Bayesian probability: models and inference

Adam Arany

16 May 2022
Lugano, Switzerland

KU LEUVEN

Overview

- Interpretation of probabilities
- Bayesian modeling
  - Parameters vs variables
  - Graphical model notation
- Bayesian inference
  - Prior conjugacy
  - Maximum a posteriori approximation
  - Predictive inference
  - Sampling methods
  - Variational inference
- Outlook
  - Structural uncertainty
  - Causality

“... a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible.”

Pierre-*Simon Laplace*

relative frequency of occurrence after repeating a process a large number of times under similar conditions

degree of reasonable belief

tendency of a given type of physical situation to yield an outcome

…

Classical

"... a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible."

Pierre-*Simon Laplace*

Frequentist

relative frequency of occurrence after repeating a process a large number of times under similar conditions

Subjective, epistemic or Bayesian

degree of reasonable belief

Propensity

tendency of a given type of physical situation to yield an outcome

...

Interpretation

Measurement of constant in nature (*e.g. the fine structure constant*) with a given measurement error (Gaussian noise with σ=0.001)

Result: 0.007114

What is the more probable value of the fine structure constant?

42 or 1/137

What is that even means? We have a single universe (what we can observe), with a definite value of this constant!

Interpretation

Measurement of constant in nature (*e.g. the fine structure constant*)
with a given measurement error (Gaussian noise with σ=0.001)

Result: 0.007114

What is the more probable value of the fine structure constant?
42 or 1/137

What is that even means? We have a single universe (what we can observe), with a definite value of this constant!

“extinction of the dinosaurs was probably caused by a large meteorite hitting the earth”

What does it mean? It was caused or it was not, there is no in between.

Bayesian Probability

Probability as

- Reasonable expectation
- Degree of belief
- State of knowledge

$$P(H|D)=\frac{P(D|H)P(H)}{P(D)}$$

Bayes' theorem

- have an epistemic / subjectivist interpretation
- is true independent of interpretations

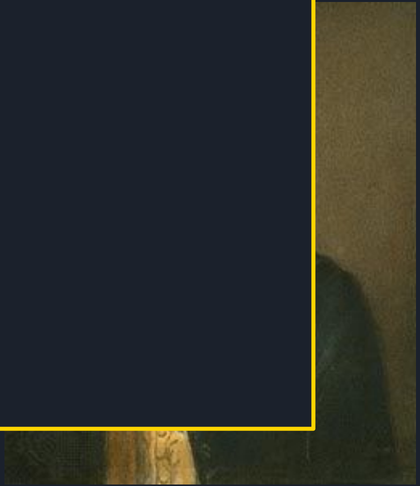
Thomas Bayes

Pierre-Simon Laplace

Bayesian Probability

$$P(H|D)=\frac{P(D|H)P(H)}{P(D)}$$

- have an epistemic / subjectivist interpretation
- true independent of interpretations



Pierre-Simon Laplace

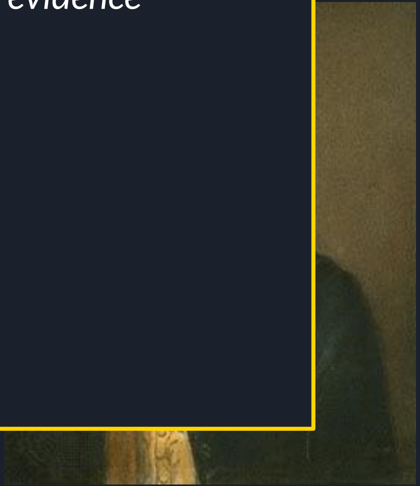
Bayesian Probability

Probability of the hypothesis *before observing the evidence*

“a priori”

$$P(H|D) = \frac{P(D|H)\overbrace{P(H)}^{\text{prior}}}{P(D)}$$

- have an epistemic / subjectivist interpretation
- true independent of interpretations



Pierre-Simon Laplace

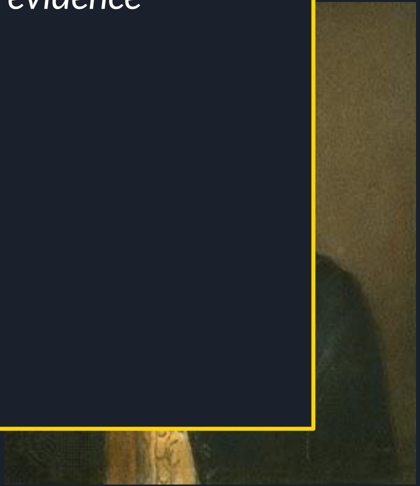
Bayesian Probability

Probability of the hypothesis ***before*** *observing the evidence*

“a priori”

$$P(H|D)=\frac{\overbrace{P(D|H)}^{\text{likelihood}}\overbrace{P(H)}^{\text{prior}}}{P(D)}$$

- have an epistemic / subjectivist interpretation
- true independent of interpretations



Pierre-Simon Laplace

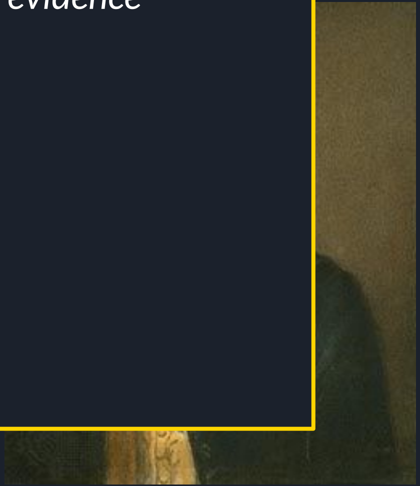
Bayesian Probability

Probability of the hypothesis
after observing the evidence
"a posteriori"

Probability of the hypothesis
before observing the evidence
"a priori"

$$\overbrace{P(H|D)}^{\text{posterior}} = \frac{\overbrace{P(D|H)}^{\text{likelihood}}\,\overbrace{P(H)}^{\text{prior}}}{P(D)}$$

- have an epistemic / subjectivist interpretation
- true independent of interpretations



Pierre-Simon Laplace

Bayesian Probability

Probability of the hypothesis
after observing the evidence
“a posteriori”

Probability of the hypothesis
before observing the evidence
“a priori”

$$\overbrace{P(H|D)}^{\text{posterior}} = \frac{\overbrace{P(D|H)}^{\text{likelihood}}\overbrace{P(H)}^{\text{prior}}}{\underbrace{P(D)}_{\text{marginal likelihood}}}$$

$$P(D) = \int P(D|H)P(H)\,dH$$

- have an epistemic / subjectivist interpretation
- true independent of interpretations

Pierre-Simon Laplace

Bayesian Probability

Probability of the hypothesis
after observing the evidence
"a posteriori"

Probability of the hypothesis
before observing the evidence
"a priori"

$$\overbrace{P(H|D)}^{\text{posterior}} = \frac{\overbrace{P(D|H)}^{\text{likelihood}}\,\overbrace{P(H)}^{\text{prior}}}{\underbrace{P(D)}_{\text{marginal likelihood}}}$$

$$P(D) = \int P(D|H)\,P(H)\,dH$$

This integral is most often intractable!

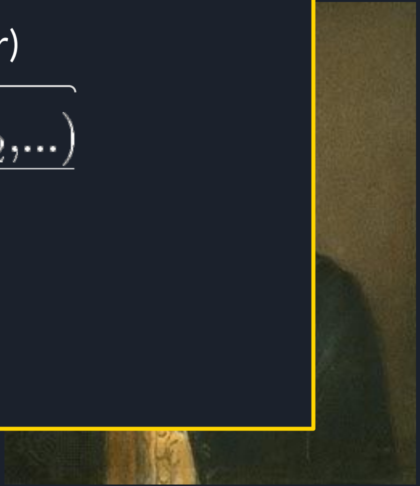
- have an epistemic / subjectivist interpretation
- true independent of interpretations

Pierre-Simon Laplace

Bayesian Probability

$$\overbrace{P(H|D_i,D_{i-1},\ldots)}^{\text{updated posterior}}=\frac{\overbrace{P(D_i|H)}^{\text{likelihood}}\overbrace{P(H|D_{i-1},D_{i-2},\ldots)}^{\text{updated prior (previous posterior)}}}{\underbrace{P(D_i)}_{\text{marginal likelihood}}}$$

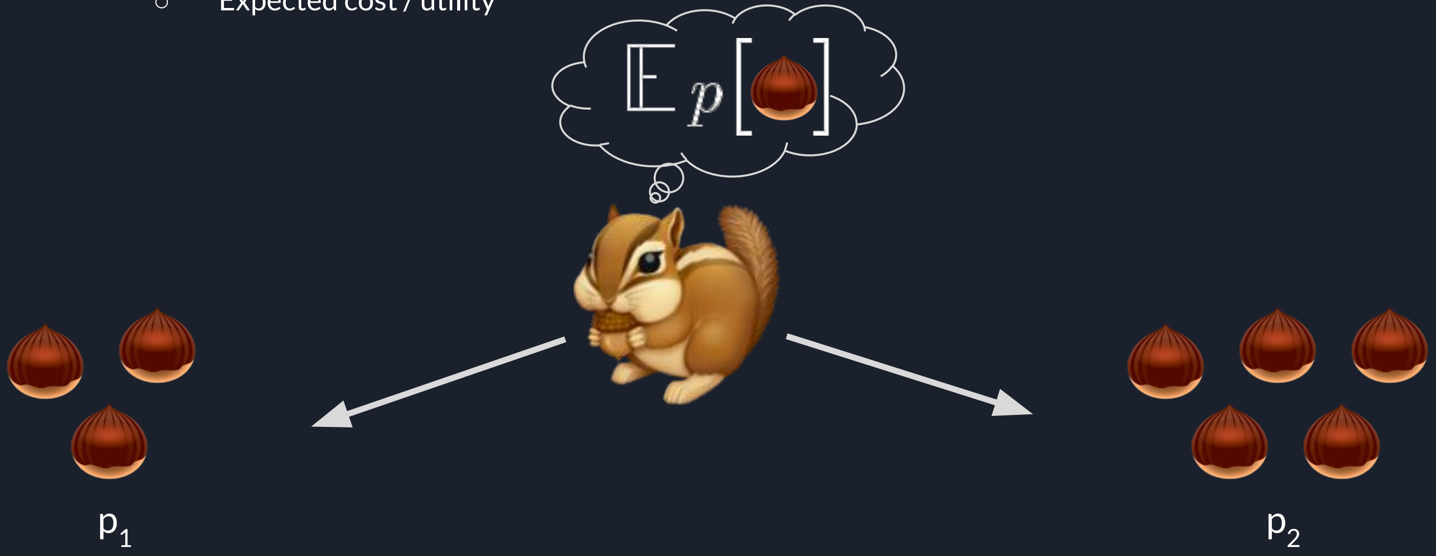
- have an epistemic / subjectivist interpretation
- true independent of interpretations



Pierre-Simon Laplace

Objectivist interpretation

- Rational agents hold consistent beliefs with reality.
- Cost/Utility function.
  - $$, hazelnuts, …
- Decision theory
  - Expected cost / utility



Steve is a random american guy.

“Steve is very shy and withdrawn, invariably helpful but with very little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.”

Is Steve more likely to be a librarian or a farmer?

/ Daniel Kahneman: *Thinking Fast, Thinking Slow* /

Importance of the prior

Steve is a random american guy.

“Steve is very shy and withdrawn, invariably helpful but with very little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.”

Is Steve more likely to be a librarian or a farmer?

/ Daniel Kahneman: *Thinking Fast, Thinking Slow* /

Far more farmers are in the USA than librarians. Even assuming extreme bias in behaviour, it is still more likely the Steve is a **farmer**.

Importance of the prior

What frequentist p-values mean?

- It does not tell you how likely your null hypothesis is! (see also Experimental Computational Work talk)
- To tell that, you need a prior!

It tells you: How likely it is that the null hypothesis would have produced the test statistics you observed.

More alike with P(D|H) than with P(H|D).

(Explainable AI talk)

Reference class problem

Assuming I am an average human being, when will the world end?

Premise: The human population grows exponentially until the end

Reference class problem

Assuming I am an average human being, when will the world end?

Premise: The human population grows exponentially until the end

-> The most people who ever lived will live in the last generations before the end.

Reference class problem

Assuming I am an average human being, when will the world end?

Premise: The human population grows exponentially until the end

-> The most people who ever lived will live in the last generations before the end.

-> Most likely I am one of these people

Reference class problem

Assuming I am an average human being, when will the world end?

Premise: The human population grows exponentially until the end

-> The most people who ever lived will live in the last generations before the end.

-> Most likely I am one of these people -> The end is near. 😨

Reference class problem

Assuming I am an average human being, when will the world end?

Premise: The human population grows exponentially until the end

-> The most people who ever lived will live in the last generations before the end.

-> Most likely I am one of these people -> The end is near.

This is known as the **Doomsday argument**, and have serious implications in cosmology.

See also: Sleeping Beauty Problem.

Bayesian Models

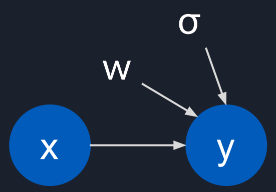
Elements of probabilistic models

Random variables

$$X \sim \mathcal{N}(0,1)$$

Dependencies

$$Y \sim \mathcal{N}(2X,0.5)$$



Parameters

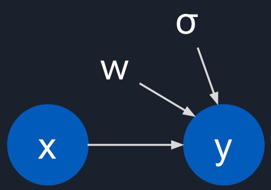
$$Y \sim \mathcal{N}(wX,\sigma)$$

Probabilistic graphical models (PGMs), Bayesian networks WARNING!

Classical and Bayesian models

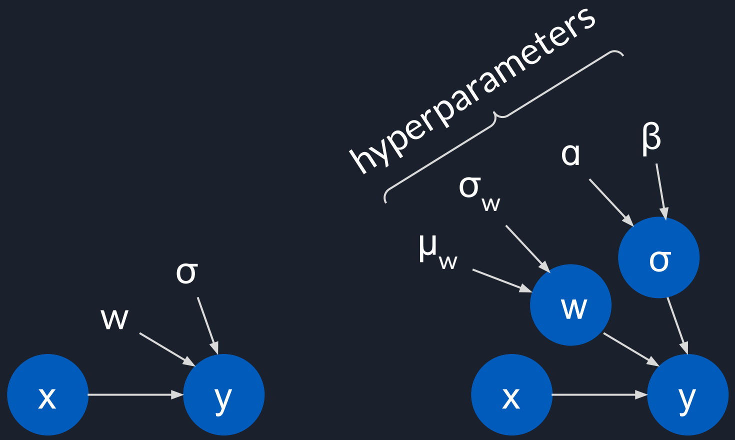
Classical point parametrization:

- Parameters and variables are distinct
- Parameters expected to have a “true” value
- Fitting the model / learning : optimizing for these parameters



point parametrization

Bayesian models



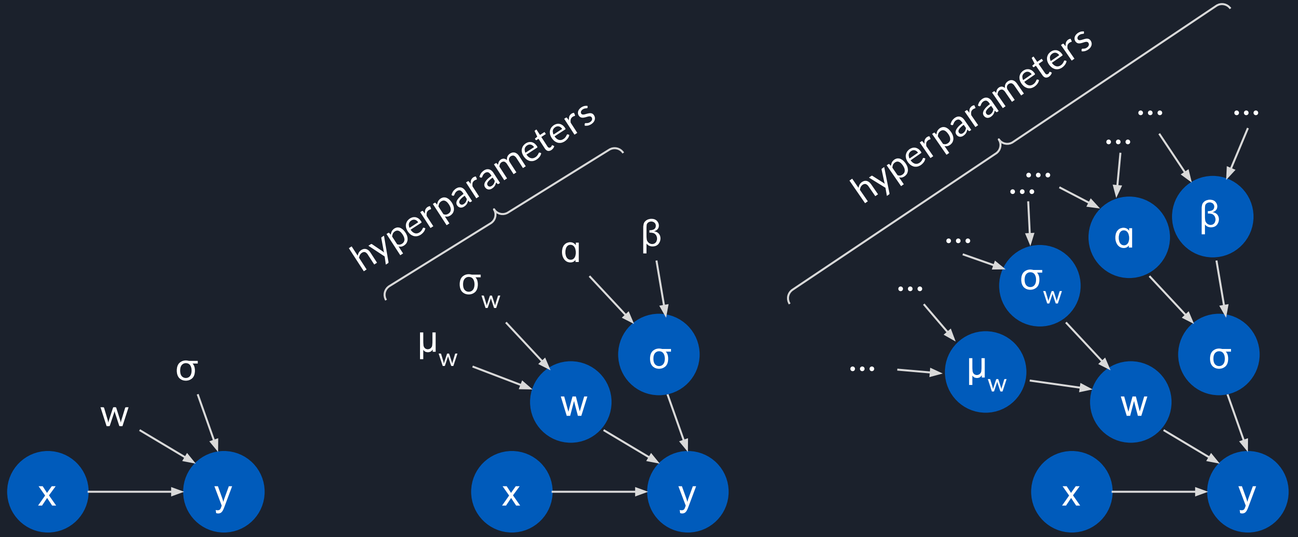
point parametrization

Bayesian

Bayesian treatment:

- Parameter is just a random variable
- We do not expect to find ‘the real’ parameter value exactly
- We search for the distribution of the parameters supported by the data.

Bayesian models



point parametrization

Bayesian

hierarchical Bayesian

Frequently used shorthand notations

1) Vector, matrix, tensor valued random variables

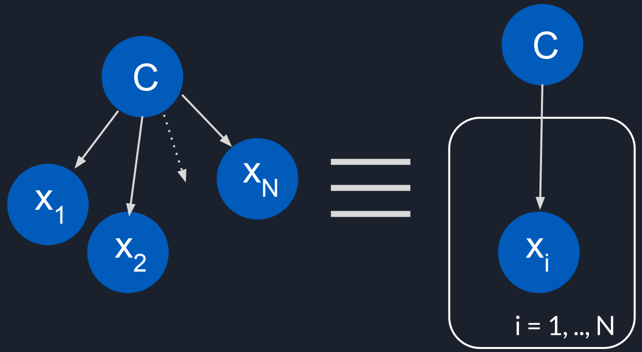
$$y \sim \mathcal{N}(W^\top x, \Sigma)$$

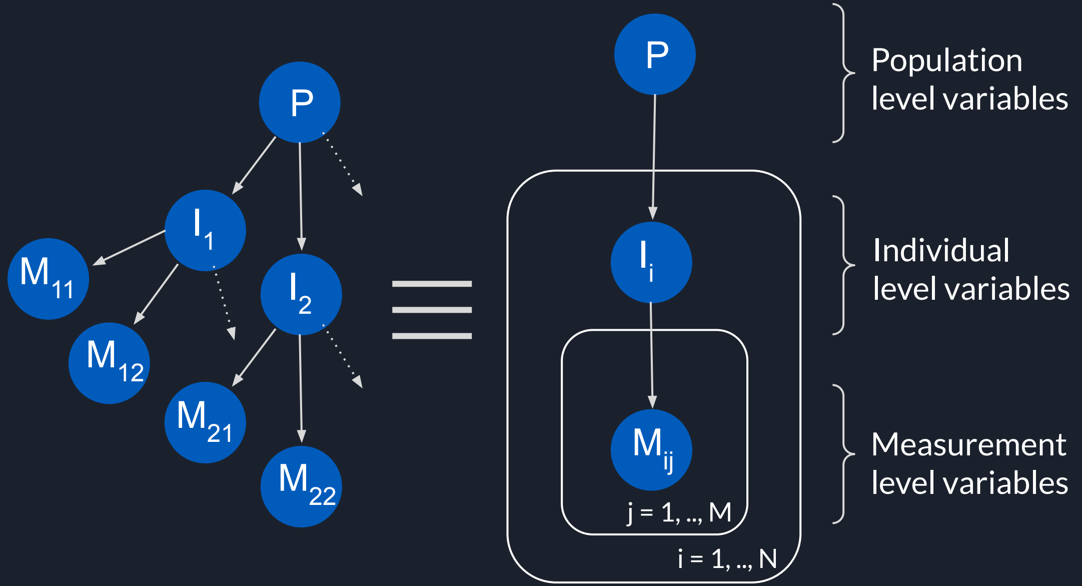
Frequently used shorthand notations

1) Vector, matrix, tensor valued random variables

$$y \sim \mathcal{N}(W^\top x, \Sigma)$$

2) Plate models





Bayesian Inference

Conjugate priors

$$p(x|\theta) = \mathcal{N}(x|\mu=0, \sigma^2=\theta)$$

We want to infer the distribution of θ given x: $p(\theta|x) \propto p(x|\theta)p(\theta)$

If we could choose the form of the prior, that the posterior have a known form, we don’t need to worry about the proportionality constant.

Conjugate priors

$$p(x|\theta)=\mathcal{N}(x|\mu=0,\sigma^2=\theta)$$

We want to infer the distribution of θ given x: $p(\theta|x) \propto p(x|\theta)p(\theta)$

If we could choose the form of the prior, that the posterior have a known form, we don’t need to worry about the proportionality constant.

$$\overbrace{c_1 \tfrac{1}{\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}^{p(x|\theta)} \overbrace{c_2(\sigma 2)^{-\alpha-1} e^{-\beta/\sigma 2}}^{p(\theta)}$$

Conjugate priors

$\theta \longrightarrow x$ $\quad p(x|\theta) = \mathcal{N}(x|\mu=0, \sigma^2=\theta)$

We want to infer the distribution of θ given x: $p(\theta|x) \propto p(x|\theta)p(\theta)$

If we could choose the form of the prior, that the posterior have a known form, we don't need to worry about the proportionality constant.

$$\overbrace{c_1 \boxed{\frac{1}{\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}^{p(x|\theta)} \overbrace{c_2 \boxed{(\sigma^2)^{-\alpha-1}} e^{-\beta/\sigma^2}}^{p(\theta)}$$

$$\alpha_N = \alpha + \frac{1}{2}$$

Conjugate priors

$\theta \rightarrow x$ $\quad p(x|\theta) = \mathcal{N}(x|\mu=0, \sigma^2=\theta)$

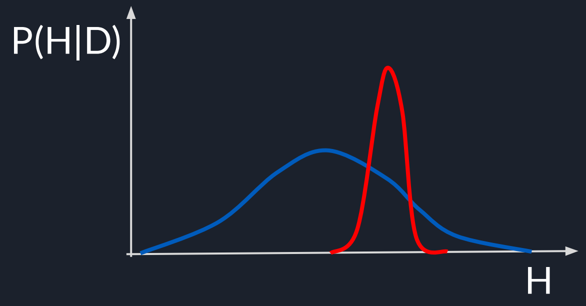
We want to infer the distribution of θ given x: $p(\theta|x) \propto p(x|\theta)p(\theta)$

If we could choose the form of the prior, that the posterior have a known form, we don't need to worry about the proportionality constant.

$$\overbrace{c_1 \boxed{\frac{1}{\sigma}} \boxed{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}}^{p(x|\theta)} \overbrace{c_2 \boxed{(\sigma^2)^{-\alpha-1}} \boxed{e^{-\beta/\sigma^2}}}^{p(\theta)}$$

$$\alpha_N = \alpha + \frac{1}{2} \qquad \beta_N = \beta + \frac{(x-\mu)^2}{2}$$

Maximum a posteriori (MAP) approximation



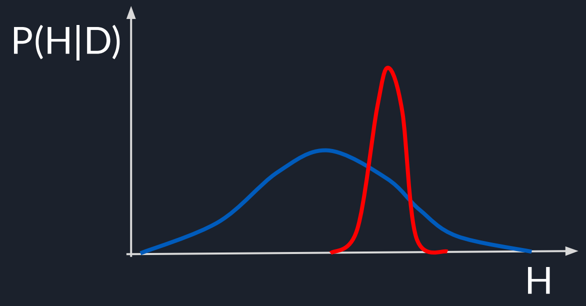
Small amount of data: flat posterior

- No point estimate is a good approximation

Lot of data: peaky posterior

- Most probable hypothesis is a good approximation

Maximum a posteriori (MAP) approximation



Small amount of data: flat posterior

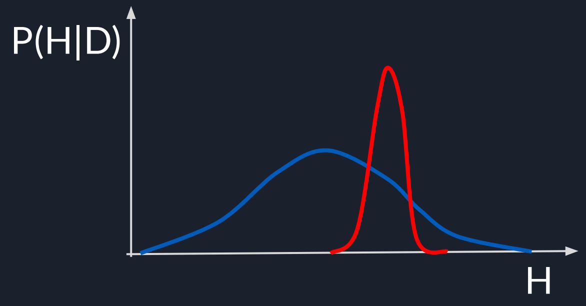
- No point estimate is a good approximation

Lot of data: peaky posterior

- Most probable hypothesis is a good approximation

Example: $\max_w \underbrace{\log p(x,y|w)}_{\text{likelihood}} + \underbrace{\log p(w)}_{\text{regularization}} - \cancel{\log p(x,y)}$ (additive constant)

Maximum a posteriori (MAP) approximation



Small amount of data: flat posterior

- No point estimate is a good approximation

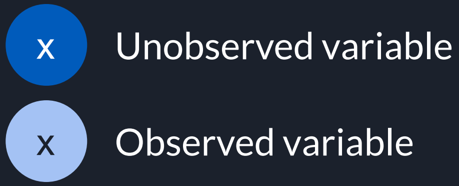
Lot of data: peaky posterior

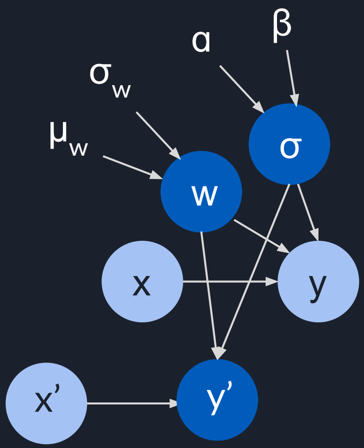
- Most probable hypothesis is a good approximation

Example: $\max_w \underbrace{\log p(x,y|w)}_{\text{likelihood}} + \underbrace{\log p(w)}_{\text{regularization}} - \cancel{\log p(x,y)}$ (additive constant)

If $p(w) = \mathcal{N}(w|0, 2/\lambda)$, then $log\, p(w) = -\lambda w^2$

Predictive inference



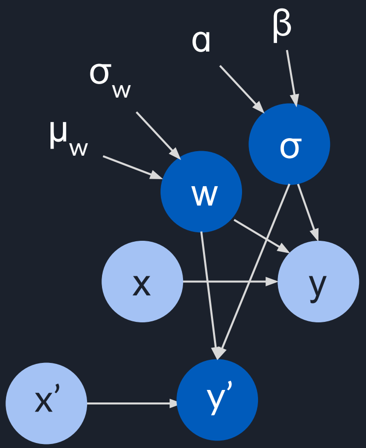


Predicted outcome? Just another random variable

Predictive inference

x Unobserved variable

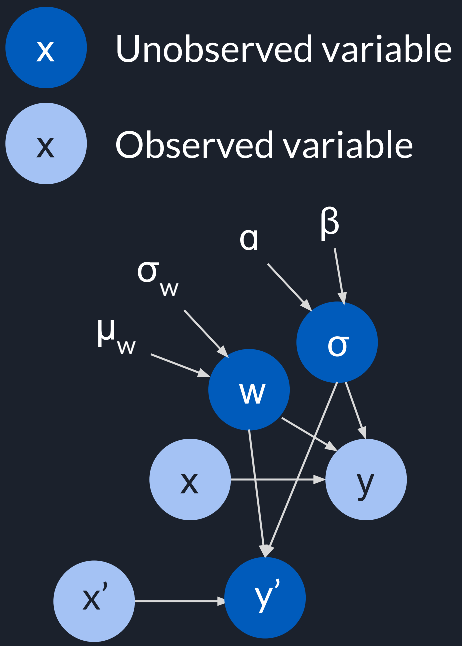
x Observed variable



Predicted outcome? Just another random variable

$$\underbrace{P(y'|x',\underbrace{x,y}_{\mathcal{D}})}_{\text{Posterior predictive distribution}} = P(y'|x',w,\sigma)\underbrace{P(w,\sigma|\underbrace{x,y}_{\mathcal{D}})}_{\text{(model) posterior}}$$

Predictive inference



Predicted outcome? Just another random variable

$$\underbrace{P(y'|x',\underbrace{x,y}_{\mathcal{D}})}_{\text{Posterior predictive distribution}}=P(y'|x',w,\sigma)\underbrace{P(w,\sigma|\underbrace{x,y}_{\mathcal{D}})}_{\text{(model) posterior}}$$

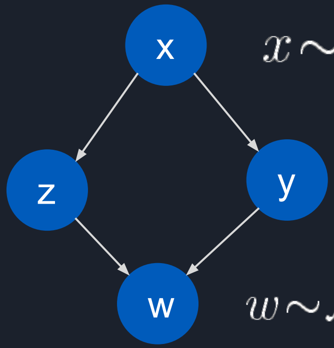
What is the mean?

$$\mathbb{E}[y'|x',\mathcal{D}]=\mathbb{E}_{P(w,\sigma|\mathcal{D})}[f_{w,\sigma}(x')]$$

"Bayesian model averaging"

Note the similarity with ensembles!

Ancestral sampling



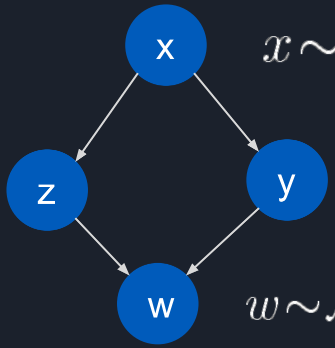
$$x \sim \mathcal{N}(0,1)$$

$$z \sim \mathcal{N}(x/2, 0.3)$$

$$y \sim \mathcal{N}(x, 0.5)$$

$$w \sim \mathcal{N}(z+y, |y|)$$

Ancestral sampling



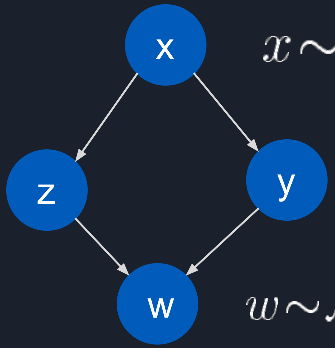
$x \sim \mathcal{N}(0,1)$ -6.44

$z \sim \mathcal{N}(x/2, 0.3)$

$y \sim \mathcal{N}(x, 0.5)$

$w \sim \mathcal{N}(z+y, |y|)$

Ancestral sampling



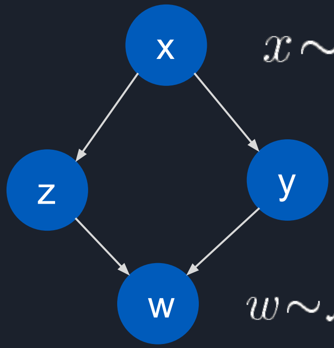
$x \sim \mathcal{N}(0,1)$ -6.44

-3.25 $z \sim \mathcal{N}(x/2, 0.3)$

$y \sim \mathcal{N}(x, 0.5)$

$w \sim \mathcal{N}(z+y, |y|)$

Ancestral sampling



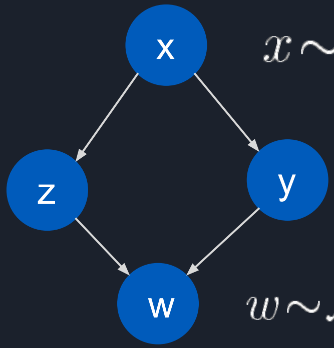
$x \sim \mathcal{N}(0,1)$ -6.44

-3.25 $z \sim \mathcal{N}(x/2, 0.3)$

$y \sim \mathcal{N}(x, 0.5)$ -6.35

$w \sim \mathcal{N}(z+y, |y|)$

Ancestral sampling



$x \sim \mathcal{N}(0,1)$ -6.44

-3.25 $z \sim \mathcal{N}(x/2, 0.3)$

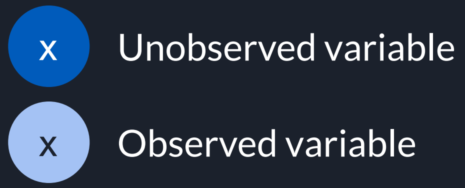
$y \sim \mathcal{N}(x, 0.5)$ -6.35

$w \sim \mathcal{N}(z+y, |y|)$ 1.54

Gibbs sampling

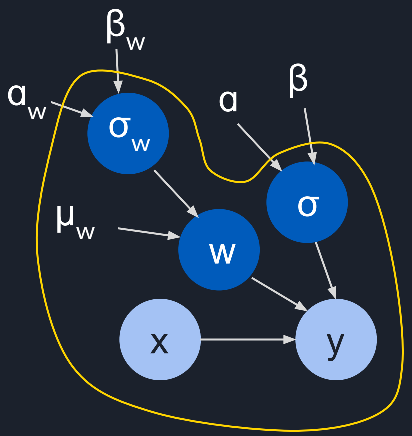
Assumption: Conditional posteriors are analytically tractable
"Imagine all but one variable observed"

Iteratively sample from the conditionals.



A variable depends on its Markov blanket:

- ancestors
- descendants
- other ancestors of the descendants
- "the parents, children and spouses"



Example: for **w** we want: $P(w|x,y,\sigma,\sigma_w)$

We have: $P(w,y|x,\sigma,\sigma_w) = P(y|w,x,\sigma)P(w|\sigma_w)$

Gibbs sampling

Assumption: Conditional posteriors are analytically tractable
“Imagine all but one variable observed”
Iteratively sample from the conditionals.

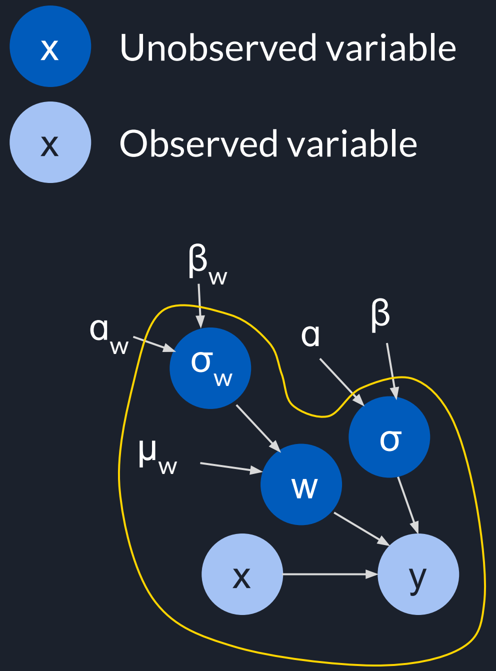
A variable depends on its Markov blanket:

- ancestors
- descendants
- other ancestors of the descendants
- “the parents, children and spouses”

Example: for **w** we want: $P(w|x,y,\sigma,\sigma_w)$

We have: $P(w,y|x,\sigma,\sigma_w) = P(y|w,x,\sigma)P(w|\sigma_w)$

$$P(w|x,\sigma,\sigma_w) = \frac{P(y|w,x,\sigma)P(w|\sigma_w)}{P(y|x,\sigma,\sigma_w)}$$

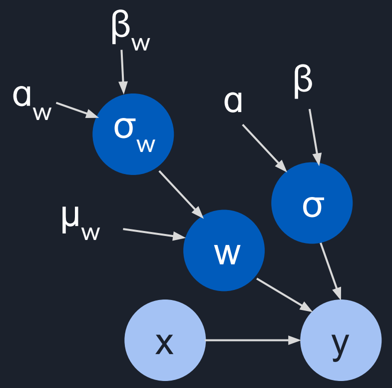


Variational inference

If we would have $P(w,\sigma|x,y)$ in analytical form, our job would be done.

- Often there is no such analytical form
- Search for a function $q(w,\sigma) \approx P(w,\sigma|x,y)$

What is our loss function?



Variational inference

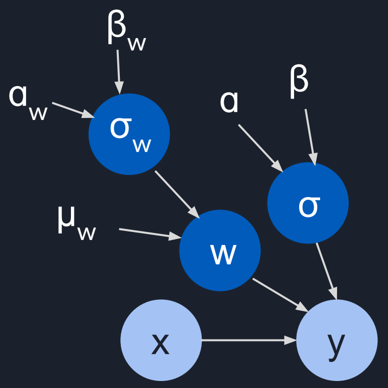
If we would have $P(w,\sigma|x,y)$ in analytical form, our job would be done.

- Often there is no such analytical form
- Search for a function $q(w,\sigma) \approx P(w,\sigma|x,y)$

What is our loss function?

$$\min_{\phi} D_{KL}(q_{\phi}(w,\sigma) \| p(w,\sigma|x,y))$$

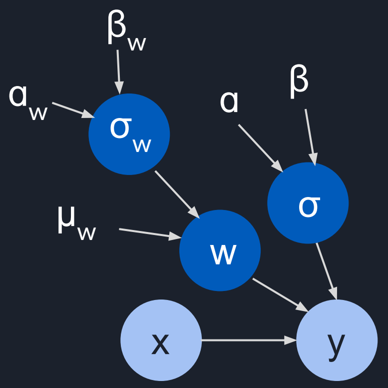
ϕ: variational parameter



Variational inference

If we would have $P(w,\sigma|x,y)$ in analytical form, our job would be done.

- Often there is no such analytical form
- Search for a function $q(w,\sigma)\approx P(w,\sigma|x,y)$



What is our loss function?

$$\min_{\phi} D_{KL}(q_\phi(w,\sigma)\|p(w,\sigma|x,y))$$

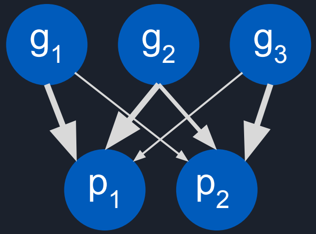
ϕ: variational parameter

It can be shown that equivalently we can take the following objective:

$$\max_{\phi} \underbrace{-\mathbb{E}_{q(w,\sigma)}[\log(q(w,\sigma))-\log(p(x,y|w,\sigma)p(w,\sigma))]}_{\text{Evidence lower bound (ELBO)}}$$

Outlook

Being Bayesian over the structure



How likely it is that gene 1 associates with phenotype 1?

$$P(\mathcal{G}|D) = \frac{P(D|\mathcal{G})P(\mathcal{G})}{P(D)}$$

$$P(D|\mathcal{G}) = \underbrace{\int P(D|\theta,\mathcal{G})P(\theta|\mathcal{G})d\theta}$$

Marginal likelihood conditioned on the structure.

"average all models"

Model class: $\mathcal{G}(V, E)$ directed acyclic graphs (DAGs).

V: Set of nodes, corresponds to set of random variables.

E: Set of edges, dependence relations.

The graph is just a random variable.

Mechanistic interpretation - towards causality

Observational equivalence

$$x \rightarrow y \rightarrow z$$

$$x \leftarrow y \leftarrow z$$

$$x \leftarrow y \rightarrow z$$

$$X \not\perp Z$$
$$X \perp Z \mid Y$$

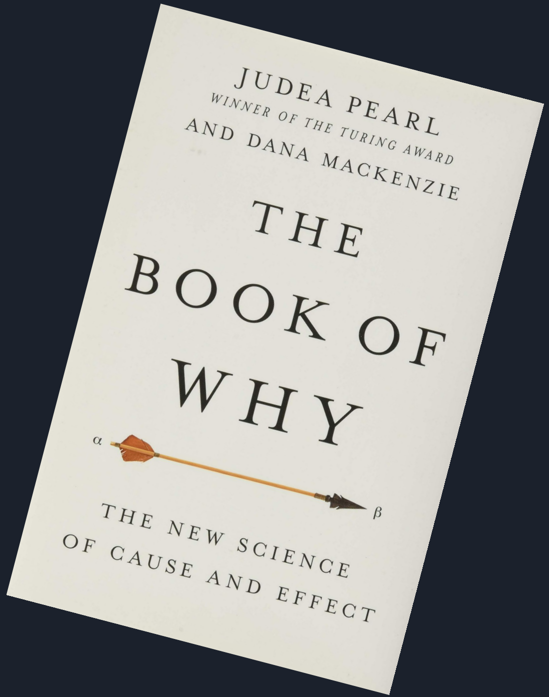
$$x \rightarrow y \leftarrow z$$

$$X \perp Z$$
$$X \not\perp Z \mid Y$$

$$p(x|do(y))$$

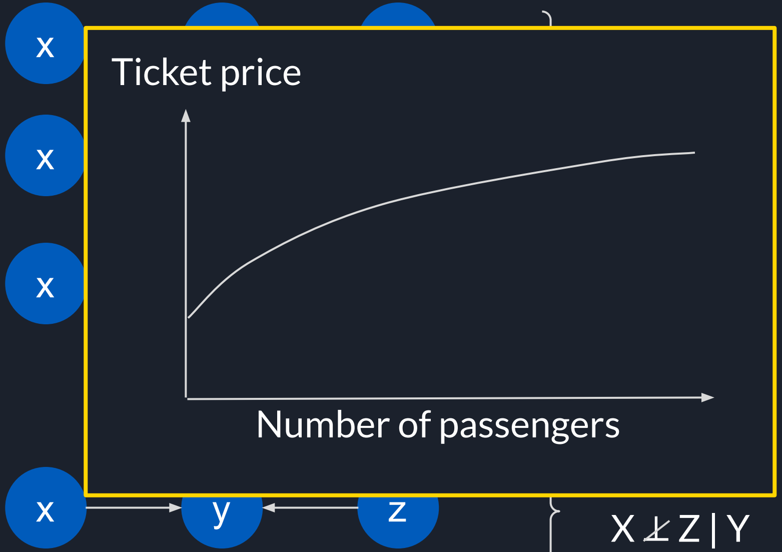
$$= p(x)$$

$$= p(x|y)$$



Mechanistic interpretation - towards causality

Observational equivalence

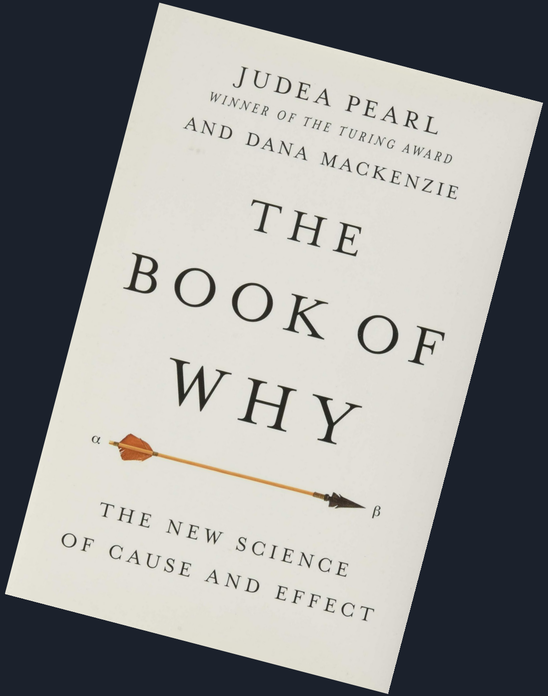


$$p(x|do(y))$$

$$= p(x)$$

$$= p(x|y)$$

$X \not\perp Z \mid Y$



Thank you for your attention!