# Basics of Model-Based Learning

Michael U. Gutmann

Probabilistic Modelling and Reasoning (INFR11134)
School of Informatics, The University of Edinburgh

Spring Semester 2024

# Recap

$$p(\mathbf{x}|\mathbf{y}_o) = \frac{\sum_{\mathbf{z}} p(\mathbf{x},\mathbf{y}_o,\mathbf{z})}{\sum_{\mathbf{x},\mathbf{z}} p(\mathbf{x},\mathbf{y}_o,\mathbf{z})}$$

Assume that $\mathbf{x}, \mathbf{y}, \mathbf{z}$ each are $d = 500$ dimensional, and that each element of the vectors can take $K = 10$ values.

- Issue 1: To specify $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$, we need to specify $K^{3d} - 1 = 10^{1500} - 1$ non-negative numbers, which is impossible.

  Topic 1: Representation What reasonably weak assumptions can we make to efficiently represent $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$?
- Directed and undirected graphical models, factor graphs
- Factorisation and independencies

# Recap

$$p(\mathbf{x}|\mathbf{y}_o) = \frac{\sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{y}_o, \mathbf{z})}{\sum_{\mathbf{x},\mathbf{z}} p(\mathbf{x}, \mathbf{y}_o, \mathbf{z})}$$

- Issue 2: The sum in the numerator goes over the order of $K^d = 10^{500}$ non-negative numbers and the sum in the denominator over the order of $K^{2d} = 10^{1000}$, which is impossible to compute.

  Topic 2: Exact inference Can we further exploit the assumptions on $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ to efficiently compute the posterior probability or derived quantities?
- Yes! Factorisation can be exploited by using the distributive law and by caching computations.
- Variable elimination and message passing algorithms
- Inference for hidden Markov models

# Recap

$$p(\mathbf{x}|\mathbf{y}_o) = \frac{\sum_{\mathbf{z}} p(\mathbf{x},\mathbf{y}_o,\mathbf{z})}{\sum_{\mathbf{x},\mathbf{z}} p(\mathbf{x},\mathbf{y}_o,\mathbf{z})}$$

- Issue 3: Where do the non-negative numbers $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ come from?

  Topic 3: Learning How can we learn the numbers from data?

# Program

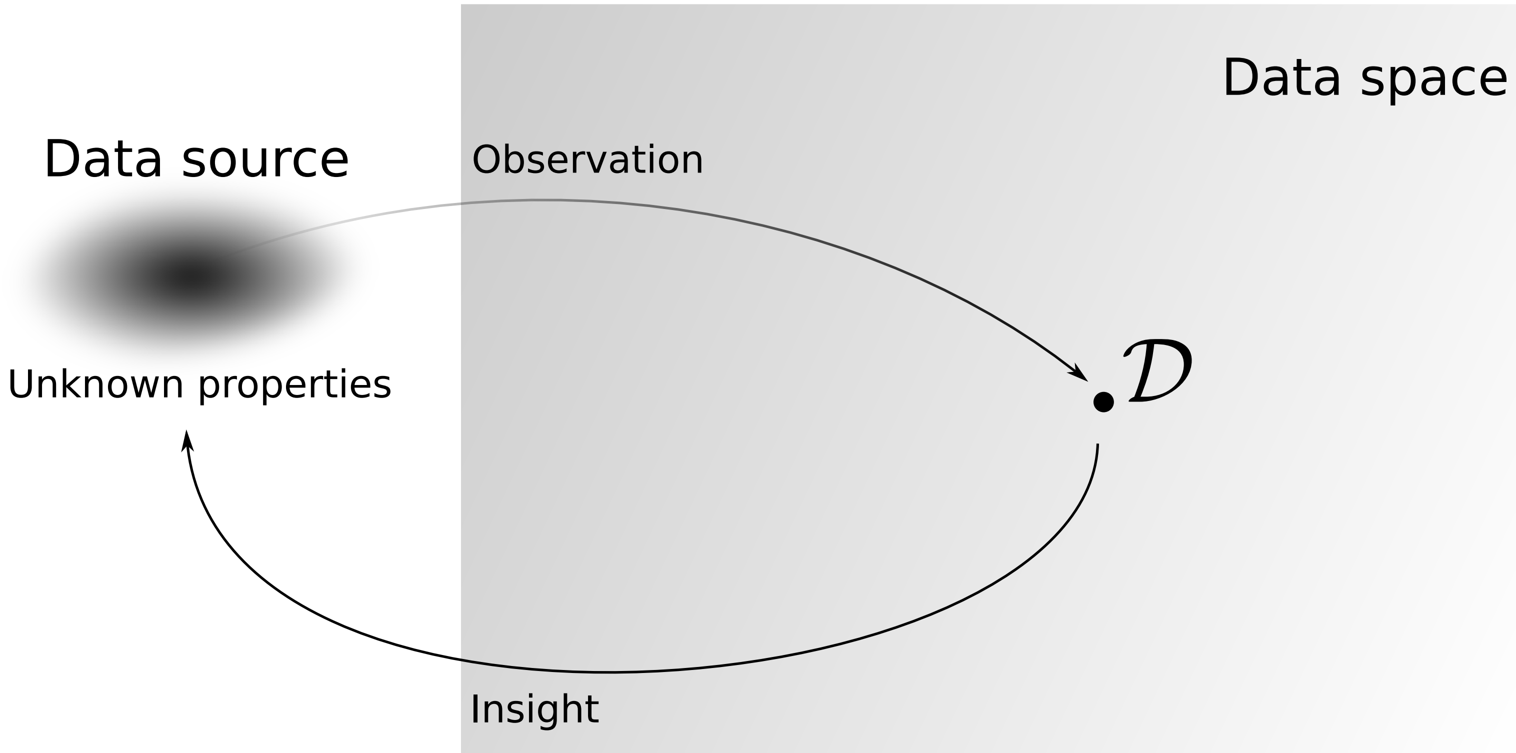
1. Basic concepts
2. Learning by maximum likelihood estimation
3. Learning by Bayesian inference

# Program

1. Basic concepts
   - Observed data as a sample drawn from an unknown data generating distribution
   - Probabilistic, statistical, and Bayesian models
   - Partition function and unnormalised statistical models
   - Learning = parameter estimation or learning = Bayesian inference
2. Learning by maximum likelihood estimation
3. Learning by Bayesian inference

# Learning from data

- Use observed data $\mathcal{D}$ to learn about their source
- Enables probabilistic inference, decision making, ...



# Data

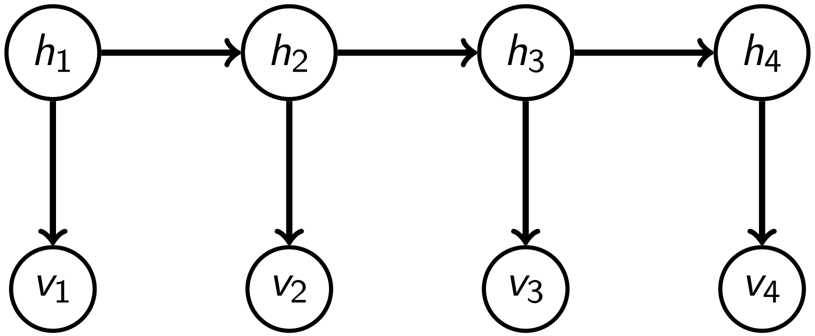
- We typically assume that the observed data $\mathcal{D}$ correspond to a random sample (draw) from an unknown distribution $p_*(\mathcal{D})$

$$\mathcal{D} \sim p_*(\mathcal{D})$$

- In other words, we consider the data $\mathcal{D}$ to be a realisation (observation) of a random variable with distribution $p_*$.

# Data

- Example: You use some transition and emission distribution and generate data from the hidden Markov model. (e.g. via ancestral sampling)




- You know the visibles $(v_1, v_2, v_3, \ldots, v_T) \sim p(v_1, \ldots, v_T)$.
- You give the generated visibles to a friend who does not know about the distributions that you used, nor possibly that you used a HMM. For your friend:

$$\mathcal{D} = (v_1, v_2, v_3, \ldots, v_T) \qquad \mathcal{D} \sim p_*(\mathcal{D})$$

# Independent and identically distributed (iid) data

- Let $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. If

$$p_*(\mathcal{D}) = \prod_{i=1}^{n} p_*(\mathbf{x}_i)$$

  then the data (or the corresponding random variables) are said to the iid. $\mathcal{D}$ is also said to be a random sample from $p_*$.
- In other words, the $\mathbf{x}_i$ were independently drawn from the same distribution $p_*(\mathbf{x})$.
- Example: $n$ time series $(v_1, v_2, v_3, \ldots, v_T)$ each independently generated with the same transition and emission distribution.

# Independent and identically distributed (iid) data

- Example: Generate $n$ samples $(x_1^{(i)}, \ldots, x_5^{(i)})$ from

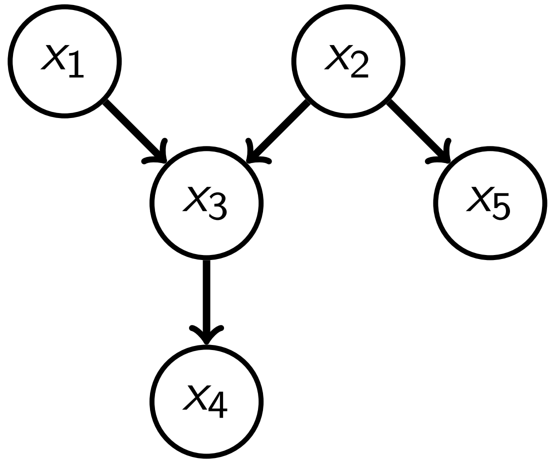
$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3)p(x_5|x_2)$$

  with known conditionals, using e.g. ancestral sampling.

- You collect the $n$ observed values of $x_4$, i.e.

$$x_4^{(1)}, \ldots, x_4^{(n)}$$

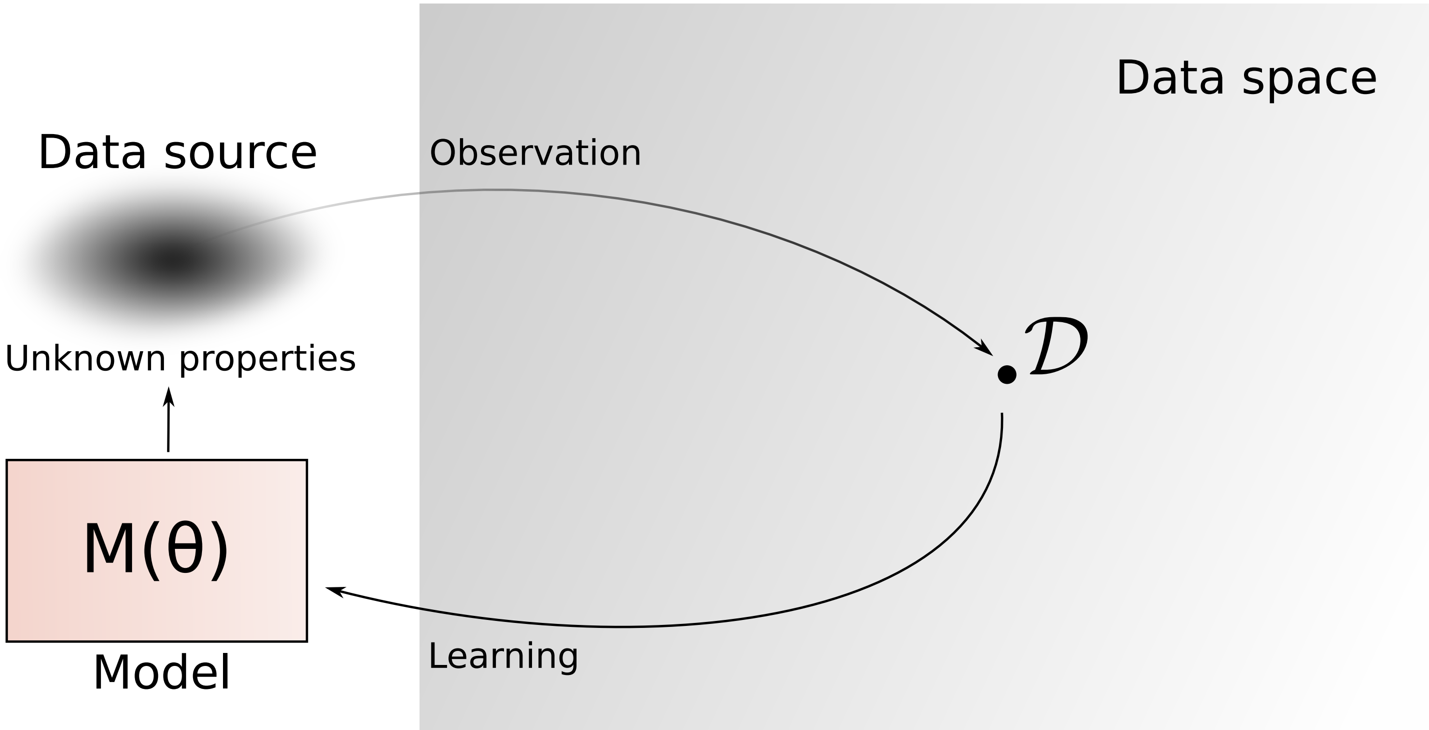
  and give them to a friend who does not know how you generated the data but that they are iid.




- For your friend, the $x_4^{(i)}$ are data points $x_i \sim p_*$.
- Remark: if the subscript index is occupied, we often use superscripts to enumerate the data points.

# Using models to learn from data

- Set up a model with properties that the unknown data source might have.
- The potential properties are the parameters $\boldsymbol{\theta}$ of the model.
- Model may include independence assumptions.
- Learning: Assess which $\boldsymbol{\theta}$ are in line with the observed data $\mathcal{D}$.



# Models

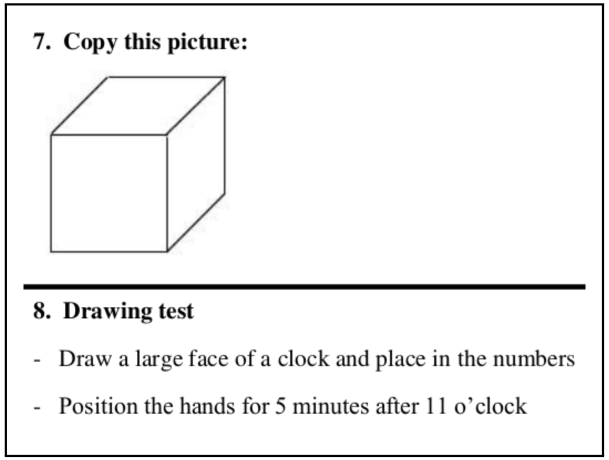
- The term "model" has multiple meanings, see e.g. `https://en.wikipedia.org/wiki/Model`
- In our course:
  - probabilistic model
  - statistical model
  - Bayesian model
- See Section 3 in the background document *Introduction to Probabilistic Modelling*
- Note: the three types are often confounded, and often just called probabilistic or statistical model, or just "model".

# Probabilistic model

Example from the first lecture: cognitive impairment test

- Sensitivity of 0.8 and specificity of 0.95 (Scharre, 2010)
- *Probabilistic* model for presence of impairment ($x = 1$) and detection by the test ($y = 1$):

$$
\begin{aligned}
\mathbb{P}(x = 1) &= 0.11 \quad \text{(prior)} \\
\mathbb{P}(y = 1 | x = 1) &= 0.8 \quad \text{(sensitivity)} \\
\mathbb{P}(y = 0 | x = 0) &= 0.95 \quad \text{(specificity)}
\end{aligned}
$$




(Example from sagetest.osu.edu)

- From first lecture:
  A probabilistic model is an abstraction of reality that uses probability theory to quantify the chance of uncertain events.

# Probabilistic model

- More technically:
  probabilistic model $\equiv$ probability distribution (pmf/pdf).
- Probabilistic model was written in terms of the probability $\mathbb{P}$.
  In terms of the pmf it is

$$
\begin{aligned}
p_x(1) &= 0.11 \\
p_{y|x}(1|1) &= 0.8 \\
p_{y|x}(0|0) &= 0.95
\end{aligned}
$$

- Commonly written as

$$
\begin{aligned}
p(x = 1) &= 0.11 \\
p(y = 1|x = 1) &= 0.8 \\
p(y = 0|x = 0) &= 0.95
\end{aligned}
$$

where the notation for probability measure $\mathbb{P}$ and pmf $p$ are confounded.

# Statistical model

- If we substitute the numbers with parameters, we obtain a (parametric) *statistical* model

$$
\begin{aligned}
p(x=1) &= \theta_1 \\
p(y=1|x=1) &= \theta_2 \\
p(y=0|x=0) &= \theta_3
\end{aligned}
$$

- For each value of the $\theta_i$, we obtain a different pmf. Dependency highlighted by writing

$$
\begin{aligned}
p(x=1; \theta_1) &= \theta_1 \\
p(y=1|x=1; \theta_2) &= \theta_2 \\
p(y=0|x=0; \theta_3) &= \theta_3
\end{aligned}
$$

- Or: $p(x, y; \boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ is a vector of parameters.
- A statistical model corresponds to a set of probabilistic models, here indexed by the parameters $\boldsymbol{\theta}$: $\{p(\mathbf{x}; \boldsymbol{\theta})\}_{\boldsymbol{\theta}}$

# Bayesian model

- In *Bayesian* models, we combine statistical models with a (prior) probability distribution on the parameters $\boldsymbol{\theta}$.
- Each member of the family $\{p(\mathbf{x}; \boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ is considered a conditional pmf/pdf of $\mathbf{x}$ given $\boldsymbol{\theta}$
- Use conditioning notation $p(\mathbf{x}|\boldsymbol{\theta})$
- The conditional $p(\mathbf{x}|\boldsymbol{\theta})$ and the pmf/pdf $p(\boldsymbol{\theta})$ for the (prior) distribution of $\boldsymbol{\theta}$ together specify the joint pmf/pdf via the product rule
$$p(\mathbf{x}, \boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$
- Bayesian model for $\mathbf{x}$ = probabilistic model for $(\mathbf{x}, \boldsymbol{\theta})$.
- The prior may be parameterised, e.g. $p(\boldsymbol{\theta}; \boldsymbol{\alpha})$. The parameters $\boldsymbol{\alpha}$ are called "hyperparameters".

# Graphical models as statistical models

- Directed or undirected graphical models are sets of probability distributions, e.g. all $p$ that factorise as

$$p(\mathbf{x}) = \prod_i k_i(x_i|\mathrm{pa}_i) \quad \text{or} \quad p(\mathbf{x}) \propto \prod_i \phi_i(\mathcal{X}_i)$$

  They are thus statistical models.
- If we consider parametric families for $k_i(x_i|\mathrm{pa}_i)$ and $\phi_i(\mathcal{X}_i)$, they correspond to parametric statistical models

$$p(\mathbf{x};\boldsymbol{\theta}) = \prod_i k_i(x_i|\mathrm{pa}_i;\boldsymbol{\theta}_i) \quad \text{or} \quad p(\mathbf{x};\boldsymbol{\theta}) \propto \prod_i \phi_i(\mathcal{X}_i;\boldsymbol{\theta}_i)$$

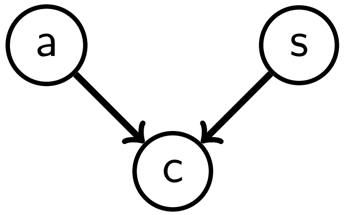
  where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots)$.

(on the next slides: will use again that $k_i(x_i|\mathrm{pa}_i) = p(x_i|\mathrm{pa}_i)$)

# Cancer-asbestos-smoking example (Barber Figure 9.4)

- Very simple toy example about the relationship between lung Cancer, Asbestos exposure, and Smoking

DAG:




Factorisation:
$p(c, a, s) = p(c|a, s)p(a)p(s)$

Parametric models: (for binary vars)

$$p(a = 1; \theta_a) = \theta_a$$
$$p(s = 1; \theta_s) = \theta_s$$

| $p(c = 1 \mid a, s; \boldsymbol{\theta}_c)$ | $a$ | $s$ |
|---|---|---|
| $\theta_c^1$ | 0 | 0 |
| $\theta_c^2$ | 1 | 0 |
| $\theta_c^3$ | 0 | 1 |
| $\theta_c^4$ | 1 | 1 |

All parameters are $\geq 0$

- Factorisation + parametric models for the factors gives the parametric statistical model

$$p(c, a, s; \boldsymbol{\theta}) = p(c|a, s; \boldsymbol{\theta}_c)p(a; \theta_a)p(s; \theta_s) \qquad \boldsymbol{\theta} = (\theta_a, \theta_s, \boldsymbol{\theta}_c)$$

# Cancer-asbestos-smoking example

- The model specification $p(a = 1; \theta_a) = \theta_a$ is equivalent to

$$
\begin{aligned}
p(a; \theta_a) &= (\theta_a)^a (1 - \theta_a)^{1-a} \\
&= \theta_a^{\mathbb{1}(a=1)} (1 - \theta_a)^{\mathbb{1}(a=0)}
\end{aligned}
$$

  Note: $(\theta_a)^a$ means parameter $\theta_a$ to the power of $a$.
- $a$ is a Bernoulli random variable with "success" probability $\theta_a$.
- Equivalently for $s$.

# Cancer-asbestos-smoking example

- Table parameterisation $p(c|a, s; \boldsymbol{\theta}_c)$, with $\boldsymbol{\theta}_c = (\theta_c^1, \ldots, \theta_c^4)$, can be written more compactly in similar form.
- Enumerate the states of the parents of $c$ so that

$$\mathrm{pa}_c = 1 \Leftrightarrow (a = 0, s = 0) \quad \ldots \quad \mathrm{pa}_c = 4 \Leftrightarrow (a = 1, s = 1)$$

- We then have

$$\begin{aligned} p(c|a, s; \boldsymbol{\theta}_c) &= \prod_{j=1}^{4} \left[ (\theta_c^j)^c (1 - \theta_c^j)^{1-c} \right]^{\mathbb{1}(\mathrm{pa}_c = j)} \\ &= \prod_{j=1}^{4} (\theta_c^j)^{\mathbb{1}(c=1, \mathrm{pa}_c = j)} (1 - \theta_c^j)^{\mathbb{1}(c=0, \mathrm{pa}_c = j)} \end{aligned}$$

  Product over the possible states of the parents and the possible states of $c$.
- Equivalent to the table but more convenient to manipulate.

# Cancer-asbestos-smoking example

- Working with the table representation does not shrink the set of probabilistic models.
- Set of $p(c, a, s)$ defined by the DAG = parametric family $\{p(c, a, s; \boldsymbol{\theta})\}_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ are the parameters in the table.
- Other parametric models are possible too:
  - As before but some parameters are tied, e.g. $\theta_c^2 = \theta_c^3$
  - $p(c = 1|a, s) = \sigma(w_0 + w_1 a + w_2 s)$ where $\sigma()$ is the sigmoid function $\sigma(u) = 1/(1 + \exp(-u))$.

In both cases, the parameterisation limits the space of possible probabilistic models.

(see slides *Basic Assumptions for Efficient Model Representation*)

# Cancer-asbestos-smoking example

- We can turn the table-based parametric model into a Bayesian model by assigning a (prior) probability distribution to $\boldsymbol{\theta}$
- Often: we assume independence of the parameters so that the prior pdf/pmf factorises, e.g.

$$p(\boldsymbol{\theta}) = p(\theta_a)p(\theta_s)\prod_{j=1}^{4} p(\theta_c^j)$$

- With correspondence $p(\mathbf{x}; \boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})$, the Bayesian model is

$$\begin{aligned}
p(\mathbf{x}, \boldsymbol{\theta}) &= p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\
&= \theta_a^{\mathbb{1}(a=1)}(1-\theta_a)^{\mathbb{1}(a=0)}p(\theta_a)\theta_s^{\mathbb{1}(s=1)}(1-\theta_s)^{\mathbb{1}(s=0)}p(\theta_s) \\
&\quad \prod_{j=1}^{4}(\theta_c^j)^{\mathbb{1}(c=1,\mathrm{pa}_c=j)}(1-\theta_c^j)^{\mathbb{1}(c=0,\mathrm{pa}_c=j)}\prod_{j=1}^{4} p(\theta_c^j)
\end{aligned}$$

- Note the factorisation.

# Program

1. Basic concepts
   - Observed data as a sample drawn from an unknown data generating distribution
   - Probabilistic, statistical, and Bayesian models
   - Partition function and unnormalised statistical models
   - Learning = parameter estimation or learning = Bayesian inference
2. Learning by maximum likelihood estimation
3. Learning by Bayesian inference

# Partition function

- pdfs/pmfs integrate/sum to one.
- Parameterised Gibbs distributions

$$p(\mathbf{x};\boldsymbol{\theta}) \propto \prod_i \phi_i(\mathcal{X}_i;\boldsymbol{\theta}_i)$$

  do typically not integrate/sum one.
- For normalisation, we can divide the unnormalised model $\tilde{p}(\mathbf{x};\boldsymbol{\theta}) = \prod_i \phi_i(\mathcal{X}_i;\boldsymbol{\theta}_i)$ by the partition function $Z(\boldsymbol{\theta})$,

$$Z(\boldsymbol{\theta}) = \int \tilde{p}(\mathbf{x};\boldsymbol{\theta})\mathrm{d}\mathbf{x} \quad \text{or} \quad Z(\boldsymbol{\theta}) = \sum_{\mathbf{x}} \tilde{p}(\mathbf{x};\boldsymbol{\theta}).$$

- By construction,

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{\tilde{p}(\mathbf{x};\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}$$

  sums/integrates to one for all values of $\boldsymbol{\theta}$.

# Unnormalised statistical models

- If each element of $\{p(\mathbf{x};\boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ integrates/sums to one

$$\int p(\mathbf{x};\boldsymbol{\theta})\mathrm{d}\mathbf{x} = 1 \quad \text{or} \quad \sum_{\mathbf{x}} p(\mathbf{x};\boldsymbol{\theta}) = 1$$

  for all $\boldsymbol{\theta}$, we say that the statistical model is normalised.
- If not, the statistical model is unnormalised.
- Undirected graphical models generally correspond to unnormalised models.
- But: partition function $Z(\boldsymbol{\theta})$ may be hard to evaluate, which is an issue for likelihood-based learning.

# Reading off the partition function from a normalised model

- Consider $\tilde{p}(\mathbf{x};\boldsymbol{\theta}) = \exp\left(-\frac{1}{2}\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}\right)$ where $\mathbf{x} \in \mathbb{R}^m$ and $\boldsymbol{\Sigma}$ is symmetric.
- Parameters $\boldsymbol{\theta}$ are the lower (or upper) triangular part of $\boldsymbol{\Sigma}$ including the diagonal.
- Corresponds to an unnormalised Gaussian.
- Partition function can be computed in closed form

$$Z(\boldsymbol{\theta}) = |\det 2\pi\boldsymbol{\Sigma}|^{1/2} \qquad p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{|\det 2\pi\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}\right)$$

- This also means that given a normalised model $p(\mathbf{x};\boldsymbol{\theta})$, you can read off the partition function as the inverse of the part that does not depend on $\mathbf{x}$, i.e. you can split a normalised $p(\mathbf{x};\boldsymbol{\theta})$ into an unnormalised model and the partition function:

$$p(\mathbf{x};\boldsymbol{\theta}) \longrightarrow p(\mathbf{x};\boldsymbol{\theta}) = \frac{\tilde{p}(\mathbf{x};\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}$$

# The domain matters

- Consider $\tilde{p}(\mathbf{x};\boldsymbol{\theta}) = \exp\left(-\frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x}\right)$ where $\mathbf{x} \in \{0,1\}^m$ and $\mathbf{A}$ is symmetric.
- Parameters $\boldsymbol{\theta}$ are the lower (or upper) triangular part of $\mathbf{A}$ including the diagonal.
- Model is known as Ising model or Boltzmann machine.
- Difference to previous slide:
  - Notation/parameterisation: $\mathbf{A}$ vs $\boldsymbol{\Sigma}^{-1}$ (does not matter)
  - $\mathbf{x} \in \{0,1\}^m$ vs $\mathbf{x} \in \mathbb{R}^m$ (does matter!)
- Partition function defined via sum rather than integral

$$Z(\boldsymbol{\theta}) = \sum_{\mathbf{x}\in\{0,1\}^m} \exp\left(-\frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x}\right)$$

- There is no analytical closed-form expression for $Z(\boldsymbol{\theta})$. Expensive to compute if $m$ is large.

# Learning

We consider two approaches to learning:

1. Learning with statistical models = parameter estimation
   (or: estimation of the model)
2. Learning with Bayesian models = Bayesian inference

# Learning with statistical models = parameter estimation

- We use use data to pick one element $p(\mathbf{x};\hat{\boldsymbol{\theta}})$ from the set of probabilistic models $\{p(\mathbf{x};\boldsymbol{\theta})\}_{\boldsymbol{\theta}}$.

$$\{p(\mathbf{x};\boldsymbol{\theta})\}_{\boldsymbol{\theta}} \qquad \xrightarrow{\text{data } \mathcal{D}} \qquad p(\mathbf{x};\hat{\boldsymbol{\theta}})$$

- In other words, we use data to select the estimate $\hat{\boldsymbol{\theta}}$ from the possible values of the parameters $\boldsymbol{\theta}$.
- Using data to pick a value of $\boldsymbol{\theta}$ corresponds to a mapping (function) from data to parameters. The mapping is called an estimator.
- Overloading of notation for the estimate and estimator:
  - $\hat{\boldsymbol{\theta}}$ as selected parameter value is the estimate of $\boldsymbol{\theta}$.
  - $\hat{\boldsymbol{\theta}}$ as mapping $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is the estimator of $\boldsymbol{\theta}$.

  This overloading of notation is often done. For example, when writing $y = x^2 + 1$, $y$ can be considered to be the output of the function ($\equiv$ estimate) or the function $y(x)$ itself ($\equiv$ estimator).

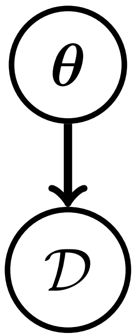
# Learning with Bayesian models = Bayesian inference

- We use data to determine the plausibility (posterior pdf/pmf) of all possible values of the parameters $\boldsymbol{\theta}$.

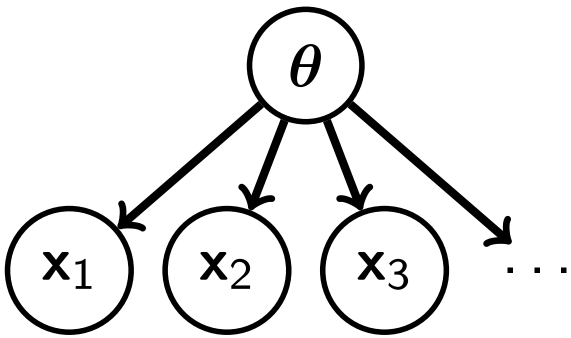
$$p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \quad \xrightarrow{\text{data } \mathcal{D}} \quad p(\boldsymbol{\theta}|\mathcal{D})$$

- Instead of picking one value from the set of possible values of $\boldsymbol{\theta}$, we here assess all of them.
- Reduces learning to inference.
- "Inverts" the data generating process

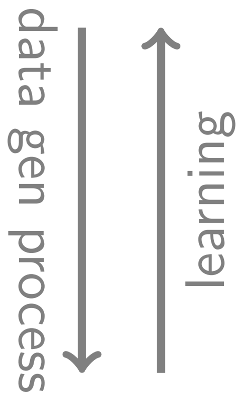
DAGs:




general case




iid data



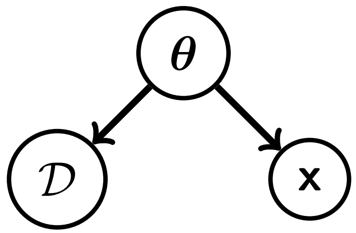
# Predictive distribution

- Given data $\mathcal{D}$, we would like to predict the next value $\mathbf{x}$.
- If we take the parameter estimation approach, the predictive distribution is $p(\mathbf{x}; \hat{\boldsymbol{\theta}})$.
- In the Bayesian inference approach, we compute

$$
\begin{aligned}
p(\mathbf{x}|\mathcal{D}) &= \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})\mathrm{d}\boldsymbol{\theta} \\
&= \int p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{D})p(\boldsymbol{\theta}|\mathcal{D})\mathrm{d}\boldsymbol{\theta} \\
&= \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\mathrm{d}\boldsymbol{\theta}
\end{aligned}
$$

(if $\mathbf{x} \perp\!\!\!\perp \mathcal{D} \mid \boldsymbol{\theta}$ as e.g. in the iid case)

Visualisation as a DAG:




Average of predictions $p(\mathbf{x}|\boldsymbol{\theta})$, weighted by $p(\boldsymbol{\theta}|\mathcal{D})$.

# Some methods for parameter estimation

- There is a multitude of methods to estimate the parameters.
- Many correspond to solving an optimisation problem, e.g. $\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}}\, J(\boldsymbol{\theta}, \mathcal{D})$ for some objective function $J$. Called M-estimation in the statistics literature.
- Maximum likelihood estimation (MLE) is popular (see next).
- Moment matching: identify the parameter configuration where the moments under the model are equal to the moments computed from the data (empirical moments).
- Maximum-a-posteriori estimation means estimating $\boldsymbol{\theta}$ by computing the maximiser of the posterior $\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}}\, p(\boldsymbol{\theta}|\mathcal{D})$.
- Score matching or noise-contrastive estimation are example methods suitable for unnormalised models (Gibbs distributions).

# Program

1. Basic concepts
   - Observed data as a sample drawn from an unknown data generating distribution
   - Probabilistic, statistical, and Bayesian models
   - Partition function and unnormalised statistical models
   - Learning = parameter estimation or learning = Bayesian inference

2. Learning by maximum likelihood estimation

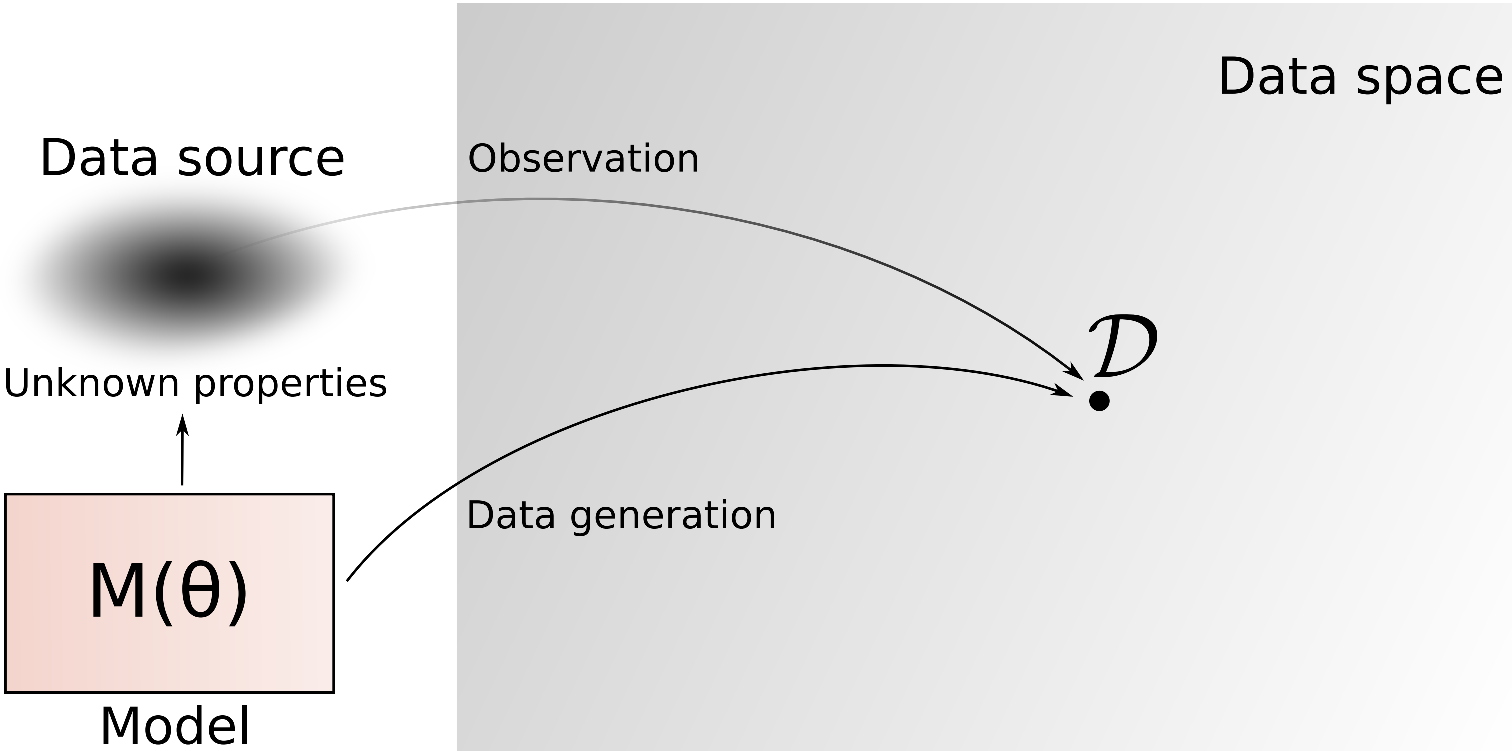
3. Learning by Bayesian inference

# Program

1. Basic concepts
2. Learning by maximum likelihood estimation
   - The likelihood function and the maximum likelihood estimate
   - MLE for Gaussian, Bernoulli, and fully observed directed graphical models of discrete random variables
   - Maximum likelihood estimation is a form of moment matching
   - The likelihood function is informative and more than just an objective function to optimise
3. Learning by Bayesian inference

# The likelihood function $L(\boldsymbol{\theta})$

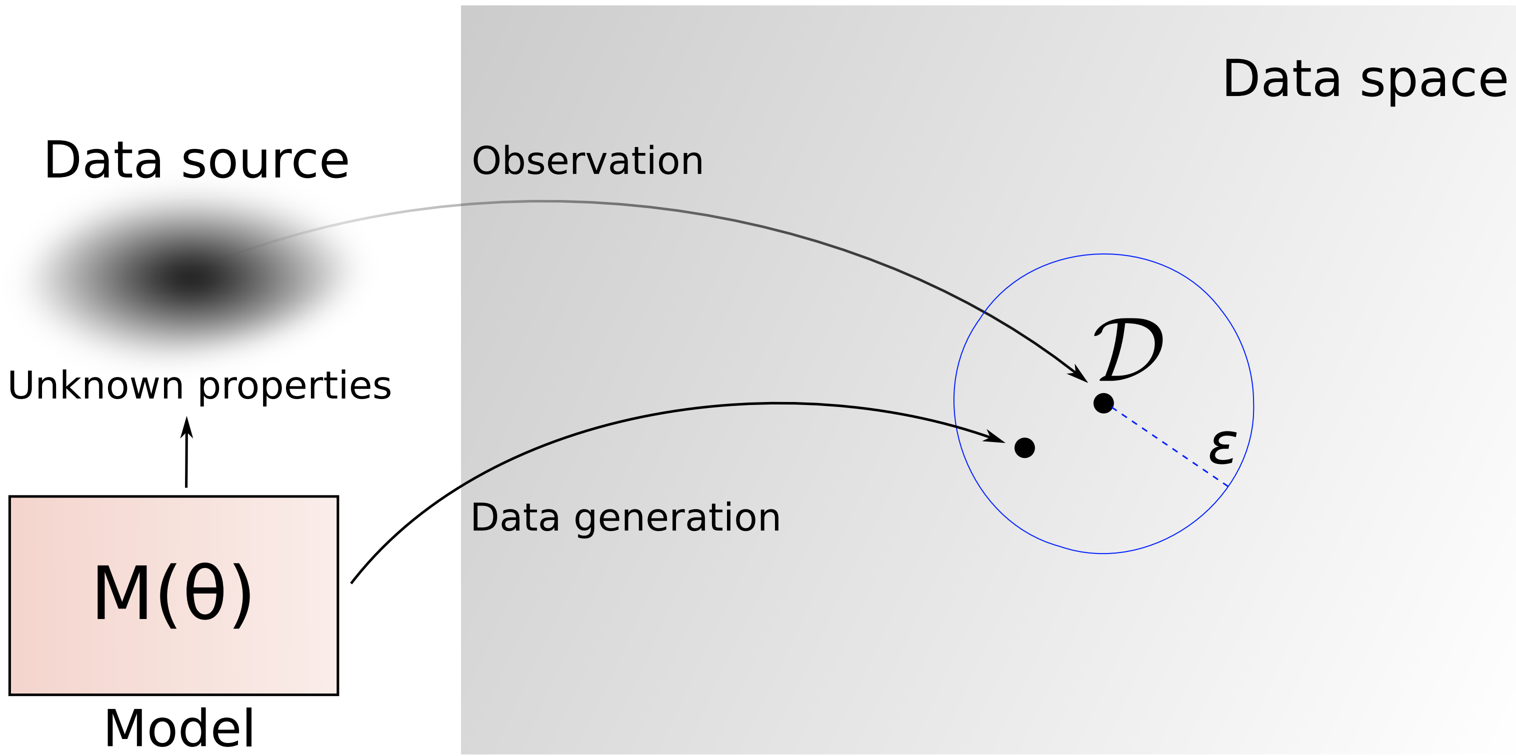
- Measures agreement between $\boldsymbol{\theta}$ and the observed data $\mathcal{D}$
- Probability that sampling from the model with parameter value $\boldsymbol{\theta}$ generates data like $\mathcal{D}$.
- Exact match for discrete random variables



# The likelihood function $L(\boldsymbol{\theta})$

- Measures agreement between $\boldsymbol{\theta}$ and the observed data $\mathcal{D}$
- Probability that sampling from the model with parameter value $\boldsymbol{\theta}$ generates data like $\mathcal{D}$.
- Small neighbourhood for continuous random variables



# The likelihood function $L(\boldsymbol{\theta})$

- Probability that the model generates data like $\mathcal{D}$ for parameter value $\boldsymbol{\theta}$,

$$L(\boldsymbol{\theta}) = p(\mathcal{D}; \boldsymbol{\theta})$$

  where $p(\mathcal{D}; \boldsymbol{\theta})$ is the parameterised model pdf/pmf.
- The likelihood function indicates the likelihood of the parameter values, and not of the data.
- For iid data $\mathbf{x}_1, \ldots, \mathbf{x}_n$

$$L(\boldsymbol{\theta}) = p(\mathcal{D}; \boldsymbol{\theta}) = p(\mathbf{x}_1, \ldots, \mathbf{x}_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i; \boldsymbol{\theta})$$

- Log-likelihood function $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$. For iid data:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\mathbf{x}_i; \boldsymbol{\theta})$$

# Maximum likelihood estimate

- The maximum likelihood estimate (MLE) is

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$$

- Numerical methods are usually needed for the optimisation.
- We typically only find local optima (sub-optimal but often useful)
- In simple cases, closed form solution possible.

# Gaussian example

- Model

$$p(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad \boldsymbol{\theta} = (\mu, \sigma^2) \qquad x \in \mathbb{R}$$

- Data $\mathcal{D}$: $n$ iid observations $x_1, \ldots, x_n$
- Log-likelihood function

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(x_i; \boldsymbol{\theta})$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{n}{2} \log(2\pi\sigma^2)$$

- Maximum likelihood estimates (see exercises)

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

# Bernoulli example

- Model
$$p(x;\theta) = \theta^x(1-\theta)^{1-x} = \theta^{\mathbb{1}(x=1)}(1-\theta)^{\mathbb{1}(x=0)}$$
with $\theta \in [0,\ 1]$, $x \in \{0, 1\}$
- Equivalent to $p(x=1;\theta) = \theta$, or the table

| $p(x;\theta)$ | $x$ |
|---|---|
| $1-\theta$ | 0 |
| $\theta$ | 1 |

- Data $\mathcal{D}$: $n$ iid observations $x_1, \ldots, x_n$
- Log-likelihood function
$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_i;\theta)$$
$$= \sum_{i=1}^{n} x_i \log(\theta) + (1-x_i)\log(1-\theta)$$

# Bernoulli example

Log-likelihood function:

$$
\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{n} x_i \log(\theta) + (1 - x_i) \log(1 - \theta) \\
&= n_{x=1} \log(\theta) + n_{x=0} \log(1 - \theta)
\end{aligned}
$$

where $n_{x=1}$ is the number of times $x_i = 1$, i.e.

$$
n_{x=1} = \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \mathbb{1}(x_i = 1)
$$

and $n_{x=0} = n - n_{x=1}$ is the number of times $x_i = 0$, i.e.

$$
n_{x=0} = \sum_{i=1}^{n} (1 - x_i) = \sum_{i=1}^{n} \mathbb{1}(x_i = 0)
$$

# Bernoulli example

- Optimisation problem:

$$\hat{\theta} = \underset{\theta \in [0,1]}{\operatorname{argmax}}\, n_{x=1} \log(\theta) + n_{x=0} \log(1-\theta)$$

  constraint optimisation problem

- Reformulation as unconstrained optimisation problem: Write

$$\eta = g(\theta) = \log\left[\frac{\theta}{1-\theta}\right] \qquad \theta = g^{-1}(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}$$

  Note: $\eta \in \mathbb{R}$

- With $\log(\theta) = \eta - \log(1+\exp(\eta))$, $\log(1-\theta) = -\log(1+\exp(\eta))$ and $n_{x=1} + n_{x=0} = n$, we have

$$\hat{\eta} = \underset{\eta}{\operatorname{argmax}}\, n_{x=1}\eta - n \log(1+\exp(\eta))$$

- Because $g(\theta)$ is an invertible function, $\hat{\theta} = g^{-1}(\hat{\eta})$ .

# Bernoulli example

- Taking the derivative with respect to $\eta$ gives necessary condition:

$$n_{x=1} - n\frac{\exp(\eta)}{1+\exp(\eta)} = 0 \qquad \frac{n_{x=1}}{n} = \frac{\exp(\eta)}{1+\exp(\eta)}$$

  Second derivative is negative for all $\eta$ so that the maximiser $\hat{\eta}$ satisfies

$$\frac{n_{x=1}}{n} = \frac{\exp(\hat{\eta})}{1+\exp(\hat{\eta})}$$

  Hence:

$$\hat{\theta} = g^{-1}(\hat{\eta}) = \frac{\exp(\hat{\eta})}{1+\exp(\hat{\eta})} = \frac{n_{x=1}}{n}$$

- Corresponds to counting: $n_{x=1}/n$ is the fraction of ones in the observed data $x_1, \ldots x_n$.
- Note: same result could here have been obtained by differentiating $\ell(\theta)$ with respect to $\theta$.

# Invariance of the MLE to re-parameterisation

- We re-parameterised the likelihood function using $\eta = \log(\theta/(1-\theta))$.
- This generalises: for $\boldsymbol{\eta} = g(\boldsymbol{\theta})$, where $g$ is invertible, we can optimise $J(\boldsymbol{\eta})$

$$J(\boldsymbol{\eta}) = \ell\left(g^{-1}(\boldsymbol{\eta})\right)$$

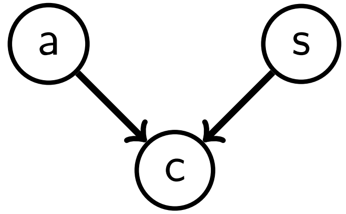
  instead of $\ell(\boldsymbol{\theta})$.
- Reason: when $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ are in a one-to-one relationship:

$$\max_{\boldsymbol{\eta}} J(\boldsymbol{\eta}) = \max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$$

$$\operatorname*{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = g^{-1}\left(\operatorname*{argmax}_{\boldsymbol{\eta}} J(\boldsymbol{\eta})\right)$$

- Sometimes simplifies the optimisation.

# Cancer-asbestos-smoking example




- Statistical model

$$p(c, a, s; \boldsymbol{\theta}) = p(c|a, s; \theta_c^1, \ldots, \theta_c^4) p(a; \theta_a) p(s; \theta_s)$$

with $p(a = 1; \theta_a) = \theta_a \quad p(s = 1; \theta_s) = \theta_s$ and

| $p(c = 1 \vert a, s; \theta_c^1, \ldots, \theta_c^4)$ | $a$ | $s$ |
|---|---|---|
| $\theta_c^1$ | 0 | 0 |
| $\theta_c^2$ | 1 | 0 |
| $\theta_c^3$ | 0 | 1 |
| $\theta_c^4$ | 1 | 1 |

- Data $\mathcal{D}$:: $n$ iid observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i = (a_i, s_i, c_i)$
- MLE of the parameters is again given by the fraction of occurrences. (see exercises)

# Maximum likelihood as moment matching

- Likelihood of $\boldsymbol{\theta}$: Probability that sampling from the model with parameter value $\boldsymbol{\theta}$ generates data like observed data $\mathcal{D}$.
- MLE: parameter configuration for which the probability to generate similar data is highest.
- Alternative interpretation: parameter configuration for which some specific moments under the model are equal to the empirical moments (moments computer from the data).
- With

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{\tilde{p}(\mathbf{x};\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}$$

we show on the next slides that the MLE $\hat{\boldsymbol{\theta}}$ satisfies:

$$\underbrace{\int \mathbf{m}(\mathbf{x};\hat{\boldsymbol{\theta}})p(\mathbf{x};\hat{\boldsymbol{\theta}})\mathrm{d}\mathbf{x}}_{\text{expected moment wrt } p(\mathbf{x};\hat{\boldsymbol{\theta}})} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbf{m}(\mathbf{x}_i;\hat{\boldsymbol{\theta}})}_{\text{empirical moment}}$$

with "moments" $\mathbf{m}(\mathbf{x};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{x};\boldsymbol{\theta})$

# Maximum likelihood as moment matching

- Gaussian example: $\log \tilde{p}(x; \mu, \sigma^2) = -\frac{(x-\mu)^2}{2\sigma^2}$
- Derivatives

$$\frac{\partial}{\partial \mu} \log \tilde{p}(x; \mu, \sigma^2) = \frac{x-\mu}{\sigma^2} \quad \frac{\partial}{\partial \sigma} \log \tilde{p}(x; \mu, \sigma^2) = \frac{(x-\mu)^2}{\sigma^3}$$

- Moment matching equations:

$$\mathbb{E}_{p(x;\hat{\mu},\hat{\sigma})}\left[\frac{x-\hat{\mu}}{\hat{\sigma}^2}\right] = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i - \hat{\mu}}{\hat{\sigma}^2}$$

$$\mathbb{E}_{p(x;\hat{\mu},\hat{\sigma})}\left[\frac{(x-\hat{\mu})^2}{\hat{\sigma}^3}\right] = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i - \hat{\mu})^2}{\hat{\sigma}^3}$$

- Two equations for two unknowns ($\hat{\mu}$ and $\hat{\sigma}^2$).

# Maximum likelihood as moment matching

Left-hand side of the first equation:

$$\mathbb{E}_{p(x;\hat{\mu},\hat{\sigma})}\left[\frac{x-\hat{\mu}}{\hat{\sigma}^2}\right] = \frac{1}{\hat{\sigma}^2}\left(\mathbb{E}_{p(x;\hat{\mu},\hat{\sigma})}\left[x-\hat{\mu}\right]\right) \tag{1}$$

$$= \frac{1}{\hat{\sigma}^2}\left(\mathbb{E}_{p(x;\hat{\mu},\hat{\sigma})}[x]-\hat{\mu}\right) \tag{2}$$

$$= \frac{1}{\hat{\sigma}^2}\left(\hat{\mu}-\hat{\mu}\right) \tag{3}$$

$$= 0 \tag{4}$$

First equation becomes:

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i-\hat{\mu}}{\hat{\sigma}^2} \tag{5}$$

Solving for $\hat{\mu}$ gives

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i \tag{6}$$

which is the maximum likelihood estimate for $\mu$.

# Maximum likelihood as moment matching

Left-hand side of the second equation:

$$\mathbb{E}_{p(x;\hat{\mu},\hat{\sigma})}\left[\frac{(x-\hat{\mu})^2}{\hat{\sigma}^3}\right] = \frac{1}{\hat{\sigma}^3}\mathbb{E}_{p(x;\hat{\mu},\hat{\sigma})}\left[(x-\hat{\mu})^2\right] \tag{7}$$

$$= \frac{\hat{\sigma}^2}{\hat{\sigma}^3} \tag{8}$$

$$= \frac{1}{\hat{\sigma}} \tag{9}$$

Second equation becomes:

$$\frac{1}{\hat{\sigma}} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\hat{\mu})^2}{\hat{\sigma}^3} \tag{10}$$

Solving for $\hat{\sigma}^2$ gives

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 \tag{11}$$

which is the maximum likelihood estimate for $\sigma^2$.

# Maximum likelihood as moment matching (proof, not examinable)

A necessary condition for the MLE $\hat{\boldsymbol{\theta}}$ to satisfy is

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})\big|_{\hat{\boldsymbol{\theta}}} = 0$$

We can write the gradient of the log-likelihood function as follows

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p(\mathbf{x}_i; \boldsymbol{\theta}) \\
&= \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log \frac{\tilde{p}(\mathbf{x}_i; \boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \\
&= \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log \tilde{p}(\mathbf{x}_i; \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} n \log Z(\boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{x}_i; \boldsymbol{\theta}) - n \nabla_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \mathbf{m}(\mathbf{x}_i; \boldsymbol{\theta}) - n \nabla_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta})
\end{aligned}$$

# Maximum likelihood as moment matching (proof, not examinable)

The gradient $\nabla_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta})$ is

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta}) &= \frac{1}{Z(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} Z(\boldsymbol{\theta}) \\
&= \frac{1}{Z(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} \int \tilde{p}(\mathbf{x};\boldsymbol{\theta}) \mathrm{d}\mathbf{x} \\
&= \frac{\int \nabla_{\boldsymbol{\theta}} \tilde{p}(\mathbf{x};\boldsymbol{\theta}) \mathrm{d}\mathbf{x}}{Z(\boldsymbol{\theta})}
\end{aligned}$$

Since $(\log f(x))' = \frac{f'(x)}{f(x)}$ we also have $f'(x) = (\log f(x))' f(x)$ so that

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta}) &= \frac{\int \nabla_{\boldsymbol{\theta}} \left[\log \tilde{p}(\mathbf{x};\boldsymbol{\theta})\right] \tilde{p}(\mathbf{x};\boldsymbol{\theta}) \mathrm{d}\mathbf{x}}{Z(\boldsymbol{\theta})} \\
&= \int \nabla_{\boldsymbol{\theta}} \left[\log \tilde{p}(\mathbf{x};\boldsymbol{\theta})\right] p(\mathbf{x};\boldsymbol{\theta}) \mathrm{d}\mathbf{x} \\
&= \int \mathbf{m}(\mathbf{x};\boldsymbol{\theta}) p(\mathbf{x};\boldsymbol{\theta}) \mathrm{d}\mathbf{x}
\end{aligned}$$

# Maximum likelihood as moment matching (proof, not examinable)

The gradient of the log-likelihood function $\ell(\boldsymbol{\theta})$ thus is

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{m}(\mathbf{x}_i; \boldsymbol{\theta}) - n \int \mathbf{m}(\mathbf{x}; \boldsymbol{\theta}) p(\mathbf{x}; \boldsymbol{\theta}) \mathrm{d}\mathbf{x}$$

The necessary condition that the gradient is zero at the MLE $\hat{\boldsymbol{\theta}}$ yields the desired result:

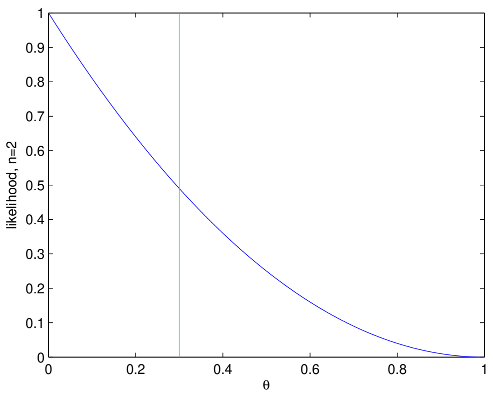
$$\int \mathbf{m}(\mathbf{x}; \hat{\boldsymbol{\theta}}) p(\mathbf{x}; \hat{\boldsymbol{\theta}}) \mathrm{d}\mathbf{x} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{m}(\mathbf{x}_i; \hat{\boldsymbol{\theta}})$$

Since the integral is the expectation of $\mathbf{m}(\mathbf{x}; \hat{\boldsymbol{\theta}})$ with respect to $p(\mathbf{x}; \hat{\boldsymbol{\theta}})$ we can write the above equation as

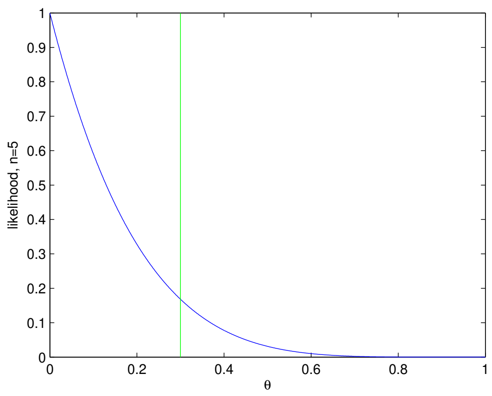
$$\mathbb{E}_{p(\mathbf{x}; \hat{\boldsymbol{\theta}})} \left[ \mathbf{m}(\mathbf{x}; \hat{\boldsymbol{\theta}}) \right] = \frac{1}{n} \sum_{i=1}^{n} \mathbf{m}(\mathbf{x}_i; \hat{\boldsymbol{\theta}})$$

# What we miss with maximum likelihood estimation

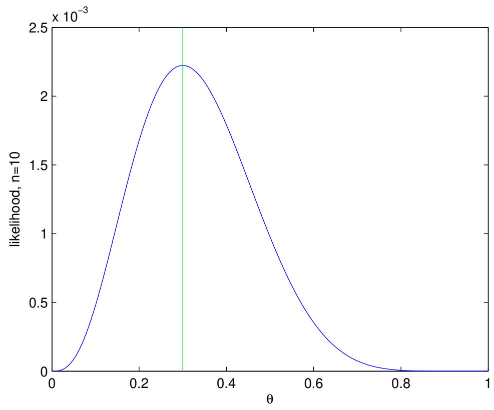
- The likelihood function indicates to which extent various parameter values are congruent with the observed data.
- Establishes an ordering of relative preferences for different parameter values, i.e. $\boldsymbol{\theta}_1$ is preferred over $\boldsymbol{\theta}_2$ if $L(\boldsymbol{\theta}_1) > L(\boldsymbol{\theta}_2)$.
- Max. lik. estimation ignores information contained in the data.
- Example: Likelihood for Bernoulli model with $\mathcal{D} = (0, 0, 0, 0, 0, 0, 0, 1, 1, 1, \ldots)$ generated with parameter value $1/3$ (green line)




(a) $n = 2$ observations




(b) $n = 5$ observations




(c) $n = 10$ observations

# What we miss with maximum likelihood estimation

- A compromise between considering the whole (log) likelihood function and only its maximum is the computation of the curvature (Hessian) at the maximum.
- strong curvature: max lik estimate clearly to be preferred
- shallow curvature: several other parameter values are nearly equally in line with the data.

# Program

1. Basic concepts

2. Learning by maximum likelihood estimation
   - The likelihood function and the maximum likelihood estimate
   - MLE for Gaussian, Bernoulli, and fully observed directed graphical models of discrete random variables
   - Maximum likelihood estimation is a form of moment matching
   - The likelihood function is informative and more than just an objective function to optimise

3. Learning by Bayesian inference

# Program

1. Basic concepts

2. Learning by maximum likelihood estimation

3. Learning by Bayesian inference
    - Bayesian approach reduces learning to probabilistic inference
    - Different views of the posterior distribution
    - Conjugate priors
    - Posterior for Gaussian, Bernoulli, and fully observed directed graphical models of discrete random variables

# Reduces learning to probabilistic inference

- We use data to determine the plausibility (posterior pdf/pmf) of all possible values of the parameters $\boldsymbol{\theta}$.

$$p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \qquad \xrightarrow{\text{data } \mathcal{D}} \qquad p(\boldsymbol{\theta}|\mathcal{D})$$

- Same framework for learning and inference.
- In some cases, closed-form solutions can be obtained (e.g. for conjugate priors).
- In some cases, exact inference methods that we discussed earlier can be used.
- If closed form solutions are not possible and exact inference is computationally too costly, we have to resort to approximate inference via e.g. sampling or variational methods.

# The posterior combines likelihood function and prior

- Bayesian inference takes the whole likelihood function into account

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &= \frac{p(\boldsymbol{\theta}, \mathcal{D})}{p(\mathcal{D})} \\ &= \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \\ &\propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &\propto L(\boldsymbol{\theta})p(\boldsymbol{\theta}) \end{aligned}$$

- $L(\boldsymbol{\theta})$ defines a change of measure from $p(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta}|\mathcal{D})$.
- For iid data $\mathcal{D} = (\mathbf{x}_1, \ldots \mathbf{x}_n)$

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto \left[\prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta})\right] p(\boldsymbol{\theta})$$

- For large $n$, likelihood dominates: $\operatorname{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{D}) \approx \text{MLE}$
(assuming the prior is non-zero at the MLE)

# The posterior distribution is a conditional

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\boldsymbol{\theta},\mathcal{D})}{p(\mathcal{D})}$$

- For simplicity, consider discrete-valued data so that

$$p(\boldsymbol{\theta}|\mathcal{D}) = p(\boldsymbol{\theta}|\mathbf{x} = \mathcal{D}) = \frac{p(\boldsymbol{\theta}, \mathbf{x} = \mathcal{D})}{p(\mathcal{D})}$$

- Assume we can sample tuples $(\boldsymbol{\theta}^{(i)}, \mathbf{x}^{(i)})$ from the joint $p(\boldsymbol{\theta}, \mathbf{x})$

$$\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta}) \qquad \mathbf{x}^{(i)} \sim p(\mathbf{x}|\boldsymbol{\theta}^{(i)})$$

- Conditioning on $\mathbf{x} = \mathcal{D}$ then corresponds to only retaining those samples $(\boldsymbol{\theta}^{(i)}, \mathbf{x}^{(i)})$ where $\mathbf{x}^{(i)} = \mathcal{D}$.
- Samples from the posterior = samples from the prior that produce data equal to the observed one.
- Remark: This view of Bayesian inference forms the basis of a class of approximate methods known as approximate Bayesian computation.

# Conjugate priors

- Assume the prior is part of a parametric family with hyperparameters $\boldsymbol{\alpha}$, i.e. the prior is an element of $\{p(\boldsymbol{\theta}; \boldsymbol{\alpha})\}_{\boldsymbol{\alpha}}$, so that

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}; \boldsymbol{\alpha}_0)$$

  for some fixed $\boldsymbol{\alpha}_0$.
- If the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ is part of the same family as the prior,
  - the prior and posterior are called conjugate distributions
  - the prior is said to be a conjugate prior for $p(\mathbf{x}|\boldsymbol{\theta})$ or for the likelihood function.
- Learning then corresponds to updating the hyperparameters.

$$\boldsymbol{\alpha}_0 \qquad \xrightarrow{\text{data } \mathcal{D}} \qquad \boldsymbol{\alpha}(\mathcal{D})$$

- Models $p(\mathbf{x}|\boldsymbol{\theta})$ that a part of the exponential family always have a conjugate prior (see Barber 8.5).

## Gaussian example (posterior of the mean for known variance)

(for more general cases, see optional reading)

- Denote pdf of a Gaussian random variable $x$ with mean $\mu$ and variance $\sigma^2$ by $\mathcal{N}(x; \mu, \sigma^2)$.
- Bayesian model

$$p(x|\theta) = \mathcal{N}(x|\theta, \sigma^2) \qquad p(\theta; \boldsymbol{\alpha}_0) = \mathcal{N}(\theta; \mu_0, \sigma_0^2)$$

  Hyperparameters $\boldsymbol{\alpha}_0 = (\mu_0, \sigma_0^2)$

- Data $\mathcal{D}$: $n$ iid observations $x_1, \ldots, x_n$
- Posterior for $\theta$ (see exercises)

$$p(\theta|\mathcal{D}) = \mathcal{N}(\theta; \mu_n, \sigma_n^2)$$

$$\mu_n = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n}\mu_0 \quad \frac{1}{\sigma_n^2} = \frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2}$$

  where $\bar{x} = 1/n \sum_i x_i$ is the sample average (the MLE).

# Gaussian example (posterior of the mean for known variance)

$$\mu_n = \frac{\sigma_0^2}{\sigma_0^2+\sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\sigma_0^2+\sigma^2/n}\mu_0$$

- Introduce
$$w_n = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n} \tag{12}$$
For $n = 0$, $w_n \to 0$. For $n \to \infty$, $w_n \to 1$
- Moreover:
$$\frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n} = 1 - w_n \tag{13}$$
- Hence
$$\mu_n = w_n \bar{x} + (1 - w_n)\mu_0 \tag{14}$$
As the number of data points increases, $\mu_n$ travels from prior mean $\mu_0$ to the MLE $\bar{x}$ along a straight line.
- The posterior mean of $\theta$ linearly interpolates between prior mean $\mu_0$ and MLE $\hat{x}$.

# Bernoulli example

- Recall: Beta distribution with parameters $\alpha, \beta$

$$\mathcal{B}(f; \alpha, \beta) \propto f^{\alpha-1}(1-f)^{\beta-1} \qquad f \in [0,1]$$

  see the background document *Introduction to Probabilistic Modelling*

- Bayesian model

$$p(x|\theta) = \theta^x(1-\theta)^{1-x} \qquad p(\theta; \boldsymbol{\alpha}_0) = \mathcal{B}(\theta; \alpha_0, \beta_0)$$

  where $x \in \{0,1\}$, $\theta \in [0,1]$, and $\boldsymbol{\alpha}_0 = (\alpha_0, \beta_0)$

- Data $\mathcal{D}$: $n$ iid observations $x_1, \ldots, x_n$
- Posterior for $\theta$ (see exercises)

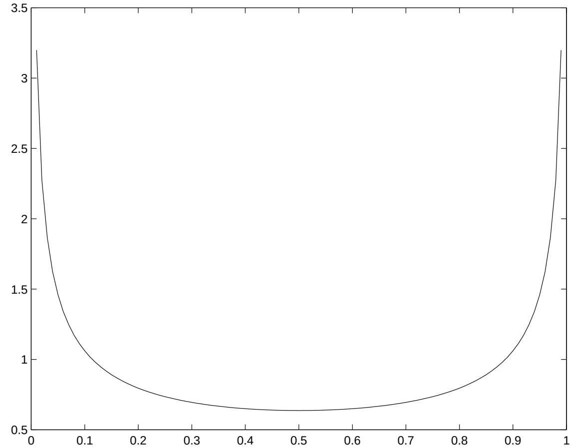
$$p(\theta|\mathcal{D}) = \mathcal{B}(\theta; \alpha_n, \beta_n)$$

$$\alpha_n = \alpha_0 + n_{x=1} \qquad \beta_n = \beta_0 + n_{x=0}$$

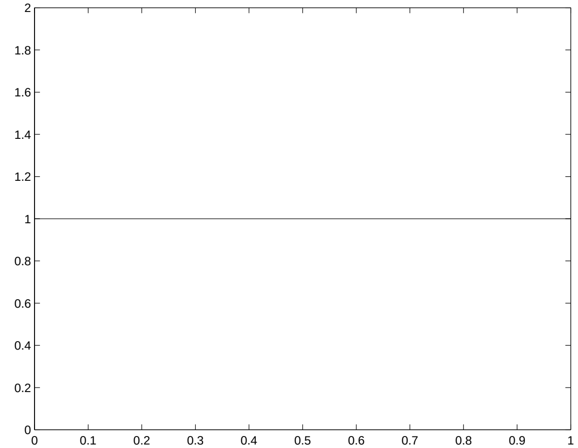
  where $n_{x=1}$ were the number of ones and $n_{x=0}$ the number of zeros in the data.

# Examples of the beta distribution $\mathcal{B}(f;\alpha,\beta)$ (Figures courtesy C. Williams)

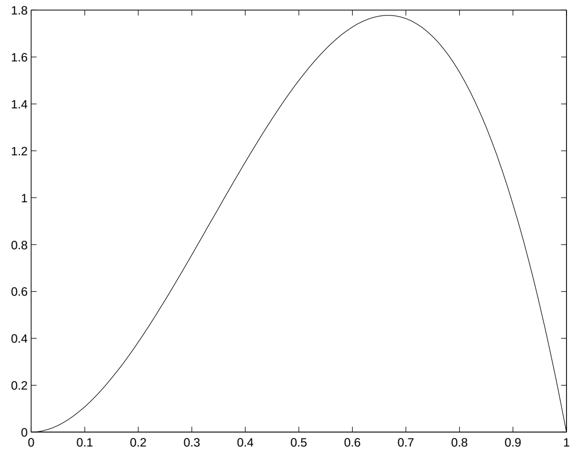
Expected value: $\frac{\alpha}{\alpha+\beta}$, Variance: $\frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\frac{1}{\alpha+\beta+1}$



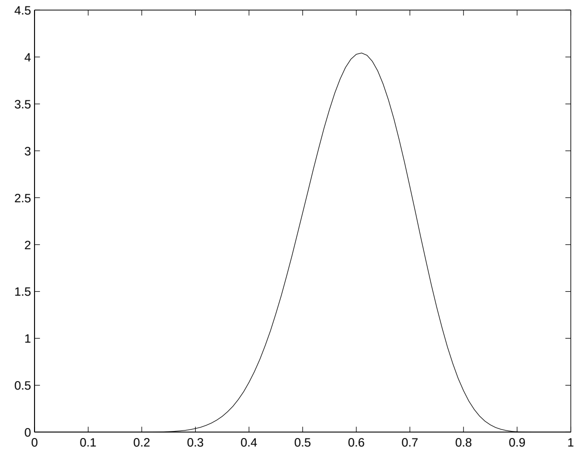

(a) $\mathcal{B}(f;0.5,0.5)$




(b) $\mathcal{B}(f;1,1)$



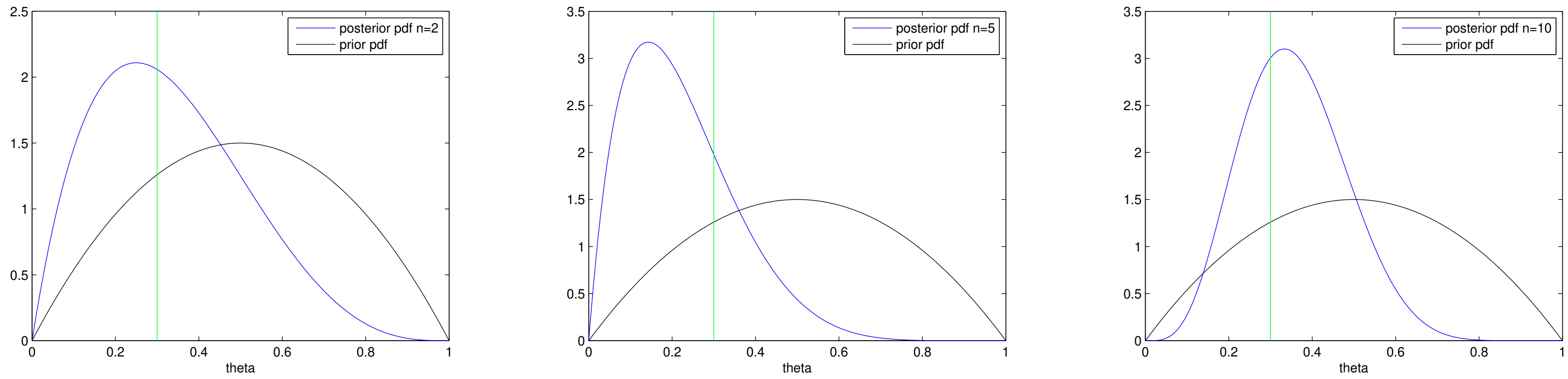

(c) $\mathcal{B}(f;3,2)$




(d) $\mathcal{B}(f;15,10)$

# Bernoulli example

- Bernoulli model with $\mathcal{D} = (0, 0, 0, 0, 0, 0, 0, 1, 1, 1, \ldots)$ generated with parameter value $1/3$ (green line)
- Posterior in blue, $\mathcal{B}(2, 2)$ prior in black
- Compare with earlier likelihood plots. Note the "pull" towards the prior when $n$ is small.




(a) $n = 2$ observations (b) $n = 5$ observations (c) $n = 10$ observations

# Cancer-asbestos-smoking example

- Bayesian model

$$
\begin{aligned}
p(c, a, s|\boldsymbol{\theta}) &= p(c|a, s, \theta_c^1, \ldots, \theta_c^4)p(a|\theta_a)p(s|\theta_s) \\
&= \prod_{j=1}^{4} (\theta_c^j)^{\mathbb{1}(c=1, \mathrm{pa}_c=j)}(1-\theta_c^j)^{\mathbb{1}(c=0, \mathrm{pa}_c=j)} \\
&\quad \theta_a^{\mathbb{1}(a=1)}(1-\theta_a)^{\mathbb{1}(a=0)}\theta_s^{\mathbb{1}(s=1)}(1-\theta_s)^{\mathbb{1}(s=0)}
\end{aligned}
$$

- Assume the prior factorises (independence assumptions):

$$
\begin{aligned}
p(\theta_a, \theta_s, \theta_c^1, \ldots, \theta_c^4; \boldsymbol{\alpha}_0) &= \prod_j \mathcal{B}(\theta_c^j; \alpha_{c,0}^j, \beta_{c,0}^j) \\
&\quad \mathcal{B}(\theta_a; \alpha_{a,0}, \beta_{a,0})\mathcal{B}(\theta_s; \alpha_{s,0}, \beta_{s,0})
\end{aligned}
$$

- Data $\mathcal{D}$: $n$ iid observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i = (a_i, s_i, c_i)$
- The parameters are independent under the posterior and follow a beta distribution (see exercises)

# Program recap

1. Basic concepts
   - Observed data as a sample drawn from an unknown data generating distribution
   - Probabilistic, statistical, and Bayesian models
   - Partition function and unnormalised statistical models
   - Learning = parameter estimation or learning = Bayesian inference
2. Learning by maximum likelihood estimation
   - The likelihood function and the maximum likelihood estimate
   - MLE for Gaussian, Bernoulli, and fully observed directed graphical models of discrete random variables
   - Maximum likelihood estimation is a form of moment matching
   - The likelihood function is informative and more than just an objective function to optimise
3. Learning by Bayesian inference
   - Bayesian approach reduces learning to probabilistic inference
   - Different views of the posterior distribution
   - Conjugate priors
   - Posterior for Gaussian, Bernoulli, and fully observed directed graphical models of discrete random variables