Adv Big Data

Professor Gu

Assignment 7

Xueyan Bai

The Summary of Assignment 7

“*Limitations of the Empirical Fisher Approximation for Natural Gradient Descent*”

Assume a probability model form as: , where is an exponential family with natural parameters in F and is a prediction function parameterized by With N *iid* training samples, the article tried to minimize:

,

which covers common scenarios such as least-squares regression or C-class classification with cross-entropy loss. Defining the Fisher information matrix of the model as a preconditioner,

.

There exists a certain approximation of the Fisher, -- empirical Fisher (EF), which defined as:

The main purpose of this article is to argue that the empirical Fisher does not generally capture second-order information as the Fisher does. The article tried to offer a critical discussion of the EF based on three arguments.

The first argument is that the empirical Fisher follows the formal definition of a generalized Gauss-Newton (GNN) matrix, which defined as: a split with convex , leads to a GNN matrix of , defined as:

,

If the split .

Using a logistic regression problem example, the results indicates that the EF is a good approximation of the Fisher at the minimum only if the assumption is fulfilled.

The second argument is that the EF converges to the true Fisher when the model is a good fit for the data. Based on the example, the author proved that the relationship between the EF and the Fisher exists at a minimum under hardly achieved conditions such as the model is correct and contains enough data relative to model capacity.

Finally, the article used the example to prove the opinion that the EF is a questionable preconditioner because the EF can lead to update directions that are opposite to the natural gradient. The magnitude of the steps may lead to poor performance even with the correct direction.

At the end of the article, the author pointed out the concept of variance adaptation and indicates how the EF could contain useful information to adapt to the gradient noise in stochastic optimization.