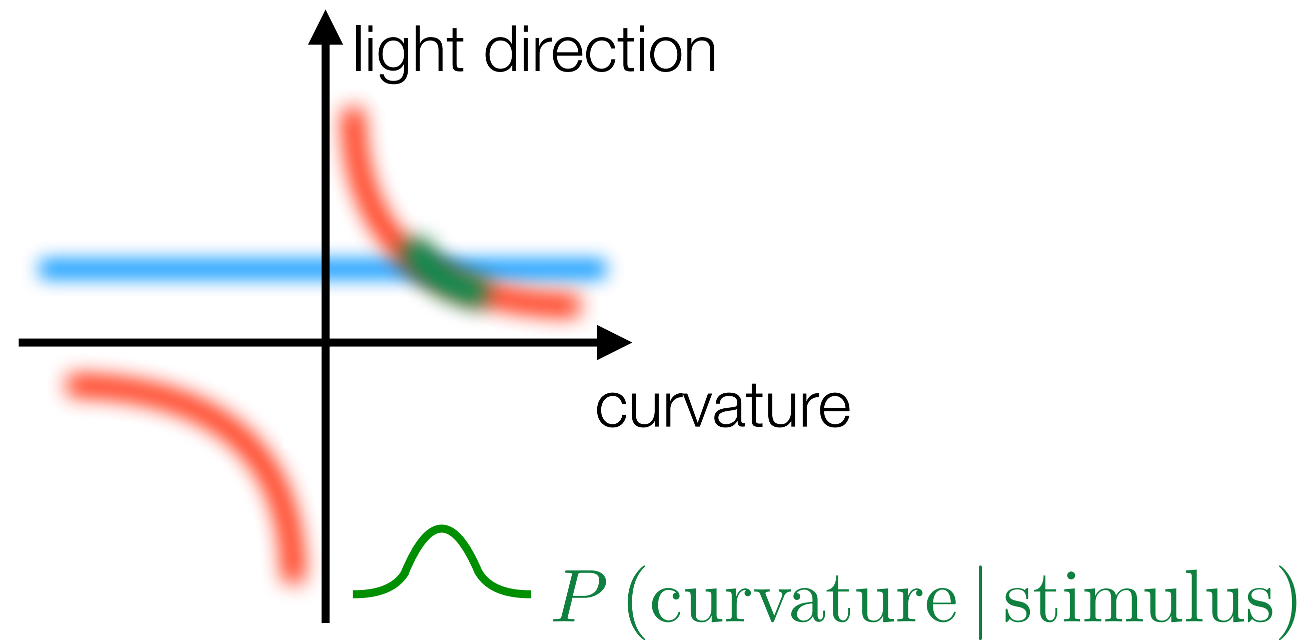
Expected value



$$\mathrm{E[curvature\,|\,stimulus]} = \int \mathrm{curvature\ P(curvature\,|\,stimulus)}\ d\mathrm{curvature}$$

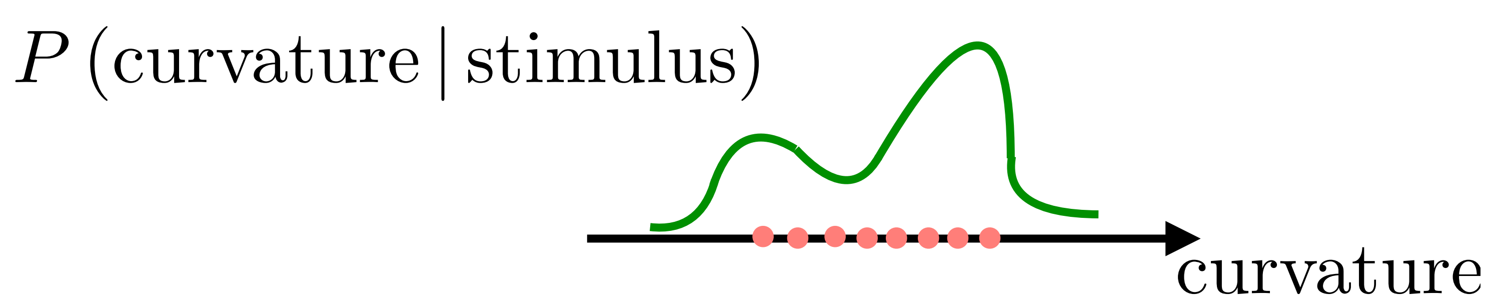
Mathematical challenges

Marginalization



$$\text{P(curvature}\,|\,\text{stimulus)} = \int \text{P(curvature, light direction}\,|\,\text{stimulus)}\; d\text{light direction}$$

Expected value

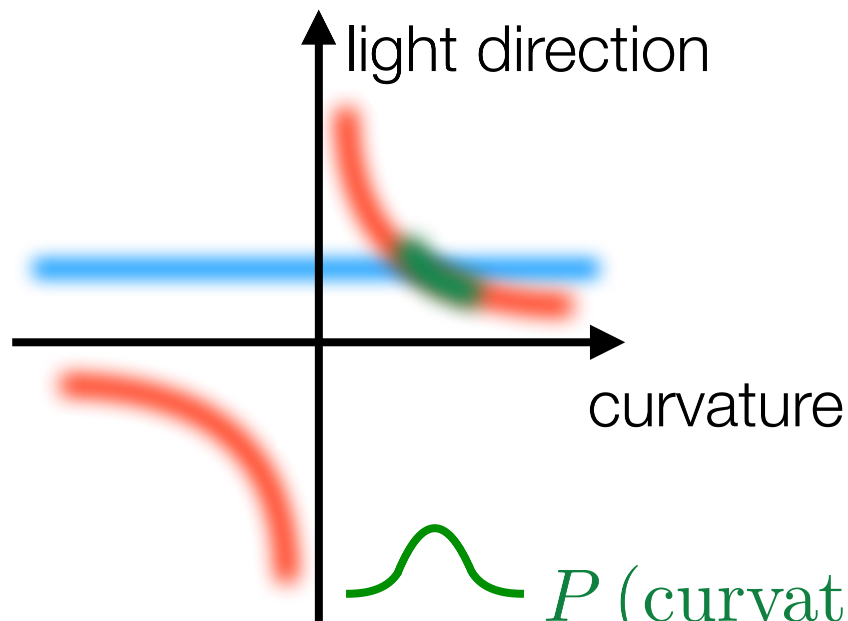


$$\text{E[curvature}\,|\,\text{stimulus]} = \int \text{curvature}\;\text{P(curvature}\,|\,\text{stimulus)}\; d\text{curvature}$$

$$\approx \sum_i \text{curvature}_i \qquad \text{such that curvature}_i \sim \text{P(curvature}\,|\,\text{stimulus)}$$

Mathematical challenges

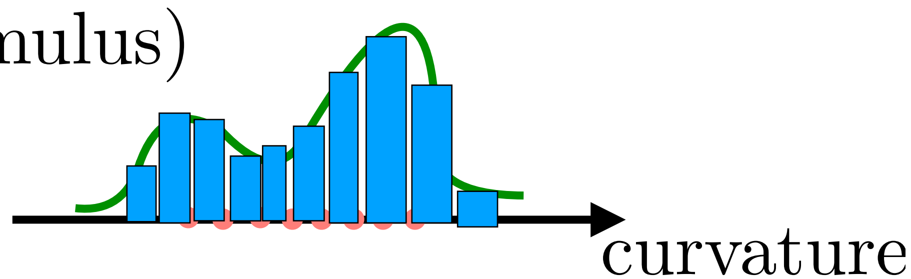
Marginalization



$$\text{P(curvature}\,|\,\text{stimulus)} = \int \text{P(curvature, light direction}\,|\,\text{stimulus)}\; d\text{light direction}$$

Expected value

$P\,(\text{curvature}\,|\,\text{stimulus})$



$$\text{E[curvature}\,|\,\text{stimulus]} = \int \text{curvature P(curvature}\,|\,\text{stimulus)}\; d\text{curvature}$$

$$\approx \sum_i \text{curvature}_i \qquad \text{such that curvature}_i \sim \text{P(curvature}\,|\,\text{stimulus)}$$

Possible remedies

Integral is intractable (a.k.a. impossible), approximation is needed

Possible remedies

Integral is intractable (a.k.a. impossible), approximation is needed

1. point estimation
   i. optimisation
   ii. Expectation Maximization

Possible remedies

Integral is intractable (a.k.a. impossible), approximation is needed

1. point estimation
   1. optimisation
   2. Expectation Maximization
2. variational approximation:
   1. pretending P( ) is a Normal distribution;
   2. find the best Normal distribution
   3. calculate the integral

Possible remedies

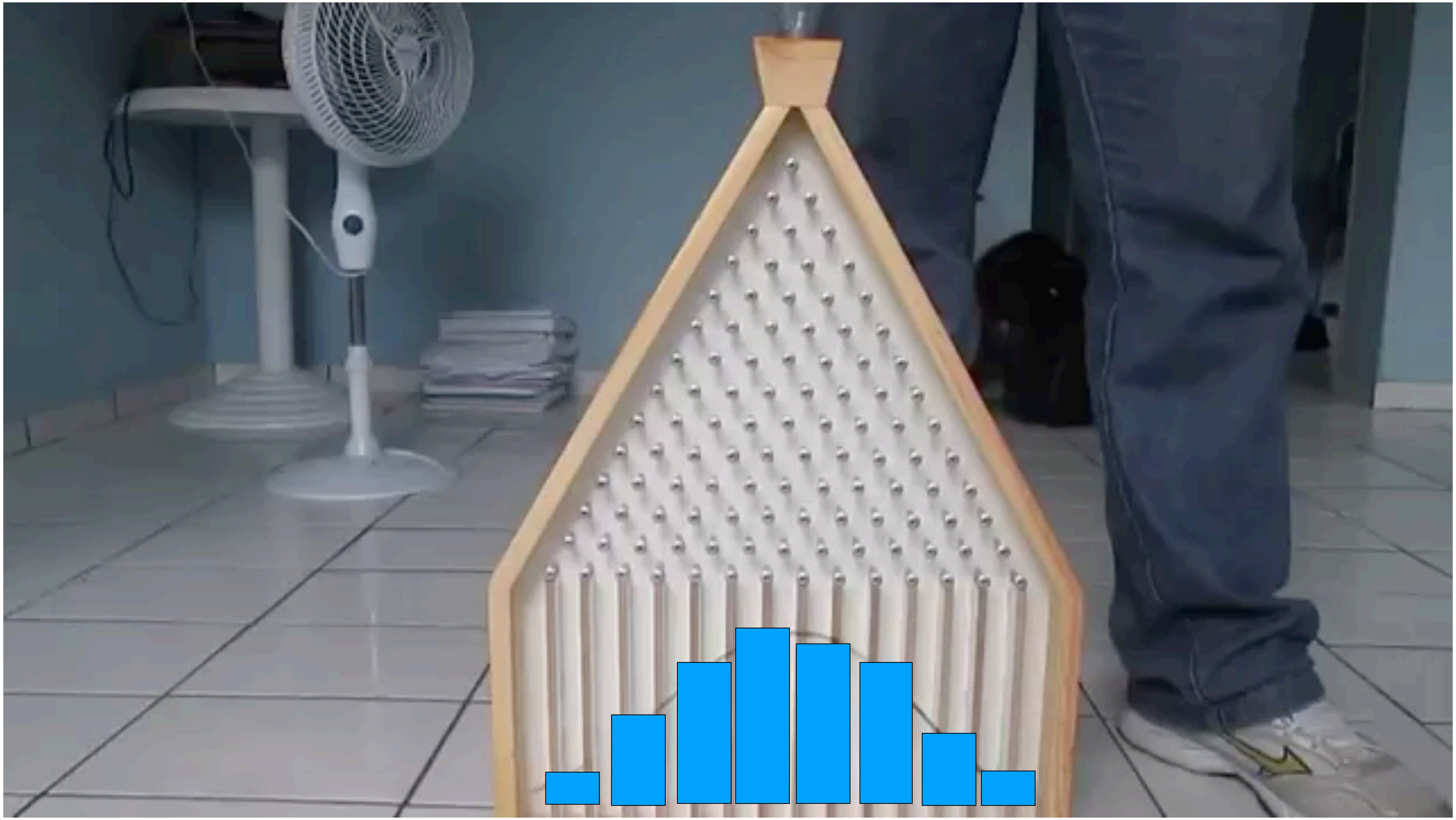
Integral is intractable (a.k.a. impossible), approximation is needed

1. point estimation
   i. optimisation
   ii. Expectation Maximization
2. variational approximation:
   a. pretending P( ) is a Normal distribution;
   b. find the best Normal distribution
   c. calculate the integral
3. sampling (Monte Carlo methods)

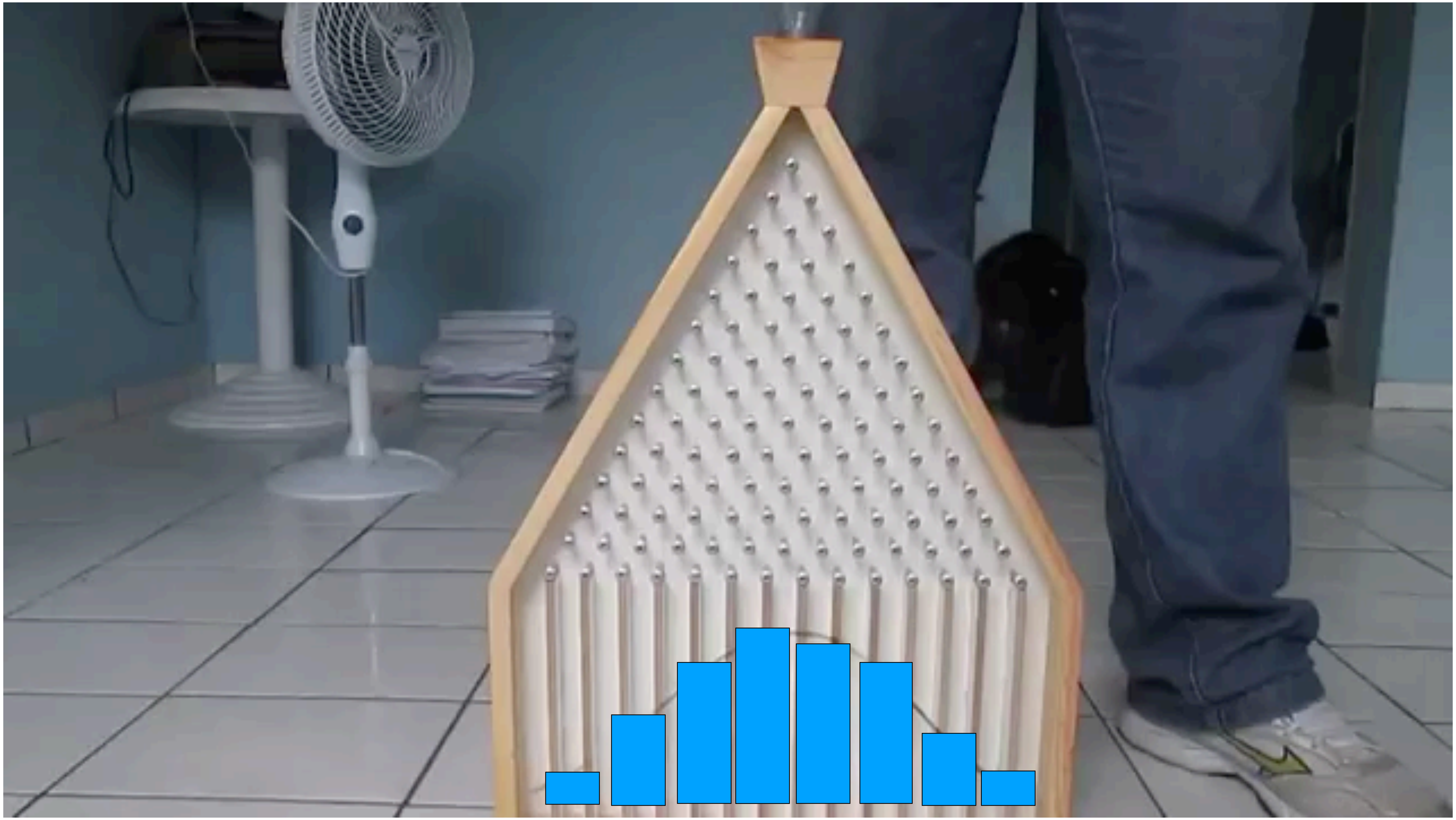
Sampling

Sampling

Sampling

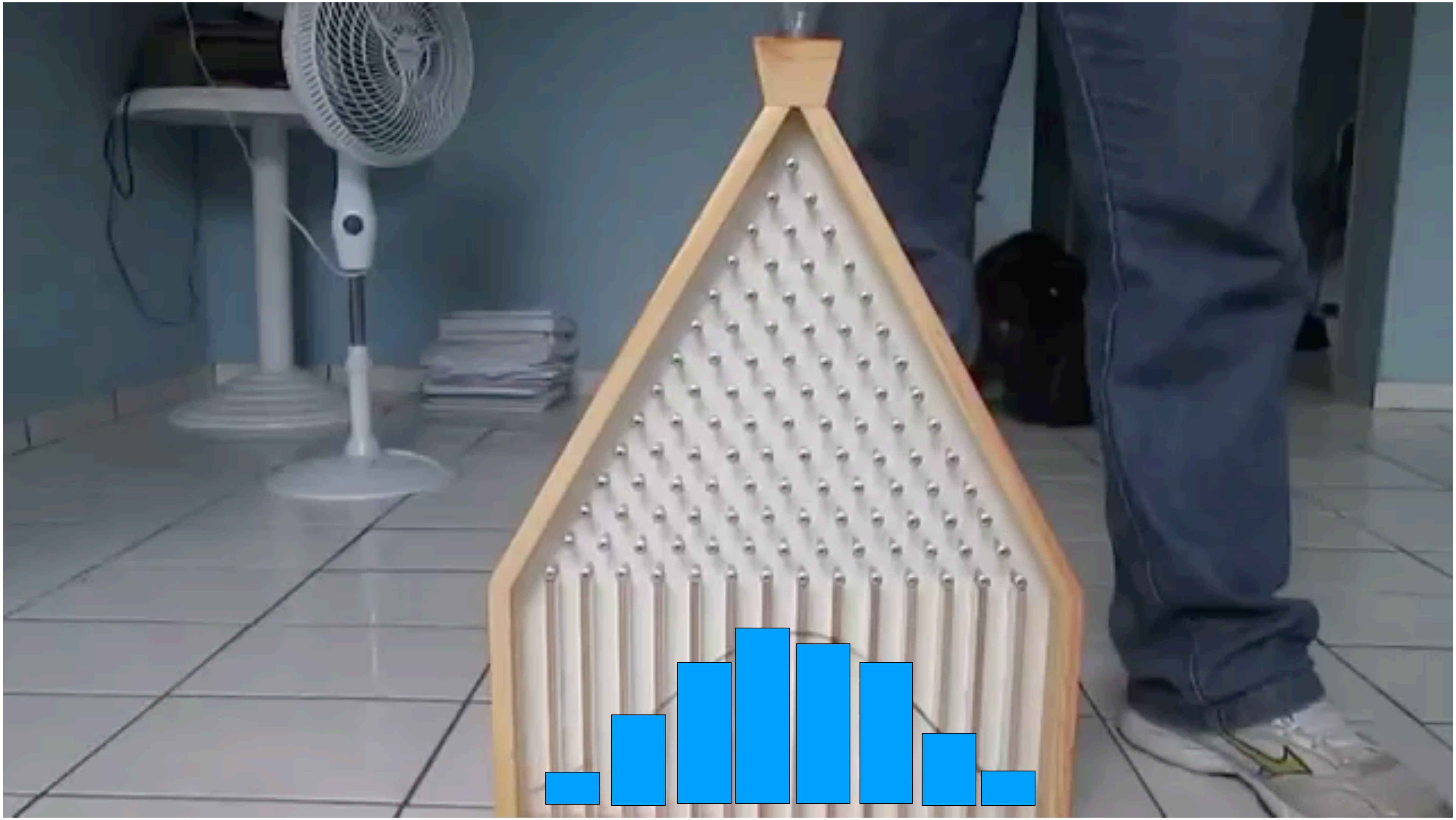


Sampling



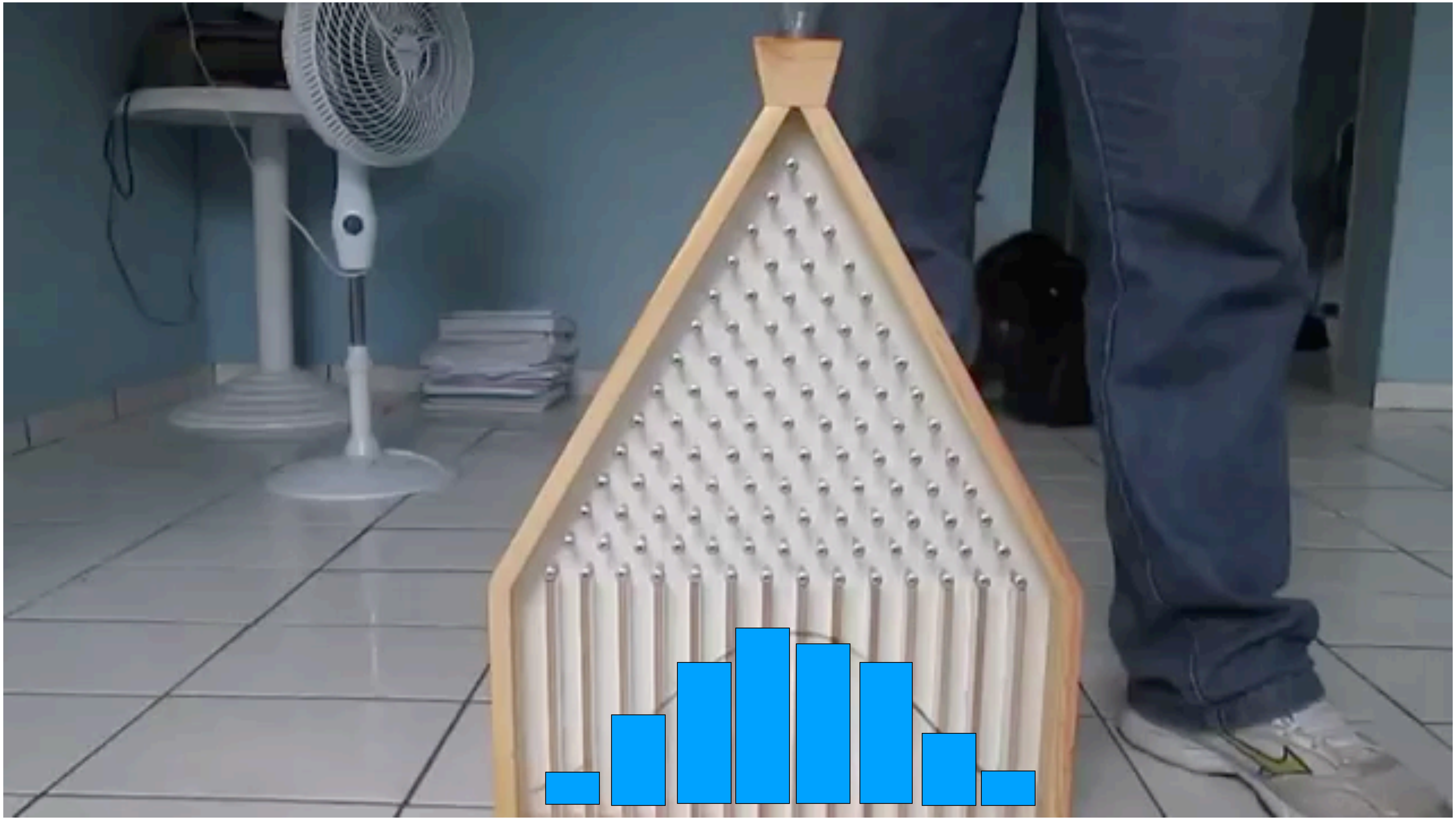
- balls are ‘examples’ from the distribution

Sampling



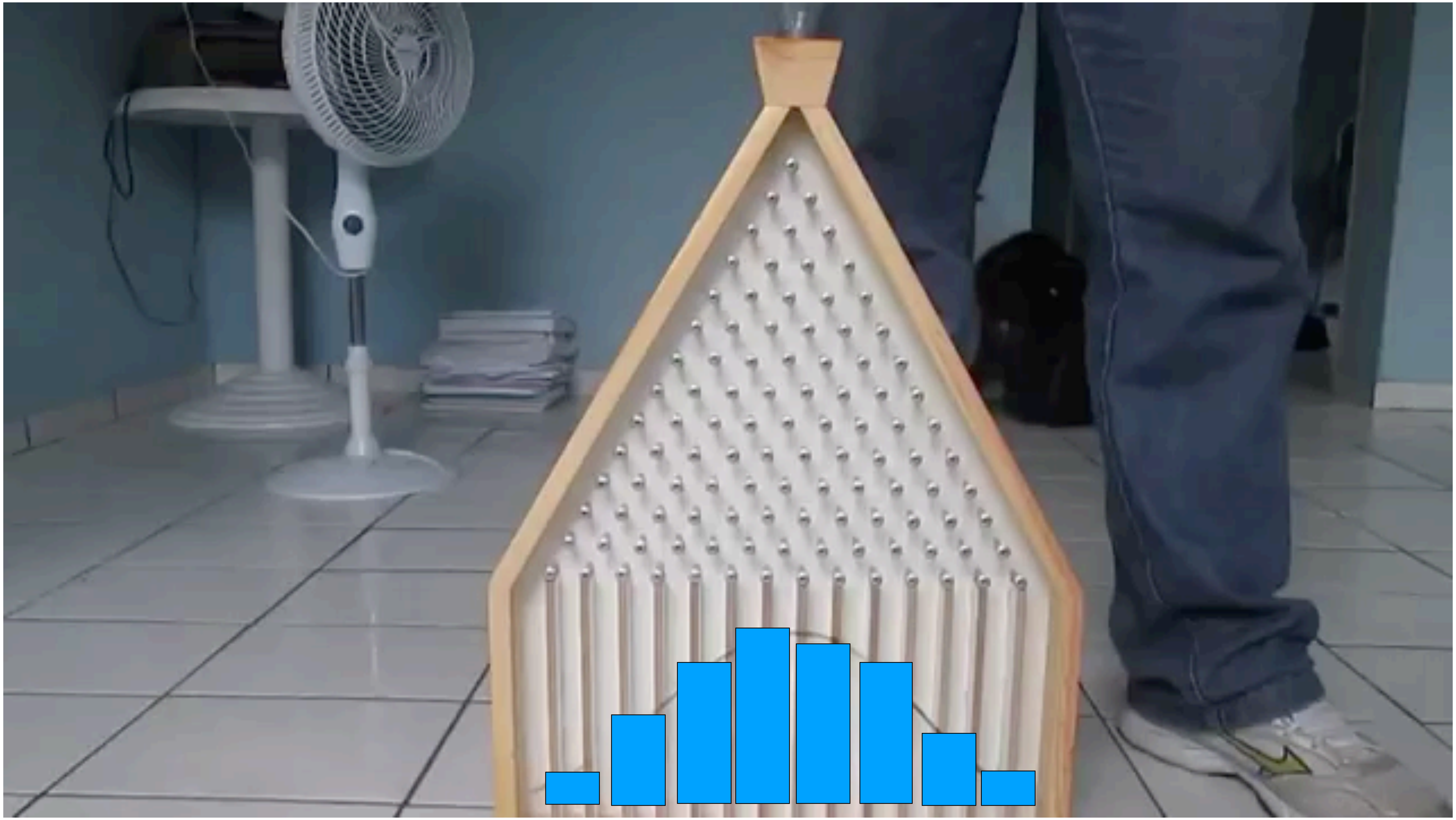
- balls are ‘examples’ from the distribution
- the proportion of balls at different possible positions is proportional to the distribution

Sampling



- balls are 'examples' from the distribution
- the proportion of balls at different possible positions is proportional to the distribution
- skimming through these examples we can approximate the distribution

Sampling



- balls are ‘examples’ from the distribution
- the proportion of balls at different possible positions is proportional to the distribution
- skimming through these examples we can approximate the distribution
- (one can think of building a histogram instead of specifying the parameters of a distribution)

Sampling methods

- Assumption: we can access a scaled version of the probability distribution: $P^*(x) = c\,P(x)$

- Motivation: inferring the posterior with Bayes rule:

$$P(x \mid \text{Data}) = \frac{P(\text{Data} \mid x)P(x)}{P(\text{Data})} \propto c \cdot P(\text{Data} \mid x)P(x)$$

Sampling methods

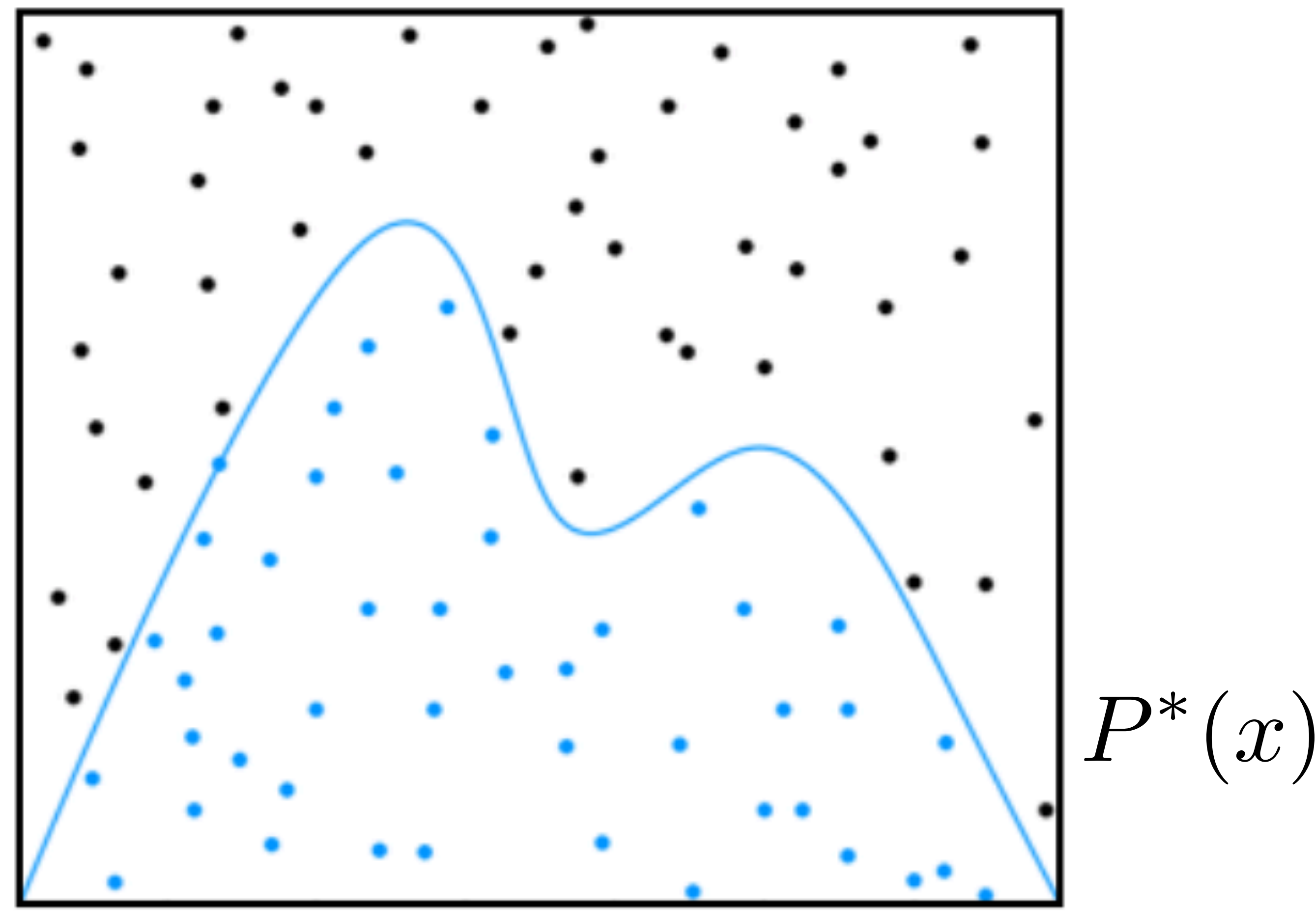
- Assumption: we can access a scaled version of the probability distribution: $P^*(x) = c\,P(x)$

- Motivation: inferring the posterior with Bayes rule:

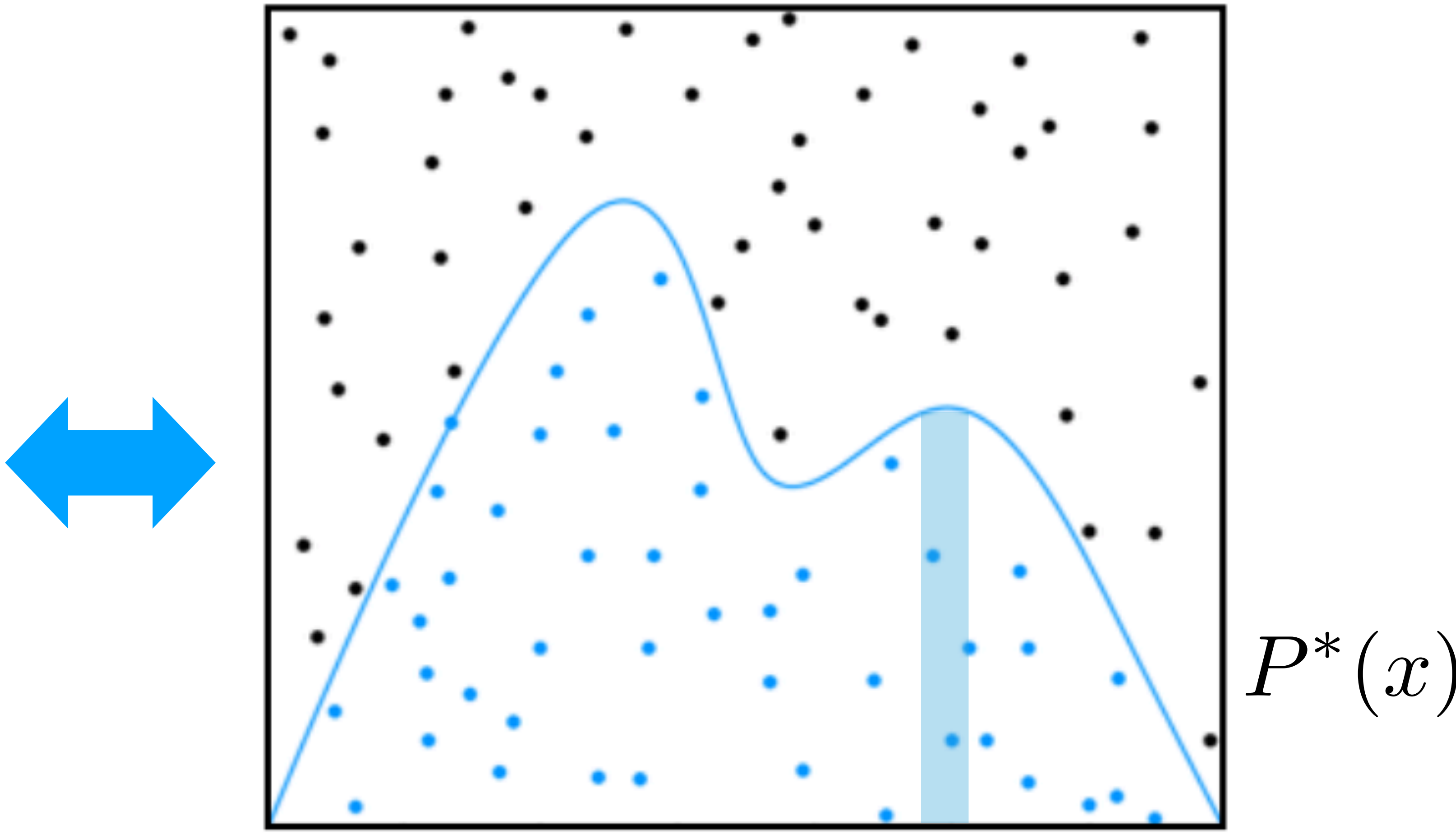
$$P(x\,|\,\text{Data}) = \frac{P(\text{Data}\,|\,x)P(x)}{P(\text{Data})} \propto c \cdot P(\text{Data}\,|\,x)P(x)$$

marginal distribution —
invokes complicated integrals,
costly to calculate

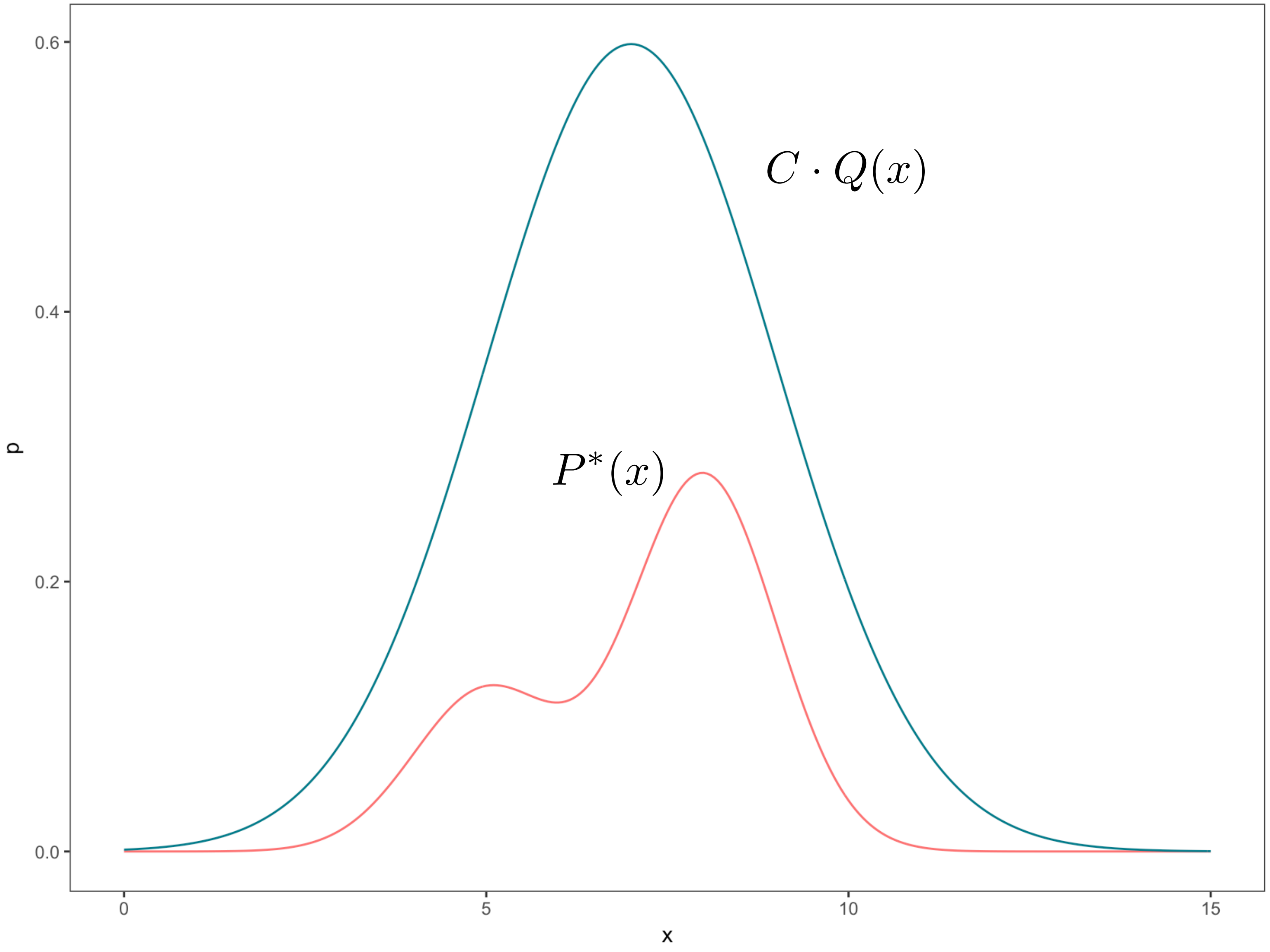
Rejection sampling



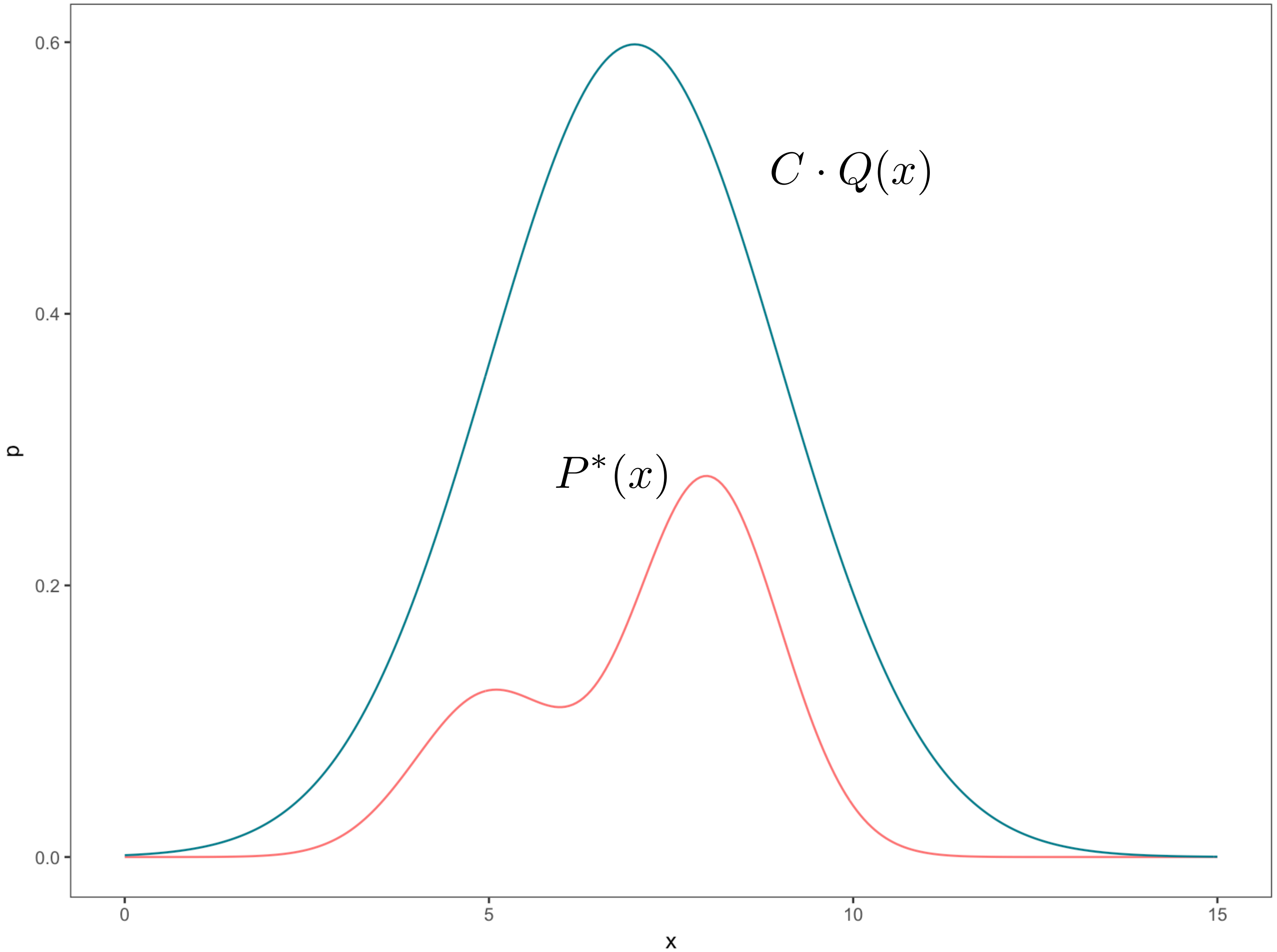
Rejection sampling



Rejection sampling



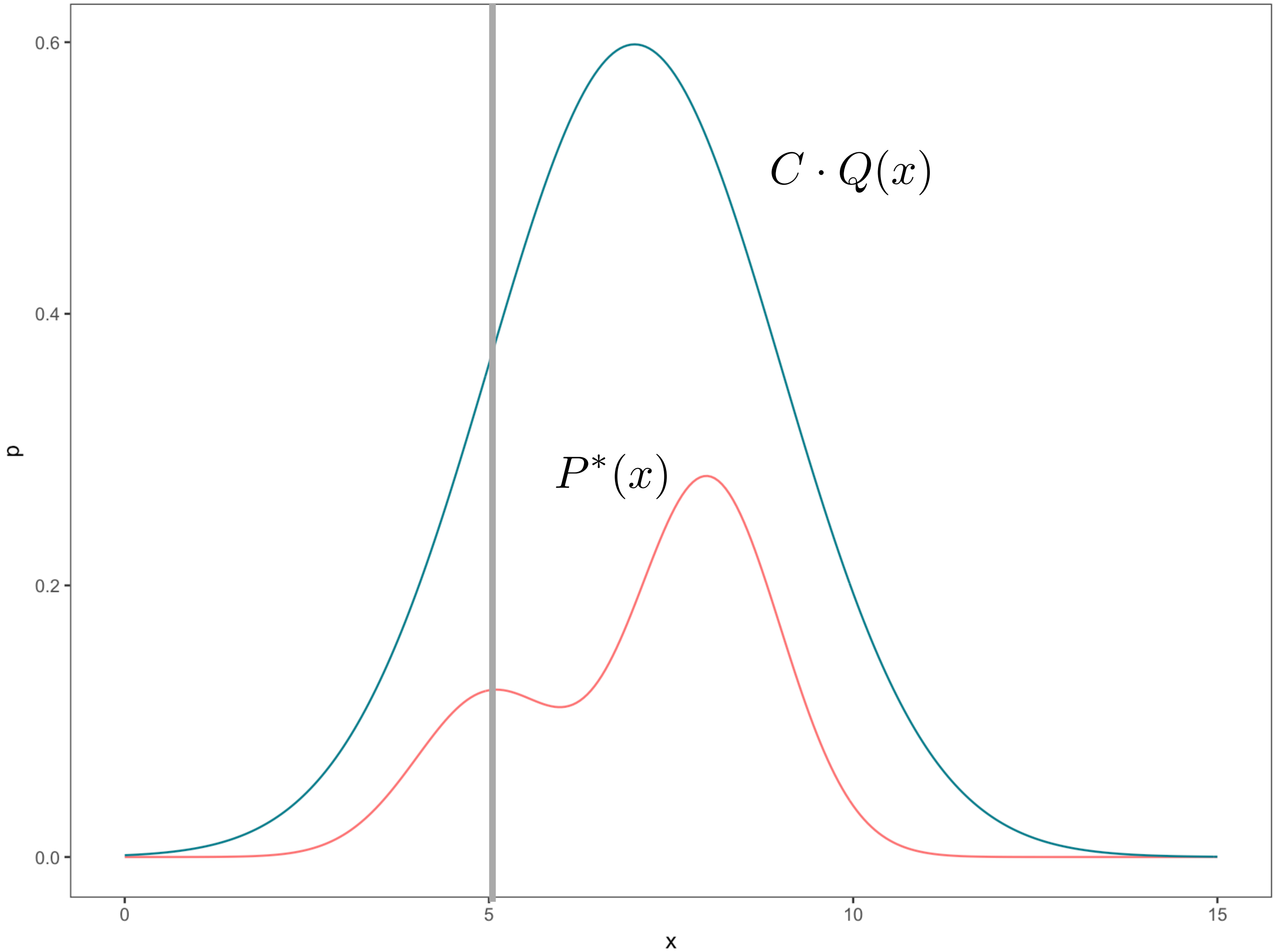
Rejection sampling



- the proposal density, Q(x), is a distribution from which we can obtain samples (have a random generator for it, e.g. Gaussian)

$$x \sim Q(x)$$

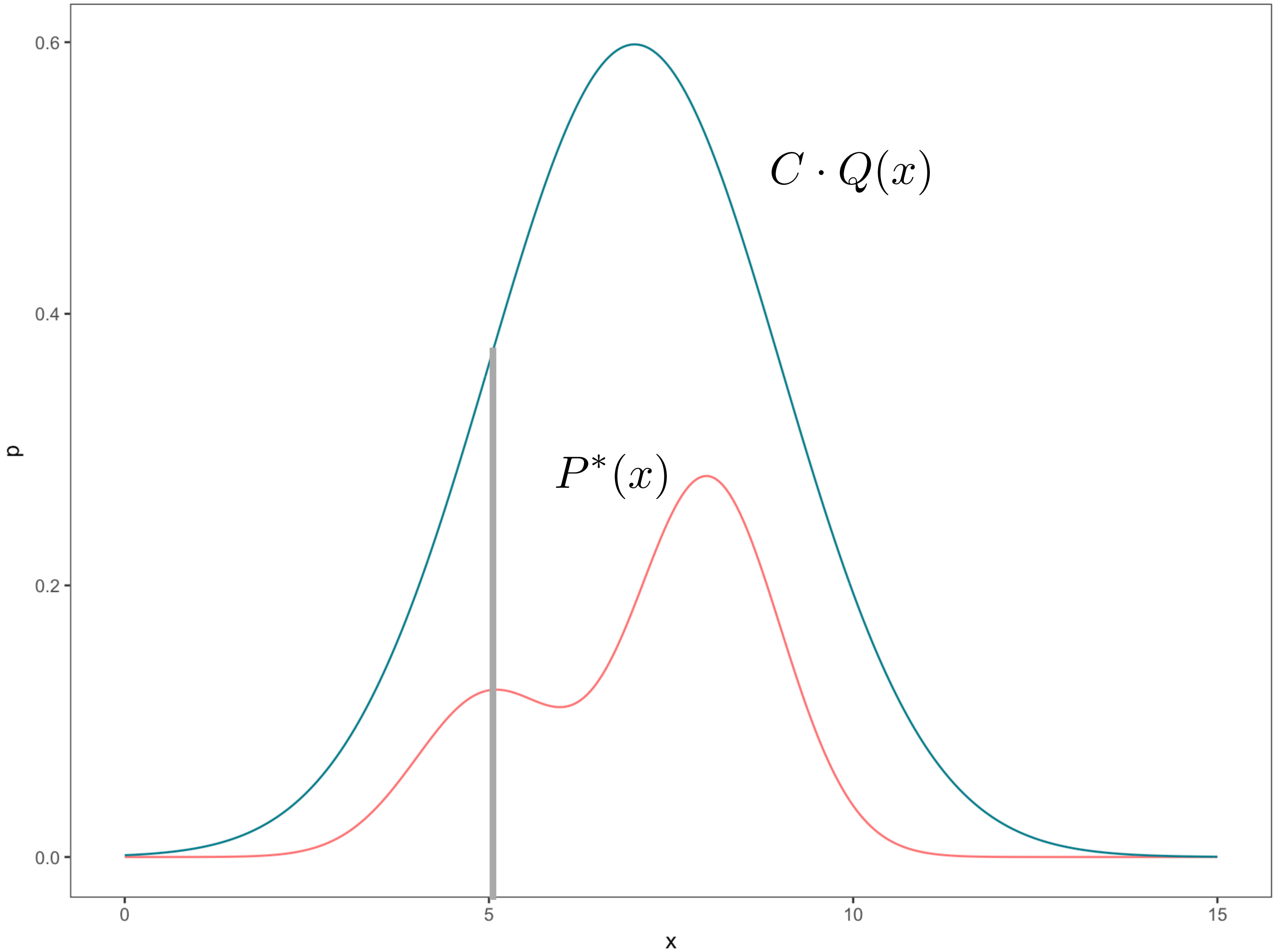
Rejection sampling



- the proposal density, Q(x), is a distribution from which we can obtain samples (have a random generator for it, e.g. Gaussian)

$$x \sim Q(x)$$

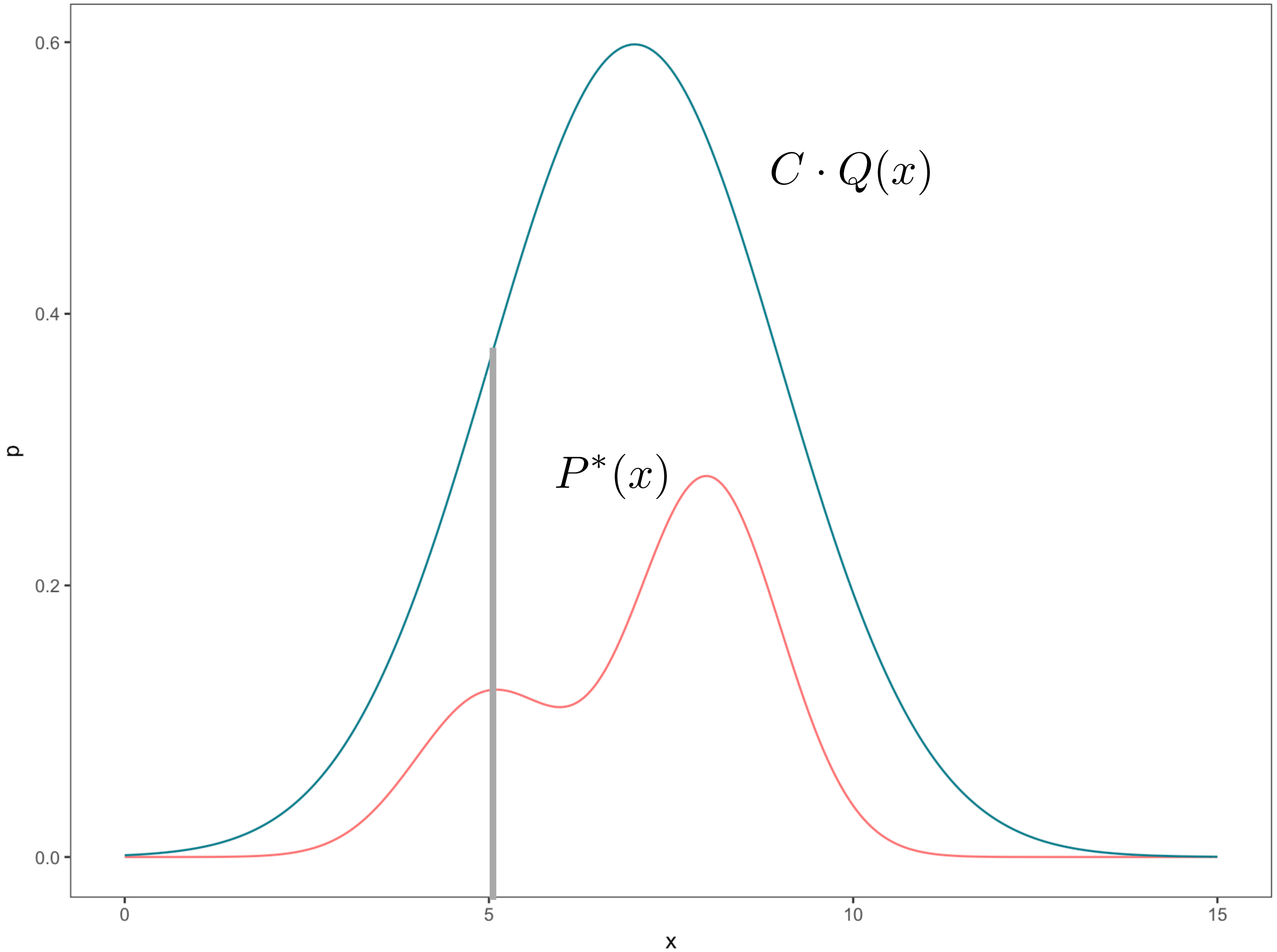
Rejection sampling



- the proposal density, Q(x), is a distribution from which we can obtain samples (have a random generator for it, e.g. Gaussian)

$$x \sim Q(x)$$

Rejection sampling



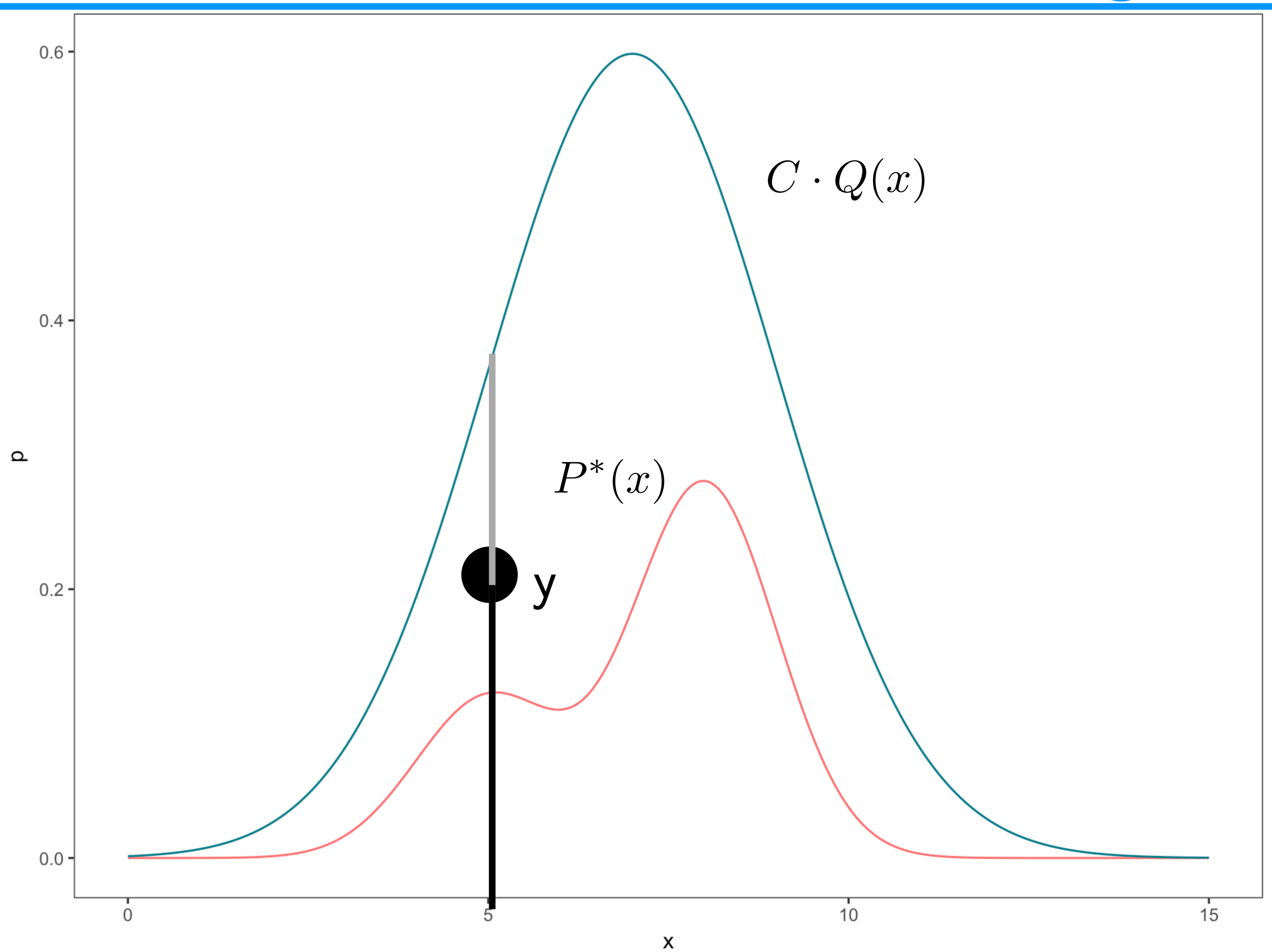
- the proposal density, Q(x), is a distribution from which we can obtain samples (have a random generator for it, e.g. Gaussian)

$$x \sim Q(x)$$

- a point along the vertical axis is sampled between 0 and the Q(x) is a distribution from which we can obtain samples (have a random generator for it, e.g. Gaussian)

$$y \sim \text{uniform}(0, cQ(x))$$

Rejection sampling



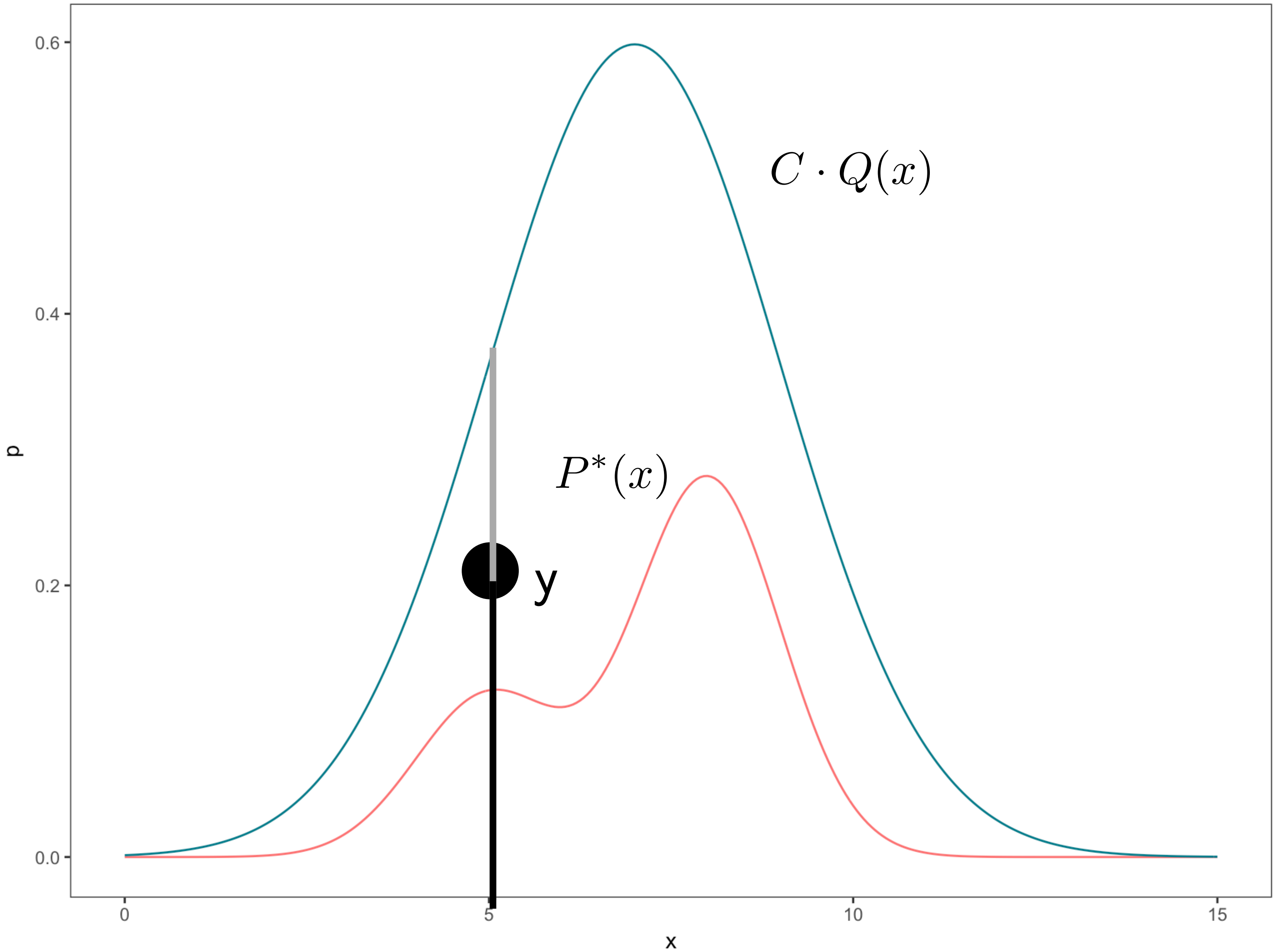
- the proposal density, Q(x), is a distribution from which we can obtain samples (have a random generator for it, e.g. Gaussian)

$$x \sim Q(x)$$

- a point along the vertical axis is sampled between 0 and the Q(x) is a distribution from which we can obtain samples (have a random generator for it, e.g. Gaussian)

$$y \sim \text{uniform}(0, cQ(x))$$

Rejection sampling



- the proposal density, Q(x), is a distribution from which we can obtain samples (have a random generator for it, e.g. Gaussian)

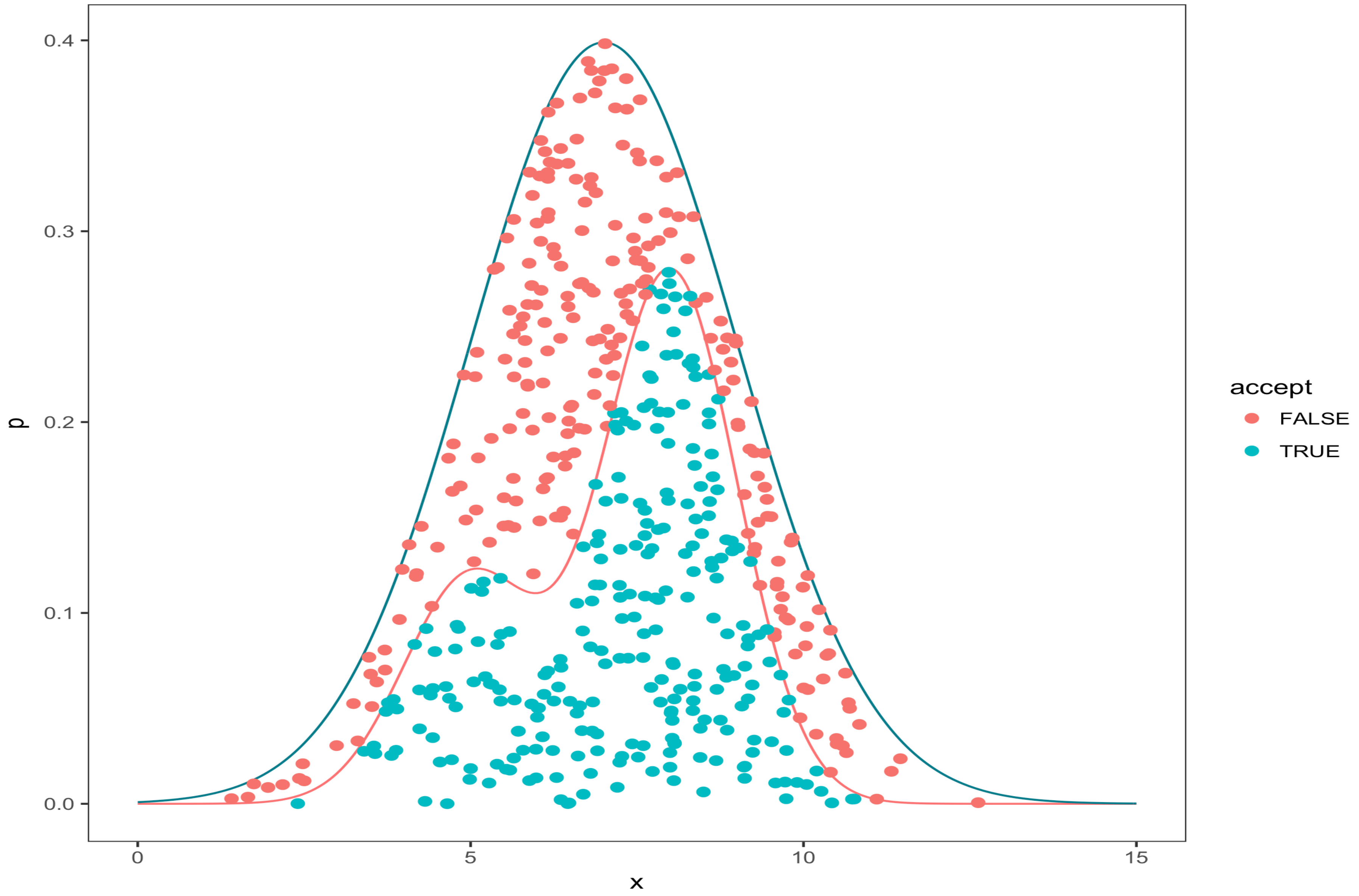
$$x \sim Q(x)$$

- a point along the vertical axis is sampled between 0 and the Q(x) is a distribution from which we can obtain samples (have a random generator for it, e.g. Gaussian)

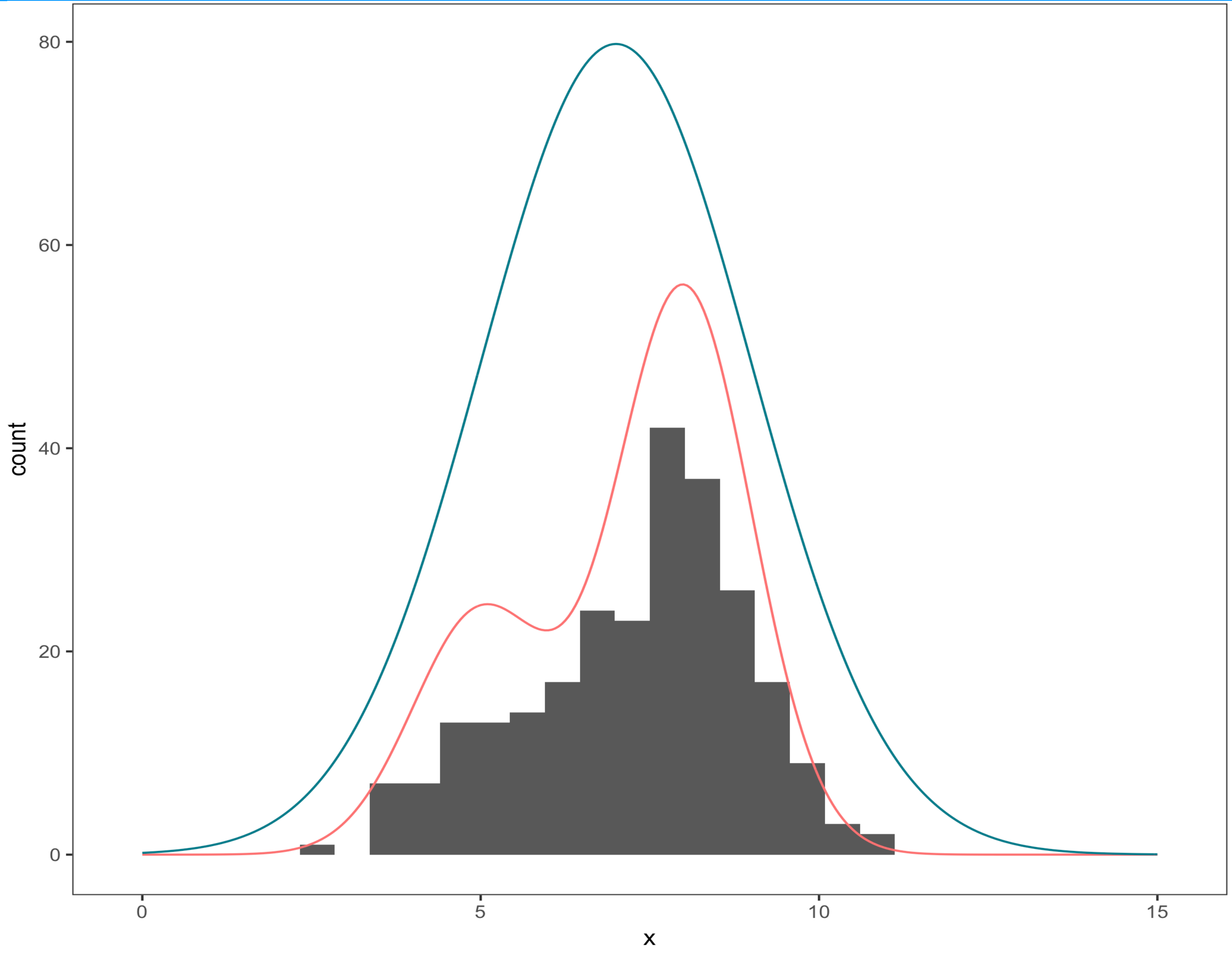
$$y \sim \text{uniform}(0, cQ(x))$$

- proposal is accepted if y is lower than P*(x)

Rejection sampling



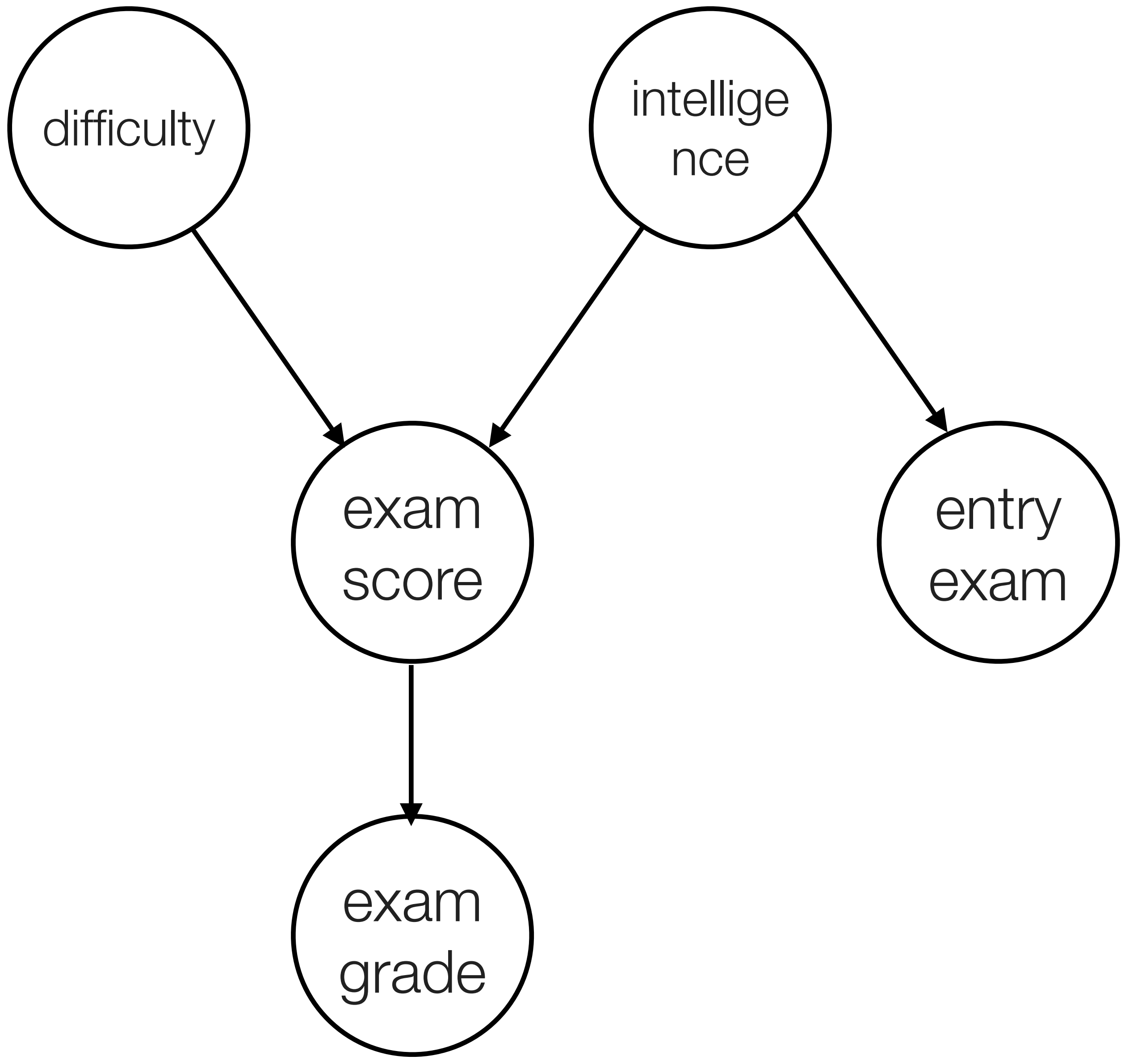
Rejection sampling



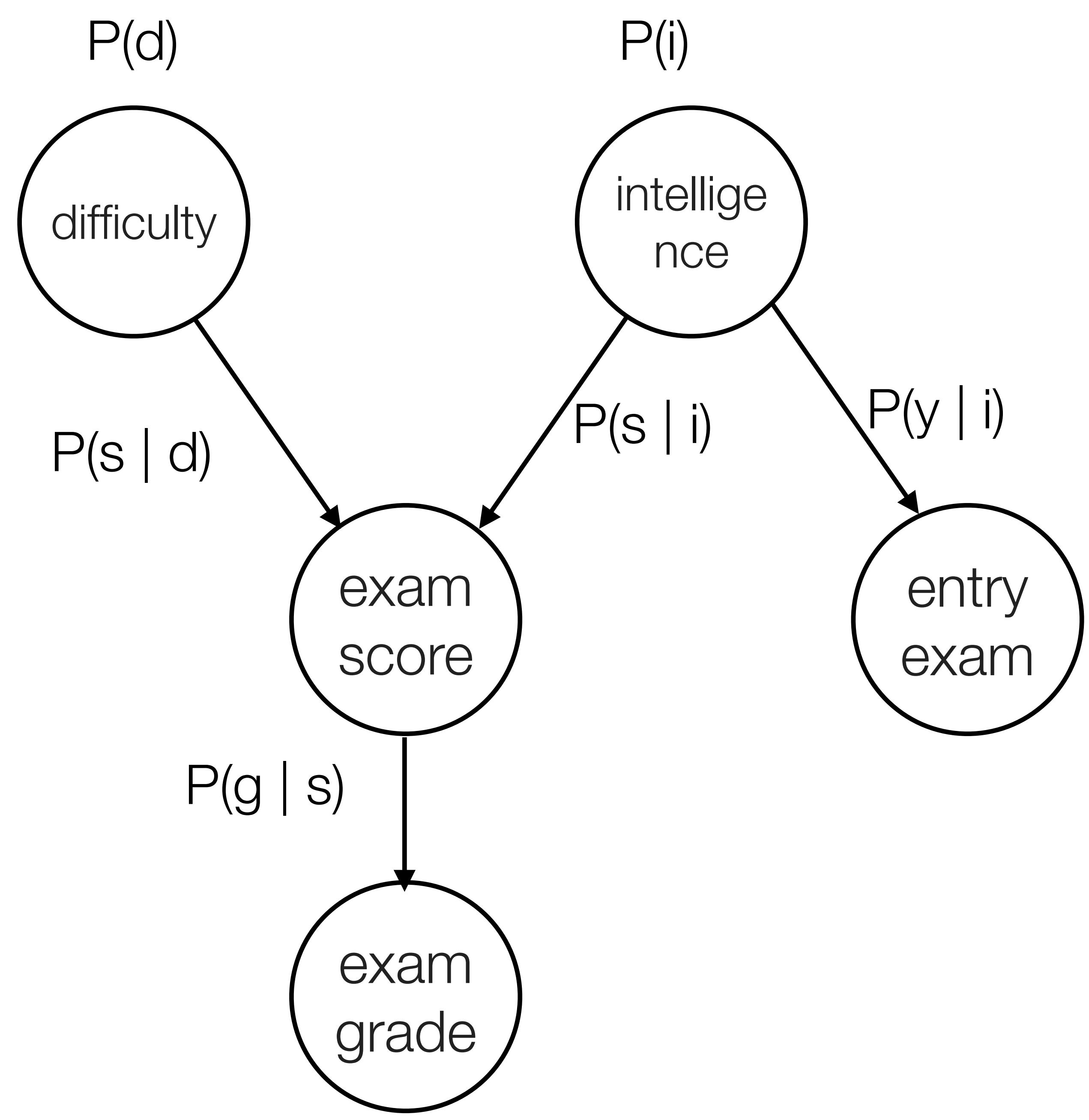
Rejection sampling problems

- $c \cdot Q(x)$ needs to be larger than $P^*(x)$, otherwise sampling will be biased (do not come from the target distribution)
- If $c \cdot Q(x)$ is too large then proportion of failed samples will increase
- It is not effective in high dimensions

Ancestral sampling

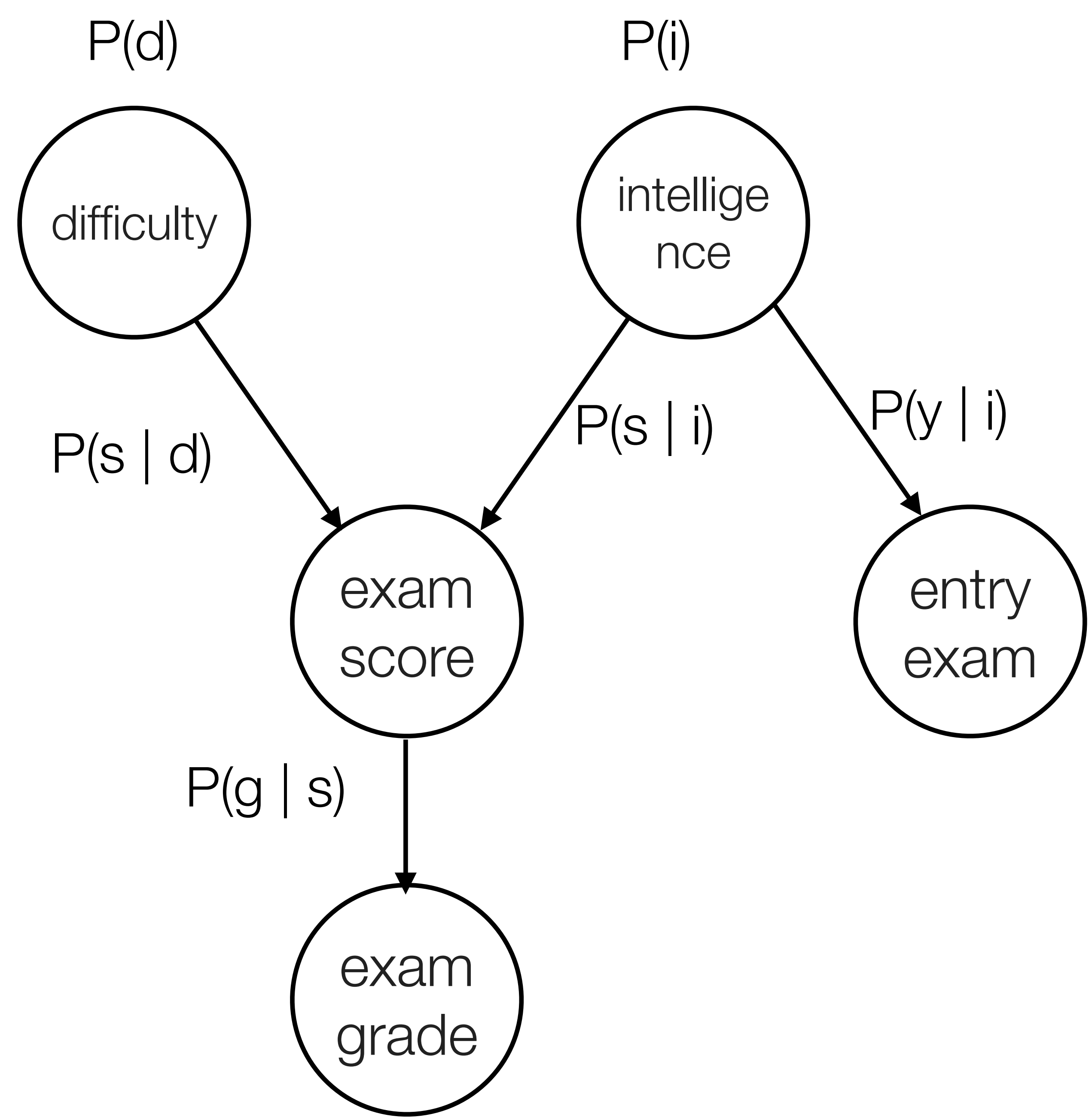


Ancestral sampling



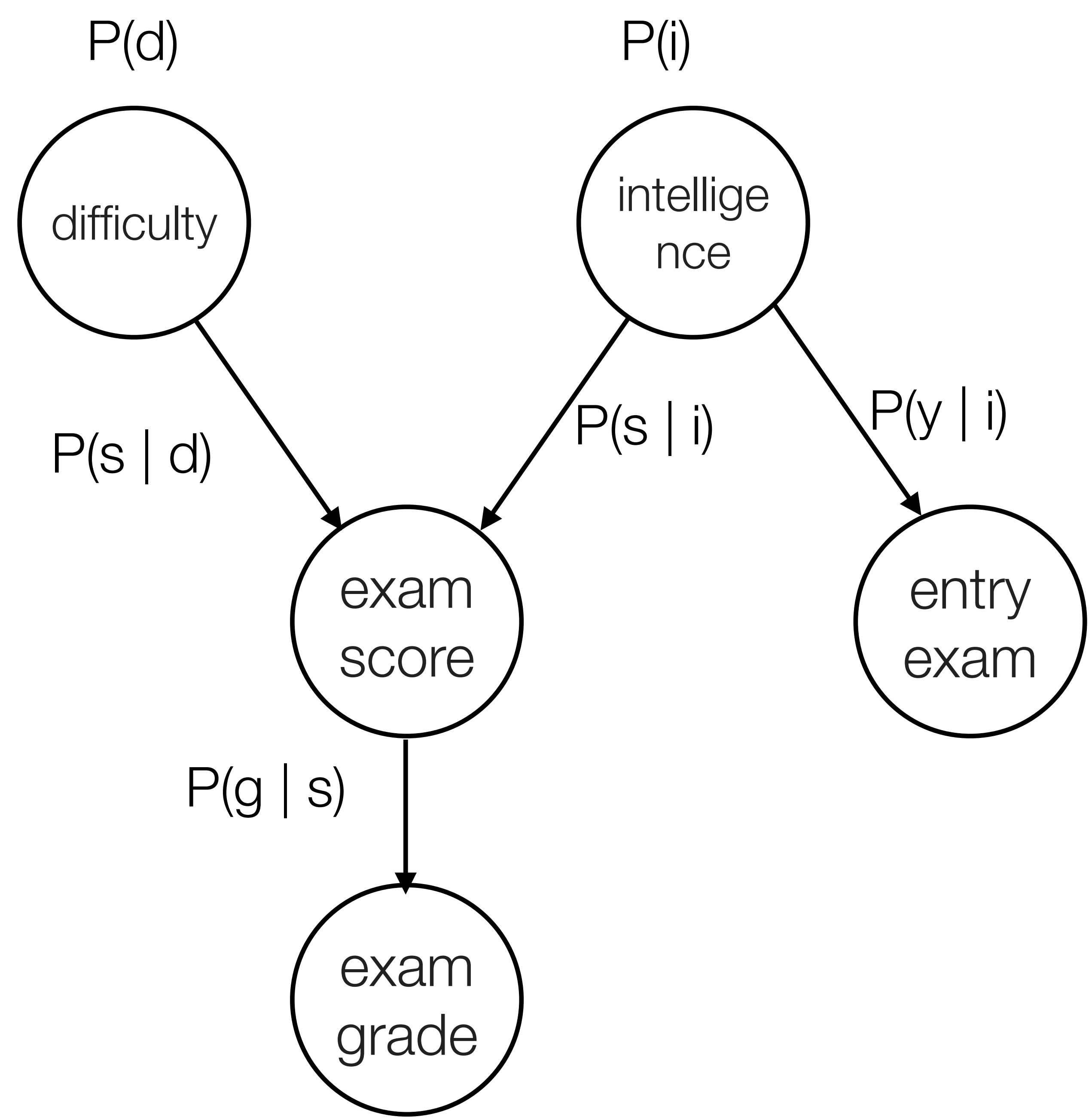
Ancestral sampling

- We are aiming to obtain a sample from the joint distribution



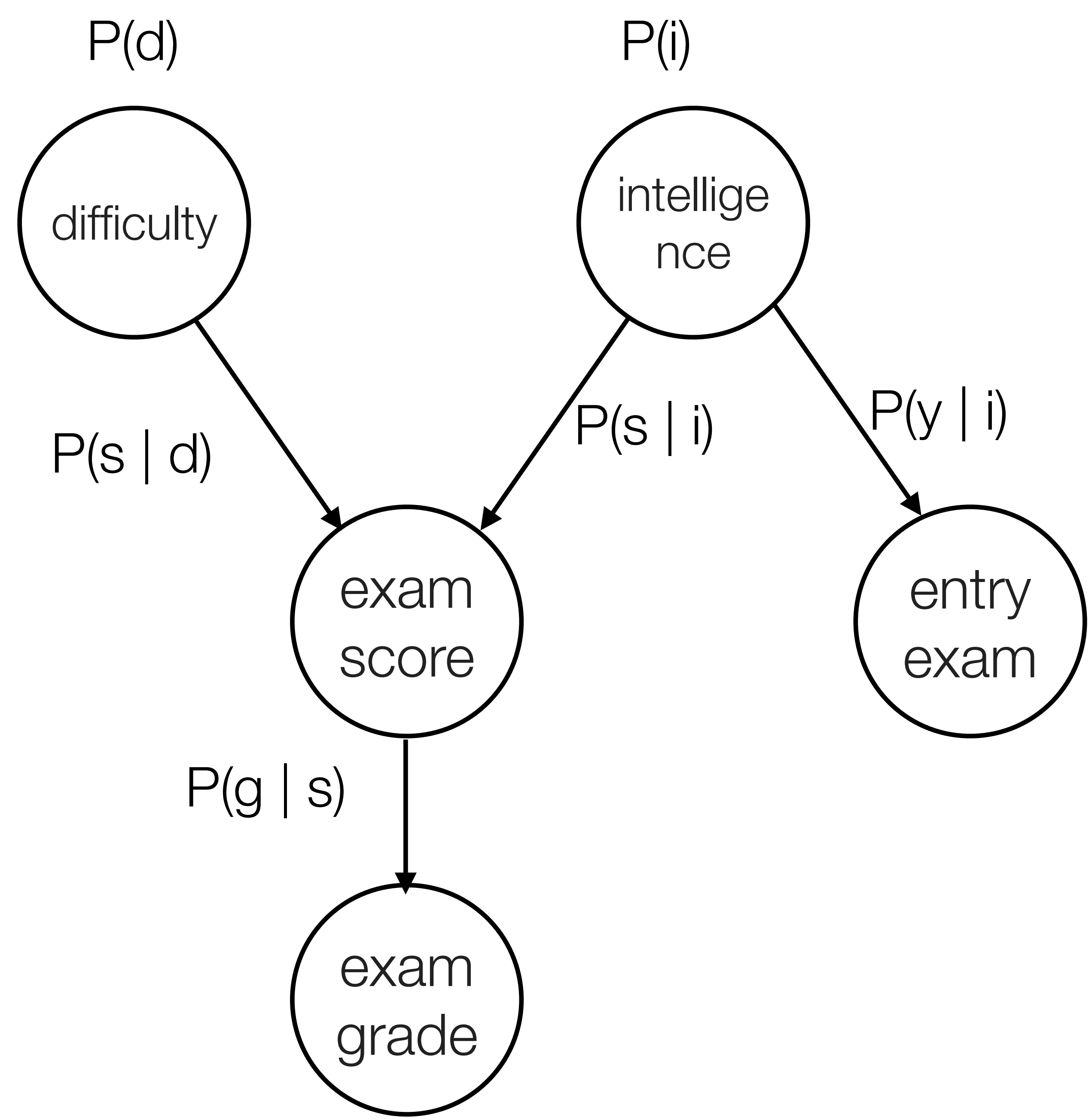
Ancestral sampling

- We are aiming to obtain a sample from the joint distribution
- We first sample the 'ancestors'



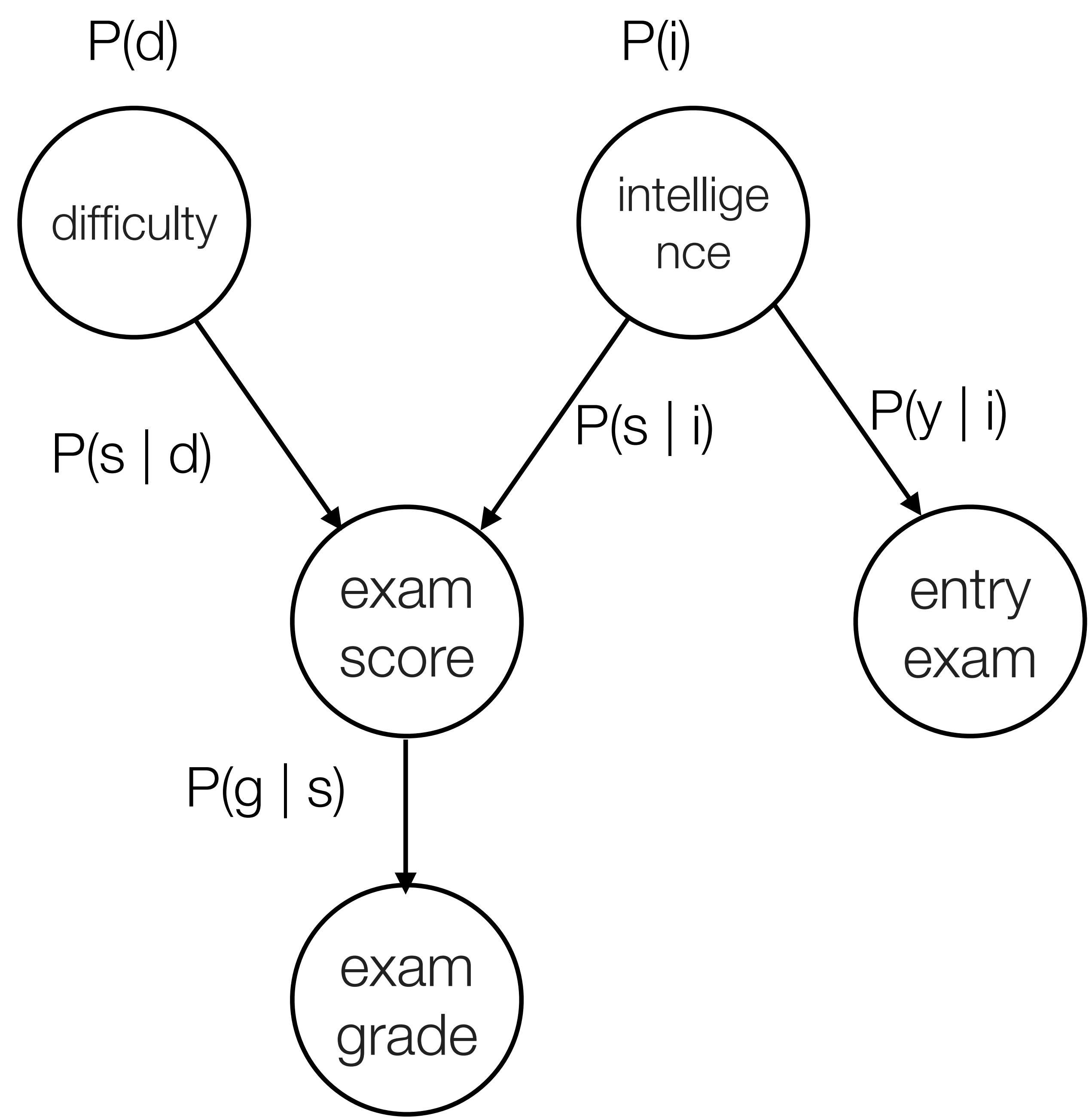
Ancestral sampling

- We are aiming to obtain a sample from the joint distribution
- We first sample the 'ancestors'
- Progressively sample descendants of sampled ancestors



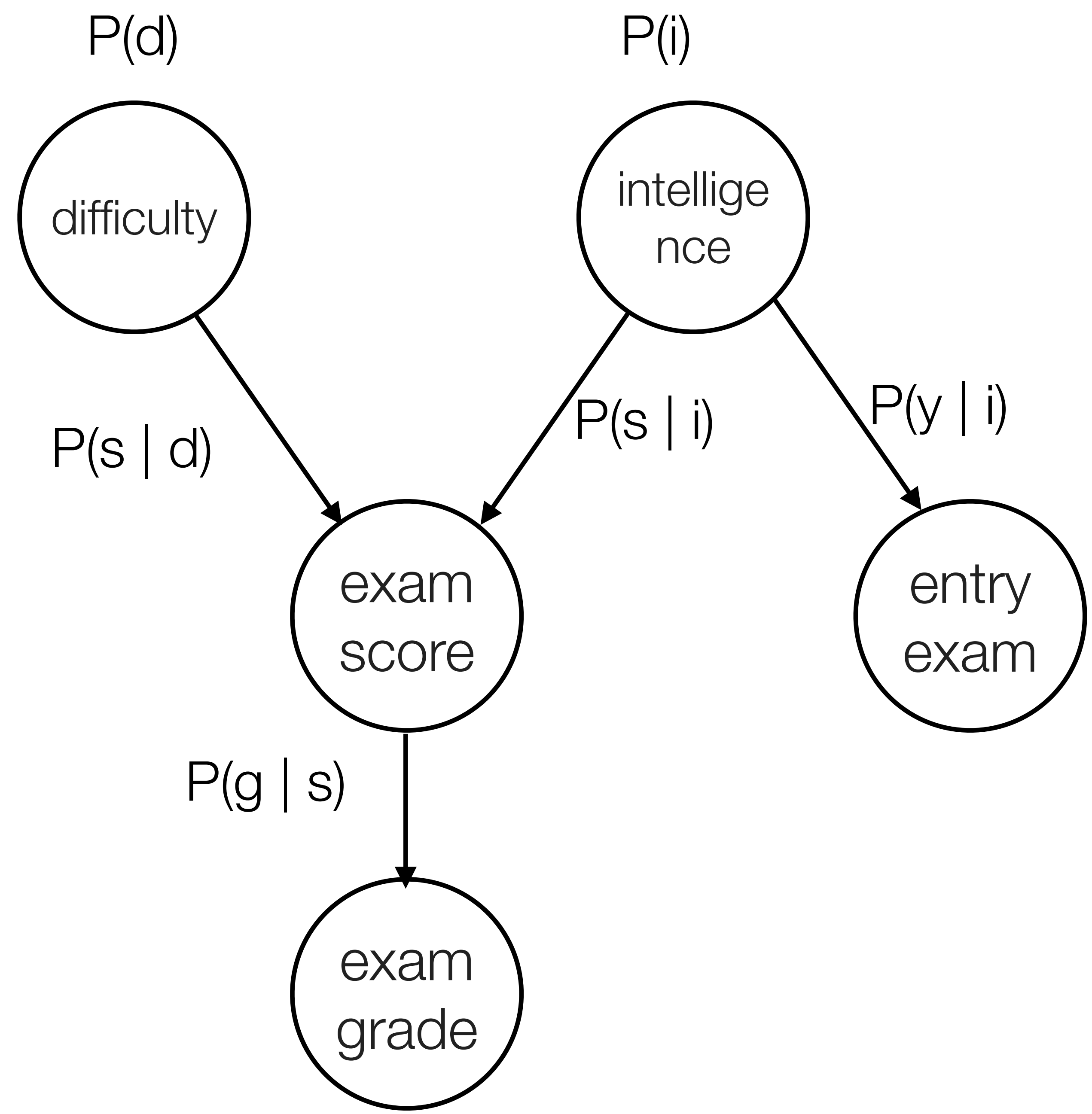
Ancestral sampling

- We are aiming to obtain a sample from the joint distribution
- We first sample the 'ancestors'
- Progressively sample descendants of sampled ancestors



How can we make inferences?
(obtain samples for arbitrary conditional distributions)

Ancestral sampling



- We are aiming to obtain a sample from the joint distribution
- We first sample the 'ancestors'
- Progressively sample descendants of sampled ancestors

How can we make inferences?
(obtain samples for arbitrary conditional distributions)

-> rejection sampling: drop those samples that are inconsistent with the conditions

Importance sampling

Importance sampling

- Instead of obtaining samples from the target distribution, P(x), we only want to calculate expectations over this distribution

$$E[f(x)] = \int f(x)P(x)\, dx$$

Importance sampling

- Instead of obtaining samples from the target distribution, P(x), we only want to calculate expectations over this distribution

$$E[f(x)] = \int f(x)P(x)\,dx$$

- We can sample a proposal distribution Q*(x)

Importance sampling

- Instead of obtaining samples from the target distribution, P(x), we only want to calculate expectations over this distribution

$$E[f(x)] = \int f(x)P(x)\,dx$$

- We can sample a proposal distribution Q*(x)

- 'Importance' of the sample from Q*(x) is set by the weight

$$w_t = \frac{P(x)}{Q^*(x)}$$

Importance sampling

- Instead of obtaining samples from the target distribution, P(x), we only want to calculate expectations over this distribution

$$E[f(x)] = \int f(x)P(x)\,dx$$

- We can sample a proposal distribution Q*(x)

- ‘Importance’ of the sample from Q*(x) is set by the weight

$$w_t = \frac{P(x)}{Q^*(x)}$$

Importance sampling

- Instead of obtaining samples from the target distribution, P(x), we only want to calculate expectations over this distribution

$$E[f(x)] = \int f(x)P(x)\,dx$$

- We can sample a proposal distribution Q*(x)

- ‘Importance’ of the sample from Q*(x) is set by the weight

$$w_t = \frac{P(x)}{Q^*(x)}$$

- The estimate is a weighted sum over samples

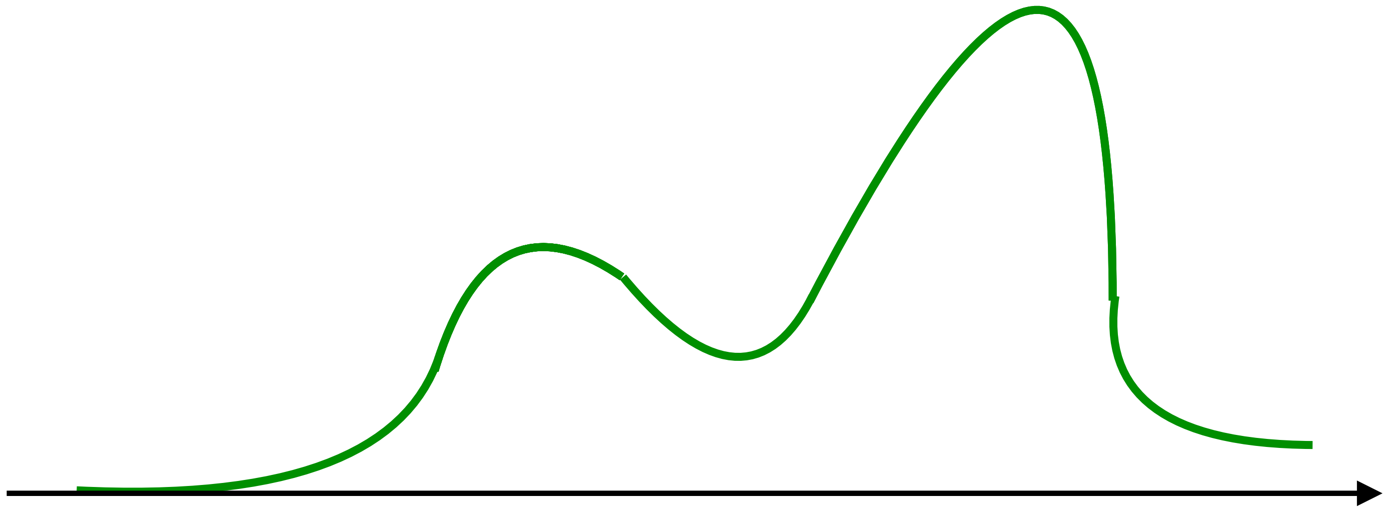
$$\hat{f}(x) = \frac{\sum_t w_t f(x)}{\sum_t w_t}$$

Importance sampling challenges

- Regions where $Q^*(x)$ is small but $P(x)$ is high are problematic
- The variance of the estimator cannot be reliably estimated
- In high dimensions (unless $Q^*(x)$ is a very good estimator) a very large number of samples is needed for a good estimate

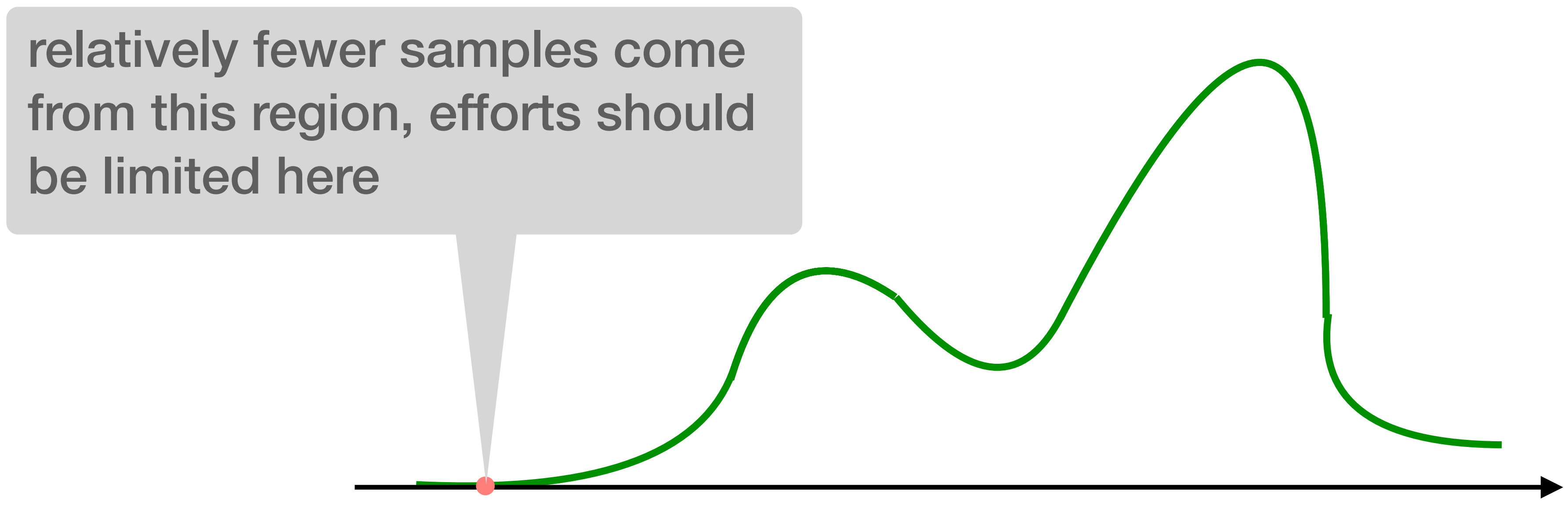
Efficient Monte Carlo methods

- Markov chain Monte Carlo (MCMC) methods:
  - Samples are generated sequentially:
  - Subsequent samples rely on earlier ones so that we sample regions where the probability mass is substantial



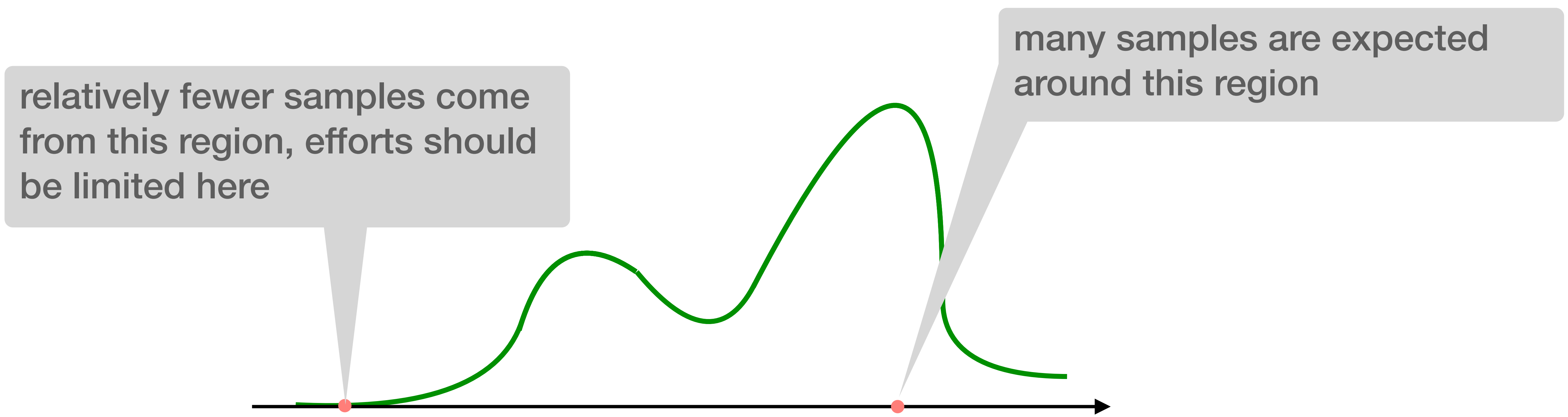
Efficient Monte Carlo methods

- Markov chain Monte Carlo (MCMC) methods:
  - Samples are generated sequentially:
  - Subsequent samples rely on earlier ones so that we sample regions where the probability mass is substantial



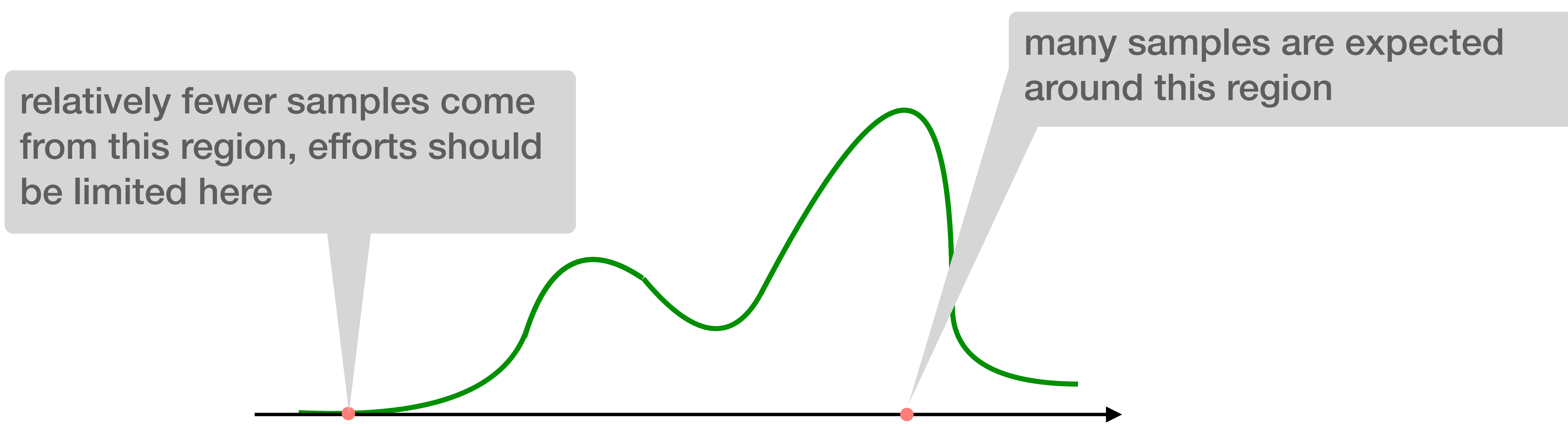
Efficient Monte Carlo methods

- Markov chain Monte Carlo (MCMC) methods:
  - Samples are generated sequentially:
  - Subsequent samples rely on earlier ones so that we sample regions where the probability mass is substantial



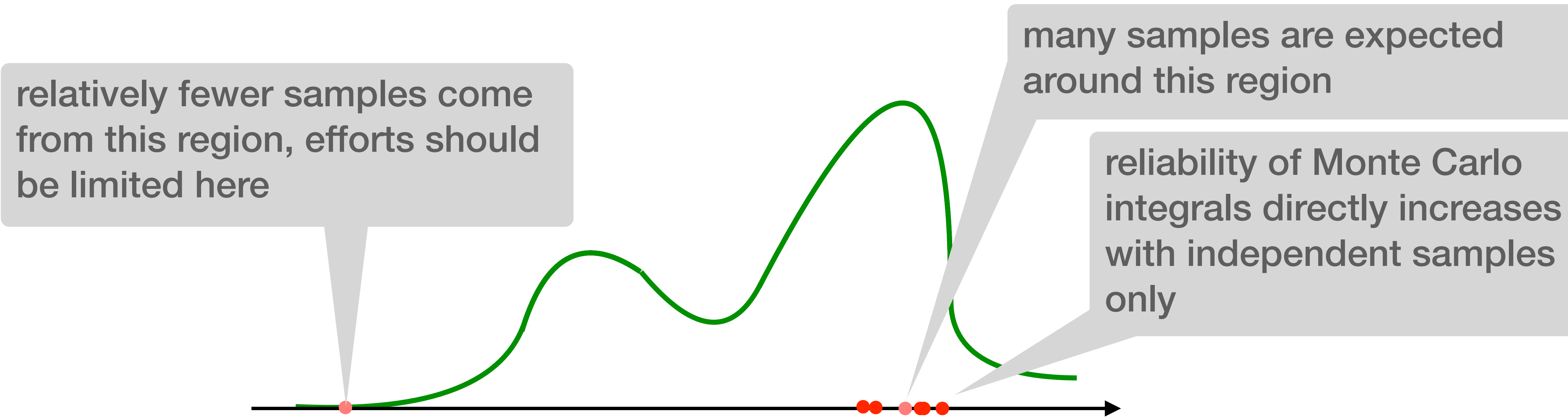
Efficient Monte Carlo methods

- Markov chain Monte Carlo (MCMC) methods:
  - Samples are generated sequentially:
  - Subsequent samples rely on earlier ones so that we sample regions where the probability mass is substantial
  - Drawback: we abandon independence of samples



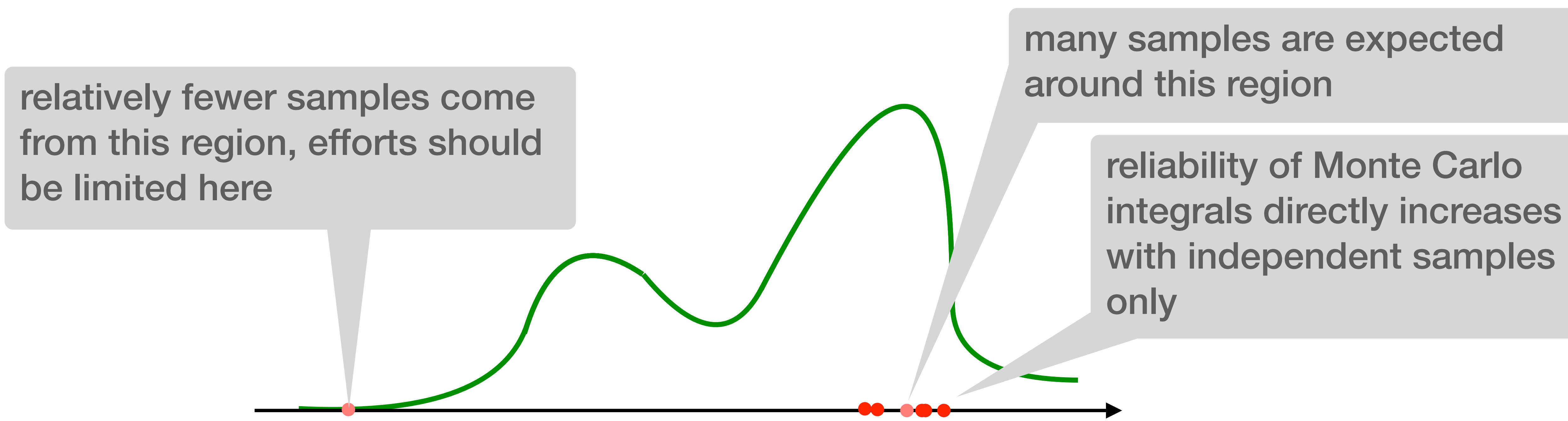
Efficient Monte Carlo methods

- Markov chain Monte Carlo (MCMC) methods:
  - Samples are generated sequentially:
  - Subsequent samples rely on earlier ones so that we sample regions where the probability mass is substantial
  - Drawback: we abandon independence of samples



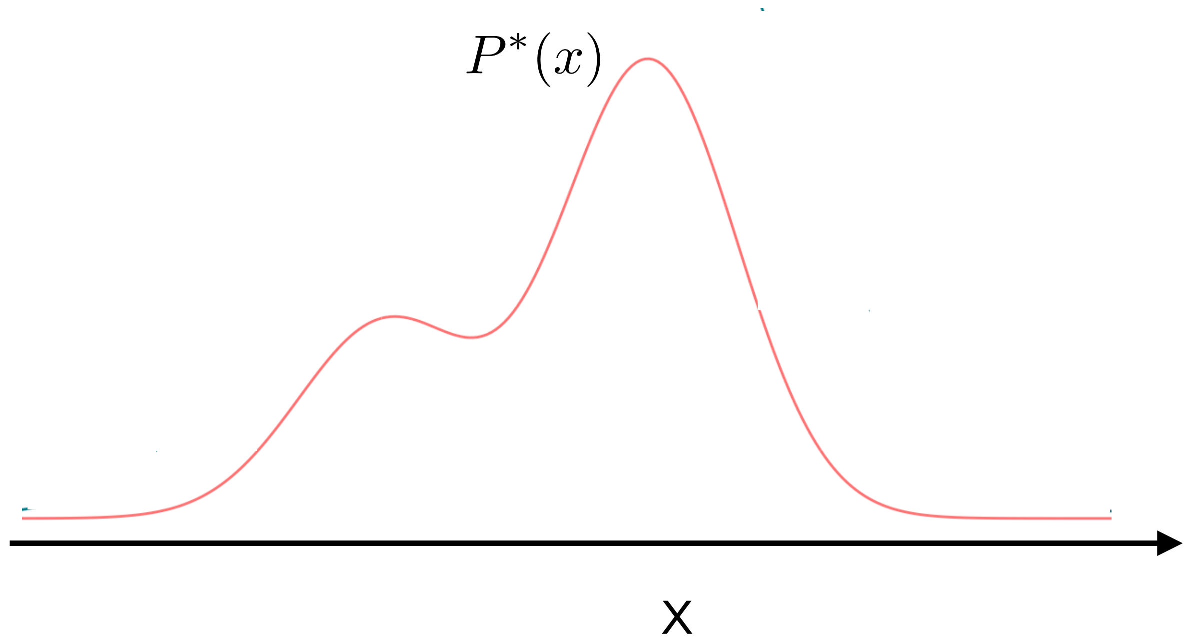
Efficient Monte Carlo methods

- Markov chain Monte Carlo (MCMC) methods:
  - Samples are generated sequentially:
  - Subsequent samples rely on earlier ones so that we sample regions where the probability mass is substantial
  - Drawback: we abandon independence of samples
  - ‘Slow mixing’: Multiple Markov chain Monte Carlo samples equal to the contribution of an independent sample



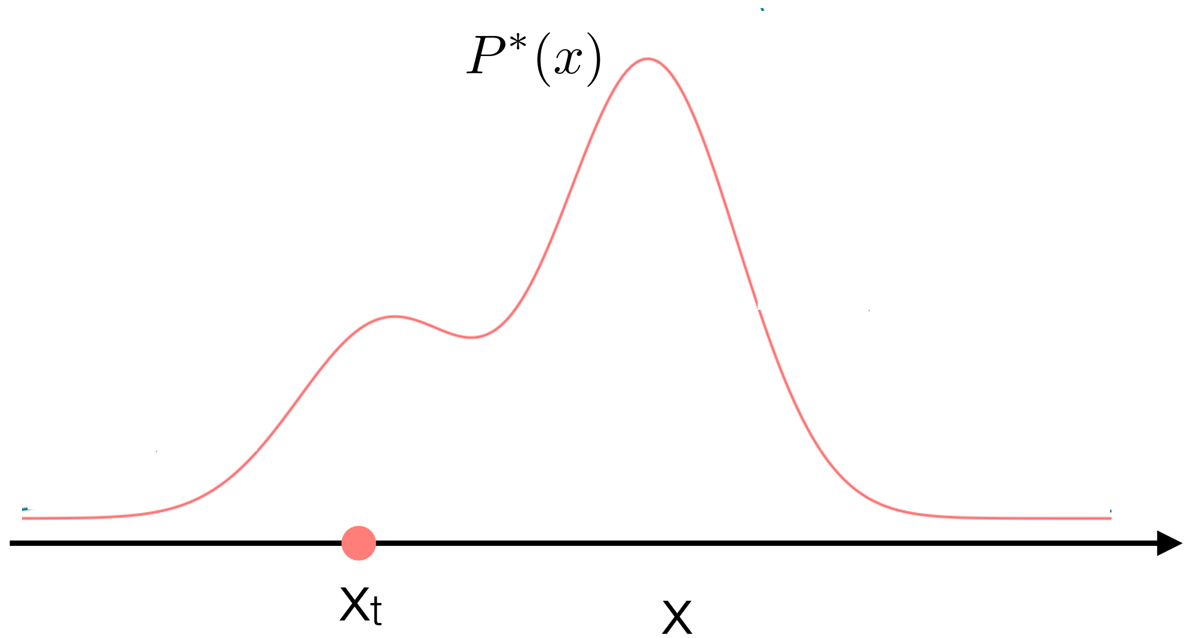
Metropolis Hastings algorithm

- Effective, general purpose algorithm to do integrals, inference, everything one needs



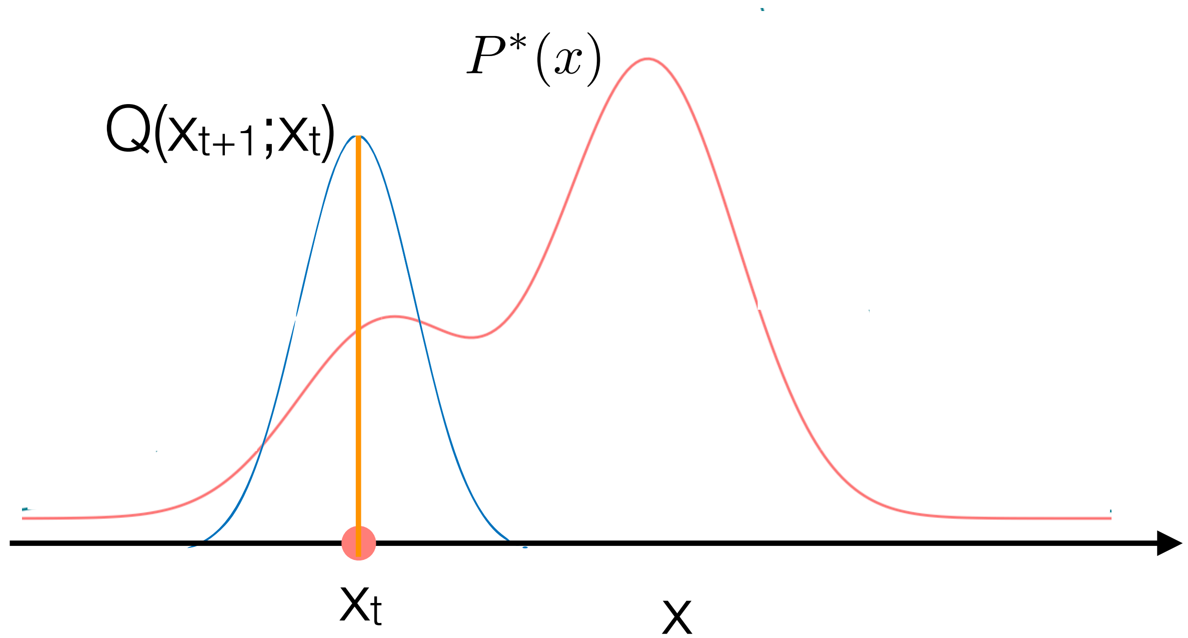
Metropolis Hastings algorithm

- Effective, general purpose algorithm to do integrals, inference, everything one needs



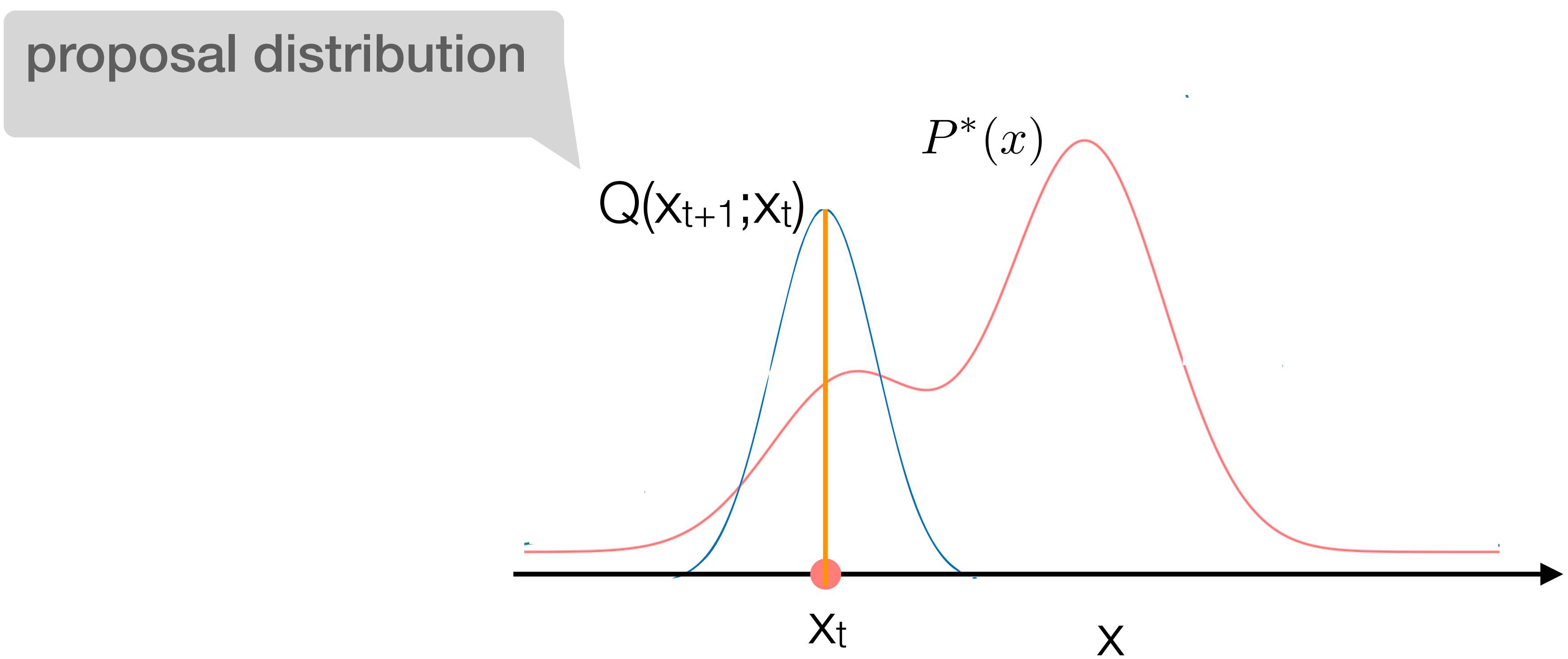
Metropolis Hastings algorithm

- Effective, general purpose algorithm to do integrals, inference, everything one needs



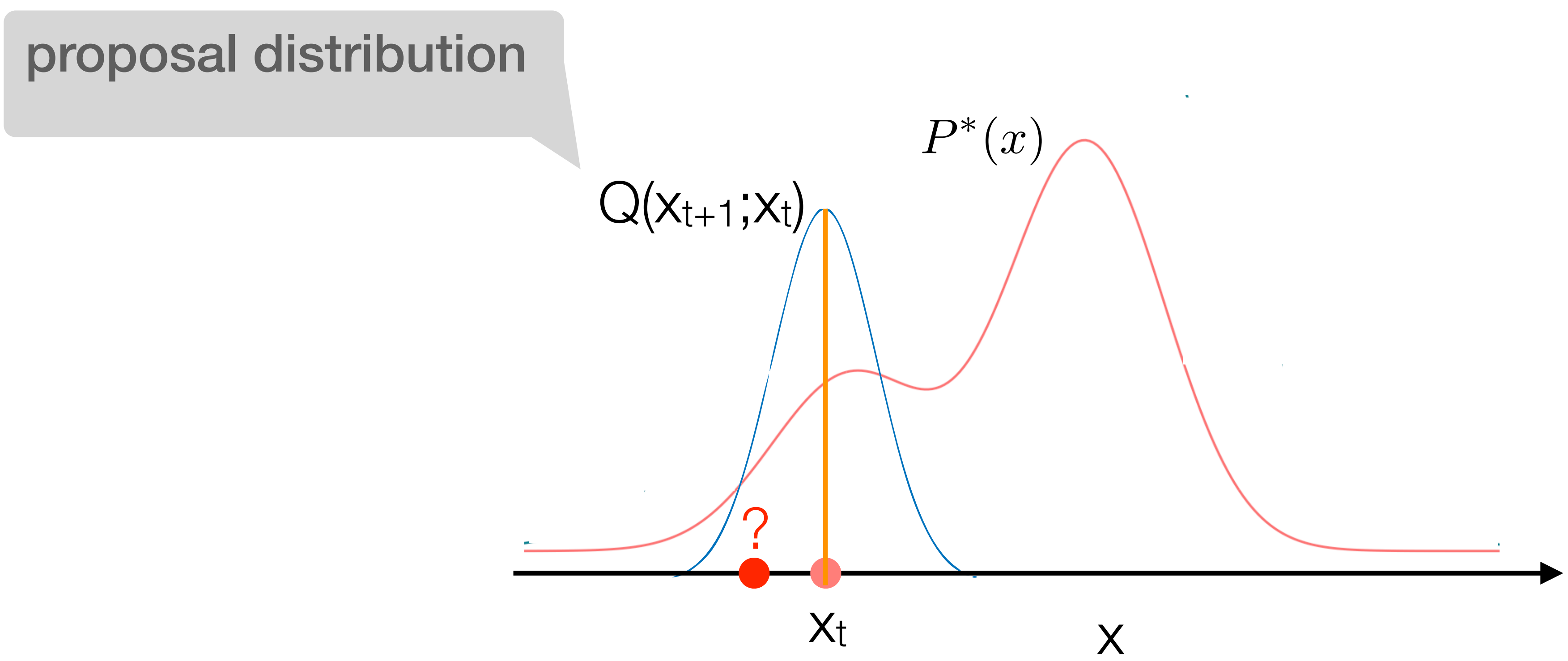
Metropolis Hastings algorithm

- Effective, general purpose algorithm to do integrals, inference, everything one needs



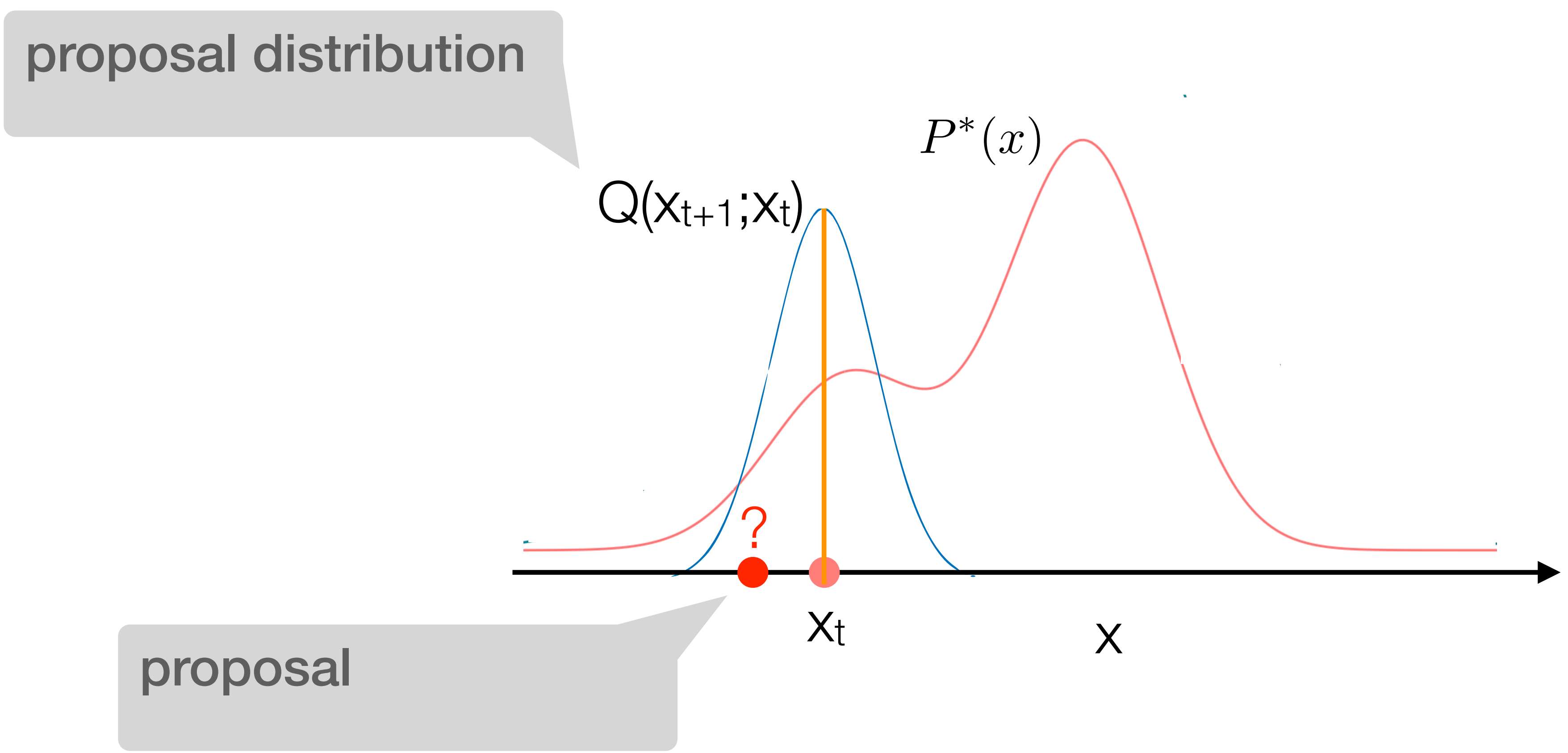
Metropolis Hastings algorithm

- Effective, general purpose algorithm to do integrals, inference, everything one needs



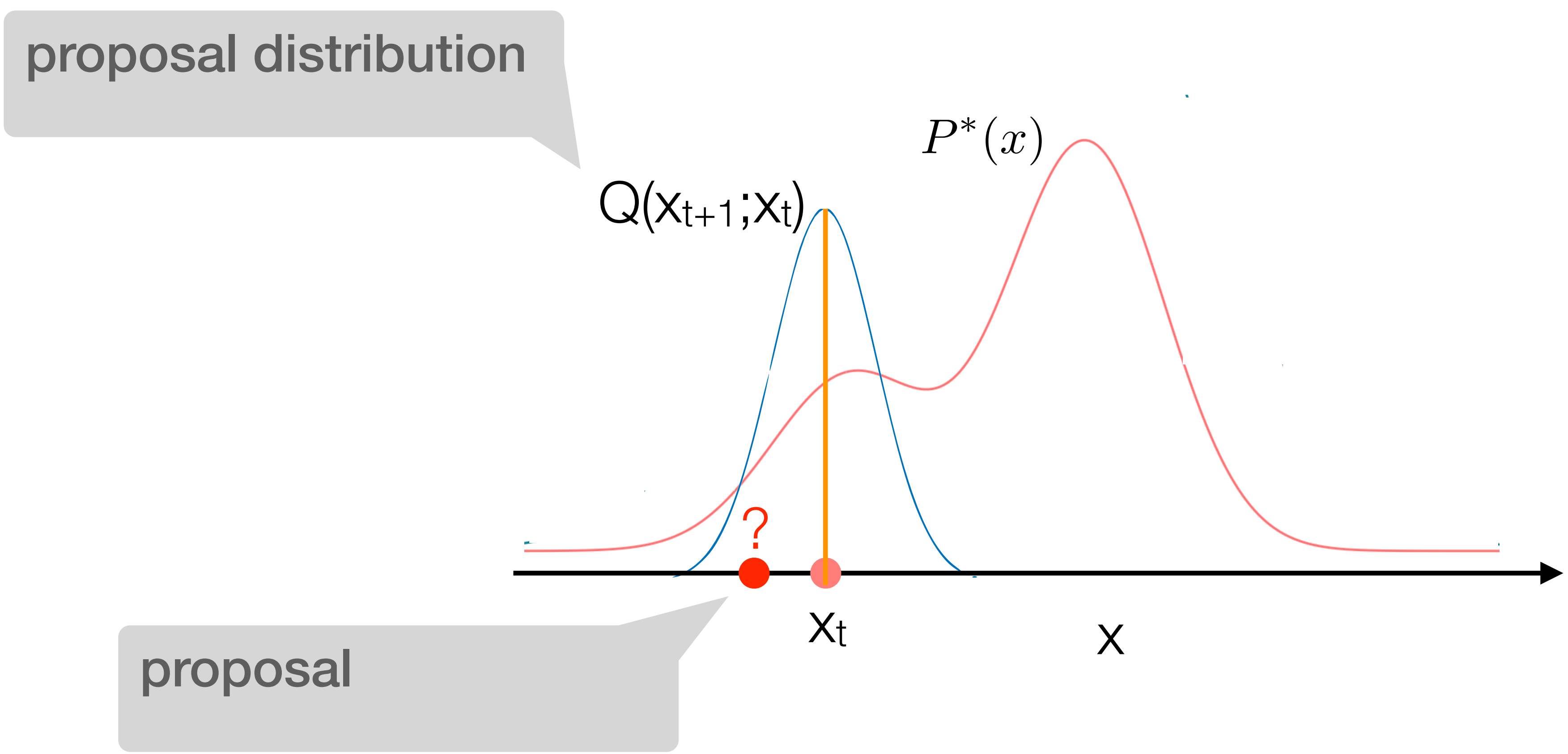
Metropolis Hastings algorithm

- Effective, general purpose algorithm to do integrals, inference, everything one needs



Metropolis Hastings algorithm

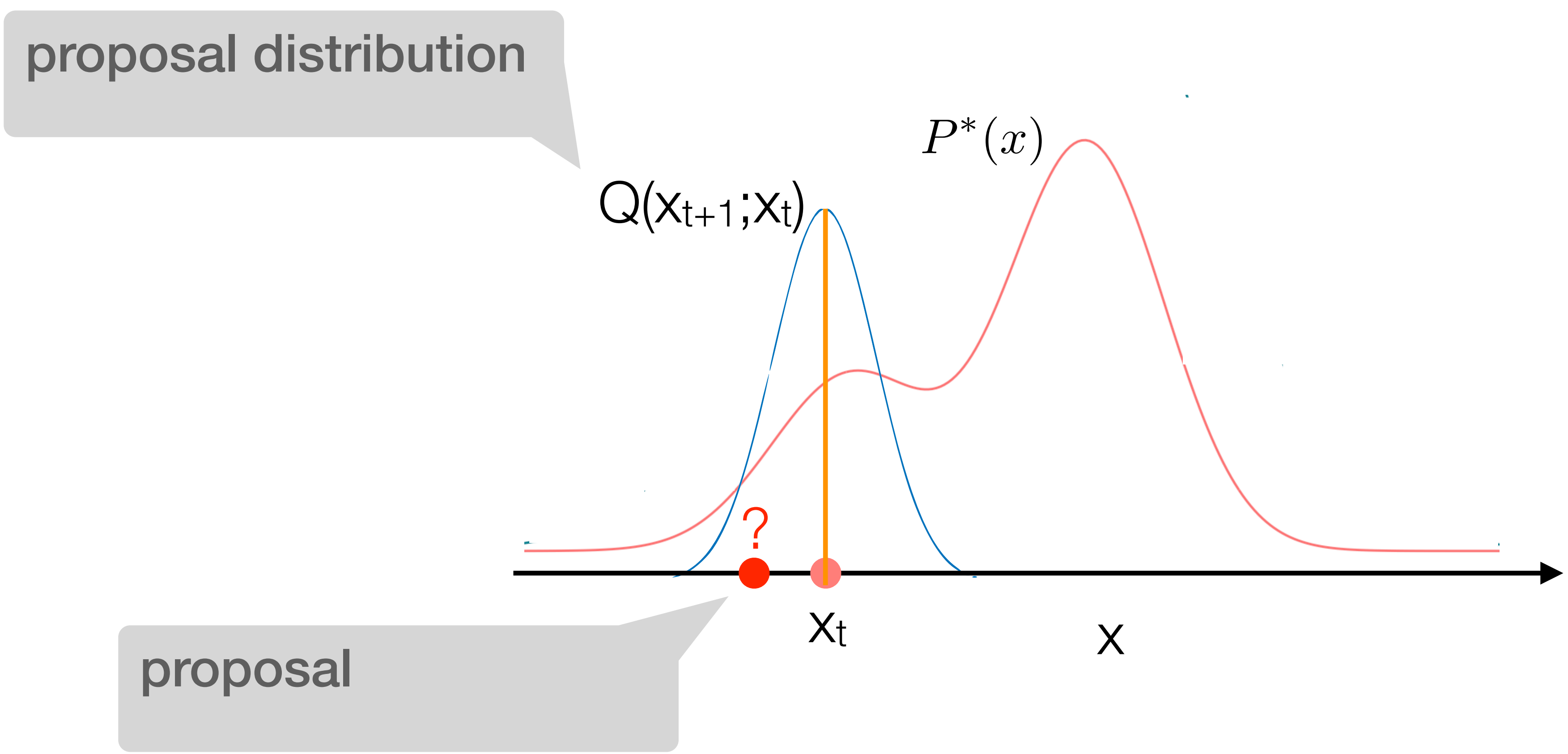
- Effective, general purpose algorithm to do integrals, inference, everything one needs



acceptance probability is proportional to $\dfrac{P^*(x_{t+1})}{P^*(x_t)}$

Metropolis Hastings algorithm

- Effective, general purpose algorithm to do integrals, inference, everything one needs

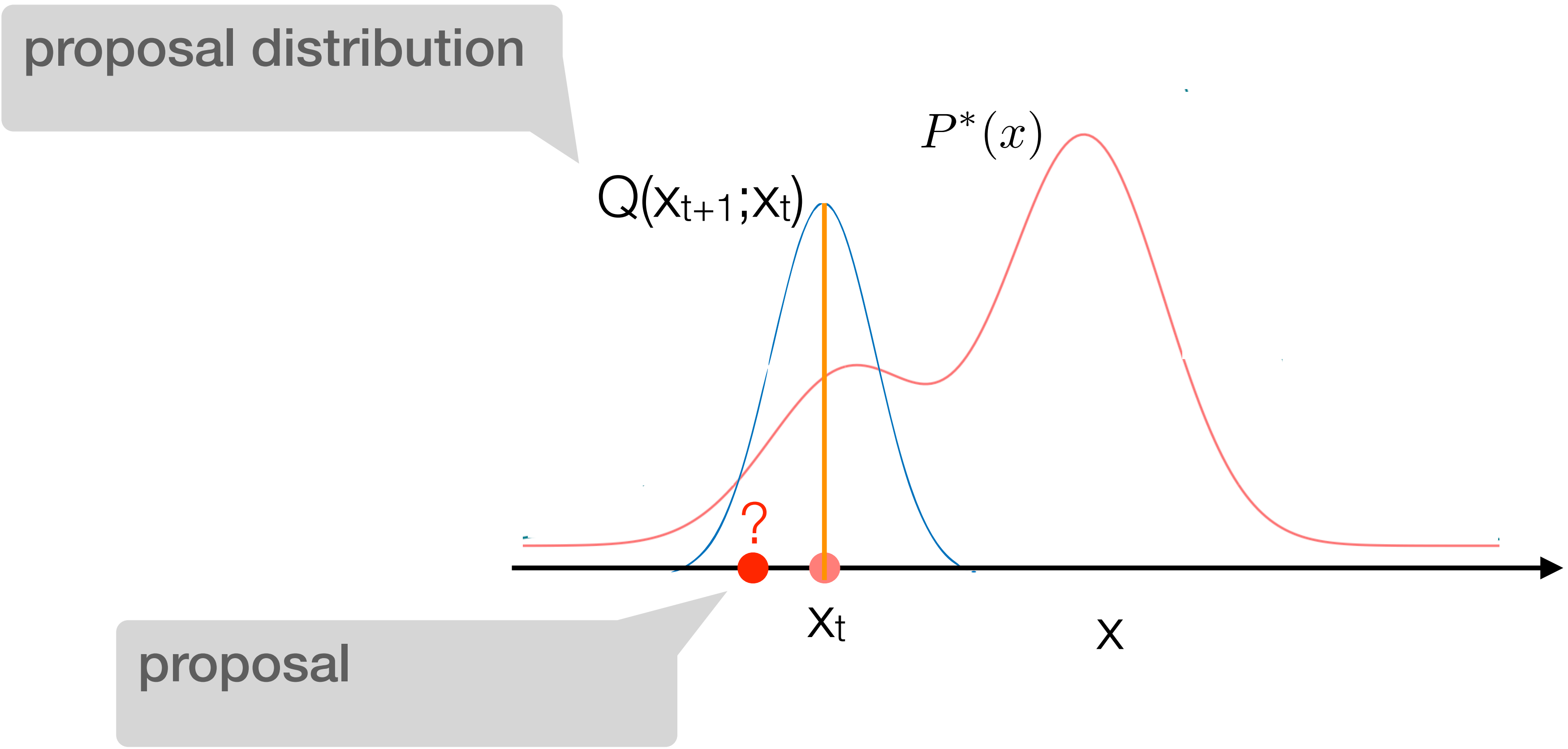


acceptance probability is proportional to $\dfrac{P^*(x_{t+1})}{P^*(x_t)}$

it is normalised by the easiness of getting from the proposed point back to the origin: $\dfrac{Q(x_t;\, x_{t+1})}{Q(x_{t+1};\, x_t)}$

Metropolis Hastings algorithm

- Effective, general purpose algorithm to do integrals, inference, everything one needs



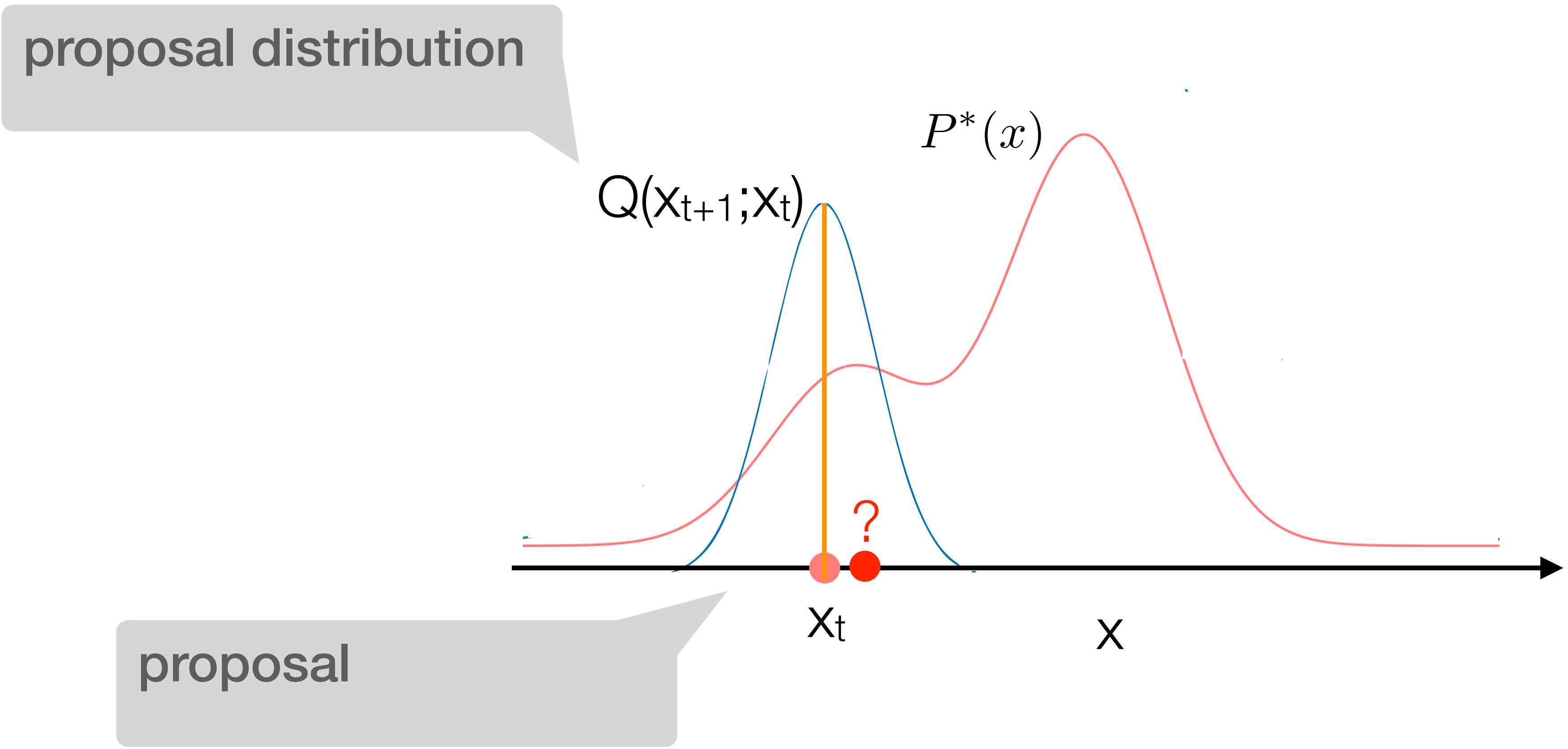
acceptance probability is proportional to $\frac{P^*(x_{t+1})}{P^*(x_t)}$

it is normalised by the easiness of getting from the proposed point back to the origin: $\frac{Q(x_t; x_{t+1})}{Q(x_{t+1}; x_t)}$

acceptance probability: $a = \frac{P^*(x_{t+1})}{P^*(x_t)} \frac{Q(x_t; x_{t+1})}{Q(x_{t+1}; x_t)}$

Metropolis Hastings algorithm

- Effective, general purpose algorithm to do integrals, inference, everything one needs



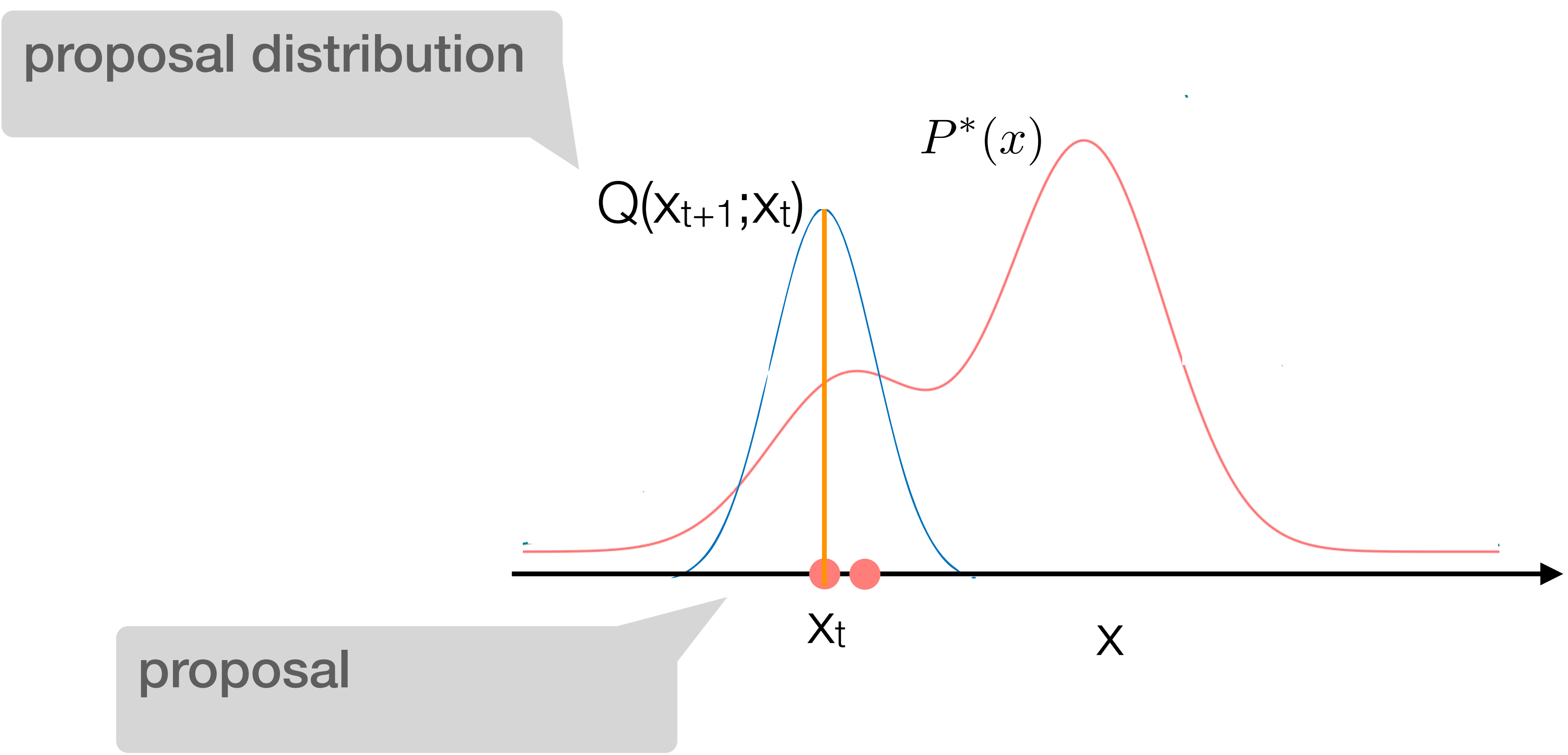
acceptance probability is proportional to $\frac{P^*(x_{t+1})}{P^*(x_t)}$

it is normalised by the easiness of getting from the proposed point back to the origin: $\frac{Q(x_t;x_{t+1})}{Q(x_{t+1};x_t)}$

acceptance probability: $a = \frac{P^*(x_{t+1})}{P^*(x_t)} \frac{Q(x_t;x_{t+1})}{Q(x_{t+1};x_t)}$

Metropolis Hastings algorithm

- Effective, general purpose algorithm to do integrals, inference, everything one needs



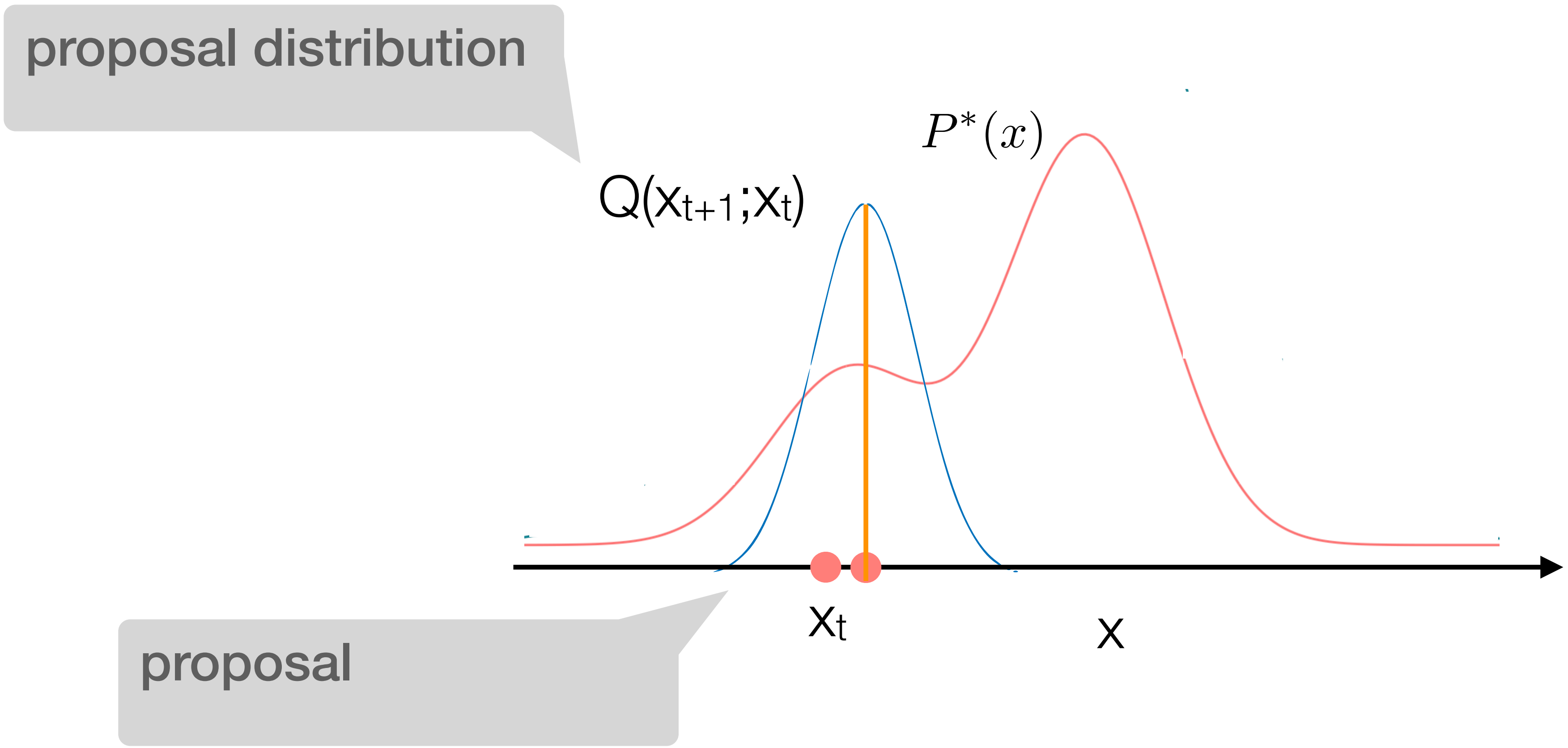
acceptance probability is proportional to $\frac{P^*(x_{t+1})}{P^*(x_t)}$

it is normalised by the easiness of getting from the proposed point back to the origin: $\frac{Q(x_t;\, x_{t+1})}{Q(x_{t+1};\, x_t)}$

acceptance probability: $a = \frac{P^*(x_{t+1})}{P^*(x_t)} \frac{Q(x_t;\, x_{t+1})}{Q(x_{t+1};\, x_t)}$

Metropolis Hastings algorithm

- Effective, general purpose algorithm to do integrals, inference, everything one needs



acceptance probability is proportional to $\frac{P^*(x_{t+1})}{P^*(x_t)}$

it is normalised by the easiness of getting from the proposed point back to the origin: $\frac{Q(x_t; x_{t+1})}{Q(x_{t+1}; x_t)}$

acceptance probability: $a = \frac{P^*(x_{t+1})}{P^*(x_t)} \frac{Q(x_t; x_{t+1})}{Q(x_{t+1}; x_t)}$

Metropolis Hastings algorithm

- After a large number of steps $x_t \sim P(x)$,
  i.e. the histogram of x_t is faithfully representing P(x)

- Initial samples depend on the initial choice:
  samples in the burn-in period need to be discarded

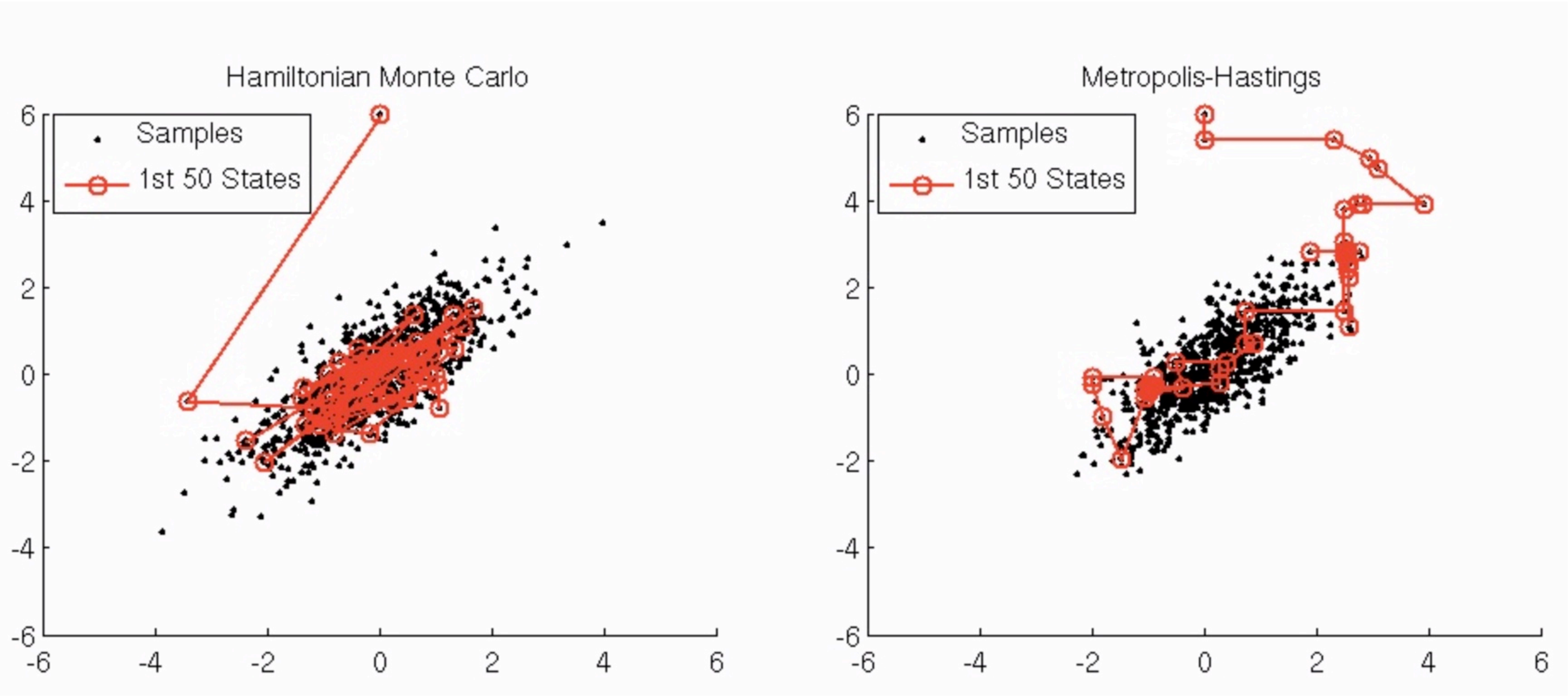
- Since samples are not independent, closely samples can be discarded: thinning

Alternative MCMC methods

- Hamiltonian Monte Carlo — exploits the shape of the probability distribution to design proposals

Alternative MCMC methods

- Hamiltonian Monte Carlo — exploits the shape of the probability distribution to design proposals



shorter burn-in & faster mixing

Alternative MCMC methods

- Hamiltonian Monte Carlo — exploits the shape of the probability distribution to design proposals

Alternative MCMC methods

- Hamiltonian Monte Carlo — exploits the shape of the probability distribution to design proposals

- Slice sampling — adjusts the properties of the proposal distributior automatically

Alternative MCMC methods

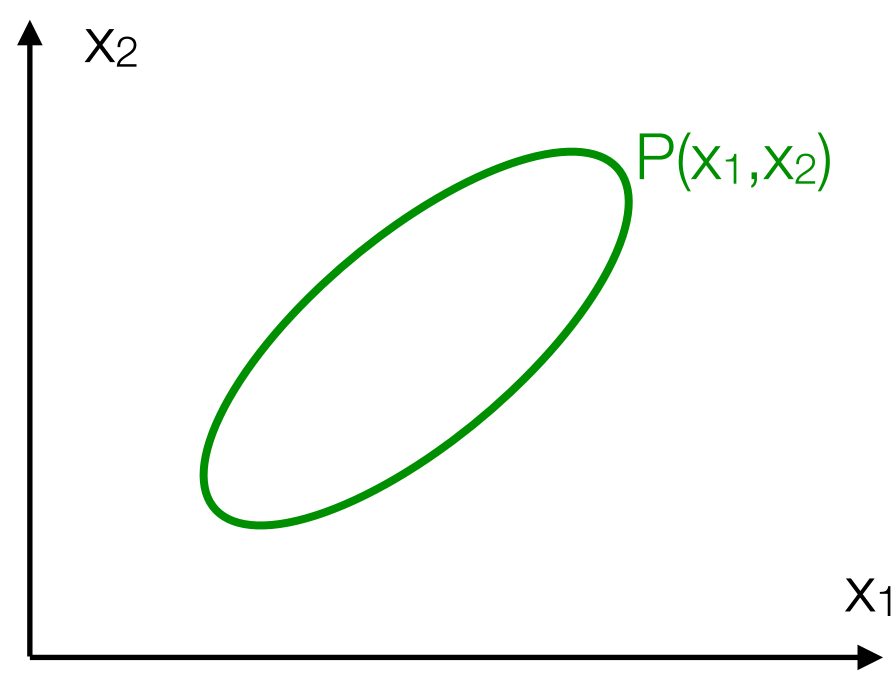
- Hamiltonian Monte Carlo — exploits the shape of the probability distribution to design proposals
- Slice sampling — adjusts the properties of the proposal distributior automatically
- Gibbs sampling — having a multi-dimensional distribution over x_1, . if conditional distributions, e.g. $P(x_1|x_2,\ldots,x_n)$, can be sampled then t are sequentially sampled

Alternative MCMC methods

- Hamiltonian Monte Carlo — exploits the shape of the probability distribution to design proposals

- Slice sampling — adjusts the properties of the proposal distributio automatically

- Gibbs sampling — having a multi-dimensional distribution over x_1, . if conditional distributions, e.g. $P(x_1|x_2,\ldots,x_n)$, can be sampled then t are sequentially sampled

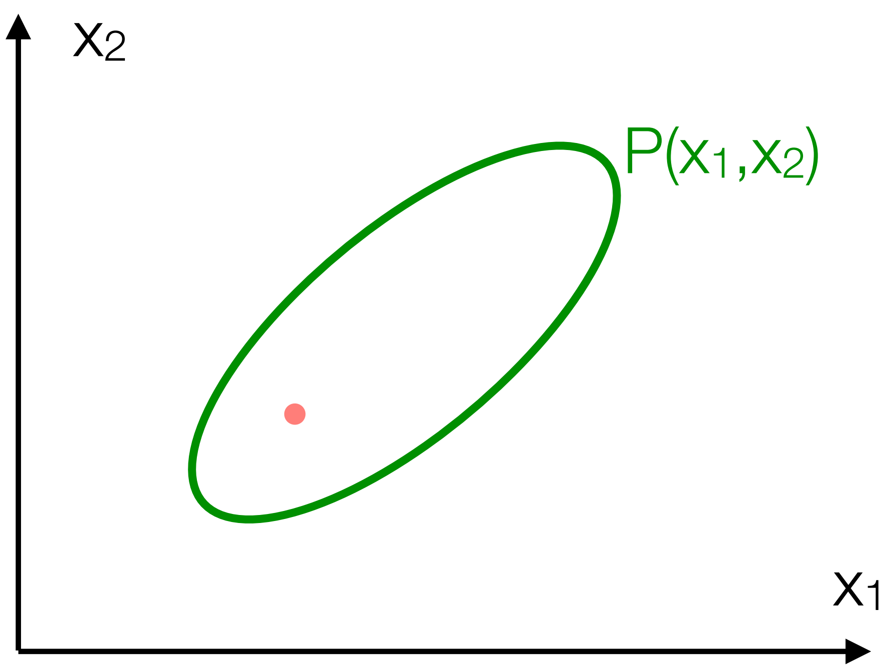


Alternative MCMC methods

- Hamiltonian Monte Carlo — exploits the shape of the probability distribution to design proposals

- Slice sampling — adjusts the properties of the proposal distribution automatically

- Gibbs sampling — having a multi-dimensional distribution over x_1, ... if conditional distributions, e.g. $P(x_1|x_2,\ldots,x_n)$, can be sampled then t... are sequentially sampled

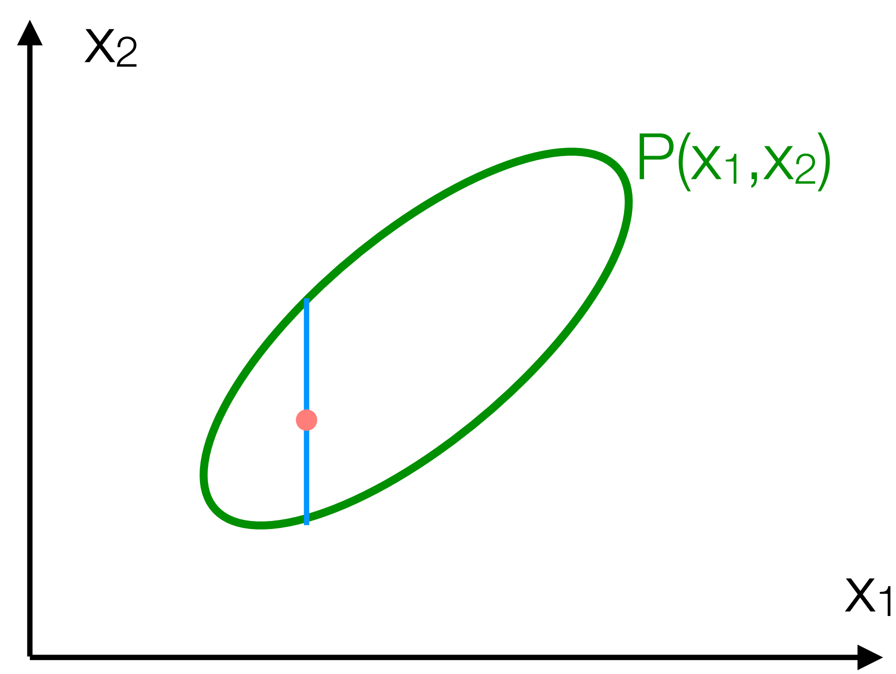


Alternative MCMC methods

- Hamiltonian Monte Carlo — exploits the shape of the probability distribution to design proposals

- Slice sampling — adjusts the properties of the proposal distribution automatically

- Gibbs sampling — having a multi-dimensional distribution over x_1, . if conditional distributions, e.g. $P(x_1|x_2,\ldots,x_n)$, can be sampled then t are sequentially sampled

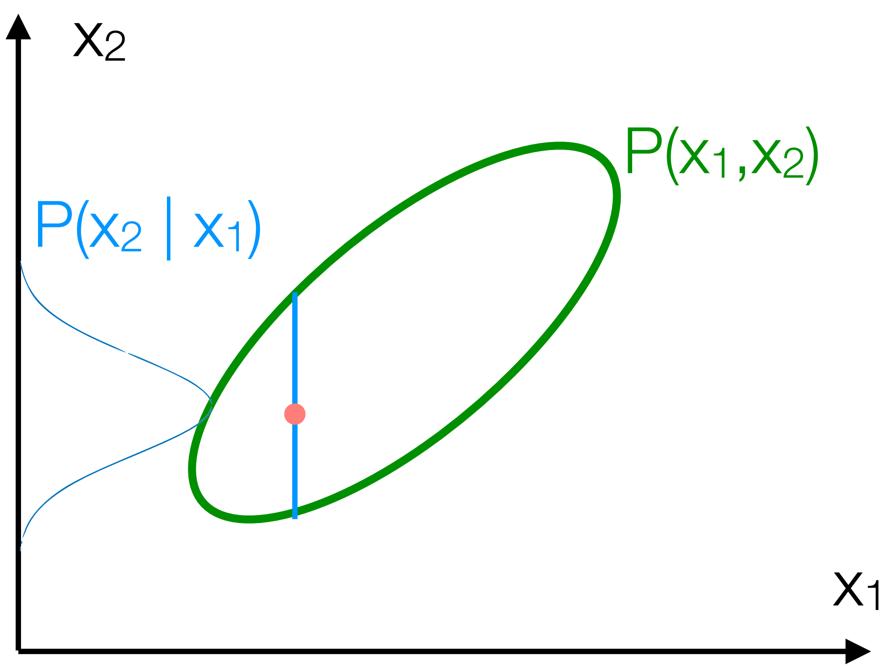


Alternative MCMC methods

- Hamiltonian Monte Carlo — exploits the shape of the probability distribution to design proposals

- Slice sampling — adjusts the properties of the proposal distributior automatically

- Gibbs sampling — having a multi-dimensional distribution over x_1, . if conditional distributions, e.g. $P(x_1|x_2,\ldots,x_n)$, can be sampled then t are sequentially sampled

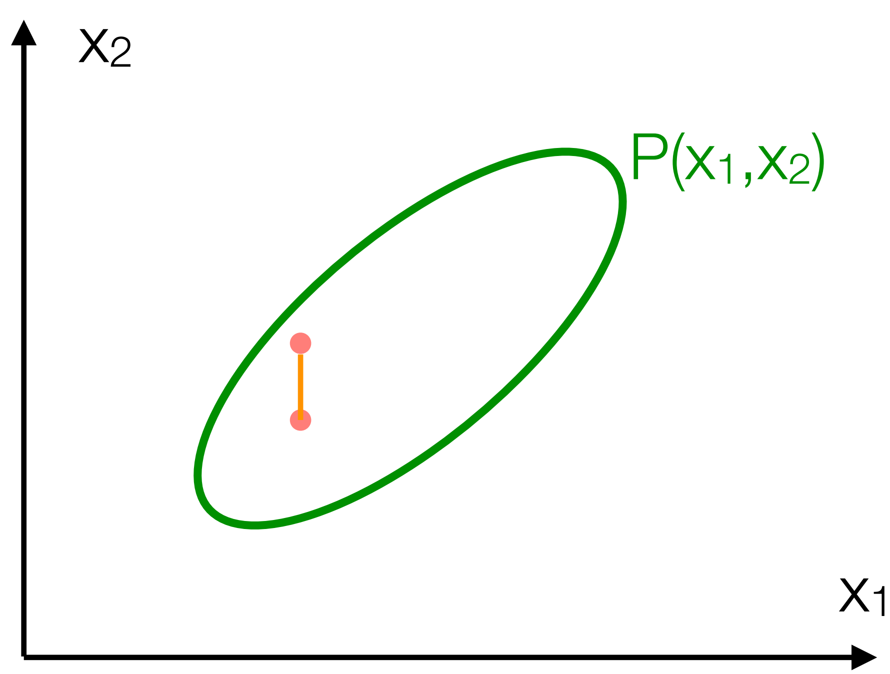


Alternative MCMC methods

- Hamiltonian Monte Carlo — exploits the shape of the probability distribution to design proposals

- Slice sampling — adjusts the properties of the proposal distributio automatically

- Gibbs sampling — having a multi-dimensional distribution over x_1, . if conditional distributions, e.g. $P(x_1|x_2,\ldots,x_n)$, can be sampled then t are sequentially sampled

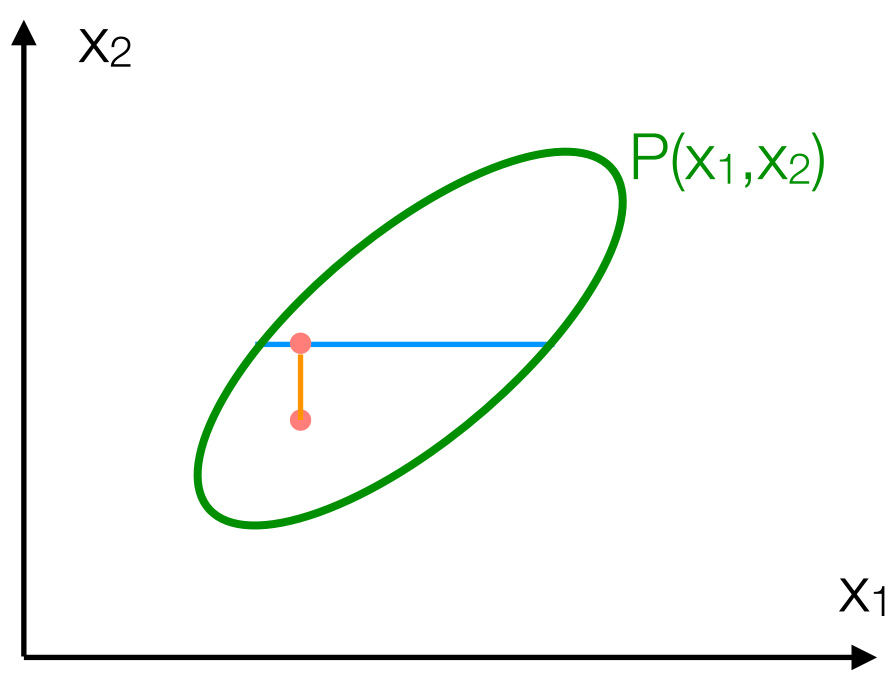


Alternative MCMC methods

- Hamiltonian Monte Carlo — exploits the shape of the probability distribution to design proposals

- Slice sampling — adjusts the properties of the proposal distributio automatically

- Gibbs sampling — having a multi-dimensional distribution over x_1, . if conditional distributions, e.g. $P(x_1|x_2,\ldots,x_n)$, can be sampled then t are sequentially sampled

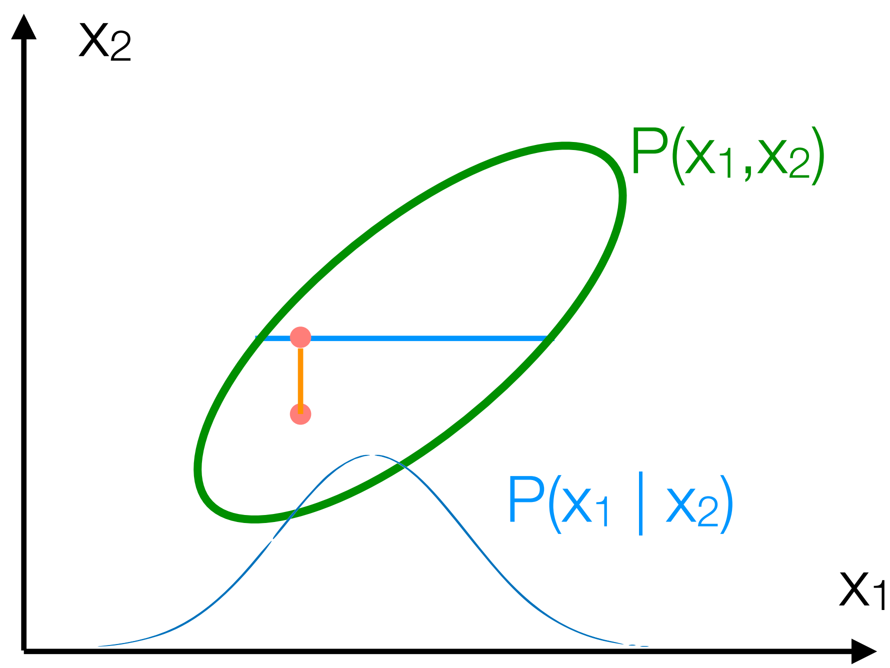


Alternative MCMC methods

- Hamiltonian Monte Carlo — exploits the shape of the probability distribution to design proposals

- Slice sampling — adjusts the properties of the proposal distributio automatically

- Gibbs sampling — having a multi-dimensional distribution over x_1, . if conditional distributions, e.g. $P(x_1|x_2,\dots,x_n)$, can be sampled then t are sequentially sampled

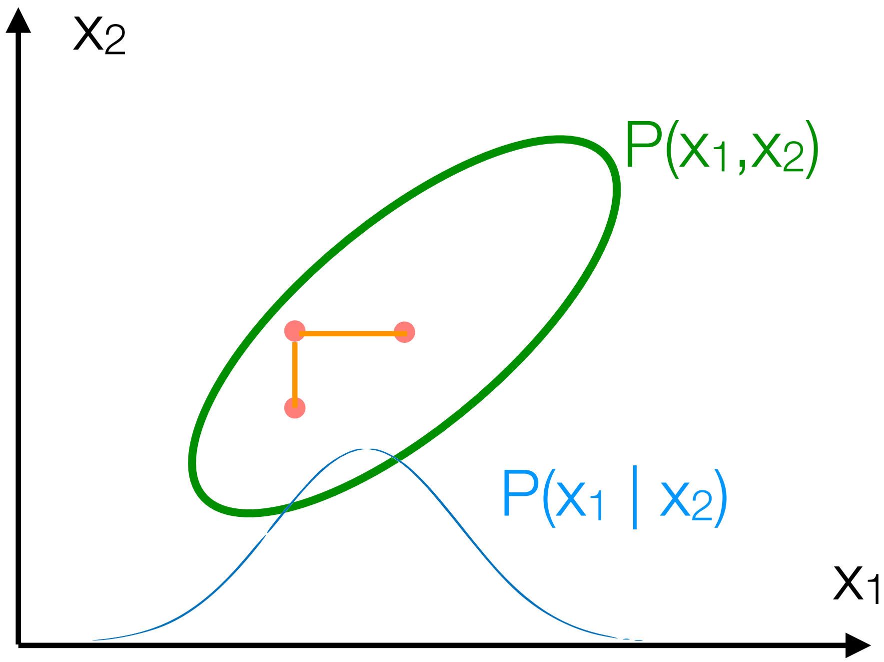


Alternative MCMC methods

- Hamiltonian Monte Carlo — exploits the shape of the probability distribution to design proposals

- Slice sampling — adjusts the properties of the proposal distributio automatically

- Gibbs sampling — having a multi-dimensional distribution over x_1, . if conditional distributions, e.g. $P(x_1|x_2,\ldots,x_n)$, can be sampled then t are sequentially sampled



Significance of sampling

- An efficient sampling architecture can save us from scary integrals: we can side step the bizarre math
- Sampling bridges the gap between the mathematical transparency of inference on discrete variables and the cumbersome inference on continuous variables
- Sampling, as an approximate strategy to perform plausible reasoning might be used by humans to make inferences