Sketchy Empirical Natural Gradient Methods for Deep Learning

Minghan Yang¹, Dong Xu², Zaiwen Wen¹, Mengyun Chen³ and Pengxiang Xu⁴

1) Beijing International Center for Mathematical Research, Peking University, China
2) School of Mathematical Sciences, Peking University, China
3) Huawei Technologies Co. Ltd, China
4) Peng Cheng Laboratory, China

Corresponding Author: Zaiwen Wen, wenzw@pku.edu.cn

ABSTRACT

In this paper, we develop an efficient sketchy empirical natural gradient method (SENG) for large-scale deep learning problems. The empirical Fisher information matrix is usually low-rank since the sampling is only practical on a small amount of data at each iteration. Although the corresponding natural gradient direction lies in a small subspace, both the computational cost and memory requirement are still not tractable due to the high dimensionality. We design randomized techniques for different neural network structures to resolve these challenges. For layers with a reasonable dimension, sketching can be performed on a regularized least squares subproblem. Otherwise, since the gradient is a vectorization of the product between two matrices, we apply sketching on the low-rank approximations of these matrices to compute the most expensive parts. A distributed version of SENG is also developed for extremely large-scale applications. Global convergence to stationary points is established under some mild assumptions and a fast linear convergence is analyzed under the neural tangent kernel (NTK) case. Extensive experiments on convolutional neural networks show the competitiveness of SENG compared with the state-of-the-art methods. On the task ResNet50 with ImageNet-1k, SENG achieves 75.9% Top-1 testing accuracy within 41 epochs. Experiments on the distributed large-batch training show that the scaling efficiency is quite reasonable.