Maximum Likelihood Estimation
&
Maximum A Posteriori Probability Estimation
Announcements

1. P1 and HW1 are due today

2. HW2 will be out today

3. No office hour (wen) this Thursday
Recap on Perceptron

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

The Perceptron Alg:
Initialize $w_0 = 0$
For $t = 0 \to \infty$
Recap on Perceptron

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

**The Perceptron Algorithm:**

Initialize $w_0 = 0$

For $t = 0 \rightarrow \infty$

User comes with feature $x_t$
Recap on Perceptron

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

**The Perceptron Alg:**

Initialize $w_0 = 0$

For $t = 0 \rightarrow \infty$

- User comes with feature $x_t$
- We make a prediction $\hat{y}_t = \text{sign}(w_t^T x_t)$
Recap on Perceptron

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

The Perceptron Alg:

Initialize $w_0 = 0$

For $t = 0 \rightarrow \infty$

- User comes with feature $x_t$
- We make a prediction $\hat{y}_t = \text{sign}(w_t^T x_t)$
- User reveals the real label $y_t$
Recap on Perceptron

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

The Perceptron Alg:
Initialize \( w_0 = 0 \)
For \( t = 0 \to \infty \)

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

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

The Perceptron Alg:
Initialize $w_0 = 0$
For $t = 0 \rightarrow \infty$

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

Theorem:
if there exists $w^*$ with $\|w^*\|_2 = 1$, such that $y_t(x_t^T w^*) \geq \gamma > 0, \forall t$, then:
$$\sum_{t=0}^{\infty} 1(\hat{y}_t \neq y_t) \leq 1/\gamma^2$$
Recap on Perceptron

The Perceptron Alg:
Initialize $w_0 = 0$

For $t = 0 \rightarrow \infty$

User comes with feature $x_t$

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

User reveals the real label $y_t$

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

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

Theorem:
if there exists $w^*$ with $||w^*||_2 = 1$, such that

$y_t(x_t^T w^*) \geq \gamma > 0, \forall t,$

then:

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

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

Binary classifier: \( \text{sign}(w^Tx) \)

The Perceptron Alg:
Initialize \( w_0 = 0 \)
For \( t = 0 \to \infty \)

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

Theorem:
if there exists \( w^* \) with \( \|w^*\|_2 = 1 \), such that 
\( y_t(x_t^Tw^*) \geq \gamma > 0, \forall t \), 
then:
\[
\sum_{t=0}^{\infty} 1(\hat{y}_t \neq y_t) \leq 1/\gamma^2
\]
Recap on Perceptron

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

**The Perceptron Alg:**
- Initialize \( w_0 = 0 \)
- For \( t = 0 \rightarrow \infty \)
  - User comes with feature \( x_t \)
  - We make a prediction \( \hat{y}_t = \text{sign}(w_t^T x_t) \)
  - User reveals the real label \( y_t \)
  - We update \( w_{t+1} = w_t + 1(\hat{y}_t \neq y_t)y_tx_t \)

**Theorem:**
If there exists \( w^* \) with \( \|w^*\|_2 = 1 \), such that \( y_t(x_t^Tw^*) \geq \gamma > 0, \forall t \),
then:
\[
\sum_{t=0}^{\infty} 1(\hat{y}_t \neq y_t) \leq 1/\gamma^2
\]

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: $\text{sign}(w^T x)$

**The Perceptron Alg:**

Initialize $w_0 = 0$

For $t = 0 \to \infty$

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

**Theorem:**

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

then:

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

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

The Perceptron Alg:

Initialize \( w_0 = 0 \)

For \( t = 0 \rightarrow \infty \)

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

Theorem:

If there exists \( w^* \) with \( \|w^*\|_2 = 1 \), such that \( y_t(x_t^T w^*) \geq \gamma > 0, \forall t \), then:

\[
\sum_{t=0}^{\infty} 1(\hat{y}_t \neq y_t) \leq \frac{1}{\gamma^2}
\]

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

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

The Perceptron Alg:

Initialize $w_0 = 0$

For $t = 0 \rightarrow \infty$

User comes with feature $x_t$

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

User reveals the real label $y_t$

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

Theorem:

if there exists $w^* \text{ with } ||w^*||_2 = 1$, such that $y_t(x_t^T w^*) \geq \gamma > 0, \forall t,$

then:

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

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

Yes! As long as margin exists!
Recap on Perceptron

When we make a mistake, i.e., $y_t(w_t^T x_t) < 0$ (e.g., $y_t = -1$, $w_t^T x_t > 0$)
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 $w^*$ 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 $w^*$ look like?
Recap on Perceptron

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

Q: What does \( w^* \) look like?
Recap on Perceptron

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

Q: What does $w^*$ 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:

\[
\cos(\theta_t) = \frac{w_t^\top w^*}{\|w_t\|_2}
\]

Q: What does \( w^* \) 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:

$$\mathcal{D} = \{y_i\}_{i=1}^n, y_i \in \{-1,1\} \quad (y_i = 1 \text{ means head in } i\text{'s trial, } -1 \text{ means tail})$$
Ex 1: Estimating the probability of a coin flip

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

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

Q: assume $y_i \sim \text{Bernoulli}(\theta^*)$, how to estimate $\theta^*$ given $\mathcal{D}$?
Ex 1: Estimating the probability of a coin flip

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

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

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

$$\hat{\theta} \approx \frac{\sum_{i=1}^{n} 1(y_i = 1)}{n}$$
Ex 1: Estimating the probability of a coin flip

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

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

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

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

Let's make this rigorous!
Maximum Likelihood Estimation

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

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

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

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

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

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

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

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

\[ P(\mathcal{D} | \theta) = \theta^{n_1}(1 - \theta)^{n-n_1} \]
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)?

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

**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:

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

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

\[ P(D \mid \theta) = \theta^{n_1}(1 - \theta)^{n-n_1} \]

**MLE Principle**: Find \( \hat{\theta}_{mle} \) that maximizes the likelihood of the data:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$= \arg \max_{\theta \in [0,1]} n_1 \ln(\theta) + (n - n_1)\ln(1 - \theta) = \frac{n_1}{n}$$
Ex 2: Estimate the mean

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

Assume data is from \( \mathcal{N}(\mu^*, I) \), want to estimate \( \mu^* \) from the data \( D \)
Ex 2: Estimate the mean

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

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

Let’s apply the MLE Principle:

Step 1: 

\[ P(\mathcal{D} | \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d}} \exp \left( -\frac{1}{2} (x_i - \mu)^\top (x_i - \mu) \right) \]
Ex 2: Estimate the mean

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

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

Let’s apply the MLE Principle:

\[ P(\mathcal{D} | \mu) = \prod_{i=1}^n \frac{1}{\sqrt{(2\pi)^d}} \exp \left( -\frac{1}{2} (x_i - \mu)^\top (x_i - \mu) \right) \]

Step 1: 

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 \( \mathcal{N}(\mu^*, I) \), want to estimate \( \mu^* \) from the data \( \mathcal{D} \)

Let's apply the MLE Principle:

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

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

\[ \arg \max_{\mu} \sum_{i=1}^n - (x_i - \mu)^\top (x_i - \mu) \Rightarrow \hat{\mu}_{mle} = \frac{1}{n} \sum_{i=1}^n x_i \]
Q: Estimate the mean and variance

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

Assume data is from \( \mathcal{N}(\mu^*, \sigma^2) \), want to estimate \( \mu^*, \sigma \) from the data \( \mathcal{D} \)
Q: Estimate the mean and variance

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

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

Let’s apply the MLE Principle:

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

\[ P(\mathcal{D} | \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2} \right) \]

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

\[ \arg \max_{\mu, \sigma > 0} \sum_{i=1}^{n} \left( -\frac{(x_i - \mu)^2}{\sigma^2} - \ln(\sigma) \right) \]
Q: Estimate the mean and variance

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

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

Let’s apply the MLE Principle:

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

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 $\hat{\theta}_{mle} \to \theta^*$, as $n \to \infty$. 
Summary of MLE

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

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)

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:

\[ \mathcal{D} = \{y_i\}_{i=1}^n, y_i \in \{-1, 1\} \quad (y_i = 1 \text{ means head in } i\text{'s trial, -1 means tail}) \]
Ex: Estimating the probability of a coin flip

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

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

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

$$\theta^* \sim P(\theta)$$
Ex: Estimating the probability of a coin flip

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

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

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

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

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

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

where $Z = \int_{\theta\in[0,1]} \theta^{\alpha-1}(1 - \theta)^{\beta-1} d\theta$
Ex: Estimating the probability of a coin flip

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

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

A Bayesian Statistician will treat the optimal parameter \( \theta^* \) being a random variable:

\[ \theta^* \sim P(\theta) \]

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 $\mathcal{D} = \{y_i\}_{i=1}^n$, define posterior distribution:

$$P(\theta | \mathcal{D})$$
The Posterior distribution over $\theta$

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:

$$P(\theta | \mathcal{D}) = \frac{P(\theta)P(\mathcal{D} | \theta)}{P(\mathcal{D})}$$
The Posterior distribution over $\theta$

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:

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

$\propto P(\theta)P(\mathcal{D} | \theta)$
The Posterior distribution over $\theta$

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:

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

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

Posterior $\propto$ Prior $\times$ Likelihood
The Posterior distribution over $\theta$

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:

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

Posterior $\propto$ Prior $\times$ Likelihood
The Posterior distribution over $\theta$

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:

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

Posterior $\propto$ Prior $\times$ Likelihood
The Posterior distribution over $\theta$

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

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

Using Bayes rule, we get:

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

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

Posterior $\propto$ Prior $\times$ Likelihood
The Posterior distribution over $\theta$

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:

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

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

Posterior $\propto$ Prior $\times$ Likelihood
Maximum A Posteriori Probability estimation (MAP)

\[ P(\theta | \mathcal{D}) \propto P(\theta)P(\mathcal{D} | \theta) \]
Maximum A Posteriori Probability estimation (MAP)

\[ P(\theta \mid \mathcal{D}) \propto P(\theta)P(\mathcal{D} \mid \theta) \]

\[ \hat{\theta}_{map} = \arg \max_{\theta \in [0,1]} P(\theta \mid \mathcal{D}) = \arg \max_{\theta \in [0,1]} P(\theta)P(\mathcal{D} \mid \theta) \]
Maximum A Posteriori Probability estimation (MAP)

\[ P(\theta \mid \mathcal{D}) \propto P(\theta)P(\mathcal{D} \mid \theta) \]

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

\[ = \arg \max_{\theta \in [0,1]} \ln P(\theta) + \ln P(\mathcal{D} \mid \theta) \]
Maximum A Posteriori Probability estimation (MAP)

\[ P(\theta \mid \mathcal{D}) \propto P(\theta)P(\mathcal{D} \mid \theta) \]

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

\[ = \arg \max_{\theta \in [0,1]} \ln P(\theta) + \ln P(\mathcal{D} \mid \theta) \]
MAP for coin flip

$$\hat{\theta}_{map} = \arg \max_{\theta \in [0,1]} \ln(P(\theta)P(\mathcal{D} | \theta))$$
MAP for coin flip

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

Step 1: specify Prior \( P(\theta) \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1} \)
MAP for coin flip

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

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

Step 2: data likelihood \( P(\mathcal{D} | \theta) = \theta^{n_1}(1 - \theta)^{n-n_1} \)
MAP for coin flip

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

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

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

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

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

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

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

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

Step 4: Compute MAP \( \hat{\theta}_{map} = \frac{n_1 + \alpha - 1}{n + \alpha + \beta - 2} \)
MAP for coin flip

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

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

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

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

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

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

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

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 \( P(y \mid x) \).
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 \( P(y | x) \)

Let us assume the ground truth has the form
\[
P(y = 1 | x; \theta^*) = \frac{\exp((\theta^*)^T x)}{1 + \exp((\theta^*)^T x)}
\]
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

\[
P(y = 1 \mid x; \theta^*) = \frac{\exp((\theta^*)^\top x)}{1 + \exp((\theta^*)^\top x)}
\]

Goal: estimate \( \theta^* \) using \( \mathcal{D} \)
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 \( P(y = 1 \mid x; \theta) = \frac{\exp(\theta^T x)}{1 + \exp(\theta^T x)} \)
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 \( P(y = 1 \mid x; \theta) = \frac{\exp(\theta^T x)}{1 + \exp(\theta^T x)} \)

Using MLE:

\[
\arg \max_{\theta} P(\mathcal{D} \mid \theta) = \arg \max_{\theta} \prod_{i=1}^{n} P(x_i, y_i \mid \theta)
\]
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 \( P(y = 1 \mid x; \theta) = \frac{\exp(\theta^\top x)}{1 + \exp(\theta^\top x)} \)

Using MLE:

\[
\arg \max_{\theta} P(\mathcal{D} \mid \theta) = \arg \max_{\theta} \prod_{i=1}^n P(x_i, y_i \mid \theta) \\
= \arg \max_{\theta} \ln \prod_{i=1}^n P(y_i \mid x_i; \theta)
\]
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 \( P(y|x) \)

Start with a parametric form \( P(y = 1|x; \theta) = \frac{\exp(\theta^T x)}{1 + \exp(\theta^T x)} \)

Using MLE:

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

\[
= \arg\max_{\theta} \ln \prod_{i=1}^n P(y_i | x_i; \theta)
\]

\[
= \arg\max_{\theta} \sum_{i} \ln P(y_i | x_i; \theta)
\]
**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 \( P(y = 1 \mid x; \theta) = \frac{\exp(\theta^\top x)}{1 + \exp(\theta^\top x)} \)

Using MLE:

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

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

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

Using MAP:

\[
\arg\max_\theta P(\theta \mid \mathcal{D}) = \arg\max_\theta P(\theta) \prod_{i=1}^n P(x_i, y_i \mid \theta)
\]
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 \( P(y | x) \)

Start with a parametric form \( P(y = 1 | x; \theta) = \frac{\exp(\theta^T x)}{1 + \exp(\theta^T x)} \)

Using MLE:
\[
\arg \max_{\theta} P(\mathcal{D} | \theta) = \arg \max_{\theta} \prod_{i=1}^n P(x_i, y_i | \theta) \\
= \arg \max_{\theta} \ln \prod_{i=1}^n P(y_i | x_i; \theta) \\
= \arg \max_{\theta} \sum_{i} \ln P(y_i | x_i; \theta)
\]

Using MAP:
\[
\arg \max_{\theta} P(\theta | \mathcal{D}) = \arg \max_{\theta} P(\theta) \prod_{i=1}^n P(x_i, y_i | \theta) \\
= \arg \max_{\theta} \ln(P(\theta) \prod_{i=1}^n P(y_i | x_i; \theta))
\]
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 \( P(y = 1 \mid x; \theta) = \frac{\exp(\theta^\top x)}{1 + \exp(\theta^\top x)} \).

Using MLE:

\[
\begin{align*}
\arg\max_{\theta} P(\mathcal{D} \mid \theta) &= \arg\max_{\theta} \prod_{i=1}^n P(x_i, y_i \mid \theta) \\
&= \arg\max_{\theta} \ln \prod_{i=1}^n P(y_i \mid x_i; \theta) \\
&= \arg\max_{\theta} \sum_{i} \ln P(y_i \mid x_i; \theta)
\end{align*}
\]

Using MAP:

\[
\begin{align*}
\arg\max_{\theta} P(\theta \mid \mathcal{D}) &= \arg\max_{\theta} P(\theta) \prod_{i=1}^n P(x_i, y_i \mid \theta) \\
&= \arg\max_{\theta} \ln(P(\theta) \prod_{i=1}^n P(y_i \mid x_i; \theta)) \\
&= \arg\max_{\theta} \ln P(\theta) + \sum_{i} \ln P(y_i \mid x_i; \theta)
\end{align*}
\]
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
\[
P(y = 1 \mid x; \theta) = \frac{\exp(\theta^\top x)}{1 + \exp(\theta^\top x)}
\]

Using MLE:
\[
\arg \max_{\theta} P(\mathcal{D} \mid \theta) = \arg \max_{\theta} \prod_{i=1}^n P(x_i, y_i \mid \theta)
= \arg \max_{\theta} \ln \prod_{i=1}^n P(y_i \mid x_i; \theta)
= \arg \max_{\theta} \sum_i \ln P(y_i \mid x_i; \theta)
\]

Using MAP:
\[
\arg \max_{\theta} P(\theta \mid \mathcal{D}) = \arg \max_{\theta} P(\theta) \prod_{i=1}^n P(x_i, y_i \mid \theta)
= \arg \max_{\theta} \ln (P(\theta) \prod_{i=1}^n P(y_i \mid x_i; \theta))
= \arg \max_{\theta} \ln P(\theta) + \sum_i \ln P(y_i \mid x_i; \theta)
\]

Independent of the data
Binary Classification

**MLE:**
\[
\arg \max_\theta \sum_i \ln P(y_i | x_i; \theta)
\]

**MAP:**
\[
\arg \max_\theta \ln P(\theta) + \sum_i \ln P(y_i | x_i; \theta)
\]
Summary for today

1 MLE (frequentist perspective):

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

$$\arg \max_{\theta} P(\mathcal{D} | \theta)$$
Summary for today

1 MLE (frequentist perspective):

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

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

2 MAP (Bayesian perspective):

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

1 MLE (frequentist perspective):
The ground truth $\theta^*$ is unknown but fixed; we search for the parameter that makes the data as likely as possible

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

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

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