# Fundamentals of Media Processing

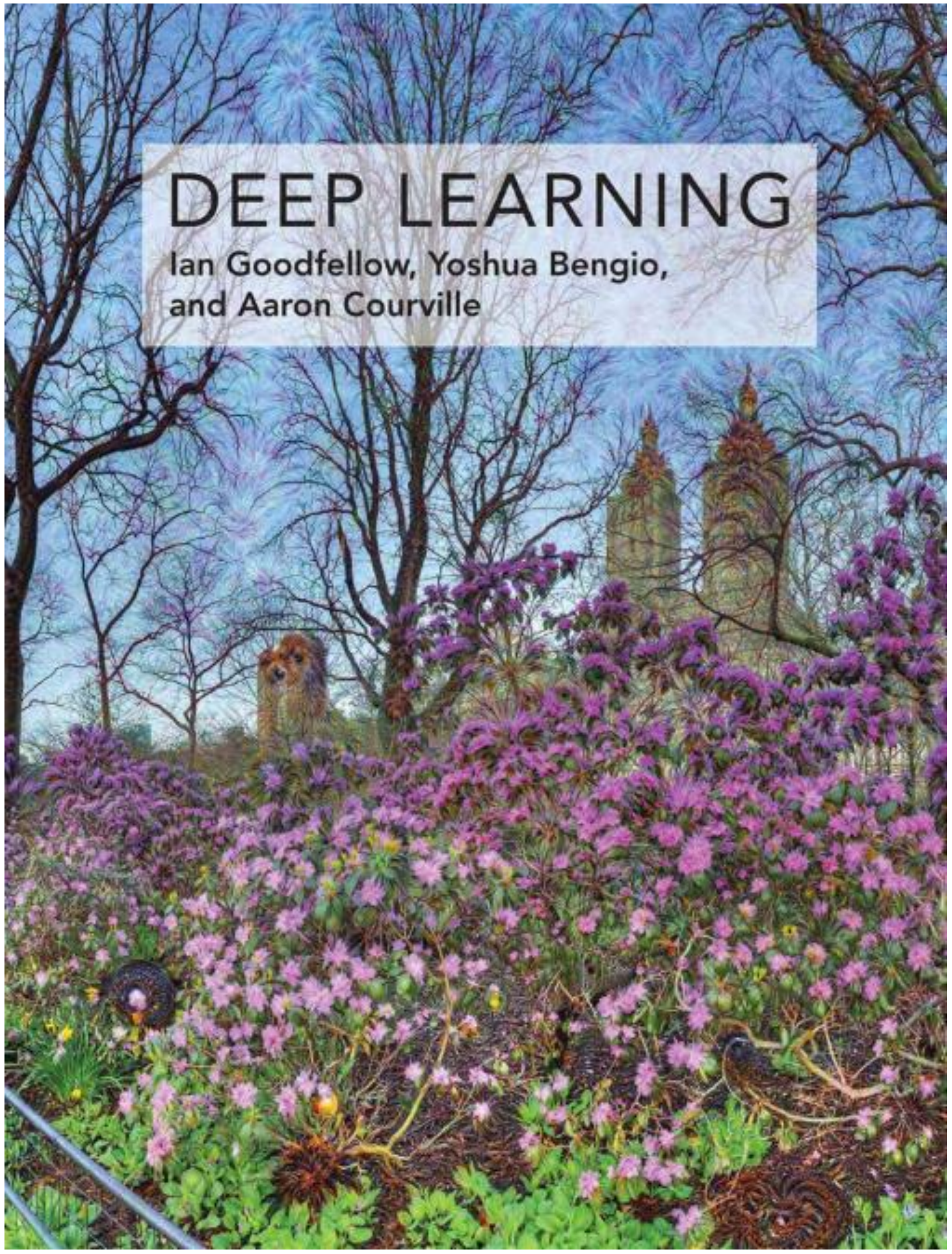
Lecturer:
池畑　諭（Prof. IKEHATA　Satoshi）
児玉　和也（Prof. KODAMA Kazuya）

Support:
佐藤　真一（Prof. SATO　Shinichi）
孟　洋（Prof. MO　Hiroshi）

# Course Overview (15 classes in total)

| | |
|---|---|
| **1-10** | **Machine Learning by Prof. Satoshi Ikehata** |
| 11-15 | Signal Processing by Prof. Kazuya Kodama |

Grading will be based on the final report.



Chapter 1-9 (out of 20)

**An introduction to a broad range of topics in deep learning, covering mathematical and conceptual background, deep learning techniques used in industry, and research perspectives.**

- Due to my background, I will mainly talk about “image”
- I will introduce some applications beyond this book

10/16 (Today) Introduction Chap. 1

## Basic of Machine Learning (Maybe for beginners)

10/23 Basic mathematics (1) (Linear algebra, probability, numerical computation) Chap. 2,3,4

10/30 Basic mathematics (2) (Linear algebra, probability, numerical computation) Chap. 2,3,4

11/6 Machine Learning Basics (1) Chap. 5

11/13 Machine Learning Basics (2) Chap. 5

## Basic of Deep Learning

11/20 Deep Feedforward Networks Chap. 6

11/27 Regularization and Deep Learning Chap. 7

12/4 Optimization for Training Deep Models Chap. 8

## CNN and its Application

12/11 Convolutional Neural Networks and Its Application (1) Chap. 9 and more

12/18 Convolutional Neural Networks and Its Application (2) Chap. 9 and more

## Linear Algebra

## Scalars, Vectors, Matrices and Tensors

Scalars

$$a = 1$$

(1-D) Vector

$$\boldsymbol{a} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} = [0,1]^T$$

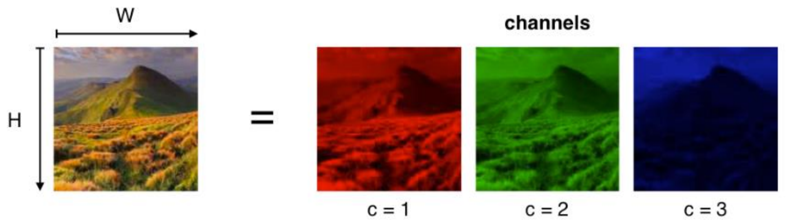
(2-D) Matrix

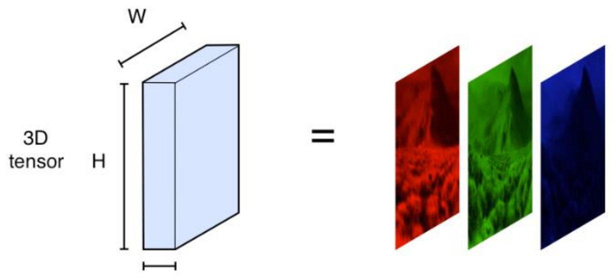
$$A = \begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix} \quad A^T = \begin{bmatrix} 0 & 2 \\ 1 & 3 \end{bmatrix} \quad A \in \mathbb{R}^{2\times 2}$$

$$A_{1,1} = 0, \quad A_{1,2} = 1, \quad A_{2,1} = 2, \quad A_{2,2} = 3$$

(N-D) Tensor

$$A_{i,j,k\ldots}$$





# Multiplying Matrices and Vectors

| | | | |
|---|---|---|---|
| $\boldsymbol{x}^T\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{x}$ | | $A(B+C) = AB + AC$ | |
| $\boldsymbol{x} * \boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{x} = \lvert\boldsymbol{x}\rvert^2$ | | $(AB)^T = B^T A^T$ | |
| $\lVert\boldsymbol{x}\rVert_2 = 1$ | Unit vector | $\boldsymbol{x}^T\boldsymbol{y} = 0$ | Orthogonal |
| $AA^{-1} = I$ | Inverse matrix | $A^T = A$ | Symmetric matrix |
| $A^T A = AA^T = I$ | Orthogonal matrix | $\mathrm{Tr}(A) = \sum_i A_{i,i}$ | Trace operator |

| | | | |
|---|---|---|---|
| L1 Norm | $\lVert\boldsymbol{x}\rVert_{\mathbf{1}} = \sum_{\boldsymbol{i}} \lvert x_i\rvert$ | L∞ Norm | $\lVert\boldsymbol{x}\rVert_\infty = \max_{\boldsymbol{i}} \lvert x_i\rvert$ |
| L2 Norm | $\lVert\boldsymbol{x}\rVert_{\mathbf{2}} = \left(\sum_{\boldsymbol{i}} \lvert x_i\rvert^{\mathbf{2}}\right)^{\frac{\mathbf{1}}{\mathbf{2}}}$ | Frobenius Norm | $\lVert A\rVert_F = \sqrt{\sum_{i,j} A_{i,j}^2}$ |
| L0 Norm | $\lVert\boldsymbol{x}\rVert_{\mathbf{0}} = \#$ of nonzero entries | | $\lVert A\rVert_F = \sqrt{\mathrm{Tr}(AA^T)}$ |

## Matrix Decomposition

- Decomposition of matrices shows us information about their functional properties that is not obvious from the representation of the matrix

- Eigendecomposition
- Singular value decomposition
- LU decomposition
- QR decomposition
- Jordan decomposition
- Cholesky decomposition
- Schur decomposition
- Rank factorization
- And more…

Differs in

- Applicable matrix
- Decomposition type
- Application

## Eigendecomposition

- ***Eigenvector*** and ***eigenvalue***:

$$A\boldsymbol{v} = \lambda\boldsymbol{v}$$

$\boldsymbol{v}$: Eigenvector
$\lambda$: Eigenvalue

- ***Eigendecomposition*** of a diagonalizable matrix :

$$A = V\mathrm{diag}(\lambda)V^{-1}$$

$V$: A set of eigenvectors
$\lambda$: A set of eigenvalues

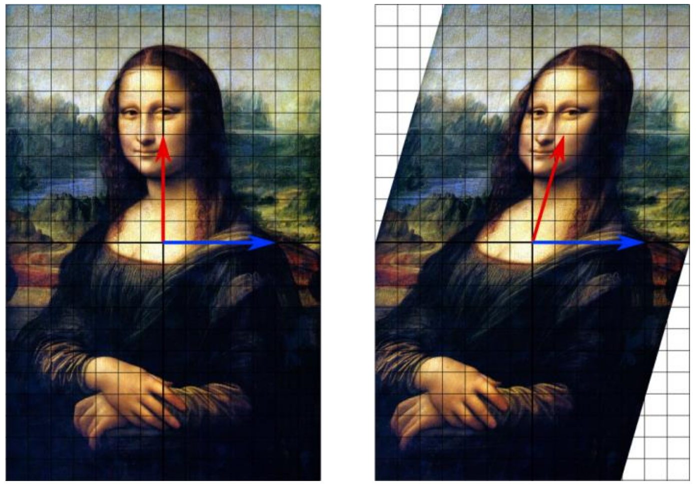
- Every symmetric matrix has a real, orthogonal eigendecomposition

$$A = Q\Lambda Q^T$$

- A matrix whose eigenvalues are all positive is ***positive definite***, positive or zero is ***positive semidefinite***

## What Eigenvalues and Eigenvectors are?

- Eigenvalues and eigenvectors feature prominently in the analysis of linear transformations. The prefix *eigen-* is adopted from the German word *eigen* for "proper", "characteristic". Originally utilized to study principal axes of the rotational motion of rigid bodies, eigenvalues and eigenvectors have a wide range of applications, for example in stability analysis, vibration analysis, atomic orbitals, facial recognition, and matrix diagonalization.


https://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors

# Calculation of Eigendecomposition

- Simply solve a linear equation

$$A\boldsymbol{v} = \lambda\boldsymbol{v} \rightarrow (A - \lambda I)\boldsymbol{v} = 0$$

- $A - \lambda I$ must not have inverse matrix, otherwise $v = 0$

$$\det(A - \lambda I) = 0 \rightarrow \quad p(\lambda) = (\lambda - \lambda_1)^{n_1}(\lambda - \lambda_2)^2\,(\lambda - \lambda_3)^{n_3} \ldots$$

- Once $\lambda_s$ are calculated, eigenvectors are calculated by:

$$(A - \lambda I)\boldsymbol{v} = 0$$

# Singular Value Decomposition (SVD)

- Similar to eigendecomposition, but matrix need not be square. ***Singular value decomposition*** of a real matrix A is

$$A = UDV^T \qquad A\boldsymbol{u} = \sigma\boldsymbol{u} \qquad A^T\boldsymbol{v} = \sigma\boldsymbol{v}$$

- The left singular vectors of $B$ are the eigenvectors of $AA^T$
- The right singular vectors of $B$ are the eigenvectors of $A^TA$
- SVD is useful for the matrix inversion of non-square matrix (***pseudo inverse matrix***) or rank approximation (E.g., find nearest matrix whose rank is k) or solving a ***homogenous system***

# Solving Linear Systems

- Number of Eq = Number of Unknown (Determined)

$a + 2b + 3c = 3$

$2a - b + c = 2$

$5a + b - 2c = 1$

$$\underset{A}{\begin{pmatrix} 1 & 2 & 3 \\ 2 & -1 & 1 \\ 5 & 1 & -2 \end{pmatrix}} \underset{\boldsymbol{x}}{\begin{pmatrix} a \\ b \\ c \end{pmatrix}} = \underset{\boldsymbol{y}}{\begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}}$$

$$\boldsymbol{x} = A^{-1}\boldsymbol{y}$$

- Number of Eq < Number of Unknown (Underdetermined)

$a + 2b + 3c = 3$

$2a - b + c = 2$

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & -1 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

?

- Number of Eq > Number of Unknown (Overdetermined)

$a + 2b + 3c = 3$

$2a - b + c = 2$

$5a + b - 2c = 1$

$8a - b + 6c = 0$

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & -1 & 1 \\ 5 & 1 & -2 \\ 8 & -1 & 6 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ 1 \\ 0 \end{pmatrix}$$

?

## Moore-Penrose Pseudoinverse

- SVD gives the ***pseudoinverse*** of A as

$$A^{+} = VD^{+}U^{T} \quad (A = UDV^{T}, D_{ii}^{+} = \frac{1}{D_{ii}})$$

- The solution to any linear systems is given by

$$\boldsymbol{x} = A^{+}\boldsymbol{y}$$

- If the linear system is:
  - Determined: this is same as the inverse
  - Underdetermined: this gives us the solution with the smallest error
  - Overdetermined: this gives us the solution with the smallest norm of x

# The Homogeneous System

- We may want to solve the homogenous system that is defined as

$$A\boldsymbol{x} = 0$$

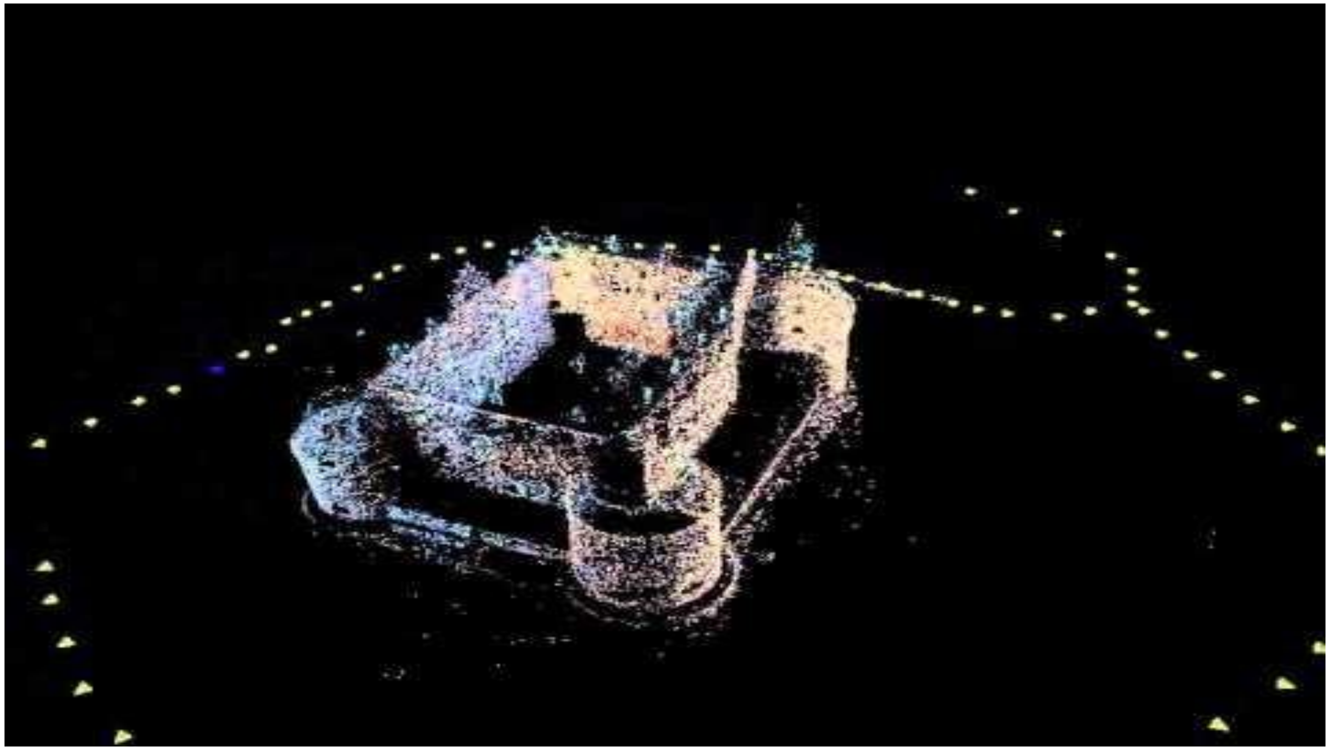
- If $\det(A^T A) = 0$, numerical solutions to a homogeneous system can be found with SVD or eigendecompositoin

$$Ax = 0 \qquad A^T A x = 0 \qquad A^T A v = \lambda v$$

  - Eigenvector corresponding to smallest eigen value (~=0)
  - Or, the column vector in V which is corresponding to the smallest singular value ($A = UDV^T$)

## Example: Structure-from-Motion

https://www.youtube.com/watch?v=i7ierVkXYa8



# Example: Structure-from-Motion

- Relative camera poses could be computed from the point correspondences on images
- Given camera pose (R,t), and camera intrinsics (K), 3D points X is given by triangulation

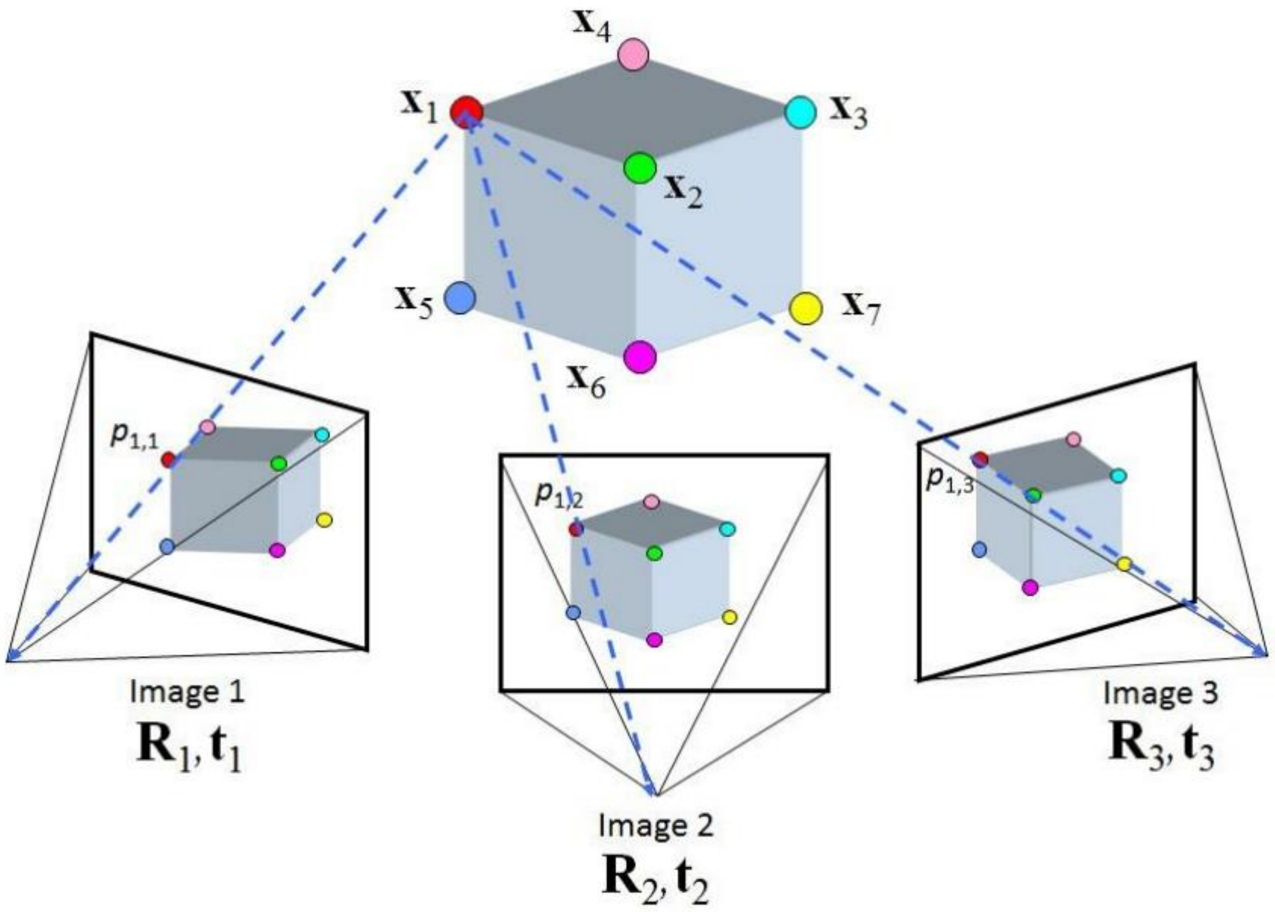
$$x_1 = K_1[R_1|T_1]\tilde{X}$$

$$x_2 = K_2[R_2|T_2]\tilde{X}$$

$$\downarrow$$

$$AX = 0$$

Homogenous system



## Probability and Information Theory

## Why probability?

Most real problem is not deterministic.

Which is this picture about Dog or Cat?

## Three possible sources of uncertainty

- Inherent stochasticity in the system

  e.g., Randomly shuffled card

- Incomplete observability

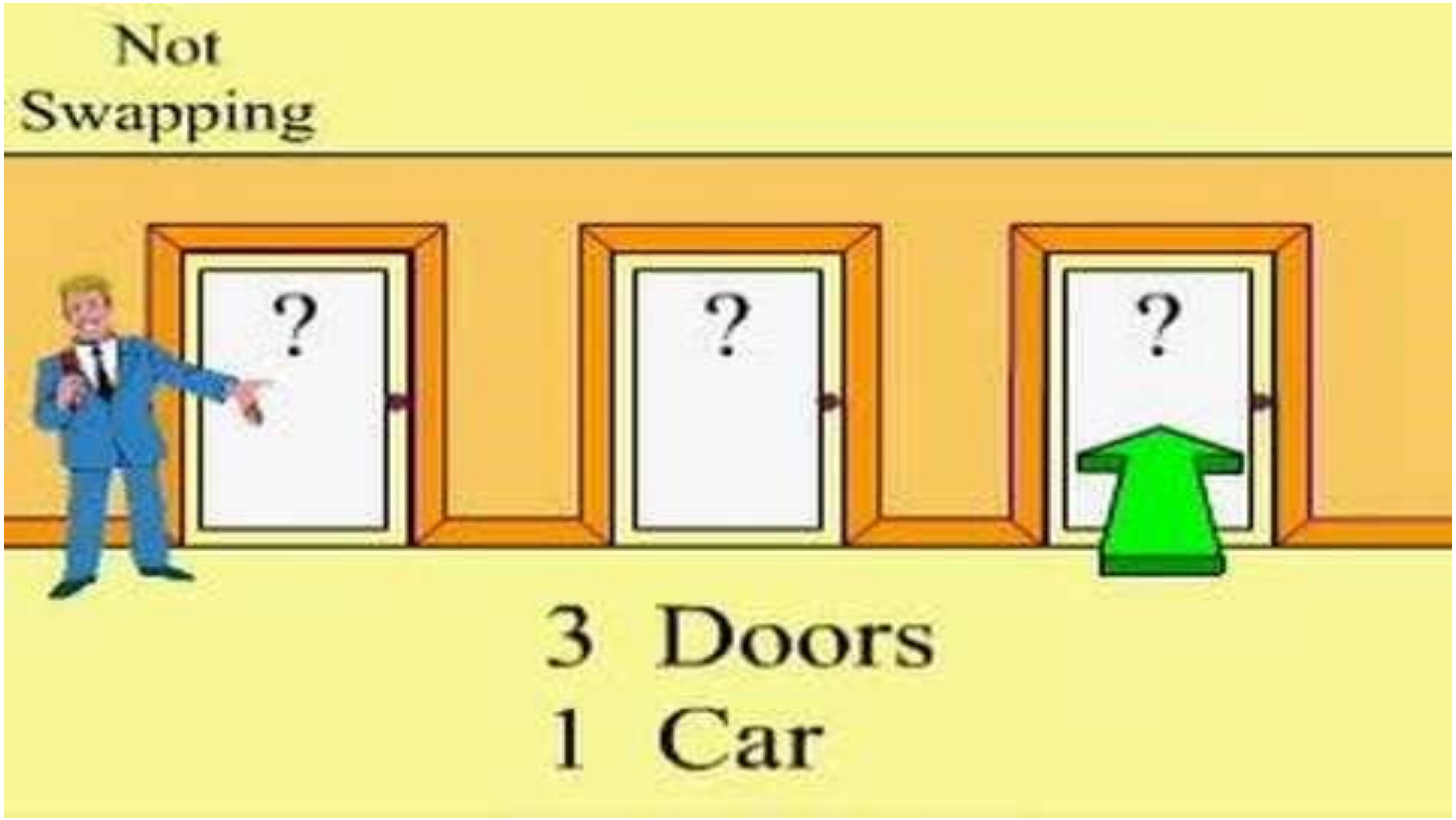
  e.g., Monty Hall problem

- Incomplete modeling

  e.g., A robot that only sees the discretized space

- A ***random variable*** is a variable that can take on different values randomly. e.g., $x \in \mathrm{x}$

# Monty Hall problem



## Discrete case: Probability Mass Function

$$P(x)$$

- The domain of *P* must be the set of all possible states of *x*
  - $\forall x \in \mathrm{x}, 0 \leq P(x) \leq 1.$ An impossible event has probability 0 and no state can be less probable than that. Likewise, an event that is guaranteed to happen has probability 1, and no state can have a greater chance of occurring
  - $\sum_{x \in \mathrm{x}} P(x) = 1.$ We refer to this property as being ***normalized***. Without this property, we could obtain probabilities greater than one by computing the probability of one of many events occurring

- Example
  - Dice : P(*x*)=1/6, *x* is an event where f(*x*)=1,2,3,4,5,6

## Continuous case: Probability Density Function

$$p(x)$$

- The domain of $p$ must be the set of all possible states of x
  - $\forall x \in \mathrm{x}, p(x) \geq 0$. Note that we do not require $p(x) \leq 1$
  - $\int p(x)dx = 1$
  - <u>Does not give the probability of a specific state directly</u>

e.g., p(0.0001) + p(0.0002) +….. >= 100%!

- Example
  - What is the probability that randomly selected value within [0,1] is more than 0.5?

# Marginal Probability

- The probability distribution over the subset
  - $\forall x \in \mathrm{x}, \mathrm{P}(x) = \sum_y P(x, y)$ (Discrete)
  - $p(x) = \int p(x, y)dy$ (Continuous)

Table: The statistics about the student

| | Male | Female |
|---|---|---|
| Tokyo | 0.4 | 0.3 |
| Outside Tokyo | 0.1 | 0.2 |

Q. How often students come from Tokyo?

# Conditional Probability

- The probability of an event, given that some other event has happened
  - $P(y|x) = P(y,x)/P(x)$
  - $P(y,x) = P(y|x)P(x)$
- Example: The boy and girl problem

Mr. Jones has two children. One is a girl. What is the probability that the other is a boy?

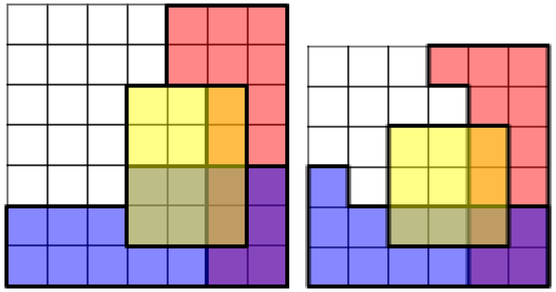
- Each child is either male or female.
- Each child has the same chance of being male as of being female.
- The sex of each child is independent of the sex of the other.

## Chain rule of Probability

- $P\left(x^{(1)}, \dots, x^{(n)}\right) = P\left(x^{(1)}\right)\Pi_{i=2}^{n} P(x^{(i)} | x^{(1)}, \dots, x^{(i-1)})$
- $P(a, b, c) = P(a|b, c)P(b, c)$
- $P(b, c) = P(b|c)P(c)$
- $P(a, b, c) = P(a|b, c)P(b|c)P(c)$

# Independence and Conditional Independence

- The random variables x and y are ***independent*** if their probability distribution can be expressed as a product of two independent factors
  - $\forall x \in \mathrm{x}, y \in \mathrm{y},\ p(x, y) = p(x)p(y)$
- Two random variables x and y are ***conditionally independent*** given random variable z if the conditional probability distribution over x and y factorizes in this way for every value of z
  - $\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z}, p(x, y|z) = p(x|z)p(y|z)$



$P(R_{ed} \cap B_{lue}|Y_{ellow}) = P(R|Y)P(B|Y)$

# Expectation, Variance and Covariance

- The ***expectation*** of some function f(*x*) with respect to P(*x*) or p(*x*) is mean value that f takes on when *x* is drawn from P or p
  - $\mathrm{E}_{x\rightarrow P}[f(x)] = \sum_x P(x)f(x)$
  - $\mathrm{E}_{x\rightarrow p}[f(x)] = \int p(x)f(x)dx$

- The ***variance*** gives how much the values of a function of a random variable *x* vary as we sample different values of *x* from its probability distribution
  - $\mathrm{Var}[f(x)] = \mathrm{E}[(f(x) - \mathrm{E}[f(x)])^2]$

- The ***covariance*** gives some sense of how much two values are linearly related to each other, as well as the scale of these variables:
  - $\mathrm{Cov}[f(x), g(y)] = \mathrm{E}[(f(x) - \mathrm{E}[f(x)])(g(y) - \mathrm{E}[g(y)])]$
  - $\mathrm{Cov}(\boldsymbol{x})_{i,j} = \mathrm{Cov}(x_i, x_j)$ Covariance matrix for $n \times n$ matrix

# Common Probability Distribution (1)

- Bernoulli distribution

$$P(1) = \Phi \qquad P(0) = 1 - \Phi \qquad P(x) = \phi^x (1 - \Phi)^{1-x}$$

- Gaussian (Normal) distribution

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

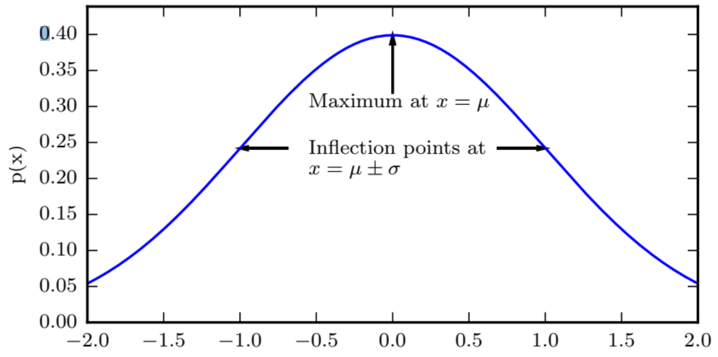


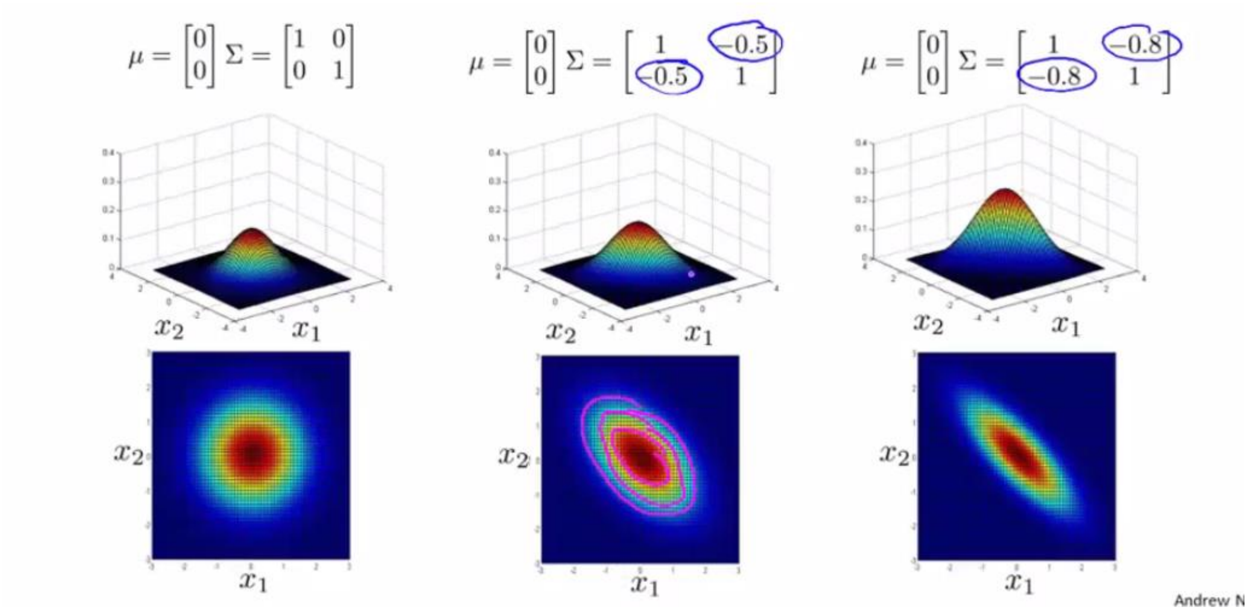
Figure 3.1

**The central limit theorem:**
The sum of many independent random variables is approximately normally distributed

# Common Probability Distribution (2)

- Multivariate normal distribution

$$N(\boldsymbol{x};\boldsymbol{\mu},\Sigma)=\sqrt{\frac{1}{(2\pi)^n\det(\Sigma)}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$



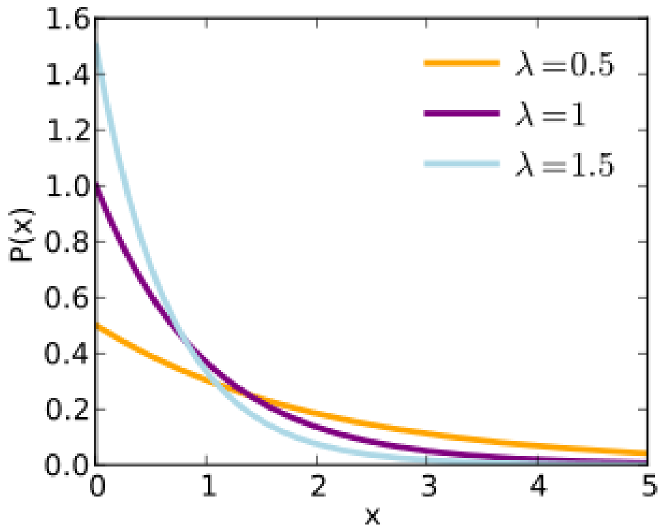
https://notesonml.wordpress.com/2015/06/30/chapter-14-anomaly-detection-part-2-multivariate-gaussian-distribution/

# Common Probability Distribution (3)

- Exponential distribution

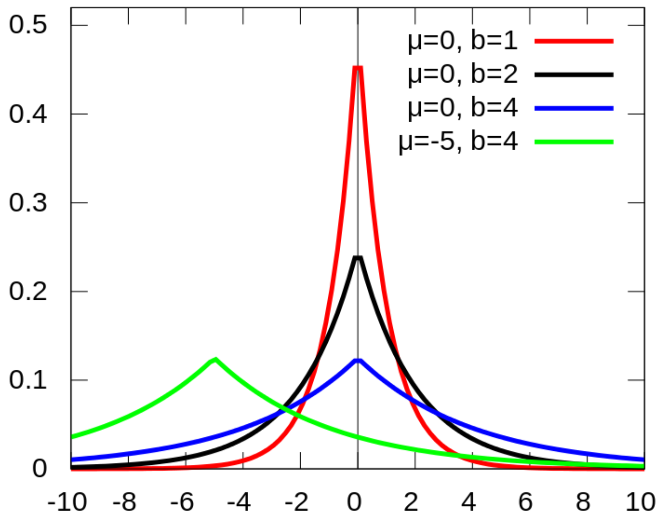
$$p(x;\lambda) = \lambda \mathbf{1}_{x\geq 0}\exp(-\lambda x)$$

$\mathbf{1}_{x\geq 0}$ assign zero to negative values of x



- Laplace distribution

$$\mathrm{Laplace}(x;\mu,\gamma) = \frac{1}{2\lambda}\exp(-\frac{|x-\mu|}{\gamma})$$



# Mixtures of Distributions

- Empirical Distribution

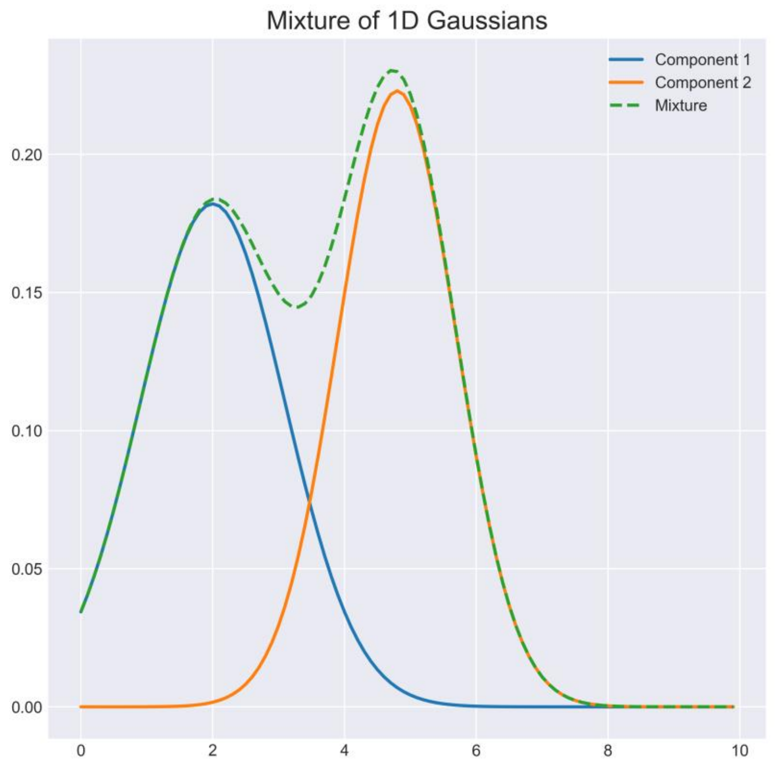
$$p(x) = \frac{1}{m}\sum \delta\left(x - x^{(i)}\right)$$

  - $\delta$ is a ***dirac delta function***

- Gaussian Mixture Model

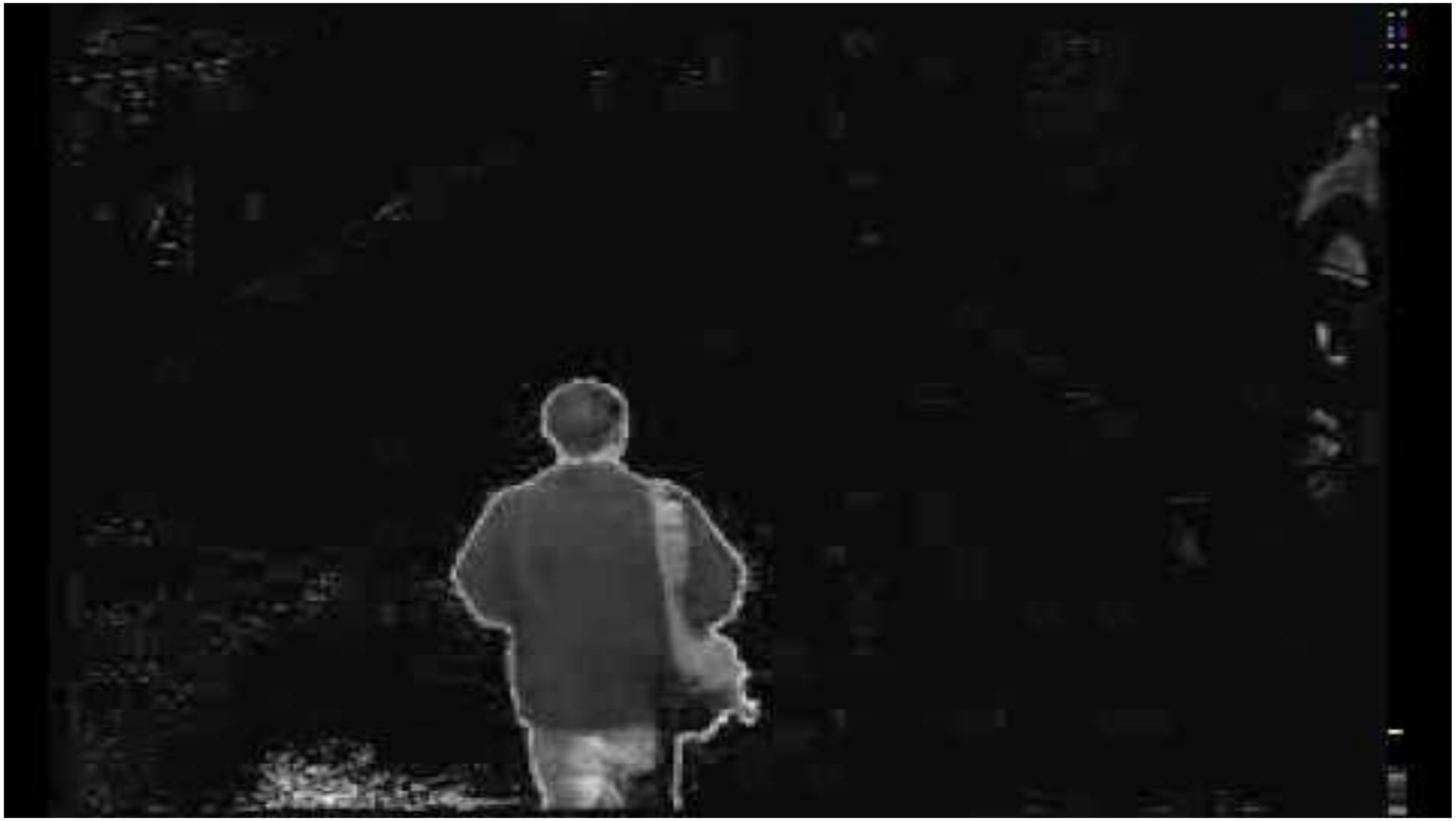
$$p(x) = \sum_i \phi_i N\left(x|\mu_i, \sigma_i^2\right)$$

$\phi_i$: latent variable (weight of gaussian)



  - GMM is a universal approximator of densities of a distribution

## Application of GMM in Computer Vision



Background subtraction by GMM

https://www.youtube.com/watch?v=KGal_NvwI7A

# Useful Properties of Common Functions

- Logistic Sigmoid

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

  - Good to produce [0,1] random values

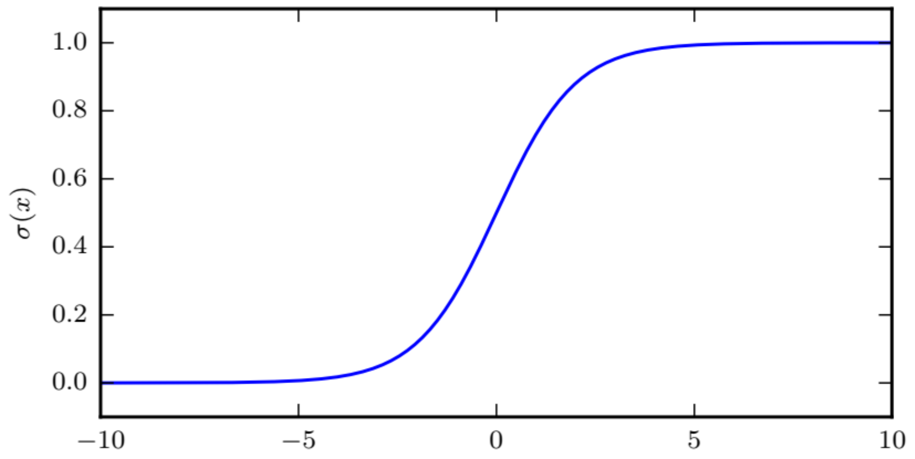


Figure 3.3: The logistic sigmoid function.

- Softplus function

$$\varsigma(x) = \log(1 + \exp(-x))$$

  - Good to prodce [0,∞] random values

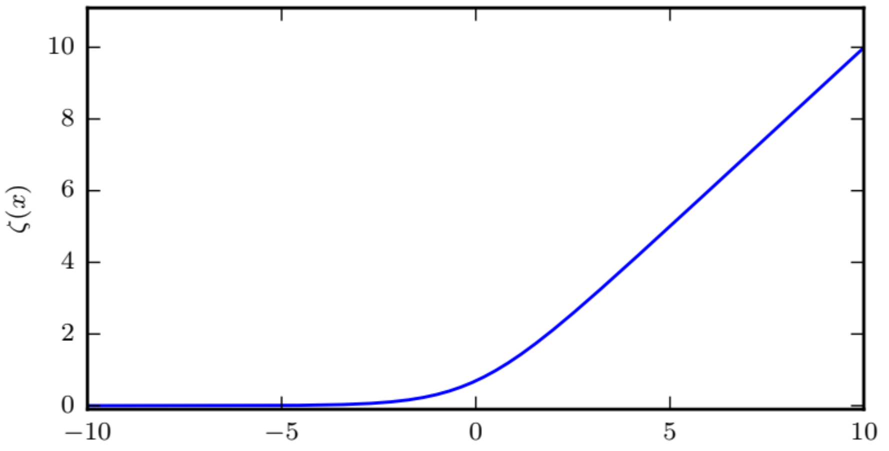


Figure 3.4: The softplus function.

## Bayes' Rule

***Prior probability*** ***likelihood***

$$P(x|y) = \frac{P(x)P(y|x)}{P(y)} = \frac{P(x,y)}{P(y)}$$

***Posterior probability*** ***Prior probability***

### Factory Problem

The entire output of a factory is produced on three machines. The three machines account for 20%, 30%, and 50% of the factory output. The fraction of defective items produced is 5% for the first machine; 3% for the second machine; and 1% for the third machine. If an item is chosen at random from the total output and is found to be defective, what is the probability that it was produced by the third machine?

# Factory Problem

The entire output of a factory is produced on three machines. The three machines account for 20%, 30%, and 50% of the factory output. The fraction of defective items produced is 5% for the first machine; 3% for the second machine; and 1% for the third machine. If an item is chosen at random from the total output and is found to be defective, what is the probability that it was produced by the third machine?

$$P(X_A) = 0.2, P(X_B) = 0.3, P(X_C) = 0.5$$

$$P(Y|X_A) = 0.05, P(Y|X_B) = 0.03, P(Y|X_C) = 0.01$$

$$P(\mathrm{Y}) = P(Y|X_A)P(X_A) + P(Y|X_B)P(X_B) + P(Y|X_C)P(X_C)$$

$$P(X_C|\mathrm{Y}) = \frac{P(X_C)P(Y|X_C)}{P(Y)} = 5/24$$

# Information Theory

- ***Self-information*** (for single outcome)
  - Likely event has low information, less likely event has higher information

$$I(x) = -\log P(x)$$

In units of ***nats*** or ***bits***: amount of information gained by observing an event of probability 1/e or 1/2

- ***Shanon entropy*** (amount of uncertainty in an entire probability distribution)

$$H(x) = E_{x\sim P}[I(x)] = -E_{x\sim P}[\log P(x)]$$

  - Known as differential entropy for p(x)

# Kullback-Leibler (KL) divergence

$$D_{KL}(P||Q) = E_{x\sim P}\left[\log\frac{P(x)}{Q(x)}\right] = E_{x\sim P}[\log P(x) - \log Q(x)]$$

- The difference of two distributions (higher is different)
  - KL divergence is positive or zero only when P and Q are the same distribution
  - Often used for model fitting (e.g., fitting GMM (Q(x)) on P(x)
  - Asymmetric measure ($D_{KL}(P||Q) \neq D_{KL}(Q||P)$)

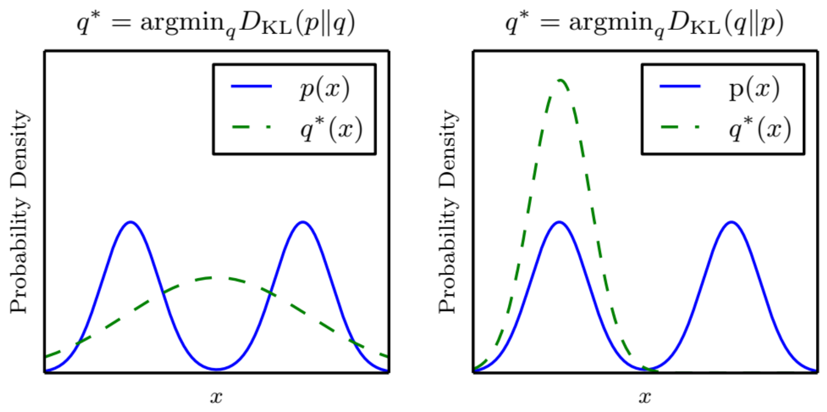


Figure 3.6

## Cross-entropy

- $H(P,Q) = H(P) + D_{KL}(P||Q) = -E_{x\sim P} \log Q(x)$

- The average number of bits needed to identify an event drawn from the set, if a coding scheme is used that is optimized for an "artificial" probability distribution Q, not true distribution P

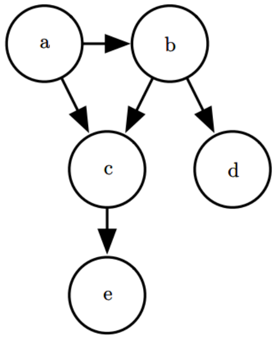
- Minimizing the cross-entropy with respect to Q is equivalent to minimizing the KL divergence

- In classification problems, the commonly used cross entropy loss, measures the cross entropy between the empirical distribution of the labels (given the inputs) and the distribution predicted by the classifier

# Structured Probabilistic Models

- In machine learning, it is inefficient to represent an entire probabilistic distribution in a single function. To reduce the parameter, the function is often factorized
- Suppose $a$ influence $b$, and $b$ influence $c$, but $a$ and $c$ are independent; $p(a,b,c) = p(a)p(b|a)p(c|b)$
- We can describe these factorizations using graphs (graphical model)

Figure 3.7



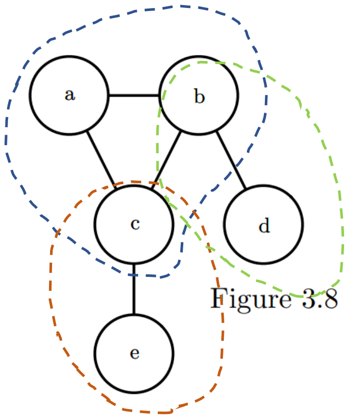


Figure 3.8

$\mathcal{G}$: graph

$\mathcal{C}^i$: clique

$\phi^{(i)}(\mathcal{C}^i)$: factor

- **Directed** model

$p(a,b,c,d,e) = p(a)p(b|a)p(c|a,b)p(d|b)p(e|c)$

- **Undirected** model (factor graph)

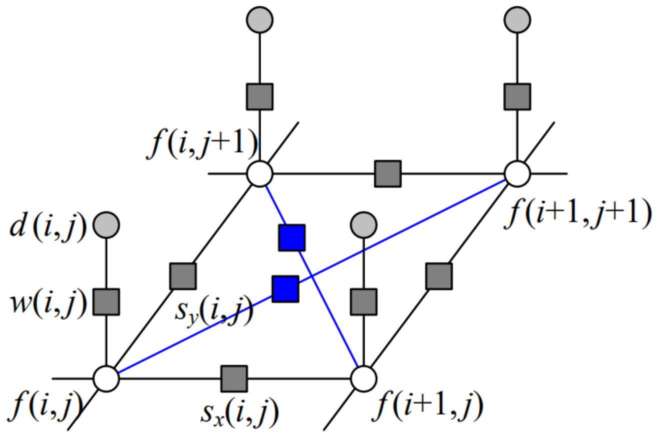
$p(a,b,c,d,e) = \frac{1}{Z}\phi^{(1)}(\mathrm{a,b,c})\phi^{(2)}(b,d)\phi^{(3)}(c,e)$

# Markov Random Field (MRF)

- In computer vision algorithm, the most common graphical model may be ***Markov Random Filed*** (MRF), whose log-likelihood can be described using local neighborhood interaction (or penalty) terms.
- MRF models can be defined over discrete variables, such as image labels (e.g., image restoration)

$$E(\boldsymbol{x},\boldsymbol{y}) = \overset{\text{Likelihood term}}{E_d(\boldsymbol{x},\boldsymbol{y})} + \overset{\text{penalty term}}{E_p(\boldsymbol{x})}$$

$$E_p(\boldsymbol{x}) = \sum_{\{(i,j),(k,l)\}\in\mathcal{N}} V_{i,j,k,l}(f(i,j),j(k,l))$$

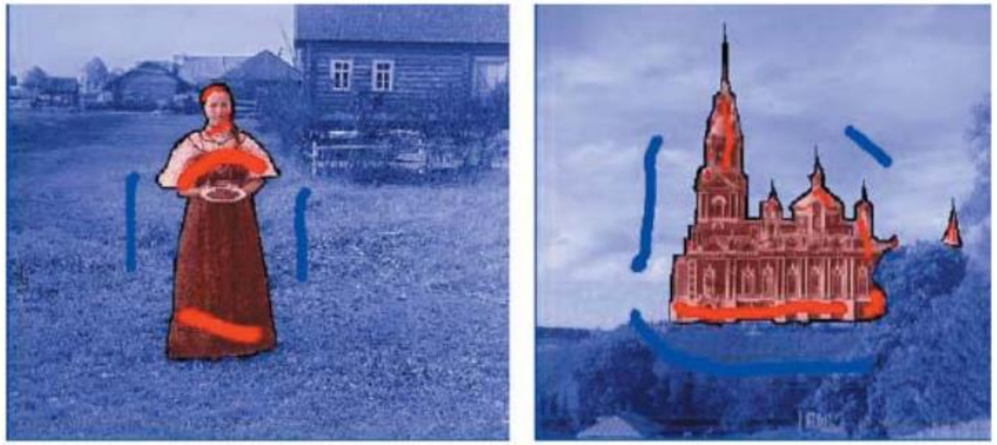


$\mathcal{N}_4$ and $\mathcal{N}_8$ neighborhood system

# Conditional Random Field (CRF)

- In classical Bayes model, prior p(x) is independent of the observation y. $p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})$
- However, it is often helpful to update the prior probability based on the observation; the pairwise term depends on the y as well as x

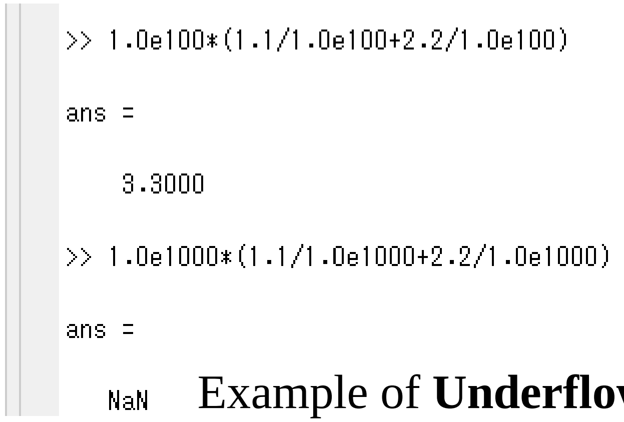
$$E(\boldsymbol{x}|\boldsymbol{y}) = E_d(\boldsymbol{x},\boldsymbol{y}) + E_s(\boldsymbol{x},\boldsymbol{y}) = \sum_p V_p(\boldsymbol{x}_p,\boldsymbol{y}) + \sum_{p,q} V_{p,q}(\boldsymbol{x}_p,\boldsymbol{x}_q,\boldsymbol{y})$$



# Numerical Computation

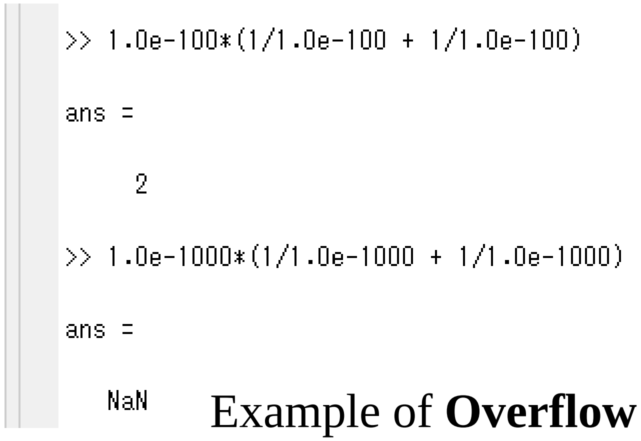
# Numerical concerns for implementations of deep learning algorithms

- Algorithms are often specified in terms of real numbers; real numbers cannot be implemented in a finite computer
  - Does the algorithm work when implemented with a finite number of bits?
- Do small changes in the input to a function cause large changes to an output?
  - Rounding errors, noise, measurement errors can cause large changes
  - Iterative search for best input is difficult



```
>> 1.0e100*(1.1/1.0e100+2.2/1.0e100)

ans =

    3.3000

>> 1.0e1000*(1.1/1.0e1000+2.2/1.0e1000)

ans =

   NaN
```

Example of **Underflow**



```
>> 1.0e-100*(1/1.0e-100 + 1/1.0e-100)

ans =

     2

>> 1.0e-1000*(1/1.0e-1000 + 1/1.0e-1000)

ans =

   NaN
```

Example of **Overflow**

# Poor Conditioning

- Conditioning refers to how rapidly a function changes with respect to small changes in its inputs

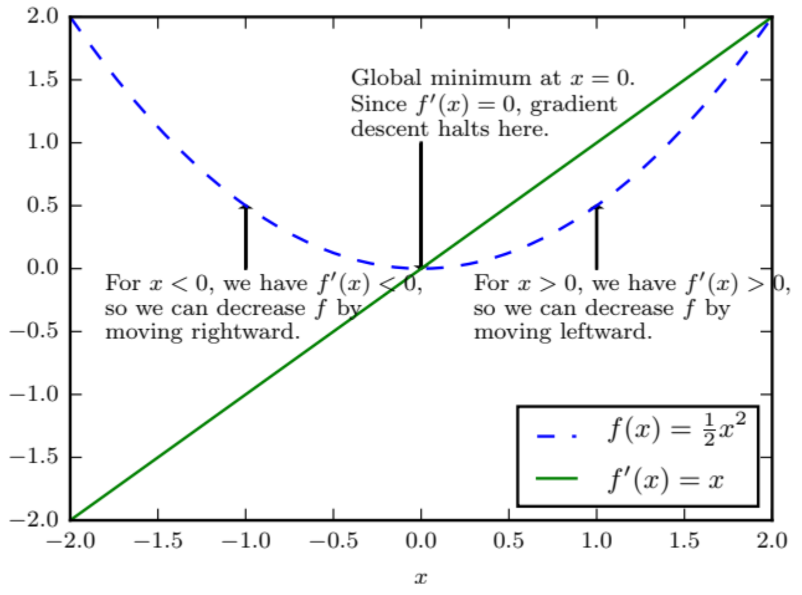
- We can evaluate the conditioning by a ***conditioning number***
  - The sensitivity is an intrinsic property of a function, not of computational error
  - For example, condition number for $f(\boldsymbol{x}) = A^{-1}\boldsymbol{x}$, where A is a positive semidefinite matrix, is

$$\max_{i,j} \left| \frac{\lambda_i}{\lambda_j} \right| ;$$ where $\lambda_s$ are eigenvalue of A

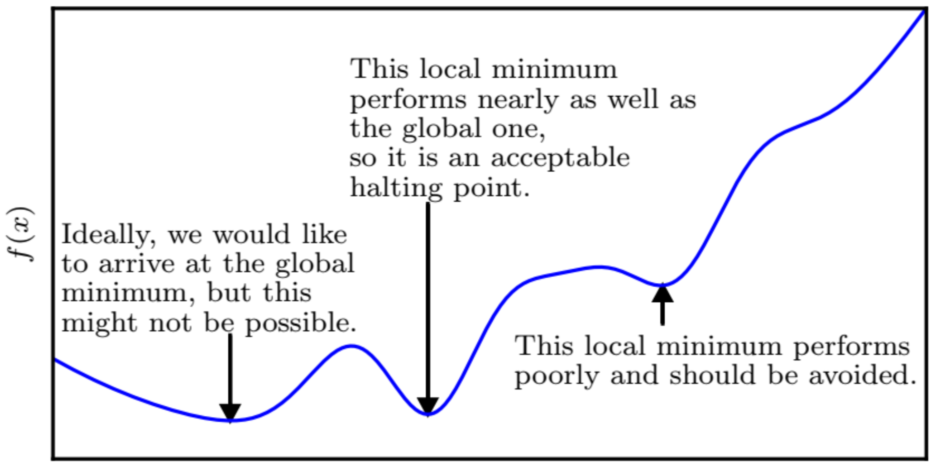
# Gradient-Based Optimization

- ***Objective function***: the function we want to minimize
- May also call it criterion, cost function, loss function, error function
- $\boldsymbol{x}^* = \operatorname{argmin} f(\boldsymbol{x})$
- The derivative of $f(x)$ is denoted as $f'(x)$ or $df/dx$

- The ***gradient descent*** is the technique to reduce $f(x)$ by moving $x$ in small steps with the opposite sign of the derivative
- Stationary points: local minima or maxima $f'(x) = 0$



Gradient descent

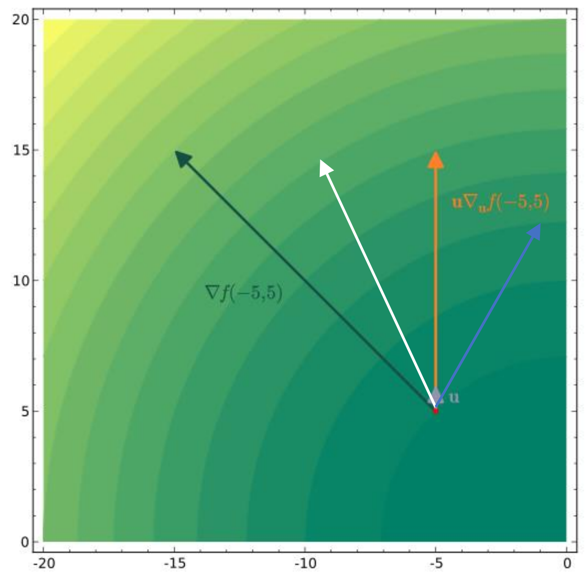


Local minima and global minimum

# Partial/Directional Derivatives for multiple inputs

$$z = f(\boldsymbol{x}) \quad \frac{\partial f}{\partial x_i} \quad \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \left[\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_m}\right]$$ Partial Derivatives

- The **directional derivatives** in direction $u$ is the slope of the function $f$ in direction $u$



https://en.wikipedia.org/wiki/Directional_derivative

To find the "steepest" direction,

$$\min_{\boldsymbol{u}} u^T \nabla_x f(x) = \min_{\boldsymbol{u}} \|\boldsymbol{u}\|_2 \|\nabla_x f(x)\|_2 \cos\theta$$
$$\cong \min_{\theta} \cos\theta$$

$$\boldsymbol{u} \longleftrightarrow f(\boldsymbol{x})$$

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \epsilon \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^t)$$

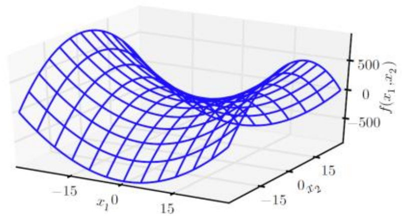
- Gradient descent for multiple inputs
- $\epsilon$ (learning rate) is fixed or adaptively selected (line search)

# Beyond the Gradient: Jacobian and Hessian Matrices

$$\boldsymbol{f}: \mathbb{R}^m \to \mathbb{R}^n \qquad \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}, \cdots, \frac{\partial f_1}{\partial x_m} \\ \vdots \\ \frac{\partial f_n}{\partial x_1}, \cdots, \frac{\partial f_n}{\partial x_m} \end{bmatrix}$$ Jacobian matrix

$$H(f)(\boldsymbol{x})_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f(\boldsymbol{x})$$ Hessian matrix

- When the function is continuous, $H(f)(\boldsymbol{x})_{ij} = H(f)(\boldsymbol{x})_{ji}$
- A real symmetric Hessian matrix has Eigendecomposition
  - When the Hessian is positive semidefinite, the point is local minimum
  - When the Hessian is negative semidefinite, the point is local maximum
  - Otherwise, the point is a **saddle** point



# Beyond the Gradient: Jacobian and Hessian Matrices

- The second derivative in a specific direction represented by a unit vector $\boldsymbol{d}$ is $\boldsymbol{d}^T H \boldsymbol{d}$

$$f(\boldsymbol{x}) \approx f\left(\boldsymbol{x}^{(0)}\right) + \left(\boldsymbol{x} - \boldsymbol{x}^{(0)}\right)^T \boldsymbol{g} + \frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{x}^{(0)}\right)^T H\left(\boldsymbol{x} - \boldsymbol{x}^{(0)}\right)$$

  - $\boldsymbol{x}^{(0)}$ is the current point, $\boldsymbol{g}$ is the gradient and $H$ is the Hessian at $\boldsymbol{x}^{(0)}$

- Then new point $\boldsymbol{x}$ will be given by $\boldsymbol{x}^{(0)} - \epsilon \boldsymbol{g}$

$$f\left(\boldsymbol{x}^{(\mathbf{0})} - \epsilon \boldsymbol{g}\right) \approx f\left(\boldsymbol{x}^{(0)}\right) - \epsilon \boldsymbol{g}^T \boldsymbol{g} + \frac{1}{2}\epsilon^2 \boldsymbol{g}^T H \boldsymbol{g}$$

- When $\boldsymbol{g}^T H \boldsymbol{g}$ is positive, solving for the optimal learning rate that decreases the function is

$$\epsilon^* = \frac{g^T g}{g^T H g}$$

# Newton's Method (Second-Order Algorithm)

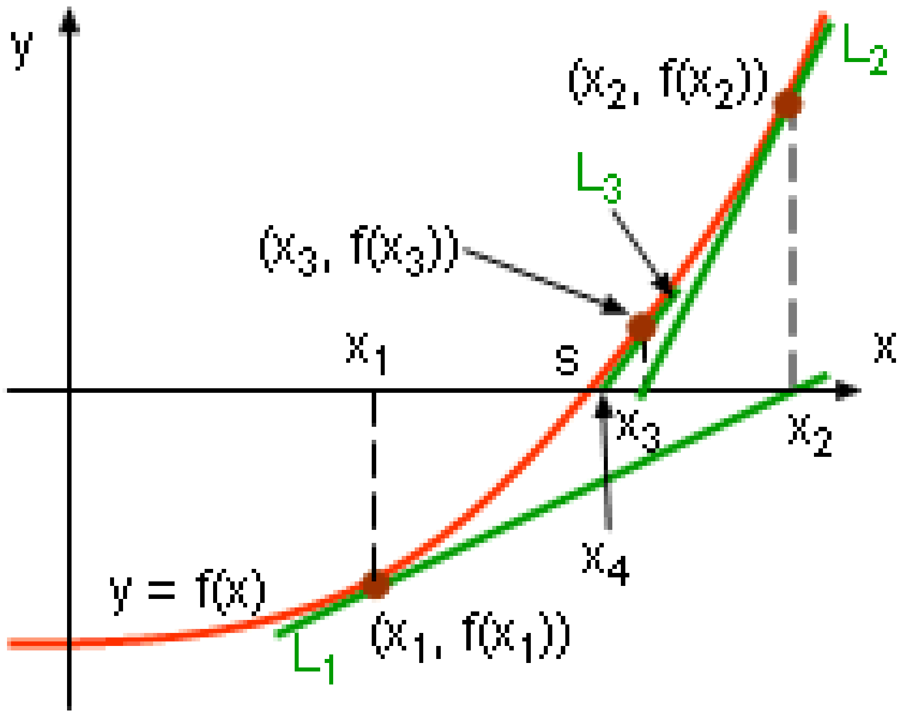
- In Gradient descent, the step size must be small enough
- ***Newton's method*** is based on using second-order Tyler Expansion

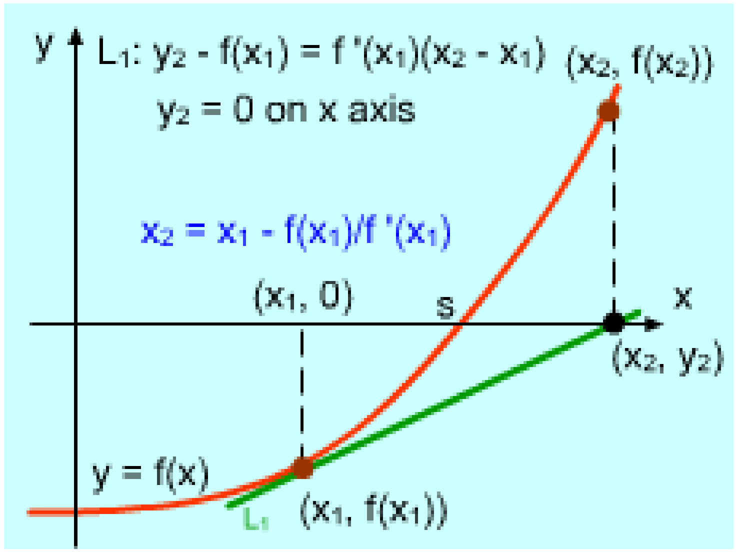
$$f(\boldsymbol{x}) \approx f\left(x^{(0)}\right) + \left(x - x^{(0)}\right)^T g + \frac{1}{2}\left(x - x^{(0)}\right)^T H\left(x - x^{(0)}\right)$$

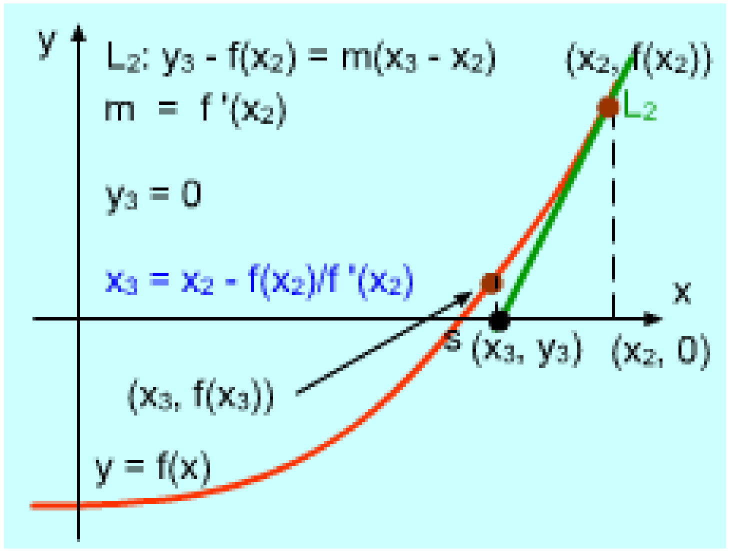
  - The critical point is $\boldsymbol{x}^* = x^{(0)} - H^{-1}g$

- When *f* is a positive definite quadratic function, Newton's method once to jump to the minimum of the function directly.
- When *f* is not truly quadratic but can be locally approximated as a positive definite quadratic, Newton's method consists of applying multiple jumping
- Jumping to the minimum of the approximation can reach the critical point much faster than gradient descent would.

# Example (For univariate function: 1st order case)







# Constrained Optimization

- Constraint optimization problem is generally written as

$$\min_{\boldsymbol{x}} f \quad \text{s.t.} \quad \underbrace{g_i(\boldsymbol{x}) = 0,}_{\text{equality constraint}} \quad \underbrace{h_i(\boldsymbol{x}) \le 0}_{\text{inequality constraint}}$$

- Karush-Kuhn-Tucker (KKT) Multiplier (Generalization of the Lagrange Multipliers)

$$\min_{x} \max_{\lambda_i} \max_{\mu_i \ge 0} L\left(x, \lambda_j, \mu_j\right)$$
$$= \min_{x} \max_{\lambda_i} \max_{\alpha \ge 0} \underbrace{f(x) + \sum \lambda_i g_i(x) + \sum \mu_i h_i(x)}_{\text{Generalized Lagrangian}}$$

# Karush-Kuhn-Tucker Condition

- ***KKT Conditions***: For a point to be optimal,

1. The gradient of the generalized Lagrangian is zero
2. All constraints on both x and the KKT multipliers are satisfied
3. The inequality constraints exhibit "complementary slackness": $\alpha \odot h(x) = 0$

$$\begin{cases} \nabla f(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i \nabla g_i(\bar{x}) + \sum_{j=1}^{l} \bar{\mu}_j \nabla h_j(\bar{x}) = 0 \\ h_j(\bar{x}) = 0 \ \ (j = 1, \ldots, l) \\ \bar{\lambda}_i \geq 0, \ g_i(\bar{x}) \leq 0, \ \bar{\lambda}_i g_i(\bar{x}) = 0 \ \ (i = 1, \ldots, m) \end{cases}$$

# Example: Linear Least Squares

- Consider an unconstrained problem of:

$$f(\boldsymbol{x}) = \frac{1}{2}||A\boldsymbol{x} - \boldsymbol{b}||_2^2$$

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = A^T(A\boldsymbol{x} - \boldsymbol{b}) = A^T A\boldsymbol{x} - A^T\boldsymbol{b}$$

$$\boldsymbol{x}^{n+1} \leftarrow \boldsymbol{x}^n - \epsilon(A^T A\boldsymbol{x}^n - A^T\boldsymbol{b})$$ (Gradient Decent)

- If subjected to $\boldsymbol{x}^T\boldsymbol{x} \leq \mathbf{1}$

$$\min_{x,\lambda\geq 0} \mathrm{L}(\boldsymbol{x},\lambda) \quad \mathrm{L}(\boldsymbol{x},\lambda) = f(\boldsymbol{x}) - \lambda(\boldsymbol{x}^T\boldsymbol{x} - 1)$$

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x},\lambda) = A^T A\boldsymbol{x} - A^T\boldsymbol{b} + 2\lambda\boldsymbol{x} = 0 \quad \boldsymbol{x} = (A^T A + 2\lambda I)^{-1}A^T\boldsymbol{b}$$

- The magnitude of $\lambda$ must obey the constraint:

  - Update $\lambda$ ($\geq 0$) until $\nabla_\lambda L(\boldsymbol{x},\lambda)$ becomes zero

$$\nabla_\lambda L(\boldsymbol{x},\lambda) = \boldsymbol{x}^T\boldsymbol{x} - 1$$