Seminar Review 4.27-5.3

Jiaxin Hu

04/29/2020

IFDS 4.27
Title: Controlling Gradient Decay in RNN using Adjoint Mechanics
Author: Liam Johnston, advised by Vivak Patel

This talk aims to combat the vanishing gradient problem in the Adjoint method of backpropagation RNN. The Adjoint method is an efficient way to compute the gradient of objective function via corresponding Lagrangian. Gradient vanishing leads the layers closest to the outcome to dominate the parameter updating. The presenter introduces a co-adjoint method to address the gradient vanishing, where a penalty term of vanishing Lagrange multiplier λ_t, $G(\lambda_1, ..., \lambda_T)$, is added to the objective function. The penalty term contains the penalty for small λ_t, $\phi(\lambda_t)$, and the variance between adjoint size. Simulations show better accuracy of this co-adjoint method over LSTM and typical Adjoint method.

Note that the gradient vanishing problem is a numerical issue because every middle term x_t contains the information of every previous input $u_1, ..., u_t$ due to the RNN nature. If we have a super process to estimate the parameters of the network, no information will be lost. Besides, the sensitivity of the penalty $\phi(\lambda_t)$ should be investigated more.

Questions:
1. Since the objective function of this co-adjoint method is no longer equal to the original objective function, how to ensure the minimizers of these two functions are close?
Possible Answer: Since the term $G(\lambda_t)$ is also a function of network parameters actually, I guess the minimizer of the co-adjoint method is a refined version of the original objective function, like the penalized likelihood methods. However, I think it is difficult to write the explicit relationship between two minimizesr as the network layers increasing.

SILO 4.29
Title: Why some robust estimators are efficiently computable
Author: Jiantao Jiao, UC Berkeley

This talk explains why we can find a robust estimate in the finite-sample corruption model theoretically and computationally. The problem is formulated as a minimization problem: $\min \|\Sigma_q\|_2,\ s.t.\ q \in \Delta_{n,\epsilon}$. First, the presenter proves that the KKT point for the program is approximate the global minimum if the proportion of the corrupted data is smaller than 1/3. Second, a gradient descent method that ignores the constrain is showed to find the KKT point efficiently, though this algorithm is not universally guaranteed in any case. Third, the presenter proposes the low-regret generalization for KKT point with respect to the constrain, which is a universal way to find the KKT.