# Maximum Likelihood Estimation
# &
# Maximum A Posteriori Probability Estimation

# Recap on Perceptron

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

**The Perceptron Alg:**

Initialize $w_0 = 0$

For $t = 0 \rightarrow \infty$

# Recap on Perceptron

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

**The Perceptron Alg:**

Initialize $w_0 = 0$

For $t = 0 \to \infty$

User comes with feature $x_t$

# Recap on Perceptron

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

**The Perceptron Alg:**

Initialize $w_0 = 0$

For $t = 0 \to \infty$

User comes with feature $x_t$

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 $t = 0 \to \infty$

User comes with feature $x_t$

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

User reveals the real label $y_t$

# Recap on Perceptron

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

**The Perceptron Alg:**

Initialize $w_0 = 0$

For $t = 0 \rightarrow \infty$

User comes with feature $x_t$

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

User reveals the real label $y_t$

We update $w_{t+1} = w_t + \mathbf{1}(\hat{y}_t \neq y_t) y_t x_t$

# Recap on Perceptron

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

**The Perceptron Alg:**

Initialize $w_0 = 0$

For $t = 0 \to \infty$

- User comes with feature $x_t$
- We make a prediction $\hat{y}_t =$ sign$(w_t^\top x_t)$
- User reveals the real label $y_t$
- We update $w_{t+1} = w_t + \mathbf{1}(\hat{y}_t \neq y_t) y_t x_t$

Theorem:

if there exists $w^\star$ with $\|w^\star\|_2 = 1$, such that $y_t(x_t^\top w^\star) \geq \gamma > 0, \forall t,$

then:

$$\sum_{t=0}^{\infty} \mathbf{1}(\hat{y}_t \neq y_t) \leq 1/\gamma^2$$

# Recap on Perceptron

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

**The Perceptron Alg:**

Initialize $w_0 = 0$

For $t = 0 \rightarrow \infty$

- User comes with feature $x_t$
- We make a prediction $\hat{y}_t = \text{sign}(w_t^\top x_t)$
- User reveals the real label $y_t$
- We update $w_{t+1} = w_t + \mathbf{1}(\hat{y}_t \neq y_t) y_t x_t$

Theorem:

if there exists $w^\star$ with $\|w^\star\|_2 = 1$, such that $y_t(x_t^\top w^\star) \geq \gamma > 0, \forall t,$

then:

$$\sum_{t=0}^{\infty} \mathbf{1}(\hat{y}_t \neq y_t) \leq 1/\gamma^2$$

Q: does the data need to be i.i.d?

# Recap on Perceptron

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

**The Perceptron Alg:**

Initialize $w_0 = 0$

For $t = 0 \to \infty$

- User comes with feature $x_t$
- We make a prediction $\hat{y}_t = \text{sign}(w_t^\top x_t)$
- User reveals the real label $y_t$
- We update $w_{t+1} = w_t + \mathbf{1}(\hat{y}_t \neq y_t) y_t x_t$

Theorem:

if there exists $w^\star$ with $\|w^\star\|_2 = 1$, such that

$$y_t(x_t^\top w^\star) \geq \gamma > 0, \forall t,$$

then:

$$\sum_{t=0}^{\infty} \mathbf{1}(\hat{y}_t \neq y_t) \leq 1/\gamma^2$$

# Recap on Perceptron

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

**The Perceptron Alg:**

Initialize $w_0 = 0$

For $t = 0 \to \infty$

- User comes with feature $x_t$
- We make a prediction $\hat{y}_t = \text{sign}(w_t^\top x_t)$
- User reveals the real label $y_t$
- We update $w_{t+1} = w_t + \mathbf{1}(\hat{y}_t \neq y_t) y_t x_t$

Theorem:

if there exists $w^\star$ with $\|w^\star\|_2 = 1$, such that

$y_t(x_t^\top w^\star) \geq \gamma > 0, \forall t,$

then:

$$\sum_{t=0}^{\infty} \mathbf{1}(\hat{y}_t \neq y_t) \leq 1/\gamma^2$$

No i.i.d assumption, and indeed data $\{x_1, y_1 \ldots, x_T, y_T\}$ can be selected by an Adversary (as long as it is separable)!!!

# Recap on Perceptron

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

**The Perceptron Alg:**

Initialize $w_0 = 0$

For $t = 0 \rightarrow \infty$

- User comes with feature $x_t$
- We make a prediction $\hat{y}_t = \text{sign}(w_t^\top x_t)$
- User reveals the real label $y_t$
- We update $w_{t+1} = w_t + \mathbf{1}(\hat{y}_t \neq y_t) y_t x_t$

Theorem:

if there exists $w^\star$ with $\|w^\star\|_2 = 1$, such that

$y_t(x_t^\top w^\star) \geq \gamma > 0, \forall t,$

then:

$$\sum_{t=0}^{\infty} \mathbf{1}(\hat{y}_t \neq y_t) \leq 1/\gamma^2$$

# Recap on Perceptron

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

**The Perceptron Alg:**

Initialize $w_0 = 0$

For $t = 0 \rightarrow \infty$

User comes with feature $x_t$

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

User reveals the real label $y_t$

We update $w_{t+1} = w_t + \mathbf{1}(\hat{y}_t \neq y_t) y_t x_t$

Theorem:

if there exists $w^\star$ with $\|w^\star\|_2 = 1$, such that

$y_t(x_t^\top w^\star) \geq \gamma > 0, \forall t,$

then:

$$\sum_{t=0}^{\infty} \mathbf{1}(\hat{y}_t \neq y_t) \leq 1/\gamma^2$$

Q: Can this be applied to infinite dimension space ($d \rightarrow \infty$))

# Recap on Perceptron

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

**The Perceptron Alg:**

Initialize $w_0 = 0$

For $t = 0 \to \infty$

- User comes with feature $x_t$
- We make a prediction $\hat{y}_t =$ sign$(w_t^\top x_t)$
- User reveals the real label $y_t$
- We update $w_{t+1} = w_t + \mathbf{1}(\hat{y}_t \neq y_t) y_t x_t$

Theorem:

if there exists $w^\star$ with $\|w^\star\|_2 = 1$, such that $y_t(x_t^\top w^\star) \geq \gamma > 0, \forall t,$

then:

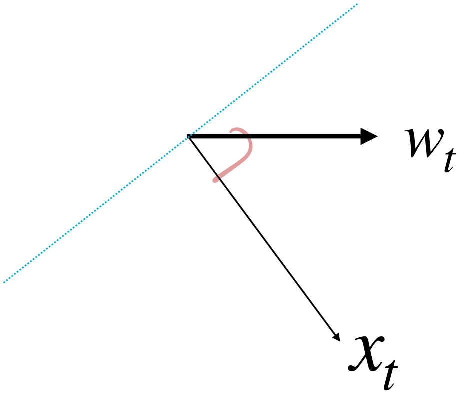
$$\sum_{t=0}^{\infty} \mathbf{1}(\hat{y}_t \neq y_t) \leq 1/\gamma^2$$

Q: Can this be applied to infinite dimension space ($d \to \infty$))

Yes! As long as margin exists!

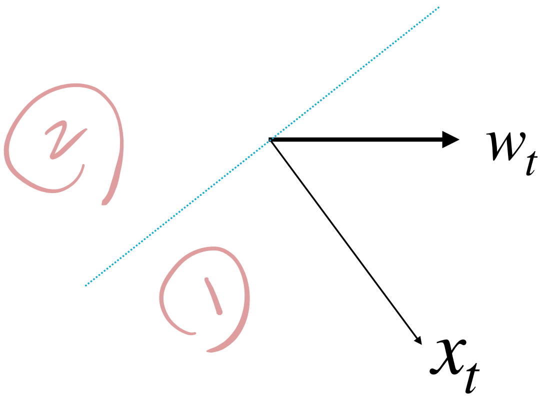
# Recap on Perceptron

When we make a mistake, i.e., $y_t(w_t^\top x_t) < 0$ (e.g., $y_t = -1$, $w_t^\top x_t > 0$)



# Recap on Perceptron

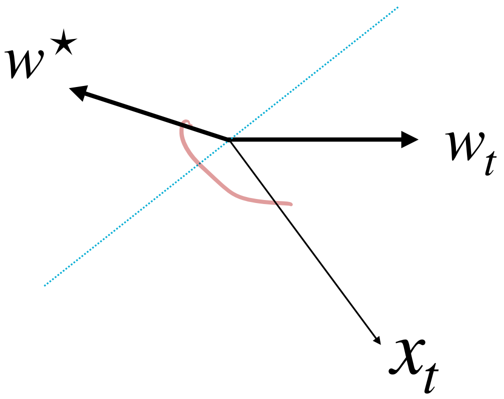
When we make a mistake, i.e., $y_t(w_t^\top x_t) < 0$ (e.g., $y_t = -1$, $w_t^\top x_t > 0$)




Q: What does $w^\star$ look like?

# Recap on Perceptron

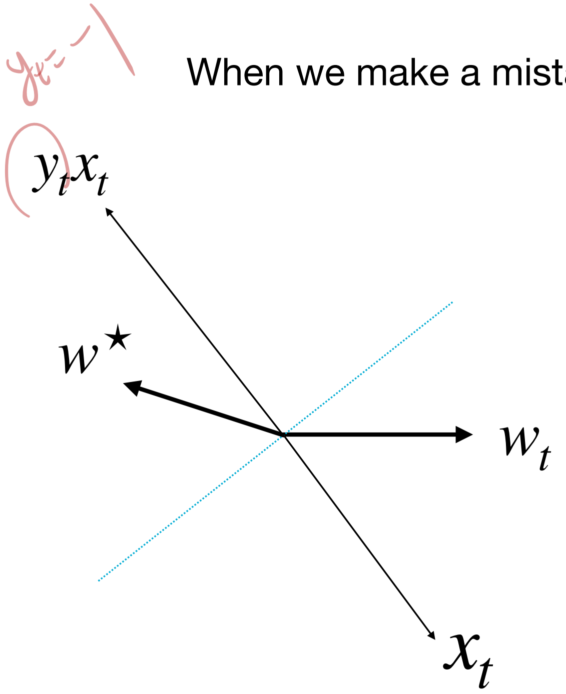
When we make a mistake, i.e., $y_t(w_t^\top x_t) < 0$ (e.g., $y_t = -1, w_t^\top x_t > 0$)




Q: What does $w^\star$ look like?

# Recap on Perceptron

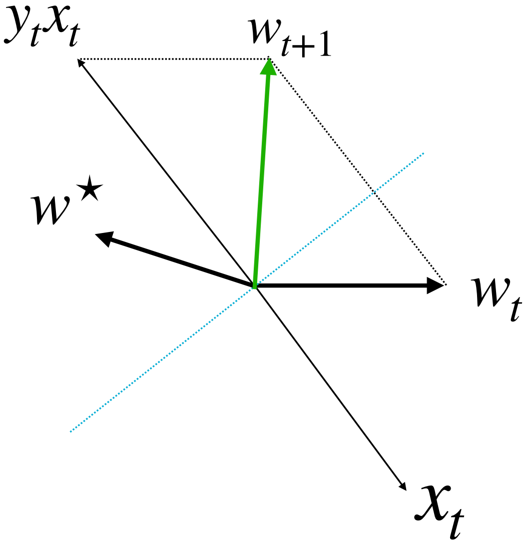
When we make a mistake, i.e., $y_t(w_t^\top x_t) < 0$ (e.g., $y_t = -1$, $w_t^\top x_t > 0$)




Q: What does $w^\star$ look like?

# Recap on Perceptron

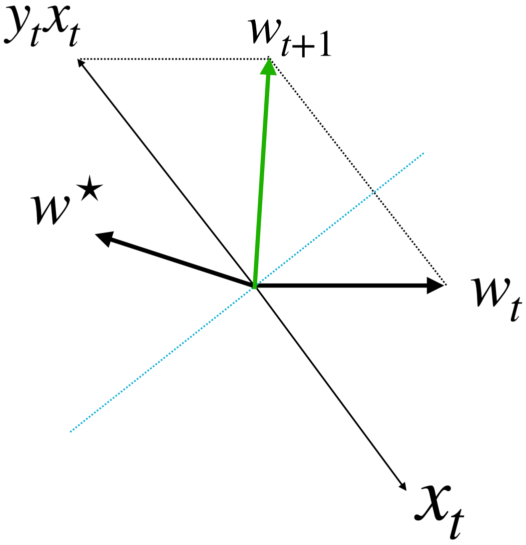
When we make a mistake, i.e., $y_t(w_t^\top x_t) < 0$ (e.g., $y_t = -1, w_t^\top x_t > 0$)




Q: What does $w^\star$ look like?

# Recap on Perceptron

When we make a mistake, i.e., $y_t(w_t^\top x_t) < 0$ (e.g., $y_t = -1, w_t^\top x_t > 0$)




We should track how the $\cos(\theta_t)$ is changing:

$$\cos(\theta_t) = \frac{w_t^\top w^\star}{\|w_t\|_2}$$

Q: What does $w^\star$ look like?

# Outline for today:

1. Maximum Likelihood estimation (MLE)

2. Maximum a posteriori probability (MAP)

3. Example: MLE and MAP for classification

# 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:

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

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

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

# 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 $\mathscr{D}$?

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

# of heads

$\hat{\theta} \to \theta^\star, n \to \infty$

# 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 $\mathscr{D}$?

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

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:

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

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

$D = \{+1, +1, -1, +1, -1\}$

↑ ↑ ↑ ↑ ↑

$\theta \quad \theta \quad 1-\theta \quad \theta \quad 1-\theta$

# 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)

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

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

$n_1 = \#(+1)$

# Maximum Likelihood Estimation

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

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

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

MLE Principle: Find $\theta$ 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 $\mathscr{D}$ (likelihood)?

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

MLE Principle: Find $\theta$ that **maximizes the likelihood** of the data:

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

likelihood

# Maximum Likelihood Estimation

We have n patients, we give each the new drug, we record whether or not it's effective

$\mathcal{D} = \{y_i\}_{i=1}^{n}, y_i \in \{-1,1\}$ ($y_1 = 1$ means effective on patient $i$, -1 otherwise)

tail head

MLE Principle: Find $\theta$ that maximizes the likelihood of the data:

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

# Maximum Likelihood Estimation

We have n patients, we give each the new drug, we record whether or not it's effective

$\mathcal{D} = \{y_i\}_{i=1}^n, y_i \in \{-1, 1\}$ ( $y_1 = 1$ means effective on patient $i$, -1 otherwise)

tail head

MLE Principle: Find $\theta$ that maximizes the likelihood of the data:

$n_1$ = # of +1 (heads)

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

# Maximum Likelihood Estimation

We have n patients, we give each the new drug, we record whether or not it's effective

$$\mathscr{D} = \{y_i\}_{i=1}^{n}, y_i \in \{-1,1\}$$ ($y_1 = 1$ means effective on patient $i$, -1 otherwise)

MLE Principle: Find $\theta$ that maximizes the likelihood of the data:

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

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

$\theta = 0.01$

$n_1 = 1000$

# Maximum Likelihood Estimation

We have n patients, we give each the new drug, we record whether or not it’s effective

$\mathscr{D} = \{y_i\}_{i=1}^n, y_i \in \{-1, 1\}$ ( $y_1 = 1$ means effective on patient $i$, -1 otherwise)

MLE Principle: Find $\theta$ that maximizes the likelihood of the data:

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

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

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

# Maximum Likelihood Estimation

We have n patients, we give each the new drug, we record whether or not it's effective

$\mathscr{D} = \{y_i\}_{i=1}^n, y_i \in \{-1,1\}$ ($y_1 = 1$ means effective on patient $i$, -1 otherwise)

MLE Principle: Find $\theta$ that maximizes the likelihood of the data:

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

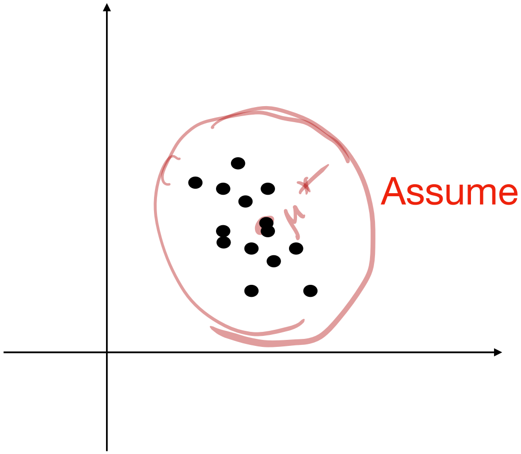
$$= \arg\max_{\theta \in [0,1]} \ln(\theta^{n_1}(1-\theta)^{n-n_1})$$

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

$$\frac{n_1}{\theta} - \frac{n-n_1}{1-\theta} = 0$$

solve for $\theta$

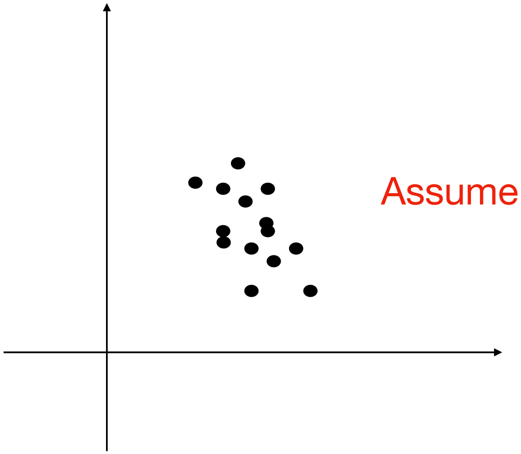
# Ex 2: Estimate the mean



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

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

# Ex 2: Estimate the mean



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

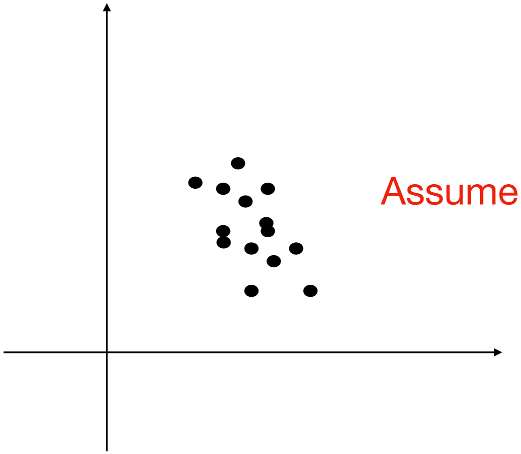
Assume data is from $\mathcal{N}(\mu^\star, I)$, want to estimate $\mu^\star$ from the data $\mathscr{D}$

Let's apply the MLE Principle:

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

likelihood of $x_i$ from $\mathcal{N}(\mu, I)$

# 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 $\mathcal{D}$

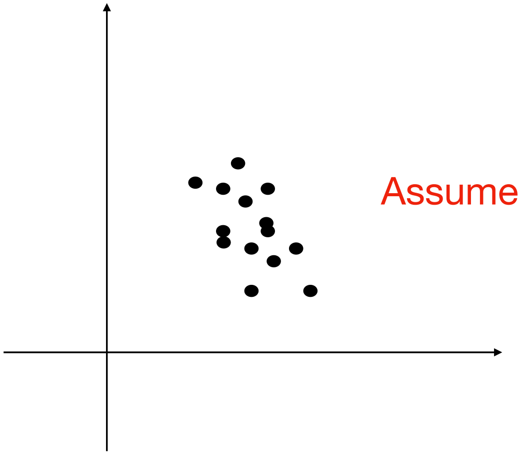
Let's apply the MLE Principle:

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:

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

# 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 $\mathscr{D}$

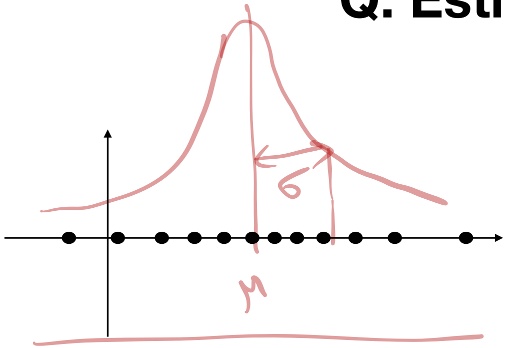
Let's apply the MLE Principle:

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:

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

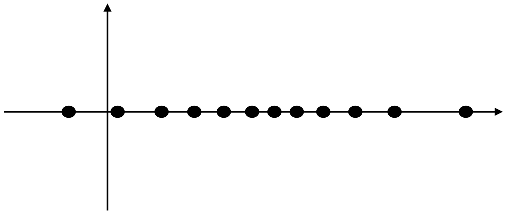
# Q: Estimate the mean and variance




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

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

# Q: Estimate the mean and variance



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

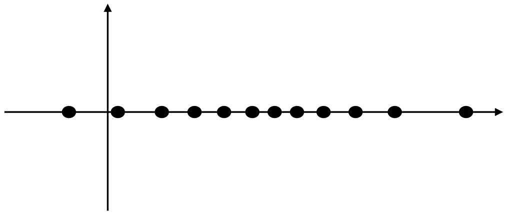
Assume data is from $\mathcal{N}(\mu^\star, \sigma^2)$, want to estimate $\mu^\star, \sigma$ from the data $\mathscr{D}$

Let's apply the MLE Principle:

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

likelihood of $x_i \sim N(\mu, \sigma^2)$

# 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:

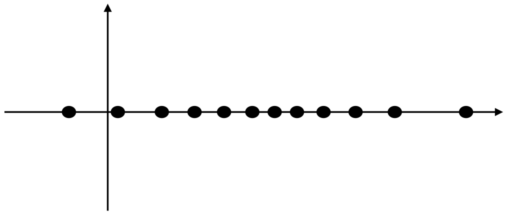
Step 1: $P(\mathscr{D} \mid \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2/\sigma^2\right)$

$$\ln P(D \mid \mu, \sigma) = \sum_{i=1}^{n}\left[\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \ln \exp\left(-\frac{1}{2}(x_i-\mu)^2/\sigma^2\right)\right]$$

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:

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

# Summary 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 $n \rightarrow \infty$

$$\left| \hat{\theta}_{mle} - \theta^{*} \right| \leq \frac{1}{\sqrt{n}}$$

# Summary 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

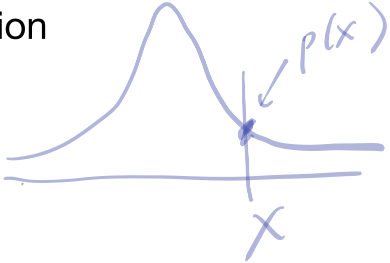
# Outline for today:

1. Maximum Likelihood estimation (MLE)

2. Maximum a Posteriori Probability (MAP)

$$\int p(x) = 1$$

3. Example: MLE and MAP for classification



# 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)

# 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:

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

prob of head

# 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-Dist

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

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

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

# 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)

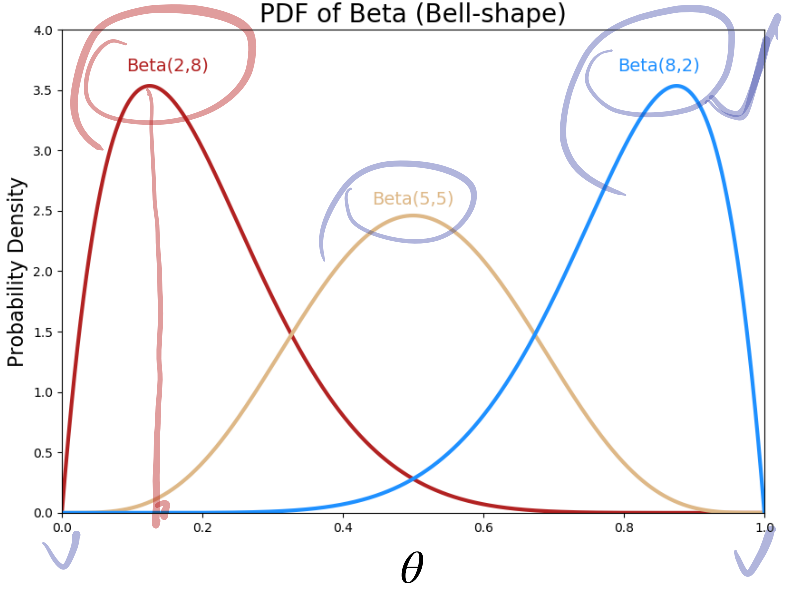
A Bayesian Statistician will treat the optimal parameter $\theta^\star$ being a random variable:

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

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 $\theta$

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

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

# The Posterior distribution over $\theta$

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

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

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

Using Bayes rule, we get:

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

# The Posterior distribution over $\theta$

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:

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

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

$P(\mathscr{D})$ is independent of $\theta$

# The Posterior distribution over $\theta$

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:

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

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

prior

Data likelihood

Posterior ∝ Prior × Likelihood

# The Posterior distribution over $\theta$

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

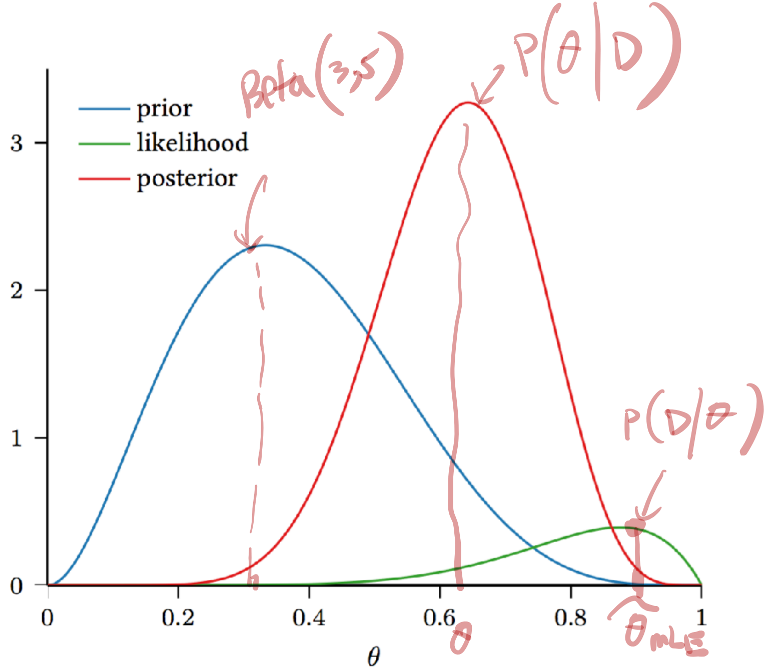
$$P(\theta \mid \mathscr{D})$$

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 $\theta$

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

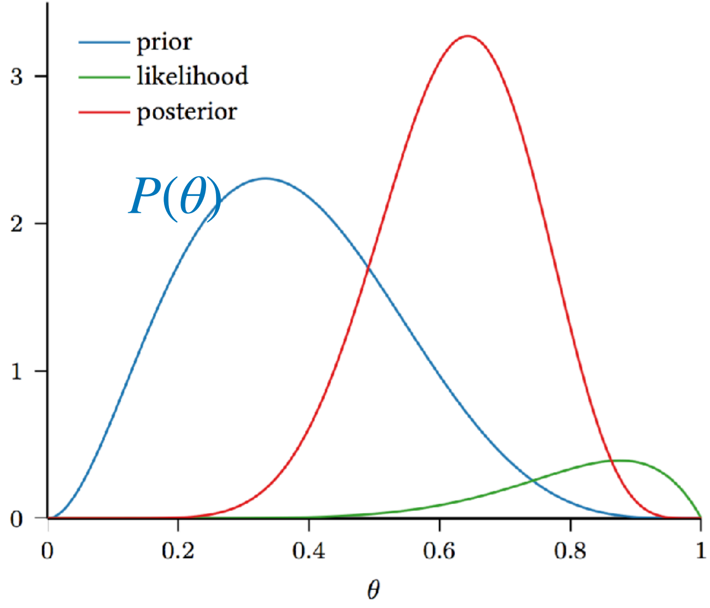
$$P(\theta \mid \mathscr{D})$$

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 $\propto$ Prior $\times$ Likelihood



# The Posterior distribution over $\theta$

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

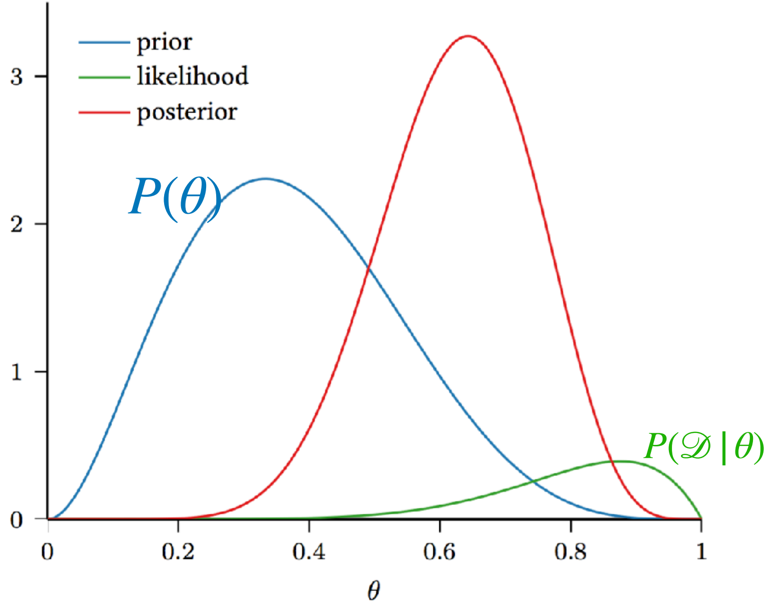
$$P(\theta \,|\, \mathscr{D})$$

Using Bayes rule, we get:

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

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

Posterior ∝ Prior × Likelihood



# The Posterior distribution over $\theta$

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

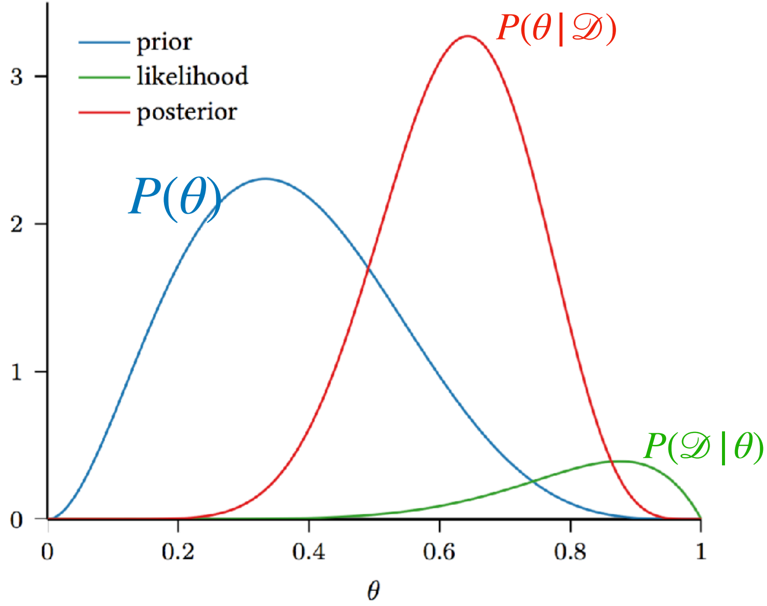
$$P(\theta \mid \mathscr{D})$$

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



# Maximum A Posteriori Probability estimation (MAP)

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

# Maximum A Posteriori Probability estimation (MAP)

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

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

# Maximum A Posteriori Probability estimation (MAP)

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

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

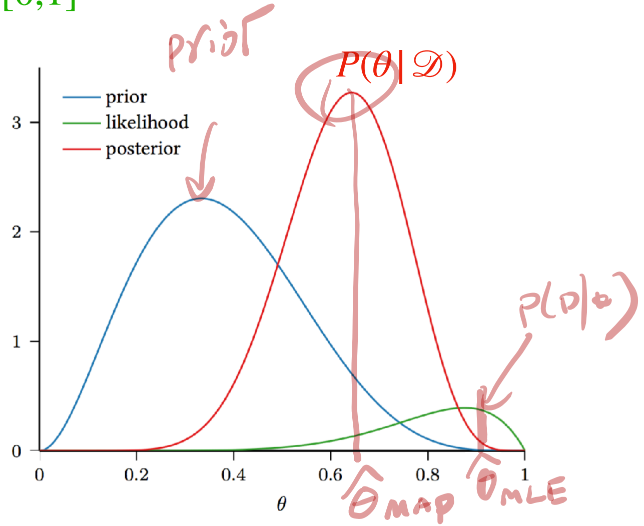
$$= \arg \max_{\theta \in [0,1]} \ln P(\theta) + \ln P(\mathscr{D} \mid \theta)$$

# 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

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

# 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 $P(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ Beta $\alpha > 0$ $\beta > 0$

# 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 $P(\mathcal{D} \mid \theta) = \theta^{n_1}(1-\theta)^{n-n_1}$ ✓ $n_1 = \#$ of $+1$ (heads)

# 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 $P(\theta \mid \mathcal{D}) \propto \theta^{n_1+\alpha-1}(1-\theta)^{n-n_1+\beta-1}$

# 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 $P(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$

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

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

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

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

$\left(MLE = n_1/n\right)$

# 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 $P(\theta \mid \mathcal{D}) \propto \theta^{n_1+\alpha-1}(1-\theta)^{n-n_1+\beta-1}$

\# of heads: $n_1 + \alpha - 1$

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

\# of Exp: $n + \alpha - 1 + \beta - 1$

$(\alpha - 1, \beta - 1)$ 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 $n \to \infty, \hat{\theta}_{map} = \frac{n_1 + \alpha - 1}{n + \alpha + \beta - 2} \to \frac{n_1}{n}$(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….

# Outline for today:

1. Maximum Likelihood estimation (MLE)

2. Maximum a posteriori probability (MAP)

3. Example: MLE and MAP for classification

# Binary Classification

$x \sim P(x) \quad y \sim P(y|x)$

Given labeled dataset $\{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$, we want to estimate $P(y\,|\,x)$

# Binary Classification

Given labeled dataset $\{x_i, y_i\}_{i=1}^{n}, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$, we want to estimate $P(y \mid x)$

Let us assume the ground truth has the form $P(y = 1 \mid x; \theta^\star) = \dfrac{\exp((\theta^\star)^\top x)}{1 + \exp((\theta^\star)^\top x)}$

# Binary Classification

Given labeled dataset $\{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$, we want to estimate $P(y \mid x)$

Let us assume the ground truth has the form $P(y = 1 \mid x; \theta^\star) = \dfrac{\exp((\theta^\star)^\top x)}{1 + \exp((\theta^\star)^\top x)}$

Goal: estimate $\theta^\star$ using $\mathscr{D}$

# Binary Classification

Given labeled dataset $\{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$, we want to estimate $P(y \mid x)$

Start with a parametric form $P(y = 1 \mid x; \theta) = \dfrac{\exp(\theta^\top x)}{1 + \exp(\theta^\top x)}$

# Binary Classification

Given labeled dataset $\{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$, we want to estimate $P(y \mid x)$

Start with a parametric form $P(y = 1 \mid x; \theta) = \dfrac{\exp(\theta^\top x)}{1 + \exp(\theta^\top x)}$

$P(a,b) = P(a) \cdot P(b|a)$

Using MLE:

$$\arg\max_\theta P(\mathscr{D} \mid \theta) = \arg\max_\theta \prod_{i=1}^n P(x_i, y_i \mid \theta)$$

$P(y_i | x_i; \theta) \cdot P(x_i)$

# Binary Classification

Given labeled dataset $\{x_i, y_i\}_{i=1}^{n}, x_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$, we want to estimate $P(y \mid x)$

Start with a parametric form $P(y = 1 \mid x; \theta) = \dfrac{\exp(\theta^\top x)}{1 + \exp(\theta^\top x)}$

Using MLE:

$$\arg\max_{\theta} P(\mathscr{D} \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{n} P(x_i, y_i \mid \theta)$$

$$= \arg\max \ln \prod_{i=1}^{n} P(y_i \mid x_i; \theta)$$

# Binary Classification

Given labeled dataset $\{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$, we want to estimate $P(y \mid x)$

Start with a parametric form $P(y = 1 \mid x; \theta) = \dfrac{\exp(\theta^\top x)}{1 + \exp(\theta^\top x)}$

Using MLE:

$$\arg\max_\theta P(\mathscr{D} \mid \theta) = \arg\max_\theta \prod_{i=1}^n P(x_i, y_i \mid \theta)$$

$$= \arg\max \ln \prod_{i=1}^n P(y_i \mid x_i; \theta)$$

$$= \arg\max_\theta \sum_i \ln P(y_i \mid x_i; \theta)$$

# Binary Classification

Given labeled dataset $\{x_i, y_i\}_{i=1}^{n}, x_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$, we want to estimate $P(y \mid x)$

Start with a parametric form $P(y = 1 \mid x; \theta) = \dfrac{\exp(\theta^\top x)}{1 + \exp(\theta^\top x)}$

Using MLE:

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

$$= \arg\max \ln \prod_{i=1}^{n} P(y_i \mid x_i; \theta)$$

$$= \arg\max_{\theta} \sum_{i} \ln P(y_i \mid x_i; \theta)$$

Using MAP:

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

prior

Data likelihood

# Binary Classification

Given labeled dataset $\{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$, we want to estimate $P(y \mid x)$

Start with a parametric form $P(y = 1 \mid x; \theta) = \dfrac{\exp(\theta^\top x)}{1 + \exp(\theta^\top x)}$

Using MLE:

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

$$= \arg\max \ln \prod_{i=1}^n P(y_i \mid x_i; \theta)$$

$$= \arg\max_\theta \sum_i \ln P(y_i \mid x_i; \theta)$$

Using MAP:

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

$$= \arg\max_\theta \ln(P(\theta) \prod_{i=1}^n P(y_i \mid x_i; \theta))$$

# Binary Classification

Given labeled dataset $\{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$, we want to estimate $P(y \mid x)$

Start with a parametric form $P(y=1 \mid x;\theta) = \dfrac{\exp(\theta^\top x)}{1+\exp(\theta^\top x)}$

Using MLE:

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

$$= \arg\max \ln \prod_{i=1}^n P(y_i \mid x_i;\theta)$$

$$= \arg\max_\theta \sum_i \ln P(y_i \mid x_i;\theta)$$

Using MAP:

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

$$= \arg\max_\theta \ln(P(\theta) \prod_{i=1}^n P(y_i \mid x_i;\theta))$$

$$= \arg\max_\theta \ln P(\theta) + \sum_i \ln P(y_i \mid x_i;\theta)$$

# Binary Classification

Given labeled dataset $\{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$, we want to estimate $P(y \mid x)$

Start with a parametric form $P(y=1 \mid x; \theta) = \dfrac{\exp(\theta^\top x)}{1+\exp(\theta^\top x)}$

Using MLE:

$$\arg\max_\theta P(\mathscr{D} \mid \theta) = \arg\max_\theta \prod_{i=1}^n P(x_i, y_i \mid \theta)$$

$$= \arg\max \ln \prod_{i=1}^n P(y_i \mid x_i; \theta)$$

$$= \arg\max_\theta \sum_i \ln P(y_i \mid x_i; \theta)$$

Using MAP:

$$\arg\max_\theta P(\theta \mid \mathscr{D}) = \arg\max_\theta P(\theta) \prod_{i=1}^n P(x_i, y_i \mid \theta)$$

$$= \arg\max_\theta \ln(P(\theta) \prod_{i=1}^n P(y_i \mid x_i; \theta))$$

$$= \arg\max_\theta \ln P(\theta) + \sum_i \ln P(y_i \mid x_i; \theta)$$

Independent of the data

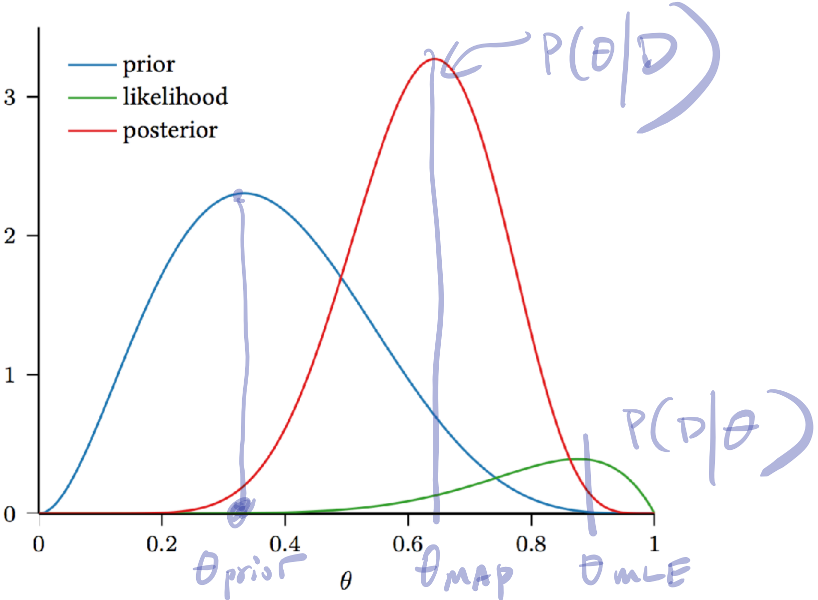
# Binary Classification

MLE:

$$\arg\max_{\theta} \sum_i \ln P(y_i \mid x_i; \theta)$$

MAP:

$$\arg\max_{\theta} \ln P(\theta) + \sum_i \ln P(y_i \mid x_i; \theta)$$



# 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

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

# 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

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