10-701: Introduction to Machine Learning Lecture 5 – MLE & MAP

Pradeep Ravikumar

Spring 2026

Front Matter

- Announcements:
  - 30 minute quiz this Friday at 2pm in Baker Hall A36 (our normal recitation time/place)
  - please submit your AWS credit request for the project, due Thursday, Jan 29 at 11:59pm
- Recommended Readings:
  - Mitchell, Estimating Probabilities

Your first consulting job

- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have a coin, if I flip it, what's the probability it will fall with the head up?
  - You say: Please flip it a few times:

Your first consulting job

- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have a coin, if I flip it, what's the probability it will fall with the head up?
  - You say: Please flip it a few times:

  - You say: The probability is: **3/5**
  - **He says: Why???**
  - You say: Because… frequency of heads in all flips

Questions

- Why frequency of heads?

- How good is this estimation?
  - Would you be willing to bet money on your guess of the probability?
  - Why not?

Model

- First we need a model that would capture the experimental data
- What is the experimental data?
- Coin Flips

Model

- First we need a model that would capture the experimental data
- What is the experimental data?
- Coin Flips

Model

- A model for coin flips
  - Bernoulli Distribution
- X is a random variable with Bernoulli distribution when:
  - X takes values in {0,1}
  - P(X = 1) = p
  - P(X = 0) = 1 - p
  - Where p in [0,1]

Model

- X is a random variable with Bernoulli distribution when:
  - X takes values in {0,1}
  - P(X = 1) = p
  - P(X = 0) = 1 - p
  - Where p in [0,1]
- X = 1 i.e. heads with probability p, and X = 0 i.e. tails with probability 1 – p
  - Coin with probability of flipping heads = p
- And we draw **independent** samples that are **identically distributed** from same distribution
  - flip the same coin multiple times

Bernoulli distribution

Data, D = 

- P(Heads) = θ, P(Tails) = 1-θ
- Flips are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Bernoulli distribution

Choose θ that maximizes the probability of observed data

Probability of one coin flip

Let's say we observe a coin flip $X \in \{0, 1\}$.

The probability of this coin flip,
given a Bernoulli distribution with parameter p:

$$p^X (1 - p)^{1-X}.$$

Equal to p when $X = 1$, and equal to $(1 - p)$ when $X = 0$.

Probability of Multiple Coin Flips

$$\text{Probability of Data} = \mathbb{P}(X_1, X_2, \ldots, X_n; \theta)$$

Probability of Multiple Coin Flips

$$\begin{aligned}\text{Probability of Data} &= \mathbb{P}(X_1, X_2, \ldots, X_n; \theta) \\ &= P(X_1)\, P(X_2) \ldots P(X_n)\end{aligned}$$

…Independence of samples

Probability of Multiple Coin Flips

$$\begin{aligned}\text{Probability of Data} &= \mathbb{P}(X_1, X_2, \ldots, X_n; \theta) \\ &= P(X_1)\, P(X_2) \ldots P(X_n) \\ &= \prod_{i=1}^{n} P(X_i)\end{aligned}$$

Probability of Multiple Coin Flips

$$\begin{aligned}\text{Probability of Data} &= \mathbb{P}(X_1, X_2, \ldots, X_n; \theta) \\ &= P(X_1)\, P(X_2) \ldots P(X_n) \\ &= \prod_{i=1}^{n} P(X_i) \\ &= \prod_{i=1}^{n} p^{X_i}\, (1-p)^{1-X_i}\end{aligned}$$

…probability of a Bernoulli sample

Probability of Multiple Coin Flips

$$
\begin{aligned}
\text{Probability of Data} &= \mathbb{P}(X_1, X_2, \ldots, X_n; \theta) \\
&= P(X_1)\, P(X_2) \ldots P(X_n) \\
&= \prod_{i=1}^{n} P(X_i) \\
&= \prod_{i=1}^{n} p^{X_i}\, (1-p)^{1-X_i} \\
&= p^{\sum_{i=1}^{n} X_i}\, (1-p)^{n-\sum_{i=1}^{n} X_i}
\end{aligned}
$$

$$\ldots\ p^a\, p^b = p^{a+b}$$

Probability of Multiple Coin Flips

$$
\begin{aligned}
\text{Probability of Data} &= \mathbb{P}(X_1, X_2, \ldots, X_n; \theta) \\
&= P(X_1)\, P(X_2) \ldots P(X_n) \\
&= \prod_{i=1}^{n} P(X_i) \\
&= \prod_{i=1}^{n} p^{X_i} \left(1-p\right)^{1-X_i} \\
&= p^{\sum_{i=1}^{n} X_i} \left(1-p\right)^{n-\sum_{i=1}^{n} X_i} \\
&= p^{n_h} \left(1-p\right)^{n-n_h}.
\end{aligned}
$$

where n_h is the number of heads,
n is the total number of coin flips

Maximum Likelihood Estimator (MLE)

The MLE solution is then given by solving the following problem:

$$\begin{aligned}\widehat{p} &= \arg\max_{p} \mathbb{P}(X_1, \ldots, X_n; p) \\ &= \arg\max_{p} \left\{ p^{n_h} \left(1-p\right)^{n-n_h} \right\}\end{aligned}$$

Maximum Likelihood Estimator (MLE)

The MLE solution is then given by solving the following problem:

$$\begin{aligned}\widehat{p} &= \arg\max_p \mathbb{P}(X_1, \ldots, X_n; p) \\ &= \arg\max_p \left\{ p^{n_h} \left(1-p\right)^{n-n_h} \right\} \\ &= \arg\max_p \left\{ n_h \log p + \left(n - n_h\right) \log(1-p) \right\}\end{aligned}$$

…argmax_x f(x) = argmax_x log f(x)

MLE for coin flips

The MLE solution is then given by solving the following problem:

$$\widehat{p} = \arg\max_{p} \{n_h \log p + (n - n_h) \log(1 - p)\}$$

$$\implies \frac{n_h}{\widehat{p}} - \frac{n - n_h}{1 - \widehat{p}} = 0$$

$$\implies \widehat{p} = \frac{n_h}{n}.$$

Maximum Likelihood Estimation

Choose θ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

MLE of probability of head:

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \quad \text{= 3/5}$$

"Frequency of heads"

How many flips do I need?

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: θ = 3/5, it is the MLE!
- He says: What if I flipped 30 heads and 20 tails?
- You say: Same answer, it is the MLE!
- **He says: If you get the same answer, would you prefer to flip 5 times or 50 times?**
- You say: Hmm… The more the merrier???
- He says: Is this why I am paying you the big bucks???

KEY QUESTION: HOW GOOD IS THE MLE (OR ANY OTHER ESTIMATOR)?

How good is this MLE?

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

How good is this MLE?

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

It would be great if it would then be equal to the "true" coin flip probability p.

How good is this MLE?

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

It would be great if it would then be equal to the "true" coin flip probability p.

More formally: as we flip more and more times, we want our estimator to converge (in probability) to the true coin flip probability.

How good is this MLE?

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

It would be great if it would then be equal to the "true" coin flip probability p.

More formally: as we flip more and more times, we want our estimator to converge (in probability) to the true coin flip probability.

This property is known as **consistency**.

How good is this MLE?

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

It would be great if it would then be equal to the "true" coin flip probability p.

More formally: as we flip more and more times, we want our estimator to converge (in probability) to the true coin flip probability.

This property is known as **consistency**.

Do we get that $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \to p$ in probability as $n \to \infty$?

How good is this MLE?

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

It would be great if it would then be equal to the "true" coin flip probability p.

More formally: as we flip more and more times, we want our estimator to converge (in probability) to the true coin flip probability.

This property is known as **consistency**.

Do we get that $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \to p$ in probability as $n \to \infty$?

By the Law of Large Numbers!

How good is this MLE?

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

It would be great if it would then be equal to the "true" coin flip probability p.

More formally: as we flip more and more times, we want our estimator to converge (in probability) to the true coin flip probability.

This property is known as **consistency**.

Do we get that $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \to p$ in probability as $n \to \infty$?

By the Law of Large Numbers!

…since the sample mean converges to E(X) = p

How good is this MLE?

If we repeated this experiment infinitely many times, i.e. flip a coin n times and calculate our estimator, and then took an average of our estimator over the infinitely many trials.

What would the average look like?

How good is this MLE?

If we repeated this experiment infinitely many times, i.e. flip a coin n times and calculate our estimator, and then took an average of our estimator over the infinitely many trials.

What would the average look like?

Formally: the estimator $\widehat{p}$ is random: it depends on the samples (i.e. coin flips) drawn from a Bernoulli distribution with parameter p.

What would the expectation of the estimator be?

How good is this MLE?

If we repeated this experiment infinitely many times, i.e. flip a coin n times and calculate our estimator, and then took an average of our estimator over the infinitely many trials.

What would the average look like?

Formally: the estimator $\widehat{p}$ is random: it depends on the samples (i.e. coin flips) drawn from a Bernoulli distribution with parameter p.

What would the expectation of the estimator be?

It would be great if this expectation be equal to the "true" coin flip probability.

How good is this MLE?

If we repeated this experiment infinitely many times, i.e. flip a coin n times and calculate our estimator, and then took an average of our estimator over the infinitely many trials.

What would the average look like?

Formally: the estimator $\widehat{p}$ is random: it depends on the samples (i.e. coin flips) drawn from a Bernoulli distribution with parameter p.

What would the expectation of the estimator be?

It would be great if this expectation be equal to the "true" coin flip probability.

This property is called **unbiasedness**.

How good is this MLE?

It would be great if this expectation be equal to the "true" coin flip probability.

This property is called **unbiasedness**.

$$\begin{aligned}\mathbb{E}(\widehat{p}) &= \mathbb{E}\left(\frac{n_h}{n}\right) \\ &= \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)\end{aligned}$$

How good is this MLE?

It would be great if this expectation be equal to the "true" coin flip probability.

This property is called **unbiasedness**.

$$\begin{aligned}\mathbb{E}(\widehat{p}) &= \mathbb{E}\left(\frac{n_h}{n}\right) \\ &= \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i)\end{aligned}$$

…linearity of expectation:
E(a X + b Y) = a E(X) + b E(Y)

How good is this MLE?

It would be great if this expectation be equal to the "true" coin flip probability.

This property is called **unbiasedness**.

$$\begin{aligned}
\mathbb{E}(\widehat{p}) &= \mathbb{E}\left(\frac{n_h}{n}\right) \\
&= \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) \\
&= \mathbb{E}X_1
\end{aligned}$$

How good is this MLE?

It would be great if this expectation be equal to the "true" coin flip probability.

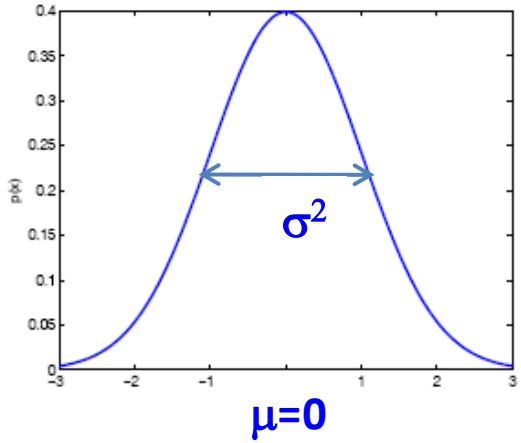
This property is called **unbiasedness**.

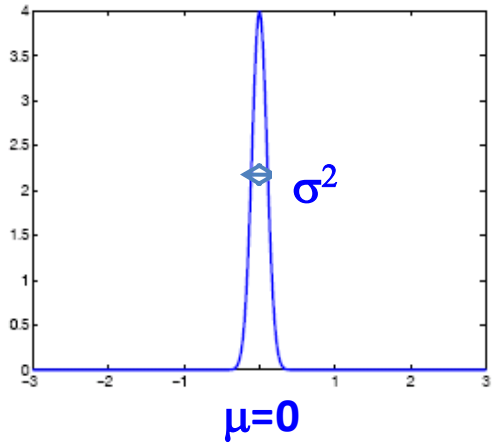
$$\begin{aligned}
\mathbb{E}(\widehat{p}) &= \mathbb{E}\left(\frac{n_h}{n}\right) \\
&= \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) \\
&= \mathbb{E}X_1 \\
&= p.
\end{aligned}$$

What about continuous variables?

- Billionaire says: If I am measuring a continuous variable, what can you do for me?
- **You say: Let me tell you about Gaussians…**

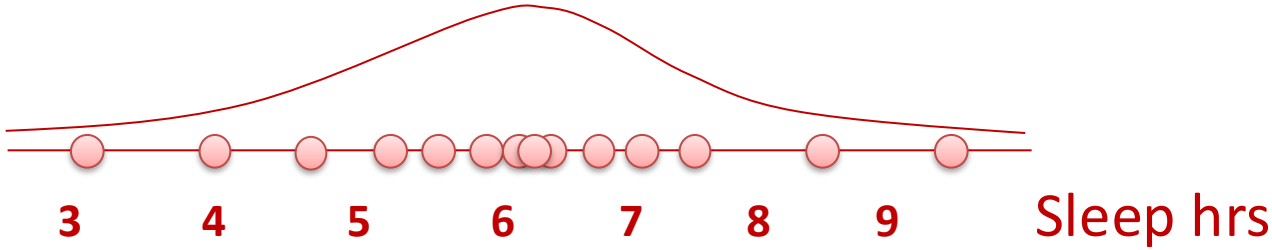
$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$ = N(μ,σ²)





Gaussian distribution

Data, D =



- Parameters: μ – mean, σ^2 - variance
- Sleep hrs are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Gaussian distribution

MLE for Gaussian mean and variance

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

Note: MLE for the variance of a Gaussian is **biased**

- Expected result of estimation is **not** true parameter!
- Unbiased variance estimator:

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

MLE for parametric models

Data: $X_1, X_2, \ldots, X_n$.

Model: $P(X; \theta)$ with parameters θ.

MLE for parametric models

Data: $X_1, X_2, \ldots, X_n$.

Model: $P(X; \theta)$ with parameters θ.

Assumption: Data drawn $i.i.d$ from distribution $P(X; \theta^*)$ for some unknown θ^*.

MLE for parametric models

Data: $X_1, X_2, \ldots, X_n$.

Model: $P(X; \theta)$ with parameters θ.

Assumption: Data drawn $i.i.d$ from distribution $P(X; \theta^*)$ for some unknown θ^*.

Mission (should you choose to accept it): recover θ^* from data $X_1, X_2, \ldots, X_n$.

MLE for parametric models

Data: $X_1, X_2, \ldots, X_n$.

Model: $P(X; \theta)$ with parameters θ.

Assumption: Data drawn $i.i.d$ from distribution $P(X; \theta^*)$ for some unknown θ^*.

Mission (should you choose to accept it): recover θ^* from data $X_1, X_2, \ldots, X_n$.

R. A. Fisher

MLE for parametric models

Data: $X_1, X_2, \ldots, X_n$.

Model: $P(X; \theta)$ with parameters θ.

Assumption: Data drawn $i.i.d$ from distribution $P(X; \theta^*)$ for some unknown θ^*.

Mission (should you choose to accept it): recover θ^* from data $X_1, X_2, \ldots, X_n$.

Likelihood Function: $L(\theta) := \prod_{i=1}^{n} P(X_i; \theta)$
The probability of seeing data $X_1, X_2, \ldots, X_n$ assuming parameters were θ.

MLE for parametric models

Data: $X_1, X_2, \ldots, X_n$.

Model: $P(X; \theta)$ with parameters θ.

Assumption: Data drawn $i.i.d$ from distribution $P(X; \theta^*)$ for some unknown θ^*.

Mission (should you choose to accept it): recover θ^* from data $X_1, X_2, \ldots, X_n$.

Likelihood Function: $L(\theta) := \prod_{i=1}^{n} P(X_i; \theta)$
The probability of seeing data $X_1, X_2, \ldots, X_n$ assuming parameters were θ.

Maximum Likelihood Estimator (MLE): find that parameter θ that would maximize the likelihood of θ

MLE for parametric models

Data: $X_1, X_2, \ldots, X_n$.

Model: $P(X; \theta)$ with parameters θ.

Assumption: Data drawn $i.i.d$ from distribution $P(X; \theta^*)$ for some unknown θ^*.

Mission (should you choose to accept it): recover θ^* from data $X_1, X_2, \ldots, X_n$.

Likelihood Function: $L(\theta) := \prod_{i=1}^{n} P(X_i; \theta)$
The probability of seeing data $X_1, X_2, \ldots, X_n$ assuming parameters were θ.

Maximum Likelihood Estimator (MLE): find that parameter θ that would maximize the likelihood of θ

i.e. pick the θ that would maximize the probability of having seen the data that we do see

Unbiasedness

An estimator $\widehat{\theta}(X_1, \ldots, X_n)$ where $X_i \sim P(X; \theta^*)$ is unbiased if

$$\mathbb{E}(\widehat{\theta}) = \theta^*.$$

MLE is ”asymptotically” unbiased i.e. there are some error terms that go to zero as a function of n, the number of samples

Consistency

An estimator $\widehat{\theta}(X_1, \ldots, X_n)$ where $X_i \sim P(X; \theta^*)$ is consistent if $\widehat{\theta} \to \theta^*$ in probability as $n \to \infty$.

MLE is consistent under some mild regularity conditions on the model, and when the model size is fixed.

How many flips?

- But recall the Billionaire's question:
  - How many flips would you prefer: 5 or 50?
  - How many flips would you need to be willing to bet money on your answer?
- Unbiasedness and Consistency do not answer this question
- We need convergence rates for our estimator
  - We will study more in Learning Theory lectures

MLE

- Well-studied question in Statistics
  - Guarantees (consistency, unbiasedness, rates)
- What has Machine Learning contributed to this statistical question:
  - Specific kinds of guarantees e.g. sample complexity
  - New tools to derive guarantees (VC Dimension, etc.)
  - Computational Issues

Computational Issues

- When number of parameters, or number of samples n is large, computing the MLE is a **large-scale optimization problem**
- Well-studied problem in optimization/operations research
- Machine Learning has contributed considerably via:
  - Better understanding of optimization problems that arise from statistical estimators such as MLE (in contrast to general optimization problems)

Recall: Your first consulting job

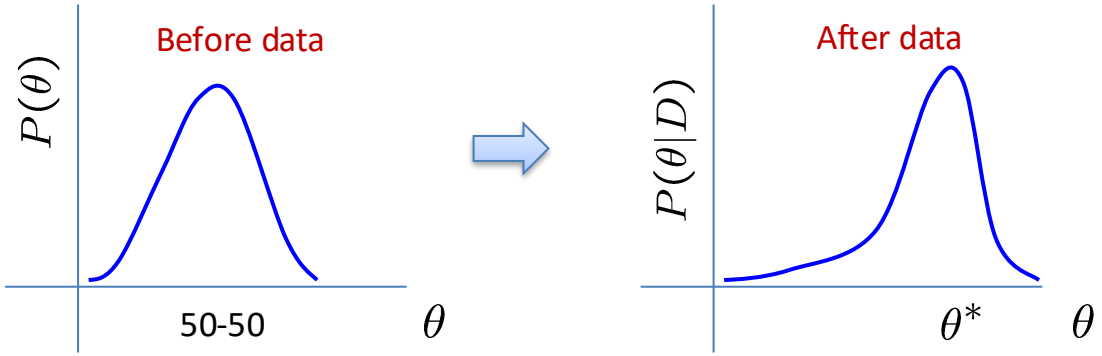
- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have a coin, if I flip it, what's the probability it will fall with the head up?
  - You say: Pl

  - You say: The probability is: 3/5 because… frequency of heads in all flips
  - **He says: But can I put money on this estimate?**
  - You say: ummm…. Maybe not.
    - Not enough flips (less than sample complexity)

What about prior knowledge?

- Billionaire says: Wait, I know that the coin is "close" to 50-50. What can you do for me now?
- **You say: I can learn it the Bayesian way…**

- Rather than estimating a single θ, we obtain a distribution over possible values of θ



Bayesian Learning

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

likelihood prior

Parameters

Data

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

Bayesian Learning

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

posterior likelihood prior

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

AIDS test (Bayes rule)

Data

- **Approximately 0.1% are infected**
- **Test detects all infections**
- **Test reports positive for 1% healthy people**

AIDS test (Bayes rule)

Data

- **Approximately 0.1% are infected**
- **Test detects all infections**
- **Test reports positive for 1% healthy people**

Probability of having AIDS if test is positive:

$$P(a=1|t=1) = \frac{P(t=1|a=1)P(a=1)}{P(t=1)}$$

$$= \frac{P(t=1|a=1)P(a=1)}{P(t=1|a=1)P(a=1) + P(t=1|a=0)P(a=0)}$$

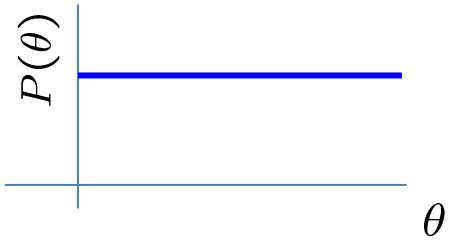
$$= \frac{1 \cdot 0.001}{1 \cdot 0.001 + 0.01 \cdot 0.999} = 0.091$$

Only 9%!...

Prior distribution

- From where do we get the prior?
  - Represents expert knowledge (philosophical approach)
  - Simple posterior form (engineer's approach)

- Uninformative priors:
  - Uniform distribution



- Conjugate priors:
  - Closed-form representation of posterior
  - $P(\theta)$ and $P(\theta|D)$ have the same algebraic form as a function of θ

Conjugate Prior

- P(θ) and P(θ|D) have the same form as a function of theta

Eg. 1 Coin flip problem

Likelihood given Bernoulli model:

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

If prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

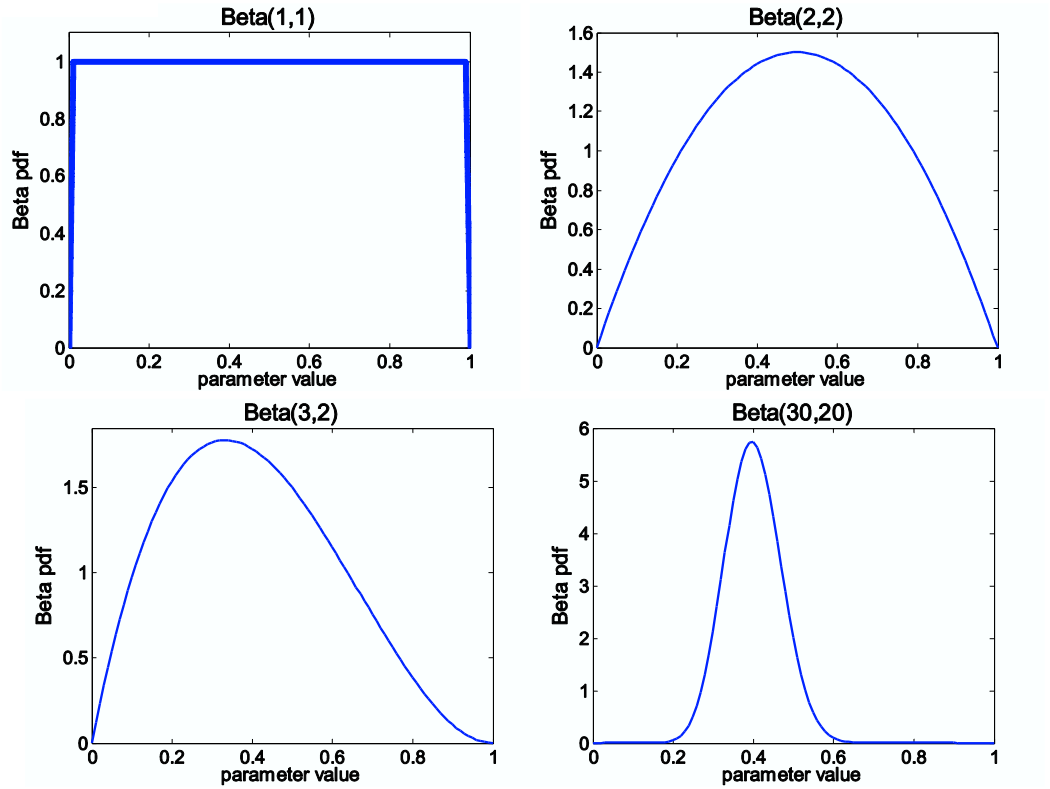
Then posterior is Beta distribution

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

For Binomial, conjugate prior is Beta distribution.

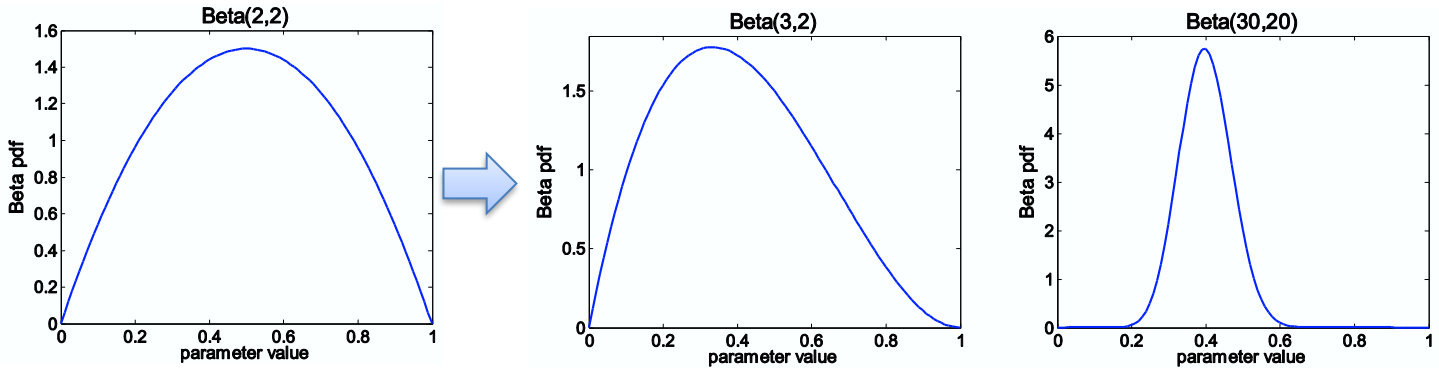
Beta distribution

$Beta(\beta_H, \beta_T)$ More concentrated as values of β_H, β_T increase



Beta conjugate prior

$$P(\theta) \sim Beta(\beta_H, \beta_T) \qquad P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



As $n = \alpha_H + \alpha_T$ increases

As we get more samples, effect of prior is "washed out"

Conjugate Prior

- P(θ) and P(θ|D) have the same form

Eg. 2 Dice roll problem (6 outcomes instead of 2)

Likelihood is ~ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

$$P(\mathcal{D} \mid \theta) = \theta_1^{\alpha_1}\theta_2^{\alpha_2} \ldots \theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\prod_{i=1}^{k} \theta_i^{\beta_i - 1}}{B(\beta_1, \ldots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_k)$$

Then posterior is Dirichlet distribution

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

For Multinomial, conjugate prior is Dirichlet distribution.

Posterior Distribution

- The approach seen so far is what is known as a **Bayesian** approach
- Prior information encoded as a **distribution** over possible values of parameter
- Using the Bayes rule, you get an updated **posterior** distribution over parameters, which you provide with flourish to the Billionaire
- But the billionaire is not impressed
  - Distribution? I just asked for one number: is it 3/5, 1/2, what is it?
  - How do we go from a distribution over parameters, to a single estimate of the true parameters?

Maximum A Posteriori Estimation

Choose θ that maximizes a posterior probability

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max_{\theta} P(\theta \mid D) \\ &= \arg\max_{\theta} P(D \mid \theta) P(\theta)\end{aligned}$$

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

MAP estimate of probability of head:

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

Mode of Beta distribution

MLE vs. MAP

- Maximum Likelihood estimation (MLE)

  Choose value that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

- Maximum *a posteriori* (MAP) estimation

  Choose value that is most probable given observed data and prior belief

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max_{\theta} P(\theta|D) \\ &= \arg\max_{\theta} P(D|\theta)P(\theta)\end{aligned}$$

When is MAP same as MLE?

MLE vs. MAP

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

What if we toss the coin too few times?

- You say: Probability next toss is a head = 0
- Billionaire says: You're fired! …with prob 1 ☺

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

- Beta prior equivalent to extra coin flips
- As *n* → infty, prior is "forgotten"
- **But, for small sample size, prior is important!**

MLE vs MAP


You are no good when sample is small
You give a different answer for different priors

MAP for Gaussian mean and variance

- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Wishart Distribution

- Prior for mean:

$$P(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{\frac{-(\mu-\eta)^2}{2\lambda^2}}$$ = $N(\eta,\lambda^2)$

MAP for Gaussian Mean

$$\widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\mu}_{MAP} = \frac{\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i + \frac{\eta}{\lambda^2}}{\frac{n}{\sigma^2} + \frac{1}{\lambda^2}}$$

Prior Information

- In the Bayesian approach, the prior information is encoded through a prior distribution over the parameters
- Seems onerous: the distribution typically seems to be obtained from convenience (conjugate distribution)
- What other ways can we encode our prior knowledge about the parameters?
- A non-Bayesian approach is via constraints

Encoding prior information via constraints

MLE:

$$\max_{\theta} \log \mathbb{P}(D; \theta).$$

Encoding prior information via constraints

MLE:

$$\max_{\theta} \log \mathbb{P}(D; \theta).$$

Constrained MLE:

$$\max_{\theta} \log \mathbb{P}(D; \theta)$$
$$\text{s.t.} \, \mathcal{R}(\theta) <= C.$$

Encoding prior information via constraints

MLE:

$$\max_{\theta} \log \mathbb{P}(D; \theta).$$

Constrained MLE:

$$\max_{\theta} \log \mathbb{P}(D; \theta)$$
$$\text{s.t.} \, \mathcal{R}(\theta) <= C.$$

When $\mathcal{R}(\theta)$ is convex, constrained MLE is equivalent to regularized MLE:

$$\max_{\theta} \left\{ \log \mathbb{P}(D; \theta) + \lambda \, \mathcal{R}(\theta) \right\}.$$

Regularized MLE

Regularized MLE:

$$\max_{\theta} \left\{ \log \mathbb{P}(D; \theta) + \lambda \mathcal{R}(\theta) \right\}.$$

Trades off maximizing the log-likelihood (i.e. fit to data), against the "prior" constraints encoded by regularization (which do not involve the data at all).

Regularized MLE

Regularized MLE:

$$\max_{\theta} \left\{ \log \mathbb{P}(D; \theta) + \lambda \mathcal{R}(\theta) \right\}.$$

Trades off maximizing the log-likelihood (i.e. fit to data), against the "prior" constraints encoded by regularization (which do not involve the data at all).

The MAP estimator can be seen to be a special case by simply setting

$$\lambda \mathcal{R}(\theta) = \log P(\theta).$$

Regularized MLE

Regularized MLE:

$$\max_{\theta} \left\{ \log \mathbb{P}(D; \theta) + \lambda \mathcal{R}(\theta) \right\}.$$

Trades off maximizing the log-likelihood (i.e. fit to data), against the "prior" constraints encoded by regularization (which do not involve the data at all).

The MAP estimator can be seen to be a special case by simply setting

$$\lambda \mathcal{R}(\theta) = \log P(\theta).$$

Here, the tradeoff between likelihood and prior is naturally captured by setting the regularization function equal to the log of the prior distribution.

Popular Regularization functions

- ℓ_2 **regularization**:

$$\mathcal{R}(\theta) = \|\theta\|_2^2 = \sum_{j=1}^{p} \theta_j^2.$$

This regularization encodes the prior information that the parameter values are not too large (where how large is determined by the regularization tradeoff parameter λ).

This regularization is thus a "general purpose" regularization function (who wants their parameters to be very large?)

Popular Regularization functions

- ℓ_1 **regularization**:

$$\mathcal{R}(\theta) = \|\theta\|_1 = \sum_{j=1}^{p} |\theta_j|.$$

This regularization encodes the prior information that the parameter values be **sparse**: i.e. with many zero values.

This is a very important prior constraint in big data settings:
with very large number of parameters, we expect the true model to depend on only a few non-zero parameters.

Widely used in high-dimensional model learning: called LASSO when used with linear regression models.

POLL: https://forms.gle/b9ajn8d5vyM5T6P46

Poll: When does MAP equal MLE?

Suppose we estimate a parameter θ using MAP:
$\theta_{MAP} = \arg\max_\theta P(D \mid \theta) P(\theta)$.

For which choice of prior $P(\theta)$ will θ_{MAP} be identical to θ_{MLE} for any dataset D?

A. A uniform (constant) prior over θ

B. A symmetric Beta(β,β) prior with $\beta > 1$

C. A prior highly concentrated near $\theta = 0.5$

D. Any prior, as long as the sample size is small

E. Any prior; MAP always equals MLE