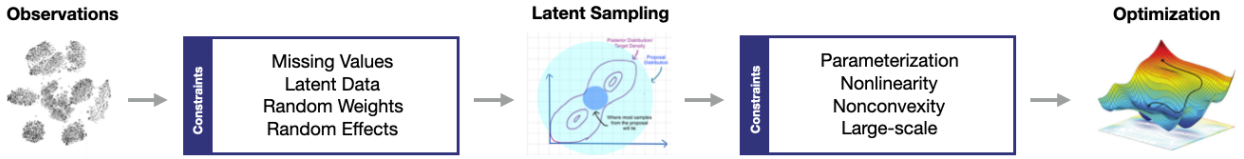
Research Statement

Belhal Karimi
Baidu Research USA

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 applications 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 of *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, the concentration of the drug after injection. While those observed covariates are necessary, they are not sufficient to describe well the biological phenomena. 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, BAYSM, 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 to the Russian leading group in Bayesian Deep Learning called *BayesGroup*.

(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 of 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 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 [16], 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. In [12], we derive a unifying framework for incremental optimization methods. Among others, our framework includes 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 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 [10], 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 [17] 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 at 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.

Fitting Latent Variable Models. The EM algorithm is one of the most popular algorithm for maximum likelihood estimation in latent data models. We propose in [13], a stochastic EM framework for exponential models encompassing incremental and several variance reduced variants. Our global and non-asymptotic bounds make the case for leveraging variance reduction techniques, borrowed from the optimization literature, to accelerate drastically the convergence to a stationary point taking $\mathcal{O}(n^{2/3}/\epsilon)$ for the latter versus $\mathcal{O}(n/\epsilon)$ for the former, with n being the number of data samples. 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 [8], 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 [6] and [14], 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.
These works were followed by our Two-Timescale scheme in [9], 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. We propose in [5, 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 [7] where we embed NLME-IMH into a stochastic variant of the EM algorithm (SAEM) for maximum likelihood estimation.
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 [15], 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.

(c) Distributed Optimization

Distributed Optimization Acceleration.

Deep Federated Learning.

(d) Generative Modeling

Improve the Sampling, improve the EBM.

EBM and its Applications.

Future Research Directions

I would like to develop here the main axis of research I am planning on conducting in the next years, driven by the *applications*, such as medical and vision, the power of *Bayesian methods*, tools like *MCMC* for posterior sampling and the leverage of a latent data structure considered as an additional layer on top of the observations. I plan on leveraging all those latter techniques on the following tasks: training complex models under the federated settings when unobserved covariates (missing values, random effects, latent labels) are at stake, designing efficient sampling-based (MCMC or VI) optimization algorithms for training energy based models or bayesian neural networks, improving the understanding of generalization in over parametrized neural networks.

Energy Based Models. EBMs are promising generative models particularly interesting to use in computer vision tasks. They are exactly at the crossroads of the sampling and the optimization domains. Developing efficient and clever MCMC proposals is key to greater exploration of the conditional distribution one needs to sample from and would lead to better mixing of the chain. Especially when short-run MCMC is involved, the hyperparameters of the proposal distribution are crucial. Mode collapse is one of the challenges encountered and given the multimodal and nonconvex nature of the target distribution, several MCMC techniques, like the one we develop in [15], are worth considering.

Federated Learning. Federated Learning has been a rapidly growing domain over the past years for its ability to handle data on million of devices while preserving their privacy. Our ability to come up with new communication efficient and private methods is utterly important in order to fit complex models, like DNNs, in order to describe human tasks in the decentralized settings. Before such progress is made, we will remain limited in the number of devices, data samples, and the nature of the tasks we can handle. [2, 10] were developed in that context. One other interesting direction, yet unexplored by the community, that I wish to contribute to is the study and adaptation of EM-like algorithms in the federated (decentralized and private) setting. More generally, latent variable models are not well studied under that setting and I believe that whenever latent inference is involved, a particular sampling procedure is needed in order to work under decentralized and private constraints. The natural and important application is the fitting on nonconvex models on hospital data where many data points are missing (thus considered as latent).

Bayesian Deep Learning. In several sensible domains, such as medical or autonomous driving, being able to derive safety guarantees that *take account of the uncertainty within the model and the input data* is mandatory before putting trained models in production. Bayesian deep learning, which assumes a prior over the weights, is capable of producing such uncertainty measures, yet their training and accuracy even on benchmark datasets are not satisfactory. In short, the current main issue I see is robustness, i.e., it should be as easy or easier to get to work than normal deep learning models, but it is not at the moment, and efficiency, i.e., it should at most be 2-3x as slow as normal deep learning training, but currently, it is slower. Developing optimization methods to tackle those challenges is important and will be one of my focus for the years to come. [8] and [14] are attempts to give solutions to those issues.

Generalization of Deep Neural Networks. While deep neural networks are working very well in practice, we have very little understanding of how they work. Except working on improving the training and the generalization properties of those surprisingly well-performing over parametrized neural networks (as in our work [17]), focus needs to be put on explaining why they work so well on

unseen data while their architectures contain more parameters than the number of training data points. One possible direction towards that goal is to extensively understand the loss landscape that one is trying to minimize. Such knowledge is important to understand how the trajectory of the parameter estimates through the epochs, and hence *a fortiori* the vector of fitted parameters, affect the generalization of our trained model. Some curvature, gradient and Hessian, based metrics are initial attempts at addressing those questions. For instance, we are developing a convergence diagnostic tool for nonconvex loss function (in particular DNNs) leveraging theoretical and empirical discoveries about stationarity metrics combined with sharpness and wideness of encountered local minima during the training procedure. Hopes are that combining saturation (widely used for convex functions) with generalization metrics, such that the curvature of the loss and the variance of the gradients, will lead to a convergence-based stepsize schedule favorable to better generalization.

Concluding Remarks

I am confident that my research and collaborations with experts in various fields, including machine learning, statistics, mathematics, and pharmacology have equipped me with the necessary background to approach the above challenging and impactful research directions. I believe the plan laid out above is coherent with my background, my skills, and most importantly my personal interests in contributing to the field. While those domains may seem highly diverse, I plan on contributing to each by providing novel *optimization* and *sampling* tools, as the core of my expertise.

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 Marc Lavielle. Efficient Metropolis-Hastings sampling for nonlinear mixed effects models. *Proceedings of BAYSM 2018*, 2018.

[5] **B. Karimi**, Marc Lavielle, and Éric Moulines. Bridging the gap between independent metropolis hastings and variational inference. In *Implicit Models Workshop (ICML)*, 2017.

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

[7] **B. 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)*, 2020.

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

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

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

[11] **B. 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.

[12] **B. Karimi**, Hoi-To Wai, and Eric Moulines. A doubly stochastic surrogate optimization scheme for non-convex finite-sum problems. *Adv. in Approx. Bayes. Inference (AABI)*, 2019.

[13] **B. 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.

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

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

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

[17] 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.