Applied Machine Learning

Maximum Likelihood and Bayesian Reasoning

Reihaneh Rabbany

McGill
School of Computer Science

COMP 551 (winter 2023)

Admin

- Study groups should be open for enrolment now
    - Join one based on your schedule and availabilities

TODO

- More details on Midterm
    - Date is March 16th
    - Closed book but your hand written notes are allowed

Model fitting

x | input features → **ML algorithm** → y | output labels

with parameters $\boldsymbol{\theta}$

$f(x^{(n)};\theta)$

the process of estimating the model parameters θ from given data $\mathcal{D}$, is the core of training ML models which often boils down into optimization of an loss function $\mathcal{L}(\theta)$

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{n=1}^{N} l(y^{(n)}, f(x^{(n)};\theta))$$

$$\theta^* = \arg\min_\theta \mathcal{L}(\theta)$$

Model fitting

x | input
features → **ML algorithm**
with parameters θ → y | output
labels

the process of estimating the model parameters θ from given data $\mathcal{D}$, is the core of training ML models which often boils down into optimization of an loss function $\mathcal{L}(\theta)$

A common approach is to use negative log probability as our loss function:

$$l(y, f(x,\theta)) = -\log p(y|f(x;\theta))$$

Objectives

learn common parameter estimation methods and understand what it means to learn a probabilistic model of the data

- using maximum likelihood principle
- using Bayesian inference
  - prior, posterior, posterior predictive
  - MAP inference
  - Beta-Bernoulli conjugate pairs

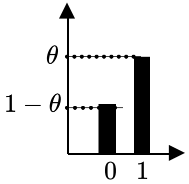
Parameter estimation

a coin's head/tail outcome has a **Bernoulli distribtion**

$$\mathrm{Bernoulli}(x|\theta) = \theta^x (1-\theta)^{(1-x)}$$

reminder: Bernoulli random variable takes values of 0 or 1, e.g. head/tail in a coin toss

$$p(x|\theta) = \begin{cases} \theta & x = 1 \\ 1-\theta & x = 0 \end{cases}$$



IID is short for *independent and identically distributed*

this is our **probabilistic model** of some head/tail IID data $\mathcal{D} = \{0,0,1,1,0,0,1,0,0,1\}$

Objective: learn the model parameter θ

since we are only interested in the counts, we can also use **Binomial distribution**

$$\mathrm{Binomial}(N, N_h|\theta) = \binom{N}{N_h} \theta^{N_h} (1-\theta)^{N-N_h}$$

$|\mathcal{D}|$

\# heads $N_h = \sum_{x \in \mathcal{D}} x$

N_t

Maximum likelihood

a coin's head/tail outcome has a **Bernoulli distribtion**

$$\text{Bernoulli}(x|\theta) = \theta^x (1-\theta)^{(1-x)}$$

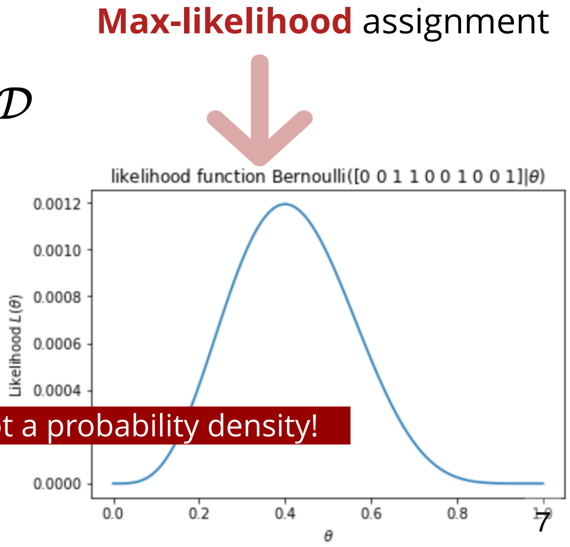
this is our **probabilistic model** of some head/tail IID data $\mathcal{D} = \{0,0,1,1,0,0,1,0,0,1\}$

Objective: learn the model parameter θ

Idea: find the parameter θ that maximizes the probability of observing $\mathcal{D}$

Likelihood $L(\theta;\mathcal{D}) = \prod_{x\in\mathcal{D}} \text{Bernoulli}(x|\theta) = \theta^4(1-\theta)^6$ is a function of θ

pick the parameters that assign the highest probability to the training data



Max-likelihood assignment

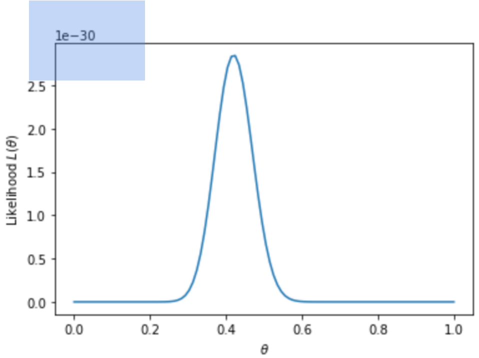
note that this is not a probability density!

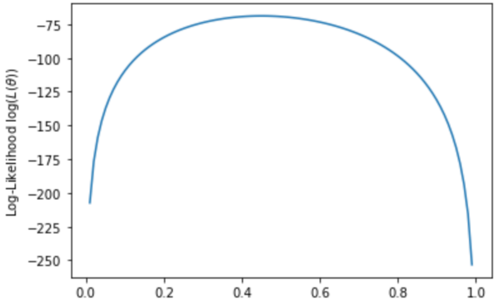
Maximizing log-likelihood

likelihood $L(\theta; \mathcal{D}) = \prod_{x \in \mathcal{D}} p(x; \theta)$

using product here creates extreme values

for 100 samples in our example, the likelihood shrinks below 1e-30





log-likelihood has the same maximum but it is well-behaved

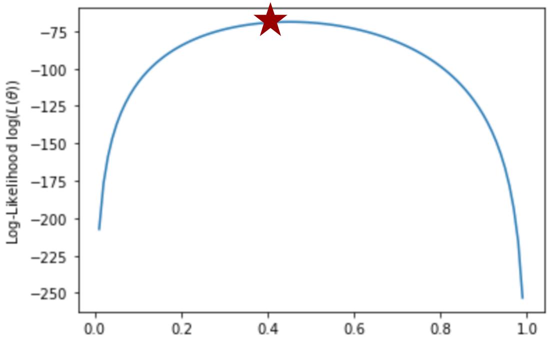
$$\ell(\theta; \mathcal{D}) = \log(L(\theta; \mathcal{D})) = \sum_{x \in \mathcal{D}} \log(p(x; \theta))$$

how do we find the max-likelihood parameter? $\theta^* = \arg\max_\theta \ell(\theta; \mathcal{D})$

for some simple models we can get the ***closed form solution***

for complex models we need to use ***numerical optimization***

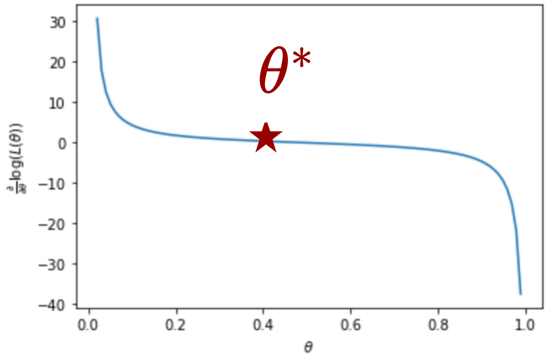
Maximizing log-likelihood



log-likelihood $\ell(\theta;\mathcal{D}) = \log(L(\theta;\mathcal{D})) = \sum_{x\in\mathcal{D}} \log(\text{Bernoulli}(x;\theta))$

observation: at maximum, the derivative of $\ell(\theta;\mathcal{D})$ is zero

idea: set the the derivative to zero and solve for θ



example max-likelihood for Bernoulli

$$\frac{\partial}{\partial\theta}\ell(\theta;\mathcal{D}) = \frac{\partial}{\partial\theta}\sum_{x\in\mathcal{D}} \log\left(\theta^{x}(1-\theta)^{(1-x)}\right)$$

$$= \frac{\partial}{\partial\theta}\sum_{x} x\log\theta + (1-x)\log(1-\theta)$$

$$= \sum_{x}\frac{x}{\theta} - \frac{1-x}{1-\theta} = 0$$

which gives $\theta^{MLE} = \frac{\sum_{x\in\mathcal{D}} x}{|\mathcal{D}|}$ is simply the portion of heads in our dataset

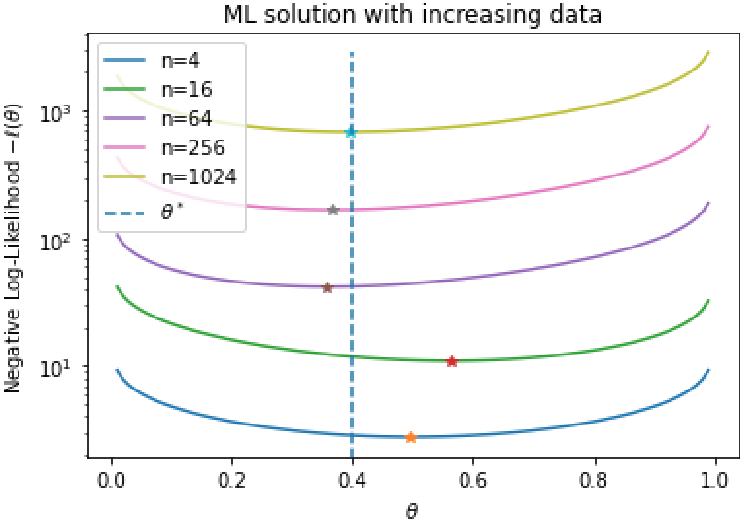
what is θ^{MLE} when $\mathcal{D} = \{0,0,1,1,0,0,1,0,0,1\}$?

Bayesian approach

max-likelihood estimate does not reflect our uncertainty:

- e.g., $\theta^{MLE} = 1$ if we observe only one head, predicts all future tosses are head!
- e.g., $\theta^{MLE} = .2$ for both 1/5 heads and 1000/5000 heads
    - in which case are we more certain of the predicted θ?

How can we quantify our uncertainty about our prediction?



Bayesian approach

How can we quantify our uncertainty about our prediction?
capture it using a conditional probability distribution instead of a single best guess

Using the Bayesian inference approach

- we maintain a *distribution* over parameters $p(\theta)$ **prior** what do we believe about θ before any observation
- after observing $\mathcal{D}$ we update this distribution $p(\theta|\mathcal{D})$ **posterior**

how to update degree of certainty given data? using **Bayes rule**

$$\overset{\text{hidden}}{p(\theta}|\underset{\text{observed}}{\mathcal{D})} = \frac{\overset{\text{prior}}{p(\theta)}p(\mathcal{D}|\theta)}{p(\mathcal{D})}$$

likelihood of the data
previously denoted by $L(\theta;\mathcal{D})$

evidence: this is a normalization, **marginal likelihood of data**

$$p(\mathcal{D}) = \int p(\theta')p(\mathcal{D}|\theta')\mathrm{d}\theta'$$

We can get a point estimate by collapsing this posterior distribution to a single point, i.e. the best guess

Bayes rule: example reminder

$c = \{\text{yes}, \text{no}\}$ patient having cancer?

$x \in \{-, +\}$ observed test results, a single binary feature

prior: .1% of population has cancer $p(\text{yes}) = .001$

likelihood: $p(+|\text{yes}) = .9$ TP rate of the test (90%)

$$p(c = yes \mid x) = \frac{p(c{=}yes)p(x|c{=}yes)}{p(x)}$$

FP rate of the test (5%)

posterior: $p(\text{yes}|+) = .0177$

evidence: $p(+) = p(\text{yes})p(+|\text{yes}) + p(\text{no})p(+|\text{no}) = .001 \times .9 + .999 \times .05 = .05$

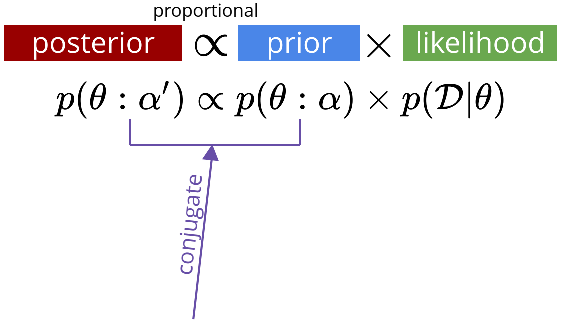
Conjugate Priors

in our coin example, we know the form of likelihood:

prior $p(\theta)?$

posterior $p(\theta|\mathcal{D})?$

likelihood $p(\mathcal{D}|\theta) = \prod_{x \in \mathcal{D}} \text{Bernoulli}(x;\theta) = \theta^{N_h}(1-\theta)^{N_t}$



proportional

posterior ∝ **prior** × **likelihood**

$$p(\theta : \alpha') \propto p(\theta : \alpha) \times p(\mathcal{D}|\theta)$$

To simplify the computation we want prior and posterior to have the **same form**

(so that we can easily update our belief with new observations)

this gives us the following form $p(\theta|a,b) \propto \theta^a(1-\theta)^b$

this means there is a normalization constant that does not depend on θ

↓

distribution of this form has a name, **Beta** distribution

we say Beta distribution is a conjugate prior to the Bernoulli likelihood

Beta distribution

Beta distribution has the following density

$$\mathrm{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

normalization

$\alpha, \beta > 0$

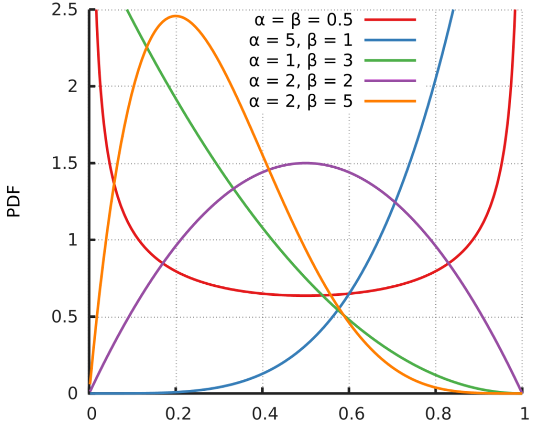
Γ is the generalization of factorial to real number $\Gamma(a+1) = a\Gamma(a)$

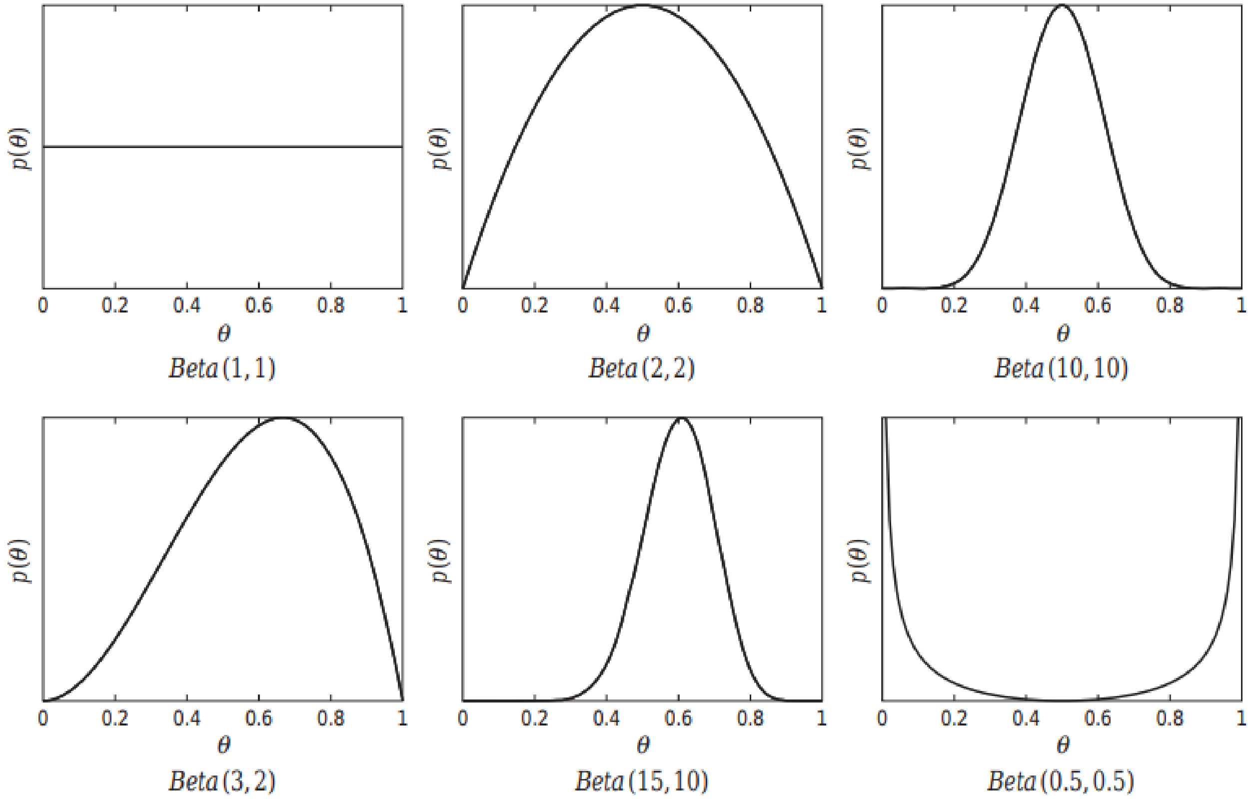
$\mathrm{Beta}(\theta|\alpha=\beta=1)$ is uniform

mean of the distribution is $\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta}$

for $\alpha, \beta > 1$ the dist. is unimodal; its mode is $\frac{\alpha-1}{\alpha+\beta-2}$

Beta distribution: more examples





$Beta(1,1)$ $Beta(2,2)$ $Beta(10,10)$

$Beta(3,2)$ $Beta(15,10)$ $Beta(0.5,0.5)$

Beta-Bernoulli conjugate pair

how to model probability of heads when we toss a coin N times

proportional

posterior $\propto$ **prior** $\times$ **likelihood**

prior $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ $\quad p(\theta) = \mathrm{Beta}(\theta|\alpha, \beta)$

likelihood $p(\mathcal{D}|\theta) = \theta^{N_h}(1-\theta)^{N_t}$ $\quad L(\theta;\mathcal{D}) = \prod \mathrm{Bernoulli}(N_h, N_t|\theta)$

product of Bernoulli likelihoods
equivalent to Binomial likelihood

posterior $p(\theta|\mathcal{D}) \propto \theta^{\alpha+N_h-1}(1-\theta)^{\beta+N_t-1}$ $\quad p(\theta|\mathcal{D}) = \mathrm{Beta}(\theta|\alpha+N_h, \beta+N_t)$

α, β are called *pseudo-counts*

their effect is similar to imaginary observation of heads (α) and tails (β)

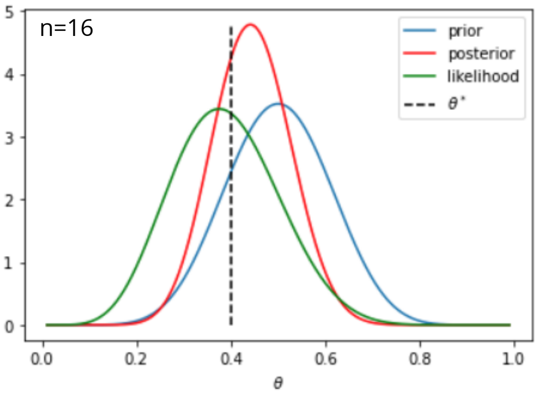
Effect of more data

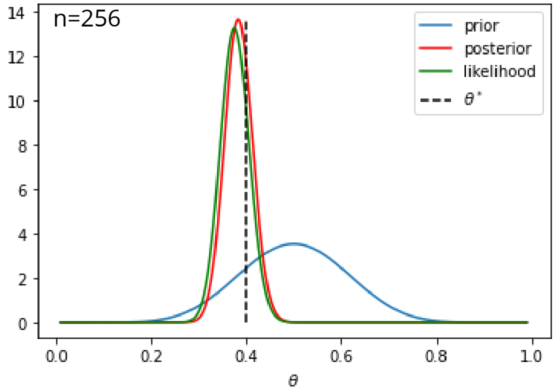
with few observations, prior has a high influence
as we increase the number of observations $N = |\mathcal{D}|$ the effect of prior diminishes
the likelihood term dominates the posterior

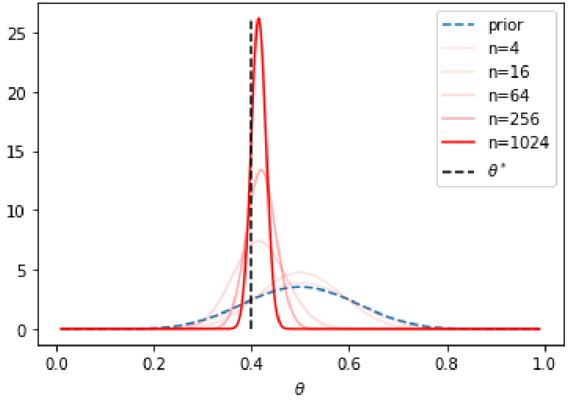
example prior $\mathrm{Beta}(\theta|10, 10)$

plot of the posterior density with **n** observations

$$p(\theta|\mathcal{D}) \propto \theta^{10+H}(1-\theta)^{10+N-H}$$







Posterior predictive

our goal was to estimate the parameters (θ) so that we can make predictions

what if we use the maximum likelihood estimste for the best parameter, θ^{MLE}, and plug it in the $p(x|\theta)$ to make the prediction?

Example:

if we see four heads in a row, what is the probability of seeing a tail next?

if $\mathcal{D} = \{1,1,1,1\}$, what is θ^{MLE}? 1.0

$$\Rightarrow 1 - \theta^{MLE} = 0.0$$

$$p(0|\theta) = \theta^0(1-\theta)^{(1-0)} = 1 - \theta$$

Next, let's use the posterior distribution we learn through Bayesian inference

Posterior predictive

our goal was to estimate the parameters (θ) so that we can make predictions

now we have a (posterior) **distribution** over parameters, $p(\theta|\mathcal{D})$, rather than a single θ^{MLE}

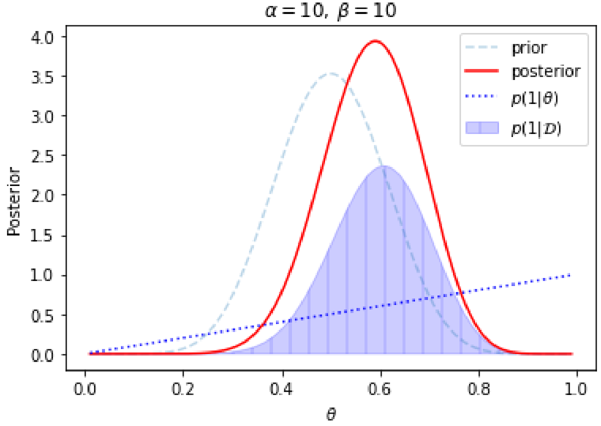
θ^{MLE} only gives a single best guess based on that parameter, $p(x|\theta)$

To make predictions, we calculate the average prediction over all possible values of θ

$$p(x|\mathcal{D}) = \int_{\theta} p(\theta|\mathcal{D}) p(x|\theta) \mathrm{d}\theta$$

for each possible θ, weight the prediction by the posterior probability of that parameter being true

posterior predictive



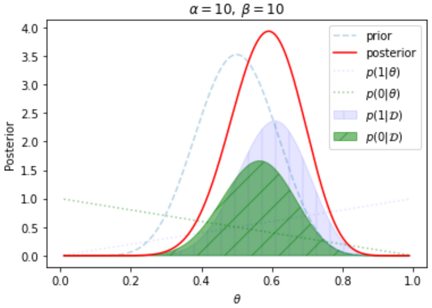
Posterior predictive

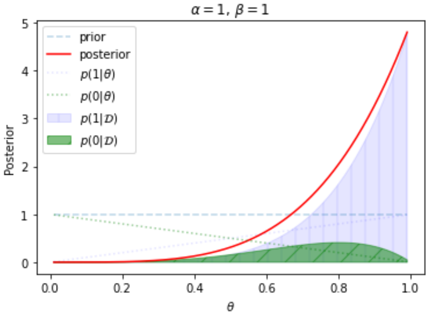
our goal was to estimate the parameters (θ) so that we can make predictions

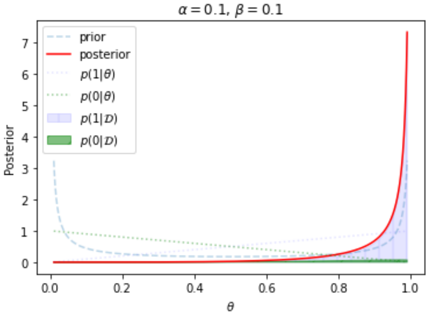
now we have a (posterior) **distribution** over parameters, $p(\theta|\mathcal{D})$

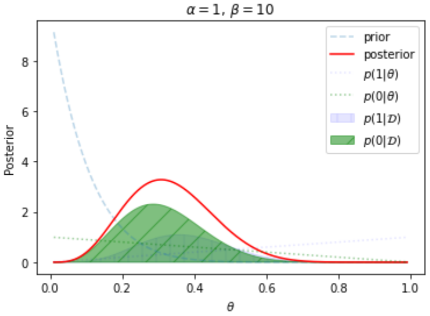
To make predictions, we calculate the average prediction over all possible values of θ

Example if we see four heads in a row, what is the probability of seeing a tail next?
if $\mathcal{D} = \{1, 1, 1, 1\}$, what is $p(0|\mathcal{D})$? depends on our prior belief







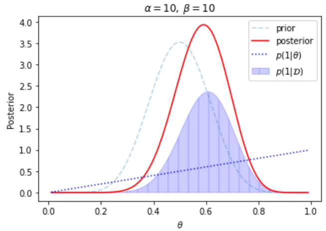


when the strenght of prior gets close to zero the prediction becomes similar to MLE

Posterior predictive for Beta-Bernoulli

start from a Beta prior $p(\theta) = \mathrm{Beta}(\theta|\alpha, \beta)$

observe N_h heads and N_t tails, the posterior is $p(\theta|\mathcal{D}) = \mathrm{Beta}(\theta|\alpha + N_h, \beta + N_t)$



Given this estimate of the parameters from training data, how can we predict the future?

what is the probability that the next coin flip is head?

marginalize over θ

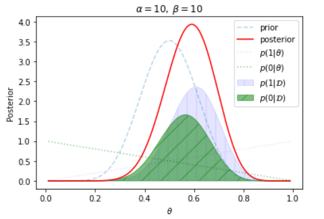
$$p(x = 1|\mathcal{D}) = \int_\theta \mathrm{Bernoulli}(x = 1|\theta)\mathrm{Beta}(\theta|\alpha + N_h, \beta + N_t)\mathrm{d}\theta$$

$$= \underbrace{\int_\theta \theta\, \mathrm{Beta}(\theta|\alpha + N_h, \beta + N_t) d\theta}_{\text{mean of Beta dist.}} = \frac{\alpha+N_h}{\alpha+\beta+N}$$

Example if we see four heads in a row, what is the probability of seeing a tail next?

if $\mathcal{D} = \{1, 1, 1, 1\}$, what is $p(1|\mathcal{D})$? $\frac{14}{24}$, $p(0|\mathcal{D})$? $\frac{10}{24}$

when we assume the prior is $\mathrm{Beta}(\alpha = 10, \beta = 10)$



compare with prediction of maximum-likelihood: $p(x = 1|\mathcal{D}) = \frac{N_h}{N} = 1,\ p(x = 1|\mathcal{D}) = 0$

Posterior predictive for Beta-Bernoulli

start from a Beta prior $p(\theta) = \mathrm{Beta}(\theta|\alpha, \beta)$

observe N_h heads and N_t tails, the posterior is $p(\theta|\mathcal{D}) = \mathrm{Beta}(\theta|\alpha + N_h, \beta + N_t)$

Given this estimate of the parameters from training data, how can we predict the future?

$$p(x = 1|\mathcal{D}) = \int_\theta \mathrm{Bernoulli}(x = 1|\theta)\mathrm{Beta}(\theta|\alpha + N_h, \beta + N_t)\mathrm{d}\theta = \frac{\alpha+N_h}{\alpha+\beta+N}$$

compare with prediction of maximum-likelihood: $p(x = 1|\mathcal{D}) = \frac{N_h}{N}$

if we assume a uniform prior, the posterior predictive is $p(x = 1|\mathcal{D}) = \frac{N_h+1}{N+2}$

Laplace smoothing

a.k.a. add-one smoothing

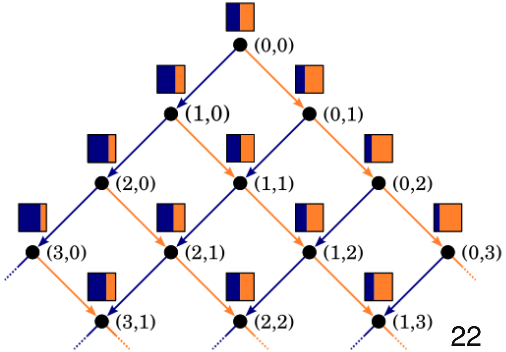
to avoid ruling out unseen cases with zero counts

Example:

sequential Baysian updating

with uniform prior

(N_h, N_t)

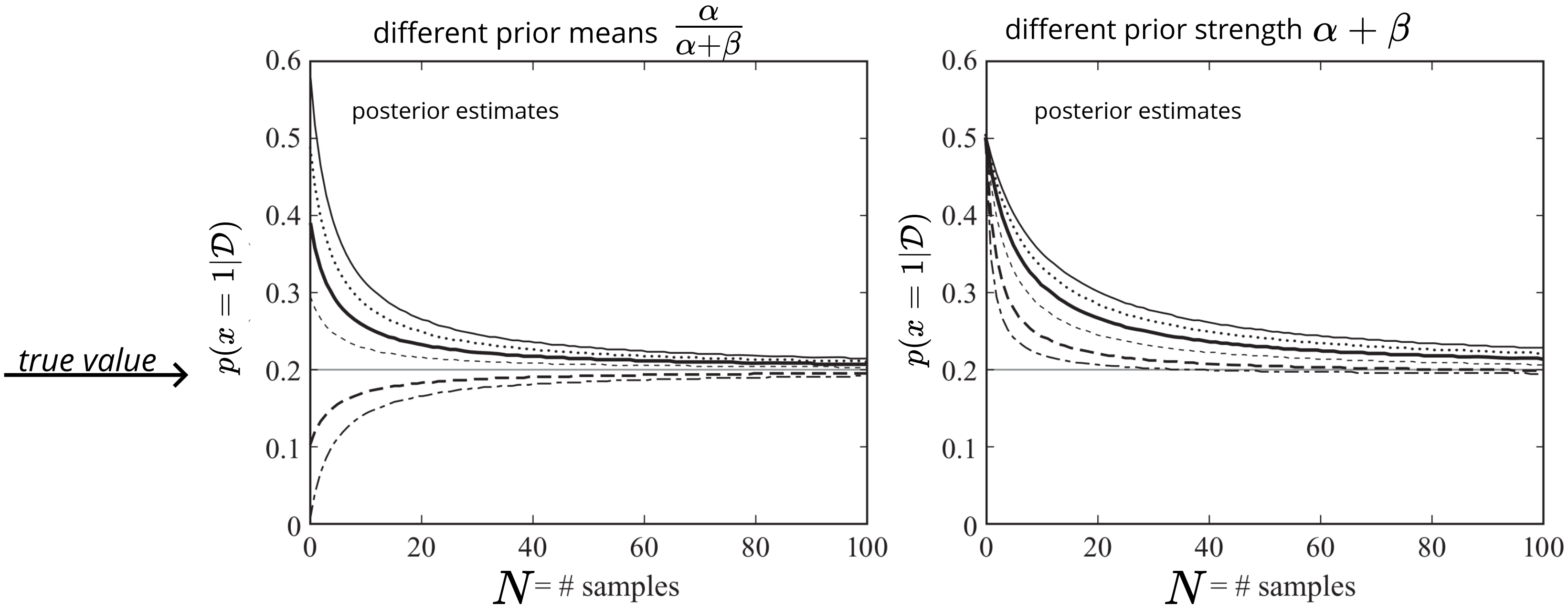


Strength of the prior

with a **strong prior** we need many samples to really change the posterior

for Beta distribution $\alpha + \beta$ decides how strong the prior is: how confident we are in our prior

example as our dataset grows our estimate becomes more accurate



different prior means $\frac{\alpha}{\alpha+\beta}$

different prior strength $\alpha + \beta$

example: PGM book by Koller & Friedman, figure 17.5

Maximum a Posteriori (MAP)

sometimes it is difficult to work with the posterior dist. over parameters

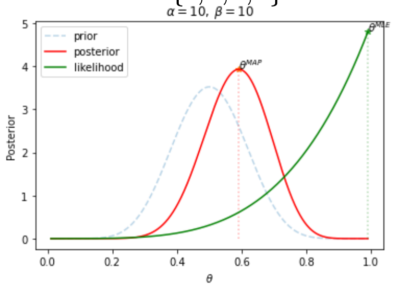
alternative: use the parameter with the highest posterior probability $p(\theta|\mathcal{D})$

MAP estimate $\theta^{MAP} = \arg\max_\theta p(\theta|\mathcal{D}) = \arg\max_\theta p(\theta)p(\mathcal{D}|\theta)$

compare with max-likelihood estimate *(the only difference is in the prior term)*

$$\theta^{MLE} = \arg\max_\theta p(\mathcal{D}|\theta)$$

$\mathcal{D} = \{1, 1, 1, 1\}$



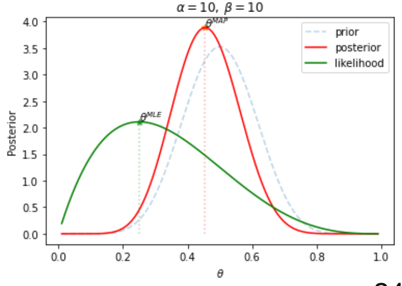
example for the posterior $p(\theta|\mathcal{D}) = \text{Beta}(\theta|\alpha + N_h, \beta + N_t)$

MAP estimate is the **mode** of posterior $\theta^{MAP} = \frac{\alpha+N_h-1}{\alpha+\beta+N_h+N_t-2}$

compare with MLE $\theta^{MLE} = \frac{N_h}{N_h+N_t}$

they are equal for uniform prior $\alpha = \beta = 1$

$\mathcal{D} = \{0, 0, 0, 1\}$



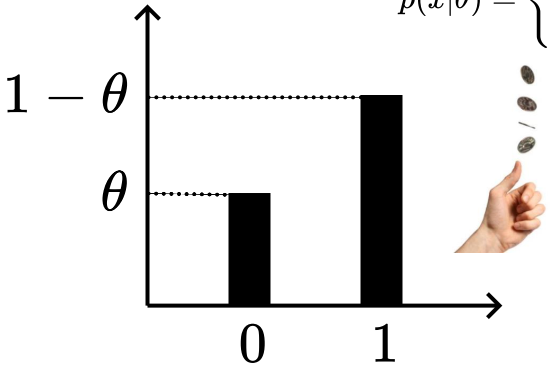
Categorical distribution

what if we have more than two categories (e.g., loaded dice instead of coin)

instead of Bernoulli we have multinoulli or **categorical** dist.

$$\text{Bernoulli}(x|\theta) = \theta^{x}(1-\theta)^{(1-x)}$$

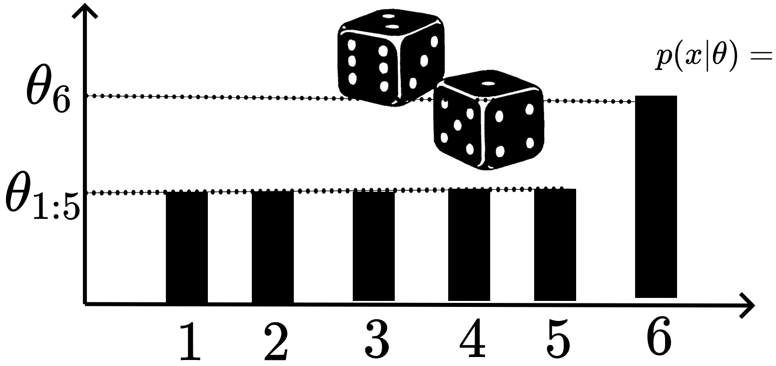
$$p(x|\theta) = \begin{cases} \theta & x = 1 \\ 1-\theta & x = 0 \end{cases}$$



categories

$$\text{Cat}(x|\theta) = \prod_{k=1}^{K} \theta_k^{\mathbb{I}(x=k)}$$

$$p(x|\theta) = \begin{cases} \theta_1 & x = 1 \\ \theta_2 & x = 2 \\ \theta_3 & x = 3 \\ \theta_4 & x = 4 \\ \theta_5 & x = 5 \\ \theta_6 & x = 6 \end{cases}$$



once:	Bernoulli distribution	categorical distribution
n times:	binomial distribution	multinomial distribution

Categorical distribution

what if we have more than two categories (e.g., loaded dice instead of coin)
instead of Bernoulli we have multinoulli or **categorical** dist.

categories

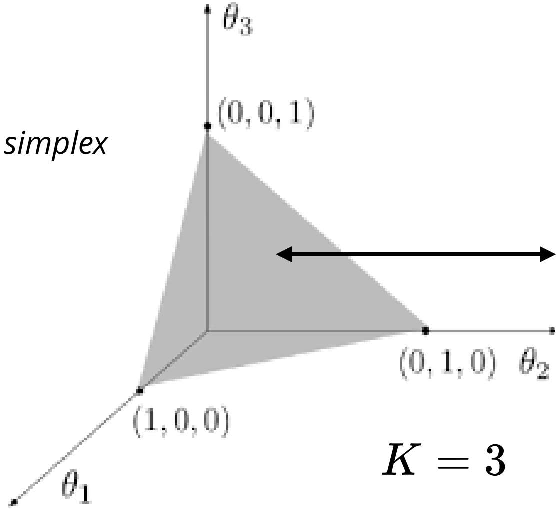
$$\mathrm{Cat}(x|\theta) = \prod_{k=1}^{K} \theta_k^{\mathbb{I}(x=k)}$$

where $\sum_k \theta_k = 1$

θ *belongs to probability simplex*

$$p(x|\theta) = \begin{cases} \theta_1 & x = 1 \\ \theta_2 & x = 2 \\ \theta_3 & x = 3 \\ \theta_4 & x = 4 \\ \theta_5 & x = 5 \\ \theta_6 & x = 6 \end{cases}$$

$\sum_k^6 \theta_k = 1$



$\theta_1 + \theta_2 + \theta_3 = 1$

$K = 3$

Maximum likelihood for categorical dist.

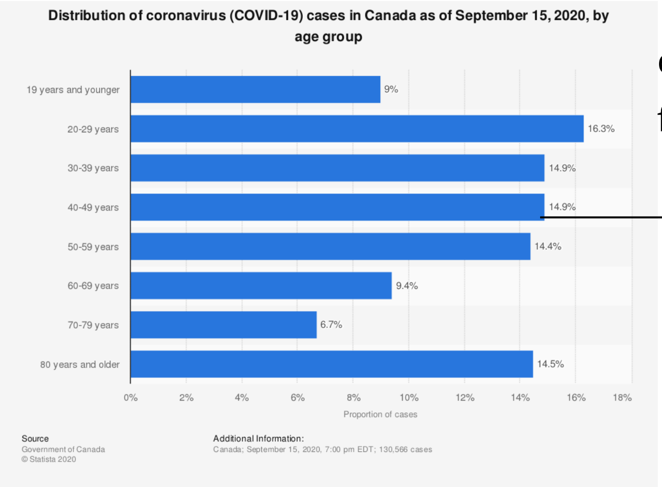
likelihood $p(\mathcal{D}|\theta) = \prod_{x \in \mathcal{D}} \mathrm{Cat}(x|\theta) = \prod_{x \in \mathcal{D}} \prod_{k=1}^{K} \theta_k^{\mathbb{I}(x=k)} = \prod_{k=1}^{K} \theta_k^{N_k}$, $N_k = \sum_{x \in \mathcal{D}} \mathbb{I}(x = k)$

log-likelihood $\ell(\theta, \mathcal{D}) = \sum_{x \in \mathcal{D}} \sum_k \mathbb{I}(x = k) \log(\theta_k) = \sum_k N_k \log(\theta_k)$

we need to solve $\frac{\partial}{\partial \theta_k} \ell(\theta, \mathcal{D}) = 0$ subject to $\sum_k \theta_k = 1$ using Lagrange multipliers

similar to the binary case, max-likelihood estimate is given by data-frequencies $\theta_k{}^{MLE} = \frac{N_k}{N}$

example



Distribution of coronavirus (COVID-19) cases in Canada as of September 15, 2020, by age group

Source
Government of Canada
© Statista 2020

Additional Information:
Canada; September 15, 2020, 7:00 pm EDT; 130,566 cases

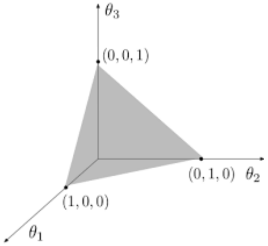
categorical distribution with K=8

frequencies are max-likelihood parameter estimates

$\theta_5^{MLE} = .149$

optional

Dirichlet distribution



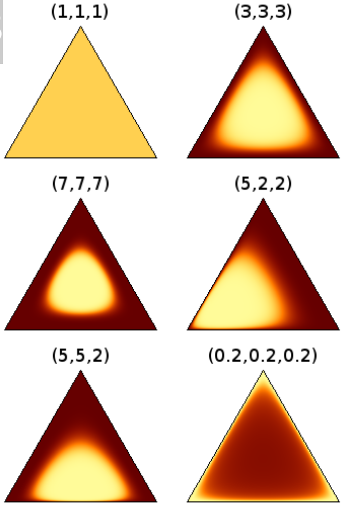
is a distribution over the parameters θ of a Categorical dist.

is a generalization of Beta distribution to K categories

this should be a dist. over prob. simplex $\sum_k \theta_k = 1$

$\alpha =$

$K = 3$



$\mathrm{Dir}(\theta, [.2, .2, .2])$

$$\mathrm{Dir}(\theta|\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}$$

normalization constant

vector of psedo-counts for K categories *(aka concentration parameters)*

$\alpha_k > 0 \; \forall k$

for $\alpha = [1, \ldots, 1]$, we get uniform distribution

for K=2, it reduces to Beta distribution optional

Dirichlet-Categorical conjugate pair

Dirichlet dist. $\mathrm{Dir}(\theta|\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}$ is a conjugate prior for Categorical dist. $\mathrm{Cat}(x|\theta) = \prod_k \theta_k^{\mathbb{I}(x=k)}$

posterior $\propto$ prior $\times$ likelihood

prior $p(\theta) = \mathrm{Dir}(\theta|\alpha) \propto \prod_k \theta_k^{\alpha_k - 1}$

likelihood $p(\mathcal{D}|\theta) = \prod_k \theta_k^{N_k}$ we observe $\overset{\eta}{N_1, \ldots, N_K}$ values from each category

posterior $p(\theta|\mathcal{D}) = \mathrm{Dir}(\theta|\alpha + \eta) \propto \prod_k \theta_k^{N_k + \alpha_k - 1}$ again, we add the real counts to pseudo-counts

posterior predictive $p(x = k|\mathcal{D}) = \frac{\alpha_k + N_k}{\sum_{k'} \alpha_{k'} + N_{k'}}$

MAP $\theta_k^{MAP} = \frac{\alpha_k + N_k - 1}{(\sum_{k'} \alpha_{k'} + N_{k'}) - K}$

Summary

in ML we often build a probabilistic model of the data $p(x;\theta)$

learning a good model could mean **maximizing the likelihood** of the data

$\max_\theta \log p(\mathcal{D}|\theta)$ | sometimes closed form solution
for more complex p, we use numerical methods

an alternative is a **Bayesian approach**:

- maintain a **distribution** over model parameters
- can specify our **prior** knowledge $p(\theta)$
- we can use **Bayes rule** to update our belief after new oabservation $p(\theta|\mathcal{D})$
- we can make predictions using **posterior predictive** $p(x|\mathcal{D})$
- can be computationally **expensive** *(not in our examples so far)*

a middle path is **MAP estimate**: $\max_\theta \log p(\mathcal{D}|\theta)p(\theta)$

- models our **prior** belief
- use a single point estimate and picks the model with highest posterior probability