ECE 595: Machine Learning I
Tutorial 02: Probability Review

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE
UNIVERSITY

Outline

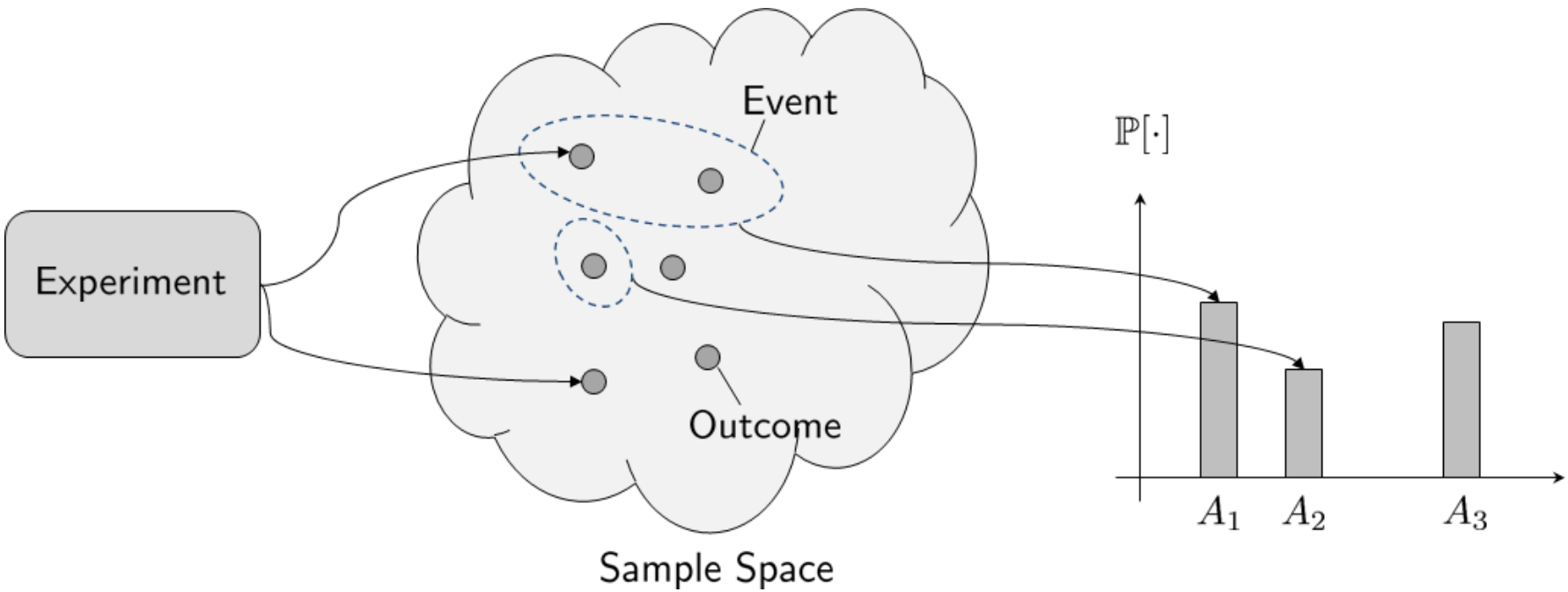
- Probability Distributions
- High-dimensional Gaussian

Reference:

- Dimitri Bertsekas, Introduction to Probability, Athena Scientific, 2008, 2nd Edition.
- Purdue ECE 302 Course Note
  `https://engineering.purdue.edu/ChanGroup/ECE302`

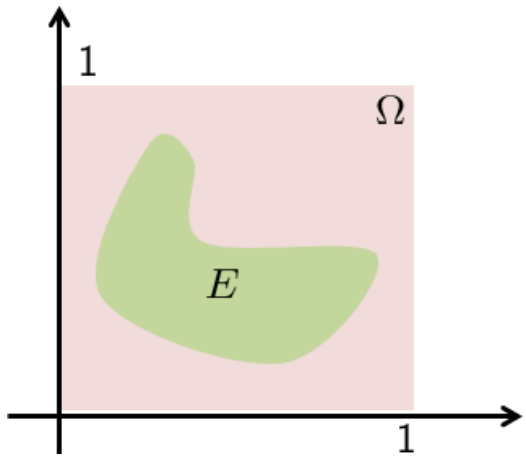
Probability Space

- Sample space Ω = set of all possible outcomes.
- Event Space $\mathcal{E}$ = set of all events. Event is a subset in Ω.
- Probability Law $\mathbb{P}$ = a mapping from $\mathcal{E}$ to $[0, 1]$.



Interpreting Probability

$\mathbb{P}[\cdot]$ is a **measure** of the size of the event.



- $\Omega = \{1, 2, 3, 4, 5, 6\}$
- $E = \{1\}$, $\mathbb{P}[E] = 1/6$
- $E = \{1, 3\}$, $\mathbb{P}[E] = 2/6$

- $\mathbb{P}[E_1] \leq \mathbb{P}[E_2]$ if $E_1 \subseteq E_2$.

- $\Omega = [0, 1] \times [0, 1]$
- $E =$ shaded region, $\mathbb{P}[E] =$ area.
- $E = \{(x_0, y_0)\}$, $\mathbb{P}[E] = 0$.

- $\mathbb{P}[E]$ can be 0 even if $E \neq \emptyset$.

Probability Axioms

- **Non-negative**:

$$\mathbb{P}[E] \geq 0$$

- **Unity**:

$$\mathbb{P}[\Omega] = 1$$

- **Additivity**: If A_n's are disjoint, then

$$\mathbb{P}\left[\bigcup_{n=1}^{N} A_n\right] = \sum_{n=1}^{N} \mathbb{P}[A_n]$$

- If A_n's are not disjoint, then Union bound holds

$$\mathbb{P}\left[\bigcup_{n=1}^{N} A_n\right] \leq \sum_{n=1}^{N} \mathbb{P}[A_n]$$

Conditional Probability

- Conditional probability of A given B is

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

- Bayes Theorem:

$$\mathbb{P}[B_i \mid A] = \frac{\mathbb{P}[A \mid B_i]\mathbb{P}[B_i]}{\underbrace{\mathbb{P}[A]}_{\uparrow}}$$

$$\mathbb{P}[A] = \sum_{n=1}^{N} \mathbb{P}[A \mid B_n]\mathbb{P}[B_n]$$

- Therefore,

$$\mathbb{P}[B_i \mid A] = \frac{\mathbb{P}[A \mid B_i]\mathbb{P}[B_i]}{\sum_{n=1}^{N} \mathbb{P}[A \mid B_n]\mathbb{P}[B_n]}$$

Random Variable

- Scalar Case: $X : \mathcal{E} \to \mathbb{R}$
  - **Encode** an event to a number.
  - E.g., $X = 1$ represents "Democrat", $X = 0$ represents "Republican".
  - Each number is called a **state**.
  - $\mathbb{P}[X = x]$ = probability of X have the state x.
- Vector Case: $\boldsymbol{X} : \mathcal{E}_1 \times \ldots \times \mathcal{E}_N \to \mathbb{R}^N$.
  - Encode a sequence of events to a sequence of numbers.

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix}$$

  - $\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}]$ = probability of $\{X_1 = x_1, \;\; X_2 = x_2, \;\; \ldots \text{ and } \; X_N = x_N\}$.

Probability Distribution

Cumulative Distribution Function:

- Defined as

$$F_X(x) = \mathbb{P}[X \leq x].$$

- For high-dimensional vectors:

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{P}[\boldsymbol{X} \leq \boldsymbol{x}].$$

- This means:

$$F_{X_1,\ldots,X_N}(x_1,\ldots,x_N) = \mathbb{P}[X_1 \leq x_1 \ldots \textbf{ and } \ldots X_N \leq x_N].$$

Probability Density Function:

- Defined as

$$p_X(x) = \frac{d}{dx} F_X(x).$$

- For high-dimensional vectors:

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{d}{d\boldsymbol{x}} F_{\boldsymbol{X}}(\boldsymbol{x})$$

- This means:

$$p_{X_1,\ldots,X_N}(x_1,\ldots,x_N) = \frac{\partial^N}{\partial x_1 \ldots \partial x_N} F(x_1,\ldots,x_N).$$

Probability Distribution

Cumulative Distribution Function:

- Defined as

$$F_X(x) = \mathbb{P}[X \leq x].$$

- For high-dimensional vectors:

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{P}[\boldsymbol{X} \leq \boldsymbol{x}].$$

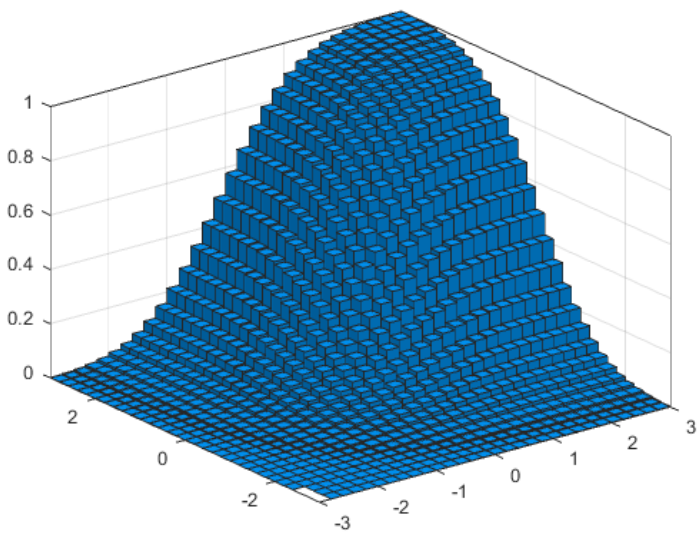
Probability Density Function:

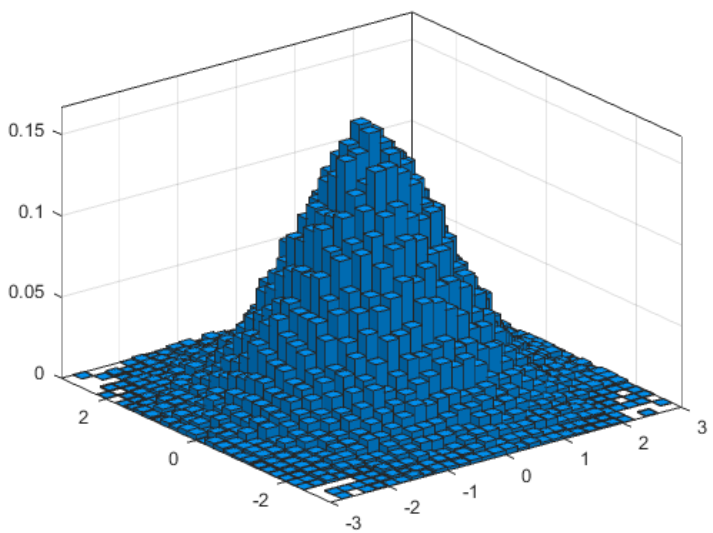
- Defined as

$$p_X(x) = \frac{d}{dx} F_X(x).$$

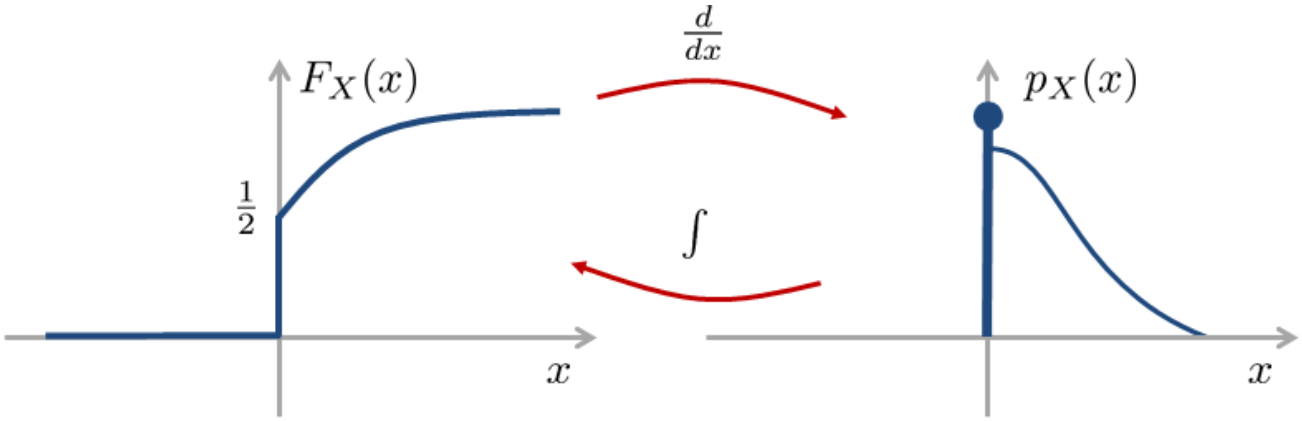
- For high-dimensional vectors:

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{d}{d\boldsymbol{x}} F_{\boldsymbol{X}}(\boldsymbol{x})$$





Linking CDF and PDF



Example. Consider a PDF

$$p_X(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{2}, & x = 0, \\ \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}, & x > 0. \end{cases}$$

Then, the CDF of X is

$$F_X(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{2} + \int_0^x \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{s^2}{2}\right\} ds, & x \geq 0. \end{cases}$$

Expectation

Definition

The **expectation** of a random variable X is

$$\mu = \mathbb{E}[X] = \int x\, p(x) dx.$$

- **Second Moment**:

$$\mathbb{E}[X^2] = \int x^2\, p(x) dx.$$

- **Variance**:

$$\mathrm{Var}[X] = \mathbb{E}[(X-\mu)^2] = \int (x-\mu)^2\, p(x) dx$$

Properties of Expectation

(i) **Function**. For any function g,

$$\mathbb{E}[g(X)] = \int g(x)p(x)dx.$$

(ii) **Linearity**. For any function g and h,

$$\mathbb{E}[g(X) + h(X)] = \mathbb{E}[g(X)] + \mathbb{E}[h(X)].$$

(iii) **Scale**. For any constant c,

$$\mathbb{E}[cX] = c\mathbb{E}[X].$$

(iv) **DC Shift**. For any constant c,

$$\mathbb{E}[X + c] = \mathbb{E}[X] + c.$$

Properties of Variance

(i) **Moment**.

$$\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

(ii) **Scale**. For any constant c,

$$\text{Var}[cX] = c^2\text{Var}[X].$$

(iii) **DC Shift**. For any constant c,

$$\text{Var}[X + c] = \text{Var}[X].$$

Moment Generating Function

Definition

The **moment generating function** (MGF) of a random variable X is

$$M_X(s) = \mathbb{E}[e^{sX}].$$

- MGF for Gaussian is

$$M_X(s) = \exp\left\{\mu s + \frac{s^2\sigma^2}{2}\right\}$$

- If X and Y are independent, then

$$\begin{aligned} M_{X+Y}(s) = \mathbb{E}[e^{s(X+Y)}] &= \mathbb{E}[e^{sX}e^{sY}] \\ &= \mathbb{E}[e^{sX}]\mathbb{E}[e^{sY}] = M_X(s)M_Y(s). \end{aligned}$$

Gaussian Random Variable

- $X \sim \mathcal{N}(\mu, \sigma^2)$ if

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

- Transforming a Gaussian. Let $Y = aX + b$, then

$$Y \sim \mathcal{N}(\ a\,\mu + b,\ \ a^2\,\sigma^2\ )$$

You may check:

$$\mathbb{E}[Y] = \mathbb{E}[aX + b] = a\mathbb{E}[X] + b = a\mu + b$$
$$\mathrm{Var}[Y] = \mathrm{Var}[aX + b] = \mathrm{Var}[aX] = a^2\mathrm{Var}[X].$$

High-dimensional Gaussian

An d-dimensional **Gaussian** has a PDF

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

where d denotes the dimensionality of the vector $\boldsymbol{x}$.

- The **mean vector** $\boldsymbol{\mu}$ is

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{bmatrix}$$

- The **covariance matrix** $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T] = \begin{bmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}(X_1, X_2) & \dots & \mathrm{Cov}(X_1, X_d) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}[X_2] & \dots & \mathrm{Cov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_d, X_1) & \mathrm{Cov}(X_N, X_2) & \dots & \mathrm{Var}[X_d] \end{bmatrix}$$

- $\boldsymbol{\Sigma}$ is always positive semi-definite. (Why?)

Special Case: Diagonal Covariance

- Suppose that X_i and X_j are independent for all $i \neq j$.
- This implies $\mathrm{Cov}(X_i, X_j) = 0$
- Simplify $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_d^2 \end{bmatrix},$$

- Then, the exponential is

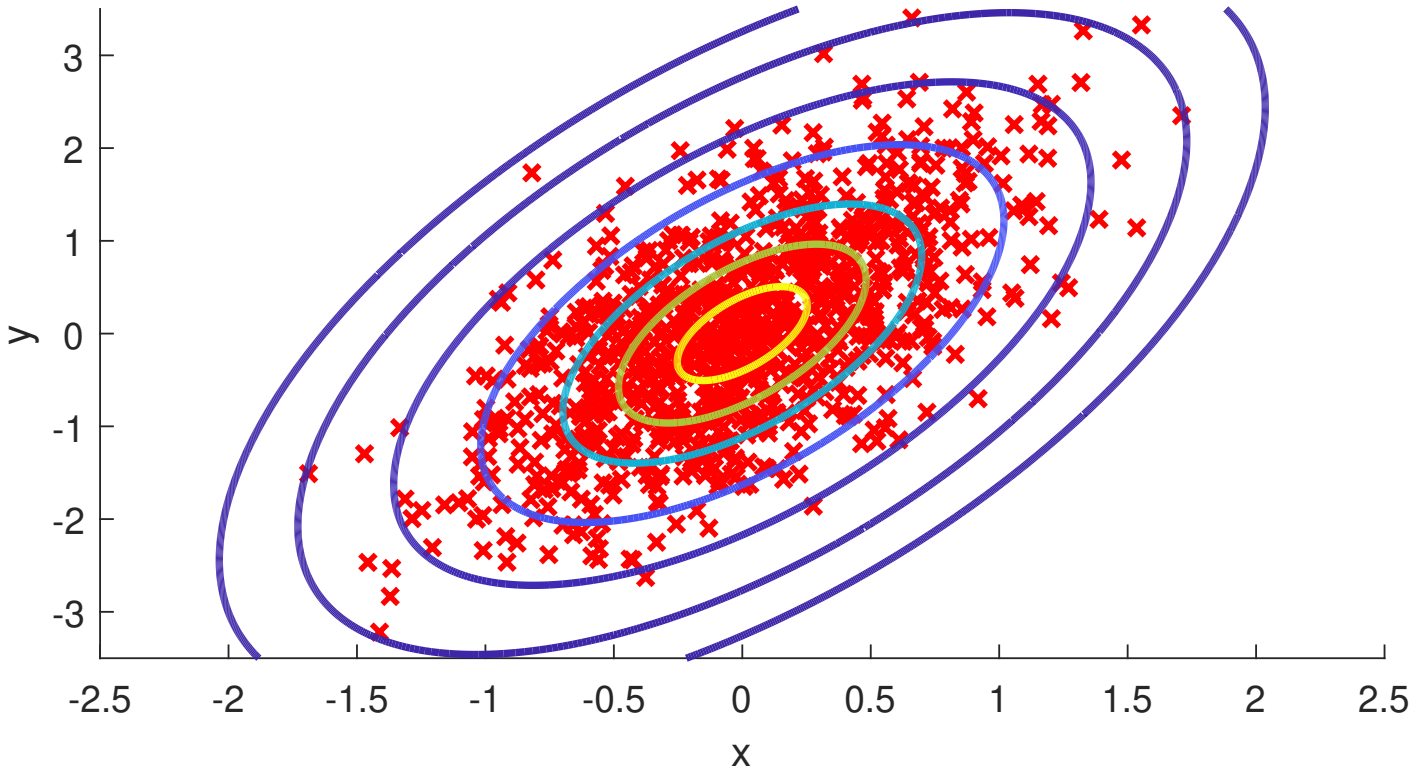
$$(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}.$$

- And hence, the PDF is

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{ -\frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right\}.$$

Visualization

- Generate 1000 random samples from a 2D Gaussian
- $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and $\boldsymbol{\Sigma} = \begin{bmatrix} 0.25 & 0.3 \\ 0.3 & 1 \end{bmatrix}$



MATLAB Code

```
% MATLAB code: Generate random numbers from multivariate Gaussian
mu    = [0 0];
Sigma = [.25 .3; .3 1];
x     = mvnrnd(mu,Sigma,1000);
```

```
% MATLAB code: Overlay random numbers with the Gaussian contour.
x1 = -2.5:.01:2.5;
x2 = -3.5:.01:3.5;
[X1,X2] = meshgrid(x1,x2);
F = mvnpdf([X1(:) X2(:)],mu,Sigma);
F = reshape(F,length(x2),length(x1));
figure(1);
scatter(x(:,1),x(:,2),'rx', 'LineWidth', 1.5); hold on;
contour(x1,x2,F,[.001 .01 .05:.1:.95 .99 .999], 'LineWidth', 2);
xlabel('x'); ylabel('y');
set(gcf, 'Position', [100, 100, 600, 300]);
```

Conditional Gaussian

- Data $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$.
- Class $Y \in \{1, 2, \ldots, K\}$.
- **Likelihood**:

$$p_{\boldsymbol{X}|Y}(\boldsymbol{x}|k) = \text{Probability of getting } \boldsymbol{X} \text{ given } Y$$

- **Prior**:

$$p_Y(k) = \text{Probability of getting } Y$$

- **Posterior**:

$$p_{Y|\boldsymbol{X}}(k|\boldsymbol{x}) = \text{Probability of getting } Y \text{ given } \boldsymbol{X}$$

- Related by

$$p_{Y|\boldsymbol{X}}(k|\boldsymbol{x}) = \frac{p_{\boldsymbol{X}|Y}(\boldsymbol{x}|k)p_Y(k)}{p_{\boldsymbol{X}}(\boldsymbol{x})} = \frac{p_{\boldsymbol{X}|Y}(\boldsymbol{x}|k)p_Y(k)}{\sum_k p_{\boldsymbol{X}|Y}(\boldsymbol{x}|k)p_Y(k)}$$

Example

- Two Gaussian $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$.
- **Prior** probability of getting a class is

$$p_Y(1) = \pi_1 \quad \text{and} \quad p_Y(2) = \pi_2.$$

- The **likelihood** term is

$$\begin{aligned} p_{\boldsymbol{X}|Y}(\boldsymbol{x}|k) &= \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ &= \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\} \end{aligned}$$

- The **posterior** is

$$\begin{aligned} p_{Y|\boldsymbol{X}}(k|\boldsymbol{x}) &= \frac{p_{\boldsymbol{X}|Y}(\boldsymbol{x}|k) p_Y(k)}{p_{\boldsymbol{X}}(\boldsymbol{x})} \\ &= \frac{\frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\} \cdot \pi_k}{\sum_{k=1}^{K} \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\} \cdot \pi_k} \end{aligned}$$

Negative Log-Likelihood

Negative Log-Likelihood for Gaussian:

$$
\begin{aligned}
& -\log p_{\boldsymbol{X}|Y}(\boldsymbol{x}|k) \\
&= -\log\left(\frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_k|}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}\right) \\
&= \underbrace{\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}_{\text{contains } \boldsymbol{x}} \underbrace{-\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_k|}_{\text{no } \boldsymbol{x}}.
\end{aligned}
$$

- $(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \geq 0$, always.
- $\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$ is called **Mahalanobis distance**.

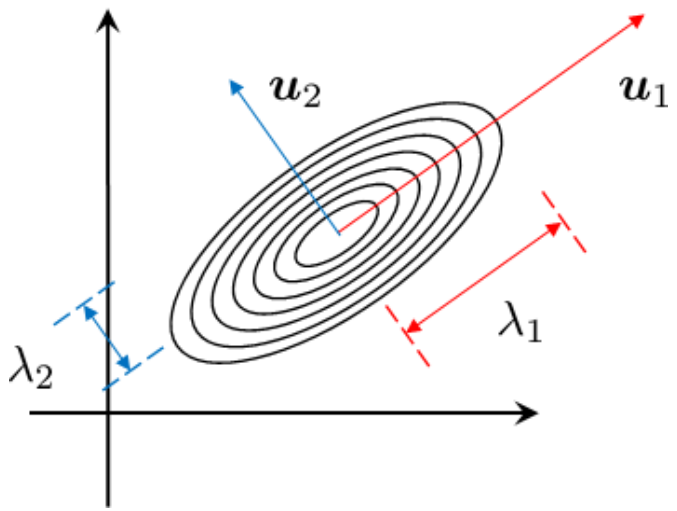
Geometry of Gaussian

- Perform eigen-decomposition

$$\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$$

$$= \begin{bmatrix} | & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} - & \boldsymbol{u}_1^T & - \\ - & \boldsymbol{u}_2^T & - \\ & \vdots & \\ - & \boldsymbol{u}_n^T & - \end{bmatrix}.$$

- $\boldsymbol{u}_i$ = orientation
- λ_i = radius

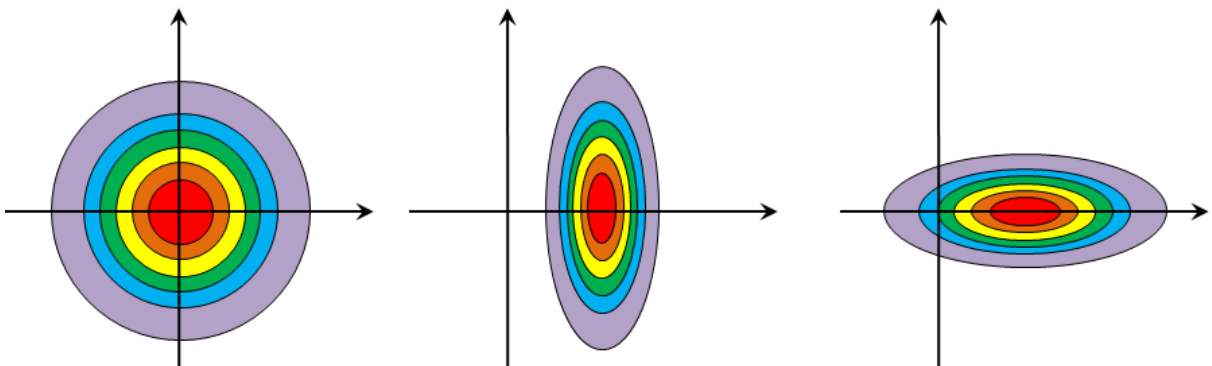


Geometry of Gaussian

- Special Case: X_i's are independent

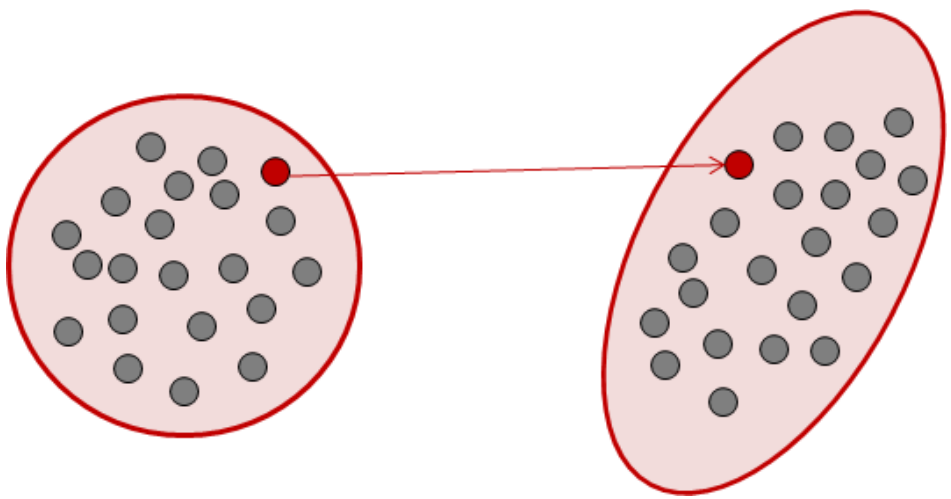
$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & 1 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \lambda_n \end{bmatrix} \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & 1 \end{bmatrix}.$$

- Ellipse; Standard bases; Different radii.



Transformation of Gaussian

- Your are given $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$. All $\boldsymbol{X}_j$ are generated from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.
- You want $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N\}$, where $\boldsymbol{Y}_j \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- But you only have $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$.



Transformation of Gaussian

How about this? Let

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$$

Can we find $\boldsymbol{A}$ and $\boldsymbol{b}$?

$$\mathbb{E}[\boldsymbol{Y}] = \mathbb{E}[\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}] = \boldsymbol{A}\underbrace{\mathbb{E}[\boldsymbol{X}]}_{=0} + \boldsymbol{b} = \boldsymbol{b}.$$

$$\begin{aligned}
\text{Cov}(\boldsymbol{Y}) &= \mathbb{E}[(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])^T] \\
&= \mathbb{E}[(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b} - \boldsymbol{b})(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b} - \boldsymbol{b})^T] \\
&= \mathbb{E}[(\boldsymbol{A}\boldsymbol{X})(\boldsymbol{A}\boldsymbol{X})^T] = \mathbb{E}[\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{A}^T] \\
&= \boldsymbol{A}\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{\Sigma}.
\end{aligned}$$

So here is the choice: $\boldsymbol{b} = \boldsymbol{\mu}$, and $\boldsymbol{A} = \boldsymbol{\Sigma}^{\frac{1}{2}}$.

Inverse Transform

If we have $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, how to transform $\boldsymbol{Y}$ to $\boldsymbol{X}$ so that $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$?

The inverse transform is

$$\boldsymbol{X} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{Y} - \boldsymbol{\mu})$$

Checking: If $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\mathbb{E}[\boldsymbol{X}] = \mathbb{E}[\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{Y} - \boldsymbol{\mu})] = \boldsymbol{\Sigma}^{-1}(\underbrace{\mathbb{E}[\boldsymbol{Y}]}_{=\boldsymbol{\mu}} - \boldsymbol{\mu}) = \boldsymbol{0}.$$

$$\begin{aligned}
\text{Cov}[\boldsymbol{X}] &= \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^T] \\
&= \mathbb{E}[(\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{Y} - \boldsymbol{\mu}))(\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{Y} - \boldsymbol{\mu}))^T] \\
&= \mathbb{E}[\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-\frac{1}{2}}] \\
&= \boldsymbol{\Sigma}^{-\frac{1}{2}}\underbrace{\mathbb{E}[(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^T]}_{=\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-\frac{1}{2}} = \boldsymbol{I}.
\end{aligned}$$