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~~ combats the vanishing gradient problem in ~~the Adjoint method of~~ backpropagation RNN using Adjoint method. The Adjoint method ~~is an efficient way to~~ computes the gradient of objective function via ~~corresponding~~ Lagrangian efficiently.~~Gradient vanishing leads the layers closest to the outcome to dominate the parameter updating.~~ However, the vanishing Lagrange multiplier leads to the gradient decay and the domination by closer layers to the outcome. The presenter introduces a co-adjoint method with penalized objective function to ~~address~~ handle 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 multipliers ~~λ_t , $\phi(\lambda_t)$,~~ and the variance between adjoint size. Simulations show the better accuracy of ~~this~~ co-adjoint method over LSTM and typical Adjoint method.

~~Note that the g~~ Gradient vanishing problem is a numerical issue. ~~because e~~ Every middle term ~~x_t~~ is supposed to contain~~s~~ all the former information ~~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, n~~ No information will ~~be lost~~ lose, if we have a super process to estimate the network parameters . Besides, we should investigate more on penalty sensitivity. ~~the sensitivity of the penalty $\phi(\lambda_t)$ should be investigated more.~~

Clean:

This talk combats the vanishing gradient problem in backpropagation RNN using Adjoint method. The Adjoint method computes the gradient of objective function via Lagrangian efficiently. However, the vanishing Lagrange multiplier leads to the gradient decay and the domination by closer layers to the outcome. The presenter introduces a co-adjoint method with penalized objective function to handle the vanishing. The penalty term contains the penalty for small multipliers and the variance between adjoint size. Simulations show the better accuracy of co-adjoint method over LSTM and typical Adjoint method.

Gradient vanishing problem is a numerical issue. Every middle term is supposed to contain all the former information due to the RNN nature. No information will lose, if we have a super process to estimate the network parameters. Besides, we should investigate more on penalty sensitivity.

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.