Research Statement

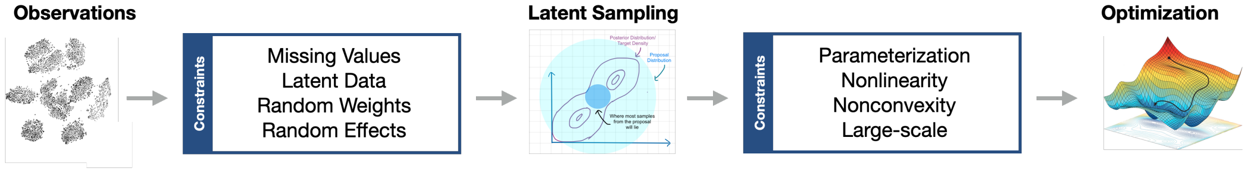
Belhal Karimi
Baidu Research

Throughout my research, I focus on developing *training*, also known as *optimization*, methods for large-scale datasets. There are several specificities to my work.

The broad panel of my work has application on various problems, datasets and domains. To name a few, such learning task as stated above is crucial while fitting complex nonlinear models (mixed models, deep neural networks, mixture models) on tabular, image, textual data to tackle problems encountered in computer vision, drug development or natural language processing.

Based on the principled approach that consists in *observing* the world, *designing* a model describing the best those observations and *training* it on the latter, my main focal point in the realm of machine learning resides in the *training*, or *learning*, phase. With the sheer size of data and the high nonconvexity of the modern models, such as mutlilayer nerual network, used to describe complex human tasks, there is a rising interest and need for scalable, faster learning methods and their rigorous theoretical understanding.

Up to some observations, either fixed or streaming, and a well designed model, the definition of a loss/cost function and its optimization (minimization) are at the heart of this training phase. Continuously improving those optimization algorithms is key for *machine learning* in order to sustain the rapid growth in dimension, compositionality of the models and the high variety of input observations (sound, image, LIDAR, etc . . .).



While my work provides *novel* methods for particularly deep neural networks (DNNs), one special case of the setting above, is when the input-output relationship of a phenomena is not completely characterized by the observations. A set of latent variables is thus needed and the loss function accepts the latter as a third argument.

Illustrative example of latent data model: During clinical trials, the kinetics and dynamics of a drug being tested are modeled using nonlinear functions (or systems of ordinary differential equations) and observations from patients which comprise for instance their gender, height, concentration of the drug after injection. While those observed covariates are necessary, they are not sufficient to describe well the biological pheonomena. A set of latent variables are used to quantify what can not be measured. In the special case of pharmacology, those latent variables describe the inter-individual variability among patients of a population (this is what makes us all different other than measurable signals). Therefore, the loss function, here the likelihood, is completed by simulations of those random effects and are then used to complete the observations before final optimization.

Thus, part of my research is at *the intersection* of **sampling** and **optimization**, bridging the gap between sampling methods such as Markov Chain Monte Carlo (MCMC) or Variational Inference and optimization method such as gradient based learning algorithms or maximum likelihood estimation.

My research has been published in top-tier conferences in machine learning such as NeurIPS, COLT, ICML and made the object of contribution in statistics Journal such as CSDA. I also received a collection of awards from those conferences and a Jacques Hadamard grant for a summer visit the Russian leading group in Bayesian Deep Learning called *BayesGroup*.

Some of my work are now implemented in the commercial modeling and simulation software for drug development *Lixoft* and in its open-source counter part *saemix*.

(a) Deep Learning: Training and Generalization

A particular interest of mine lies in the practical training and theoretical understanding of deep neural networks, widely used for most learning tasks in the past decade. Scaling, speeding and improving existing training algorithms is of utmost importance and drive most of my existing publications. Recently, I have been also interested in the generalization properties of such training algorithms. Speeding training and making sure the output parameter estimates lead to models generalizing well on unseen data are the two main challenges I am tackling today.

Training Acceleration

Dealing with the speed of convergence of a given training algorithm is a classical problem in modern machine learning. From a theoretical perspective, we define the convergence of an algorithm when this latter reaches a so-called ϵ-stationary point. In deep learning, and more generally in stochastic nonconvex optimization, the chosen suboptimality condition is the second order moment of the gradient of the objective function. Then, deriving the algorithm convergence rate simply consists in finding the number of iterations until that quantity is bounded by ϵ. In [11], we establish that the classical Stochastic Gradient Descent (SGD) algorithm reaches an ϵ-stationary point in $\mathcal{O}\left(c_0 + \log(n)/\sqrt{n}\right)$ iterations. The results also hold when the stochastic gradient is biased, i.e., its expectation is not equal to the full gradient. This setting has not been studied before our contribution and yet is presented in numerous applications such as the online EM algorithm or the policy-gradient method for average reward maximization in reinforcement 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 [15], 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 accelerate the convergence and increase sample efficiency. In [11], we derive a unifying framework for the incremental optimization methods. Among others, our framework include stochastic variational inference and MISO.

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 device, 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 [9], 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 Generalizaiton

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 [16] 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

Mostly driven by the potential applications and as stated in the beginning of this statement, the models I am considering in my work are comprised of some latent variables. Indeed either in medical applications, where latent variables may be missing values uninformed by the patients or random effects in the special case of pharmacology, or in computer vision applications, and more specifically generative modeling, where layers of latent variables are used to disentangle a better representation of the input data, being able to *sample/infer* those latent variables is key during the *optimization* phase. I detail below different contributions where this setting is respected.

Hierarchical Latent Structure Based Models

[12] [3] [7]
Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Aenean nonummy turpis id odio. Integer euismod imperdiet turpis. Ut nec leo nec diam imperdiet lacinia. Etiam eget lacus eget mi ultricies posuere. In placerat tristique tortor. Sed porta vestibulum metus. Nulla iaculis sollicitudin pede. Fusce luctus tellus in dolor. Curabitur auctor velit a sem. Morbi sapien. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Donec adipiscing urna vehicula nunc. Sed ornare leo in leo. In rhoncus leo ut dui. Aenean dolor quam, volutpat nec, fringilla id, consectetuer vel, pede.

Two-level Stochastic Optimization Methods

[6] [13] [8]
Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Aenean nonummy turpis id odio. Integer euismod imperdiet turpis. Ut nec leo nec diam imperdiet lacinia. Etiam eget lacus eget mi ultricies posuere. In placerat tristique tortor. Sed porta vestibulum metus. Nulla iaculis sollicitudin pede. Fusce luctus tellus in dolor. Curabitur auctor velit a sem. Morbi sapien. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Donec adipiscing urna vehicula nunc. Sed ornare leo in leo. In rhoncus leo ut dui. Aenean dolor quam, volutpat nec, fringilla id, consectetuer vel, pede.

MCMC Based Optimization

We propose in [4] an efficient MCMC procedure, namely NLME-IMH, for posterior sampling in nonlinear mixed effects models, based on the Laplace approximation. This work was followed by [5] where we embed NLME-IMH into a stochastic variant of the EM algorithm (SAEM) for maximum likelihood estimation.
[13] [14]

Future Research Directions

dajndaundadzandzanidazni

Energy Based Models. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Aenean nonummy turpis id odio. Integer euismod imperdiet turpis. Ut nec leo nec diam imperdiet lacinia. Etiam eget lacus eget mi ultricies posuere. In placerat tristique tortor. Sed porta vestibulum metus. Nulla iaculis sollicitudin pede. Fusce luctus tellus in dolor. Curabitur auctor velit a sem. Morbi sapien. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Donec adipiscing urna vehicula nunc. Sed ornare leo in leo. In rhoncus leo ut dui. Aenean dolor quam, volutpat nec, fringilla id, consectetuer vel, pede.

Federated Learning. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Aenean nonummy turpis id odio. Integer euismod imperdiet turpis. Ut nec leo nec diam imperdiet lacinia. Etiam eget lacus eget mi ultricies posuere. In placerat tristique tortor. Sed porta vestibulum metus. Nulla iaculis sollicitudin pede. Fusce luctus tellus in dolor. Curabitur auctor velit a sem. Morbi sapien. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Donec adipiscing urna vehicula nunc. Sed ornare leo in leo. In rhoncus leo ut dui. Aenean dolor quam, volutpat nec, fringilla id, consectetuer vel, pede.

Bayesian Deep Learning. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Aenean nonummy turpis id odio. Integer euismod imperdiet turpis. Ut nec leo nec diam imperdiet lacinia. Etiam eget lacus eget mi ultricies posuere. In placerat tristique tortor. Sed porta vestibulum metus. Nulla iaculis sollicitudin pede. Fusce luctus tellus in dolor. Curabitur auctor velit a sem. Morbi sapien. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Donec adipiscing urna vehicula nunc. Sed ornare leo in leo. In rhoncus leo ut dui. Aenean dolor quam, volutpat nec, fringilla id, consectetuer vel, pede.

Stochastic Optimization for DNNs. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Aenean nonummy turpis id odio. Integer euismod imperdiet turpis. Ut nec leo nec diam imperdiet lacinia. Etiam eget lacus eget mi ultricies posuere. In placerat tristique tortor. Sed porta vestibulum metus. Nulla iaculis sollicitudin pede. Fusce luctus tellus in dolor. Curabitur auctor velit a sem. Morbi sapien. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Donec adipiscing urna vehicula nunc. Sed ornare leo in leo. In rhoncus leo ut dui. Aenean dolor quam, volutpat nec, fringilla id, consectetuer vel, pede.

References

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

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

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

[4] **Belhal Karimi** and Marc Lavielle. Efficient Metropolis-Hastings sampling for nonlinear mixed effects models. *Proceedings of BAYSM 2018*, 2018.

[5] **Belhal Karimi**, Marc Lavielle, and Eric Moulines. f-SAEM: A fast stochastic approximation of the EM algorithm for nonlinear mixed effects models. *Computational Statistics and Data Analysis, (CSDA)*, 2018.

[6] **Belhal Karimi**, Marc Lavielle, and Éric Moulines. On the convergence properties of the mini-batch em and mcem algorithms. *HAL preprint hal: 02334485*, 2019.

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

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

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

[10] **Belhal Karimi**, Blazej Miasojedow, Eric Moulines, and Hoi-To Wai. Non-asymptotic analysis of biased stochastic approximation scheme. In *Proceedings of the Thirty-Second Conference on Learning Theory (COLT) 2019*. PMLR, 2019.

[11] **Belhal Karimi**, Hoi-To Wai, and Eric Moulines. A doubly stochastic surrogate optimization scheme for non-convex finite-sum problems. *1st Symposium on Advances in Approximate Bayesian Inference (AABI)*, 2019.

[12] **Belhal Karimi**, Hoi-To Wai, Eric Moulines, and Marc Lavielle. On the global convergence of (fast) incremental expectation maximization methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2837–2847, 2019.

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

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

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

[16] Yingxue Zhou, **Belhal 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.