Probabilistic Graphical Models

Parameter Estimation

Tomer Galanti

December 14, 2015

Overview

What did we have so far?

1. **Representations**: how do we **model the problem**? (directed/undirected).
2. **Inference**: given a model and **partially observed data**, how can we **recognize the rest of the data**? (VE/MCMC/Gibbs sampling/etc').

Suruhanjaya Komunikasi dan Multimedia Malaysia
Malaysian Communications and Multimedia Commission

Motivation

A complete discussion of graphical models includes

- **Representation**: the conditions between the variables.
- **Inference**: the ability to make observations within the model.
- **Parameters**: specifying the probabilities.

Motivation

Why estimating the parameters of a graphical model is important to us?

This problem rises very often in practice, since numerical parameters are harder to elicit from human experts than the structure is.

Motivation

1. Consider a **Bayesian network**.
2. The network's **structure** is **fixed**.
3. We assume a data set D of **fully observed** network variables $D = \{\zeta[1], ..., \zeta[M]\}$.
4. How do we **estimate the parameters** of the network?

Solutions

We will have two different approaches!

1. One is based on the **maximum likelihood estimation**.
2. The other is based on **Bayesian perspectives**.

Solutions

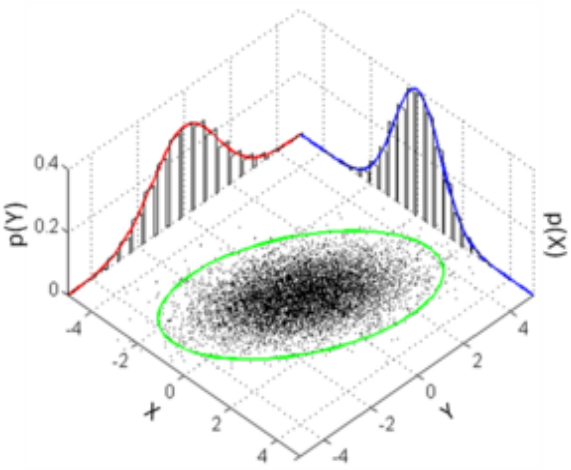
We will have two different approaches!

1. One is based on the **maximum likelihood estimation**.
2. The other is based on **Bayesian perspectives**.

Maximum likelihood: estimating a normal distribution

Before we give a formal definition for this method...
Let's start with a few simple examples...!

Assume the points $D = \{x_1, ..., x_M\}$ are i.i.d samples of a 1D Gaussian distribution $\mathcal{N}(\mu, 1)$.



How do we **select μ that best fit the data**?

Maximum likelihood: estimating a normal distribution

A reasonable approach: select μ that maximizes the probability for sampling D from $\mathcal{N}(\mu, 1)$.

Formally, solve the following program:

$$\mu^* = \arg\max_{\mu} P[D; \mu]$$

Maximum likelihood: estimating a normal distribution

By i.i.dness: $P[D;\mu] = \prod_{i=1}^{M} P[x_i;\mu]$.

In addition, $x_i \sim \mathcal{N}(\mu, 1)$ and therefore,

$$\mu^* = \arg\max_{\mu} \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi}} \exp(-(x_i - \mu)^2/2)$$

Taking log of the inner argument yields the same solution, since log is a strictly increasing function,

$$\mu^* = \arg\max_{\mu} \log \left[\prod_{i=1}^{M} \frac{1}{\sqrt{2\pi}} \exp(-(x_i - \mu)^2/2) \right]$$

Maximum likelihood: estimating a normal distribution

Equivalently,

$$\mu^* = \arg\max_{\mu} \sum_{i=1}^{M} -(x_i - \mu)^2/2$$

Differentiate and hope for good:

$$\sum_{i=1}^{M}(x_i - \mu) = 0 \implies \mu^* = \frac{1}{M}\sum_{i=1}^{M} x_i$$

The second derivative is $-M < 0$ and therefore, this is indeed the **maximum** of this function.

Maximum likelihood: estimating a multinomial distribution

What about coin flips? This will be our **running example**.

Provided with a data set $D = \{x_1, ..., x_M\}$ of i.i.d samples $x_i \sim Bernoulli(\theta)$, we want to estimate θ.

head "=" 1 and tail "=" 0.

Maximum likelihood: estimating a multinomial distribution

Again we maximize the probability for sampling the data,

$$P[D;\theta] = \theta^{M[head]} \cdot (1-\theta)^{M[tail]}$$

Where: $M[head]$ = number of heads in D
and $M[tail]$ = number of tails in D.

Maximum likelihood: estimating a multinomial distribution

By taking log we have:

$$M[head] \log \theta + M[tail] \log(1 - \theta)$$

Which is maximized by:

$$\hat{\theta} = \frac{M[head]}{M[head] + M[tail]}$$

Reasonable right?

Maximum likelihood

This is just a special case of a much more general concept..

Definition (Maximum likelihood)

For a data set $D = \{\zeta[1], ..., \zeta[M]\}$ and a family of distributions $P[\cdot; \theta]$, the likelihood of D for a given choice of the parameters θ is

$$L(\theta : D) = \prod_m P[\zeta[m]; \theta]$$

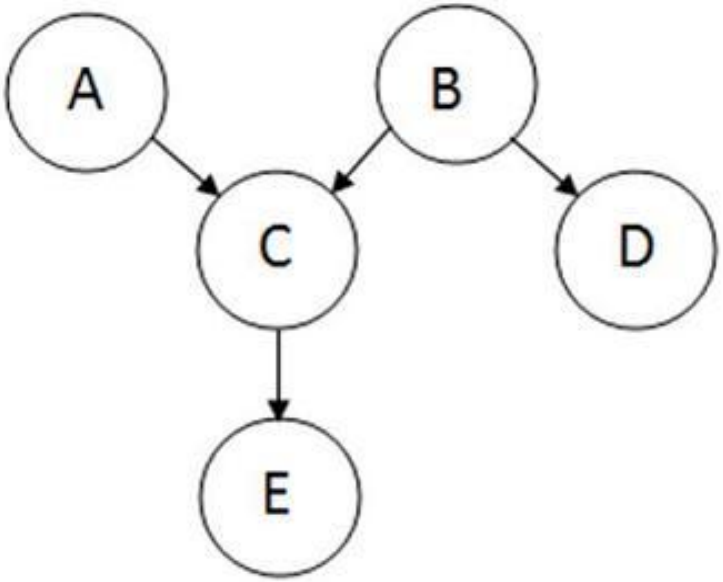
The maximum likelihood estimator (MLE) returns θ that maximizes this quantity.

In many cases, we apply the "**taking log**" trick to **simplify** the problem.

Maximum likelihood: Bayesian networks

So far, we designed a rule for selecting the parameters of a statistical model.

It is very interesting to see **how it works out** in the case of **Bayesian networks**!



Maximum likelihood: Bayesian networks

Let's start with the simplest non-trivial network..

Consider a graph of **two boolean variables** $X \to Y$.

In this case, the parameterization θ consists of 6 parameters:

- $\theta_{x^0} := P[X = 0]$ and $\theta_{x^1} := P[X = 1]$.
- $\theta_{y^0|x^0} := P[Y = 0|X = 0]$ and $\theta_{y^1|x^0} := P[Y = 1|X = 0]$.
- $\theta_{y^0|x^0} := P[Y = 0|X = 1]$ and $\theta_{y^1|x^0} := P[Y = 1|X = 1]$.

Maximum likelihood: Bayesian networks

Given a data set $D = \{(x[m], y[m])\}_{m=1}^{M}$ the likelihood becomes:

$$\begin{aligned} L(\theta : D) &= \prod_m P[(x[m], y[m]); \theta] \\ &= \prod_m P[x[m]; \theta] \cdot P[y[m]|x[m]; \theta] \\ &= \prod_m P[x[m]; \theta] \prod_m P[y[m]|x[m]; \theta] \\ &= \theta_{x^0}^{M[0]} \cdot \theta_{x^1}^{M[1]} \prod_m P[y[m]|x[m]; \theta] \end{aligned}$$

Where $M[x]$ counts the number of samples such that $x[m] = x$.

Maximum likelihood: Bayesian networks

It is left to represent the other product:

$$\prod_m P[y[m]|x[m];\theta] = \prod_m P[y[m]|x[m];\theta_{Y|X}]$$

$$= \prod_{m:x[m]=0} P[y[m]|x[m];\theta_{Y|X}] \cdot \prod_{m:x[m]=1} P[y[m]|x[m];\theta_{Y|X}]$$

$$= \prod_{m:x[m]=0} P[y[m]|x[m];\theta_{Y|X=0}] \cdot \prod_{m:x[m]=1} P[y[m]|x[m];\theta_{Y|X=1}]$$

$$= \theta_{y^0|x^0}^{M[0,0]} \cdot \theta_{y^1|x^0}^{M[0,1]} \cdot \theta_{y^0|x^1}^{M[1,0]} \cdot \theta_{y^1|x^1}^{M[1,1]}$$

Where $M[x, y]$ counts the number of samples $(x[m], y[m]) = (x, y)$.

Maximum likelihood: Bayesian networks

Finally, the likelihood decomposes very nicely.

$$L(\theta : D) = \theta_{x^o}^{M[0]} \cdot \theta_{x_1}^{M[1]} \cdot \theta_{y^0|x^0}^{M[0,0]} \cdot \theta_{y^1|x^0}^{M[0,1]} \cdot \theta_{y^0|x^1}^{M[1,0]} \cdot \theta_{y^1|x^1}^{M[1,1]}$$

We have three sets of **separable terms**: $\theta_{x^o}^{M[0]} \cdot \theta_{x_1}^{M[1]}$, $\theta_{y^0|x^0}^{M[0,0]} \cdot \theta_{y^1|x^0}^{M[0,1]}$ and $\theta_{y^0|x^1}^{M[1,0]} \cdot \theta_{y^1|x^1}^{M[1,1]}$.
Therefore, we can **maximize** each one **separately**.

Maximum likelihood: Bayesian networks

By the **same analysis used for the coin flip example**, we arrive to the following conclusions:

1. $\hat{\theta}_{x^0} = \frac{M[0]}{M[0]+M[1]}$ and $\hat{\theta}_{x^1} = 1 - \hat{\theta}_{x^0}$.
2. $\hat{\theta}_{y^0|x^0} = \frac{M[0,0]}{M[0,0]+M[0,1]}$ and $\hat{\theta}_{y^1|x^0} = 1 - \hat{\theta}_{y^0|x^0}$.
3. $\hat{\theta}_{y^0|x^1} = \frac{M[1,0]}{M[1,0]+M[1,1]}$ and $\hat{\theta}_{y^1|x^1} = 1 - \hat{\theta}_{y^0|x^1}$.

(By log-likelihood maximization).

Intuitive, right?

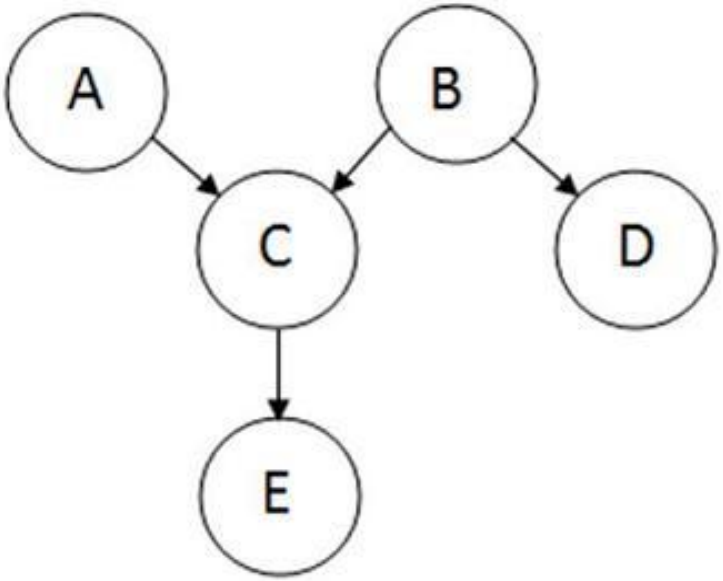
Maximum likelihood: Bayesian networks

What about the general case?

Actually, there is no much difference...

Maximum likelihood: Bayesian networks

- Consider a Bayesian network G.
- For each variable X_i we have a set of parameters specifying the probabilities for it's values given it's parents $\theta_{X_i|Pa_{X_i}}$ (i.e, it's CPD).
- The total set of parameters is: $\theta = \cup_i \theta_{X_i|Pa_{X_i}}$.



Maximum likelihood: Bayesian networks

The likelihood decomposes into local likelihoods.

$$\begin{aligned} L(\theta : D) &= \prod_m P[\zeta[m]; \theta] \\ &= \prod_m \prod_i P[X_i[m] | Pa_{X_i}[m]; \theta] \\ &= \prod_i \prod_m P[X_i[m] | Pa_{X_i}[m]; \theta] \\ &= \prod_i L_i(\theta_{X_i|Pa_{X_i}} : D) \end{aligned}$$

Each $L_i(\theta_{X_i|Pa_{X_i}} : D) = \prod_m P[X_i[m] | Pa_{X_i}[m]; \theta]$ is the **local likelihood of** X_i.

Maximum likelihood: Bayesian networks

The local likelihoods are parameterized by disjoint sets of parameters.
Therefore, **maximizing the total likelihood** is **equivalent to maximizing each local likelihood separately**.

Maximum likelihood: Bayesian networks

Here is how to maximize a local likelihood.

Let X be a variable and U it's parents. We have a parameter $\theta_{x|u}$ for each combination $x \in Val(X)$ and $u \in Val(U)$.

$$L_X(\theta_{X|U} : D) = \prod_m \theta_{x[m]|u[m]}$$

$$= \prod_{u \in Val(U)} \left[\prod_{x \in Val(X)} \theta_{x|u}^{M[u,x]} \right]$$

Where $M[u, x]$ is the number of times $x[m] = x$ and $u[m] = u$.

Maximum likelihood: Bayesian networks

For each u we maximize $\prod_{x \in Val(X)} \theta_{x|u}^{M[u,x]}$ separately and obtain:

$$\hat{\theta}_{x|u} := \frac{M[u,x]}{M[u]} = \text{“ratio of } X = x \text{ between all examples that satisfy } U = u \text{ in the data set”}$$

Where $M[u] := \sum_x M[u,x]$.

Maximum likelihood estimation as M-Projection

The MLE principle gives a **recipe how to construct estimators for diferent statistical models** (for example, multinomials and Gaussians). As we have seen, for simple examples the resulting estimators are quite intuitive.

However, the same principle can be applied in a much broader range of parametric models.

Maximum likelihood estimation as M-Projection

Recall the notion of **projection**: **finding the distribution**, within a specified class, that is **closest to a given target distribution**.

Parameter estimation is similar in the sense that we **select a distribution** from a given class that is **closest to our data**.

As we show next, the MLE aims to find the distribution that is closest to the empirical distribution $\hat{P}_D$.

Maximum likelihood estimation as M-Projection

Theorem

The MLE $\hat{\theta}$ in a parametric family of distributions relative to data set D is the M-Projection of $\hat{P}_D$ on the parametric family.

$$\hat{\theta} = \arg\min_{\theta \in \Theta} KL(\hat{P}_D || P_\theta)$$

Here, $\hat{P}_D(x) = \frac{|\{z \in D : z = x\}|}{|D|}$
And $KL(p||q) = \mathbb{E}_{x \sim p}[\log(p(x)/q(x))]$.

Maximum likelihood estimation as M-Projection

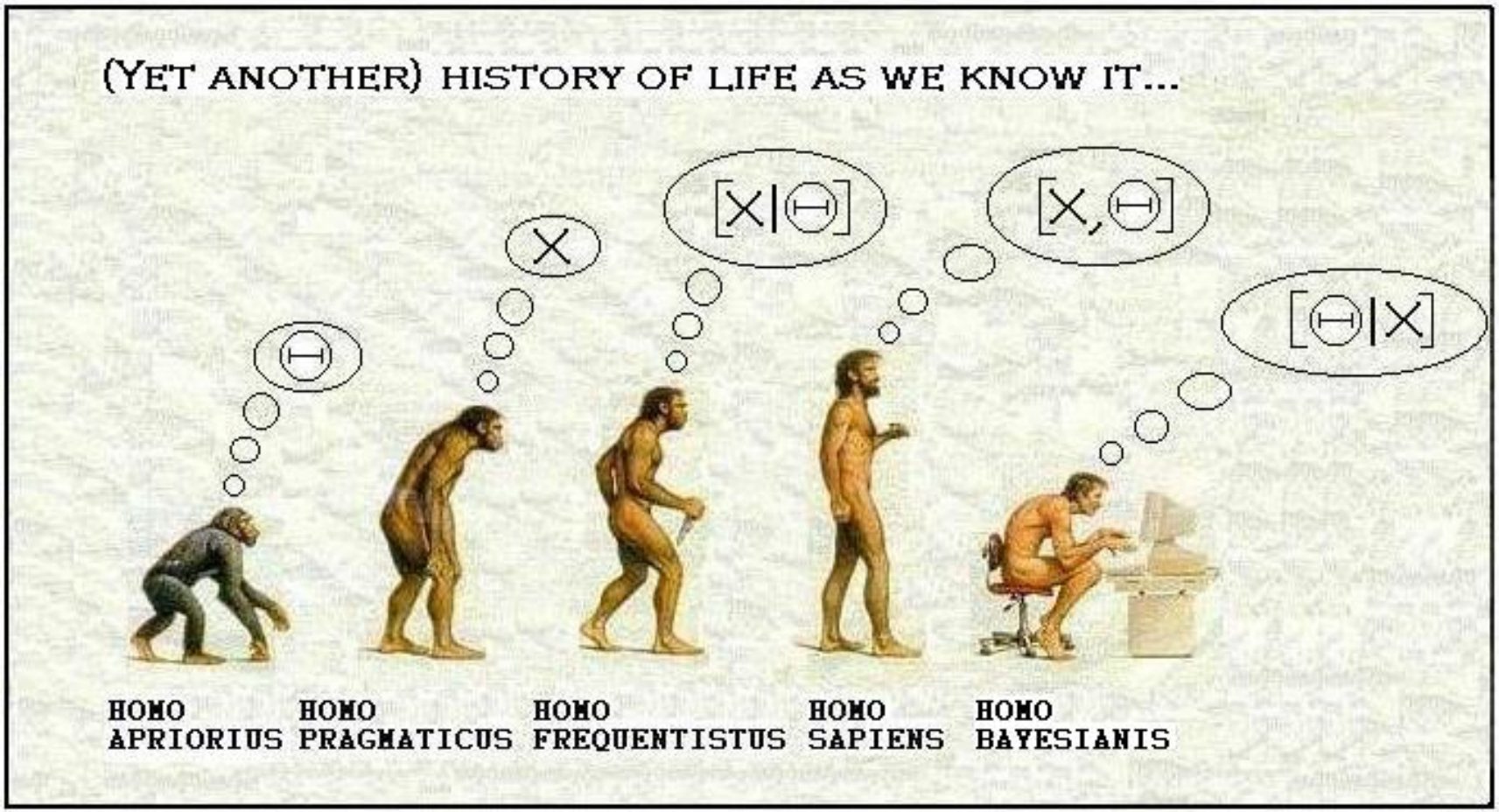
Proof.

$$\begin{aligned}
\log L(\theta : D) &= \sum_m \log P[\zeta[m]; \theta] \\
&= \sum_\zeta \log P[\zeta; \theta] \cdot \sum_m I[\zeta[m] = \zeta] \\
&= \sum_\zeta M \hat{P}_D(\zeta) \log P[\zeta; \theta] \\
&= M \cdot \mathbb{E}_{\hat{P}_D}[\log P[\zeta; \theta]] \\
&= M \cdot \Big(H(\hat{P}_D) - KL(\hat{P}_D || P_\theta) \Big)
\end{aligned}$$

Maximizing this object is equivalent to minimizing $KL(\hat{P}_D||P_\theta)$ since $H(\hat{P}_D)$ is fixed. □

Bayesian parameter estimation

Bayesian parameter estimation.



Bayesian parameter estimation

The **MLE** seems plausible, but it can be **overly simplistic** in many cases.

Assume we perform an experiment of tossing a coin and get 3 heads out of 10. A reasonable conclusion would be that the probability for head is ≈ 0.3, right?

What if we have a great experience with tossing random coins? What if a probable coin is very close to a fair coin?

This is called, **prior knowledge**. We do not want the prior knowledge to be the absolute guide, but rather a **reasonable starting point**.

Bayesian parameter estimation

Bayesian parameter estimation: a full discussion of parameter estimation includes **both analysis of the data** and a **prior knowledge** about the parameters.

Bayesian parameter estimation: joint probabilistic model

How do we model the prior knowledge..?

One approach is to **encode the prior knowledge** about θ with a **distribution**. Think of it as a hierarchy between the different choices of θ.

This is called a **prior distribution** and is denoted by $P(\theta)$.

Bayesian parameter estimation: joint probabilistic model

In the previous model, the samples were i.i.d according to a distribution $P[\cdot;\theta]$. In the current setup the samples are **conditionally independent given** θ, since θ is itself a random variable.

Bayesian parameter estimation: joint probabilistic model

Visually, the sampling process looks like this.

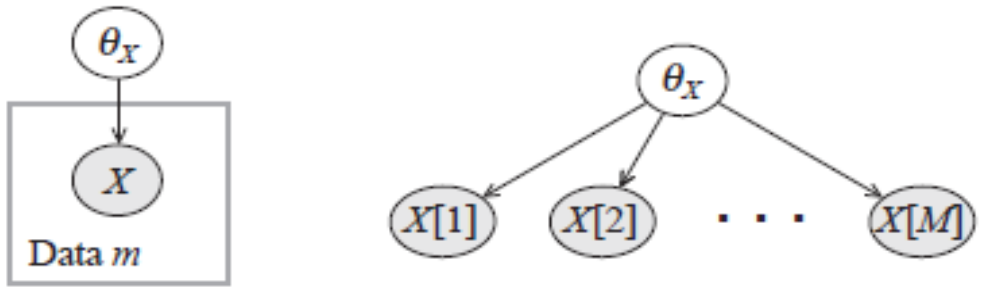


Figure 17.3 Meta-network for IID samples of a random variable X. (a) Plate model; (b) Ground Bayesian network.

Bayesian parameter estimation: joint probabilistic model

In this model, the **samples** are taken **along to the parameter** of the model.

The joint probability of the data and the parameter factorizes as follows:

$$\begin{aligned} P[D, \theta] &= P[D|\theta] \cdot P(\theta) \\ &= P(\theta) \cdot \prod_m P[x[m]|\theta] \cdot \end{aligned}$$

Where $D = \{x[1], ..., x[M]\}$.

Bayesian parameter estimation: the posterior

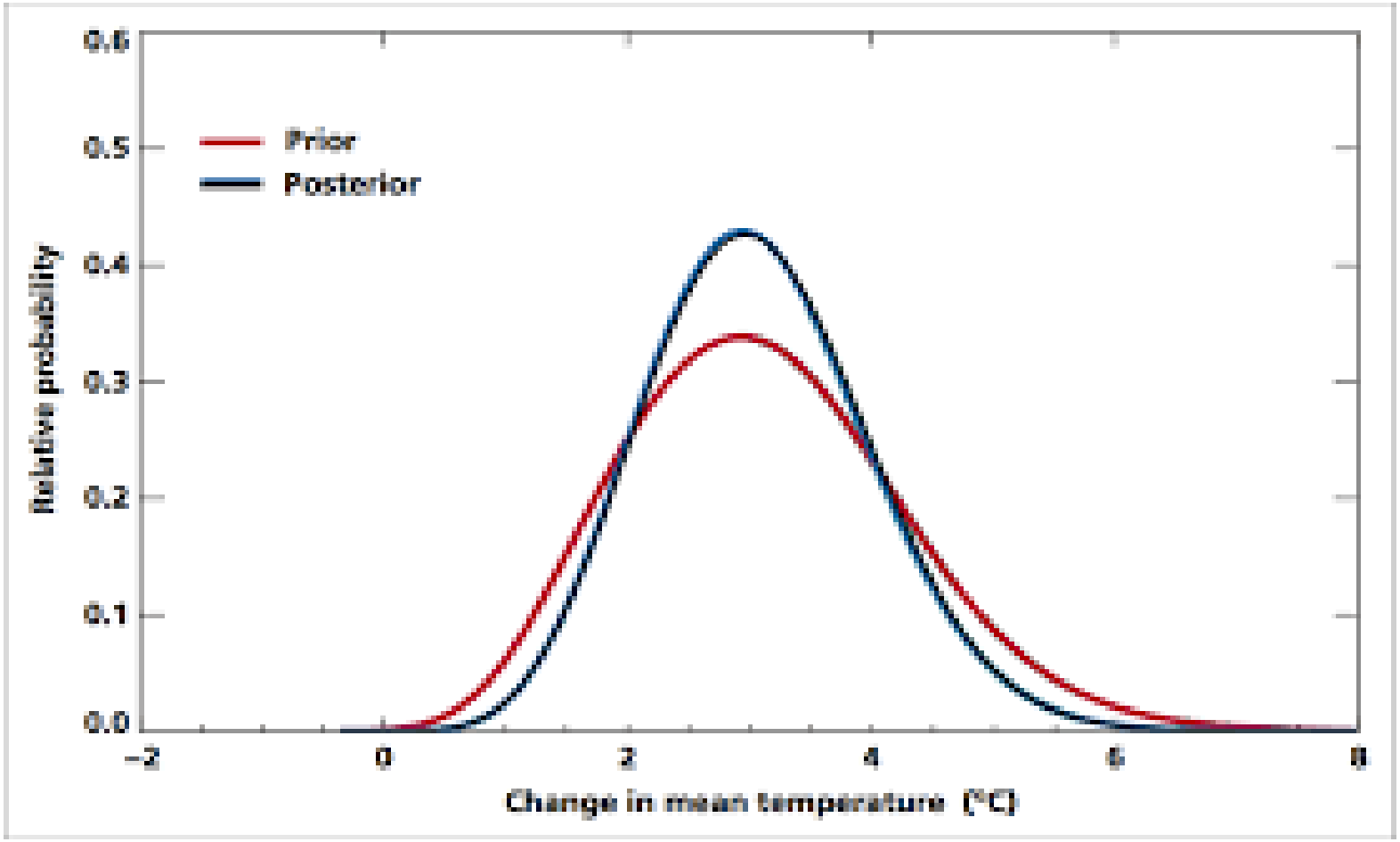
We define the **posterior distribution** as follows.

$$P[\theta|D] = \frac{P[D|\theta] \cdot P(\theta)}{P[D]}$$

The posterior is actually what we are interested in, right?

It **encodes** our **posterior knowledge** about the choices of the parameter **given the data and the prior knowledge**.

Bayesian parameter estimation: the posterior



Bayesian parameter estimation: the posterior

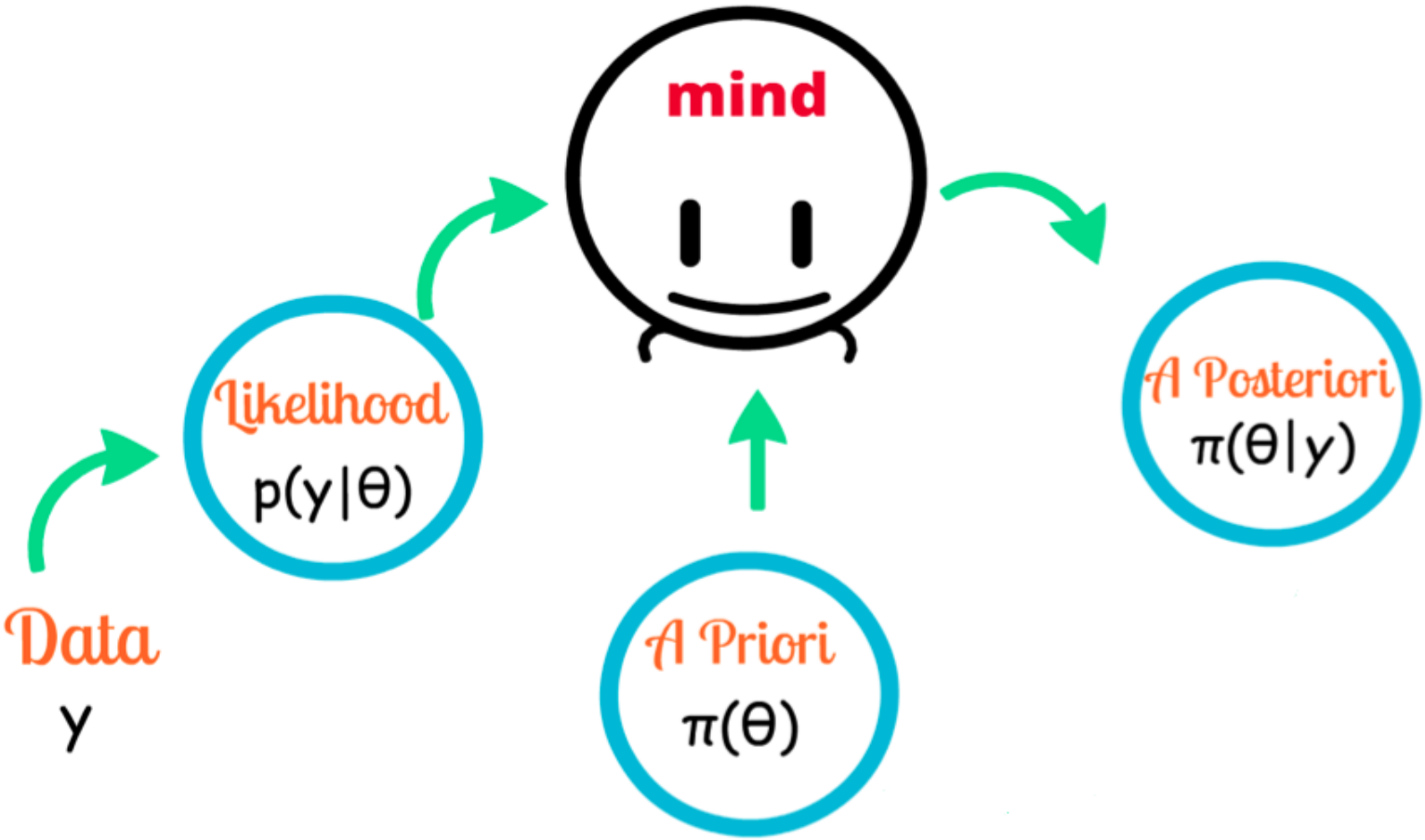
Let's take a further look on the posterior.

$$P[\theta|D] = \frac{P[D|\theta] \cdot P(\theta)}{P[D]}$$

- The first term in the numerator is the **likelihood**.
- The second is the **prior distribution**.
- The denominator is a normalizing constant.

Posterior distribution $\propto$ **likelihood** $\times$ **prior distribution**.

Bayesian parameter estimation: the posterior



Bayesian parameter estimation: prediction

We are also very interested in **making predictions**.

This is done through the distribution for a new sample given the data set:

$$\begin{aligned} & P[x[M+1]|D] \\ &= \int P[x[M+1]|\theta, D] \cdot P[\theta|D] d\theta \\ &= \int P[x[M+1]|\theta] \cdot P[\theta|D] d\theta \end{aligned}$$

Bayesian parameter estimation: prediction

Let's revisit the coin flip example!

Assume that the prior is uniform over $\theta \in [0, 1]$.

What is the probability of a new sample given the data set?

$$P[x[M+1]|D] = \int P[x[M+1]|\theta] \cdot P[\theta|D] d\theta$$

Bayesian parameter estimation: prediction

Since θ is uniformly distributed,

$$P[\theta|D] = P[D|\theta]/P[D]$$

In addition, $P[x[M+1] = head|\theta] = \theta$.

Therefore,

$$\begin{aligned} P[x[M+1] = head|D] &= \int P[x[M+1]|\theta] \cdot P[\theta|D] d\theta \\ &= \frac{1}{P[D]} \int \theta \cdot \theta^{M[head]} \cdot (1-\theta)^{M[tail]} d\theta \\ &= \frac{1}{P[D]} \frac{(M[head]+1)!M[tail]!}{(M[head]+M[tail]+2)!} \end{aligned}$$

(See Beta functions)

Bayesian parameter estimation: prediction

Similarily:

$$P[x[M+1] = tail|D] = \frac{1}{P[D]} \cdot \frac{(M[tail]+1)!M[head]!}{(M[head]+M[tail]+2)!}$$

Their sum is:
$\frac{1}{P[D]} \int \theta^{M[head]} \cdot (1-\theta)^{M[tail]} d\theta = \frac{M[head]!M[tail]!}{(M[head]+M[tail]+1)!}$.

Bayesian parameter estimation: prediction

We normalize and obtain:

$$
\begin{aligned}
&P[x[M+1] = head|D] \\
&= \frac{(M[head]+1)!M[tail]!}{P[D](M[head]+M[tail]+2)!} \Big/ \frac{M[head]!M[tail]!}{P[D](M[head]+M[tail]+1)!} \\
&= \frac{M[head]+1}{M[head]+M[tail]+2}
\end{aligned}
$$

Similar to the MLE prediction except that it **adds one imaginary sample** to each count. As the number of samples grows the Bayesian estimator and the maximum likelihood estimator **converge to the same value**.

Bayesian parameter estimation: Bayesian networks

We now turn to Bayesian estimation in the context of a Bayesian network. Recall that the Bayesian framework **requires us to specify a joint distribution over the unknown parameters and the data instances**.

Bayesian parameter estimation: Bayesian networks

We would like to introduce a simplified formula for the posterior distribution.

For this purpose, we take two simplifying assumptions on the decomposition of the prior.

Bayesian parameter estimation: Bayesian networks

The first assumption asserts global decomposition of the prior.

Definition (Global parameter indepdence)

Let G be a Bayesian network with parameters $\theta = (\theta_{X_1|Pa_{X_1}}, ..., \theta_{X_n|Pa_{X_n}})$. A prior distribution satisfies the global parameter independence if it has the form:

$$P(\theta) = \prod_i P(\theta_{X_i|Pa_{X_i}})$$

Bayesian parameter estimation: Bayesian networks

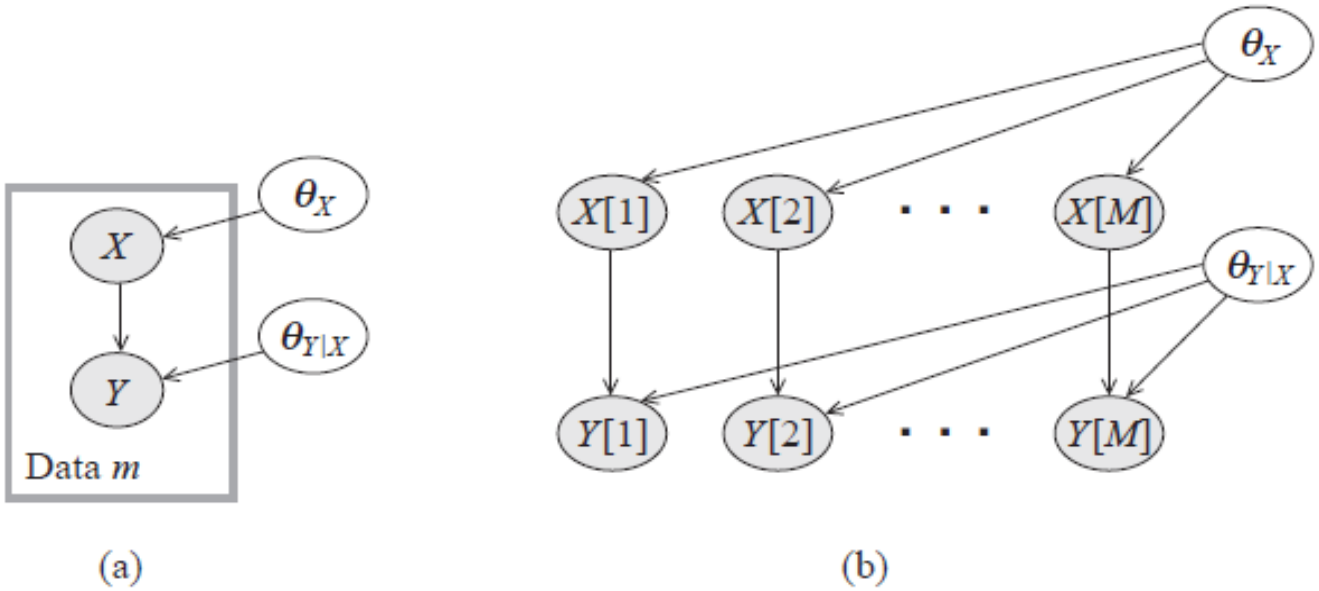


Figure 17.7 Meta-network for IID samples from a network $X \rightarrow Y$ with global parameter independence. (a) Plate model; (b) Ground Bayesian network.

Bayesian parameter estimation: Bayesian networks

The second assumption asserts local decomposition of the prior.

Definition (Local parameter independence)

Let X be a variable with parents U. We say that the prior $P(\theta_{X|U})$ satisfies local parameter independence if

$$P(\theta_{X|U}) = \prod_u P(\theta_{X|u})$$

Bayesian parameter estimation: Bayesian networks

First, we decompose the distribution.

$$P[\theta|D] = \frac{P[D|\theta] \cdot P(\theta)}{P[D]}$$

Recall the decomposition of the likelihood:

$$P[D|\theta] = \prod_i L_i(\theta_{X_i|Pa_{X_i}} : D)$$

And by the global parameter independence:

$$P[\theta|D] = \frac{1}{P[D]} \cdot \prod_i L_i(\theta_{X_i|Pa_{X_i}} : D) \cdot P(\theta_{X_i|Pa_{X_i}})$$

Bayesian parameter estimation: Bayesian networks

By definition of local likelihoods:

$$P[\theta|D] = \frac{1}{P[D]} \cdot \prod_i L_i(\theta_{X_i|Pa_{X_i}} : D) \cdot P(\theta_{X_i|Pa_{X_i}})$$

$$= \frac{1}{P[D]} \cdot \prod_i \prod_m P[x_i[m]|Pa_{X_i}[m]; \theta_{X_i|Pa_{X_i}}] \cdot P(\theta_{X_i|Pa_{X_i}})$$

Bayesian parameter estimation: Bayesian networks

And by applying bayes rule

$$P(\theta_{X_i|Pa_{X_i}}|D) = \frac{P[x_i[m]|Pa_{X_i}[m]; \theta_{X_i|Pa_{X_i}}] \cdot P(\theta_{X_i|Pa_{X_i}})}{P[x_i[m]|Pa_{X_i}[m]]}$$

We obtain,

$$P[\theta|D] = \frac{1}{P[D]} \cdot \prod_i \left(P(\theta_{X_i|Pa_{X_i}}|D) \cdot \prod_m P[x_i[m]|Pa_{X_i}[m]] \right) = \prod_i P(\theta_{X_i|Pa_{X_i}}|D)$$

Finally, by the local parameter independence:

$$P[\theta|D] = \prod_i \prod_{p \text{ instantiation of } Pa_{X_i}} P(\theta_{X_i|p}|D)$$

Bayesian parameter estimation: selecting a prior

It is left to choose the prior distribution.

We arrived to a very pleasing decomposition of the prior,

$$P(\theta) = \prod_i \prod_p P(\theta_{X_i|p})$$

Each $\theta_{X_i|p}$ behaves as a discrete distribution, i.e, a vector that sums to 1.

How would you choose the prior on each $\theta_{X_i|p}$?

Bayesian parameter estimation: selecting a prior

A widely applicable distribution over discrete distributions is the Dirichlet distribution.

$$(a_1, ..., a_k) \sim Dir(\alpha_1, ..., \alpha_k) \text{ s.t } \sum_i a_i = 1 \text{ and } a_i \in (0, 1)$$

The PDF is:

$$\frac{1}{B(\alpha_1, ..., \alpha_k)} \prod_i a_i^{\alpha_i - 1} \text{ where } B(\alpha_1, ..., \alpha_k) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)}$$

Here $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$ is the well known Gamma function.

Bayesian parameter estimation: selecting a prior

The following theorem shows that it is convenient to choose dirichlet distributions as priors.

Theorem

If $\theta \sim Dir(\alpha_1, ..., \alpha_k)$ *then* $\theta|D \sim Dir(\alpha_1 + M[1], ..., \alpha_k + M[k])$ *where* $M[k]$ *counts the number of occurances of* k *in* D.

[See h.w]

Bayesian parameter estimation: selecting a prior

Finally we have:

$$P[\theta|D] = \prod_{i} \prod_{p \text{ instantiation of } Pa_{X_i}} P(\theta_{X_i|p}|D)$$

Where each prior is:

$$\theta_{X|p} \sim Dir(\alpha_{x^1|p}, ..., \alpha_{x^K|p})$$

And therefore the posterior is:

$$\theta_{X_i|p}|D \sim Dir(\alpha_{x^1|p} + M[p, x^1], ..., \alpha_{x^K|p} + M[p, x^K])$$

Bayesian parameter estimation: making predictions

The **predictive model** is dictated by the following distribution.

$$
\begin{aligned}
&P[X_1[M+1], ..., X_n[M+1]|D] \\
&= \prod_i P[X_i[M+1]|Pa_{X_i}[M+1], D] \\
&= \prod_i \int P[X_i[M+1]|Pa_{X_i}[M+1], \theta_{X_i|Pa_{X_i}}] \cdot P[\theta_{X_i|Pa_{X_i}}|D] d\theta_{X_i|Pa_{X_i}}
\end{aligned}
$$

Bayesian parameter estimation: making predictions

notice that:

$$(X_i[M+1]|Pa_{X_i}[M+1] = p, \theta_{X_i|p}) \sim Categorical(\theta_{X_i|p})$$

And

$$\theta_{X_i|p}|D \sim Dir(\alpha_{x^1|p} + M[p, x^1], ..., \alpha_{x^K|p} + M[p, x^K])$$

In analogue to the coin flips example, the posterior induces a predictive model in which:

$$P[X_i[M+1] = x_i|Pa_{X_i}[M+1] = p, D] = \frac{\alpha_{x_i|p} + M[p, x_i]}{\sum_i \alpha_{x_i|p} + M[p, x_i]}$$

Generalization analysis

One intuition that permeates our discussion is that more training instances give rise to more accurate parameter estimates.

Next, we provide some formal analysis that supports this intuition.

All generalizations are false,
including this one.
Marc Twain.

Generalization analysis: Almost surely convergence

Theorem

Let P^* be the generating distribution, let $P(\cdot;\theta)$ be a parametric family of distributions, and let $\theta^* = \arg\min_\theta KL(P^*||P(\cdot;\theta))$ be the M-projection on P^* on this family. In addition, $\hat{\theta} = \arg\min_\theta KL(\hat{P}_D||P(\cdot;\theta))$. Then,

$$\lim_{M\to\infty} P(\cdot;\hat{\theta}) = P(\cdot;\theta^*)$$

Almost surely.

Generalization analysis: Convergence of multinomials

We revisit the coin flip example.

This time we are interested to measure the number of samples required to approximate the probability for head/tail.

Generalization analysis: Convergence of multinomials

Assume we have a data set $D = \{x_1, ..., x_M\}$ of i.i.d samples $x_i \sim Binomial(\theta)$.
We would like to estimate **how large** M should be in order to **suffice** that the **MLE is close to** θ.

Generalization analysis: Convergence of multinomials

Theorem

Let $\epsilon, \delta > 0$ *and let* $M > \frac{1}{2\epsilon^2} \log(2/\delta)$, *then:*

$$P[|\hat{\theta} - \theta| < \epsilon] \geq 1 - \delta$$

The probability is over D, *as set of* M *i.i.d samples. Here,* $\hat{\theta}$ *is the MLE with respect to* D.

[See h.w]