Optimisation for Deep Learning

1 Introduction

This course will be offered in Semester 2 2021 (ACE).

Lecturers:

- Dr. Vera Roshchina, UNSW (v.roshchina@unsw.edu.au)
- Dr. Nadezda (Nadia) Sukhorukova, Swinburne University of Technology (nsukhorukova@swin.edu.au)
- Dr. Julien Ugon, Deakin University (julien.ugon@deakin.edu.au)

Textbooks:

1. Optimisation part: Convex Optimization by Stephen Boyd and Lieven Vandenberghe, Cambridge University Press;
2. Deep learning part: Deep learning by Yoshua Bengio, Ian Goodfellow, Aaron Courville.

Both books are available online.

2 Motivation

A number of problems in machine learning and, in particular, deep learning, can be formulated as optimisation problems and solved using a suitable optimisation method. Most modern software packages on deep learning use a default optimisation method ("black box"). For many applications, these methods are suitable and the users only need to change the number of layers, nodes, change the activation function, etc. In other cases, however, users need to open the "black box", in order to improve the performance. The main purpose of this course is to provide a good theoretical foundation to optimisation theory, apply it to formulate optimisation problems appearing in deep learning models and provide a range of optimisation methods that can be used to solve these problems. This unit can be seen as a mathematical unit focused on deep learning as its important potential application.

3 Short history

The term "Deep Learning" was introduced by Rina Dechter in 1986. Most models are based on the so called Universal Approximation Theorem.

Theorem 3.1 *Universal approximation theorem, George Cybenko, 1989*
Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a nonconstant, bounded, and continuous function (called the activation function). Let I_m be a compact set in $\mathbb{R}^m$. The space of real-valued

continuous functions on I_m is denoted by $C(I_m)$. Then, given any $\varepsilon > 0$ and any function $f \in C(I_m)$, there exist an integer N, real constants $v_i,\ b_i \in \mathbb{R}$ and real vectors

$$w_i \in \mathbb{R}^m,\ i = 1, \ldots, N,$$

such that we may define:

$$F(x) = \sum_{i=1}^{N} v_i \varphi\left(w_i^T x + b_i\right)$$

and $|F(x) - f(x)| \leq \varepsilon$ for all $x \in I_m$. In other words, functions of the form $F(x)$ are dense in $C(I_m)$.

Cybenko proved this theorem for the case when the activation function φ is a sigmoid function. Later these results were improved by Kurt Hornek (1991) and Leshno et. al. (1993).

Generally speaking, neural network is not the first discovered "universal approximators". Classical polynomials were the first proved "universal approximators" (so-called StoneWeierstrass approximation theorem, 1885).

When the activation function is not used (that is, this function is an affine function) the problem becomes a simple linear regression and this is a well-studied convex optimisation problem (least squares). When the activation functions are more complex, the optimisation problems are nonconvex and require advanced optimisation skills to understand the "black box" output and how this result can be improved.

4 Preliminary structure

- Weeks 1-2: Introduction to optimisation, machine learning and deep learning: general terminology and convention. Overview of optimisation problems appearing in deep learning.
- Weeks 3-4: Linear optimisation and elements of linear integer optimisation.
- Weeks 5-6: Convex optimisation.
- Weeks 7-8: Non-convex optimisation.
- Weeks 9-12: Guided study and project submission.