Chair Data Science and Artificial Intelligence for Digitalized Industry and Services

Internship subject : Asymptotics of Deep Residual Networks

Possibility to continue as a PhD candidate : Yes

About the chair : The Chair Data Science and Artificial Intelligence for Digitalized Industry and Services (DSAIDIS), lead by Florence d'Alché-Buc, a Professor in the department Image, Data, Signal of Telecom Paris, unites five industrial partners : Airbus Defence & Space, Engie, Idemia, Safran et Valeo. Its objective is to develop teaching and research of the international level, in collaboration with the partners.

Supervision :

— Pascal Bianchi, Prof., Telecom Paris/LTCI, Institut Polytechnique de Paris
— Walid Hachem, Dir. CNRS, LIGM, Univ. Gustave Eiffel
— François-Xavier Vialard, Prof., LIGM, Univ. Gustave Eiffel

Location and dates of the internship
Telecom Paris, 19 Place Marguerite Perey, 91120 Palaiseau
Department IDS, Team Signal, Statistique et Apprentissage (S2A)
Start date : Early 2024

Keywords : Stochastic optimization, deep learning, residual networks, functional optimization.

Detailed subject

What are ResNets ? : Residual Networks is a type of neural network architecture that was introduced in 2016 to address the challenges of training very deep neural networks. The key innovation is the use of residual blocks, which incorporate skip connections. In a standard neural network without skip connections, each layer is expected to learn a direct mapping from input to output. However, as the network becomes deeper, it becomes increasingly difficult to train, and the risk of vanishing/exploding gradients arises.

The residual blocks in ResNet address this issue by allowing the network to learn the difference between the current layer's input and output. This residual learning provide a shortcut for the gradient to flow directly through the network, mitigating the vanishing gradient problem, and making it easier to train very deep networks. ResNet architectures have been widely adopted and adapted for various applications beyond image recognition, including natural language processing, speech recognition, and more, making them a significant contribution to the field of deep learning.

Mathematical model : Consider a ResNet architecture with K layers. Each layer k is governed by some free parameters $\theta_k \in \mathbb{R}^d$, so that the total number of free parameters is $K \times d$. The input-output relation $y = h_{\theta_1,\dots,\theta_K}(x)$ is given by $y = x(K)$, where $x(K)$ is iteratively defined by :

$$x(k+1) = x(k) + \varphi(\theta_k, x(k))\,,$$

and $x(0) = x$. Here, the function φ represents the residual block. Let (X,Y) be a random couple, where Y is a label, and X is a feature vector. Given a loss function $\ell(\cdot,\cdot)$, the aim is to identify the Resnets coefficient which minimize the risk

$$R(\theta_1,\dots,\theta_K) = \mathbb{E}(\ell(Y, h_{\theta_1,\dots,\theta_K}(X)))\,.$$

The learner uses stochastic gradient descent (SGD) in order to learn the coefficients from a data set.

Proposed work : The aim is to analyze the convergence of SGD applied to ResNet architectures, as the number of iterations goes to infinity. The main challenge is that the risk $R(\theta_1, \dots, \theta_K)$ has many spurious critical points. Thus, in the standard setting, SGD may converge to local minimizers about which little can be said, making the convergence result non informative. An alternative is to assume that the number K of layers tends to infinity. In this limiting regime, the input-output relation of the ResNet can be approximated by a ordinary differential equation (ODE), called the Neural-ODE :

$$\frac{dx(t)}{dt} = \varphi_t(x(t)), \quad (t \in [0,1])$$

with $x(0) = x$ is the input, and $x(1) = y$ is the output. Here, the vector field φ_t plays the role of an infinitesimal residual block of the ResNet architecture. The risk minimization problem becomes :

$$\min_{\varphi:[0,1]\times\mathbb{R}^d \to \mathbb{R}^d} \mathbb{E}(\ell(Y, h_\varphi(X))), \tag{1}$$

where $h_\varphi(X)$ is the output of the Neural-ODE with input X. Such type of problems have been widely explored in optimal control theory.

The aim of the project is 1) to review the literature on optimal control theory related to the control problem (1), and 2) to adapt SGD to the case of infinite number of layers, and analyze its convergence.

Candidate profile : Master 2 in one of the these domains : probability, statistics, optimization, datascience

Application : To send at pascal.bianchi@telecom-paris.fr