# ECE595 / STAT598: Machine Learning I

## Lecture 11.1: Maximum-Likelihood Estimation - Basic Principles

Spring 2020

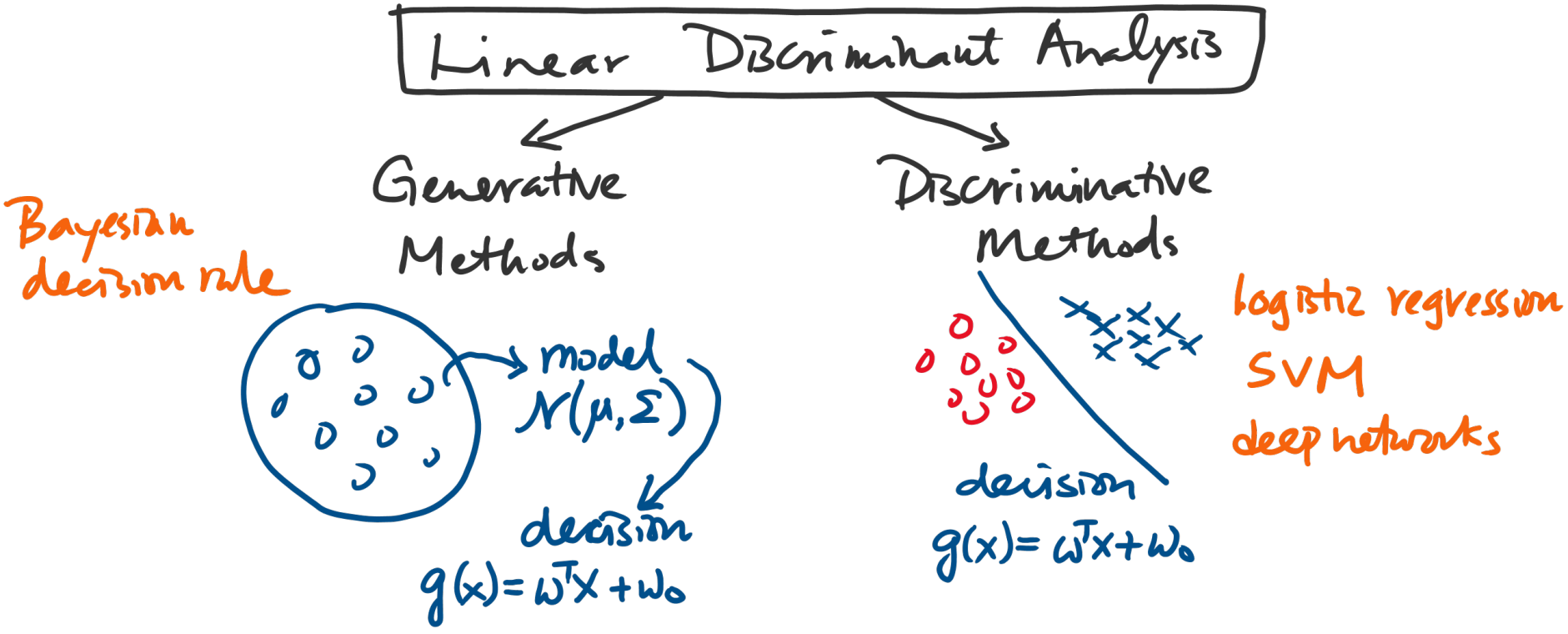
Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE
UNIVERSITY

©Stanley Chan 2020. All Rights Reserved.

# Overview




- In linear discriminant analysis (LDA), there are generally two types of approaches
- **Generative approach**: Estimate model, then define the classifier
- **Discriminative approach**: Directly define the classifier

©Stanley Chan 2020. All Rights Reserved.

# Outline

**Generative Approaches**

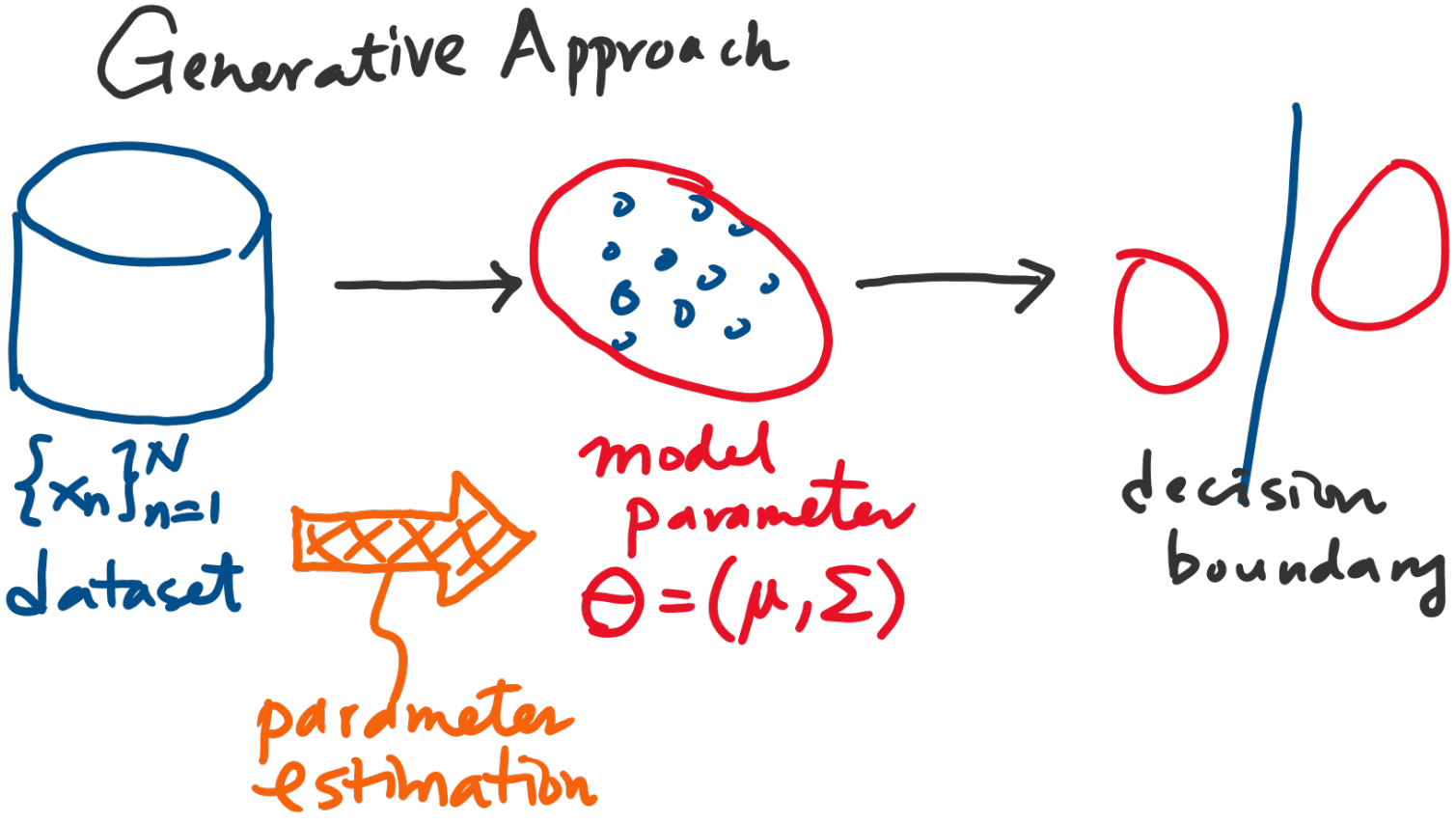
- Lecture 9 Bayesian Decision Rules
- Lecture 10 Evaluating Performance
- Lecture 11 Parameter Estimation
- Lecture 12 Bayesian Prior
- Lecture 13 Connecting Bayesian and Linear Regression

**Today's Lecture**

- Basic Principles
  - Likelihood Function
  - Maximum Likelihood Estimate
  - 1D Illustration
  - Gaussian Distributions
- Examples
  - Non-Gaussian Distributions
  - Biased and Unbiased Estimators
  - From MLE to MAP

©Stanley Chan 2020. All Rights Reserved.

# What is Parameter Estimation?

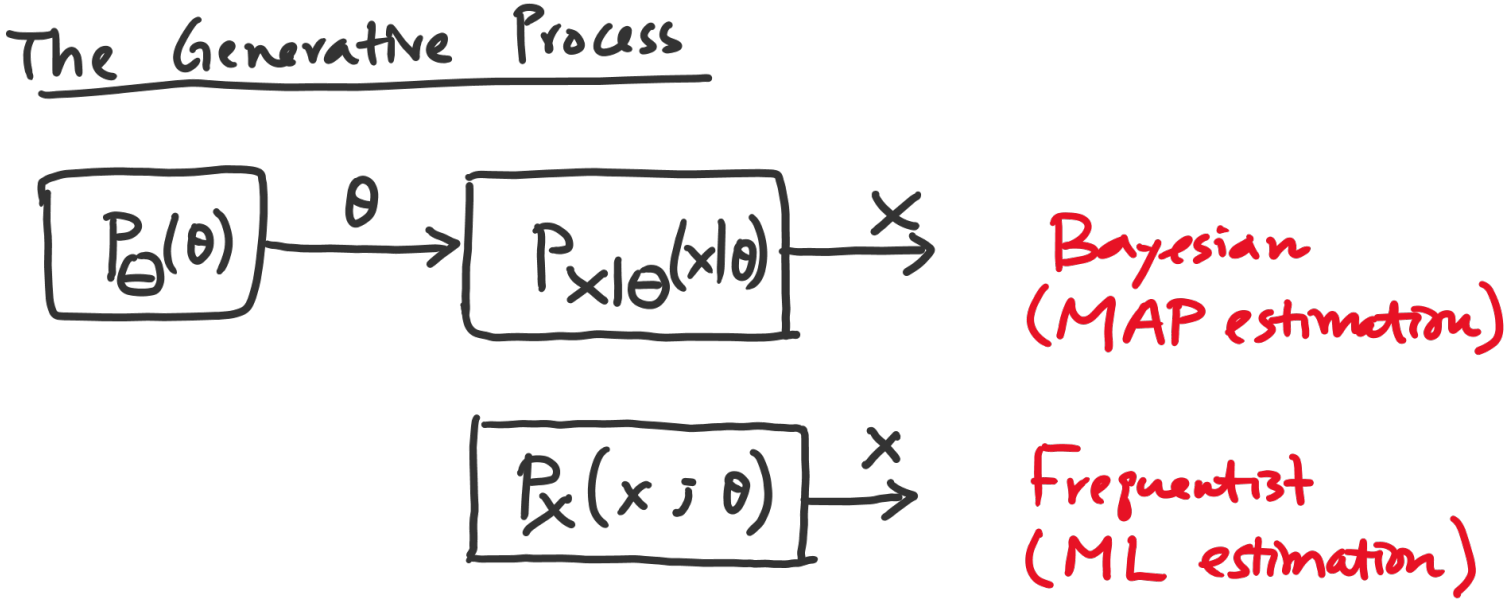



- The goal of parameter estimation is to determine $\boldsymbol{\Theta} = (\mu, \boldsymbol{\Sigma})$ from dataset
- This is *the step* where you use data

©Stanley Chan 2020. All Rights Reserved.

# MLE and MAP

There are two typical ways of estimating parameters.




- Maximum-likelihood estimation (MLE): $\boldsymbol{\theta}$ is deterministic.
- Maximum-a-posteriori estimation (MAP): $\boldsymbol{\theta}$ is random and has a prior distribution.

©Stanley Chan 2020. All Rights Reserved.

# Maximum Likelihood Estimation

Given the dataset $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^N$, how to estimate the model parameters?

- We are going to use Gaussian as an illustration.
- Denote $\boldsymbol{\theta}$ as the model parameter.
- In Gaussian

$$\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$$

- The likelihood for one data point $\boldsymbol{x}_n$ is

$$p(\boldsymbol{x}_n \,|\, \overbrace{\boldsymbol{\theta}}^{=\{\boldsymbol{\mu},\boldsymbol{\Sigma}\}}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu})\right\}$$

- $\boldsymbol{\theta}$ is a deterministic quantity, not a random variable.
- $\boldsymbol{\theta}$ does not have a distribution.
- $\boldsymbol{\theta}$ is fixed but unknown.

©Stanley Chan 2020. All Rights Reserved.

# Likelihood for the Entire Dataset

- Likelihood for the entire dataset $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ is

$$p(\mathcal{D} \,|\, \boldsymbol{\theta}) = \prod_{n=1}^{N} \left\{ \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right\} \right\}$$

$$= \left( \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \right)^N \exp\left\{ \sum_{n=1}^{N} -\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right\}$$

- The Negative Log-Likelihood is

$$-\log p(\mathcal{D} \,|\, \boldsymbol{\theta}) = \frac{N}{2} \log |\boldsymbol{\Sigma}| + \frac{N}{2} \log (2\pi)^d + \sum_{n=1}^{N} \left\{ \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right\}.$$

©Stanley Chan 2020. All Rights Reserved.

# Maximum Likelihood Estimation

- Goal: Find $\boldsymbol{\theta}$ that maximizes the likelihood:

$$\begin{aligned}
\widehat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\text{argmax}}\ p(\mathcal{D}\,|\,\boldsymbol{\theta}) \\
&= \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{n=1}^{N} \left\{ \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right\} \right\} \\
&= \underset{\boldsymbol{\theta}}{\text{argmin}} - \log(\cdots) \\
&= \underset{\boldsymbol{\theta}}{\text{argmin}}\ \frac{N}{2} \log |\boldsymbol{\Sigma}| + \frac{N}{2} \log(2\pi)^d \\
&\qquad + \sum_{n=1}^{N} \left\{ \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right\}.
\end{aligned}$$

- This optimization is called the **maximum likelihood estimation** (MLE).

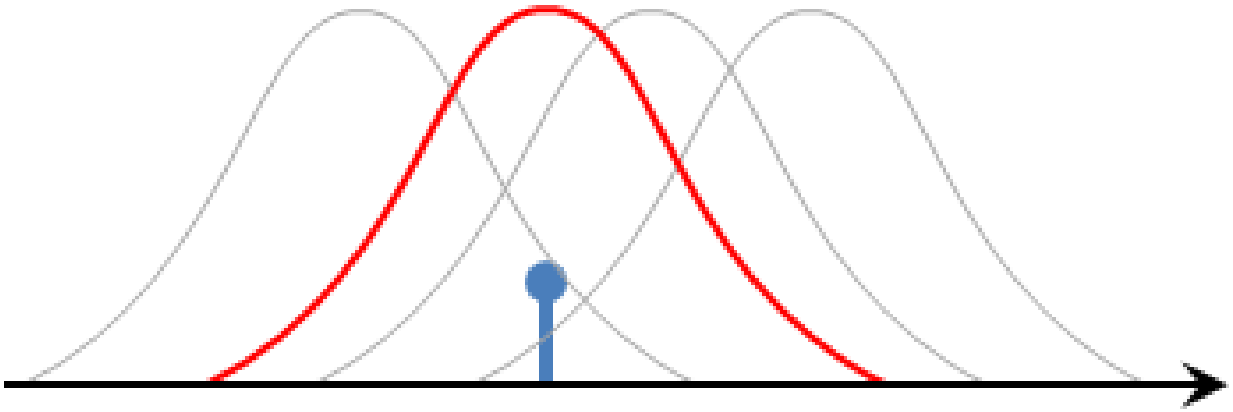
©Stanley Chan 2020. All Rights Reserved.

# Illustrating MLE when $N = 1$. Known $\sigma$.

When $N = 1$: The MLE solution is

$$\widehat{\mu} = \underset{\mu}{\text{argmax}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_1-\mu)^2}{2\sigma^2}\right\}$$
$$= \underset{\mu}{\text{argmin}} \ (x_1-\mu)^2 = x_1.$$

- Which $\mu$ will give you the best Gaussian?
- When $\mu = x_1$, the probability of obtaining $x_1$ is the highest.



©Stanley Chan 2020. All Rights Reserved.

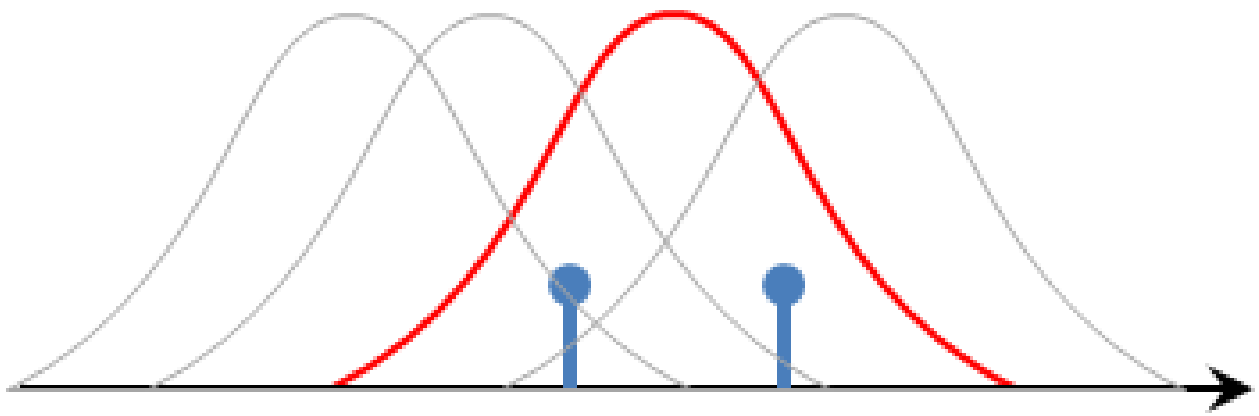
# Illustrating MLE when $N = 2$. Known $\sigma$.

When $N = 2$: The MLE solution is

$$\widehat{\mu} = \underset{\mu}{\text{argmax}} \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^2 \exp\left\{ -\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2}{2\sigma^2} \right\}$$

$$= \underset{\mu}{\text{argmin}} \ (x_1 - \mu)^2 + (x_2 - \mu)^2 = \frac{x_1 + x_2}{2}.$$

- Which $\mu$ will give you the best Gaussian?
- When $\mu = (x_1 + x_2)/2$, the prob. of obtaining $x_1$ and $x_2$ is highest.



©Stanley Chan 2020. All Rights Reserved.

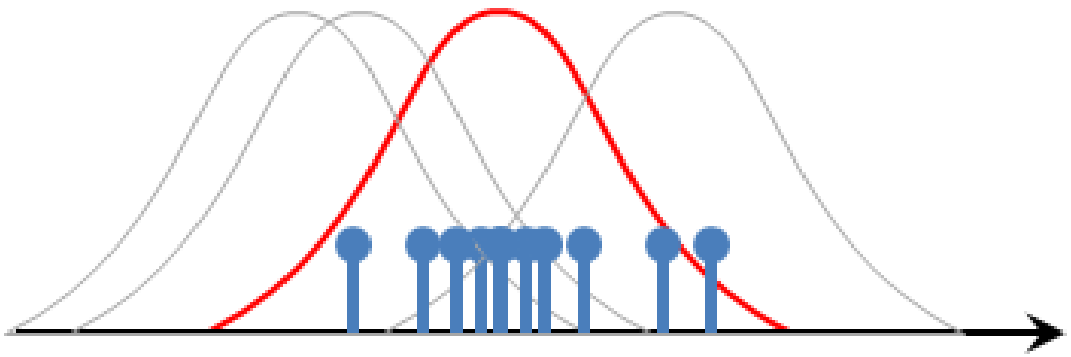
# Illustrating MLE when $N$ = arbitrary integer

The MLE solution is

$$\widehat{\mu} = \underset{\mu}{\text{argmax}} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^2 \exp\left\{-\sum_{n=1}^{N} \frac{(x_n-\mu)^2}{2\sigma^2}\right\}$$

$$= \underset{\mu}{\text{argmin}} \sum_{n=1}^{N}(x_n-\mu)^2 = \frac{1}{N}\sum_{n=1}^{N} x_n.$$

- Which $\mu$ will give you the best Gaussian?
- When $\mu = \frac{1}{N}\sum_{n=1}^{N} x_n$, the prob. of obtaining $\{x_n\}$ is highest.



©Stanley Chan 2020. All Rights Reserved.

# Estimation in High-dimension

- Assume $\boldsymbol{\Sigma}$ is known and fixed.
- Thus, $\boldsymbol{\theta} = \boldsymbol{\mu}$. Estimate $\boldsymbol{\mu}$

$$\widehat{\boldsymbol{\mu}} = \underset{\boldsymbol{\mu}}{\text{argmin}} \cancel{\frac{N}{2}\log|\boldsymbol{\Sigma}|} + \cancel{\frac{N}{2}\log(2\pi)^d} + \sum_{n=1}^{N}\left\{\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\right\}$$

$$= \underset{\boldsymbol{\mu}}{\text{argmin}} \sum_{n=1}^{N}\left\{(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\right\}$$

- Take derivative, setting to zero:

$$\nabla_{\boldsymbol{\mu}}\left\{\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\right\} = 2\sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}) = \mathbf{0}.$$

©Stanley Chan 2020. All Rights Reserved.

# Estimation in High-dimension

- Let us do some algebra

$$\sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}) = \mathbf{0} \implies \sum_{n=1}^{N} \boldsymbol{x}_n = \sum_{n=1}^{N} \boldsymbol{\mu}$$

- Then we can show that the MLE solution is

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n.$$

- This is just the **empirical average** of the entire dataset!
- You can show that if $\mathbb{E}[\boldsymbol{x}_n] = \boldsymbol{\mu}$ for all $n$, then

$$\mathbb{E}[\widehat{\boldsymbol{\mu}}] = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[\boldsymbol{x}_n] = \boldsymbol{\mu}.$$

- We say that $\widehat{\boldsymbol{\mu}}$ is a **unbiased estimator** of $\boldsymbol{\mu}$ since $\mathbb{E}[\widehat{\boldsymbol{\mu}}] = \boldsymbol{\mu}$.

©Stanley Chan 2020. All Rights Reserved.

# When both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are Unknown

What will be the MLE when both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown?

$$(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = \underset{\boldsymbol{\mu}, \boldsymbol{\Sigma}}{\text{argmin}} \ \frac{N}{2} \log |\boldsymbol{\Sigma}| + \frac{N}{2} \log(2\pi)^d + \underbrace{\sum_{n=1}^{N} \left\{ \frac{1}{2} (\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right\}}_{\varphi(\boldsymbol{\mu}, \boldsymbol{\Sigma})}.$$

You need to take derivative with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and solve

$$\begin{aligned} \nabla_{\boldsymbol{\mu}} \varphi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \mathbf{0} \\ \nabla_{\boldsymbol{\Sigma}} \varphi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \mathbf{0} \end{aligned}$$

With some (tedious) matrix calculus, we can show that

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n, \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^T.$$

Exercise: Prove this result when $\boldsymbol{x}_n$ is a 1D scalar.

©Stanley Chan 2020. All Rights Reserved.