Foundations of Machine Learning (ECE 5984)

- Probabilistic Perspective -

Eunbyung Park

Assistant Professor

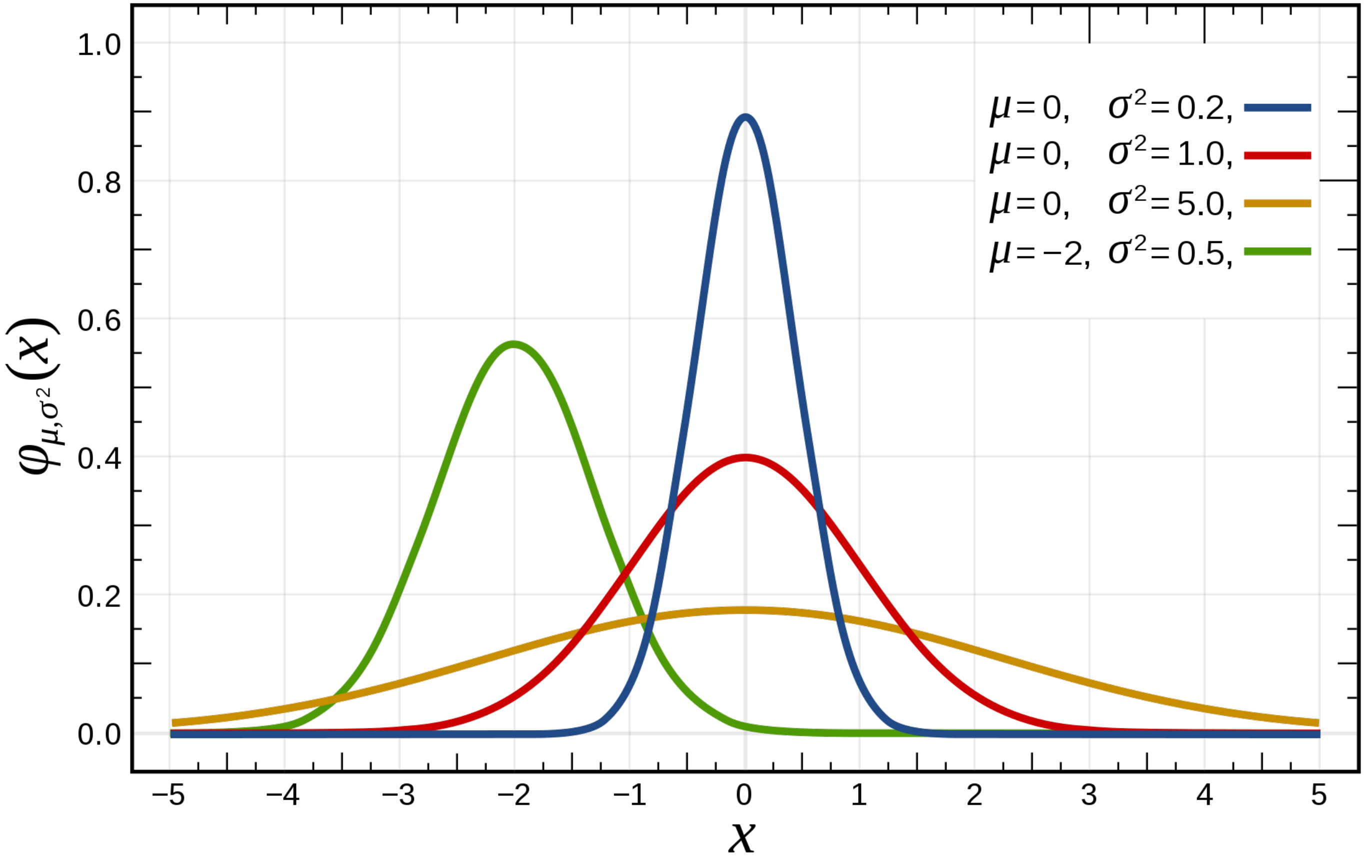
School of Electronic and Electrical Engineering

Eunbyung Park (silverbottlep.github.io)

Gaussian Distribution

- Normal distribution
- Widely used model for the distribution of continuous variable

$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

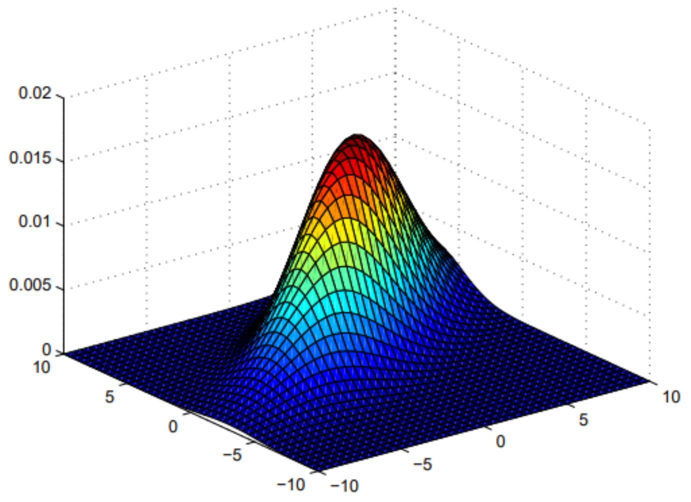


Multivariate Gaussian Distribution

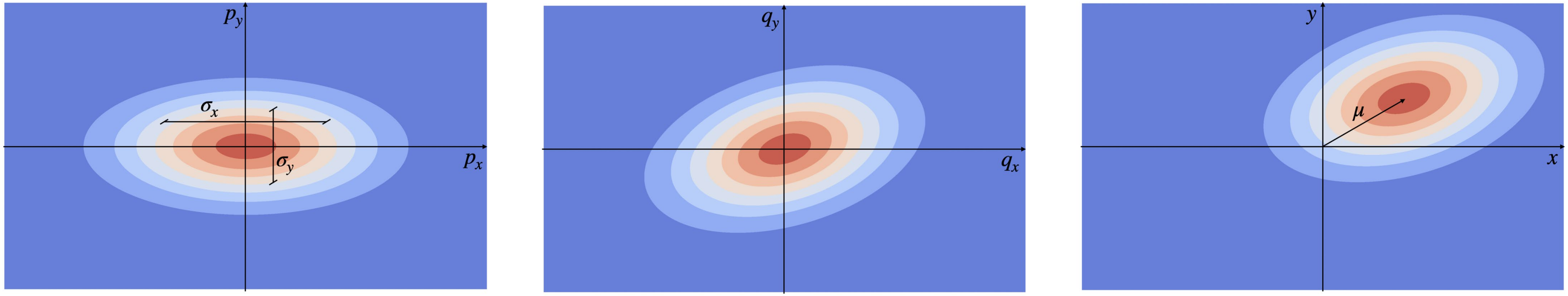
$x, \mu \in \mathbb{R}^d$

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

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu)\right)$$



Multivariate Gaussian Distribution



2D Multivariate Gaussian Distribution

$$x, \mu \in \mathbb{R}^2$$

$$\Sigma \in \mathbb{R}^{2 \times 2}$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

$$p(x; \mu, \Sigma) =$$

2D Multivariate Gaussian Distribution (Diagonal)

$$x, \mu \in \mathbb{R}^2 \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

$$\Sigma \in \mathbb{R}^{2\times 2}$$

$$p(x;\mu,\Sigma) = \frac{1}{2\pi \begin{vmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{vmatrix}^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^\top \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right)$$

$$= \frac{1}{2\pi (\sigma_1^2 \sigma_2^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^\top \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix}^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right)$$

2D Multivariate Gaussian Distribution (Diagonal)

$$x, \mu \in \mathbb{R}^2 \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

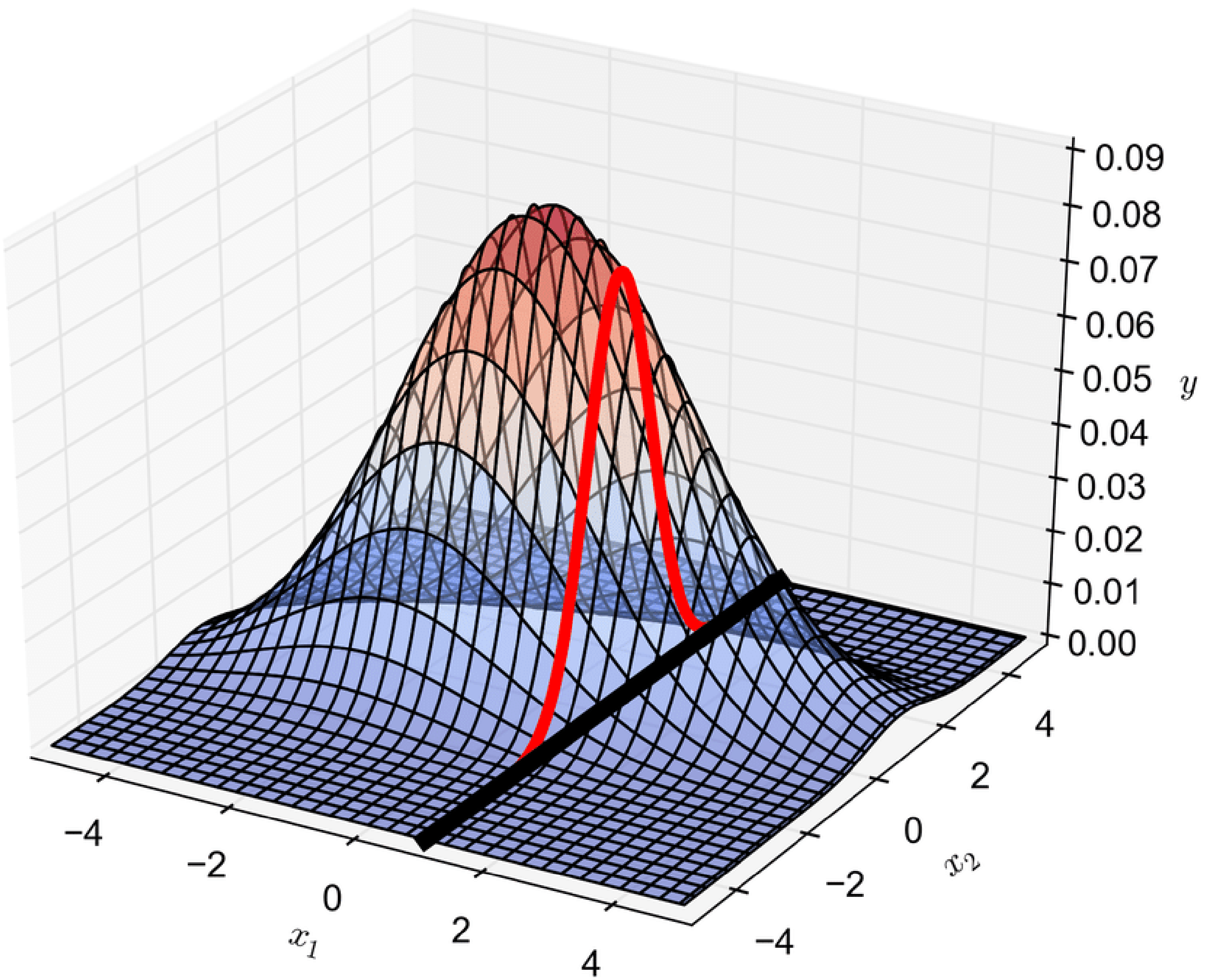
$$\Sigma \in \mathbb{R}^{2\times 2}$$

$$p(x; \mu, \Sigma) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^\top \begin{bmatrix} \frac{1}{\sigma_1^2}(x_1 - \mu_1) \\ \frac{1}{\sigma_2^2}(x_2 - \mu_2) \end{bmatrix} \right)$$

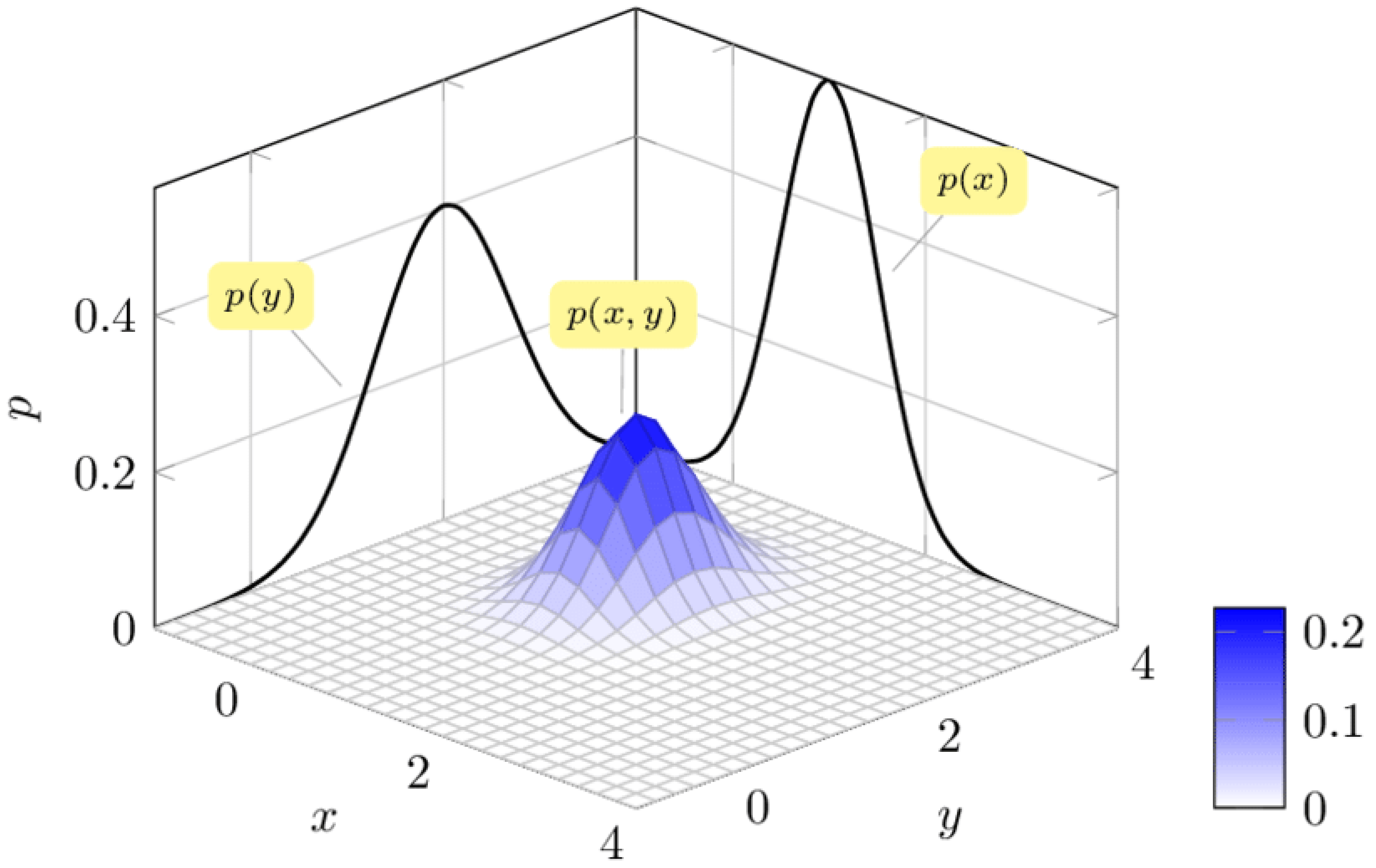
$$= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right)$$

$$= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 \right) \exp\left(-\frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right)$$

Conditional Gaussian



Marginal Gaussian



Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE)

Probability

A (probability density/mass) function of the data given the fixed parameters

$$p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

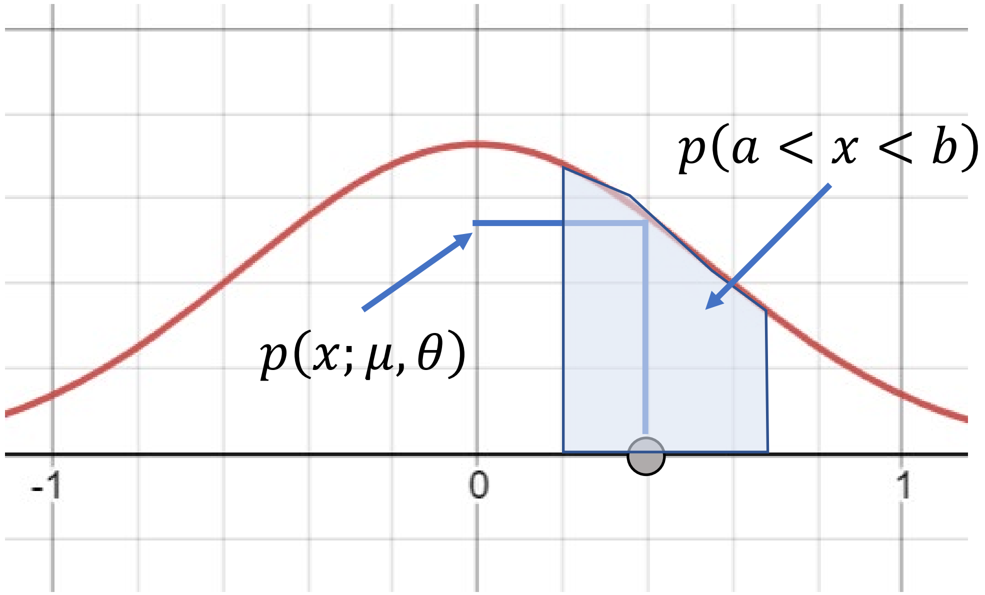
Likelihood

A (probability density /mass) function of parameters given the data

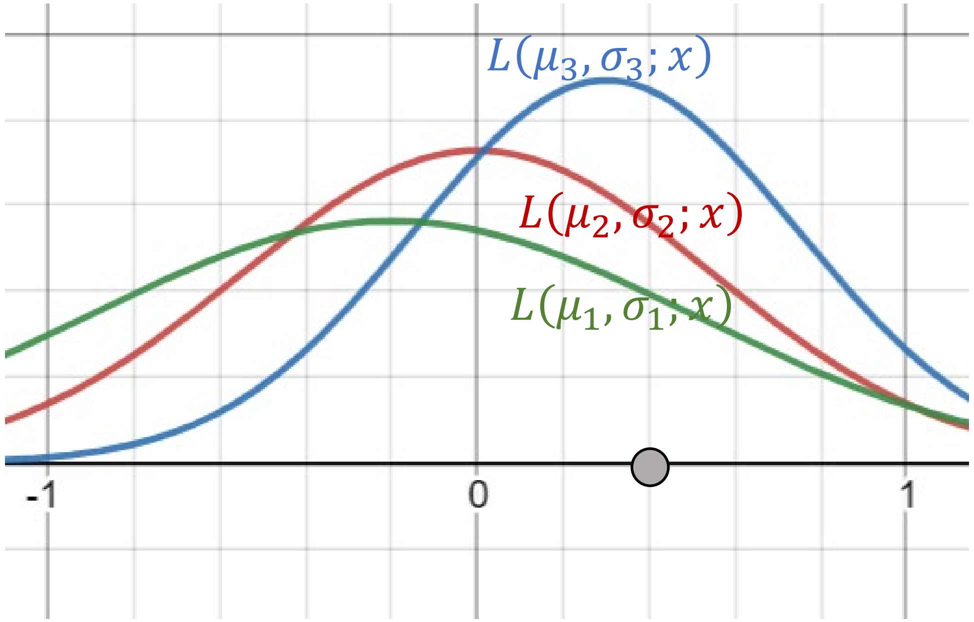
$$L(\mu, \sigma; x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Maximum Likelihood Estimation (MLE)

Probability Density Function



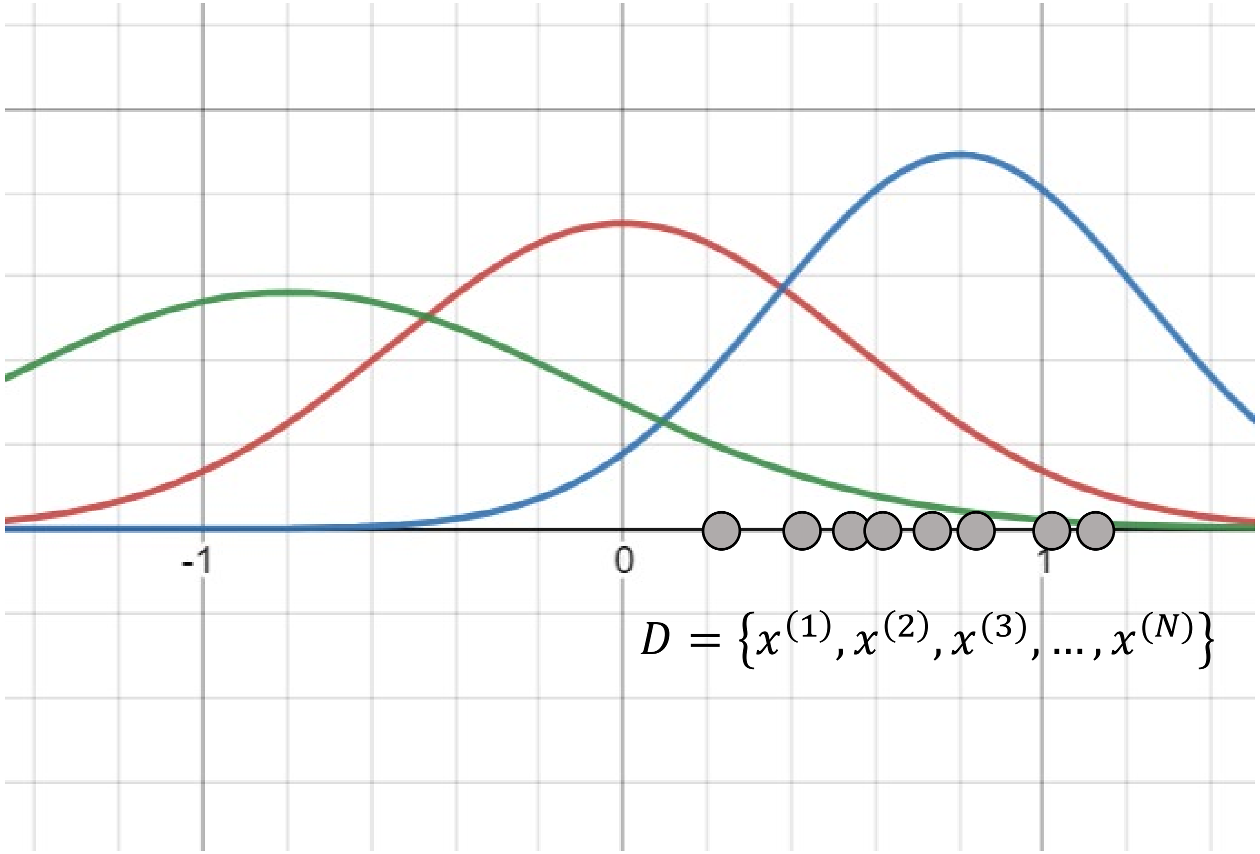
Likelihood



Maximum Likelihood Estimation (MLE)

- Finding the parameters that maximize the probability (density/mass) function

$$\arg\max_{\theta} L(\theta; D)$$



$D = \{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(N)}\}$

Maximum Likelihood Estimation (MLE)

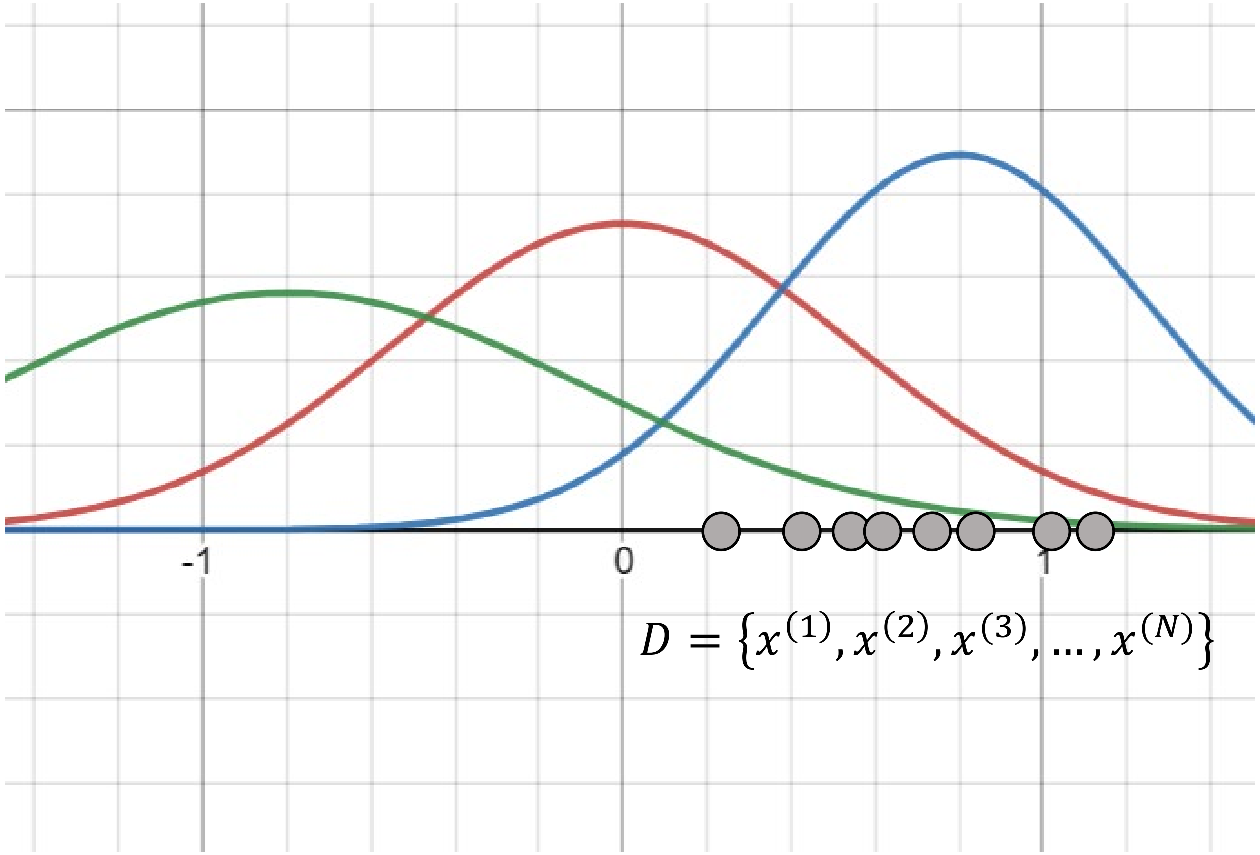
- Finding the parameters that maximize the probability (density/mass) function

I.I.D assumption

$$\arg\max_{\theta} L(\theta; D) = \arg\max_{\theta} \prod_{i=1}^{N} p(x^{(i)}; \theta)$$

$$= \arg\max_{\theta} \log \prod_{i=1}^{N} p(x^{(i)}; \theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{N} \log p(x^{(i)}; \theta)$$



$D = \{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(N)}\}$

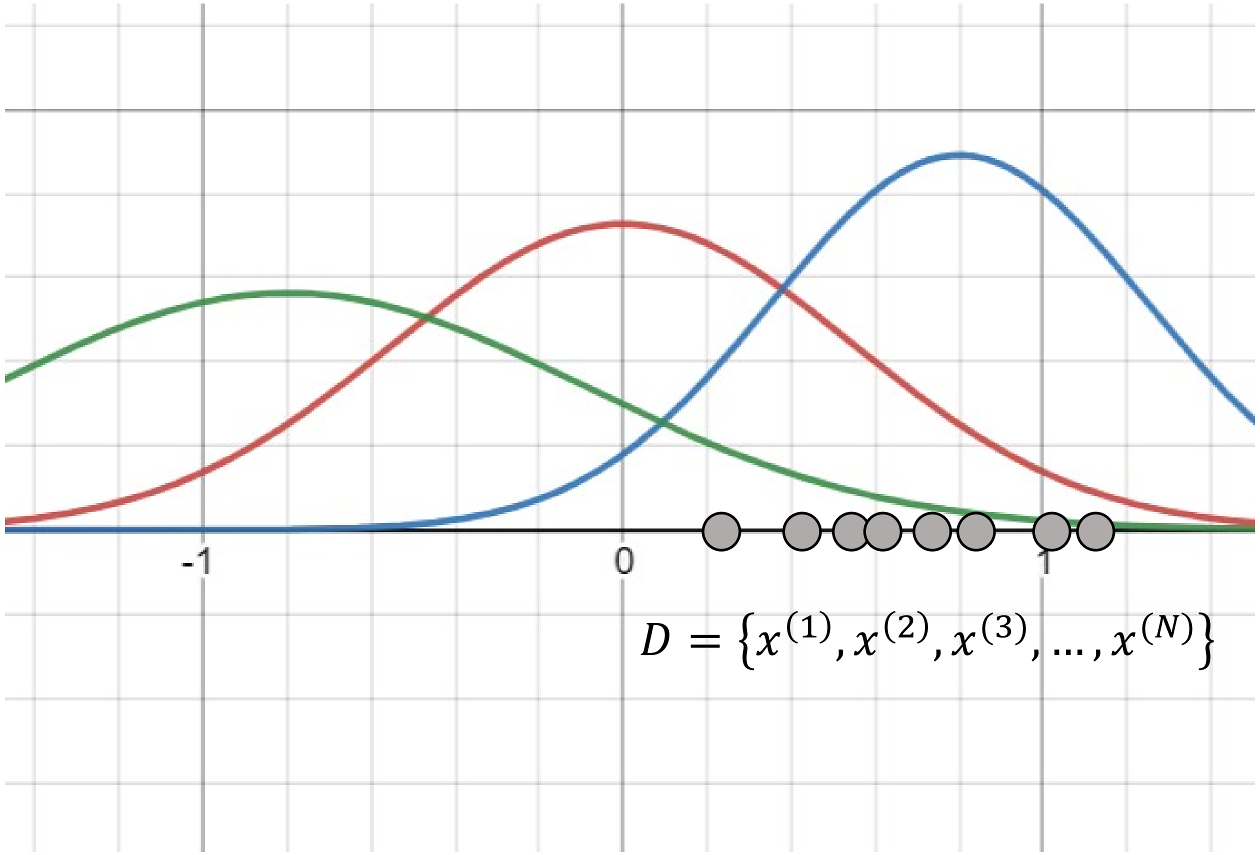
Maximum Likelihood Estimation (MLE)

- Finding the parameters that maximize the probability (density/mass) function

$$\arg\max_{\theta} \sum_{i=1}^{N} \log p(x^{(i)}; \theta)$$

$$= \arg\max_{\mu,\sigma} \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x^{(i)}-\mu\right)^2}{2\sigma^2}}\right)$$

$$= \arg\max_{\mu,\sigma} \sum_{i=1}^{N} -\frac{\left(x^{(i)}-\mu\right)^2}{2\sigma^2} - \log\left(\sqrt{2\pi\sigma^2}\right)$$



$$D = \left\{x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(N)}\right\}$$

Maximum Likelihood Estimation (MLE)

- Finding the parameters that maximize the probability (density/mass) function

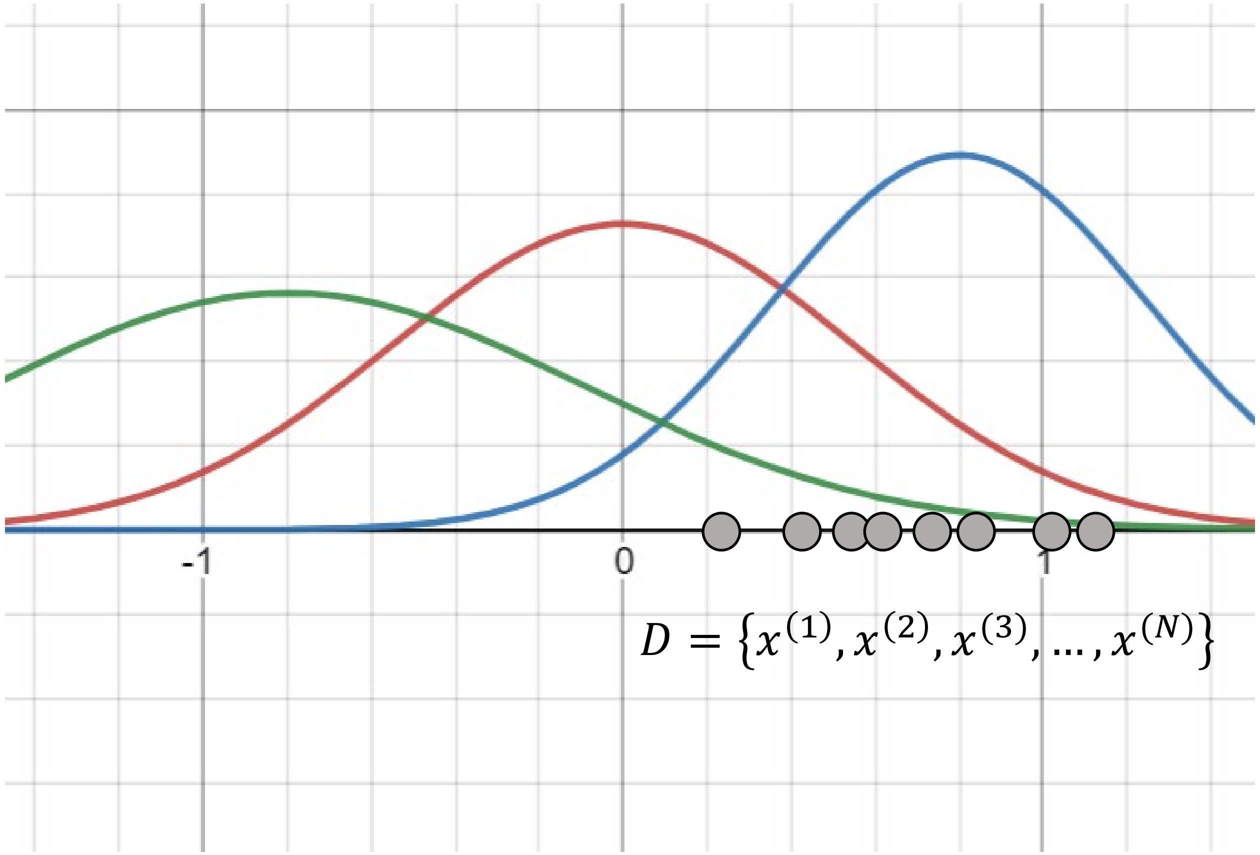
$$\arg\max_{\mu} \sum_{i=1}^{N} -\frac{\left(x^{(i)}-\mu\right)^2}{2\sigma^2} - \log\left(\sqrt{2\pi\sigma^2}\right)$$

$$\frac{\partial}{\partial\mu} \sum_{i=1}^{N} -\frac{\left(x^{(i)}-\mu\right)^2}{2\sigma^2} - \log\left(\sqrt{2\pi\sigma^2}\right)$$

$$= \sum_{i=1}^{N} \frac{\left(x^{(i)}-\mu\right)}{\sigma^2} = 0$$

$$\sum_{i=1}^{N} x^{(i)} - N\mu = 0$$

$$\mu^* = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$$



MLE - Linear Regression

MLE for Linear Regression

- Finding the parameters that the errors are distributed from $N(0, \sigma^2)$

Assumption1: $\epsilon = y - w^\top x, \;\; \epsilon \sim N(0, \sigma^2)$

Assumption2: I.I.D

$$y = w^\top x + \epsilon \qquad\qquad \epsilon \sim N(0, \sigma^2)$$

"We are going to predict y except for the white noise"

MLE for Linear Regression

- Finding the parameters that the errors are distributed from $N(0, \sigma^2)$

Assumption1: $\epsilon = y - w^\top x, \quad \epsilon \sim N(0, \sigma^2)$

Assumption2: I.I.D

$$L(w) = \prod_{i=1}^{N} p(y^{(i)}|x^{(i)}; w) = \sum_{i=1}^{N} \log p(y^{(i)}|x^{(i)}; w)$$

$$= \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(\left(y^{(i)} - w^\top x^{(i)}\right) - 0\right)^2}{2\sigma^2}}\right)$$

$\epsilon \sim N(0, \sigma^2)$

MLE for Linear Regression

- Finding the parameters that the errors are distributed from $N(0, \sigma^2)$

Assumption1: $\epsilon = y - w^\top x, \;\; \epsilon \sim N(0, \sigma^2)$

Assumption2: I.I.D

$$L(w) = \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(\left(y^{(i)} - w^\top x^{(i)}\right) - 0\right)^2}{2\sigma^2}}\right)$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y^{(i)} - w^\top x^{(i)}\right)^2 - N\log\left(\sqrt{2\pi\sigma^2}\right)$$

$\sigma = 1$, we recover MSE Loss

MLE for Linear Regression

- Finding the parameters that maximize 'conditional likelihood'