Estimation

Pre-lecture 3 video

Estimation

Finding the right parameters θ to describe a distribution

- Example: I flip a coin 3 times and get HHH. What is the probability of heads for this coin?
  - You could say P(H)=1 (frequentist)
  - or you might have a prior belief about the coin and integrate observation and prior together (Bayes)

One of the most important operations involving probabilities is that of finding weighted averages of functions. The average value of some function $f(x)$ under a probability distribution $p(x)$ is called the *expectation* of $f(x)$ and will be denoted by $\mathbb{E}[f]$. For a discrete distribution, it is given by

$$\mathbb{E}[f] = \sum_x p(x) f(x) \tag{1.33}$$

so that the average is weighted by the relative probabilities of the different values of x. In the case of continuous variables, expectations are expressed in terms of an integration with respect to the corresponding probability density

$$\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x. \tag{1.34}$$

Bernoulli distribution is the probability of a single coin flip turning up heads (x=1) with probability θ and tails (x=0) with probability $(1-\theta)$

$$P(x \mid \theta) = \theta^{x}(1-\theta)^{1-x}$$

So to have a run of n coin flips in a dataset $D = \{x_1, x_2, \ldots x_n\}$

$$P(D \mid \theta) = \prod_{k=1}^{n} P(x_k \mid \theta)$$

Take a log to get rid of the product and make it a sum

$$l(D) = \ln(P(D \mid \theta)) = \sum_{k=1}^{n} x_k \ln(\theta) + (1 - x_k)\ln(1 - \theta)$$

$$\frac{\partial l(D)}{\partial D} = \sum_{k=1}^{n} \frac{x_k}{\theta} + \frac{1 - x_k}{1 - \theta}$$

Set $\nabla l(D) = 0$ and solve to find the Maximum Likelihood Estimate

Which it turns out is just the $\mathbb{E}(l(D)) = \frac{\sum_{k=1}^{n} x_k}{n}$

Maximum likelihood estimation

- **Maximum likelihood estimation (MLE):**
  find θ which maximizes likelihood $P(D \mid \theta)$.

$$\begin{aligned}
\theta^* &= \arg\max_\theta P(D|\theta) \\
&= \arg\max_\theta \theta^H (1-\theta)^T \\
&= \frac{H}{H+T}
\end{aligned}$$

H: # of heads

T: # of tails

θ: probability of coin producing heads

Bayesian estimation

Finding the right parameters θ to describe a distribution

- Data D provides evidence for or against our beliefs.
  We update our belief θ based on the evidence we see:

$$\underbrace{P(\theta|D)}_{\text{Posterior}} = \frac{\overbrace{P(\theta)}^{\text{Prior}}\overbrace{P(D|\theta)}^{\text{Likelihood}}}{\underbrace{\int P(\theta)P(D|\theta)d\theta}_{\text{Marginal Likelihood } (=P(D))}}$$

The posterior is proportional to prior x likelihood:

$P(\theta|D) \propto P(\theta)\ P(D|\theta)$

In search of a prior

- If the posterior distribution $P(\theta \mid D)$ is in the same family of distributions as the prior probability distribution $P(\theta)$ the prior and posterior are then said to be *conjugate distributions*
- It makes the math and computation both much easier if we start and end with the same kind of distribution
- All members of the exponential family have conjugate priors

Likelihood	Prior	Posterior
Binomial	Beta	Beta
Negative Binomial	Beta	Beta
Poisson	Gamma	Gamma
Geometric	Beta	Beta
Exponential	Gamma	Gamma
Normal (mean unknown)	Normal	Normal
Normal (variance unknown)	Inverse Gamma	Inverse Gamma
Normal (mean and variance unknown)	Normal/Gamma	Normal/Gamma
Multinomial	Dirichlet	Dirichlet

Likelihood	Prior	Posterior
Binomial	Beta	Beta

As an example of a discrete exponential family, consider the Binomial distribution with known number of trials n. The pmf for this distribution is

$$p(x|\theta) = \text{Binomial}(n, \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, x \in \{0, 1, .., n\} \quad (6)$$

This can equivalently be written as

$$p(x|\theta) = \binom{n}{x}\exp(x\log(\frac{\theta}{1-\theta}) + n\log(1-\theta)) \quad (7)$$

which shows that the Binomial distribution is an exponential

A random variable X $(0 < \mathrm{x} < 1)$ has a Beta distribution with (hyper)parameters α ($\alpha > 0$) and β ($\beta > 0$) if X has a continuous distribution with probability density function

$$P(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

The first term is a normalization factor (to obtain a distribution)

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

Expectation: $\frac{\alpha}{\alpha+\beta}$

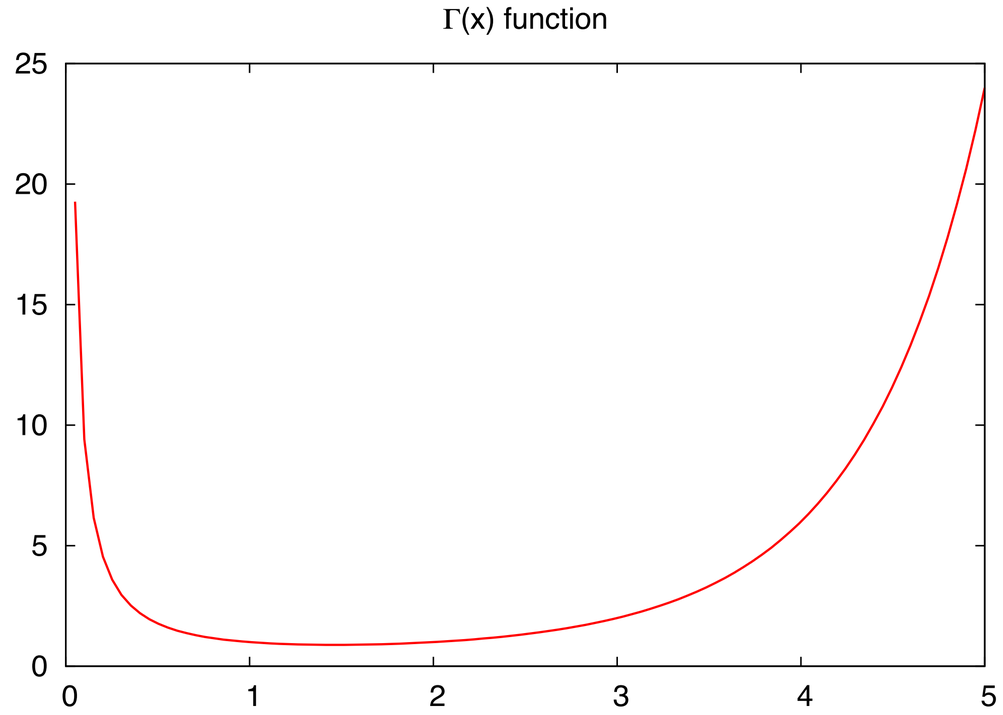
The Gamma function

The Gamma function $\Gamma(x)$ is the generalization of the factorial $x!$ (or rather $(x-1)!$) to the reals:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \quad \text{for } \alpha > 0$$

For $x > 1$, $\Gamma(x) = (x-1)\Gamma(x-1)$.

For positive integers, $\Gamma(x) = (x-1)!$



The Beta distribution

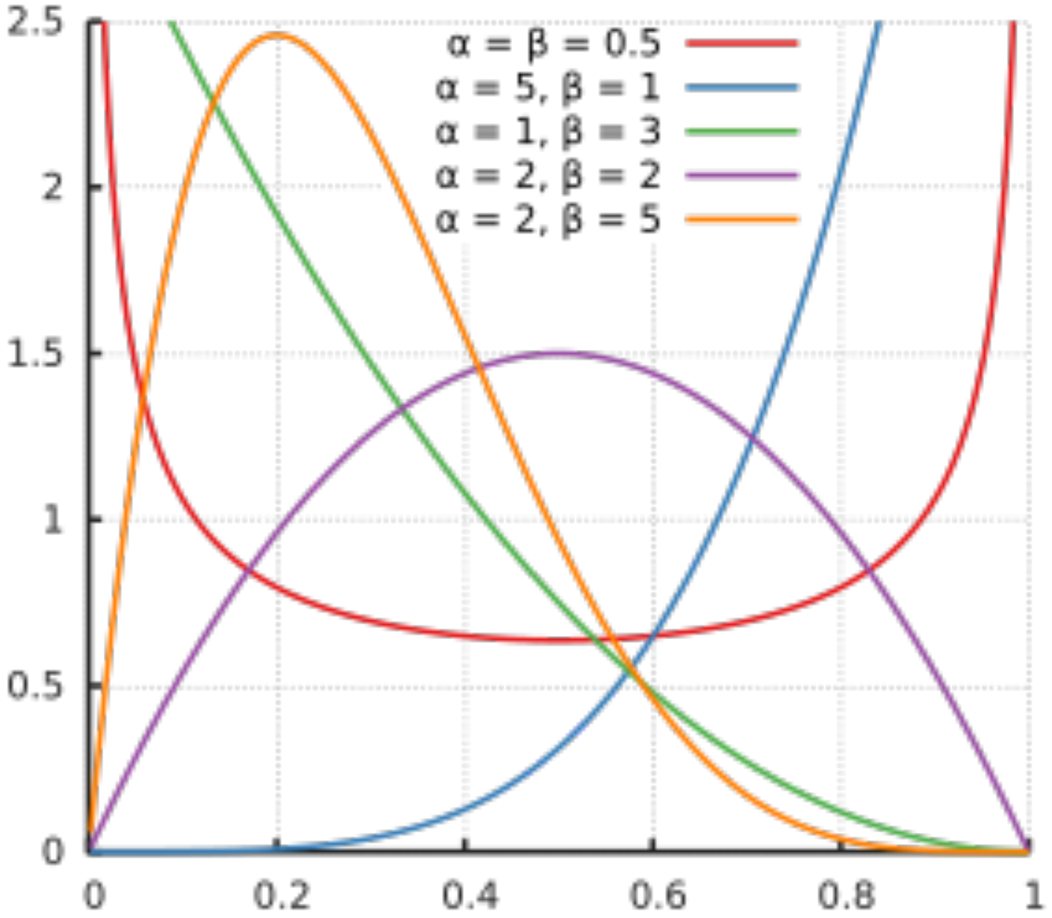
A random variable X $(0 < x < 1)$ has a Beta distribution with (hyper)parameters α $(\alpha > 0)$ and β $(\beta > 0)$ if X has a continuous distribution with probability density function

$$P(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

The first term is a normalization factor (to obtain a distribution)

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

Expectation: $\frac{\alpha}{\alpha+\beta}$



Beta(1,1) => uniform distribution!

The Beta distribution is conjugate to the Binomial distribution.

$$p(\theta|x) = p(x|\theta)p(\theta) = \text{Binomial}(n, \theta) * \text{Beta}(a, b) =$$

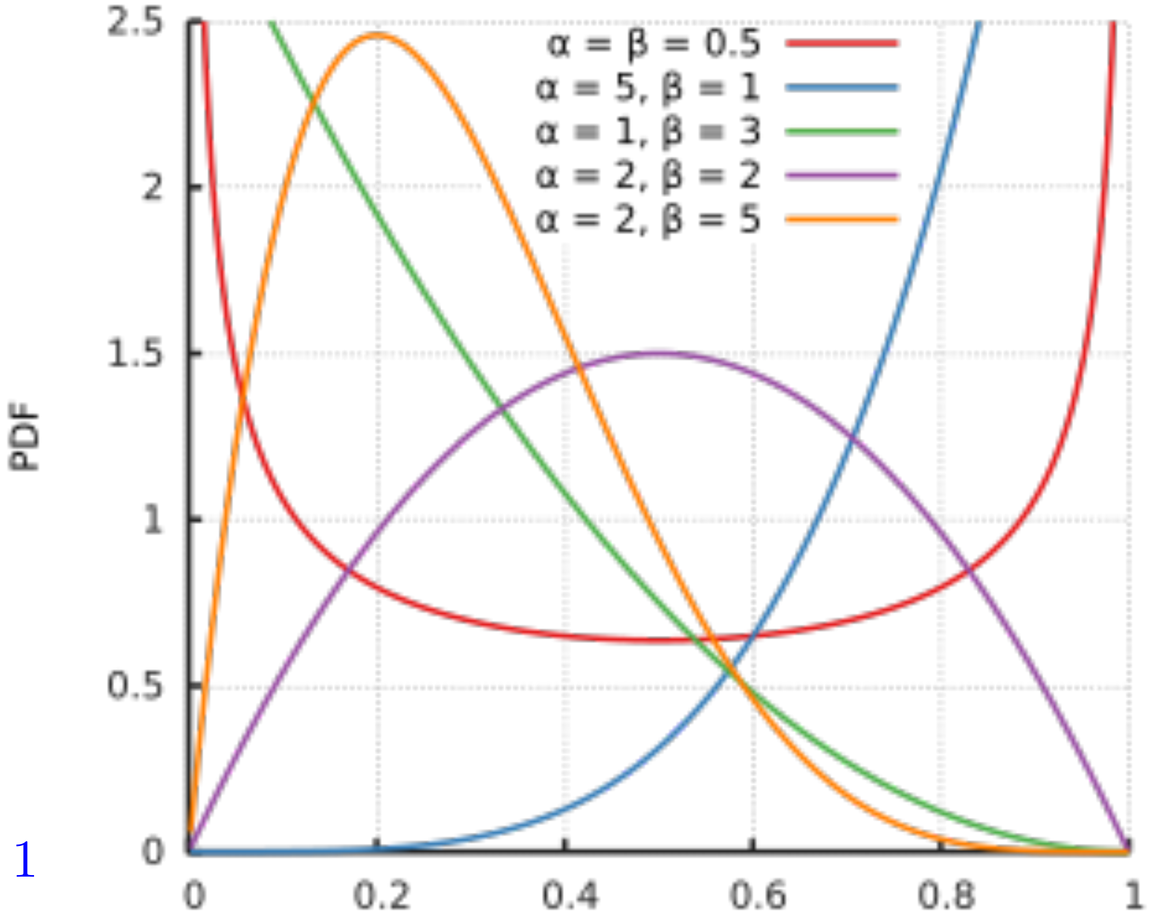
$$\binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{(a-1)}(1-\theta)^{b-1}$$

$$\propto \theta^x(1-\theta)^{n-x}\theta^{(a-1)}(1-\theta)^{b-1}$$

$$p(\theta|x) \propto \theta^{(x+a-1)}(1-\theta)^{n-x+b-1}$$

The posterior distribution is simply a Beta$(x + a, n - x + b)$ distribution.

Effectively, our prior is just adding $a - 1$ successes and $b - 1$ failures to the dataset.



Suppose that the likelihood follows a binomial(N, θ) distribution where N is known and θ is the (unknown) parameter of interest. We also have that the data x from one trial is an integer between 0 and N. Then for a beta prior we have the following table:

hypothesis	data	prior	likelihood	posterior
θ	x	$\text{beta}(a,b)$ $= c_1\theta^{a-1}(1-\theta)^{b-1}$	$\text{binomial}(N,\theta)$ $= c_2\theta^{x}(1-\theta)^{N-x}$	$\text{beta}(a+x,b+N-x)$ $= c_3\theta^{a+x-1}(1-\theta)^{b+N-x-1}$

The table is simplified by writing the normalizing coefficients as c1, c2 and c3 respectively. If needed, we can recover the values of the c1 and c2 by recalling (or looking up) the normalizations of the beta and binomial distributions.

$$c_1 = \frac{(a+b-1)!}{(a-1)!\,(b-1)!} \qquad c_2 = \binom{N}{x} = \frac{N!}{x!\,(N-x)!} \qquad c_3 = \frac{(a+b+N-1)!}{(a+x-1)!\,(b+N-x-1)!}$$

Different priors ==> different posteriors

- For Beta prior - Binomial likelihood estimation our prior is adding a-1 heads and b-1 tails to our posterior
- Beta(1,1) => uniform distribution… this is (one kind of) uninformative prior
- Beta(20,20) => strong preference for a 50/50 coin
- Beta(10,1) => preference for a biased heads coin

Estimation

COGS 118B Winter 2024

Lecture 2

Jason G. Fleischer, PhD
Department of Cognitive Science
University of California San Diego

https://jgfleischer.com
Book a slot in my office hours

$$P(H_i \mid D) = \frac{P(D \mid H_i)P(H_i)}{\sum_{j=1}^{m} P(D \mid H_j)P(H_j)}$$

A very useful form of Bayes rule

"Initial belief plus new evidence equals a new improved belief"

In particular, we are accumulating evidence about multiple hypotheses and comparing them to each other

$P(H_i)$	Prior probability of hypothesis H_i
$P(D \mid H_i)$	Probability of observing data D under H_i <— "Likelihood"
$P(H_i \mid D)$	Probability of H_i given the data <— "Posterior probability"

Estimation

Finding the right parameters θ to describe a distribution

- Example: I flip a coin 3 times and get HHH. What is the probability of heads for this coin?
  - You could say P(H)=1 (frequentist)
    - Maximum Likelihood Estimation (MLE) is what happens when you use scipy.stats.rv_continuous.fit()
  - or you might have a prior belief about the coin and integrate observation and prior together (Bayes)
    - Maximum A Posteriori (MAP) is a method for using the prior to estimate the optimal parameter $\theta^* = \arg\max_\theta P(\theta \mid D)$
    - Bayesian Inference is a method for estimating the entire *distribution of parameters* $P(\theta \mid D)$, integrating over parameters
    - Bayesian Predictive is a method for estimating the next data given previous observations, integrating over parameters
    - You can often calculate MAP / Predictive approaches yourself exactly
    - full Bayesian inference can be computationally intensive, done approximately instead of exactly, and might use specialized tools like PyStan

Estimation

Finding the right parameters θ to describe a distribution

Example: I flip a coin 3 times and get HHH. What is the probability of heads for this coin?

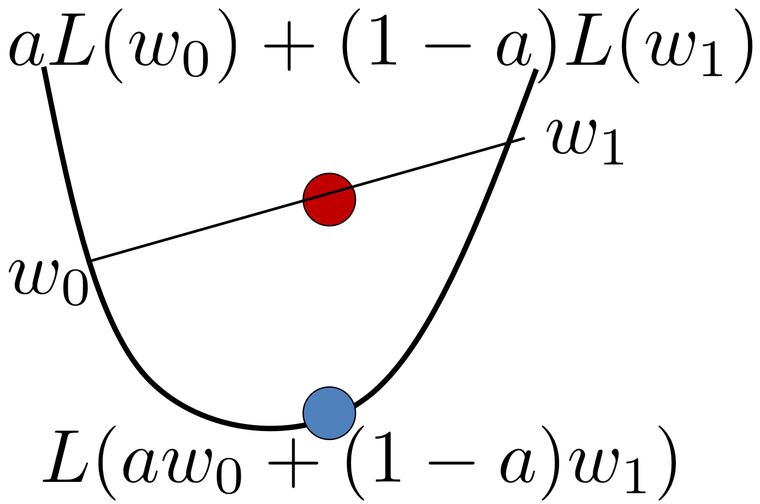
- You could say P(H)=1 (frequentist)
  - Maximum Likelihood Estimation uses only the likelihood term to estimate the parameter $\theta^* = \arg\max_{\theta} P(D \mid \theta)$
  - MLE is what happens when you use scipy.stats.rv_continuous.fit()
- Or you might have a prior belief about the coin and integrate observation and prior together (Bayes)
  - Maximum A Posteriori (MAP) is a method for using both the prior and the likelihood to estimate the optimal parameter $\theta^* = \arg\max_{\theta} P(\theta \mid D)$

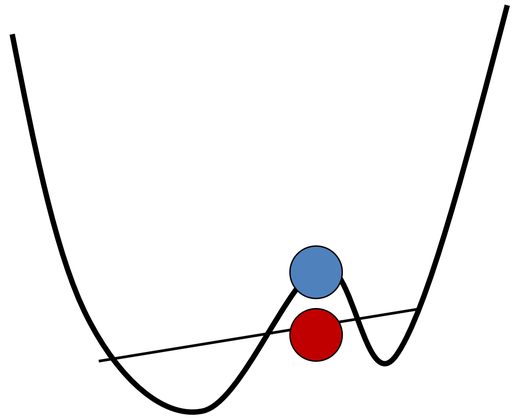
Maximum Likelihood Estimation

Frequentists love likelihood and hate priors

- MLE is (usually) optimal as $\lim n \to \infty$
- Can be very sensitive to overfitting if your model is complicated
- "Find the parameters that make the observed data the most likely"
  $\theta^* = \arg\max_{\theta} P(D \mid \theta)$ (MAP has θ and D reversed!)
- This leads to overfitting because sometimes the data is UNLIKELY given the parameters

Defining convexity with Jensen's inequality





NOT convex!

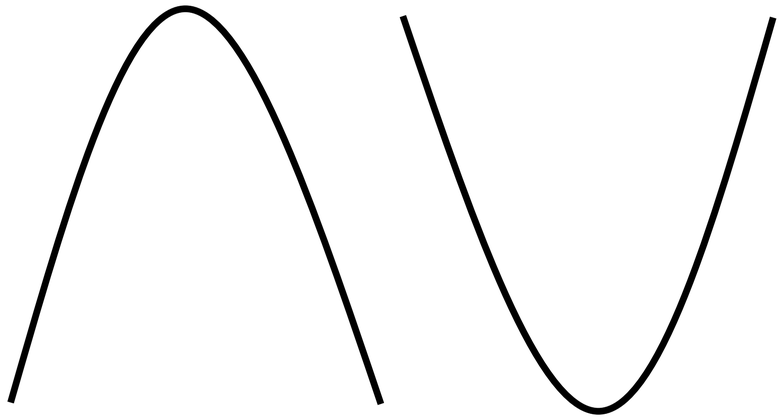
Definition:

$$\forall w_0, w_1, a \in [0, 1]$$

$$aL(w_0) + (1-a)L(w_1) \geq L(aw_0 + (1-a)w_1)$$

Convexity

Is this a convex function?



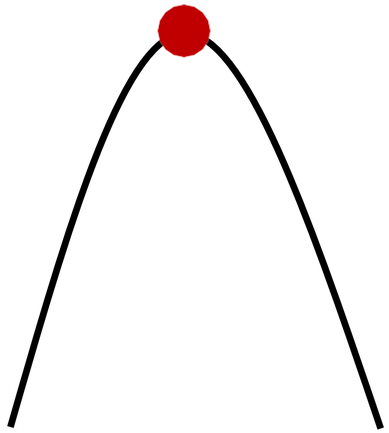
A. Yes

B. No

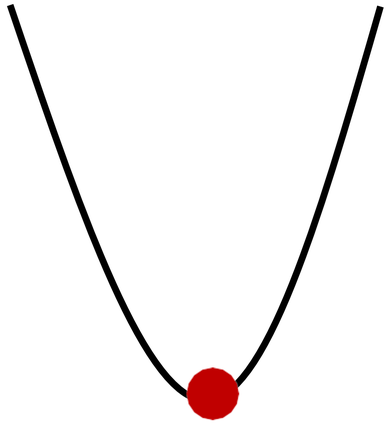
★ C. It depends

It is concave! 

But for a concave function $L(w)$, $-L(w)$ is convex, and vice versa.

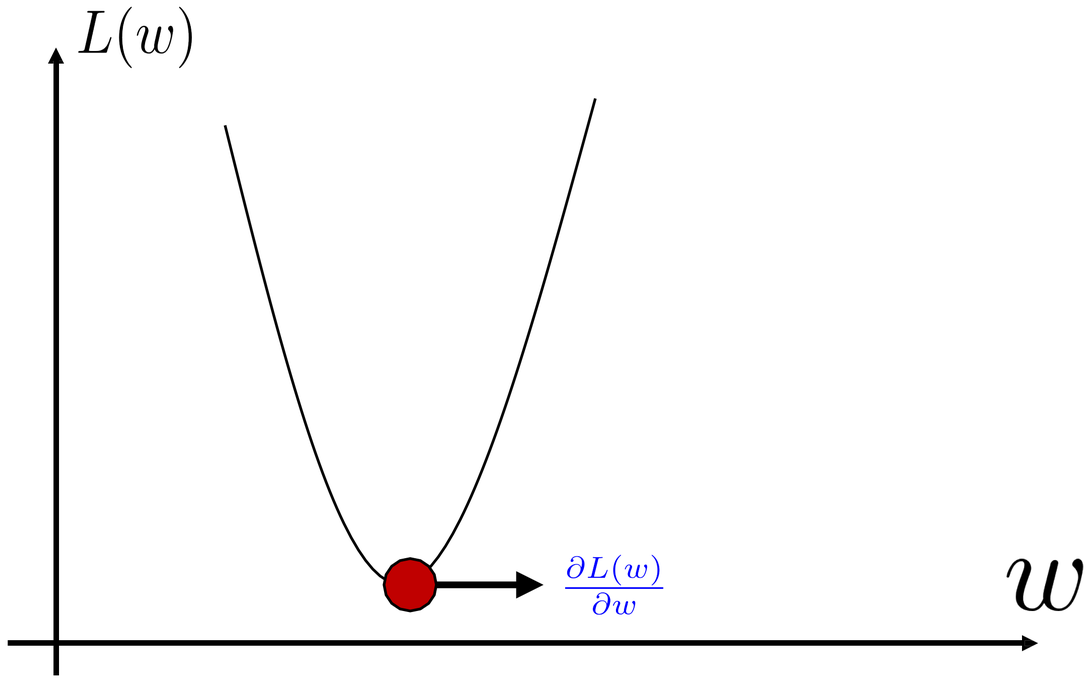


=



$$\arg\max_{\theta} L(\theta) \qquad \arg\min_{\theta} -L(\theta)$$

Convex functions that are differentiable



$$w^* = \arg min_\theta L(w)$$

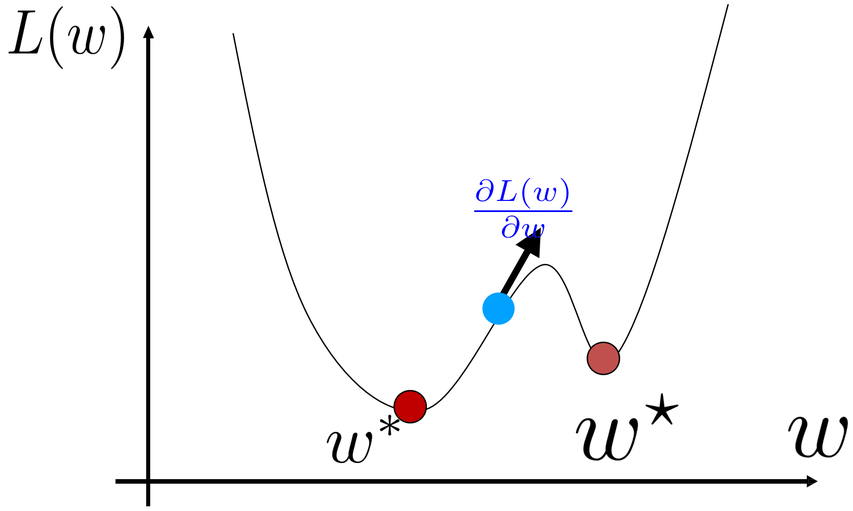
1. (Convex) Function
2. Set Derivative to 0
3. Solve for w

$$L(w) = (w-3)^2 + 4$$

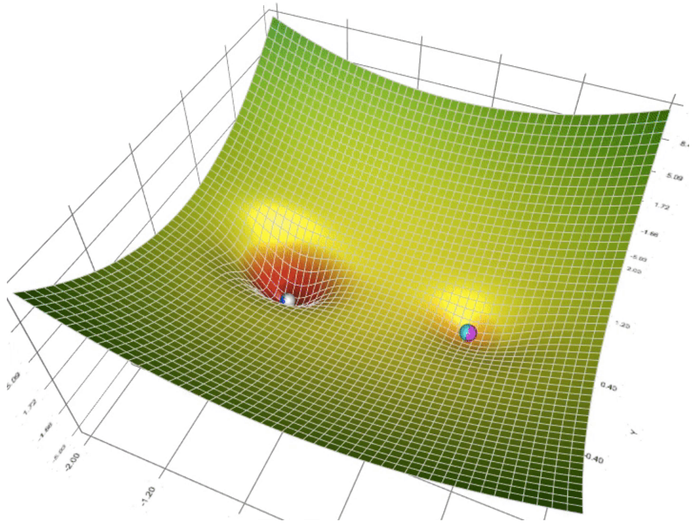
$$\frac{dL(w)}{dw} = 2 \times (w-3) \qquad \frac{dL(w)}{dw} = 0$$

$$2 \times (w-3) = 0 \rightarrow w = 3$$

Non-convex functions that are differentiable



Compute $\frac{\partial L(w)}{\partial w}$ to solve the problem iteratively through gradient descent



This isn't related to today. Just taking an opportunity to educate

Bernoulli distribution is the probability of a single coin flip turning up heads (x=1) with probability θ and tails (x=0) with probability $(1-\theta)$

$$P(x\,|\,\theta) = \theta^x(1-\theta)^{1-x}$$

So to have a run of n coin flips in a dataset $D = \{x_1, x_2, \ldots x_n\}$

$$P(D\,|\,\theta) = \prod_{k=1}^{n} P(x_k\,|\,\theta)$$

Take a log to get rid of the product and make it a sum

$$\begin{aligned} l(D) = \ln(P(D\,|\,\theta)) &= \sum_{k=1}^{n} \ln(\prod P(x_k\,|\,\theta)) \\ &= \sum_{k=1}^{n} \ln(\theta^x) + \sum_{k=1}^{n} \ln((1-\theta)^{1-x}) \\ &= \sum_{k=1}^{n} x_k \ln(\theta) + \sum_{k=1}^{n} (1-x_k)\ln(1-\theta) \end{aligned}$$

$$l(D) = \ln(P(D \mid \theta)) = \sum_{k=1}^{n} x_k \ln(\theta) + \sum_{k=1}^{n} (1 - x_k)\ln(1 - \theta)$$

$$\frac{\partial l(D)}{\partial D} = \sum_{k=1}^{n} \frac{x_k}{\theta} - \sum_{k=1}^{n} \frac{1 - x_k}{1 - \theta}$$ (the negative sign comes from chain rule)

Natural logs are convex and differentiable! So set $\nabla l(D) = 0$ and solve to find the Maximum Likelihood Estimate

$$\frac{\sum_{k=1}^{n} x_k}{\theta} = \frac{\sum_{k=1}^{n} 1 - x_k}{1 - \theta}$$

$$\frac{\sum_{k=1}^{n} x_k}{\theta} = \frac{n - \sum_{k=1}^{n} x_k}{1 - \theta}$$

$$(1 - \theta) \sum_{k=1}^{n} x_k = \theta(n - \sum_{k=1}^{n} x_k)$$

$(1 - \theta) \sum_{k=1}^{n} x_k + \theta \sum_{k=1}^{n} x_k = \theta n$ which simplifies to $\theta = \frac{\sum_{k=1}^{n} x_k}{n}$ or the expected value of the distribution

Examples

MLE estimation

- It’s just “calculate the empirical mean” but we showed you the proof
- HHH -> Bern: $\theta = 1$
- HHHTH -> Bern: $\theta = \frac{4}{5} = 0.8$

Maximum A Posteriori

Finding the right parameters θ to describe a distribution

$$P(H_i \mid D) = \frac{P(D \mid H_i)P(H_i)}{\sum_{j=1}^{m} P(D \mid H_j)P(H_j)}$$

A very useful form of Bayes rule

$$P(\theta_i \mid D) \propto P(D \mid \theta_i)P(\theta_i)$$

Our hypothesis is a particular set of parameters. And we can get rid of the marginal denominator (a constant) and talk about proportional to instead of equals

$$\theta^* = \arg\max_{\theta} P(\theta \mid D)$$

MLE has θ and D reversed because it cares about likelihood, MAP cares about the posterior

Completely off topic

This is why NHST is at best an approximation

$$P(H_i \mid D) \propto P(D \mid H_i)P(H_i)$$

P-values for any statistical test are $P(D \mid H_0)$, but we pretend like they are $P(H_0 \mid D)$. Frequentist approaches use likelihood and Bayesian approaches use the posterior.

They are identical only under a limited number of conditions. In fact p-values are often an order of magnitude larger than the posterior

Likelihood	Prior	Posterior
Binomial	Beta	Beta

As an example of a discrete exponential family, consider the Binomial distribution with known number of trials n. The pmf for this distribution is

$$p(x|\theta) = \text{Binomial}(n, \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, x \in \{0, 1, .., n\} \quad (6)$$

This can equivalently be written as

$$p(x|\theta) = \binom{n}{x}\exp(x\log(\frac{\theta}{1-\theta}) + n\log(1-\theta)) \quad (7)$$

which shows that the Binomial distribution is an exponential

Coin flipping example: MAP

Our model is:

$$X \sim \text{Ber}(\theta),$$

If we observe 'HHH' then our posterior distribution is

$$\begin{aligned} p(\theta|HHH) &= \frac{p(HHH|\theta)p(\theta)}{p(HHH)} && \text{(Bayes' rule)} \\ &\propto p(HHH|\theta)p(\theta) && (p(HHH) \text{ is constant}) \\ &= \theta^3(1-\theta)^0 p(\theta) && \text{(likelihood def'n)} \end{aligned}$$

But what's the prior?

Coin flipping example: MAP

Our model is:

$$X \sim \text{Ber}(\theta), \quad \theta \sim \text{Beta}\ (a, b).$$

If we observe 'HHH' then our posterior distribution is

$$\begin{aligned}
p(\theta|HHH) &= \frac{p(HHH|\theta)p(\theta)}{p(HHH)} && \text{(Bayes' rule)} \\
&\propto p(HHH|\theta)p(\theta) && (p(HHH) \text{ is constant}) \\
&= \theta^3(1-\theta)^0 p(\theta) && \text{(likelihood def'n)} \\
&= \theta^3(1-\theta)^0\theta^{a-1}(1-\theta)^{b-1} && \text{(prior def'n)} \\
&= \theta^{(3+a)-1}(1-\theta)^{b-1}.
\end{aligned}$$

Coin flipping example: MAP

Our model is:

$$X \sim \text{Ber}(\theta), \quad \theta \sim \text{Beta}\ (a, b).$$

If we observe 'HHH' then our posterior distribution is

$$p(\theta|HHH) = \frac{p(HHH|\theta)p(\theta)}{p(HHH)} \qquad \text{(Bayes' rule)}$$

$$\propto \theta^{(3+a)-1}(1-\theta)^{b-1}$$

A beta prior + Bernoulli likelihood = a beta posterior. The posterior is effectively adding together two samples: the empirical one and a fake "prior" sample.

This is a case where we can simply solve analytically without computing the integrals!

Examples

MAP estimation

- $P(\theta \mid D) \propto \theta^{(H+a-1)}(1-\theta)^{(T+b-1)}$ where H is empirical number of heads and T is number of tails
- If you take the derivative of this and set it equal to zero you find $\hat{\theta} = \frac{H+a-1}{H+a+T+b-2}$ (left as an exercise!)
- HHH + Beta(1,1) prior —> Bern: $\theta = 1$ **(UNINFORMATIVE PRIORS = MLE)**
- HHH + Beta(20,20) prior —> ????