

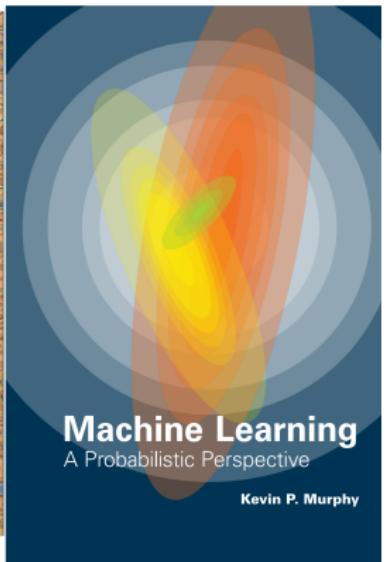
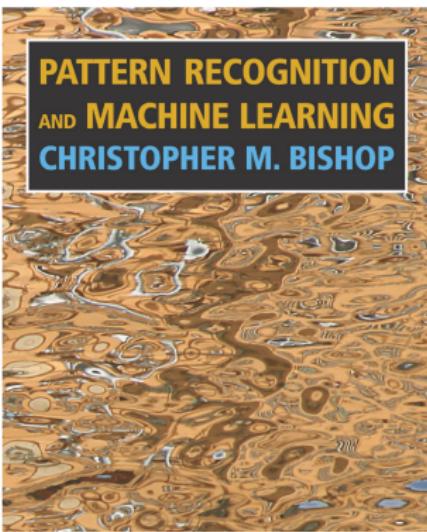
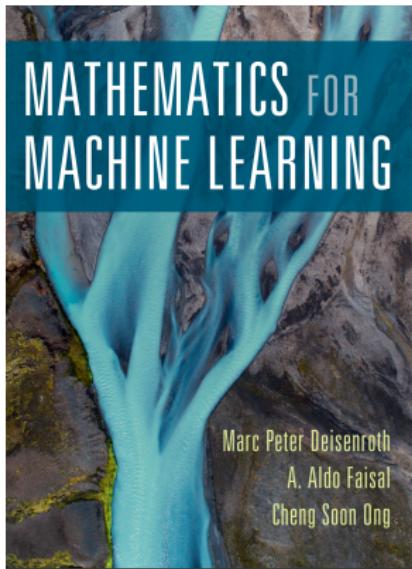
# Artificial Intelligence II (CS4442B & CS9542B)

## A Brief Review of Mathematics for Machine Learning

Boyu Wang  
Department of Computer Science  
University of Western Ontario

# Outline

- ▶ Linear Algebra
- ▶ Probability
- ▶ Vector Calculus & Optimization



# Why worry about the math?

- ▶ There are lots of easy-to-use machine learning packages out there.
- ▶ **However**, to get really useful results, you need good mathematical intuitions about certain machine learning principles, as well as the inner workings of the individual algorithms.
  - Choose the right algorithm(s) for the problem
  - Make good choices on parameter settings, validation strategies
  - Troubleshoot poor / ambiguous results
  - Do a better job of coding algorithms
  - Apply for a PhD at a top-tier university

# Linear Algebra

# Notation Reference

Table: Table of Notations

Notation	Meaning
$\mathbb{R}$	set of real numbers (one-dimensional space)
$\mathbb{R}^n$	set of $n$ -tuples of real numbers, ( $n$ -dimensional space)
$\mathbb{R}^{m \times n}$	set of $m \times n$ matrix space
$a$	a scalar or vector (i.e., $x \in \mathbb{R}$ or $x \in \mathbb{R}^n$ )
$A$	a matrix (i.e., $X \in \mathbb{R}^{m \times n}$ )
$I$	identity matrix
$A^{-1}$	inverse of a <i>square</i> matrix $A$ : $AA^{-1} = A^{-1}A = I$
$a^\top, A^\top$	transpose of a vector/matrix
$\langle a, b \rangle$	<b>dot product</b> of vectors: $\langle a, b \rangle = a^\top b = \sum_{i=1}^n a_i b_i$
$\ a\ _2, \ a\ _1$	$\ell_2, \ell_1$ -norm of $a$
$ A $	determinant of a <i>square</i> matrix $A$
$\text{tr}(A)$	trace of a <i>square</i> matrix $A$

# Linear algebra applications

- ▶ Operations on or between vectors and matrices.
- ▶ Dimensionality reduction.
- ▶ Linear regression.
- ▶ Many others.

# Why vectors and matrices?

Most common form of data organization for machine learning is a 2D array, where

- *rows* represent samples (records, items, datapoints)
- *columns* represent attributes (features, variables)

Natural to think of each sample as a *vector* of attributes, and whole array as a *matrix*

vector

matrix

Refund	Marital Status	Taxable Income	Cheat
Yes	Single	125K	No
No	Married	100K	No
No	Single	70K	No
Yes	Married	120K	No
No	Divorced	95K	Yes
No	Married	60K	No
Yes	Divorced	220K	No
No	Single	85K	Yes
No	Married	75K	No
No	Single	90K	Yes

# Vectors

- ▶ Definition: an  $n$ -tuple of values (usually real numbers).
- ▶ Can be written in column form or row form (**column form is conventional**). Vector elements referenced by subscript.

$$\mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = [a_1, \dots, a_n]^\top$$

- ▶ Can think of a vector as: a **point** in space or a **directed line segment** with a magnitude and direction.

# Vector arithmetic

- ▶ Addition of two vectors
  - add corresponding elements:  $a + b = [a_1 + b_1, \dots, a_n + b_n]^\top$
  - result is a vector
- ▶ Scalar multiplication of a vector
  - multiply each element by scalar:  $\lambda a = [\lambda a_1, \dots, \lambda a_n]^\top$
  - result is a vector.
- ▶ Inner/Dot product of two vectors
  - multiply corresponding elements, then add products:  
$$\langle a, b \rangle = a \cdot b = a^\top b = b^\top a = \sum_{i=1}^n a_i b_i$$
  - result is a **scalar**.
- ▶  $\ell_2$ -norm of a vector
  - $\|a\| = \sqrt{\langle a, a \rangle} = \sqrt{a^\top a} = \sqrt{\sum_{i=1}^n a_i^2}$
  - $a^\top b = \|a\| \|b\| \cos(\theta)$
  - Euclidean distance between two vectors:  $\|a - b\|$

# Matrices

A vector can be regarded as special case of a matrix, where one of matrix dimensions = 1.

- ▶ Matrix transpose (denoted  $\top$ ): swap columns and rows.  $m \times n$  matrix becomes  $n \times m$  matrix
- ▶ Addition of two matrices
- ▶ Scalar multiplication of a matrix

# Matrices multiplication

If  $A$  is an  $m \times n$  matrix (i.e.,  $A \in \mathbb{R}^{m \times n}$ ), and  $B$  is an  $n \times p$  matrix (i.e.,  $B \in \mathbb{R}^{n \times p}$ )

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{np} \end{bmatrix}$$

the matrix product  $C = AB$  (denoted without multiplication dots) is defined to be the  $m \times p$  matrix

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mp} \end{bmatrix},$$

where  $c_{ij} = \sum_{k=1}^n a_{ik} b_{kj}$ :  $c_{ij}$  is given by the vector product between the  $i$ -th row of  $A$  and the  $j$ -th column of  $B$ .

# Matrices multiplication

Properties:

- ▶  $A(BC) = (AB)C$
- ▶  $AB \neq BA$  (generally)
- ▶  $(AB)^\top = B^\top A^\top$
- ▶ **RULE:** In any chain of matrix multiplications, the column dimension of one matrix in the chain must match the row dimension of the following matrix in the chain.

# Square matrices and symmetric matrices

$A$  is a **square matrix** if it has the same number of rows and columns (i.e.,  $A \in \mathbb{R}^{n \times n}$ ). If  $A = A^\top$ , then  $A$  is also a **symmetric matrix**.

- ▶ Special kinds
  - diagonal matrix
  - identity matrix
  - positive-definite matrix:  $x^\top Ax > 0$  for any  $x$
  - invertible matrix and its **inverse**:  $A$  is invertible if or non-singular if there exists a matrix  $B$  such that  $AB = BA = I$ . If  $B$  exists, it is **unique** and is called the inverse matrix of  $A$ , denoted  $A^{-1}$ .
  - orthogonal matrix:  $A$  is an orthogonal matrix if  $A^\top = A^{-1}$ , which entails  $AA^\top = A^\top A = I$

# Eigenvalues and eigenvectors

Let  $A$  be a  $n \times n$  square matrix, if we can find a scalar  $\lambda$  and a unit vector  $v$ , such that

$$Av = \lambda v$$

then,  $\lambda$  is an eigenvalue of  $A$ , and  $v$  is its corresponding eigenvector.

- ▶  $A = V\Lambda V^{-1}$  is the **eigendecomposition** of  $A$ , where  $V$  is the  $n \times n$  matrix whose  $i$ -th column is the  $i$ -th eigenvector of  $A$ , and  $\Lambda$  is the diagonal matrix whose diagonal elements are the corresponding eigenvalues.
- ▶ If  $A$  is positive-definite  $\Rightarrow$  all the eigenvalues are positive
- ▶ If  $A$  is symmetric  $\Rightarrow$   $V$  is an orthogonal matrix:  $V^{-1} = V^T$

# Probability

# Why probability

To characterize the uncertainties of the world!

- ▶ Uncertain **data**
  - Missing data
  - Noisy data
- ▶ Uncertain **knowledge**
  - Incomplete enumeration of conditions or effects
  - Incomplete knowledge of causality in the domain
  - Stochastic effects

Probability theory provides powerful tools for modeling and dealing with uncertainty.

# Interpretations of probability

*The probability that a coin will land heads is 0.5*

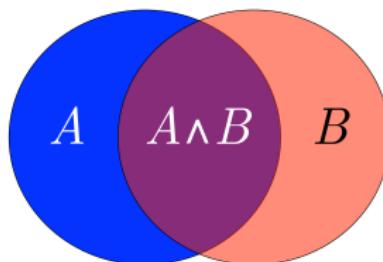
- ▶ **Frequentist** interpretation: to represent long run frequencies of events.
  - If we flip the coin many times, we expect it to land heads about half the time.
- ▶ **Bayesian** interpretation: to quantify our uncertainty about something – related to information rather than repeated trials.
  - We believe the coin is equally likely to land heads or tails on the next toss.

# Random variables

The expression  $p(A)$  denotes the probability that the event  $A$  is true.

- ▶  $0 \leq p(A) \leq 1$
- ▶  $p(\bar{A})$  denotes the probability that the event  $A$  is false:  
 $p(\bar{A}) = 1 - p(A)$ .
- ▶ The probability of a disjunction is given by:

$$p(A \vee B) = p(A) + p(B) - p(A \wedge B)$$



# Fundamental concepts

- ▶ A **union** of two events – the probability of  $A$  **or**  $B$ :  
 $p(A \vee B) = p(A) + p(B) - p(A \wedge B)$ . If  $A$  and  $B$  are mutually exclusive,  $p(A \vee B) = p(A) + p(B)$
- ▶ **Joint probability** – the probability of the joint event  $A$  **and**  $B$ :  
 $p(A, B) = p(A \wedge B) = p(A|B)p(B)$ , where  $p(A|B)$  is the **conditional probability** of event  $A$  given  $B$ :

$$p(A|B) = \frac{p(A, B)}{p(B)}$$

If  $A$  and  $B$  are independent to each other, we have  
 $p(A|B) = P(A)$ .

- ▶ Chain rule:

$$p(X_{1:N}) = p(X_1)p(X_2|X_1)p(X_3|X_1, X_2), \dots, p(X_N|X_{1:N-1})$$

# Bayes Rule

## Bayes Rule

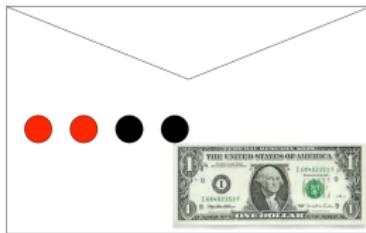
$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)}$$

$$p(\text{hypothesis}|\text{evidence}) = \frac{p(\text{evidence}|\text{hypothesis}) \times p(\text{hypothesis})}{p(\text{evidence})}$$

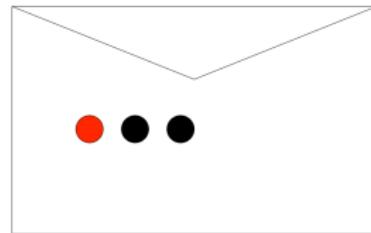
$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal distribution}}$$

- ▶ The most important formula in probabilistic machine learning
- ▶ Allows us to reason from **evidence** to **hypotheses**
  - Example:  $p(\text{headache}) = \frac{1}{10}$ ,  $p(\text{flu}) = \frac{1}{40}$ ,  
 $p(\text{headache}|\text{flu}) = \frac{1}{2}$ . Given the evidence of headache,  
what is  $p(\text{flu}|\text{headache})$ ?

# Using Bayes Rule to Gamble



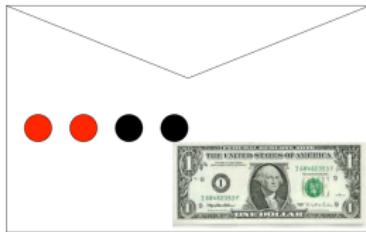
The “Win” envelope has a dollar and four beads in it



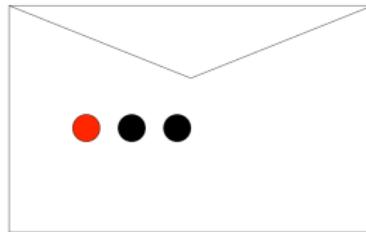
The “Lose” envelope has three beads and no money

**Trivial question:** Someone draws an envelope at random and offers to sell it to you.  
How much should you pay?

# Using Bayes Rule to Gamble



The “Win” envelope has a dollar and four beads in it



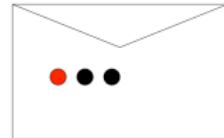
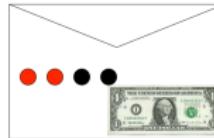
The “Lose” envelope has three beads and no money

**Interesting question:** Before deciding, you are allowed to see one bead drawn from the envelope.

Suppose it's black: How much should you pay?

Suppose it's red: How much should you pay?

# Calculation...



Suppose it's black: How much should you pay?

$$P(b \mid \text{win}) = 1/2 \quad P(b \mid \text{lose}) = 2/3$$

$$P(\text{win}) = 1/2$$

$$\begin{aligned} P(\text{win} \mid b) &= \alpha P(b \mid \text{win}) P(\text{win}) \\ &= \alpha 1/2 \times 1/2 = 0.25\alpha \end{aligned}$$

$$\begin{aligned} P(\text{lose} \mid b) &= \alpha P(b \mid \text{lose}) P(\text{lose}) \\ &= \alpha 2/3 \times 1/2 = 0.3333\alpha \end{aligned}$$

$$1 = P(\text{win} \mid b) + P(\text{lose} \mid b) = 0.25\alpha + 0.3333\alpha \rightarrow \alpha = 1.714$$

$$P(\text{win} \mid b) = 0.4286 \quad P(\text{lose} \mid b) = 0.5714$$

Based on example by Andrew Moore

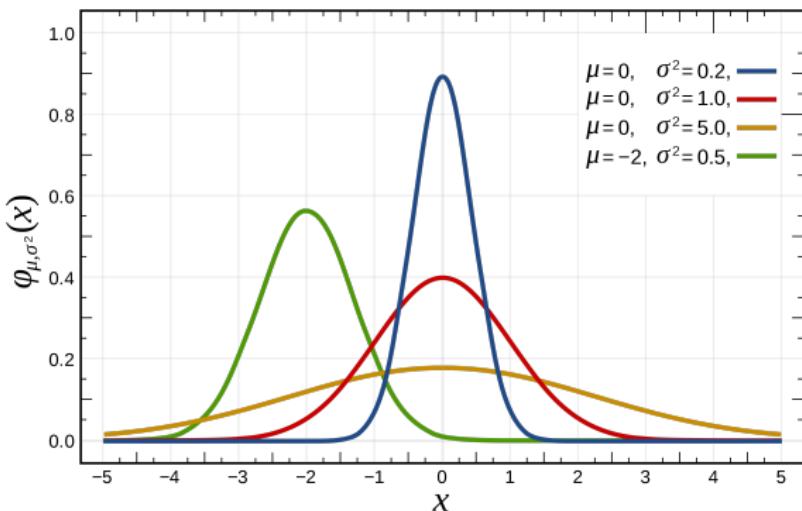
# Probability distributions

- ▶ Discrete distributions
  - binomial and Bernoulli distributions
  - multinomial and multinoulli distributions
  - Poisson distribution
- ▶ Continuous distributions
  - Gaussian distribution
  - Laplace distribution
  - gamma distribution

# Gaussian distribution

1-dimensional Gaussian distribution

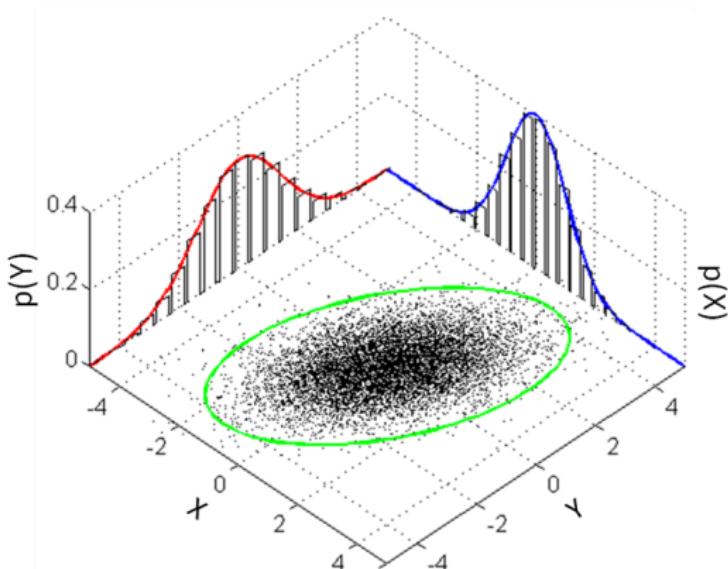
$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



# Gaussian distribution

$n$ -dimensional (multivariate) Gaussian distribution

$$\mathcal{N}(\mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}$$



# Vector Calculus & Optimization

# Fundamental concepts

- ▶ **derivative:** the sensitivity to change of the function value (output value) with respect to a change in its argument (input value).

$$f'(x) = \lim_{a \rightarrow 0} \frac{f(x + a) - f(x)}{a}$$

**second derivative –  $f''(x)$ :** the derivative of  $f'(x)$

- ▶ **convex function:**

$$\forall x_1, x_2, \forall t \in [0, 1] : f(tx_1 + (1 - t)x_2) \leq tf(x_1) + (1 - t)f(x_2)$$

- $f''(x) \geq 0 \Leftrightarrow f(x)$  is a convex function
- If  $f(x)$  is a convex function,  $f'(x_0) = 0 \Rightarrow x_0$  is the global minimum point (i.e.,  $f(x_0) \leq f(x)$ )

- ▶ **chain rule:** Let  $F = f(g(x))$ , then  $F'(x) = f'(g(x))g'(x)$

# Vector Calculus

- ▶ **gradient:** the derivative of a multi-variable function  $f$  with respect to  $x = [x_1, \dots, x_n]^\top$

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

where  $\frac{\partial f}{\partial x_i}$  is the partial derivative of  $f$  with respect to  $x_i$ .

- ▶  $f(x) = x_1^2 + 3x_2 + x_2x_3$ :

$$\frac{\partial f}{\partial x_1} = 2x_1, \quad \frac{\partial f}{\partial x_2} = 3 + x_3, \quad \frac{\partial f}{\partial x_3} = x_2$$
$$\Rightarrow \nabla f(x) = [2x_1, 3 + x_3, x_2]^\top$$

- ▶ Let  $f(x) = a^\top x$ ,  $f(x) = x^\top Ax$ , what is  $\nabla f(x)$ ?

- ▶ **Hessian matrix** : the second derivative of a multi-variable function  $f$  with respect to  $x = [x_1, \dots, x_n]^\top$

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

where  $\frac{\partial^2 f}{\partial x_i \partial x_j}$  is the mixed partial derivative of  $f$ . The order of differentiation does not matter (Schwarz's theorem).

- ▶  $f(x) = x_1^2 + 3x_2 + x_2x_3$ :

$$\frac{\partial f}{\partial x_1} = 2x_1, \quad \frac{\partial f}{\partial x_2} = 3 + x_3, \quad \frac{\partial f}{\partial x_3} = x_2$$

$$\frac{\partial f}{\partial x_1^2} = 2, \quad \frac{\partial f}{\partial x_2^2} = \frac{\partial f}{\partial x_3^2} = 0, \quad \frac{\partial f}{\partial x_1 \partial x_2} = \frac{\partial f}{\partial x_1 \partial x_3} = 0, \quad \frac{\partial f}{\partial x_2 \partial x_3} = 1$$

$$\Rightarrow H = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

# Convex multi-variable function

$$\forall x_1, x_2, \forall t \in [0, 1] : f(tx_1 + (1 - t)x_2) \leq tf(x_1) + (1 - t)f(x_2)$$

- $H$  is positive-definite  $\Leftrightarrow f(x)$  is a convex function
- If  $f(x)$  is a convex function,  $\nabla f(x_0) = 0 \Rightarrow x_0$  is the global minimum point (i.e.,  $f(x_0) \leq f(x)$ )
- $f(x) = x_1^2 + 3x_2 + x_2x_3$ :

$$H = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

The eigenvalues of  $H$  are -1, 1, 2  $\Rightarrow f(x)$  is not convex

# Function minimization

- ▶ Most machine learning problems can be formulated as a function minimization problem

# Function minimization

- ▶ Most machine learning problems can be formulated as a function minimization problem
- ▶ Solve the equation:

$$\nabla f(x) = 0 \tag{1}$$

Done!

# Function minimization

- ▶ Most machine learning problems can be formulated as a function minimization problem
- ▶ Solve the equation:

$$\nabla f(x) = 0 \tag{1}$$

Done!

- ▶ Really?

# Function minimization

- ▶ Most machine learning problems can be formulated as a function minimization problem
- ▶ Solve the equation:

$$\nabla f(x) = 0 \tag{1}$$

Done!

- ▶ Really?
  - If  $f$  is not convex, we obtain a suboptimal solution.

# Function minimization

- ▶ Most machine learning problems can be formulated as a function minimization problem
- ▶ Solve the equation:

$$\nabla f(x) = 0 \tag{1}$$

Done!

- ▶ Really?
  - If  $f$  is not convex, we obtain a suboptimal solution.
  - We don't have an analytical solution for (1).  
Example:  $f(x) = x_1^2 + e^{x_2} + \sin(x_3 + \log(x_1))$ .

# Function minimization

- ▶ Most machine learning problems can be formulated as a function minimization problem
- ▶ Solve the equation:

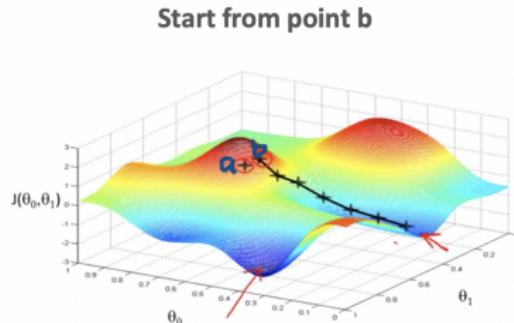
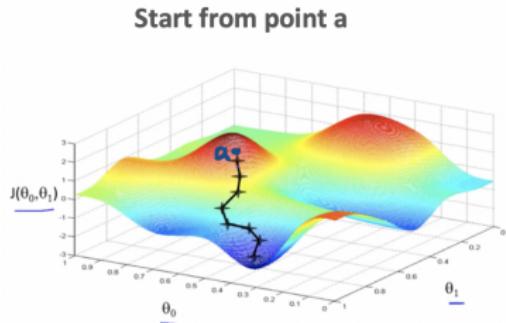
$$\nabla f(x) = 0 \tag{1}$$

Done!

- ▶ Really?
  - If  $f$  is not convex, we obtain a suboptimal solution.
  - We don't have an analytical solution for (1).  
Example:  $f(x) = x_1^2 + e^{x_2} + \sin(x_3 + \log(x_1))$ .
- ▶ Gradient descent!

# Gradient descent

An optimization algorithm used to minimize a function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient.



Starting from 2 slightly separated points we reached 2 different pits or minimas or local optimum.

(Property of Gradient Descent)