Research Statement

Lijun Ding, Cornell Univeristy

October 19, 2020

My work lies at the intersection of optimization, statistics, and machine learning, where I works on solving large-scale and high dimensional optimization problems. My research has been broadly motivated by the technology trend of "high dimensional but structured data": instead of having a few accurate data, we now face a huge amount of data that are often indirect, incomplete, have more variables and features than the number of samples, but with special structures, e.g., sparsity and low-rankness. This trend has resulted in large-scale optimization problems for decision making, statistical inference and predictions. However, these problems are traditionally considered intractable to solve due to (i) the high computational cost in handling the data and the optimization variables, and (ii) the potential ill-conditioning and NP hardness coming from nonconvexity due to modeling and from noisy/corrupted data generating processes.

My work resolve the above issues by (i) justifying theoretically the good conditioning, the implicit regularization, and the benign geometry for these large scale optimization problems under mild and standard statistical assumptions, and (ii) developing robust, computational efficient and theoretically sound algorithms that can learn and utilize these special structure to solve these large-scale problems. To achieve both items, my research brings together fundamental ideas and sophisticated tools from a broad spectrum of areas including high-dimensional statistics, convex and non-convex optimization, and statistical learning theory. This work could have far-reaching impacts on both practitioners and theoreticians: (i) to practitioners, this new set of optimization algorithms allows for faster and yet robust decision making and predictions, (ii) to theoreticians, the lens of statistics and probability brings new insights to the fundamental researches of optimization in their application to large scale problems.

In what follows, I shall first present my prior work on solving large-scale optimization problems, including classical convex optimization problems such as semidefinite programming, and newly arising nonconvex problems, based on the idea and techniques such as Frank-Wolfe, strict complementarity, and the leave-one-out argument. Then I present my plans for several important further directions including (i) bridging convex and nonconvex optimization algorithmically and theoretically ; (ii) efficient sampling and optimization in reinforcement learning.

1 Prior work

1.1 Semidefintie programming: regularity, conditioning, and efficient algorithm

Semidefinite programming form a class of convex optimization problems with remarkable modeling power. Its application ranges from combinatorical problems such as max-cut, optics problems such as phase retrieval, graph network problem such as stochastic block model, to modern machine learning problem such as matrix completion and robust PCA. In all these applications, SDP has demonstrated the best theoretical guarantees once it is exactly solved. However, SDPs are challenging to solve because they involve a matrix variable $X \in \mathbb{R}^{n \times n}$ whose dimension n can rise into the millions or billions. Most existing algorithms suffer from both high storage, $\Omega(n^2)$, of storing the matrix variable, and the extraordinary high time complexity $\Omega(n^3)$. Such high complexity forbids their usage in modern large-scale problems. Even worse, even assuming we can run these algorithms, it is not clear that these problems are well-conditioned or enough regular so that the traditional guarantees on the objective value can be translated to guarantees on metrics such as distance to solution which are more concerned in SDP applications.

My research close the above gaps by (i) demonstrating that the traditionally considered regularity conditions do hold in these applications [DU20], (ii) design storage and time efficient algorithms for solving SDPs [DYC+19, DG20, DFU20], which converges merely under strong duality, and speed up to linear convergence if the further regularity conditions are satisfied. Hence they are well-suited for solving large scale SDPs.

More precisely, in the work [DU20], I demonstrate that the traditionally considered regularity conditions such as strict complementarity, non-degeneracy of SDPs are satisfied for $\mathbb{Z}_2$ synchronization, stochastic block model, and matrix completion. Hence, accuracy measured by objective value does translated to the distance metric as shown in my work [DU]. In the work [DYC$^+$19], I designed a new *storage optimal* (and potentially *the first* to my best knowledge) algorithm that provably solves SDP with sublinear convergence rate and only requires one eigenvector computation per iteration and solving one small-scale SDP with decision variable of size r, where r is the rank of the optimal solution which is expected to be much smaller than n in most of applications. In the followup works [DG20, DFU20], I further refined the algorithm, by exploration of the idea from classical spectral bundle method and Frank-Wolfe, so that it still keeps the sublinear rate yet achieve linear convergence rate once the regularity conditions are satisfied.

1.2 Statistical nonconvex optimization: analysis framework and efficient algorithm

Modern optimization problems are usually nonconvex, either due to the modeling or the nature of the underlying mechanism. One important approach to solving these problems is convex relaxation, and it has achieved great theoretical success in terms of finding the underlying signal in the signal processing regime or its generalization power in the statistics and machine learning regime, once the convex relation is solved exactly. However, it is observed that most algorithms for solving the relaxations are slow in general, that is, they usually only enjoy a sublinear rate of convergence in both theory and practice. Thus, people turn to directly solving nonconvex low-rank optimization (i.e., optimization problems with rank constraints) via simple gradient based algorithm. Despite the remarkable empirical success achieved by these simple algorithms, there is a huge gap between this and the theory side of nonconvex algorithms. These simple algorithms lack theoretical guarantees, whereas existing algorithms with guarantees are often highly artificial and too complicated to be implemented in practice.

To close the above gap, I have proposed a framework to analyze and explain the fast convergence of the simple gradient based algorithms. Applying this framework allowed me to solve an open problem posted back in 2009 [JMD10]. Moreover, inspired by the analysis of these simple algorithms, I have designed efficient, simple to implement and provably linearly/quadratically convergent algorithms. Let me explain my work on this line in more detail in the following two sections.

Stochastic iteration procedure analysis framework   Based on a novel technique, the "leave one out technique", Professor Chen and I [DC20] introduced a general and powerful framework for analyzing an iterative stochastic procedure arising in nonconvex algorithms. This powerful framework enables us to solve an open problem from 2009 [JMD10], by establishing the convergence of singular value projection (SVP), a simple gradient based algorithm for solving low-rank matrix completion. In fact we show that SVP actually obtains a linear convergence rate. The power of our framework is not limited to just simple gradient algorithms, it can also be applied to other stochastic iteration procedures. Moreover, we also show that the sample complexity for a convex relaxation of low-rank matrix completion is near optimal (which is the best result in the literature) via this framework.

Convex optimization based algorithms   Inspired by the analysis of gradient based algorithms for solving nonconvex problems, one of my works [CCD$^+$19] developed algorithms that in practice only need to solve a few (three to five) small scale convex optimization problems, where each of these problems is not much costlier than a couple of iterations of gradient based algorithms. Under a statistical setting, by utilizing tools from random matrix theory, variational analysis, and spectral theory, we show our algorithm is almost free of tuning and achieves quadratic convergence when initialized properly. Empirically, this method not only converges fast but also can be initialized randomly when sample size is reasonably large.

Averaged projected gradient method (AVPG)   Having established algorithms that are well-suited for quadratic or norm based objective with nonconvex constraints, I further advanced my research to non-quadratic objective where familiar concepts such as restricted isometry are no longer satisfied, and standard nonconvex approaches such as gradient descent may perform badly in practice and have no satisfactory theory in guaranteeing global and efficient convergence. I demonstrate that the critical component in dealing with non-quadratic objective is a regularity projection oracle, which enables global and linear convergence of our proposed algorithm, AVPG [DZC20]. Our results apply to a wide range of non-quadratic problems including rank aggregation, one bit matrix sensing/completion, and more broadly generalized linear models with rank constraint.

2 Future Research

2.1 Bridging convex and nonconvex optimization: algorithm and theory

Having explored both the convex and nonconvex worlds, I want to leverage the insights and algorithms from convex optimization and use them in solving nonconvex optimization problems.

Algorithmically, I would like to design a method which gets the best from both worlds even under the setting of limited observations or weak signals: (1) it generates an output which generalizes both theoretically and empirically (as some convex methods) (2) it is simple to implement and runs very fast (like some nonconvex methods).

To achieve this, I propose a hybrid approach based on convex relaxation and nonconvex methods, using approximate solutions from convex relaxations as the initialization for algorithms solving statistical nonconvex problems. The rationale behind this hybrid approach is two fold. First, in term of generalizability, there is hope that for various important settings, the approximate solution is just close enough to the true signal to be a good initialization point for nonconvex algorithms. Second, the proposed approach also enjoys low computation cost because in each phase, the initialization phase and the nonconvex algorithm phase, does not consume much computational budget. Indeed, in the initialization phase, though obtaining high accuracy in the optimal solution is usually not affordable, approximate solutions can be obtained with reasonable cost, e.g., an early stopping rule for iterative algorithms. In the second phase, many noncvonex algorithms are found to be very fast in practice and in theory when initialized appropriately.

Theoretically, I want to understand when effective nonconvex algorithms exist if we already know that the convex relaxation admits generalizable solutions. The rational behind this is that the theory for convex relaxation is well-developed, while analyzing nonconvex problems usually requires lengthy, technical analysis, and certain *ad-hoc* arguments. More importantly, this should help us in build intuitive and robust algorithms for previously unexplored nonconvex problems. I would like to further explore this direction based on my previous work, which has identified a few conditions and algorithm oracles that ensure the translation of knowledge of convex optimization to nonconvex optimization, such as the regular projection oracle [DZC20], and the sharpness and growhth condition [].

2.2 Efficient sampling and optimization in reinforcement learning

Tons of real-world problems are best modeled by reinforcement learning, where an agent has to sequentially makes decisions and receives rewards from the environment so that she could maximize the cumulative reward. The nonconvexity comes from the functional approximation to the value function or the parametrization of the policy space. The convergence of algorithms such as policy gradient and Q-learning and the landscape of the objective, the cumulative reward are complicated due to the functional approximation error and the intricate dependency along the trajectory of data. However, these heuristics has gained remarkable empirical success in games, robotics, and even traditional operations research area such as scheduling. My plan is to understand the intricate the statistical dependence and the geometry of the landscape via certain low rank structures of the Q-function, the transition function, or the reward function, and further design efficient sampling and optimization schemes.

References

[CCD+19] Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Díaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *arXiv preprint arXiv:1904.10020*, 2019.

[DC20] Lijun Ding and Yudong Chen. Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Transactions on Information Theory*, 2020.

[DFU20] Lijun Ding, Jicong Fan, and Madeleine Udell. k fw: A frank-wolfe style algorithm with stronger subproblem oracles. *arXiv preprint arXiv:2006.16142*, 2020.

[DG20] Lijun Ding and Benjamin Grimmer. Revisit of spectral bundle methods: Primal-dual (sub) linear convergence rates. *arXiv preprint arXiv:2008.07067*, 2020.

[DU] Lijun Ding and Madeleine Udell. A strict complementary slackness approach to growth, error bound, and sensitivity of solution in conic programming. *To appear soon.*

[DU20]   Lijun Ding and Madeleine Udell. On the regularity and conditioning of low rank semidefinite programs. *arXiv preprint arXiv:2002.10673*, 2020.

[DYC⁺19] Lijun Ding, Alp Yurtsever, Volkan Cevher, Joel A Tropp, and Madeleine Udell. An optimal-storage approach to semidefinite programming using approximate complementarity. *arXiv preprint arXiv:1902.03373*, 2019.

[DZC20]   Lijun Ding, Yuqian Zhang, and Yudong Chen. Low-rank matrix recovery with non-quadratic loss: projected gradient method and regularity projection oracle. *arXiv preprint arXiv:2008.13777*, 2020.

[JMD10]   Prateek Jain, Raghu Meka, and Inderjit S Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pages 937–945, 2010.