Maximum Likelihood Estimation
&
Maximum A Posteriori Probability Estimation

Announcements

1. HW2 (Perceptron, PCA, K-means) will be out today

Recap on Perceptron

Binary classifier: sign$(w^\top x)$

The Perceptron Alg:

Initialize $w_0 = 0$

$\vec{0} \in \mathbb{R}^d$

For $t = 0 \rightarrow \infty$

Recap on Perceptron

Binary classifier: sign$(w^\top x)$

The Perceptron Alg:

Initialize $w_0 = 0$

For $t = 0 \rightarrow \infty$

feature x_t shows up

Recap on Perceptron

Binary classifier: sign$(w^\top x)$

The Perceptron Alg:

Initialize $w_0 = 0$

For $t = 0 \to \infty$

feature x_t shows up

We make a prediction $\hat{y}_t = \text{sign}(w_t^\top x_t)$

Recap on Perceptron

Binary classifier: sign$(w^\top x)$

The Perceptron Alg:

Initialize $w_0 = 0$

For $t = 0 \rightarrow \infty$

feature x_t shows up

We make a prediction $\hat{y}_t = \text{sign}(w_t^\top x_t)$

Check if $\hat{y}_t$ equal to y_t

Recap on Perceptron

Binary classifier: $\text{sign}(w^\top x)$

The Perceptron Alg:

Initialize $w_0 = 0$

For $t = 0 \rightarrow \infty$

- feature x_t shows up
- We make a prediction $\hat{y}_t = \text{sign}(w_t^\top x_t)$
- Check if $\hat{y}_t$ equal to y_t
- We update $w_{t+1} = w_t + \mathbf{1}(\hat{y}_t \neq y_t) y_t x_t$

$= \begin{cases} 1 & y_t \neq \hat{y}_t \\ 0 & \text{else} \end{cases}$

Recap on Perceptron

Binary classifier: $\text{sign}(w^\top x)$

The Perceptron Alg:

Initialize $w_0 = 0$

For $t = 0 \rightarrow \infty$

- feature x_t shows up
- We make a prediction $\hat{y}_t = \text{sign}(w_t^\top x_t)$
- Check if $\hat{y}_t$ equal to y_t
- We update $w_{t+1} = w_t + \mathbf{1}(\hat{y}_t \neq y_t) y_t x_t$

Q: how to apply this on a static dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$?

Recap on Perceptron

Binary classifier: sign$(w^\top x)$

The Perceptron Alg:

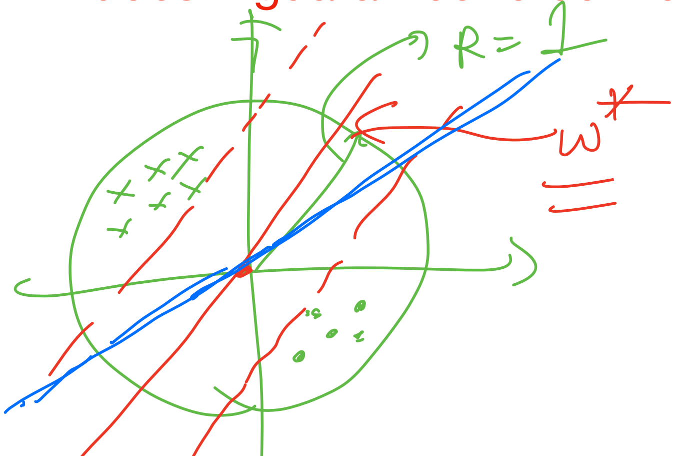
Initialize $w_0 = 0$

For $t = 0 \rightarrow \infty$

- feature x_t shows up
- We make a prediction $\hat{y}_t = \text{sign}(w_t^\top x_t)$
- Check if $\hat{y}_t$ equal to y_t
- We update $w_{t+1} = w_t + \mathbf{1}(\hat{y}_t \neq y_t) y_t x_t$

Q: how to apply this on a static dataset $\mathscr{D} = \{x_i, y_i\}_{i=1}^n$?

Q: If data has margin $y_i(x_i^\top w^\star) \geq \gamma$, does it guarantee to converge to $w^\star$?



Objective for today:

Understand the two common statistical learning framework: MLE and MAP

Outline for today:

1. Maximum Likelihood estimation (MLE)

2. Maximum a posteriori probability (MAP)

Ex 1: Estimating the probability of a coin flip

We toss a coin n times (independently), we observe the following outcomes:

$$\mathscr{D} = \{y_i\}_{i=1}^{n}, y_i \in \{-1, 1\}$$

($y_i = 1$ means head in i's trial, -1 means tail)

Ex 1: Estimating the probability of a coin flip

We toss a coin n times (independently), we observe the following outcomes:

$$\mathscr{D} = \{y_i\}_{i=1}^n, y_i \in \{-1, 1\} \quad (y_i = 1 \text{ means head in } i\text{'s trial, -1 means tail})$$

Q: assume $y_i \sim \text{Bernoulli}(\theta^\star)$, how to estimate $\theta^\star$ given $\mathscr{D}$?

$$\begin{cases} y_i = +1 & \text{wp } \theta^\star \\ = -1 & \text{wp } 1-\theta^\star \end{cases}$$

Ex 1: Estimating the probability of a coin flip

We toss a coin n times (independently), we observe the following outcomes:

$$\mathscr{D} = \{y_i\}_{i=1}^n, y_i \in \{-1,1\}$$ ($y_i = 1$ means head in i's trial, -1 means tail)

Q: assume $y_i \sim$ Bernoulli($\theta^\star$), how to estimate $\theta^\star$ given $\mathscr{D}$?

$$\hat{\theta} = \frac{\sum_{i=1}^n \mathbf{1}(y_i = 1)}{n}$$

⟷ $\theta^\star$ when $n \to \infty$

Ex 1: Estimating the probability of a coin flip

We toss a coin n times (independently), we observe the following outcomes:

$\mathscr{D} = \{y_i\}_{i=1}^n, y_i \in \{-1,1\}$ ($y_i = 1$ means head in i's trial, -1 means tail)

Q: assume $y_i \sim$ Bernoulli($\theta^\star$), how to estimate $\theta^\star$ given $\mathscr{D}$?

$$\hat{\theta} = \frac{\sum_{i=1}^n \mathbf{1}(y_i = 1)}{n}$$

Let's make this rigorous!

Maximum Likelihood Estimation

We toss a coin n times (independently), we observe the following outcomes:

$\mathscr{D} = \{y_i\}_{i=1}^n, y_i \in \{-1,1\}$ ($y_i = 1$ means head in i's trial, -1 means tail)

Maximum Likelihood Estimation

We toss a coin n times (independently), we observe the following outcomes:

$$\mathscr{D} = \{y_i\}_{i=1}^n, y_i \in \{-1,1\}$$ ($y_i = 1$ means head in i's trial, -1 means tail)

If the probability of getting head is $\theta \in [0,1]$, what is the probability of observing the data $\mathscr{D}$ (i.e., likelihood)?

$$P(D|\theta) = \prod_{i=1}^{n} P(y_i|\theta)$$

$$= \begin{cases} y_i = 1, & \text{wp } \theta \\ y_i = -1 & \text{wp } 1-\theta \end{cases}$$

Maximum Likelihood Estimation

We toss a coin n times (independently), we observe the following outcomes:

$$\mathscr{D} = \{y_i\}_{i=1}^n, y_i \in \{-1,1\}$$ ($y_i = 1$ means head in i's trial, -1 means tail)

If the probability of getting head is $\theta \in [0,1]$, what is the probability of observing the data $\mathscr{D}$ (i.e., likelihood)?

$$P(\mathscr{D} \mid \theta) = \theta^{n_1}(1-\theta)^{n-n_1}$$

$$n_1 = \sum_{i=1}^{n} 1(y_i = 1)$$

# of heads

Maximum Likelihood Estimation

We toss a coin n times (independently), we observe the following outcomes:

$$\mathscr{D} = \{y_i\}_{i=1}^n, y_i \in \{-1,1\}$$ ($y_i = 1$ means head in i's trial, -1 means tail)

If the probability of getting head is $\theta \in [0,1]$, what is the probability of observing the data $\mathscr{D}$ (i.e., likelihood)?

$$P(\mathscr{D} \mid \theta) = \theta^{n_1}(1-\theta)^{n-n_1}$$

MLE Principle: Find θ that **maximizes the likelihood** of the data:

Maximum Likelihood Estimation

We toss a coin n times (independently), we observe the following outcomes:

$$\mathscr{D} = \{y_i\}_{i=1}^{n}, y_i \in \{-1,1\}$$ ($y_i = 1$ means head in i's trial, -1 means tail)

If the probability of getting head is $\theta \in [0,1]$, what is the probability of observing the data $\mathscr{D}$ (i.e., likelihood)?

$$P(\mathscr{D} \mid \theta) = \theta^{n_1}(1-\theta)^{n-n_1}$$

$$P(D \mid \theta) = \prod_{i=1}^{n} P(y_i \mid \theta)$$

MLE Principle: Find θ that **maximizes the likelihood** of the data:

$$\hat{\theta}_{mle} = \arg\max_{\theta \in [0,1]} P(\mathscr{D} \mid \theta)$$

Maximum Likelihood Estimation

We toss a coin n times (independently), we observe the following outcomes:

$$\mathscr{D} = \{y_i\}_{i=1}^n, y_i \in \{-1, 1\}$$

($y_i = 1$ means head in i's trial, -1 means tail)

MLE Principle: Find θ that maximizes the likelihood of the data:

$$\hat{\theta}_{mle} = \arg\max_{\theta \in [0,1]} P(\mathscr{D} \mid \theta)$$

Maximum Likelihood Estimation

We toss a coin n times (independently), we observe the following outcomes:

$$\mathscr{D} = \{y_i\}_{i=1}^n, y_i \in \{-1,1\}$$ ($y_i = 1$ means head in i's trial, -1 means tail)

MLE Principle: Find θ that maximizes the likelihood of the data:

$$\hat{\theta}_{mle} = \arg\max_{\theta \in [0,1]} P(\mathscr{D} \mid \theta) = \arg\max_{\theta \in [0,1]} \theta^{n_1}(1-\theta)^{n-n_1}$$

Maximum Likelihood Estimation

We toss a coin n times (independently), we observe the following outcomes:

$$\mathscr{D} = \{y_i\}_{i=1}^n, y_i \in \{-1,1\} \quad (y_i = 1 \text{ means head in } i\text{'s trial, -1 means tail})$$

MLE Principle: Find θ that maximizes the likelihood of the data:

$$\hat{\theta}_{mle} = \arg\max_{\theta \in [0,1]} P(\mathscr{D} \mid \theta) \quad = \arg\max_{\theta \in [0,1]} \theta^{n_1}(1-\theta)^{n-n_1}$$

$$= \arg\max_{\theta \in [0,1]} \ln(\theta^{n_1}(1-\theta)^{n-n_1})$$

$$\ln \theta^{n_1} + \ln(1-\theta)^{n-n_1} \Rightarrow n_1 \ln \theta + (n-n_1)\ln(1-\theta)$$

Maximum Likelihood Estimation

We toss a coin n times (independently), we observe the following outcomes:

$$\mathscr{D} = \{y_i\}_{i=1}^n, y_i \in \{-1, 1\}$$

($y_i = 1$ means head in i's trial, -1 means tail)

MLE Principle: Find θ that maximizes the likelihood of the data:

$$\hat{\theta}_{mle} = \arg\max_{\theta \in [0,1]} P(\mathscr{D} \mid \theta) = \arg\max_{\theta \in [0,1]} \theta^{n_1}(1-\theta)^{n-n_1}$$

$$= \arg\max_{\theta \in [0,1]} \ln(\theta^{n_1}(1-\theta)^{n-n_1})$$

$$= \arg\max_{\theta \in [0,1]} n_1 \ln(\theta) + (n - n_1)\ln(1-\theta)$$

Maximum Likelihood Estimation

We toss a coin n times (independently), we observe the following outcomes:

$$\mathscr{D} = \{y_i\}_{i=1}^n, y_i \in \{-1,1\}$$ ($y_i = 1$ means head in i's trial, -1 means tail)

MLE Principle: Find θ that maximizes the likelihood of the data:

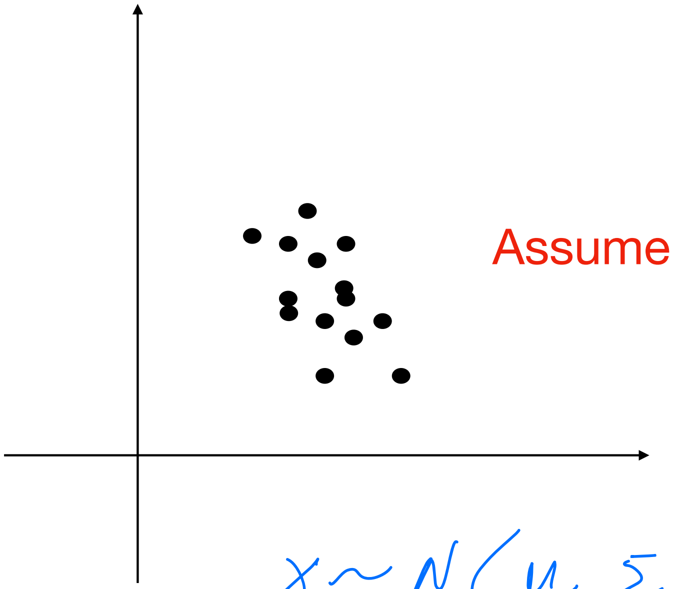
$$\hat{\theta}_{mle} = \arg\max_{\theta\in[0,1]} P(\mathscr{D} \mid \theta) = \arg\max_{\theta\in[0,1]} \theta^{n_1}(1-\theta)^{n-n_1}$$

$$= \arg\max_{\theta\in[0,1]} \ln(\theta^{n_1}(1-\theta)^{n-n_1})$$

$$= \arg\max_{\theta\in[0,1]} n_1 \ln(\theta) + (n-n_1)\ln(1-\theta) = \frac{n_1}{n}$$

$z = u^T x \sim N(\mu^T u, u^T \Sigma u)$

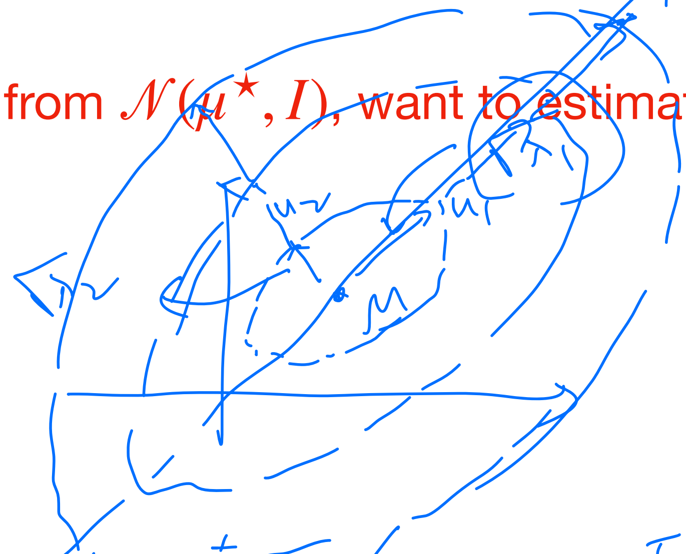
Ex 2: Estimate the mean



$x \sim N(\mu, \Sigma)$

$$\mathscr{D} = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}^d$$

Assume data is from $\mathcal{N}(\mu^\star, I)$, want to estimate $\mu^\star$ from the data $\mathscr{D}$



$\Sigma = U \Lambda U^T$

$x \in R^d$

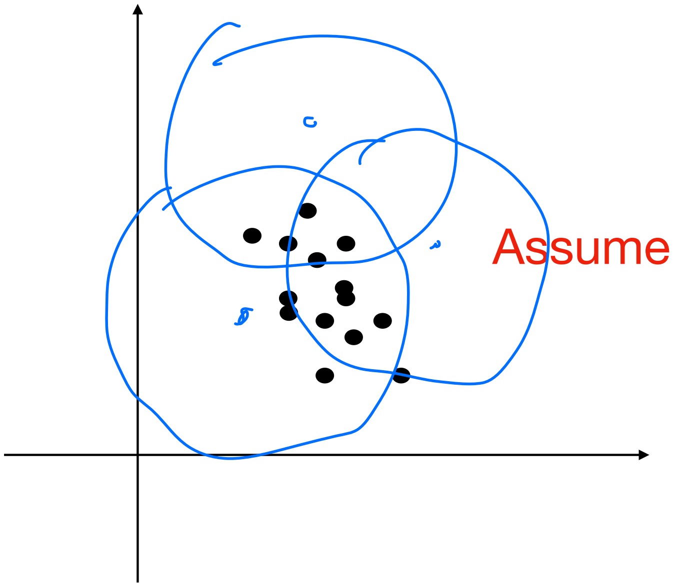
$\mu \in R^d$

$\Sigma \in R^{d \times d}$

PSD

pdf: $\frac{1}{\sqrt{(2\pi)^d}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$

Ex 2: Estimate the mean



$$\mathscr{D} = \{x_i\}_{i=1}^{n}, x_i \in \mathbb{R}^d$$

Assume data is from $\mathcal{N}(\mu^\star, I)$, want to estimate $\mu^\star$ from the data $\mathscr{D}$

$\in \mathbb{R}^{d \times d}$

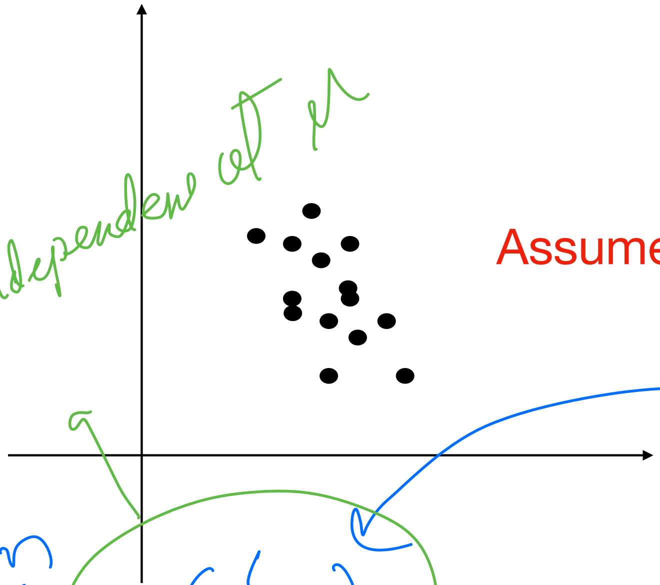
Let's apply the MLE Principle:

Step 1: $$P(\mathscr{D} \mid \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2}(x_i - \mu)^\top (x_i - \mu)\right)$$

$$= \prod_{i=1}^{d} P(x_i \mid \mu)$$

$P(x_i \mid \mu)$

Ex 2: Estimate the mean



$$\mathscr{D} = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}^d$$

Assume data is from $\mathscr{N}(\mu^\star, I)$, want to estimate $\mu^\star$ from the data $\mathscr{D}$

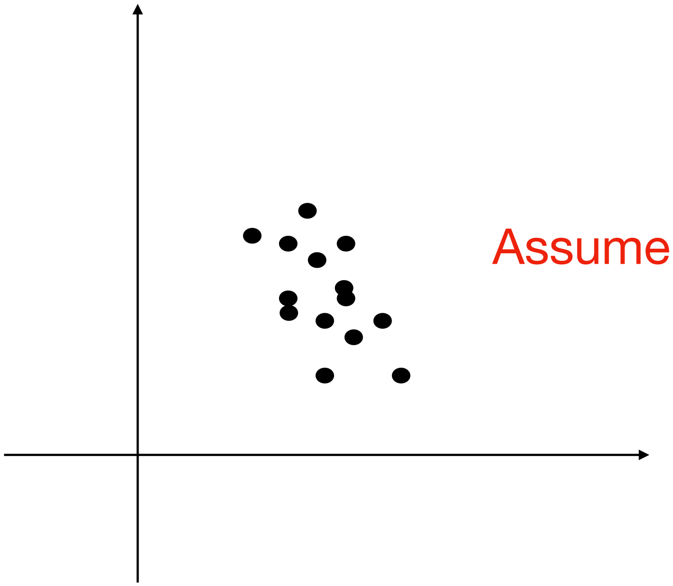
Let's apply the MLE Principle:

Step 1: $P(\mathscr{D} \mid \mu) = \prod_{i=1}^n \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2}(x_i - \mu)^\top (x_i - \mu)\right)$

Step 2: apply log and maximize the log-likelihood:

$$\arg\max_{\mu} \sum_{i=1}^n -(x_i - \mu)^\top (x_i - \mu)$$

Ex 2: Estimate the mean



$$\mathcal{D} = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}^d$$

Assume data is from $\mathcal{N}(\mu^\star, I)$, want to estimate $\mu^\star$ from the data $\mathcal{D}$

Let's apply the MLE Principle:

Step 1: $P(\mathcal{D} \mid \mu) = \prod_{i=1}^n \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2}(x_i - \mu)^\top (x_i - \mu)\right)$

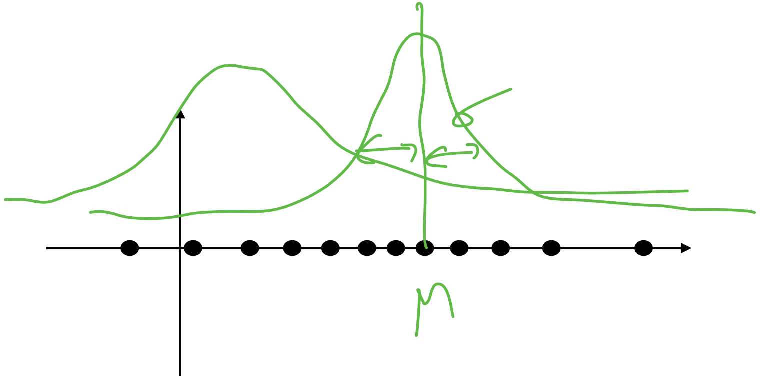
Step 2: apply log and maximize the log-likelihood:

$$\arg\max_\mu \sum_{i=1}^n -(x_i - \mu)^\top (x_i - \mu) \Rightarrow \hat{\mu}_{mle} = \sum_{i=1}^n x_i / n$$

Q: Estimate the mean and variance

$\mu \in \mathbb{R}$

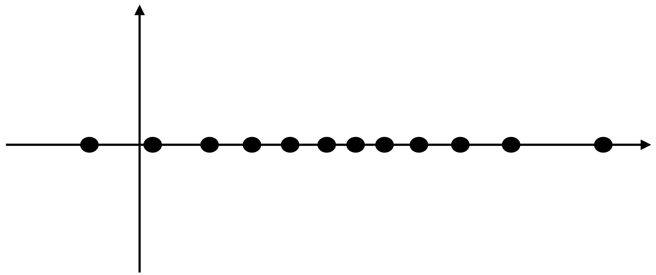
$\sigma \in \mathbb{R}^+$



$$\mathscr{D} = \{x_i\}_{i=1}^{n}, x_i \in \mathbb{R}$$

Assume data is from $\mathcal{N}(\mu^\star, \sigma^2)$, want to estimate $\mu^\star, \sigma$ from the data $\mathscr{D}$

Q: Estimate the mean and variance



$$\mathscr{D} = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}$$

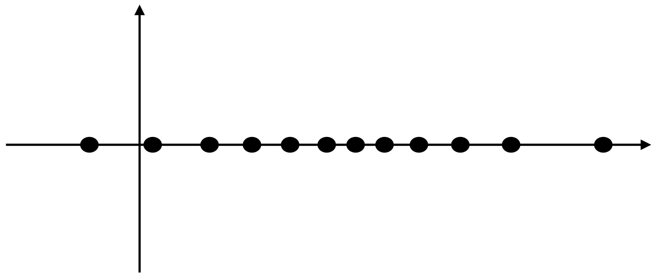
Assume data is from $\mathcal{N}(\mu^\star, \sigma^2)$, want to estimate $\mu^\star, \sigma$ from the data $\mathscr{D}$

Let's apply the MLE Principle:

Step 1: $P(\mathscr{D} \mid \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2/\sigma^2\right)$

$P(x_i \mid \mu, \sigma)$

Q: Estimate the mean and variance



$$\mathscr{D} = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}$$

Assume data is from $\mathcal{N}(\mu^\star, \sigma^2)$, want to estimate $\mu^\star, \sigma$ from the data $\mathscr{D}$

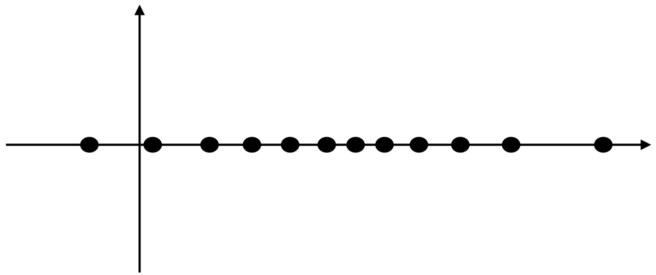
Let's apply the MLE Principle:

Step 1: $P(\mathscr{D} \mid \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2/\sigma^2\right)$

Step 2: apply log and maximize the log-likelihood:

$$\arg\max_{\mu,\sigma>0} \sum_{i=1}^{n} \left(-(x_i-\mu)^2/\sigma^2 - \ln(\sigma)\right)$$

Q: Estimate the mean and variance



$$\mathcal{D} = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}$$

Assume data is from $\mathcal{N}(\mu^\star, \sigma^2)$, want to estimate $\mu^\star, \sigma$ from the data $\mathcal{D}$

Let's apply the MLE Principle:

Step 1: $P(\mathcal{D} \mid \mu, \sigma) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2/\sigma^2\right)$

Step 2: apply log and maximize the log-likelihood:

$$\arg\max_{\mu,\sigma>0} \sum_{i=1}^n \left(-(x_i - \mu)^2/\sigma^2 - \ln(\sigma)\right) = ??$$

Some properties of MLE

1. MLE is consistent: if our model assumption is correct (e.g., coin flip follows some Bernoulli distribution), then $\hat{\theta}_{mle} \rightarrow \theta^{\star},$ as $n \rightarrow \infty$

Some properties of MLE

1. MLE is consistent: if our model assumption is correct (e.g., coin flip follows some Bernoulli distribution), then $\hat{\theta}_{mle} \rightarrow \theta^\star,$ as $n \rightarrow \infty$

2. When our model assumption is wrong (e.g., we use Gaussian to model data which is from some more complicated distribution), then MLE loses such guarantee

Outline for today:

1. Maximum Likelihood estimation (MLE)

2. Maximum a Posteriori Probability (MAP)

Ex: Estimating the probability of a coin flip

We toss a coin n times (independently), we observe the following outcomes:

$$\mathcal{D} = \{y_i\}_{i=1}^n, y_i \in \{-1,1\}$$

($y_i = 1$ means head in i's trial, -1 means tail)

Ex: Estimating the probability of a coin flip

We toss a coin n times (independently), we observe the following outcomes:

$$\mathcal{D} = \{y_i\}_{i=1}^n, y_i \in \{-1,1\}$$ ($y_i = 1$ means head in i's trial, -1 means tail)

A Bayesian Statistician will treat the optimal parameter $\theta^\star$ being a random variable:

$$\theta^\star \sim P(\theta)$$

prior

Ex: Estimating the probability of a coin flip

We toss a coin n times (independently), we observe the following outcomes:

$$\mathscr{D} = \{y_i\}_{i=1}^n, y_i \in \{-1,1\}$$ ($y_i = 1$ means head in i's trial, -1 means tail)

A Bayesian Statistician will treat the optimal parameter $\theta^\star$ being a random variable:

$$\theta^\star \sim P(\theta)$$

Example: $P(\theta)$ being a Beta distribution:

$$P(\theta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}/Z,$$ (α, β)

where $Z = \int_{\theta\in[0,1]} \theta^{\alpha-1}(1-\theta)^{\beta-1} d_\theta$

normalizer

Ex: Estimating the probability of a coin flip

We toss a coin n times (independently), we observe the following outcomes:

$$\mathscr{D} = \{y_i\}_{i=1}^n, y_i \in \{-1,1\}$$ ($y_i = 1$ means head in i's trial, -1 means tail)

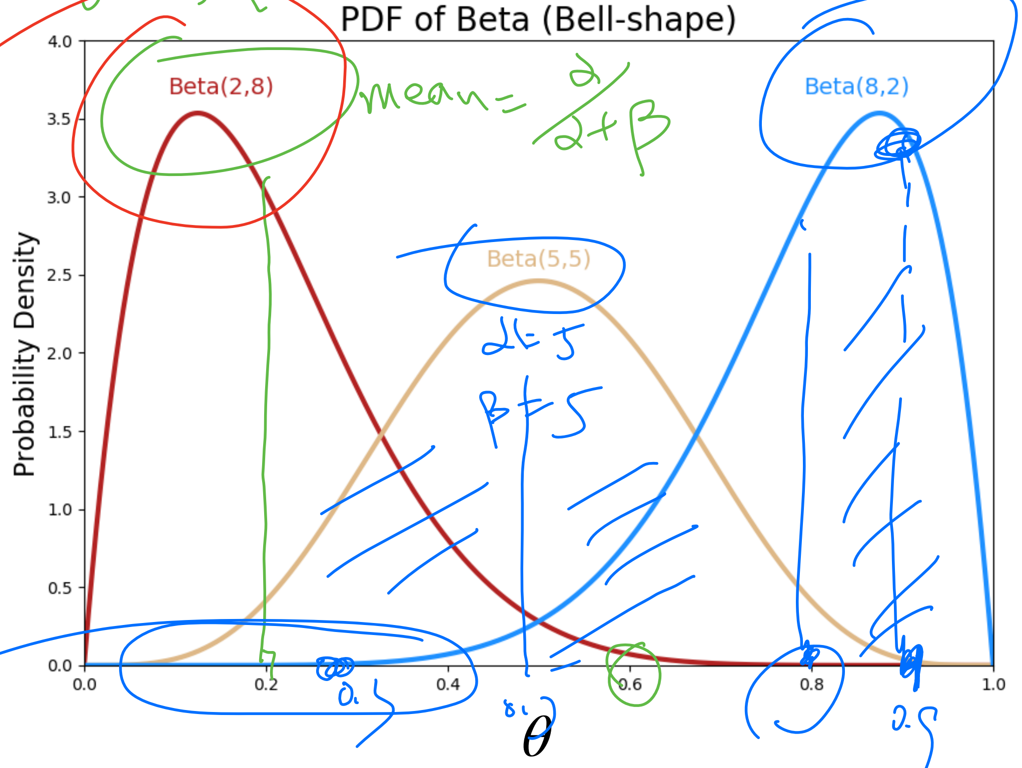
A Bayesian Statistician will treat the optimal parameter $\theta^\star$ being a random variable:

$$\theta^\star \sim P(\theta)$$

Example: $P(\theta)$ being a Beta distribution:

$$P(\theta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}/Z,$$

where $Z = \int_{\theta\in[0,1]} \theta^{\alpha-1}(1-\theta)^{\beta-1} d_\theta$



The Posterior distribution over θ

Now, we have a prior $P(\theta)$, and we have a dataset $\mathscr{D} = \{y_i\}_{i=1}^n$, define posterior distribution:

$$P(\theta \mid \mathscr{D})$$

MLE: $P(D \mid \theta)$

$$P(a, b) = P(b \mid a) P(a)$$

The Posterior distribution over θ

Now, we have a prior $P(\theta)$, and we have a dataset $\mathscr{D} = \{y_i\}_{i=1}^{n}$, define posterior distribution:

$$P(\theta \,|\, \mathscr{D})$$

Using Bayes rule, we get:

$$P(\theta \,|\, \mathscr{D}) = P(\theta)P(\mathscr{D} \,|\, \theta)/P(\mathscr{D})$$

$$P(\theta | D) \cdot P(D) = P(D | \theta) \cdot P(\theta) = P(\theta, D)$$

The Posterior distribution over θ

Now, we have a prior $P(\theta)$, and we have a dataset $\mathscr{D} = \{y_i\}_{i=1}^n$, define posterior distribution:

$$P(\theta \mid \mathscr{D})$$

Using Bayes rule, we get:

$$P(\theta \mid \mathscr{D}) = P(\theta)P(\mathscr{D} \mid \theta)/P(\mathscr{D})$$

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

$f(x) \propto g(x)$

$\iff \frac{f(x)}{g(x)} = c, \forall x$

independent of θ

The Posterior distribution over θ

Now, we have a prior $P(\theta)$, and we have a dataset $\mathscr{D} = \{y_i\}_{i=1}^n$, define posterior distribution:

$$P(\theta \mid \mathscr{D})$$

prior

Using Bayes rule, we get:

$$P(\theta \mid \mathscr{D}) = P(\theta)P(\mathscr{D} \mid \theta)/P(\mathscr{D})$$

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

← Data likelihood

Posterior ∝ Prior × Likelihood

The Posterior distribution over θ

Now, we have a prior $P(\theta)$, and we have a dataset $\mathscr{D} = \{y_i\}_{i=1}^{n}$, define posterior distribution:

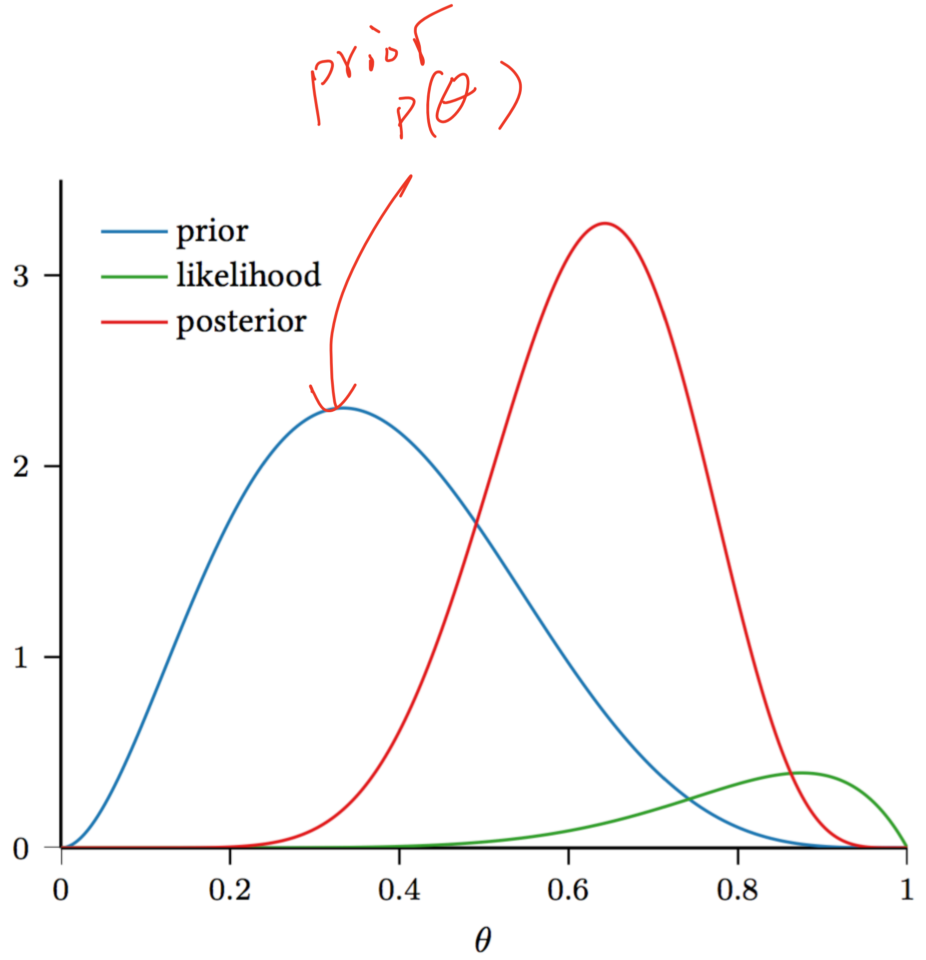
$$P(\theta \mid \mathscr{D})$$

Using Bayes rule, we get:

$$P(\theta \mid \mathscr{D}) = P(\theta)P(\mathscr{D} \mid \theta)/P(\mathscr{D})$$

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

Posterior ∝ Prior × Likelihood



The Posterior distribution over θ

Now, we have a prior $P(\theta)$, and we have a dataset $\mathscr{D} = \{y_i\}_{i=1}^n$, define posterior distribution:

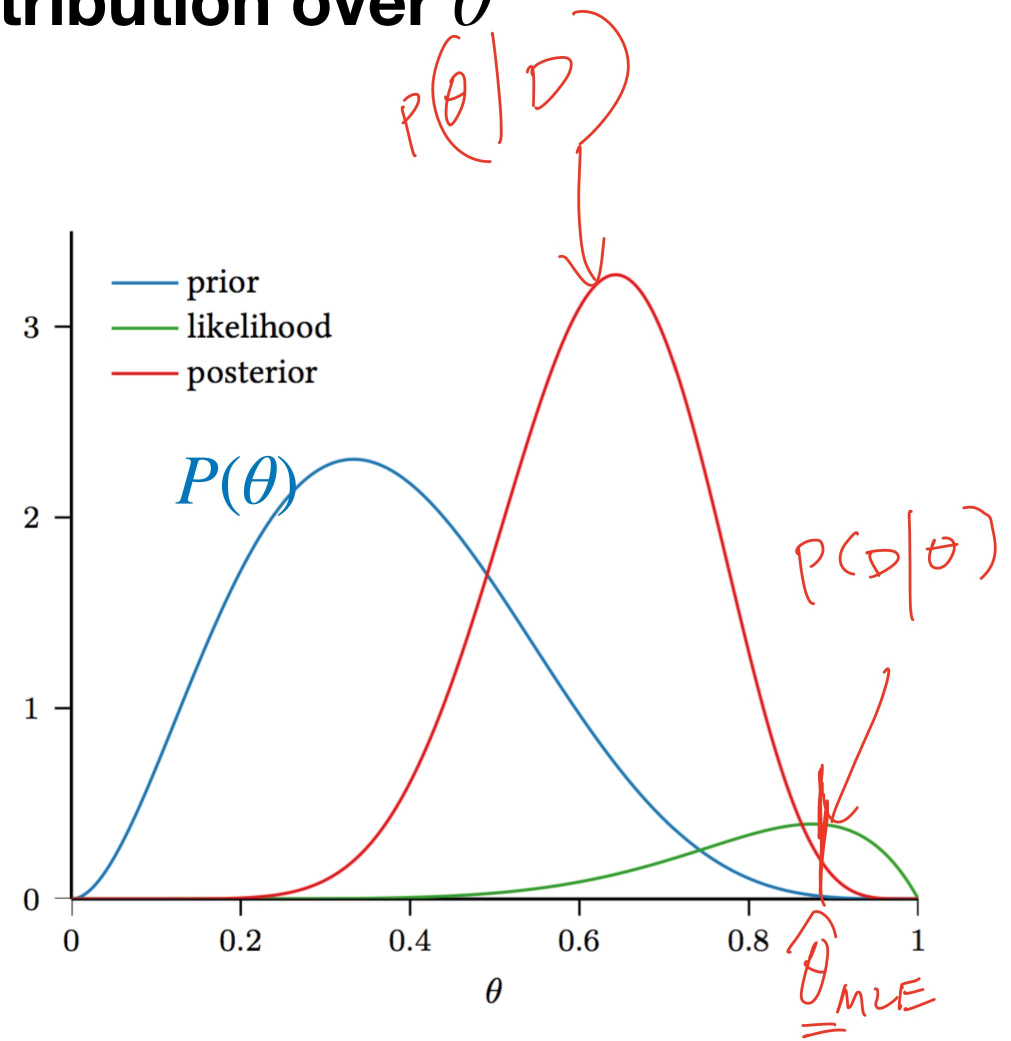
$$P(\theta \mid \mathscr{D})$$

Using Bayes rule, we get:

$$P(\theta \mid \mathscr{D}) = P(\theta)P(\mathscr{D} \mid \theta)/P(\mathscr{D})$$

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

Posterior ∝ Prior × Likelihood



The Posterior distribution over θ

Now, we have a prior $P(\theta)$, and we have a dataset $\mathscr{D} = \{y_i\}_{i=1}^n$, define posterior distribution:

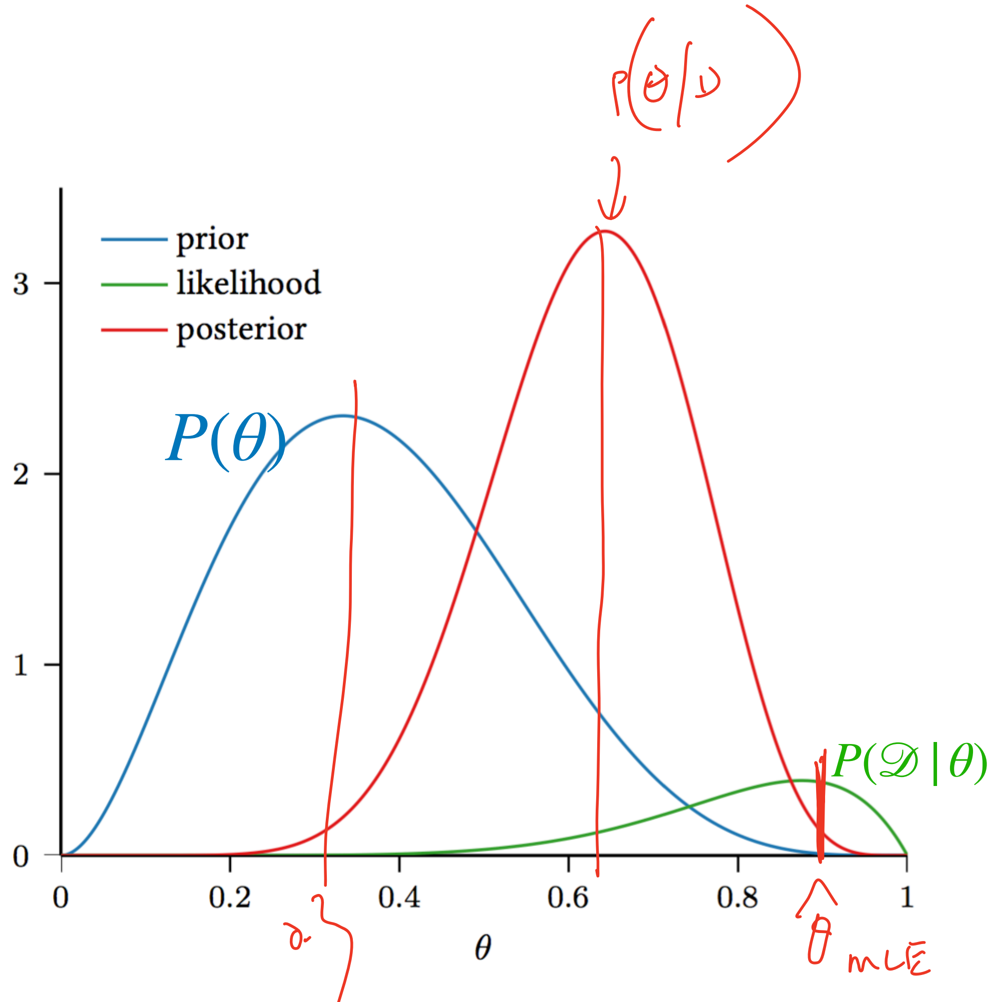
$$P(\theta \mid \mathscr{D})$$

Using Bayes rule, we get:

$$P(\theta \mid \mathscr{D}) = P(\theta)P(\mathscr{D} \mid \theta)/P(\mathscr{D})$$

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

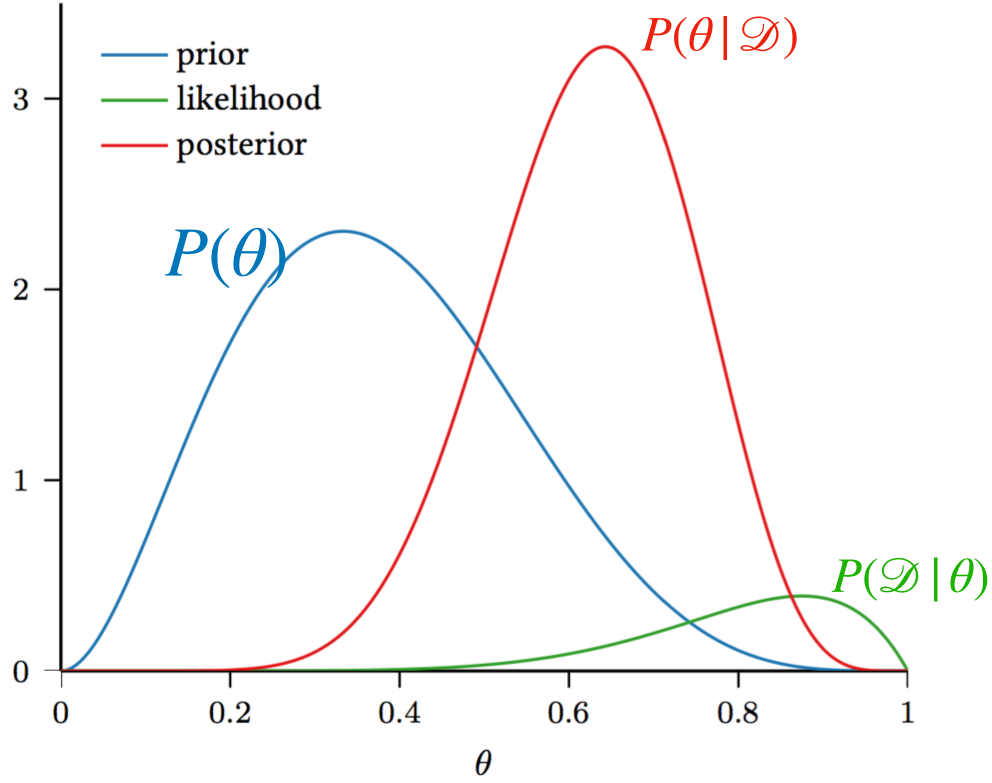
Posterior ∝ Prior × Likelihood



The Posterior distribution over θ

Now, we have a prior $P(\theta)$, and we have a dataset $\mathcal{D} = \{y_i\}_{i=1}^{n}$, define posterior distribution:

$$P(\theta \,|\, \mathcal{D})$$



Using Bayes rule, we get:

$$P(\theta \,|\, \mathcal{D}) = P(\theta)P(\mathcal{D} \,|\, \theta)/P(\mathcal{D})$$

$$\propto P(\theta)P(\mathcal{D} \,|\, \theta)$$

Posterior ∝ Prior × Likelihood

Maximum A Posteriori Probability estimation (MAP)

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

$$P(\theta) = \frac{1}{Z} \theta^{\alpha-1}(1-\theta)^{\beta-1} \leftarrow \text{prior}$$

$$P(\mathscr{D} \mid \theta) = \prod_{i=1}^{n} \theta^{n_1}(1-\theta)^{n-n_1}$$

Maximum A Posteriori Probability estimation (MAP)

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

$$\hat{\theta}_{map} = \arg\max_{\theta \in [0,1]} P(\theta \mid \mathscr{D}) = \arg\max_{\theta \in [0,1]} P(\theta)P(\mathscr{D} \mid \theta)$$

Maximum A Posteriori Probability estimation (MAP)

$$P(\theta | \mathscr{D}) \propto P(\theta)P(\mathscr{D} | \theta)$$

$$\hat{\theta}_{map} = \arg\max_{\theta \in [0,1]} P(\theta | \mathscr{D}) = \arg\max_{\theta \in [0,1]} P(\theta)P(\mathscr{D} | \theta)$$

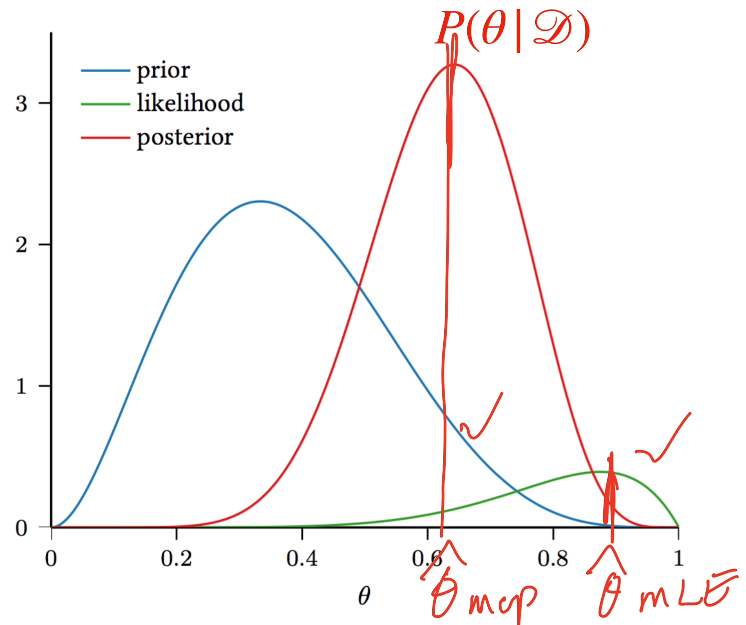
$$= \arg\max_{\theta \in [0,1]} \ln P(\theta) + \ln P(\mathscr{D} | \theta)$$

Maximum A Posteriori Probability estimation (MAP)

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

$$\hat{\theta}_{map} = \arg\max_{\theta \in [0,1]} P(\theta \mid \mathscr{D}) = \arg\max_{\theta \in [0,1]} P(\theta)P(\mathscr{D} \mid \theta)$$

$$= \arg\max_{\theta \in [0,1]} \ln P(\theta) + \ln P(\mathscr{D} \mid \theta)$$



MAP for coin flip

$$\hat{\theta}_{map} = \arg\max_{\theta \in [0,1]} \ln(P(\theta)P(\mathcal{D} \mid \theta))$$

$$P(\theta) = \frac{1}{Z} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$P(\mathcal{D} \mid \theta) = \prod_{i=1}^{n} P(y_i \mid \theta)$$

← Bernoulli(θ)

MAP for coin flip

$$\hat{\theta}_{map} = \arg\max_{\theta \in [0,1]} \ln(P(\theta)P(\mathscr{D} \mid \theta))$$

Step 1: specify Prior $P(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ ✓ Prior Beta Dis

Step 2: data likelihood $P(\mathscr{D} \mid \theta) = \theta^{n_1}(1-\theta)^{n-n_1}$ ✓

Step 3: Compute posterior $P(\theta \mid \mathscr{D}) \propto \theta^{n_1+\alpha-1}(1-\theta)^{n-n_1+\beta-1}$

Step 4: Compute MAP $\hat{\theta}_{map} = \dfrac{n_1+\alpha-1}{n+\alpha+\beta-2}$

$$\left(\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)\cdot\left(\theta^{n_1}(1-\theta)^{n-n_1}\right)$$

$$\arg\max_{\theta \in [0,1]} P(\theta / D)$$

MAP for coin flip

$$\hat{\theta}_{map} = \arg\max_{\theta \in [0,1]} \ln(P(\theta)P(\mathcal{D} \mid \theta))$$

Step 1: specify Prior $P(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$

Step 2: data likelihood $P(\mathcal{D} \mid \theta) = \theta^{n_1}(1-\theta)^{n-n_1}$

Step 3: Compute posterior $P(\theta \mid \mathcal{D}) \propto \theta^{n_1+\alpha-1}(1-\theta)^{n-n_1+\beta-1}$

Step 4: Compute MAP $\hat{\theta}_{map} = \dfrac{n_1 + \alpha - 1}{n + \alpha + \beta - 2}$

$$\frac{n_1 + (\alpha - 1)}{(n_1 + \alpha - 1) + (n - n_1 + \beta - 1)}$$

$(\alpha - 1, \beta - 1)$ can be understood as some fictions flips: we had $\alpha - 1$ hallucinated heads, and $\beta - 1$ hallucinated tails

Some considerations on prior distributions

1. In coin flip example, when $n \to \infty$, $\hat{\theta}_{map} = \frac{n_1 + \alpha - 1}{n + \alpha + \beta - 2} \to \frac{n_1}{n}$ (i.e., $\hat{\theta}_{mle}$)

Some considerations on prior distributions

1. In coin flip example, when $n \to \infty, \hat{\theta}_{map} = \dfrac{n_1 + \alpha - 1}{n + \alpha + \beta - 2} \to \dfrac{n_1}{n}$(i.e.,$\hat{\theta}_{mle}$)

2. When n is small and our prior is accurate, MAP can work better than MLE

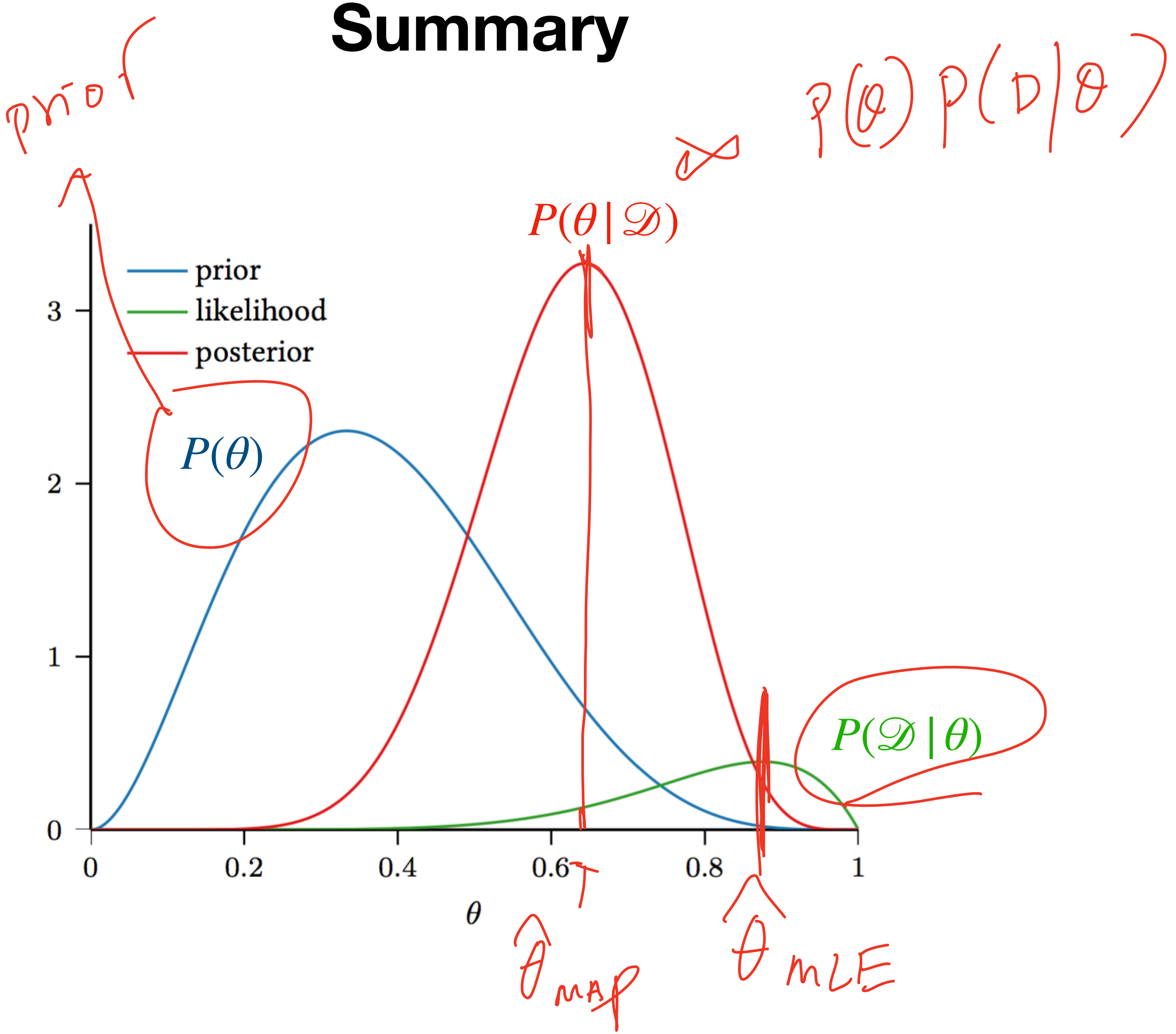
Some considerations on prior distributions

1. In coin flip example, when $n \to \infty, \hat{\theta}_{map} = \dfrac{n_1 + \alpha - 1}{n + \alpha + \beta - 2} \to \dfrac{n_1}{n}$(i.e.,$\hat{\theta}_{mle}$)

2. When n is small and our prior is accurate, MAP can work better than MLE

3. In general, not so easy to set up a good prior….

Summary


prior
likelihood
posterior
$P(\theta)$
$P(\theta | \mathscr{D})$
$P(\mathscr{D} | \theta)$
θ
0
0.2
0.4
0.6
0.8
1
0
1
2
3
prior
$\propto P(\theta) P(D|\theta)$
$\hat{\theta}_{MAP}$
$\hat{\theta}_{MLE}$

Summary for today

1 MLE (frequentist perspective):

The ground truth $\theta^\star$ is unknown but fixed; we search for the parameter that makes the data as likely as possible

Summary for today

1 MLE (frequentist perspective):

The ground truth $\theta^\star$ is unknown but fixed; we search for the parameter that makes the data as likely as possible

$$\arg\max_{\theta} P(\mathscr{D} \mid \theta)$$

$$\arg\max_{\theta\in[0,1]} \theta^{x}(1-\theta)^{n-x}$$

$$\arg\max_{\theta\in[0,1]} x\ln\theta + (n-x)\ln(1-\theta)$$

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

$$\frac{x}{\theta} = \frac{n-x}{1-\theta}$$

$$\Rightarrow \theta = \frac{x}{n}$$

Summary for today

1 MLE (frequentist perspective):

The ground truth $\theta^\star$ is unknown but fixed; we search for the parameter that makes the data as likely as possible

$$\arg\max_\theta P(\mathscr{D} \mid \theta)$$

2 MAP (Bayesian perspective):

The ground truth $\theta^\star$ treated as a random variable, i.e., $\theta^\star \sim P(\theta)$; we search for the parameter that maximizes the posterior

Summary for today

1 MLE (frequentist perspective):

The ground truth $\theta^\star$ is unknown but fixed; we search for the parameter that makes the data as likely as possible

$$\arg\max_\theta P(\mathscr{D} \mid \theta)$$

2 MAP (Bayesian perspective):

The ground truth $\theta^\star$ treated as a random variable, i.e., $\theta^\star \sim P(\theta)$; we search for the parameter that maximizes the posterior

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