Baidu, Beijing

Belhal Karimi

I've started working in the Cognitive Computing Lab (CCL), led by Professor Ping Li, in Baidu Beijing on February 17th 2020. In this document I am listing the contributions I've made within the team during my time in this team. The entirety of my research is deferred to the research statement attached to this document. Besides, the references cited in this document, along with the names of my collaborators (past or present members of the CCL) are given in the last page of this short note. My research done at CCL has been published in top-tier conferences in machine learning such as NeurIPS (2020) and is under review for the most, constantly receiving peer reviews for it. The work I have been able to execute with the team can be decomposed into two parts: (a) research towards the training and generalization of deep neural networks and (b) algorithms at the junction of sampling and nonconvex optimization.

(a) Deep Learning: Training and Generalization

Speeding training and making sure the output parameter estimates lead to models generalizing well on unseen data are the two main challenges I have been tackling in Baidu

Training Acceleration. Dealing with the speed of convergence of a given training algorithm is a classical problem in modern machine learning.
From a practical perspective, we propose a variant of the known AMSGrad algorithm, a popular adaptive gradient method, in order to facilitate its acceleration. In [9], we add prior knowledge about the sequence of consecutive mini-batch gradients and leverages its underlying structure making the gradients sequentially predictable. By exploiting the predictability and ideas from optimistic online learning, our proposed algorithm accelerates the convergence and increases sample efficiency.

Decentralized Training. Given the need for distributed training procedures, distributed optimization algorithms are at the center of attention. With the growth of computing power and the need for using machine learning models on mobile devices, the communication cost of distributed training algorithms needs careful consideration. In that regard, more and more attention is shifted from the traditional parameter server training paradigm to the decentralized one, which usually requires lower communication costs. We develop, in [1], a general algorithmic framework that can convert existing adaptive gradient methods to their decentralized counterparts and thoroughly analyze the convergence behavior of the proposed algorithmic framework showing that if a given adaptive gradient method converges, under some specific conditions, then its decentralized counterpart is also convergent. Apart from the focus on communication complexity, the privacy of the data stored on the devices on which distributed learning occurs is also critical. In [2], we derive FedSKETCH, a method based on the compression of the accumulation of local gradients using count sketch. Due to the lower dimension of sketching used, our method exhibits communication-efficiency property. We also deal with the case where the data is heterogeneous across devices, which is commonly faced in federated learning, by developing FedSKETCHGATE. In particular, we establish a communication complexity of order $\mathcal{O}(\log(d))$ per round, where d is the dimension of the vector of parameters compared to $\mathcal{O}(d)$ complexity per round of baseline mini-batch SGD. Another focus on the federated learning setting is made in our work [6], where we develop a local variant of AMSGrad by using layerwise and dimensionwise adaptive learning rates. The main contribution of the paper lies in the embedding of the LARS method in the local AMSGrad method.

Towards Better Generalization. The final aspect of my work on training DNNs pertain to improving their generalization performances. Adaptive gradient methods have been optimizers of choice for deep learning due to their fast training speed, yet, their generalization performance is often worse than that of SGD for over-parameterized neural networks. To tackle this flaw, we propose in [10] Stable Adaptive Gradient Descent (SAGD) which leverages differential privacy to boost the generalization performance of adaptive gradient methods. Empirical runs on image classification or language modeling are backed with theoretical justifications to highlight the improved generalization properties of SAGD.

(b) When Sampling meets Optimization

Being able to *sample/infer* those latent variables is key during the *optimization* phase. I detail below different contributions I have participated in durign my time in Baidu, Beijing where sampling is occuring.

Fitting Latent Variable Models. From a modeling perspective, we propose in [3] a novel approach to embed flow-based models with hierarchical latent data structures. Integrating normalizing flows in variational graphs leads to a better recovery of the latent relational structures of high dimensional data. Moreover, a particularly interesting class of latent variable models is Bayesian Neural Networks (BNNs). BNNs attempt to combine the strong predictive performance of neural networks with formal quantification of uncertainty of the predicted output in the Bayesian framework. Yet, today, training those networks is slow and inefficient. Thus, we propose in [4], a simple averaging method in the space of the hyperparameters of the random weights leading to faster training and better empirical generalization.

Two-level Stochastic Optimization Methods. The EM algorithm, when used on highly non-convex models, is intractable. A natural solution is to alleviate the intractable expectations with Monte Carlo (MC) approximations. In [7], we analyze those variants when two levels of stochasticity are involved. The first one being the MC approximation and the second one the index sampling for stochastic updates. This work was followed by our Two-Timescale scheme in [5], where Robbins-Monro type of update is combined with stochastic variance reduction. Thus, two dynamics are progressing iteratively, one being driven by the stochastic approximation stepsize (slow) and the other one driven by the variance reduction stepsize (fast). Our framework displays better convergence performances for various applications from fitting pharmacological models to training deformable template for image analysis.

MCMC Based Optimization. When MC approximation is involved, as stated above, sampling from the posterior distribution is not always direct. For complex models, we have recourse to sampling techniques such as VI or MCMC.
In Energy Based Models (EBMs), this sampling procedure, aiming at drawing samples from the potential of the EBM, is crucial for the ultimate task of training a generative model. In [8], we improve current state-of-the-art samplers for EBMs by introducing an anisotropic stepsize in our Langevin updates. The drift term of the Langevin diffusion is not only depending on the dimension of the posterior landscape, but the covariance of the Brownian motion is also gradient informed. Making the proposal empirically efficient to explore a larger space of the posterior distribution and thus avoiding in practice mode collapse.

References

[1] Xiangyi Chen, **B. Karimi**, Weijie Zhao, and Ping Li. Convergent adaptive gradient methods in decentralized optimization. *Submitted*, 2020.

[2] Farzin Haddadpour, **B. Karimi**, Ping Li, and Xiaoyun Li. FedSKETCH: Communication-efficient federated learning via sketching. *Submitted*, 2020.

[3] Shaogang Ren, Yang Zhao, **B. Karimi**, and Ping Li. VFG: Variational flow graphical model with hierarchical latent structure. *Submitted*, 2020.

[4] **B. Karimi** and Ping Li. HWA: Hyperparameters weight averaging bayesian neural networks. *Submitted*, 2020.

[5] **B. Karimi** and Ping Li. Two timescale stochastic em algorithms. *Submitted*, 2020.

[6] **B. Karimi**, Xiaoyun Li, and Ping Li. Layerwise and dimensionwise adaptive local ams method for federated learning. *Work in progress*, 2020.

[7] **B. Karimi**, Hoi-To Wai, Eric Moulines, and Ping Li. MISSO: Minimization by incremental stochastic surrogate optimization for large scale nonconvex problems. *Submitted*, 2020.

[8] **B. Karimi**, Jianwen Xie, and Ping Li. Anila: Anisotropic langevin dynamics for training energy-based models. *Work in progress*, 2020.

[9] Jun-Kun Wang, Xiaoyun Li, **B. Karimi**, and Ping Li. An optimistic acceleration of amsgrad for nonconvex optimization. *Submitted*, 2020.

[10] Yingxue Zhou, **B. Karimi**, Jinxing Yu, Zhiqiang Xu, and Ping Li. Towards better generalization of adaptive gradient methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–10, 2020.