Robotics

Estimation and Learning
with Dan Lee

Basic Intro to Probability

Penn Engineering

Why Learn About Probability?

- The real world has huge aspects of randomness and uncertainty.
- Still, we hope to make useful predictions and inferences.
- Randomness often follows reliable laws.
- The language of these laws is the language of probability (and statistics).

Contents

1. Definition of Probability

- Consider an (potentially abstract) experiment
  - A **sample space** Ω is the set of *all possible* outcomes of that experiment
  - An **elementary event** ω is a single outcome of the set.
- Example (Rolling two dice):
  - Sample space $\Omega = \{(1,1), (1,2), (1,3), \ldots, (6,5), (6,6)\}$
  - Each member $\omega \in \Omega$ is an elementary event
  - Example of non-elementary event : Rolling doubles

    $B = \{(1,1), (2,2), (3,3), \ldots, (6,6)\}$

1. Definition of Probability

- Consider a finite* set Ω
- A **probability space** (Ω, P) is a sample space Ω, together with a function P, satisfying the following:
  i. $0 \leq P(\omega) \leq 1$ for all $\omega \in \Omega$
  ii. $\sum_{\omega \in \Omega} P(\omega) = 1$
  iii. For any event $A \subseteq \Omega$, $P(A) = \sum_{\omega \in \mathrm{A}} P(\omega)$
- The function P is called **probability measure**.

*To deal with countably infinite or uncountable space, we need third element called sigma algebra, but here we are simplifying.

1. Definition of Probability

- Basic Consequences

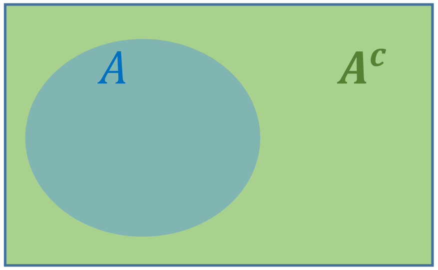
$$P(\emptyset) = 0$$

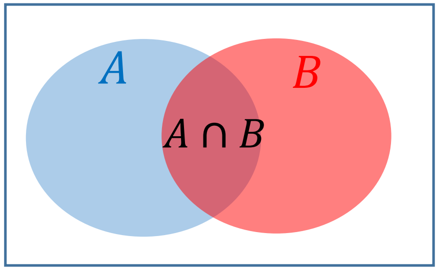
$$P(\Omega) = 1$$

$$P(A^C) = 1 - P(A)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A) = P(A \cap B) + P(A \cap B^c)$$





2. Independence

- Given a probability space, two events A and B are **independent** if and only if:

$$P(A \cap B) = P(A)P(B)$$

- Two events are dependent if they are not independent.

2. Independence

- Example (two coin flip): $\Omega = \{HH, HT, TH, TT\}$
  - Assume $\forall \omega \in \Omega$, $P(\omega) = \frac{1}{|\Omega|} = \frac{1}{4}$.
  - Define event A: "First flip is H." $A = \{HH, HT\}$
  - Define event B: "Second flip is H." $B = \{HH, TH\}$
  - Are A and B Independent?

    i) $P(A \cap B) = P(\{HH\}) = 1/4$

    ii) $P(A)P(B) = 1/2 * 1/2 = 1/4$ → Yes

2. Independence

- Example (two coin flip): $\Omega = \{HH, HT, TH, TT\}$
    - Assume $\forall \omega \in \Omega$, $P(\omega) = \frac{1}{|\Omega|} = \frac{1}{4}$.
    - Define event A: "First flip is H." $A = \{HH, HT\}$
    - Define event B: "Contains a T." $B = \{HT, TH, TT\}$
    - Are A and B Independent?

        i) $P(A \cap B) = P(\{HT\}) = 1/4$

        ii) $P(A)P(B) = 1/2 * 3/4 = 3/8$ → No

3. Conditional Probability

- Given some probability space (Ω, P), for any two events A and B, if $P(B) \neq 0$, then we define the **conditional probability** $P(A|B)$ that A occurs given that B occurs as,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- From this follows **Chain Rule**:

$$P(A \cap B) = P(A|B)\,P(B)$$

$$\begin{aligned} P(A \cap B \cap C) &= P(A|B \cap C)\,P(B \cap C) \\ &= P(A|B \cap C)\,P(B|C)P(C) \end{aligned}$$

and so on..

3. Conditional Probability

- Example (two coin flip): What is the probability that both are head, GIVEN at least one is head?
- Probability Problem:

$\Omega = \{HH, HT, TH, TT\}$
$B = \{HH, HT, TH\}$ → "At least one is head."
$A = \{HH\}$ → "Both are heads."
$P(A \cap B) = P(\{HH\}) = 1/4$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/4}{3/4} = \frac{1}{3}$$

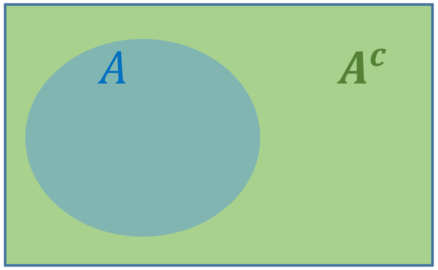
3. Conditional Probability

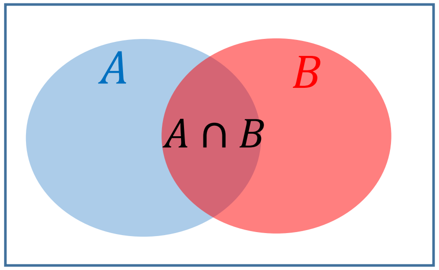
- Consequences

$$P(\emptyset|B) = 0$$

$$P(\mathrm{B}|\mathrm{B}) = 1$$

$$P(A|B) = 1 - P(A^C|B)$$





4. Bayes Rule

- From chain rule,
  $P(A \cap B) = P(A|B)\, P(B)$
  $P(B \cap A) = P(B|A)\, P(A)$
- Intersections are commutative, $A \cap B = B \cap A$.

$$
\begin{array}{ccccc}
P(A \cap B) & = & P(A|B) & P(B) \\
\| & & \| & \\
P(B \cap A) & = & P(B|A) & P(A)
\end{array}
$$

- As a result, we have **Bayes Rule**:

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

4. Bayes Rule

- Each term is often called:

$$\underbrace{P(A|B)}_{\text{Posterior}} = \frac{\overbrace{P(B|A)}^{\text{Likelihood}}\overbrace{P(A)}^{\text{Prior}}}{\underbrace{P(B)}_{\text{Evidence}}}$$

5. Random Variables

- Given some probability space (Ω, P), a **random variable** $X: \Omega \to R$ is a *function* that maps the sample space to the reals.
- When we say $P(X = a)$, we actually mean the probability of the inverse image $X^{-1}(a)$. That is,

$$P(X = a) = P\left(X^{-1}(a)\right) = P(\{\omega \in \Omega | X(\omega) = a\})$$

- Example (single coin flip): $\Omega = \{Head, Tail\}$

$$X(Head) = 1, X(Tail) = 0$$

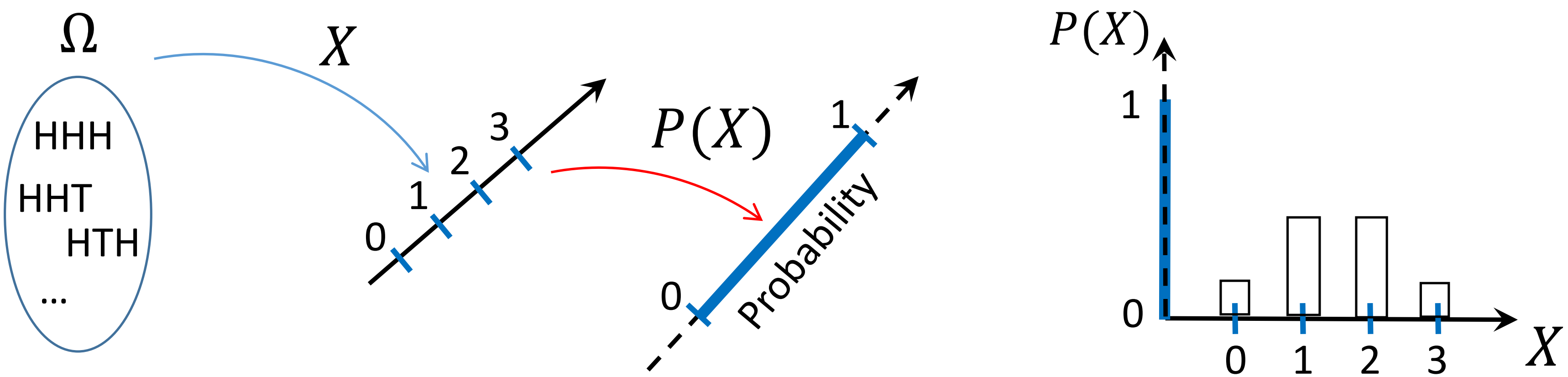
$$P(X = 1) = P\left(X^{-1}(1)\right) = P(Head)$$

5. Random Variables

- Example: 3 coin flips
  - $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
  - Let us define $X(\omega)$ to be the number of Heads in a given flip. Then,

    $X(HHH) = 3, X(HHT) = X(HTH) = 2, \ldots, X(TTT) = 0$

    $P(X = 3) = 1/8, P(X = 2) = P(X = 1) = 3/8, P(X = 0) = 1/8$

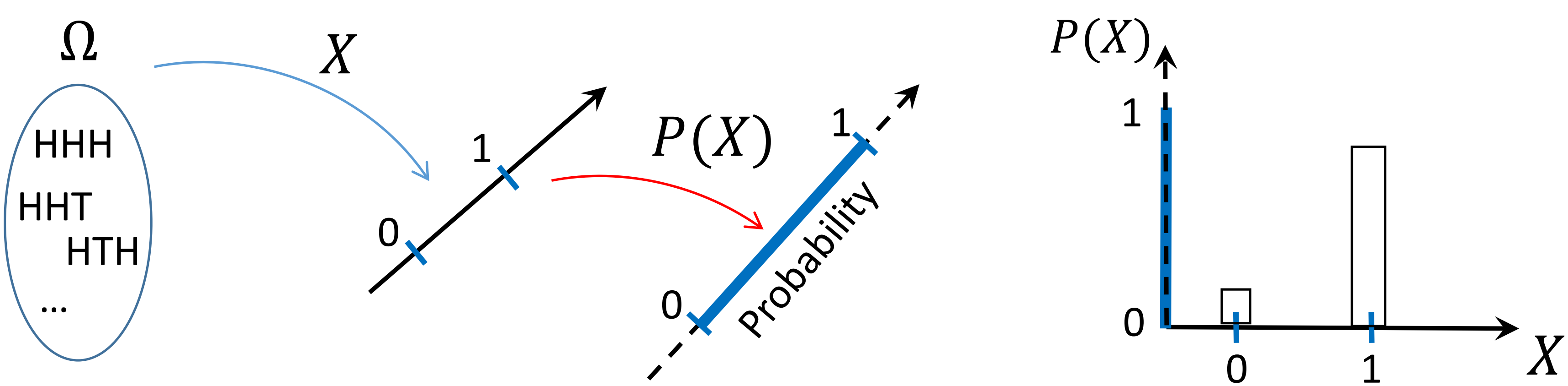


5. Random Variables

- Example: 3 coin flips
  - $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
  - This time, let us define $X(\omega)$ to be 1 if H appears in a given flip, otherwise is 0. Then,

    $X(HHH) = X(HHT) = \cdots = 1\,, X(TTT) = 0$

    $P(X = 1) = 7/8, P(X = 0) = 1/8$

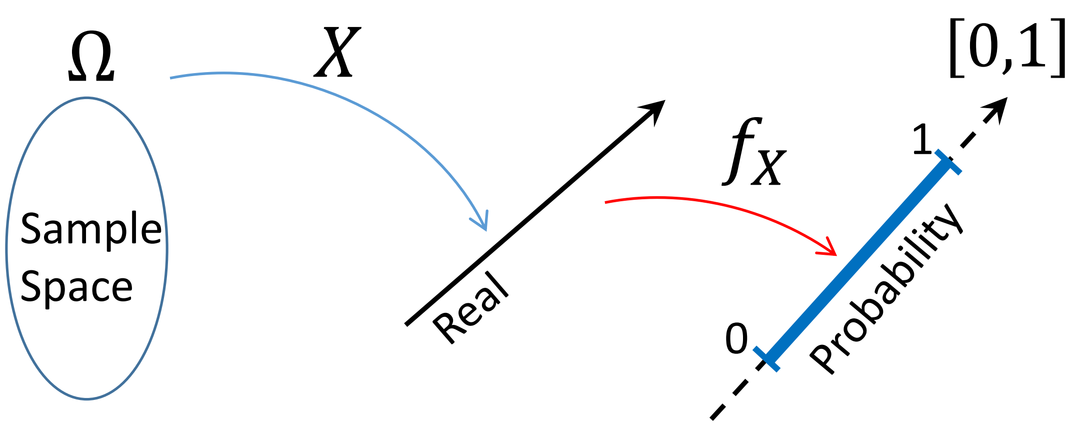


6. Density/Distribution Functions

- **Probability mass function** (pmf) of **discrete** RVs
- **Probability density function** (pdf) of **continuous** RVs

$f: R \rightarrow [0,1]$

$\forall a \in R, f_X(a) = P(X = a)$



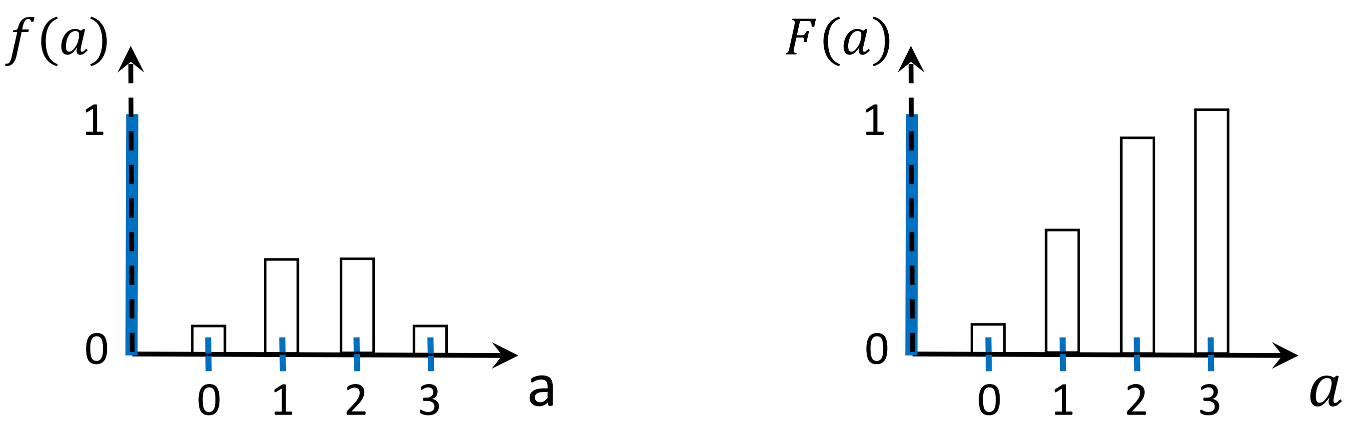
6. Density/Distribution Functions

- **Cumulative distribution functions** (cdf)

  $F: R \to [0,1]$

  $\forall a \in R, F_X(a) = P(X \leq a)$
- A cdf is a monotonic nondecreasing function, i.e.,

  $\forall x \leq y, F(x) \leq F(y)$



Acknowledgement

- Thanks to Daniel Moroz, Dan Lee's master student at the University of Pennsylvania, for allowing us to take parts of contents from his slides.

Robotics

Estimation and Learning
with Dan Lee

Week 1.
Gaussian Model Learning

1.2.1 1D Gaussian Distribution

Gaussian Distribution

Why Gaussian?

- The two parameters (mean and variance) are easy to compute and interpret.

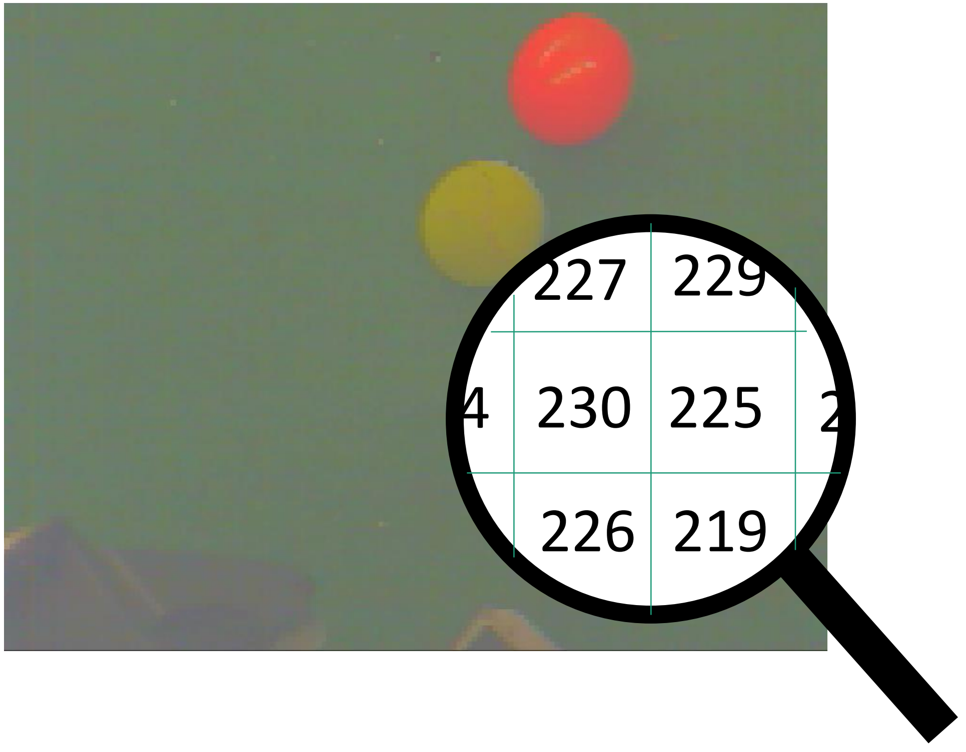
- Good mathematical properties:
  e.g., product of Gaussian distributions forms Gaussian.

- Central limit theorem:
  Expectation of the mean of any random variables converges to Gaussian.

Gaussian Distribution : Example

Ball color distribution

Color Image

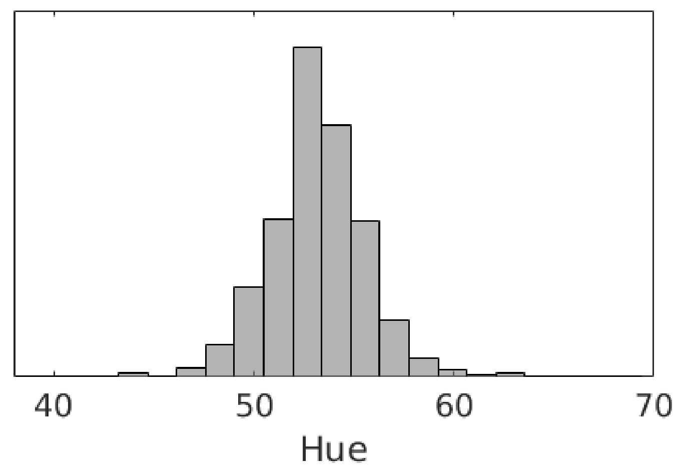


“Yellow”?
“Red”?

Gaussian Distribution : Example

Ball color distribution

Color Image



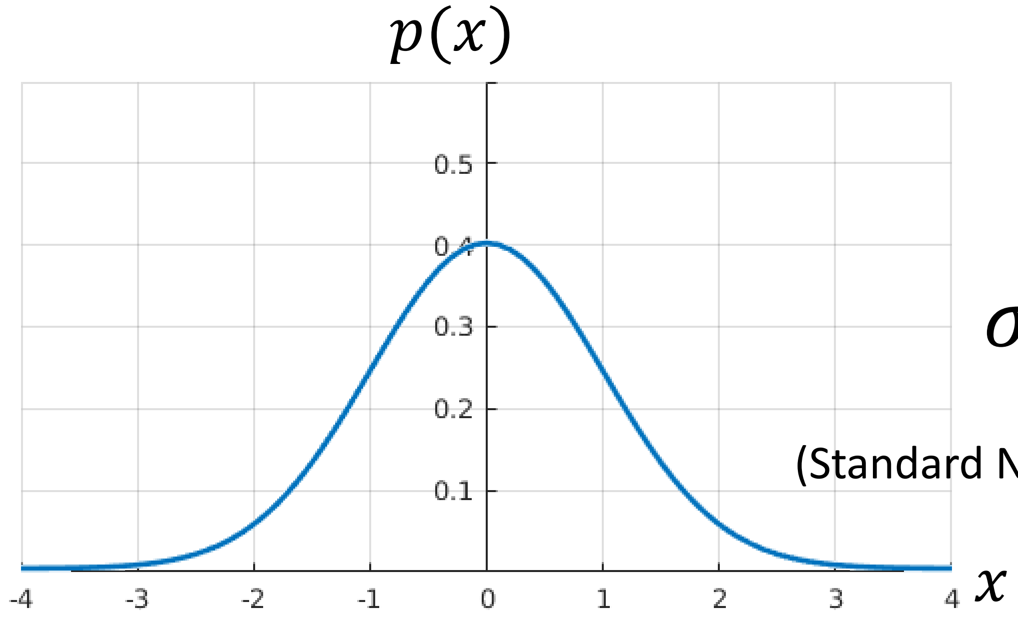
Gaussian Distribution (1D)

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

x	Variable
μ	Mean
σ^2	Variance
σ	Standard deviation

Gaussian Distribution (1D)

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



$$\mu = 0$$

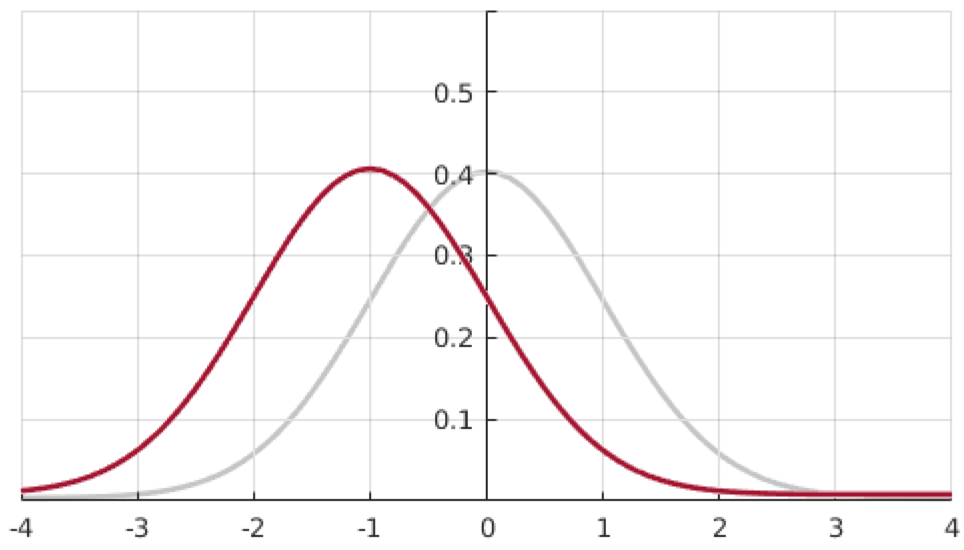
$$\sigma^2 = 1$$

(Standard Normal Distribution)

Gaussian Distribution (1D)

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x+1)^2}{2}}$$

$p(x)$



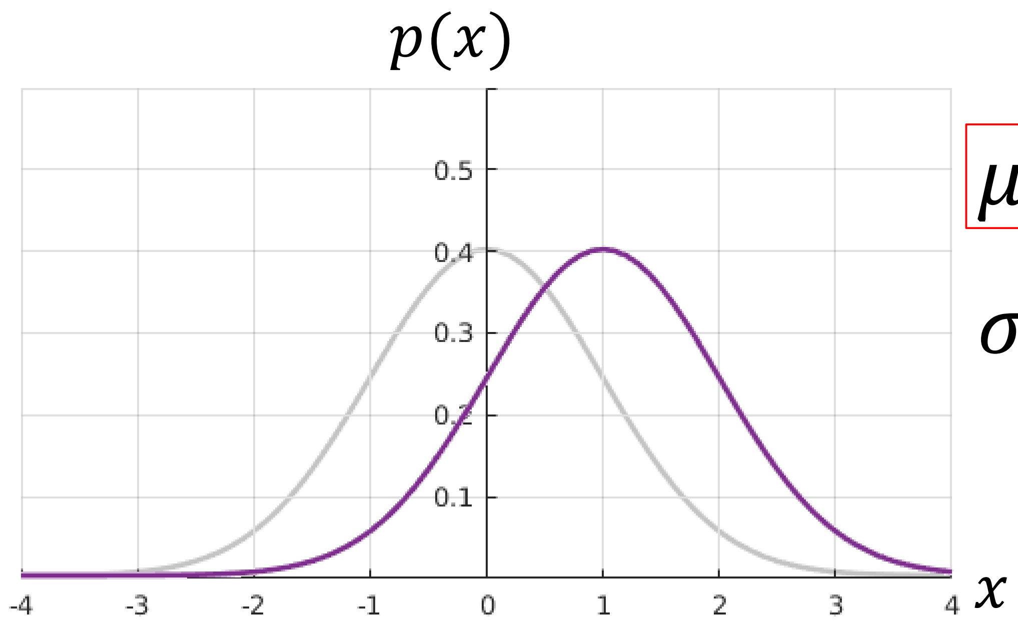
x

$\boxed{\mu = -1}$

$\sigma^2 = 1$

Gaussian Distribution (1D)

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

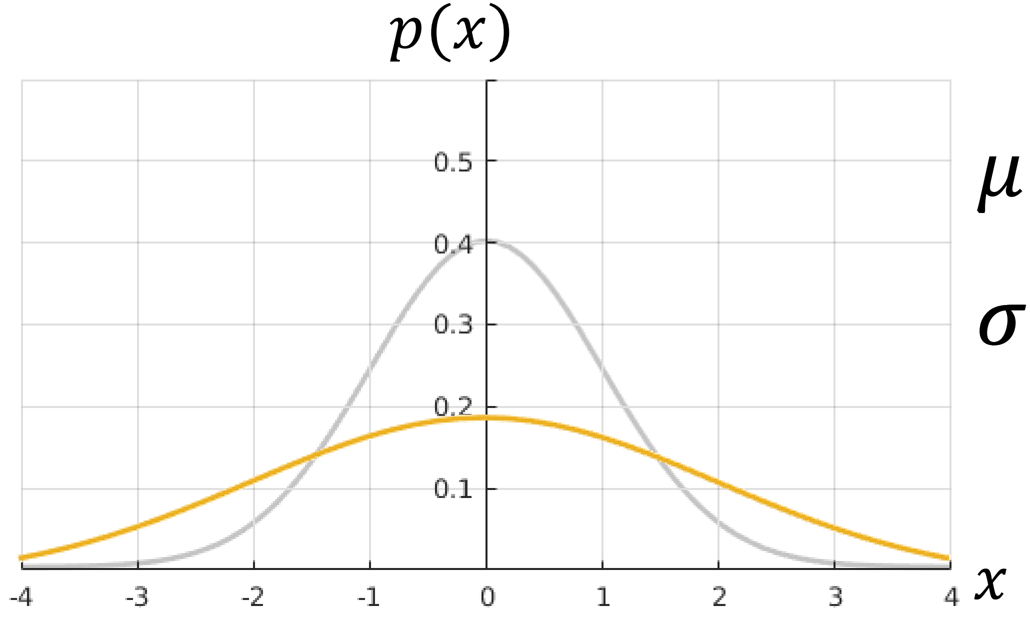


$\boxed{\mu = 1}$

$\sigma^2 = 1$

Gaussian Distribution (1D)

$$p(x) = \frac{1}{2\sqrt{\pi}} e^{-\frac{x^2}{4}}$$

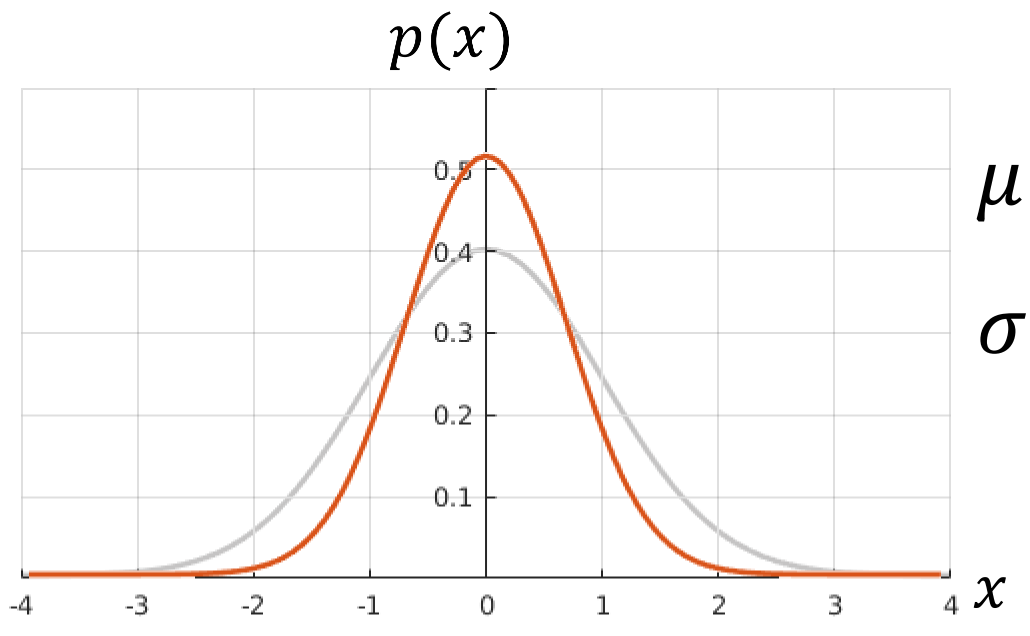


$\mu = 0$

$\sigma^2 = 2$

Gaussian Distribution (1D)

$$p(x) = \frac{1}{\sqrt{1.4\pi}} e^{-\frac{x^2}{1.4}}$$

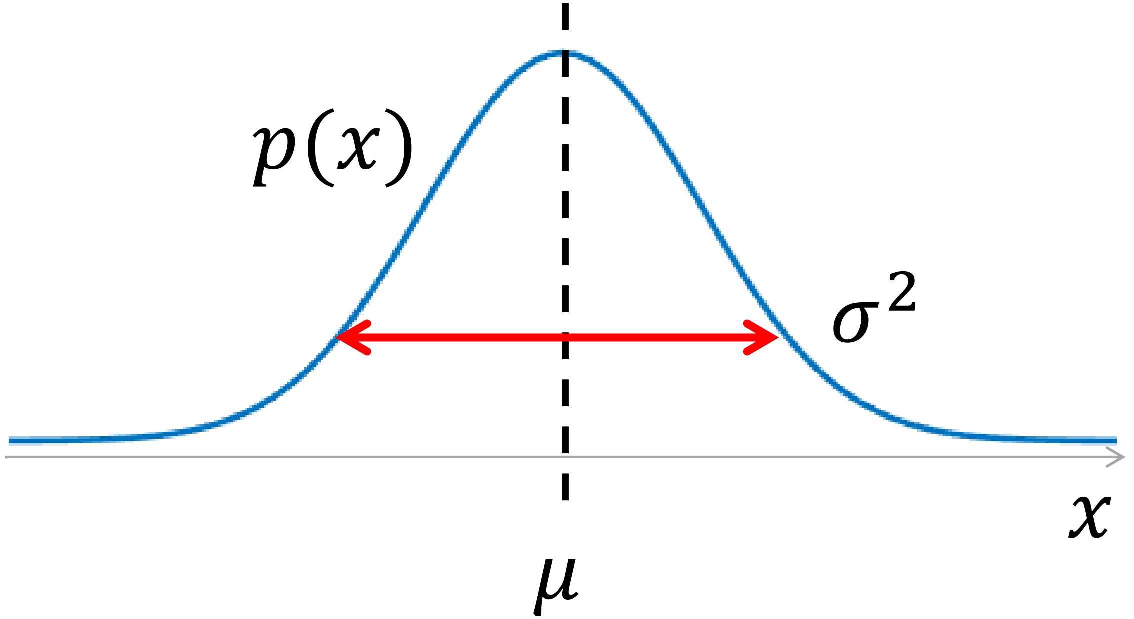


$\mu = 0$

$\sigma^2 = 0.7$

Gaussian Distribution (1D)

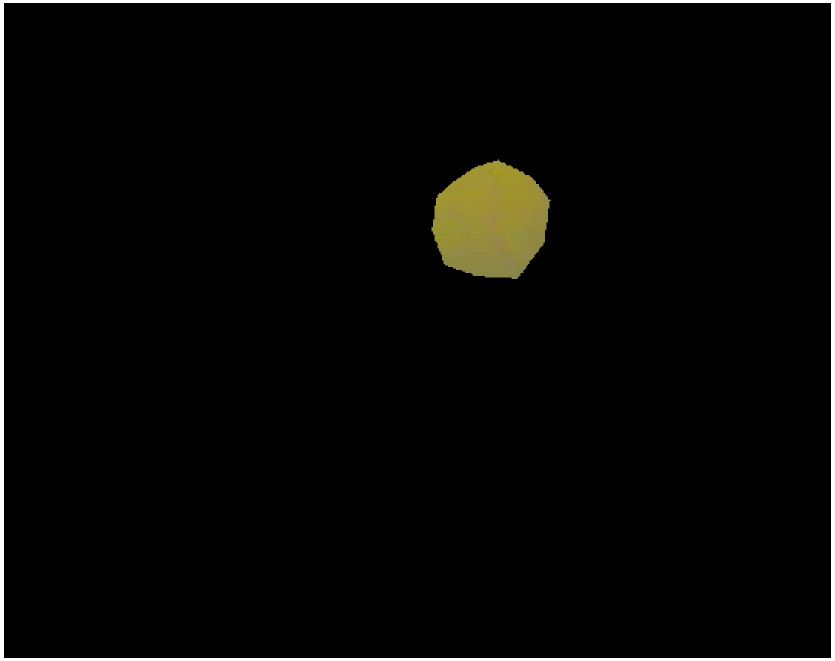
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

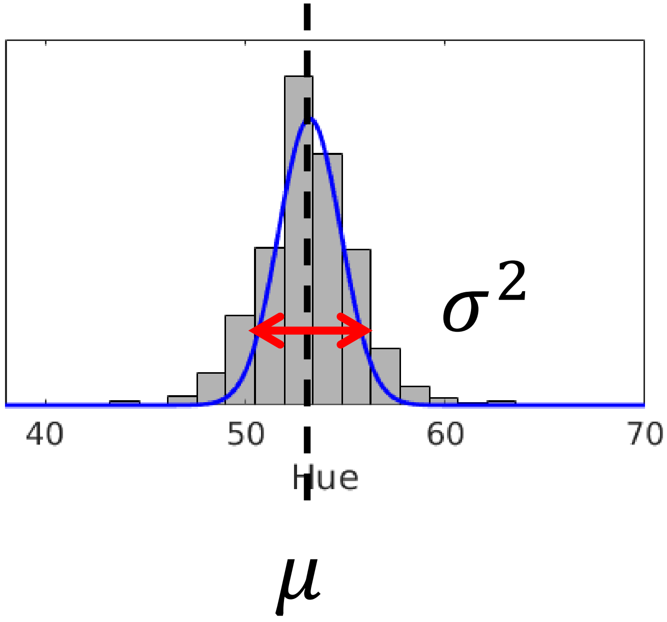


Gaussian Distribution : Example

Ball color distribution

Color Image





Acknowledgement

- Thanks to Rei Suzuki, Dan Lee’s master student at the University of Pennsylvania, for helping us create the lectures for WEEK 1.

Robotics

Estimation and Learning
with Dan Lee

Week 1.
Gaussian Model Learning

1.2.2 Maximum Likelihood Estimate

Penn Engineering

Maximum Likelihood Estimate of Gaussian Model Parameters

- Objective

  Estimate the mean and the variance given observed data

Likelihood: $p(\{x_i\}|\mu, \sigma)$

Observed data Unknown parameters

Maximum Likelihood Estimate of Gaussian Model Parameters

- Objective

$$\hat{\mu}, \hat{\sigma} = \arg\max_{\mu,\sigma} p(\{x_i\}|\mu, \sigma)$$

Maximum Likelihood Estimate of Gaussian Model Parameters

- Objective

$$\hat{\mu}, \hat{\sigma} = \arg\max_{\mu,\sigma} p(\{x_i\}|\mu, \sigma)$$

Assuming independence of observations,

$$p(\{x_i\}|\mu, \sigma) = \prod_{i=1}^{N} p(x_i|\mu, \sigma)$$

$$\boxed{\hat{\mu}, \hat{\sigma} = \arg\max_{\mu,\sigma} \prod_{i=1}^{N} p(x_i|\mu, \sigma)}$$

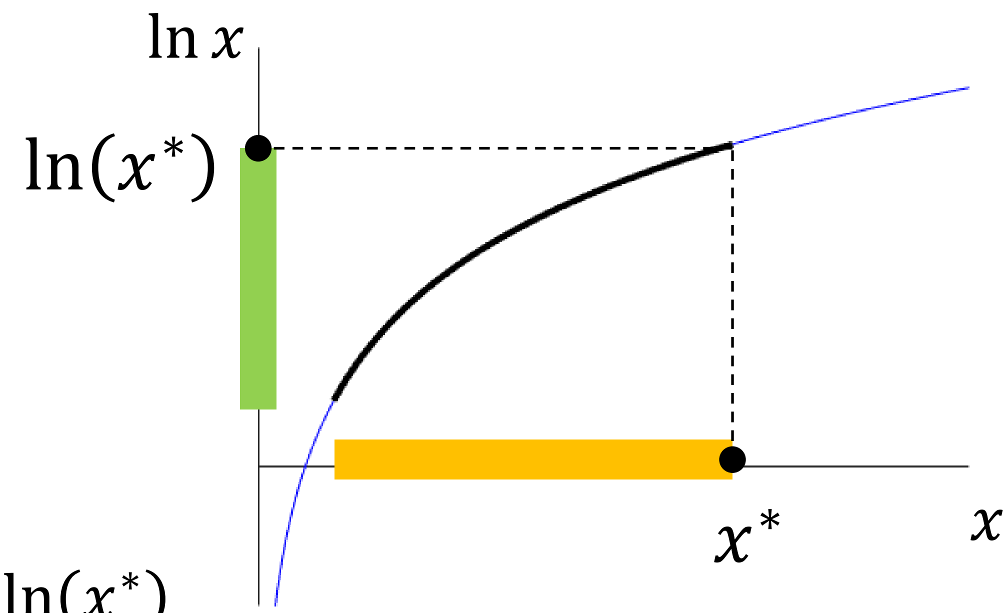
Maximum Likelihood Estimate of Gaussian Model Parameters

$$\hat{\mu}, \hat{\sigma} = \arg\max_{\mu,\sigma} \prod_{i=1}^{N} p(x_i | \mu, \sigma)$$

(1) Take the log!

Maximum Likelihood Estimate of Gaussian Model Parameters

$$\hat{\mu}, \hat{\sigma} = \arg\max_{\mu,\sigma} \prod_{i=1}^{N} p(x_i | \mu, \sigma)$$



NOTE 1:

$$x \leq x^* \leftrightarrow \ln(x) \leq \ln(x^*)$$

Maximum Likelihood Estimate of Gaussian Model Parameters

$$\hat{\mu}, \hat{\sigma} = \arg\max_{\mu,\sigma} \prod_{i=1}^{N} p(x_i|\mu,\sigma)$$

(1) $$\arg\max_{\mu,\sigma} \prod_{i=1}^{N} p(x_i|\mu,\sigma) = \arg\max_{\mu,\sigma} \ln\left\{\prod_{i=1}^{N} p(x_i|\mu,\sigma)\right\}$$

NOTE 1:

$$x \leq x^* \leftrightarrow \ln(x) \leq \ln(x^*)$$

Maximum Likelihood Estimate of Gaussian Model Parameters

$$\hat{\mu}, \hat{\sigma} = \arg\max_{\mu,\sigma} \prod_{i=1}^{N} p(x_i | \mu, \sigma)$$

(1) $$\arg\max_{\mu,\sigma} \prod_{i=1}^{N} p(x_i | \mu, \sigma) = \arg\max_{\mu,\sigma} \ln \left\{ \prod_{i=1}^{N} p(x_i | \mu, \sigma) \right\}$$

$$= \arg\max_{\mu,\sigma} \boxed{\sum_{i=1}^{N} \ln p(x_i | \mu, \sigma)}$$

NOTE 2:

$$\log(x_1 \times x_2 \times \cdots \times x_k) = \log(x_1) + \log(x_2) + \cdots + \log(x_k)$$

Maximum Likelihood Estimate of Gaussian Model Parameters

$$\hat{\mu}, \hat{\sigma} = \arg\max_{\mu,\sigma} \sum_{i=1}^{N} \ln \underline{p(x_i|\mu, \sigma)}$$

(2) Gaussian!

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\}$$

Maximum Likelihood Estimate of Gaussian Model Parameters

$$\hat{\mu}, \hat{\sigma} = \arg\max_{\mu,\sigma} \sum_{i=1}^{N} \ln p(x_i|\mu,\sigma)$$

$$\ln \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$$

$$= \left\{-\frac{(x_i-\mu)^2}{2\sigma^2} - \ln\sigma - \ln\sqrt{2\pi}\right\}$$

Maximum Likelihood Estimate of Gaussian Model Parameters

$$\hat{\mu}, \hat{\sigma} = \arg\max_{\mu,\sigma} \sum_{i=1}^{N} \left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} - \ln \sigma \right\}$$

$$\updownarrow$$

$$\boxed{\hat{\mu}, \hat{\sigma} = \arg\min_{\mu,\sigma} \sum_{i=1}^{N} \left\{ \frac{(x_i - \mu)^2}{2\sigma^2} + \ln \sigma \right\}}$$

Maximum Likelihood Estimate of Gaussian Model Parameters

$$\hat{\mu}, \hat{\sigma} = \arg\min_{\mu,\sigma} \underbrace{\sum_{i=1}^{N} \left\{ \frac{(x_i - \mu)^2}{2\sigma^2} + \ln \sigma \right\}}_{J(\mu,\sigma)}$$

- At optimum,

$$\boxed{\frac{\partial J}{\partial \mu} = 0 \longrightarrow \hat{\mu}} \qquad \boxed{\frac{\partial J(\hat{\mu}, \sigma)}{\partial \sigma} = 0 \rightarrow \hat{\sigma}}$$

Maximum Likelihood Estimate of Gaussian Model Parameters

- The MLE Solution:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

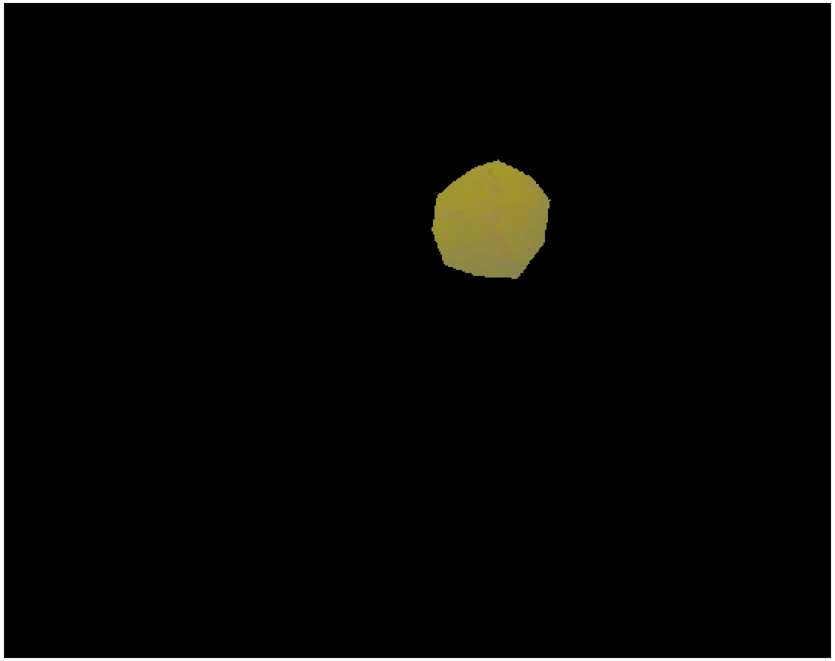
MLE Estimate: Example

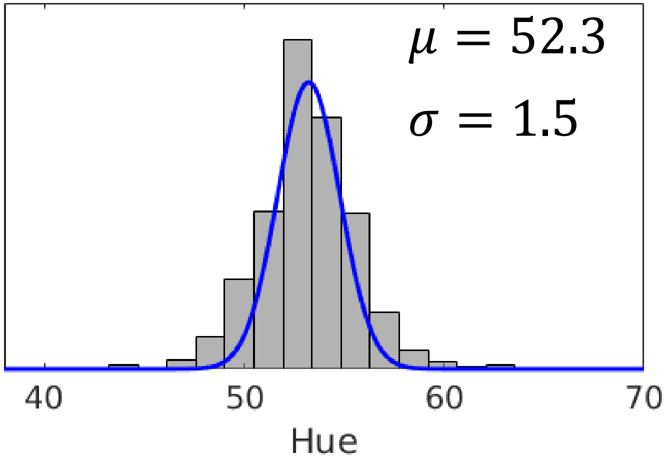
Ball color distribution

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$$

Segmented Ball Image





Robotics

Estimation and Learning
with Dan Lee

Supplementary Notes on
MLE for Univariate Gaussian

Penn
Engineering

Maximum Likelihood Estimate of Gaussian Model Parameters

- Objective

  Estimate the mean and the variance given observed data

$$\hat{\mu}, \hat{\sigma} = \arg\max_{\mu,\sigma} p(\{x_i\}|\mu, \sigma)$$

- Assuming independence of observations,

$$\hat{\mu}, \hat{\sigma} = \arg\max_{\mu,\sigma} \prod_{i=1}^{N} p(x_i|\mu, \sigma)$$

Maximum Likelihood Estimate of Gaussian Model Parameters

- To obtain $\hat{\mu}, \hat{\sigma} = \arg\max_{\mu,\sigma} \prod_{i=1}^{N} p(x_i|\mu,\sigma)$

(1) $\arg\max_{\mu,\sigma} \prod_{i=1}^{N} p(x_i|\mu,\sigma) = \arg\max_{\mu,\sigma} \sum_{i=1}^{N} \ln p(x_i|\mu,\sigma)$

(2) $p(x_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$ (Gaussian)

Maximum Likelihood Estimate of Gaussian Model Parameters

$$\hat{\mu}, \hat{\sigma} = \arg\max_{\mu,\sigma} \prod_{i=1}^{N} p(x_i | \mu, \sigma)$$

$$= \arg\max_{\mu,\sigma} \sum_{i=1}^{N} \ln p(x_i | \mu, \sigma) \qquad \text{from (1)}$$

$$= \arg\max_{\mu,\sigma} \sum_{i=1}^{N} \ln \left\{ \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} \right\} \quad \text{from (2)}$$

Maximum Likelihood Estimate of Gaussian Model Parameters

(continued)

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

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

$$= \arg\min_{\mu,\sigma} \sum_{i=1}^{N} \left\{\frac{(x_i-\mu)^2}{2\sigma^2} + \ln\sigma + const\right\}$$

Maximum Likelihood Estimate of Gaussian Model Parameters

Let $$J(\mu, \sigma) = \sum_{i=1}^{N} \left\{ \frac{(x_i - \mu)^2}{2\sigma^2} + \ln\sigma \right\}$$

Then $$\hat{\mu}, \hat{\sigma} = \arg\min_{\mu,\sigma} J(\mu, \sigma)$$

① $$\frac{\partial J}{\partial \mu} = 0 \longrightarrow \hat{\mu}$$

② $$\frac{\partial J(\hat{\mu}, \sigma)}{\partial \sigma} = 0 \longrightarrow \hat{\sigma}$$

Maximum Likelihood Estimate of Gaussian Model Parameters

① $$\boxed{\frac{\partial J}{\partial \mu} = 0 \longrightarrow \hat{\mu}}$$

$$\frac{\partial J}{\partial \mu} = \frac{\partial}{\partial \mu}\sum_{i=1}^{N}\left\{\frac{(x_i-\mu)^2}{2\sigma^2} + \ln\sigma\right\}$$

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

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N}x_i \longleftarrow = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu) = 0$$

Maximum Likelihood Estimate of Gaussian Model Parameters

② $$\boxed{\frac{\partial J(\hat{\mu},\sigma)}{\partial \sigma} = 0 \longrightarrow \hat{\sigma}}$$

$$\frac{\partial J}{\partial \sigma} = \frac{\partial}{\partial \sigma}\sum_{i=1}^{N}\left\{\frac{(x_i-\hat{\mu})^2}{2\sigma^2} + \ln\sigma\right\}$$

$$= \left(\frac{\partial}{\partial \sigma}\frac{1}{2\sigma^2}\right)\left(\sum_{i=1}^{N}(x_i-\hat{\mu})^2\right) - \frac{N}{\sigma}$$

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

Maximum Likelihood Estimate of Gaussian Model Parameters

- In summary, we have

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

Robotics

Estimation and Learning with Dan Lee

Week 1. Gaussian Model Learning

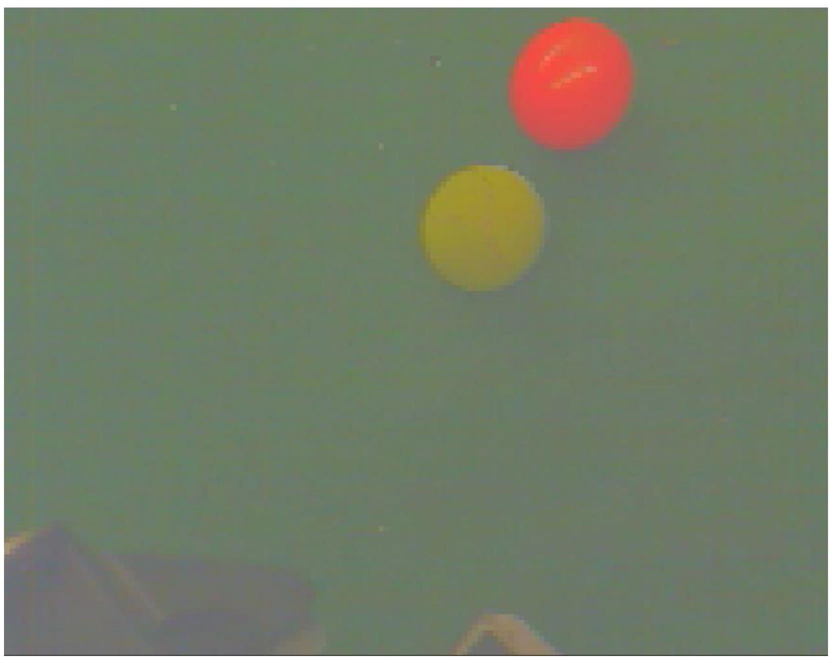
1.3.1 Multivariate Gaussian Distribution

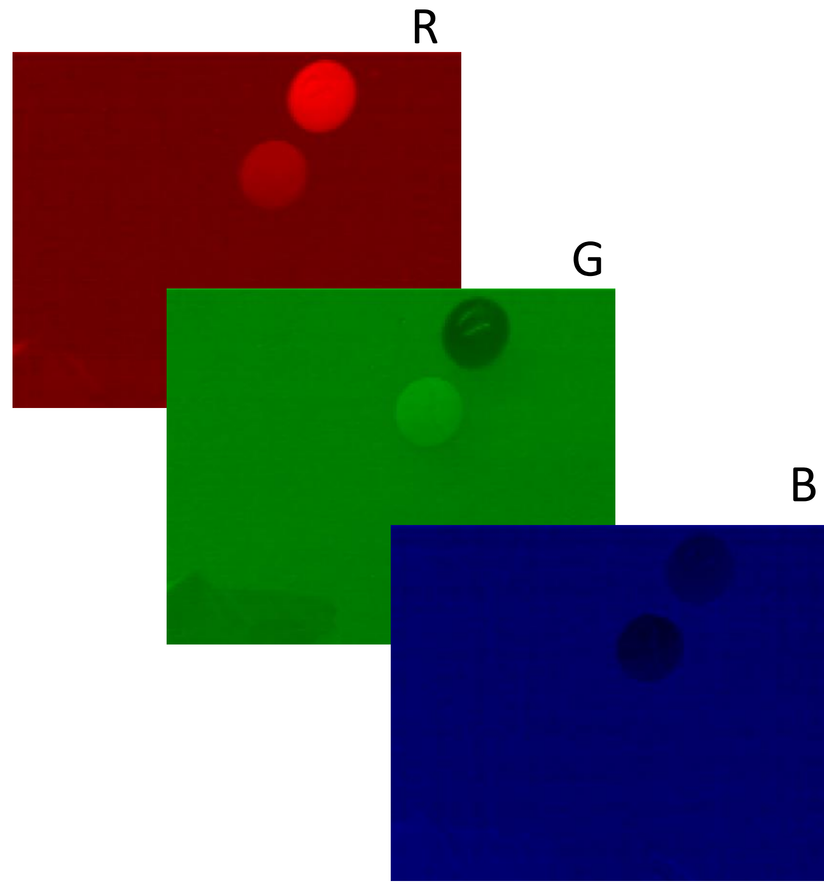
Penn Engineering

Multivariate Gaussian : Example

Ball color in multi-channels

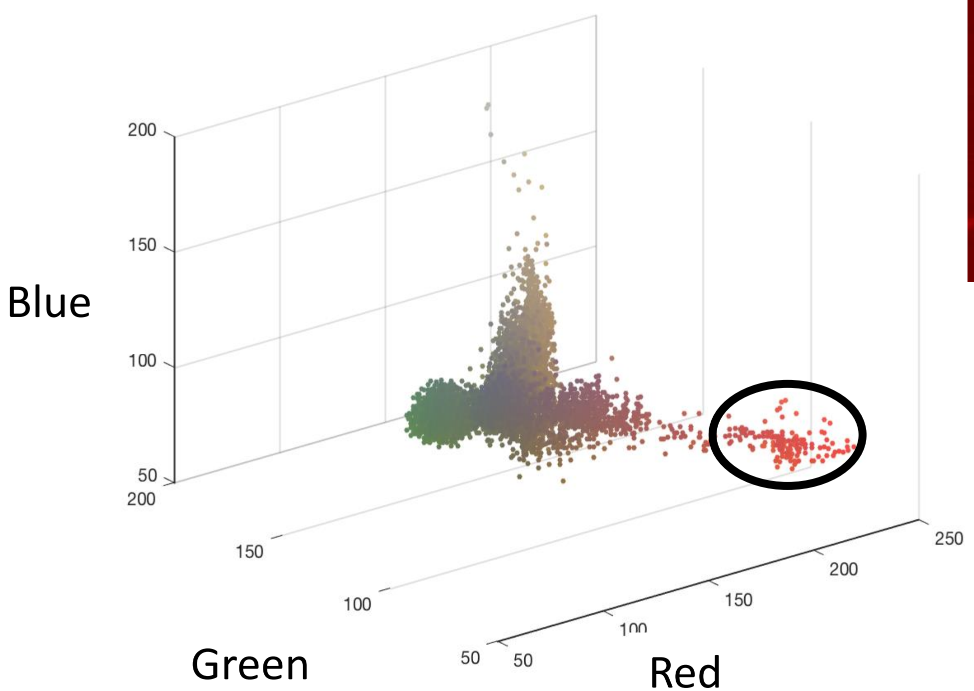
RGB Image

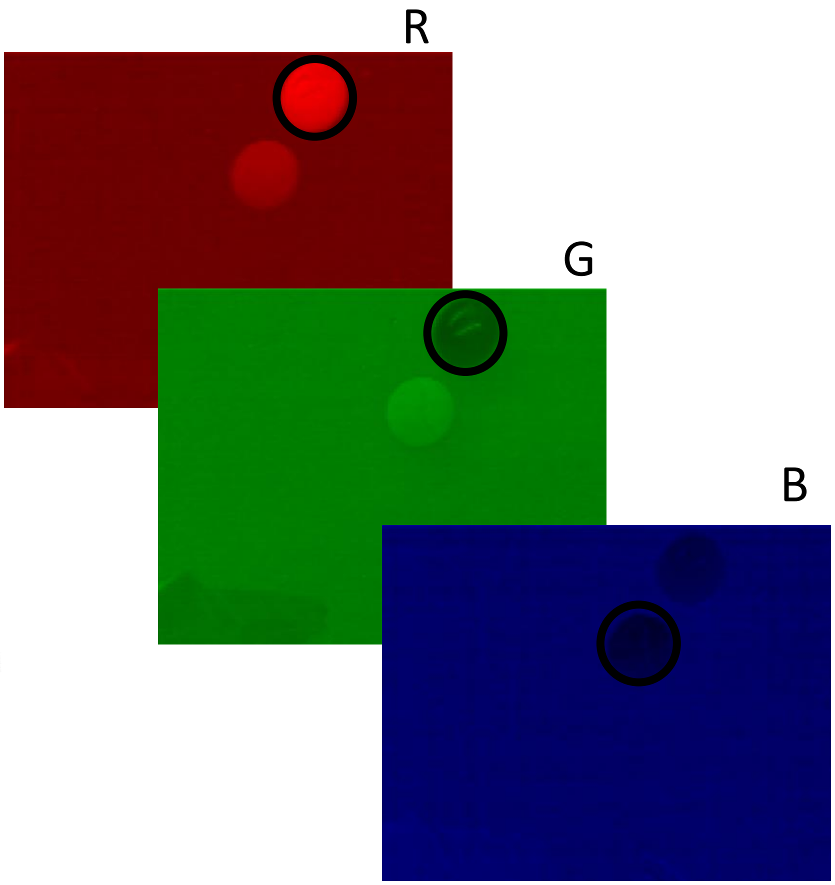




Multivariate Gaussian : Example

Ball color in multi-channels





Multivariate Gaussian

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\boldsymbol{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

D Number of Dimensions

$\mathbf{x}$ Variable

$\boldsymbol{\mu}$ Mean *vector*

Σ Covariance *matrix*

(Dimension = 1)

$$\boxed{p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}}$$

Multivariate Gaussian

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{T}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Σ Covariance *matrix*

* Diagonal terms: variance
* Off-diagonal terms: correlation

(Dimension = 2)

$$\boxed{\Sigma = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1x_2}^2 \\ \sigma_{x_2x_1}^2 & \sigma_{x_2}^2 \end{bmatrix}} \quad (\sigma_{x_1x_2}^2 = \sigma_{x_2x_1}^2)$$

Multivariate Gaussian

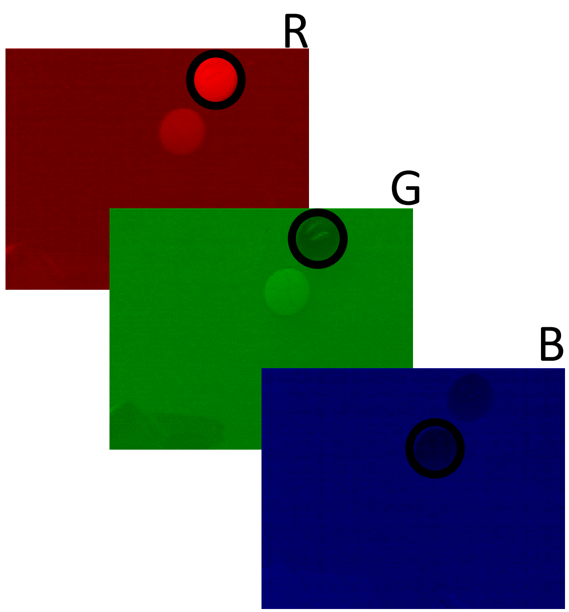
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{T}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Determinant of Σ

Multivariate Gaussian

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{T}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Ball color in multi-channels



$$D = 3$$

$$\mathbf{x} = [x_R \quad x_G \quad x_B]$$

$$\boldsymbol{\mu} = [\mu_R \quad \mu_G \quad \mu_B]$$

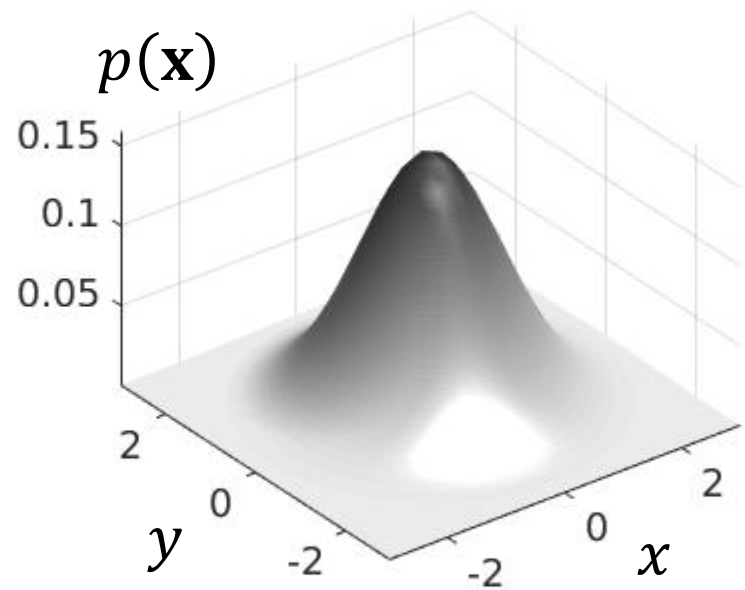
$$\Sigma = \begin{bmatrix} \sigma_{x_R}^2 & \sigma_{x_R x_G}^2 & \sigma_{x_R x_B}^2 \\ \sigma_{x_R x_G}^2 & \sigma_{x_G}^2 & \sigma_{x_G x_B}^2 \\ \sigma_{x_R x_B}^2 & \sigma_{x_G x_B}^2 & \sigma_{x_B}^2 \end{bmatrix}$$

Multivariate Gaussian: 2D

$$p(\mathbf{x}) = \frac{1}{2\pi} \exp\left\{-\frac{x^2 + y^2}{2}\right\}$$

- 2D Zero-mean Spherical Case

$$D = 2$$
$$\mathbf{x} = [x \quad y]^T$$
$$\boldsymbol{\mu} = [0 \quad 0]^T$$
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

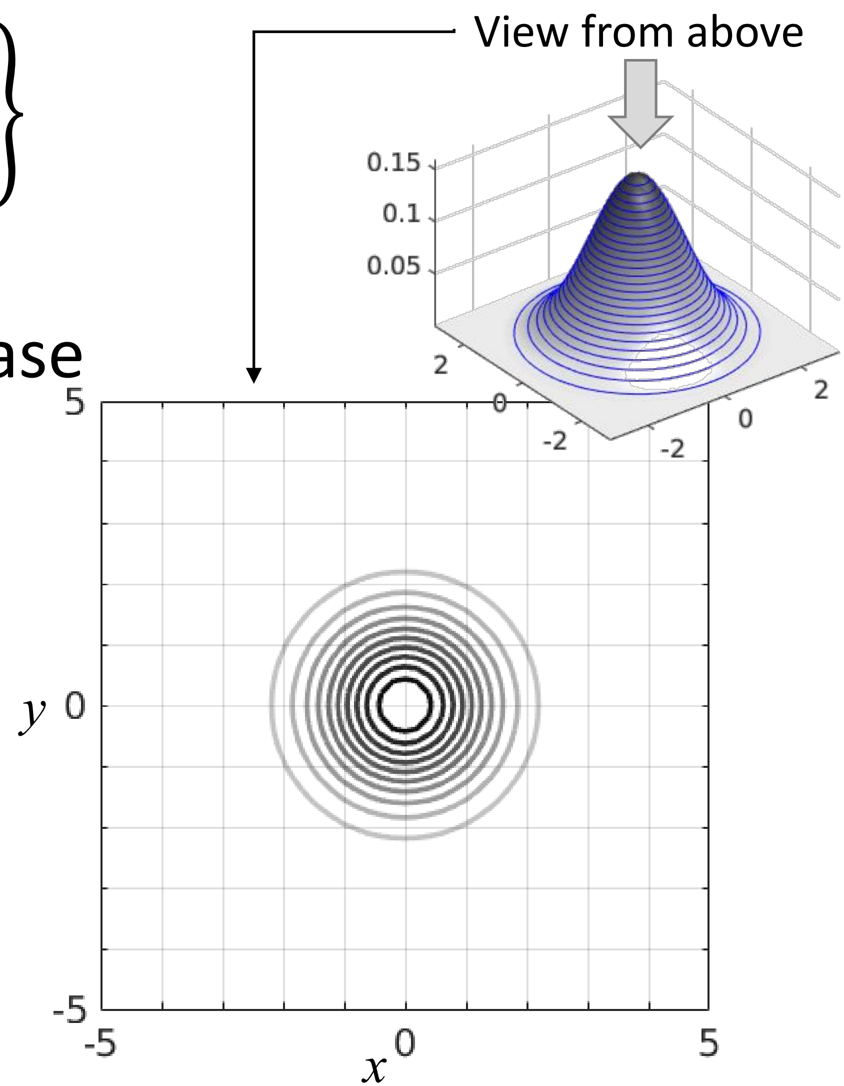


Multivariate Gaussian: 2D

$$p(\mathbf{x}) = \frac{1}{2\pi}\exp\left\{-\frac{x^2+y^2}{2}\right\}$$

- 2D Zero-mean Spherical Case

$$D = 2$$
$$\mathbf{x} = [x \quad y]^T$$
$$\boldsymbol{\mu} = [0 \quad 0]^T$$
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



Multivariate Gaussian: 2D

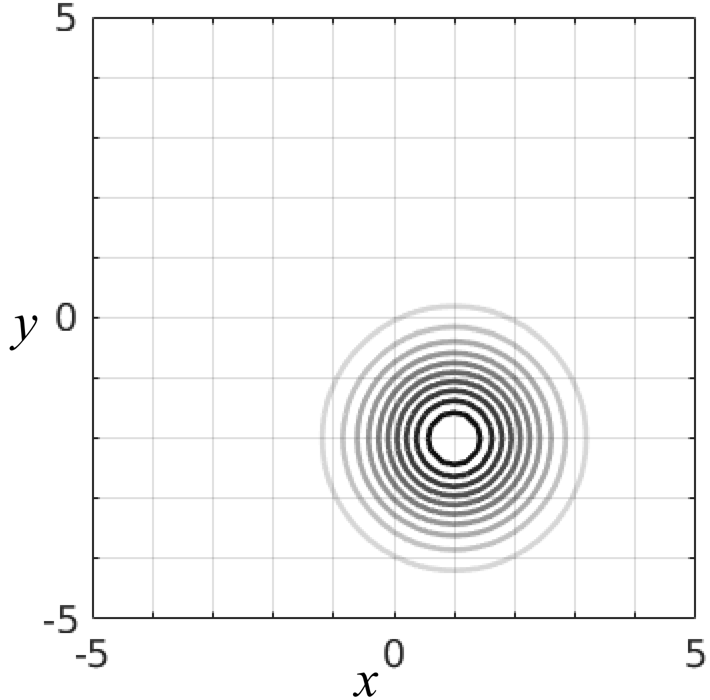
$$p(\mathbf{x}) = \frac{1}{2\pi} \exp\left\{-\frac{(x-\mu_x)^2+(y-\mu_y)^2}{2}\right\}$$

$$D = 2$$

$$\mathbf{x} = [x \quad y]^T$$

$$\boxed{\boldsymbol{\mu} = [\mu_x \quad \mu_y]^T}$$

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



Multivariate Gaussian: 2D

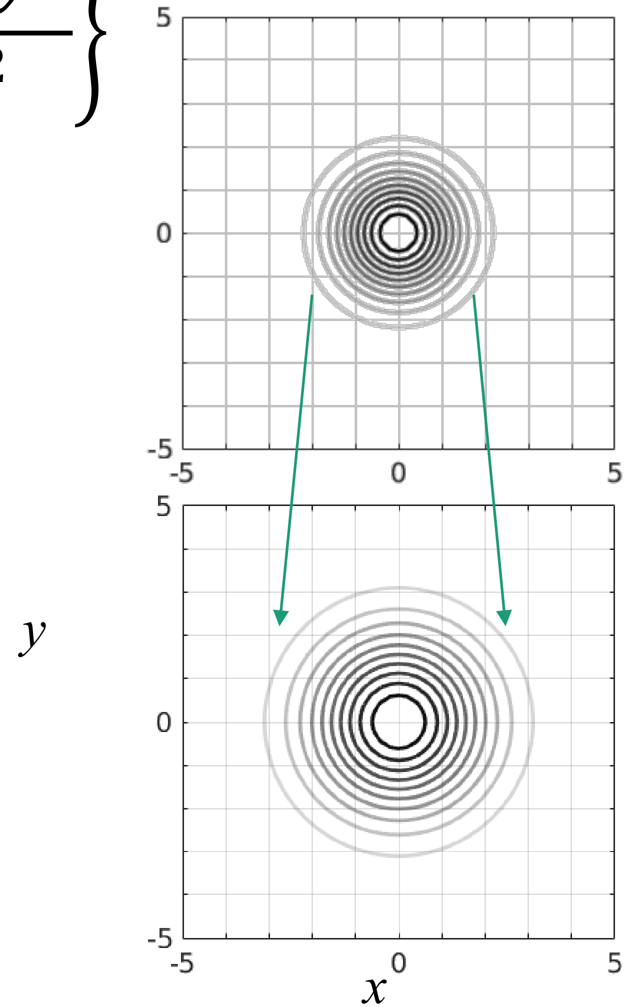
$$p(\mathbf{x}) = \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{x^2+y^2}{2\sigma^2}\right\}$$

$$D = 2$$

$$\mathbf{x} = [x \quad y]^T$$

$$\boldsymbol{\mu} = [0 \quad 0]^T$$

$$\boxed{\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}}$$



Larger σ^2

Multivariate Gaussian: 2D

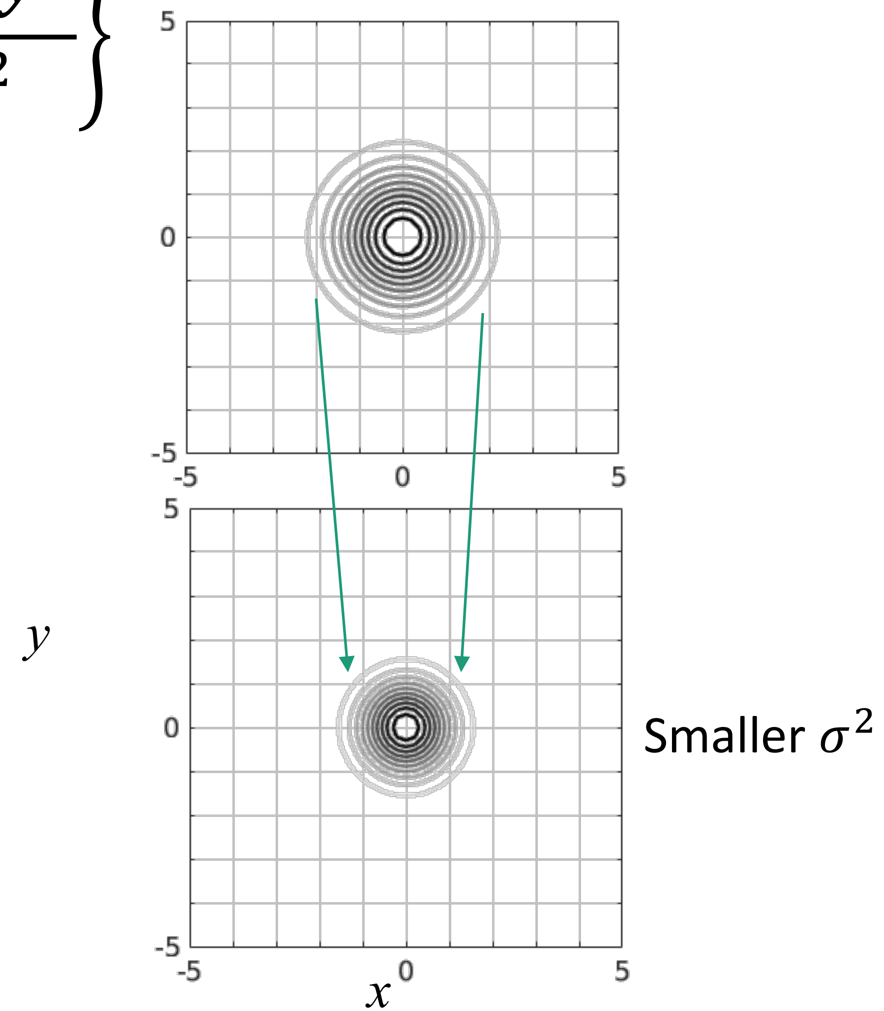
$$p(\mathbf{x}) = \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{x^2+y^2}{2\sigma^2}\right\}$$

$$D = 2$$

$$\mathbf{x} = [x \quad y]^T$$

$$\boldsymbol{\mu} = [0 \quad 0]^T$$

$$\boxed{\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}}$$



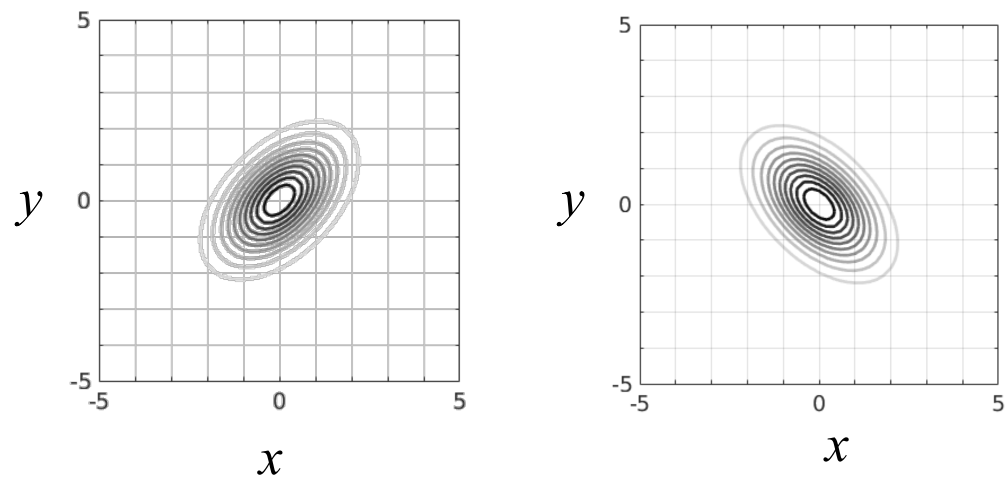
Multivariate Gaussian: 2D

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\boldsymbol{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- 2D General Case

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy}^2 \\ \sigma_{xy}^2 & \sigma_y^2 \end{bmatrix}$$

$\sigma_{xy}^2 > 0$ $\sigma_{xy}^2 < 0$



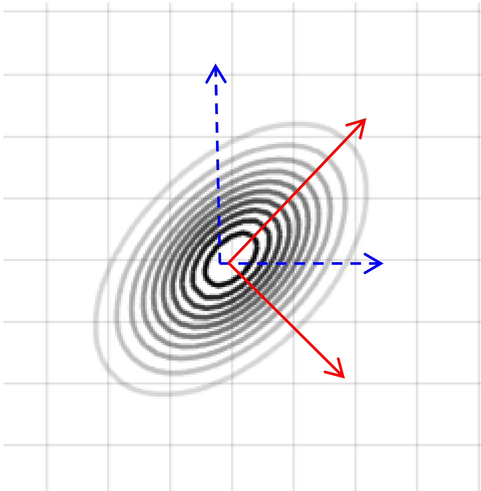
Multivariate Gaussian: 2D

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\boldsymbol{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Properties of Covariance Matrix Σ

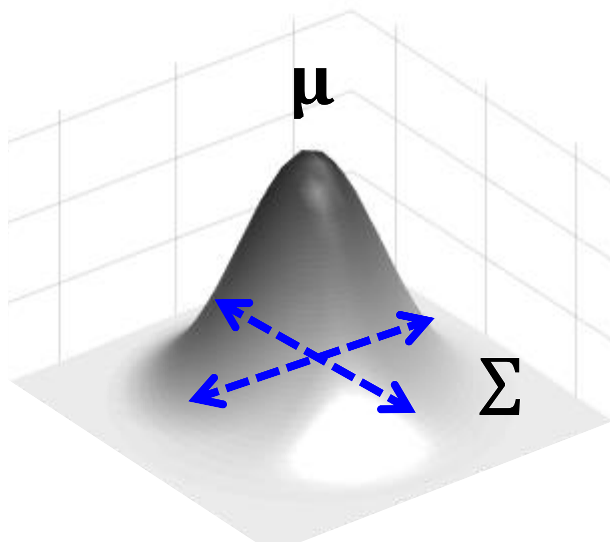
1) Σ is Symmetric and Positive Definite.

2) Diagonalization: Σ can be decomposed in the form of UDU^T.
(D is a Diagonal matrix.)



Multivariate Gaussian: 2D

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{T}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$



Robotics

Estimation and Learning
with Dan Lee

Week 1.
Gaussian Model Learning

1.3.2 MLE of Multivariate Gaussian

Penn Engineering

Maximum Likelihood Estimate of Multivariate Gaussian Parameters

- Objective

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} p(\{\mathbf{x}_i\}|\boldsymbol{\mu}, \Sigma)$$

Maximum Likelihood Estimate of Multivariate Gaussian Parameters

- Objective

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} p(\{\mathbf{x}_i\}|\boldsymbol{\mu}, \Sigma)$$

Assuming independence of observations,

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\mu}, \Sigma)$$

Maximum Likelihood Estimate of Multivariate Gaussian Parameters

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\max_{\boldsymbol{\mu}, \Sigma} \prod_{i=1}^{N} p(\mathbf{x}_i | \boldsymbol{\mu}, \Sigma)$$

(1) Take the log!

$$\arg\max \, likelihood \leftrightarrow \arg\max \ln(likelihood)$$

$$\log(x_1 \times x_2 \times \cdots \times x_k) = \log(x_1) + \log(x_1) + \cdots + \log(x_k)$$

$$\boxed{\arg\max_{\boldsymbol{\mu}, \Sigma} \prod_{i=1}^{N} p(\mathbf{x}_i | \boldsymbol{\mu}, \Sigma) \Rightarrow \arg\max_{\boldsymbol{\mu}, \Sigma} \sum_{i=1}^{N} \ln p(\mathbf{x}_i | \boldsymbol{\mu}, \Sigma)}$$

Maximum Likelihood Estimate of Multivariate Gaussian Parameters

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\max_{\boldsymbol{\mu},\Sigma} \sum_{i=1}^{N} \underline{\ln p(\mathbf{x}_i | \boldsymbol{\mu}, \Sigma)}$$

(2) Gaussian!

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{T}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

$$\boxed{\ln p(\mathbf{x}_i | \boldsymbol{\mu}, \Sigma)} \longrightarrow \left\{-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^{T}\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) - \frac{1}{2}\ln|\Sigma| + c\right\}$$

$c = -D/2 \ln(2\pi)$

Maximum Likelihood Estimate of Multivariate Gaussian Parameters

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\max_{\boldsymbol{\mu},\Sigma} \sum_{i=1}^{N} \left\{ -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^{\boldsymbol{T}} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) - \frac{1}{2} \ln|\Sigma| + c \right\}$$

$$\boxed{\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\min_{\boldsymbol{\mu},\Sigma} \sum_{i=1}^{N} \left\{ \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^{\boldsymbol{T}} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) + \frac{1}{2} \ln|\Sigma| \right\}}$$

Maximum Likelihood Estimate of Multivariate Gaussian Parameters

$$\underbrace{\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\min_{\boldsymbol{\mu},\Sigma} \sum_{i=1}^{N} \left\{ \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^{\boldsymbol{T}} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) + \frac{1}{2} \ln|\Sigma| \right\}}_{J(\boldsymbol{\mu}, \Sigma)}$$

- At optimum,

① $\dfrac{\partial J}{\partial \boldsymbol{\mu}} = \mathbf{0} \longrightarrow \hat{\boldsymbol{\mu}}$ ② $\dfrac{\partial J(\hat{\boldsymbol{\mu}}, \Sigma)}{\partial \Sigma} = 0 \longrightarrow \hat{\Sigma}$

Maximum Likelihood Estimate of Multivariate Gaussian Parameters

- In summary, we have

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

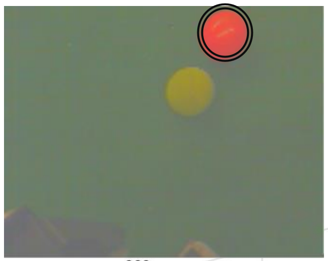
$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\intercal$$

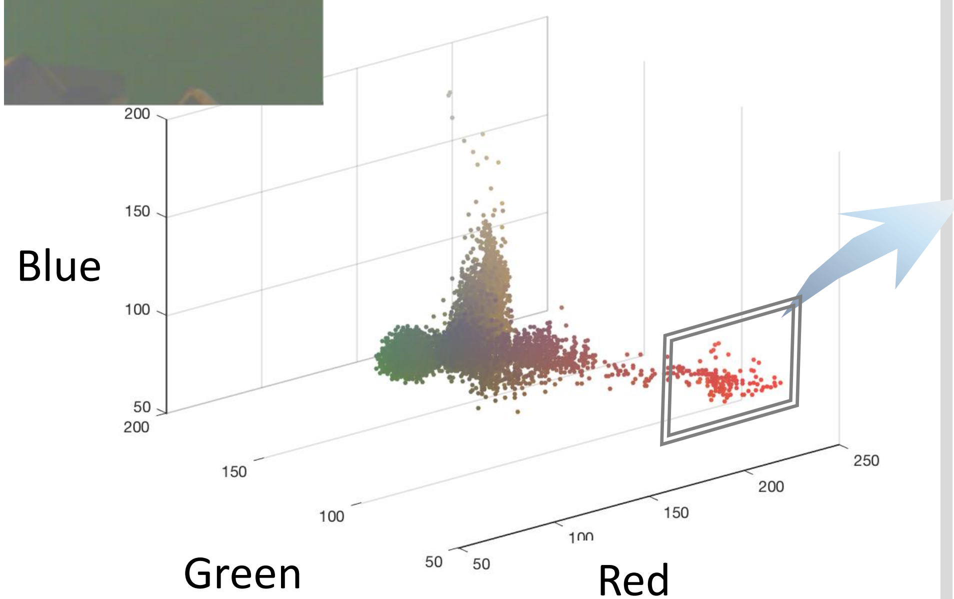
Multi-dimension Distribution: Example

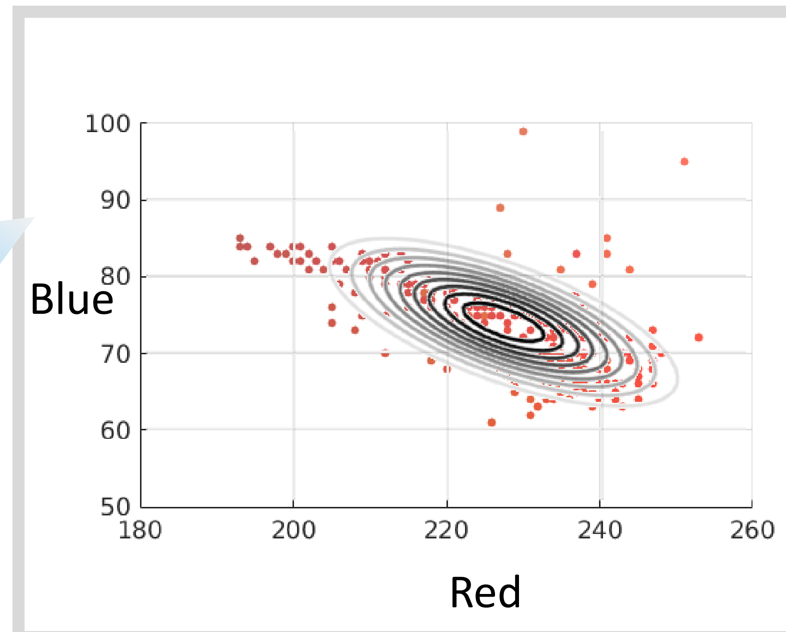
- Ball color in multi-channels

$$\hat{\boldsymbol{\mu}} = [227 \quad 74]$$

$$\hat{\Sigma} = \begin{bmatrix} 150 & -48 \\ -48 & 46 \end{bmatrix}$$







Robotics

Estimation and Learning
with Dan Lee

Supplementary Notes on
MLE for Multivariate Gaussian

Penn Engineering

Maximum Likelihood Estimate of Multivariate Gaussian Parameters

- Objective

  Estimate the mean and the variance given observed data

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\max_{\boldsymbol{\mu}, \Sigma} p(\{\mathbf{x}_i\} | \boldsymbol{\mu}, \Sigma)$$

- Assuming independence of observations,

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\max_{\boldsymbol{\mu}, \Sigma} \prod_{i=1}^{N} p(\mathbf{x}_i | \boldsymbol{\mu}, \Sigma)$$

Maximum Likelihood Estimate of Multivariate Gaussian Parameters

- To obtain $\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\max_{\boldsymbol{\mu},\Sigma} \prod_{i=1}^{N} p(\mathbf{x}_i | \boldsymbol{\mu}, \Sigma)$

(1) $\arg\max_{\boldsymbol{\mu},\Sigma} \prod_{i=1}^{N} p(\mathbf{x}_i | \boldsymbol{\mu}, \Sigma) = \arg\max_{\boldsymbol{\mu},\Sigma} \sum_{i=1}^{N} \ln p(\mathbf{x}_i | \boldsymbol{\mu}, \Sigma)$

(2) $p(\mathbf{x}) = \dfrac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left\{ -\dfrac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$

Maximum Likelihood Estimate of Multivariate Gaussian Parameters

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\max_{\boldsymbol{\mu},\Sigma} \prod_{i=1}^{N} p(\mathbf{x}_i | \boldsymbol{\mu}, \Sigma)$$

$$= \arg\max_{\mu,\sigma} \sum_{i=1}^{N} \ln p(\mathbf{x}_i | \boldsymbol{\mu}, \Sigma)$$

from (1)

$$= \arg\max_{\mu,\sigma} \sum_{i=1}^{N} \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^{\boldsymbol{T}} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) - \frac{1}{2} \ln|\Sigma| + c \right\}$$

from (2)

$$= \arg\min_{\mu,\sigma} \sum_{i=1}^{N} \left\{ \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^{\boldsymbol{T}} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) + \frac{1}{2} \ln|\Sigma| \right\}$$

Maximum Likelihood Estimate of Multivariate Gaussian Parameters

Let $J(\boldsymbol{\mu}, \Sigma) = \sum_{i=1}^{N} \left\{ \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) + \frac{1}{2} \ln|\Sigma| \right\}$

Then $\widehat{\boldsymbol{\mu}}, \widehat{\Sigma} = \arg\min_{\boldsymbol{\mu}, \Sigma} J(\boldsymbol{\mu}, \Sigma)$

① $\frac{\partial J}{\partial \boldsymbol{\mu}} = \mathbf{0} \longrightarrow \widehat{\boldsymbol{\mu}}$ ② $\frac{\partial J(\widehat{\boldsymbol{\mu}}, \Sigma)}{\partial \Sigma} = 0 \longrightarrow \widehat{\Sigma}$

Maximum Likelihood Estimate of Multivariate Gaussian Parameters

① $$\boxed{\frac{\partial J}{\partial \boldsymbol{\mu}} = \mathbf{0} \longrightarrow \hat{\boldsymbol{\mu}}}$$

$$\frac{\partial J}{\partial \boldsymbol{\mu}} = \frac{\partial}{\partial \boldsymbol{\mu}} \sum_{i=1}^{N} \left\{ \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) + \frac{1}{2} |\Sigma| \right\}$$

$$= \frac{\partial}{\partial \boldsymbol{\mu}} \sum_{i=1}^{N} \left\{ \frac{1}{2} \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} - \boldsymbol{\mu}^T \Sigma^{-1} \mathbf{x}_i \right\}$$

$$= \Sigma^{-1} \sum_{i=1}^{N} \{ \boldsymbol{\mu} - \mathbf{x}_i \} = \mathbf{0} \longrightarrow \hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

Maximum Likelihood Estimate of Multivariate Gaussian Parameters

② $$\boxed{\frac{\partial J(\hat{\boldsymbol{\mu}}, \Sigma)}{\partial \Sigma} = 0 \longrightarrow \hat{\Sigma}}$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^{\top}$$

$$\frac{\partial J}{\partial \Sigma} = \frac{\partial}{\partial \Sigma}\sum_{i=1}^{N}\left\{\frac{1}{2}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^{T}\Sigma^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}) + \frac{1}{2}\ln|\Sigma|\right\}$$

$$= \frac{1}{2}\sum_{i=1}^{N}\{-\Sigma^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^{T}\Sigma^{-1} + \Sigma^{-1}\}$$

$$= \frac{1}{2}\Sigma^{-1}\left[-\left\{\sum_{i=1}^{N}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^{T}\right\}\Sigma^{-1} + N\cdot \mathrm{I}\right] = 0$$

Maximum Likelihood Estimate of Multivariate Gaussian Parameters

- In summary, we have

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^{\intercal}$$

Robotics

Estimation and Learning
with Dan Lee

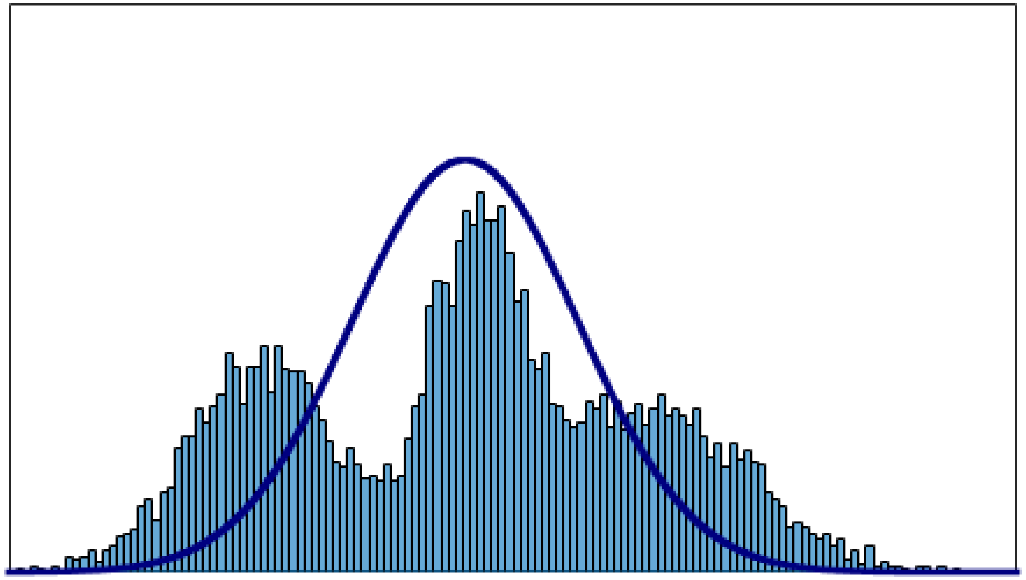
Week 1.
Gaussian Model Learning

1.4.1 Gaussian Mixture Model

Penn Engineering

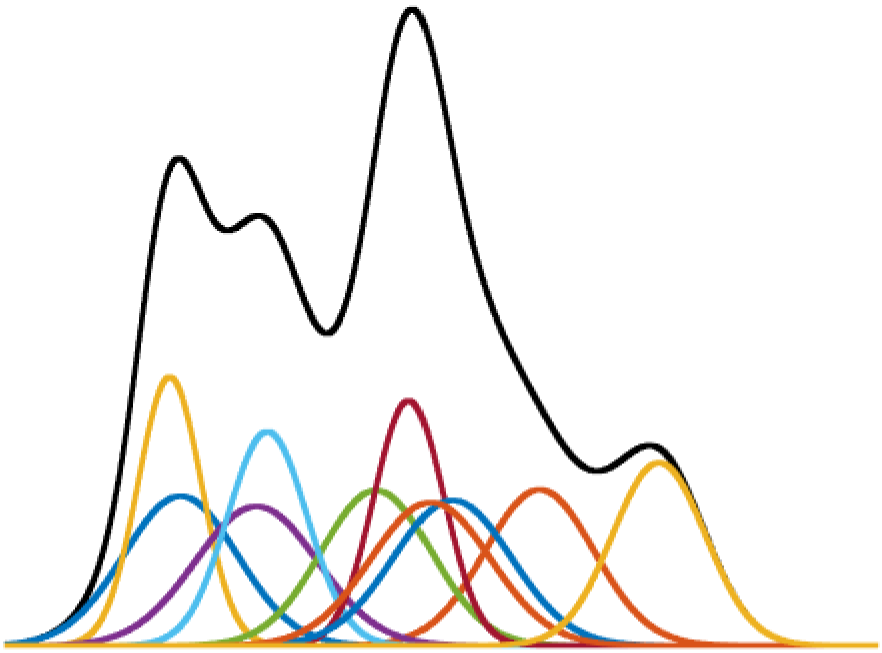
Limitations of Single Gaussian

- Single Mode
- Symmetric



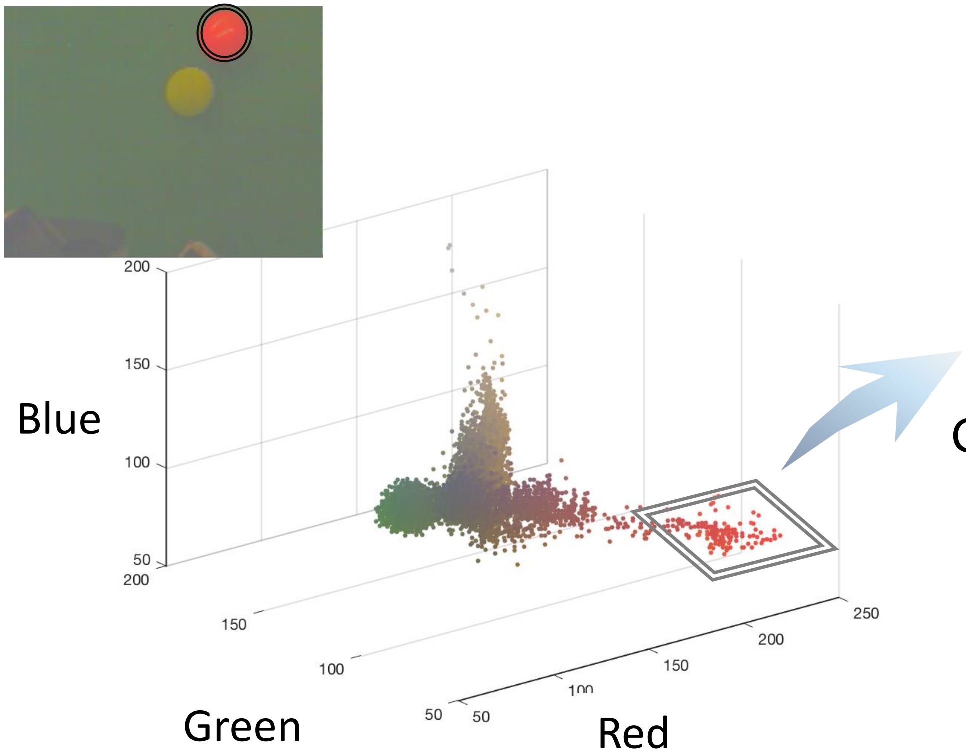
Gaussian Mixture Model

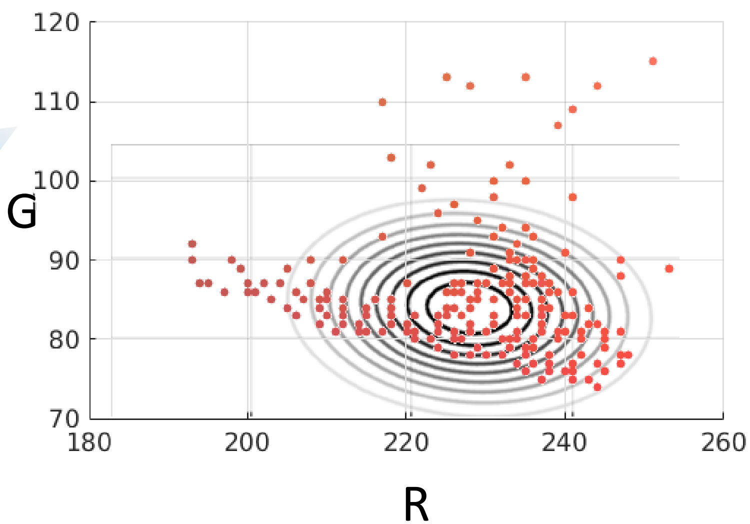
- Mixture (=**Sum**) of Gaussians



Multi-dimension Distribution: Example

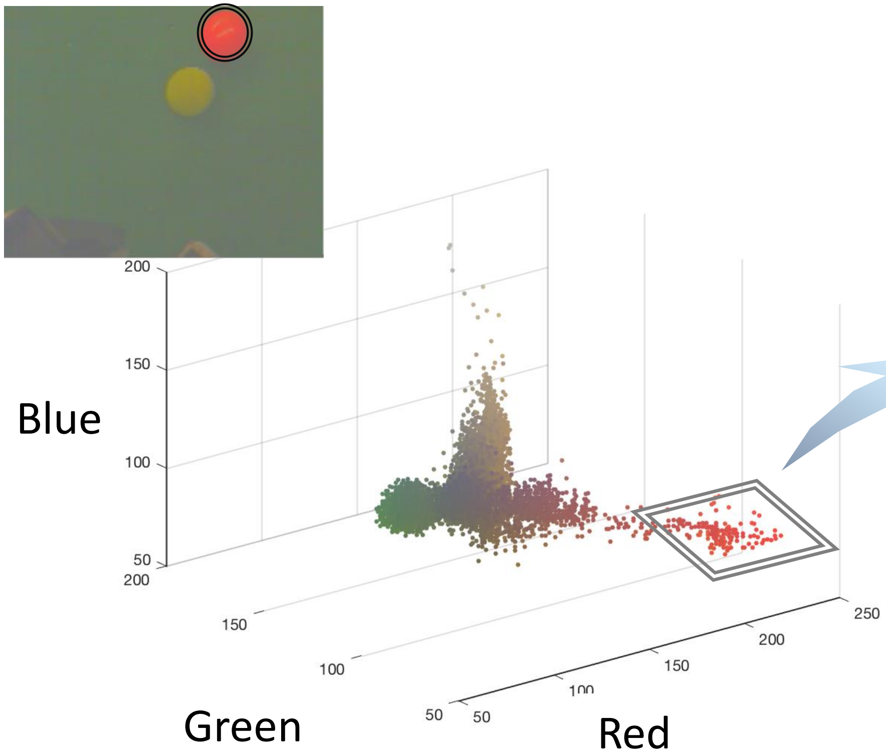
- Ball color in multi-channels

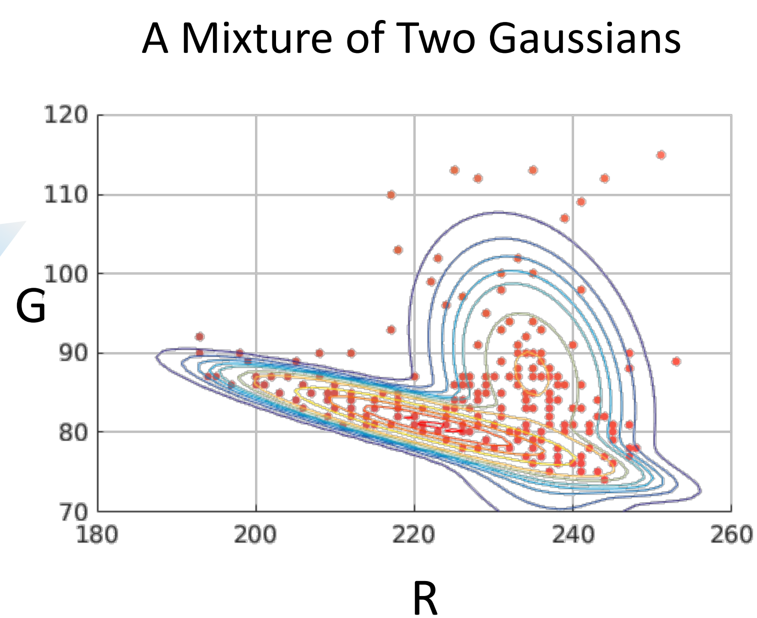




Multi-dimension Distribution: Example

- Ball color in multi-channels





Gaussian Mixture Model

- Mixture of Gaussians

$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k g_k(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

g_k: Gaussian with $\boldsymbol{\mu}_k$ and Σ_k

w_k: mixing coefficient (weight, a prior) $w_k > 0, \sum_{k=1}^{K} w_k = 1$

Using GMM

Flexibility

Parameters ↑

- No analytic solution
- Overfitting

$$\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K\}$$

$$\Sigma = \{\Sigma_1, \Sigma_2, \dots, \Sigma_K\}$$

$$\boldsymbol{w} = \{w_1, w_2, \dots, w_K\}$$

K: Number of Components

Using GMM

Flexibility

\# Parameters ↑

- No analytic solution
- Overfitting

$$\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K\}$$

$$\Sigma = \{\Sigma_1, \Sigma_2, \dots, \Sigma_K\}$$

$$\boldsymbol{w} = w = 1/K$$

K: given number of Components

Robotics

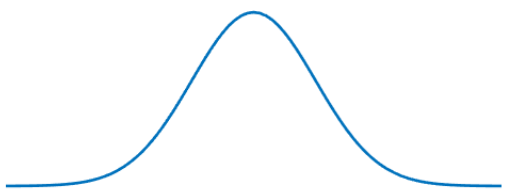
Estimation and Learning
with Dan Lee

Week 1.
Gaussian Model Learning

1.4.2 GMM Parameter Estimation via EM

Penn Engineering

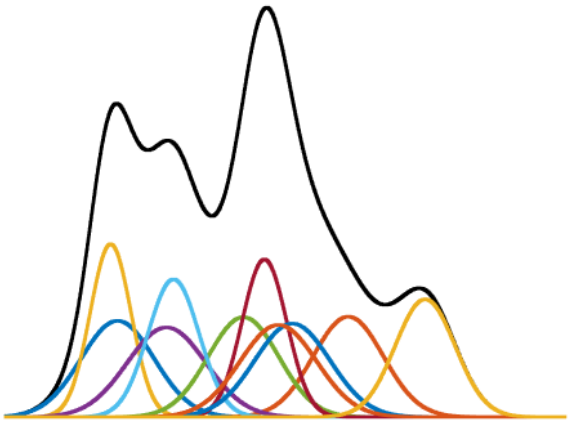
- Single Gaussian



$\boldsymbol{\mu}$

Σ

- Mixture of Gaussians



$$\boxed{\begin{aligned}\boldsymbol{\mu} &= \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K\} \\ \Sigma &= \{\Sigma_1, \Sigma_2, \dots, \Sigma_K\}\end{aligned}}$$

$w = 1/K$

K: Given

Learning GMM Parameters

Likelihood: $p(\{\mathbf{x}_i\}|\boldsymbol{\mu}, \Sigma)$

Observed data　　Unknown parameters

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\max_{\boldsymbol{\mu}, \Sigma} p(\{\mathbf{x}_i\}|\boldsymbol{\mu}, \Sigma)$$

$$\boldsymbol{\mu} = \{\boldsymbol{\mu}_k\}$$

$$\Sigma = \{\Sigma_k\} \quad k = 1, 2, \ldots, K$$

Learning GMM Parameters

- Objective

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} p(\{\mathbf{x}_i\} | \boldsymbol{\mu}, \Sigma)$$

Assuming independence of observations,

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} \prod_{i=1}^{N} p(\mathbf{x}_i | \boldsymbol{\mu}, \Sigma)$$

Learning GMM Parameters

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\max_{\boldsymbol{\mu},\Sigma} \prod_{i=1}^{N} p(\mathbf{x}_i | \boldsymbol{\mu}, \Sigma)$$

(1) Take the log!

$$\arg\max \, likelihood \leftrightarrow \arg\max \ln(likelihood)$$

$$\log(x_1 \times x_2 \times \cdots \times x_k) = \log(x_1) + \log(x_1) + \cdots + \log(x_k)$$

$$\boxed{\arg\max_{\boldsymbol{\mu},\Sigma} \prod_{i=1}^{N} p(\mathbf{x}_i | \boldsymbol{\mu}, \Sigma) \Rightarrow \arg\max_{\boldsymbol{\mu},\Sigma} \sum_{i=1}^{N} \ln p(\mathbf{x}_i | \boldsymbol{\mu}, \Sigma)}$$

Learning GMM Parameters

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \underset{\boldsymbol{\mu},\Sigma}{\arg\max} \sum_{i=1}^{N} \ln \underline{p(\mathbf{x}_i | \boldsymbol{\mu}, \Sigma)}$$

(2) Gaussian Mixture Model!

$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k g_k(\mathbf{x} | \boldsymbol{\mu}_k, \Sigma_k)$$

g_k: Gaussian with $\boldsymbol{\mu}_k$ and Σ_k

$$\boxed{\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \underset{\boldsymbol{\mu},\Sigma}{\arg\max} \sum_{i=1}^{N} \ln \left\{ \frac{1}{K} \sum_{k=1}^{K} g_k(\mathbf{x}_i | \boldsymbol{\mu}_k, \Sigma_k) \right\}}$$

Learning GMM Parameters

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \underset{\boldsymbol{\mu}, \Sigma}{\arg\max} \sum_{i=1}^{N} \ln \left\{ \frac{1}{K} \sum_{k=1}^{K} g_k(\mathbf{x}_i | \boldsymbol{\mu}_k, \Sigma_k) \right\}$$

where

$$g_k(\mathbf{x}) = \frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^{\boldsymbol{T}} {\Sigma_k}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\}$$

→ No closed form solution exist.

Expectation-Maximization (EM)

1) Special case : EM for GMM Parameter Estimation
2) General EM Algorithm

EM for GMM

Initial $\boldsymbol{\mu}$ and Σ

Latent variable z

EM for GMM

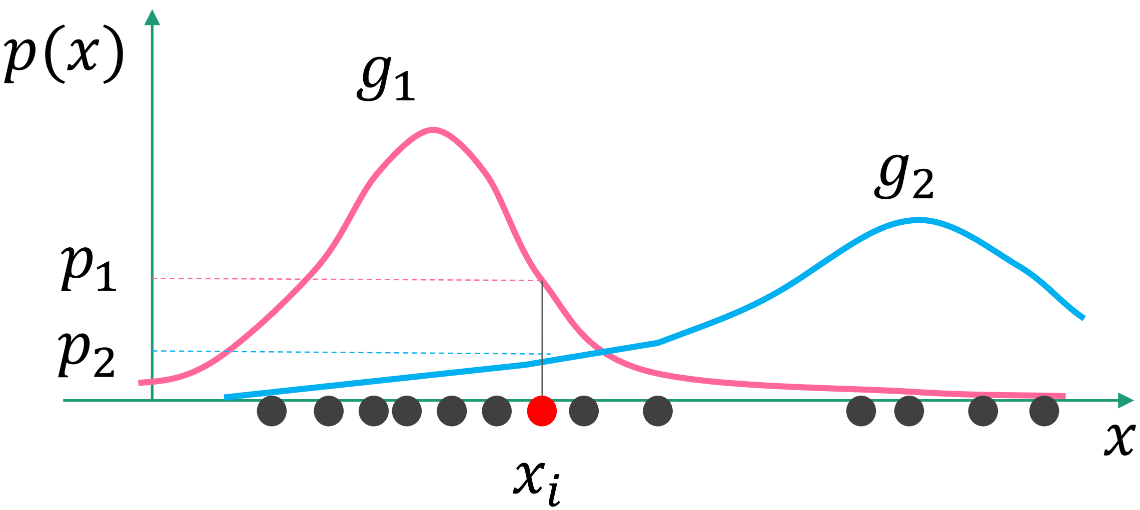
- Latent Variable

$$z_k^i = \frac{g_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{k=1}^{K} g_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)}$$

EM for GMM

- Latent Variable Example

$$z_k^i = \frac{g_k(\mathbf{x}_i | \boldsymbol{\mu}_k, \Sigma_k)}{g_1(\mathbf{x}_i | \boldsymbol{\mu}_1, \Sigma_1) + g_2(\mathbf{x}_i | \boldsymbol{\mu}_2, \Sigma_2)}$$



$$z_1^i = \frac{p_1}{p_1 + p_2}$$

$$z_2^i = \frac{p_2}{p_1 + p_2}$$

EM for GMM

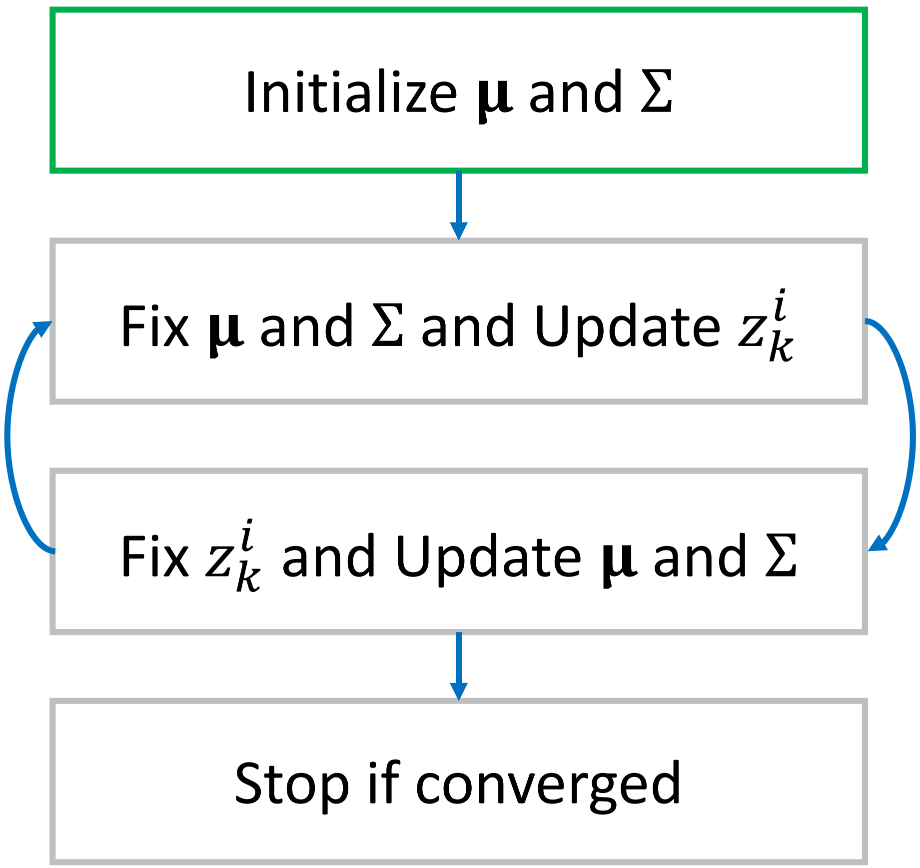
- Mean and Covariance Matrix

$$\widehat{\boldsymbol{\mu}}_k = \frac{1}{z_k} \sum_{i=1}^{N} z_k^i \mathbf{x}_i$$

$$\widehat{\Sigma}_k = \frac{1}{z_k} \sum_{i=1}^{N} z_k^i (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)^{\top}$$

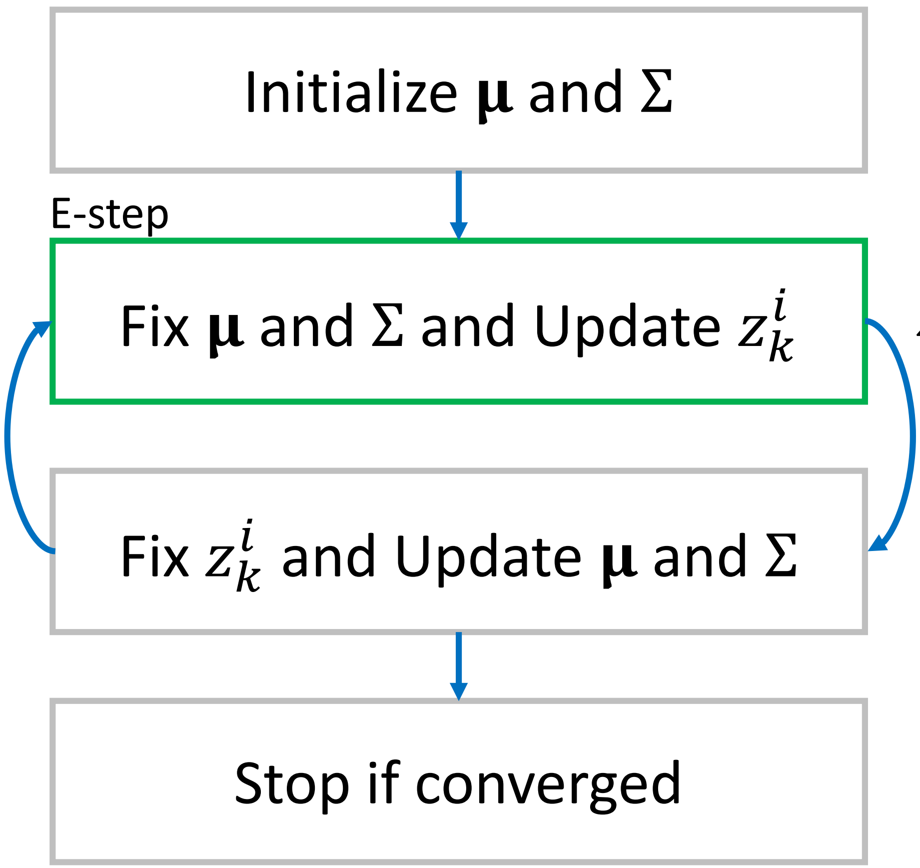
$$z_k = \sum\nolimits_{i=1}^{N} z_k^i$$

EM for GMM



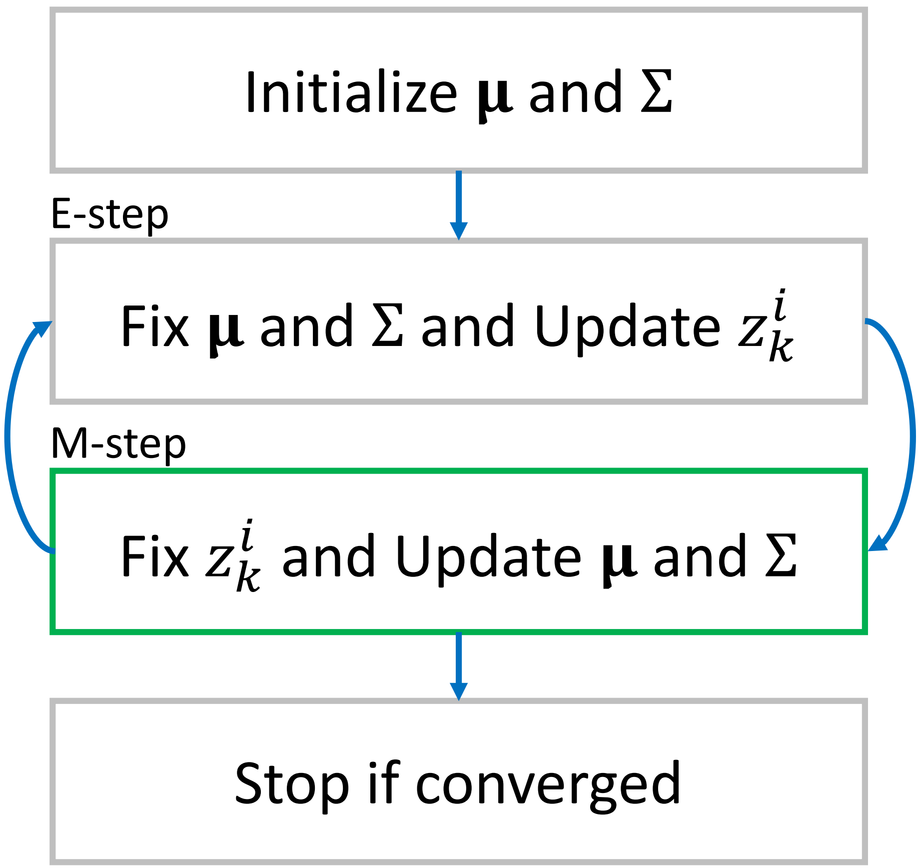
Initialize $\boldsymbol{\mu}$ and Σ
Fix $\boldsymbol{\mu}$ and Σ and Update z_k^i
Fix z_k^i and Update $\boldsymbol{\mu}$ and Σ
Stop if converged

EM for GMM



$$z_k^i = \frac{g_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{k=1}^{K} g_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)}$$

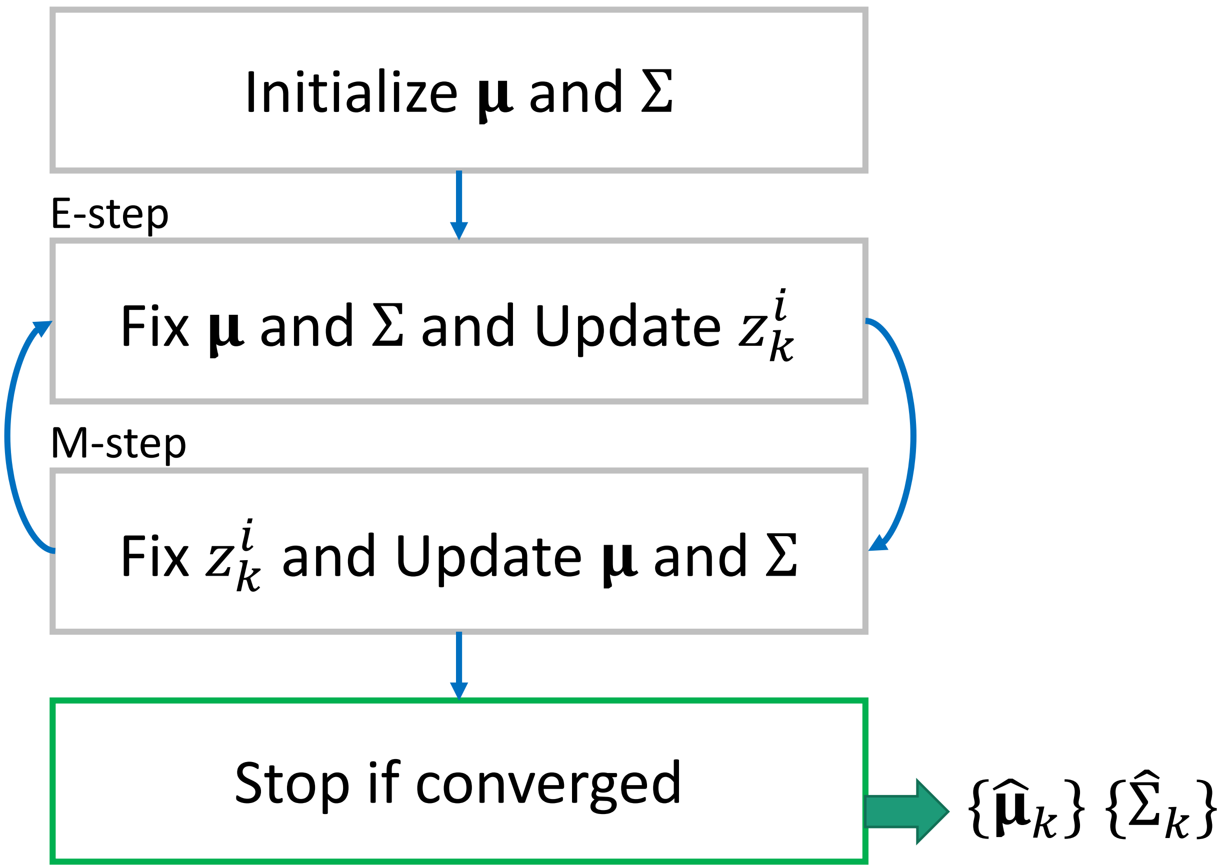
EM for GMM



$$\hat{\boldsymbol{\mu}}_k = \frac{1}{z_k}\sum_{i=1}^{N} z_k^i \mathbf{x}_i$$

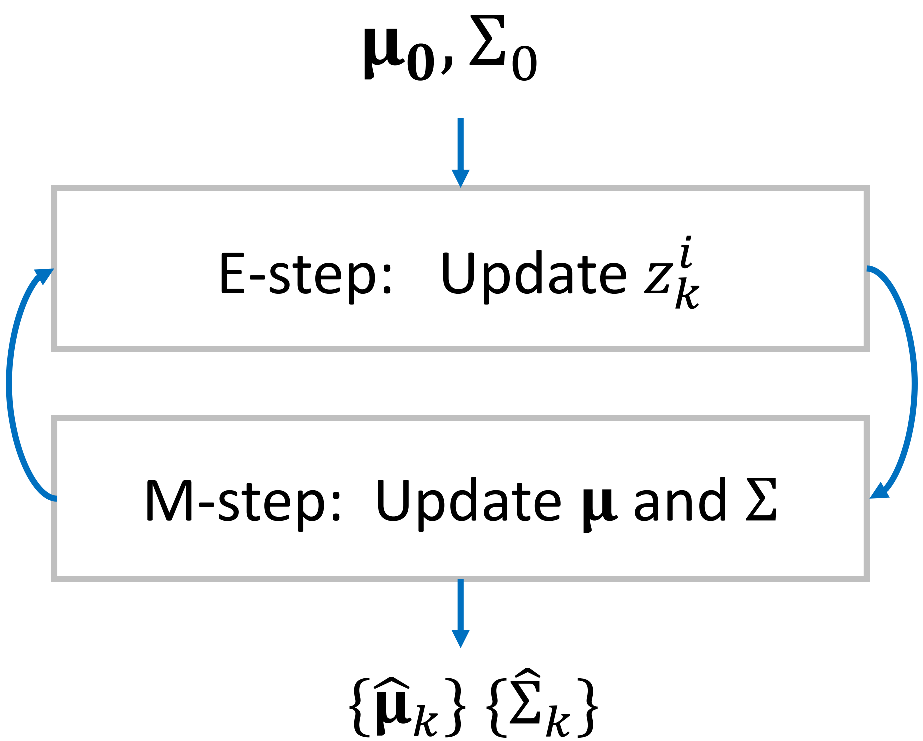
$$\hat{\Sigma}_k = \frac{1}{z_k}\sum_{i=1}^{N} z_k^i (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^{\top}$$

EM for GMM



Initialize $\boldsymbol{\mu}$ and Σ
E-step
Fix $\boldsymbol{\mu}$ and Σ and Update z_k^i
M-step
Fix z_k^i and Update $\boldsymbol{\mu}$ and Σ
Stop if converged
$\{\hat{\boldsymbol{\mu}}_k\}$ $\{\hat{\Sigma}_k\}$

EM for GMM


$\boldsymbol{\mu}_{\mathbf{0}}, \Sigma_0$
E-step: Update z_k^i
M-step: Update $\boldsymbol{\mu}$ and Σ
$\{\widehat{\boldsymbol{\mu}}_k\}$ $\{\widehat{\Sigma}_k\}$

Week 1.
Gaussian Model Learning

1.4.3. EM Algorithm [Advanced]

Robotics

Estimation and Learning
with Dan Lee

Week 1.
Gaussian Model Learning

1.4.3 EM Algorithm [Advanced]

Penn Engineering

Expectation-Maximization (EM)

- EM as lower-bound maximization

$$\arg\max_{\theta} \sum_i \ln p(x_i|\theta)$$

θ : All parameters

(1) Jensen's inequality

(2) Latent variable and marginal probability

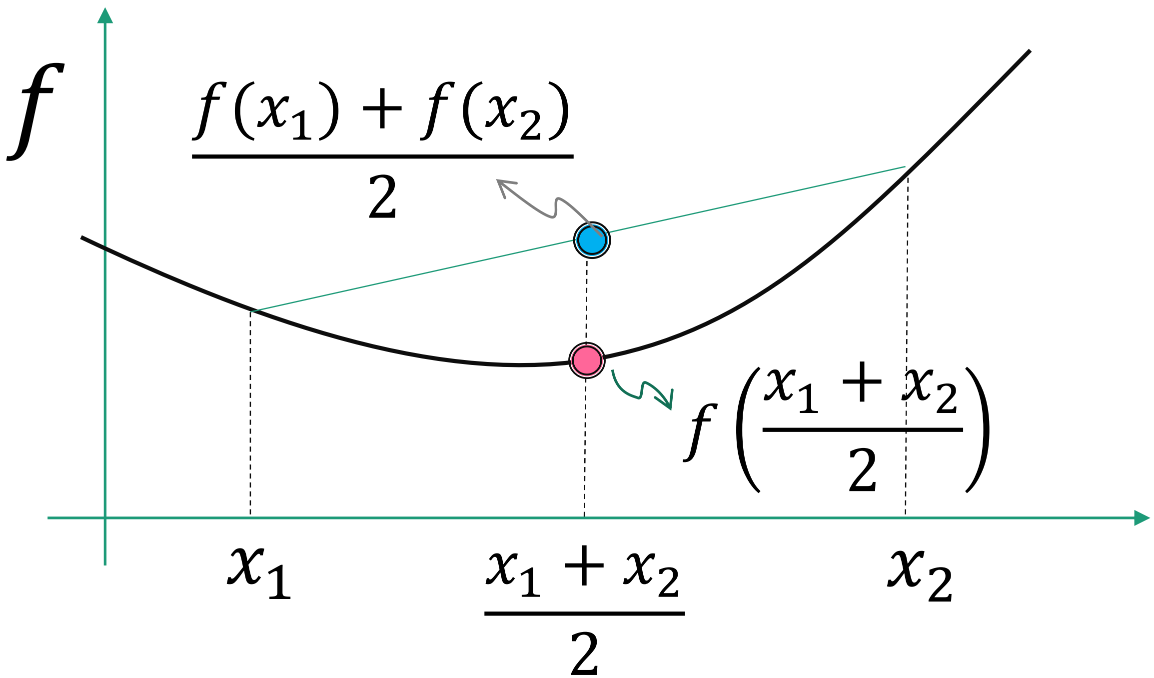
(3) Procedure : E-step and M-step.

Expectation-Maximization (EM)

(1) Jensen's inequality

f : **convex** function

$$\boxed{f\left(\frac{x_1+x_2}{2}\right) \le \frac{f(x_1)+f(x_2)}{2}}$$



Expectation-Maximization (EM)

(1) Jensen's inequality

f : **convex** function

$$f\left(\sum a_i x_i\right) \leq \sum a_i f(x_i)$$

$$\sum a_i = 1$$

$$a_i \geq 0$$

Expectation-Maximization (EM)

(1) Jensen's inequality

f : **concave** function

$$f\left(\sum a_i x_i\right) \geq \sum a_i f(x_i)$$

$$\sum a_i = 1$$

$$a_i \geq 0$$

Expectation-Maximization (EM)

(1) Jensen's inequality

$\ln$ is *concave*

$$\ln\left(\sum a_i p_i\right) \geq \sum a_i \ln p_i$$

$$\sum a_i = 1$$

$$a_i \geq 0$$

Expectation-Maximization (EM)

(2) latent variable z

$$p(X|\theta) = \sum_{Z} p(X, Z|\theta)$$

(From definition of marginal probability)

Expectation-Maximization (EM)

(2) latent variable

$$\ln p(X|\theta) = \ln \sum_{Z} p(X, Z|\theta)$$

Log-likelihood

$$= \ln \sum_{Z} q(Z) \frac{p(X, Z|\theta)}{q(Z)} \geq \sum_{Z} q(Z) \ln \frac{p(X, Z|\theta)}{q(Z)}$$

Lower bound

Note: $q(Z)$ is a valid probability distribution over Z.

Expectation-Maximization (EM)

(2) latent variable

$$\ln p(X|\theta) = \ln \sum_{Z} p(X, Z|\theta) \;\geq \sum_{Z} q(Z) \ln \frac{p(X, Z|\theta)}{q(Z)}$$

Log-likelihood

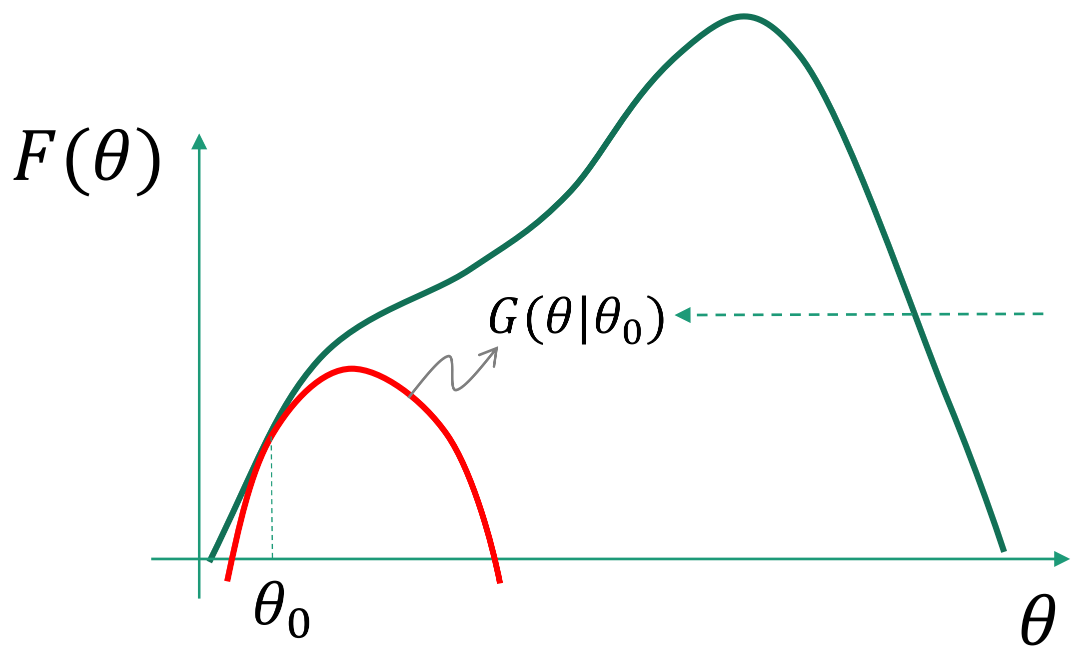
Lower bound

Find q!

Expectation-Maximization (EM)

(3a) Find a lower bound G with an initial guess

$$\underbrace{\ln p(X|\theta)}_{F} \geq \underbrace{\sum_{Z} q(Z) \ln \frac{p(X,Z|\theta)}{q(Z)}}_{G}$$

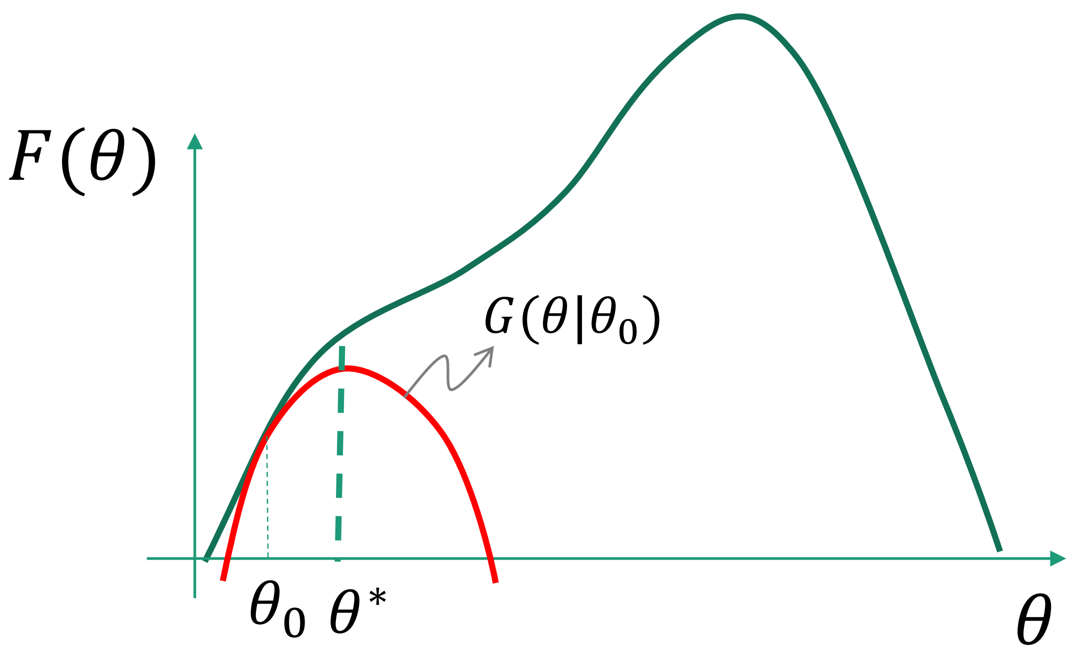


This Implies that $q(Z)$ is found based on θ_0.

Expectation-Maximization (EM)

(3b) Find $\theta^* = \arg\max_{\theta} G(\theta|\theta_0)$

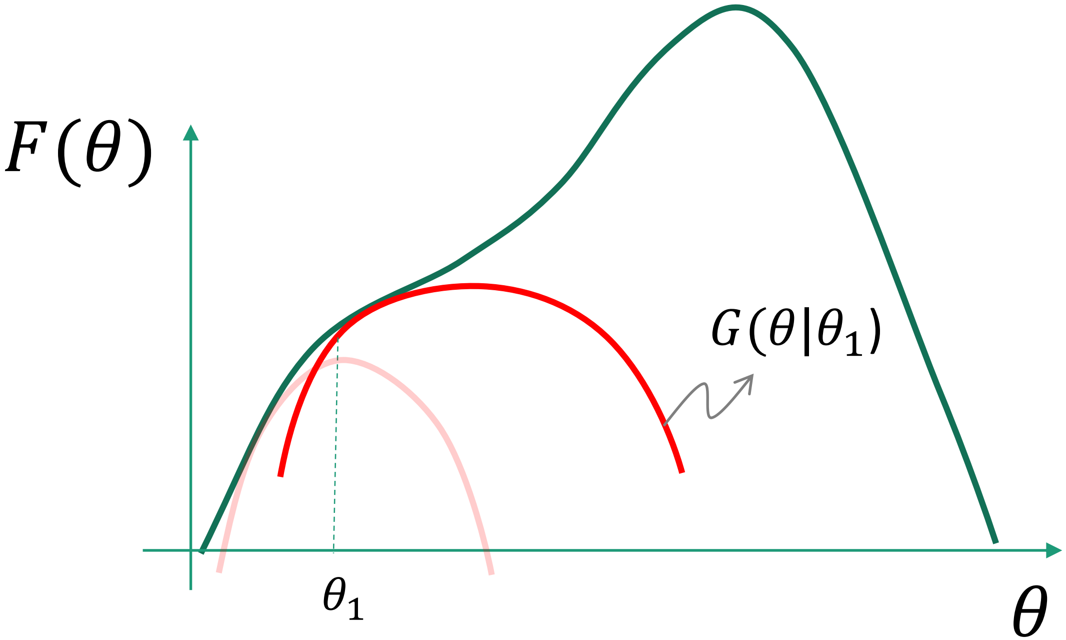
$$\underbrace{\ln p(X|\theta)}_{F} \geq \underbrace{\sum_{Z} q(Z) \ln \frac{p(X,Z|\theta)}{q(Z)}}_{G}$$



Expectation-Maximization (EM)

(3c) Find a new lower bound G with $\theta_1 \leftarrow \theta^*$

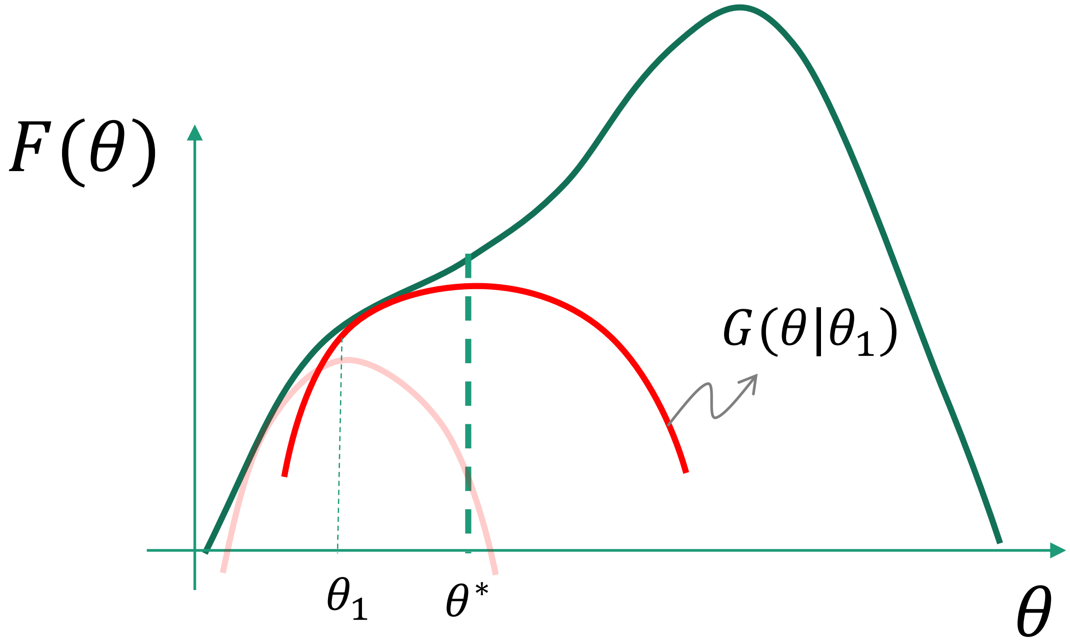
$$\underbrace{\ln p(X|\theta)}_{F} \geq \underbrace{\sum_{Z} q(Z) \ln \frac{p(X,Z|\theta)}{q(Z)}}_{G}$$



Expectation-Maximization (EM)

(3d) Find $\theta^* = \arg\max_{\theta} G(\theta|\theta_1)$

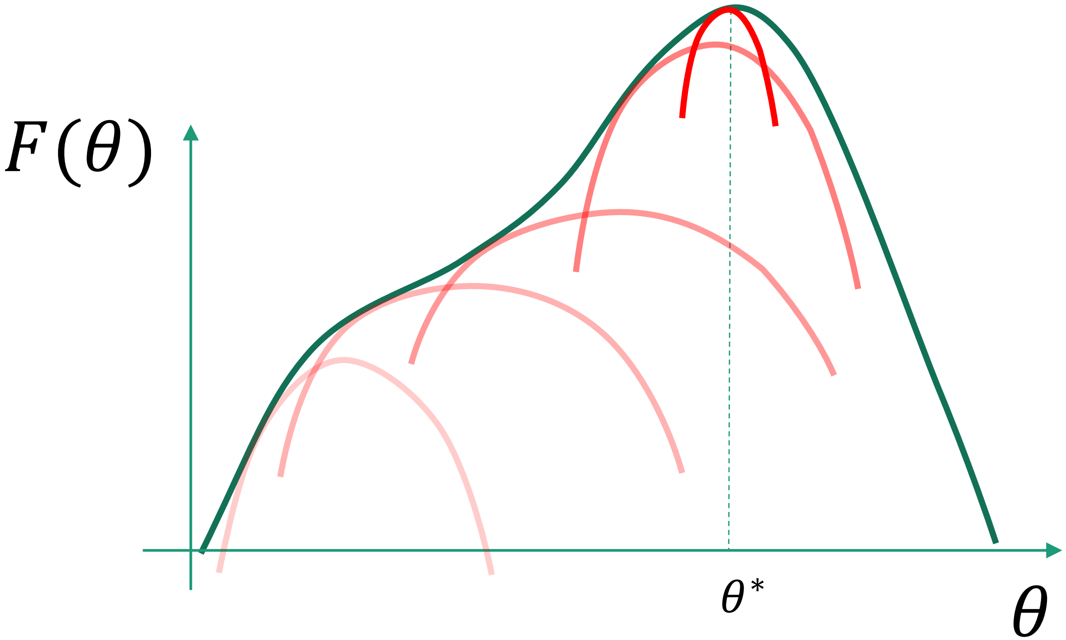
$$\underbrace{\ln p(X|\theta)}_{F} \geq \underbrace{\sum_{Z} q(Z) \ln \frac{p(X,Z|\theta)}{q(Z)}}_{G}$$



Expectation-Maximization (EM)

(4e) Repeat (until converged)

$$\underbrace{\ln p(X|\theta)}_{F} \geq \underbrace{\sum_{Z} q(Z) \ln \frac{p(X,Z|\theta)}{q(Z)}}_{G}$$



Expectation-Maximization (EM)

Initialize $\hat{\theta}$

E-step

Find a lower bound $G(\theta|\hat{\theta})$

This Implies that $q(Z)$ is found based on $\hat{\theta}$.

$$\hat{\theta}_{new} = \arg\max_{\theta} G(\theta|\hat{\theta})$$

Stop if converged

Expectation-Maximization (EM)

Initialize $\hat{\theta}$

Find a lower bound $G(\theta|\hat{\theta})$

M-step

$$\hat{\theta}_{new} = \arg\max_{\theta} G(\theta|\hat{\theta})$$

Stop if converged