Dmitrii Avdiukhin

Research Statement

Overview

My research interests lie primarily in the areas of continuous optimization, machine learning, and algorithms. A more detailed overview is provided below.

- **First-order methods in continuous optimization**. Gradient descent (GD) is a fundamental algorithm in continuous optimization and is crucial in modern machine learning. Its basic properties are well understood: it's known that GD and its variations converge to first-order stationary points (i.e. points with small gradient), and its perturbed variations converge to second-order stationary points.

  However, some properties of GD, especially in non-convex settings, require more research. For example, there is still a large gap between the practical behavior of GD and our theoretical understanding of GD. Second-order convergence is not proven in important constrained and distributed settings.

- **Federated Learning**. In modern machine learning, due to a large amount of data and required computational resources, models are often trained on multiple devices, with the server aggregating and distributing model updates. In Federated Learning (FL), model updates are often computed on the end users' devices, which can compute and communicate their updates asynchronously.

  Due to a large number of devices, communication becomes the main bottleneck. Common approaches include compressed communication and delayed communication, which complicate the analysis of FL algorithms. The situation is further complicated by asynchronous computation/communication and the fact that the data on different devices belongs to different distributions. A unified framework for these algorithms is yet to be established, and in many settings, there is a gap between known lower and upper bounds on convergence rates.
- **Generalization.** In machine learning, an important question is how different is the empirical loss from the population loss. While for convex settings, good generalization bounds are known, the bounds for non-convex, distributed, and many other settings can probably be improved.

- **Unsupervised learning**. A common unsupervised learning problem is Hierarchical Clustering (HC), where the goal is to find a good tree-like representation of the data. A line of research originating from Dasgupta [2016] shows that, when either similarities or dissimilarities between all pairs of elements are known, HC can be treated as an optimization problem, for which popular practical algorithms provide provable approximation bounds.

  However, the existing results are non-exhaustive: certain HC objectives are barely studied, and for others, there is a large gap between known lower and upper bounds on the approximation factors. Moreover, the approximation factors of popular algorithms differ greatly in theory and practice.

  A different line of research considers recovering hierarchical trees satisfying certain structural constraints (e.g., given 3 elements, which pair of the elements is most similar). There are many open problems in this direction.

When optimizing HC objectives, the need often arises to solve certain constraint satisfaction problems. While such problems are in general NP-hard, in certain cases, it's possible to solve them either exactly or approximately.

- **Other research areas** include submodular optimization, graph partitioning, and graph convolutional networks. Before my PhD, I worked on model checking and syntax analysis.

Previous Research

Constrained Optimization

In Avdiukhin et al. [2019] we presented a solution for the graph partitioning problem with multiple balance constraints. The idea is to solve a quadratic optimization problem under linear inequality constraints using projected gradient descent. Since then, I was interested in **convergence of gradient descent**, in particular **in constrained settings**. It is known that finding a local minimum of a non-convex function is NP-hard, and hence the common goal is to find an approximate first- or second-order stationary point (FOSP and SOSP respectively). In the unconstrained setting, it's possible to find a FOSP or a SOSP efficiently Jin et al. [2021]. On the other hand, in constrained settings, while it's possible to a FOSP efficiently, finding a SOSP is NP-hard, even for linear inequality constraints [Murty and Kabadi, 1987]. In Avdiukhin and Yaroslavtsev [2022], we show polynomial-time convergence to a SOSP when the number of constraints is logarithmic.

Distributed and Federated Learning

In Avdiukhin and Kasiviswanathan [2021], we consider federated learning settings where we allow **arbitrary communication between clients and the server**, with the only (unavoidable) assumption that the number of rounds when a client doesn't communicate with the server is bounded. This was in contrast with most of the research which considered only specific communication patterns, such as full or random client participation at certain iterations.

In Avdiukhin and Yaroslavtsev [2021], we analyze second-order convergence of compressed stochastic gradient descent: at every iteration, each client computes local stochastic gradient; the gradients are compressed, aggregated on the server, and the averaged gradient is broadcasted to the clients. In this setting, our paper was the first result showing **convergence of compressed SGD to a SOSP**, in contrast to most of the work which only considered convergence to a FOSP. We show that for a certain choice of gradient compression, the required communication is provably better compared with full communication.

Hierarchical Clustering and Approximation Algorithms

Background Dasgupta [2016] introduces the following hierarchical clustering (HC) objective: given a hierarchical tree T on a set of elements as well as similarities w_{ij} between each pair of elements i and j, the goal is to minimize $\sum_{i,j} w_{ij}|\,\mathrm{LCA}_T(i,j)|$, where $|\,\mathrm{LCA}_T(i,j)|$ is the number of elements under the least common ancestor of i and j in tree T. The objective can be approximated within $O(\sqrt{\log n})$ factor [Charikar and Chatziafratis, 2017]. Moseley and Wang [2017] introduce the complementary objective $\sum_{i,j} w_{ij}(n - |\,\mathrm{LCA}_T(i,j)|)$, where n is the number of elements, and Alon et al. [2020] show 0.585 approximation.

Previous Research Similar objectives can be defined for the case when dissimilarities (e.g. distances) d_{ij} between the elements are known. For the **maximization dissimilarity-based objective $\sum_{i,j} d_{ij}|\,\mathrm{LCA}_T(i,j)|$, we prove 0.74 approximation** [Naumov et al., 2021]. We are working on improving the approximation factor; while with incremental modifications of Naumov et al. [2021] we can achieve 0.77 approximation factor, we are looking for a more substantial improvement.

We are currently working on approximating the **minimization objective** $\sum_{i,j} d_{ij}(n - |\mathrm{LCA}_T(i,j)|)$. For in the ℓ_1-metric case, Rajagopalan et al. [2021] show 2-approximation, and we believe that the result can be generalized to $O(\log n)$ approximation for arbitrary metric. However, we are also interested in the non-metric case, which is substantially harder.

Another research direction involves construction of a hierarchical tree satisfying known structural constraints, e.g. of form "in the tree, a should be separated first from b and c". In Avdiukhin et al. [2023], given the set of structural constraints, we construct a hierarchical tree, based on which we can predict the unseen structural constraints with the required error rate.

Current and Future Research

- **Generalization.** Hardt et al. [2016] analyze the generalization of models trained with SGD based on stability bounds, showing dimension-independent and hence substantially improved generalization bounds. Their results for convex settings are further refined in Lei and Ying [2020]. Unfortunately, currently known generalization bounds for non-convex objectives are much worse compared with convex ones, and we are working on improving them. Then, I plan to work on generalization in various settings, including federated learning.
- **Federated Learning.** I plan to work on first- and second-order convergence of first-order methods in general FL settings, including compressed communication, arbitrarily asynchronous computation and communication, and potentially arbitrary communication topologies. I'm especially interested in lower bounds on communication and computation, as well as the weakest assumptions necessary for convergence. I'm also interested in studying FL under realistic assumptions which allow faster convergence and less communications.
- **First-order methods.** To generalize the above, I'm broadly interested in working on properties of gradient descent and its variations in various settings.
- **Theoretical foundations of machine learning.** I'm interested in the theoretical analysis of commonly used machine learning models. In particular, I plan to work on the analysis of their loss landscape (e.g. understanding its smoothness and behavior of its local minima) and their generalization on the unseen data.
- **Approximation algorithms.** I'm interested in approximation algorithms in general, and, in particular, in approximation algorithms for HC objectives. Since approximating such objectives often requires finding approximate solutions to certain constraint satisfaction problem, I'm also particularly interested in their approximation algorithms, as well as their applications.

References

N. Alon, Y. Azar, and D. Vainstein. Hierarchical clustering: A 0.585 revenue approximation. In *Conference on Learning Theory*, pages 153–162. PMLR, 2020.

D. Avdiukhin and S. Kasiviswanathan. Federated learning under arbitrary communication patterns. In *International Conference on Machine Learning*, pages 425–435. PMLR, 2021.

D. Avdiukhin and G. Yaroslavtsev. Escaping saddle points with compressed sgd. *Advances in Neural Information Processing Systems*, 34:10273–10284, 2021.

D. Avdiukhin and G. Yaroslavtsev. Houdini: Escaping from moderately constrained saddles. *arXiv preprint arXiv:2205.13753*, 2022.

D. Avdiukhin, S. Pupyrev, and G. Yaroslavtsev. Multi-dimensional balanced graph partitioning via projected gradient descent. In *Proc. VLDB Endow.*, volume 12, pages 906–919, 2019.

D. Avdiukhin, G. Yaroslavtsev, D. Vainstein, O. Fischer, S. Das, and F. Mirza. Tree learning: Optimal sample complexity and algorithms. In *AAAI*, 2023.

M. Charikar and V. Chatziafratis. Approximate hierarchical clustering via sparsest cut and spreading metrics. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 841–854. SIAM, 2017.

S. Dasgupta. A cost function for similarity-based hierarchical clustering. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 118–127, 2016.

M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68 (2):1–29, 2021.

Y. Lei and Y. Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020.

B. Moseley and J. Wang. Approximation bounds for hierarchical clustering: Average linkage, bisecting k-means, and local search. *Advances in neural information processing systems*, 30, 2017.

K. G. Murty and S. N. Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39:117–129, 1987.

S. Naumov, G. Yaroslavtsev, and D. Avdiukhin. Objective-based hierarchical clustering of deep embedding vectors. In *AAAI*, pages 9055–9063, 2021.

A. Rajagopalan, F. Vitale, D. Vainstein, G. Citovsky, C. M. Procopiuc, and C. Gentile. Hierarchical clustering of data streams: scalable algorithms and approximation guarantees. In *International conference on machine learning*, pages 8799–8809. PMLR, 2021.