

# ECBM E4040 Neural Networks and Deep Learning

## Lecture #3: Elements of Probability and Information Theory, and Numerical Computation

Instructor Zoran Kostic

Columbia University  
Department of Electrical Engineering

September 15, 2016

# Outline of Part I

- 2 Summary of the Previous Lecture
  - Topics Covered
  - Learning Objectives

# Outline of Part II

- 3 Probability Spaces and Random Variables
  - Probability Spaces
  - Random Variables and Stochastic Processes
  - Conditional Probability

# Outline of Part III

- ④ Gradient-Based Optimization
- ⑤ Constraint Optimization
  - Karush-Kuhn-Tucker Approach
- ⑥ Linear Least Squares

# Part I

## Review of Previous Lecture

# Topics Covered

## Elements of Linear Algebra

- Finite Dimensional Vector Spaces
- Eigendecomposition and SVD
- Principal Component Analysis

# Learning Objectives

- Reviewing key elements of Singular Value Decomposition
- Definition of the Psedo-Inverse and characterizing its properties
- Formulating and deriving Principal Component Analysis

## Part II

# Elements of Probability and Information Theory



# Basic Overview

Probability theory is an axiomatic branch of measure theory that provides

- means of quantifying uncertainty,
- a formal calculus to derive statements about uncertain events.

The theory of probability constitutes a formidable body of knowledge that provides a set of tools indispensable in machine learning.

# The Notion of Probability

The basic notion in probability theory is that of a **random experiment** whose outcome cannot be determined in advance. A set of all possible outcomes of an experiment is called a **sample space** and it is usually denoted by  $\Omega$ .

An **event** is a subset of a sample space. An event  $A \subset \Omega$  is set to occur iff the observed outcome  $\omega$  is an element of the set  $A$ .

## Example

Consider an experiment that consists of counting the number of traffic accidents at an intersection during rush hour. Here  $\Omega = \{0, 1, 2, \dots\}$  and  $A = \{0, 1, \dots, 7\}$  is the event describing that the number of accidents is less than or equal to 7. The event  $A = \{5, 6, 7, \dots\}$  occurs iff the number of accidents is 5 or 6 or ...

# Frequently Encountered Events

Given  $A \subset \Omega$ , the **complement**  $A^c$  of  $A$  is defined to be the event which occurs iff  $A$  does not occur, that is,

$$A^c = \{\omega \in \Omega \mid \omega \notin A\}.$$

Given two events  $A$  and  $B$ , their **union** is the event which occurs iff either  $A$  or  $B$  (or both) occurs, that is

$$A \cup B = \{\omega \in \Omega \mid \omega \in A \text{ or } \omega \in B\}.$$

The intersection of  $A$  and  $B$  is the event which occurs iff both  $A$  and  $B$  occur, that is

$$A \cap B = \{\omega \in \Omega \mid \omega \in A \text{ and } \omega \in B\}.$$

## Frequently Encountered Events (cont'd)

The operations of taking unions, intersections, and complements can be combined to obtain new events. For example,

$$(A \cup B)^c = A^c \cap B^c, \quad (A \cap B)^c = A^c \cup B^c.$$

Two events are set to be **disjoint** if they have no element in common, that is

$$A \cap B = \emptyset.$$

If two events are disjoint, the occurrence of one implies that the other has not occurred. A family of events is called disjoint if every pair of them are disjoint.

## Definition

Let  $\Omega$  be a sample space and  $\mathbb{P}$  a function which associates a number with each event. Then  $\mathbb{P}$  is called a **probability measure** provided that

- (i) for every event  $A$ ,  $0 \leq \mathbb{P}(A) \leq 1$  ;
- (ii)  $\mathbb{P}(\Omega) = 1$  ;
- (iii) for any sequence  $A_1, A_2, \dots$  of disjoint events

$$\mathbb{P}(\cup_{i \in \mathbb{N}} A_i) = \sum_{i \in \mathbb{N}} \mathbb{P}(A_i).$$

## Remark

*It might not be practically possible to assign a probability to each event in an explicit fashion. Often, therefore, the probabilities of only a few key events are specified. The remaining probabilities are computed from the axioms above.*

# Computing Event Probabilities

Here are a some of the key results that can be directly obtained from the three axioms.

If  $A_1, \dots, A_n$  are disjoint events

$$\mathbb{P}(A_1 \cup \dots \cup A_n) = \mathbb{P}(A_1) + \dots + \mathbb{P}(A_n).$$

If  $A \subset B$ , then  $\mathbb{P}(A) \leq \mathbb{P}(B)$ .

For any event  $A$ ,  $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$ .

If  $B_1, B_2, \dots$  are disjoint events with  $\cup_{i \in \mathbb{N}} B_i = \Omega$ , then for any event  $A$ ,

$$\mathbb{P}(A) = \sum_{i \in \mathbb{N}} \mathbb{P}(A \cap B_i).$$

# Random Variables

## Definition

A random variable  $\mathbf{x}$  with values in a set  $E$  is a function which assigns a value  $\mathbf{x}(\omega)$  in  $E$  to each outcome  $\omega \in \Omega$ .

## Definition

A stochastic process with state space  $E$  is a collection  $\mathbf{x}_t, t \in \mathbb{R}$ , defined on the same probability space and taking values in  $E$ .

# Independence and Conditional Probability

## Definition

The discrete random variables  $\mathbf{x}_1, \dots, \mathbf{x}_n$  are said to be independent if

$$\mathbb{P}(\mathbf{x}_1 = a_1, \dots, \mathbf{x}_n = a_n) = \mathbb{P}(\mathbf{x}_1 = a_1) \cdots \mathbb{P}(\mathbf{x}_n = a_n)$$

## Definition

Let  $A$  and  $B$  be two events. The conditional probability of  $A$  given  $B$ , written  $\mathbb{P}(A \mid B)$ , is a number satisfying

- (i)  $0 \leq \mathbb{P}(A \mid B) \leq 1$ ,
- (ii)  $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B)$ .



## Conditional Probabilities (cont'd)

If  $B_1, B_2, \dots$  are disjoint events with  $\cup_{i \in \mathbb{N}} B_i = \Omega$ , then for any event  $A$ ,

$$\mathbb{P}(A) = \sum_{i \in \mathbb{N}} \mathbb{P}(A|B_i)Pr(B_i).$$

### Lemma (Bayes' Formula)

*If  $B_1, B_2, \dots$  are disjoint events with union  $\Omega$ , then for any event  $A$  with  $\mathbb{P}(A) > 0$ , and any  $j$ ,*

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_{i \in \mathbb{N}} \mathbb{P}(A|B_i)\mathbb{P}(B_i)}.$$

## Part III

# Elements of Numerical Computation

# Optimization in Deep Learning

Most deep learning algorithms involve the minimization of maximization of a function called **objective function** or **criterion**. In minimization, these functions are also called **cost functions**, **loss functions** or **error functions**.

The value that minimizes or maximizes a function is denoted by a superscript  $*$ :

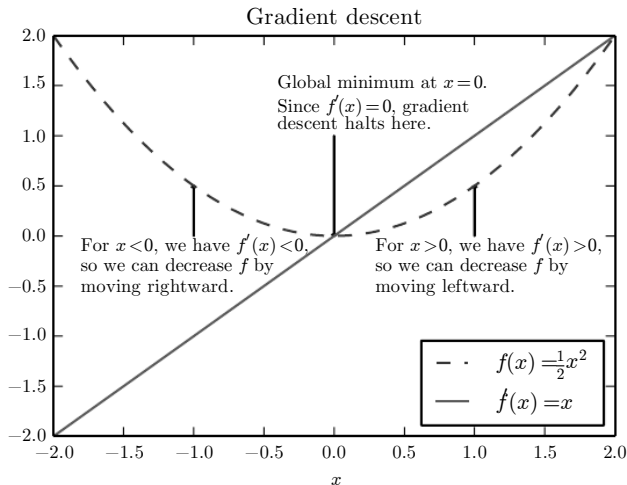
$$\mathbf{x}^* = \arg \min_{\mathbf{x}} f(\mathbf{x}).$$

Assume that  $y = f(x)$  with  $x, y \in \mathbb{R}$ . Points where

$$\frac{df(x)}{dx} = 0$$

are called **critical or stationary points**. These can be local minima, local maxima, saddle points, or global minima/maxima.

# The Intuition Behind the Gradient Descent



The derivative of the function can be used to follow the function downhill to a minimum. This technique is called gradient descent.

# Gradient Descent for Functions with Multiple Inputs

Here we consider functions of the form  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ . The partial derivative  $\frac{\partial}{\partial x_i} f(\mathbf{x})$  measures how  $f$  changes as a function of  $x_i$  at point  $\mathbf{x}$ . The gradient generalizes to the notion of derivative with respect to a vector: the gradient of  $f$  is the vector containing all the partial derivatives, denoted by  $\nabla f(\mathbf{x})$ .

The **directional derivative** in direction  $\mathbf{u}$  (a unit vector) is the slope of the function  $f$  in direction  $\mathbf{u}$ , i.e., the derivative of the function  $f(\mathbf{x} + \alpha\mathbf{u})$  with respect to  $\alpha$ , evaluated at  $\alpha = 0$ .

Using the chain rule, the directional derivative amounts to

$$\mathbf{u}^T \nabla f(\mathbf{x}).$$

# Gradient Descent: Functions with Multiple Inputs (cont'd)

To minimize  $f$ , we have to find the direction in which  $f$  decreases the fastest. Using the directional derivative we have

$$\min_{\mathbf{u}, \mathbf{u}^T \mathbf{u} = 1} \mathbf{u}^T \nabla f(\mathbf{x}) = \min_{\mathbf{u}, \mathbf{u}^T \mathbf{u} = 1} \|\mathbf{u}\|_2 \|\nabla f(\mathbf{x})\|_2 \cos \theta,$$

where  $\theta$  is the angle between  $\mathbf{u}$  and the gradient. Since  $\|\mathbf{u}\|_2 = 1$  the minimization above can be reduced to

$$\min_{\mathbf{u}} \cos \theta.$$

The minimum of the cos function above is  $-1$  and is achieved when  $\mathbf{u}$  points in the opposite direction of the gradient.

# Gradient Descent: Functions with Multiple Inputs (cont'd)

## The Method of Steepest Descent

Steepest descent chooses a new point

$$\mathbf{x}' = \mathbf{x} - \varepsilon \nabla f(\mathbf{x}),$$

where  $\varepsilon$  is the size of the step. There are many ways to choose  $\varepsilon$  (it is an art ...).

# Problem Formulation and the KKT Approach

Often, we may wish to find the maximal or minimal value of  $f(\mathbf{x})$  for values of  $\mathbf{x}$  in some set  $\mathbb{S}$ . This is known as constrained optimization. Points  $\mathbf{x}$  that lie within the set  $\mathbb{S}$  are called **feasible points** in constrained optimization terminology.

A very general solution to constrained optimization problem above is provided by the KarushKuhnTucker (KKT) approach. The KKT approach, is based on introducing a new function called the generalized Lagrangian or generalized Lagrange function.

To define the Lagrangian, we'll first describe  $\mathbb{S}$  in terms of equations and inequalities. We want a description of  $\mathbb{S}$  in terms of  $m$  functions  $g_i$  and  $n$  functions  $h_j$  so that

$$\mathbb{S} = \{\mathbf{x} | \forall i, g_i(\mathbf{x}) = 0 \text{ and } \forall j, h_j(\mathbf{x}) \leq 0\}.$$

The equations involving  $g_i$  are called the equality constraints and the inequalities involving  $h_j$  are called inequality constraints.



# Problem Formulation and the KKT Approach

The generalized Lagrangian is then defined as

$$L(\mathbf{x}, \lambda, \alpha) = f(\mathbf{x}) + \sum_i \lambda_i g_i(\mathbf{x}) + \sum_j \alpha_j h_j(\mathbf{x}).$$

We solve the constrained minimization problem using unconstrained optimization of the generalized Lagrangian. Observe that, so long as at least one feasible point exists and  $f(\mathbf{x})$  is not permitted to have the value  $\infty$ , then

$$\min_{\mathbf{x}} \max_{\lambda} \max_{\alpha, \alpha \geq 0} L(\mathbf{x}, \lambda, \alpha),$$

has the same optimal objective function value and set of optimal points  $\mathbf{x}$  as

$$\min_{\mathbf{x} \in \mathbb{S}} f(\mathbf{x}).$$

# Problem Formulation and the KKT Approach

The above follows because any time the constraints are satisfied,

$$\max_{\lambda} \max_{\alpha, \alpha \geq 0} L(\mathbf{x}, \lambda, \alpha) = f(\mathbf{x}),$$

while any time a constraint is violated,

$$\max_{\lambda} \max_{\alpha, \alpha \geq 0} L(\mathbf{x}, \lambda, \alpha) = \infty,$$

These properties guarantee that no infeasible point will ever be optimal, and that the optimum within the feasible points is unchanged.

# Unconstrained Linear Least Squares

Find the value of  $\mathbf{x}$  that minimizes

$$f(\mathbf{x}) = \frac{1}{2} \|\mathbf{Ax} - \mathbf{b}\|_2^2.$$

To apply the gradient-based optimization, we derive the gradient

$$\nabla f(\mathbf{x}) = \mathbf{A}^T(\mathbf{Ax} - \mathbf{b}) = \mathbf{A}^T \mathbf{Ax} - \mathbf{A}^T \mathbf{b}.$$

We now just apply the standard steepest decent algorithm

$$\mathbf{x}' = \mathbf{x} - \varepsilon(\mathbf{A}^T \mathbf{Ax} - \mathbf{A}^T \mathbf{b}).$$

# Constrained Linear Least Squares

## An Example

We will minimize the same function  $f(\mathbf{x})$ , but subject to the constraint  $\mathbf{x}^T \mathbf{x} \leq 1$ . We introduce the Lagrangian

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda(\mathbf{x}^T \mathbf{x} - 1).$$

We can now solve the problem

$$\min_{\mathbf{x}} \max_{\lambda, \lambda \geq 0} L(\mathbf{x}, \lambda).$$

The solution to the unconstrained least squares problem is given by  $\mathbf{x} = \mathbf{A}^+ \mathbf{b}$ . If this point is feasible, then it is the solution to the constrained problem. Otherwise, we must find a solution where the constraint is active. By differentiating the Lagrangian with respect to  $\mathbf{x}$ , we obtain the equation

$$\mathbf{A}^T \mathbf{A} \mathbf{x} - \mathbf{A}^T \mathbf{b} + 2\lambda \mathbf{x} = 0.$$

# Constrained Linear Least Squares (cont'd)

## An Example

This tells us that the solution will take the form

$$\mathbf{x} = (\mathbf{A}^T \mathbf{A} + 2\lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{b}.$$

The magnitude of  $\lambda$  must be chosen such that the result obeys the constraint. We can find this value by performing gradient ascent on  $\lambda$ . To do so, observe that

$$\frac{\partial}{\partial \lambda} L(\mathbf{x}, \lambda) = \mathbf{x}^T \mathbf{x} - 1.$$

When the norm of  $\mathbf{x}$  exceeds 1, this derivative is positive, so to ascend the gradient and increase the Lagrangian with respect to  $\lambda$ , we increase  $\lambda$ . This will in turn shrink the optimal  $\mathbf{x}$ . The process continues until  $\mathbf{x}$  has the correct norm and the derivative on  $\lambda$  is 0.