Proposal to Alibaba

1 Main Research Directions

Let's consider the following problem

$$F(x) = \frac{1}{m}\sum_{k=1}^{m} f_k(x) + g(x) \to \min_{x \in \mathcal{R}^n}. \tag{1}$$

This problem arises in Machine Learning applications [1], i.e. in Deep Learning [2]. Typically, this problem is not convex. Since that one can try to find the solution x_* (stationary point) only in terms of $\|\nabla F(x^N)\|_2 \leq \epsilon$. For example, if $F(x)$ has L-Lipschitz continuous gradient then one can solve (1) by classical gradient descent

$$x^{t+1} = x^t - \frac{1}{L}\nabla F(x^t)$$

with the number of oracle calls $O(L(F(x^0) - F(x_*))/\epsilon^2)$ – the number of $\nabla F(x)$ calculations. One can easily obtain this estimate by using

$$F(x^t) - F(x^{t+1}) \geq \frac{1}{2L}\|\nabla F(x^t)\|_2^2.$$

And this estimate is unimprovable in general [3]. One can also generalize this result to intermediate level of smoothness of $F(x)$, one can also consider $g(x)$ to be non-smooth composite-friendly term and, finally, one can also propose adaptive (universal) composite gradient method that doesn't require any information of smoothness and works according to the lower bounds in any case [4].

The project is devoted to the investigation of the following three global problems:

1. Unfortunately, by using of gradient-type methods we can find only a critical point, not even a local minima and especially global minima [5, 6]. The lower complexity bound for global minimum in general non-convex case is $O(\epsilon^{-(n-1)})$, which is very pessimistic.

2. The other problem is slow rate of convergence to the stationary point (in comparison to the convex case).

3. The last problem: in many applications m is too big and in gradient-type procedures the cost of one oracle call (one iteration) is proportional to m.

As for the first global problem, we propose to develop three sub-directions:

1.1 The first one is based on the investigation of the nature of the problem. In several statistics and engineering problems, including regression models with non-convex penalties and objectives, phase retrieval, non-convex (low-rank) reformulations of semidefinite programs and matrix completion it is possible to show that all stationary points are (near) global minima (see [7] for the references). So we plan to make the class of such problems broader.

1.2 The second one goes back to Nesterov–Polyak [8]. The idea is to use second-order schemes directly (using Hessian) [8] or indirectly [9] (the last scheme is much more faster). The last scheme means that we estimate Hessian–vector product from the first-order information by using automatic differentiation or representation:

$$\nabla^2 F(x)v = \frac{\nabla F(x+hv) - \nabla F(x)}{h}.$$

Simple explanation how to use this Hessian-vector product one can find at the end of the survey [10]. We also return to this sub-direction and describe how we plan to use it further, when we try to solve second global problem.

1.3 The third one is based on the observation that this isn't typical to gradient descent to get stuck in a saddle-point. And if we introduce some small noise to our procedure (one can easily do it in a cheap manner) this noise allows to avoid almost sure not only simple saddle-points as attractors [11, 12] but also high-order saddle-points [13]. We plan to develop this sub-direction by using randomized approach [14] (see also below). We assume that natural (sum-type) randomization can play the same role as axillary noise that was introduced in gradient-type procedure in [11, 12].

For the second global problem, we propose to develop six sub-directions.

2.1 The first one is to use high-order (p-order) information with additional (high-order) assumption of smoothness $F(x)$. This allows us to reduce the complexity [7] $O(\epsilon^{-2})$ (see above) to $O(\epsilon^{-\frac{p+1}{p}})$. Unfortunately, these methods have very high iteration complexity (if n is not small enough) and we don't plan to consider these methods too much time. But even for the class of first-order schemes one can in principle reduce $O(\epsilon^{-2})$ to $O(\epsilon^{-\frac{12}{7}})$ with additional assumption about smoothness of $F(x)$ [3]. We plan to concentrate in this sub-direction on the possibility to reach the last estimate and on practically implementability of such kind of methods [9], especially by utilizing sum-type structure of the objective function (1) [14, 15].

2.2 The second one is based on the idea that in the vicinity of local minima the behavior of $F(x)$ seems to be convex. So it is natural to develop such methods that utilize convexity sites of method's trajectory for (Nesterov's) acceleration on these sites. There exists such a method (uniformly optimal) that automatically (that is without any advises from our side) do it [16, 17, 18]. We plan to combine the ideas of this sub-direction with distributed and randomized tricks we'll use under consideration in the third (the last one) global problem. We consider this sub-direction very perspective. The general line here is formulation numerical method for non-convex optimization problem based on existing method that was proposed for the convex one in such a manner that this method inherit all good properties in convex sites of the trajectory.

2.3 The third sub-direction makes an additional assumption about the smallest (negative) eigenvalue of the Hessian (to this situation one can reduce more general cases [14]). If the global lower bound on this quantity is small enough in absolute value than one can consider the problem formulation (1) to be close enough to convex one. And this gives the right to hope that one can also accelerate the convergence in this case [9, 14, 19, 20, 21]. Moreover, in the works [14, 19, 20, 21] one can find the possible approach (based on randomization) to solve the third global problem (utilizing the sum-type structure of (1) to reduce the complexity of each iteration $O(m)$). It seems that this approaches strongly matches to what is interesting to Alibaba so we plan to work hard in this direction by trying to combine it with distributed context [22, 23].

2.4 The fourth sub-direction is based on the additional assumption that in (1) $F(x)$ is convex but $f_k(x)$ can be non-convex. In this case one can apply different convex-type tricks to obtain better convergence rate [24]. This is very specific situation, but we plan to involve it into consideration because of utility for the general line: using the specificity of the problem (1) to reduce the situation to well developed convex case. This line can be expressed in a bit different manner.

2.5 The fifth sub-direction is different generalization of the convexity: like Polyak–Lojasiewicz condition, star-convexity e.t.c. [8, 25]. For example, if instead of standard convexity we have only weak quasi-convexity [26]: for all x we have $F(x) - F(x_*) \leq < \nabla F(x), x - x_* >$, where x_* is the solution of (1), than there exists such a version of fast gradient method (FGM) [27] (with fixed step size) that converges like the standard FGM. Moreover, one can generalize this results to α-weak quasi-convexity [28]. We plan to introduce more appropriate class of relaxations convexity condition that will better suite for (1).

2.6 Important role in practical implementation of different methods plays the choice of the parameters (like the size of the step). We've already mentioned at the very beginning universal approach [4, 17, 29]. For example, in Deep Learning there exist special tricks that allows to choose step adaptively even for different stochastic (randomized) procedures [30]. But all these tricks don't assume acceleration by smoothness. Based on [27] we plan to propose adaptive accelerated-type procedure for stochastic (randomized) case.

The third (and the last one) global problem is the cost of iteration. For full gradient methods it's too costly to calculate whole gradient of $F(x)$. For example, in Alibaba's applications m could be of order 10^8. Here we propose to develop three sub-directions.

3.1 The first one is just sum-randomization trick. These type of methods we've been previously mentioned in other contexts [14, 15, 19, 20]. At the same time they particular reduce the complexity of each iteration.

3.2 The second sub-direction additionally assume that n is not too big (say 10^3 like in some Alibaba's applications). In this case we can also consider second-order sum-randomized schemes, where Hessian is replaced by its mini-batched estimate [31]. We consider this direction to be rather perspective theoretically. We plan to investigate the practicality of this approach.

3.3 The third (and the main) sub-direction is distributed approach [22, 23, 32]. In particular we try to build non-convex generalization of optimal convex-case methods [33]. The problem is the optimal methods in convex case significantly use primal-duality that take place only in the convex case. We try to develop different sub-directions for different global problems mentioned above for distributed context. One of the ideas is to consider penalty reformulation [34] of initial problem (1) (for simplicity here we assume $g = 0$, $n = 1$)

$$\frac{1}{m}\sum_{k=1}^{m} f_k(x_k) \to \min_{x_1=...=x_m},$$

$$\frac{1}{m}\sum_{k=1}^{m} f_k(x_k) \to \min_{Wx=0},$$

$$\frac{1}{m}\sum_{k=1}^{m} f_k(x_k) \to \min_{<x,Wx>=\|\sqrt{W}x\|_2^2=0},$$

$$F_\gamma(x_1,...,x_n) = \frac{1}{m}\sum_{k=1}^{m} f_k(x_k) + \frac{\gamma}{2} < x, Wx > \to \min, \tag{2}$$

where $\gamma = O(1/\epsilon)$ and for Laplacian matrix W of communication network we have that $(x = (x_1, ..., x_m)^T)$: $Wx = 0$ is equivalent to $\sqrt{W}x = 0$ and to $x_1 = ... = x_m$. Gradient of $F_\gamma(x) = F_\gamma(x_1, ..., x_m)$ can be calculated in a distributed manner. The solution x_γ of (2) guarantee that $F(x_\gamma) - F(x_*) = O(\epsilon)$ and $\|\sqrt{W}x_\gamma\|_2 = O(\epsilon)$. In centralized case we plan to investigate asynchronous delays [35] and possibility of reduction of this delays in smooth case for randomized procedures via splitting smooth term (sensitive to asynchronism) and stochastic term (non sensitive to asynchronism) [36]. Also we plan to consider asynchronous parallelization delays [37].

So the basic idea of our approach is searching the better combinations of mentioned above sub-directions (and its generalizations) that will suite for different concrete non-convex problems of type (1) Alibaba interested in.

2 Other Directions

1. Multiobjective Optimization

Among other possible directions for future joint research is multiobjective optimization which finds itself adequate to numerous applications.

The formulation of the problem is as follows. Given a compact *feasible set* $Q \subset \mathbb{R}^n$, and m scalar *objective functions* $f_i \colon \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$, minimize all of them simultaneously. Obviously, in general, the solution cannot be attained at a single point in Q.

Instead, consider the image of Q under the mapping specified by the f_is: $F = f(Q) = \{z \in \mathbb{R}^m \colon z_i = f_i(x) \text{ for some } x \in Q\}$. Next, introduce the notion of dominance. Namely, $z^* \in F$ is said to *dominate* $z \in F$, $z^* \preccurlyeq z$, if $z_i^* \leqslant z_i$, $i = 1, \ldots, m$ and there exists at least one j such that $z_j^* < z_j$. A typical approach to multiobjective optimization is to describe the *Pareto set (front)* defined as $\mathcal{P} = \{p \in F \colon \nexists\, z \in F \colon z \preccurlyeq p\}$ and its pre-image.

By definition, the points in $\mathcal{P}$ are "equally optimal" in the sense that, improving one of the objectives leads to unavoidable degradation of (some of the) others. Hence, this whole set is adopted as a solution to the problem.

In applications, the dimensionality of the set Q may be large (thousands) and it is usually poorly scaled, the image F is highly nonconvex, since the objectives are not convex (and may be nondifferentiable), their closed-form description is not available (instead, a zero-order oracle only exists which returns values of the f_is). Moreover, the Pareto front may happen to be non-connected. These features make the problem difficult to solve. Also, the final *decision making*—choosing a unique preferable point in $\mathcal{P}$—is a nontrivial problem.

Our group has certain experience in solving such kind of problems; in particular, we have developed a randomization-based iterative method that monotonically improves approximations to the Pareto front.

2. In case of constrained optimization a problem can be non-convex due to constraint set, not the function. There are algorithms can be applied for the optimization of such sets, i.e. smooth manifolds. The idea is to stay on constraint set for each steps rather than to use approximations [38].

3. To cope with large-scale non-convex problems we propose to apply local and global search methods. We plan to use Fast Automatic Differentiation [39, 40] techniques to compute first and second order derivatives of an objective. This approach will enable the use of efficient local search methods based on derivatives, e.g. gradient methods, Newton-type methods and quasi-Newton methods [41]. An obvious drawback of Newton methods is an expensive Hessian evaluation. This issue becomes critical for large scale problems. It is planned to develop new methods based on partial Hessian evaluation and Hessian approximation. Such methods converge to a minimum with approximately the same rate as Newton methods but do not require full Hessian evaluation. To overcome trapping in local minima we propose to use the stochastic methods [42] proved to efficiently "jump" between basins of attractions. Contemporary variants of pattern search [43] and coordinate descent [44] methods are also worth a try.

References

[1] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[2] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[3] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, "Lower bounds for finding stationary points ii: First-order methods," *arXiv preprint arXiv:1711.00841*, 2017.

[4] P. Dvurechensky, "Gradient method with inexact oracle for composite non-convex optimization," *arXiv preprint arXiv:1703.09180*, 2017.

[5] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013, vol. 87.

[6] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, A. Singh, and B. Poczos, "Gradient descent can take exponential time to escape saddle points," in *Advances in Neural Information Processing Systems*, 2017, pp. 1067–1077.

[7] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, "Lower bounds for finding stationary points i," *arXiv preprint arXiv:1710.11606*, 2017.

[8] Y. Nesterov and B. T. Polyak, "Cubic regularization of newton method and its global performance," *Mathematical Programming*, vol. 108, no. 1, pp. 177–205, 2006.

[9] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, "Accelerated methods for non-convex optimization," *arXiv preprint arXiv:1611.00756*, 2016.

[10] S. J. Wright, "Optimization algorithms for data analysis," *IAS/Park City Mathematics Series, to appear*, 2017.

[11] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *Conference on Learning Theory*, 2016, pp. 1246–1257.

[12] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, "First-order methods almost always avoid saddle points," *arXiv preprint arXiv:1710.07406*, 2017.

[13] A. Anandkumar and R. Ge, "Efficient approaches for escaping higher order saddle points in non-convex optimization," in *Conference on Learning Theory*, 2016, pp. 81–102.

[14] Z. Allen-Zhu and Y. Li, "Neon2: Finding local minima via first-order oracles," *arXiv preprint arXiv:1711.06673*, 2017.

[15] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma, "Finding approximate local minima faster than gradient descent," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2017, pp. 1195–1199.

[16] S. Ghadimi and G. Lan, "Accelerated gradient methods for nonconvex nonlinear and stochastic programming," *Mathematical Programming*, vol. 156, no. 1-2, pp. 59–99, 2016.

[17] S. Ghadimi, G. Lan, and H. Zhang, "Generalized uniformly optimal methods for nonlinear programming," *arXiv preprint arXiv:1508.07384*, 2015.

[18] Y. Carmon, O. Hinder, J. C. Duchi, and A. Sidford, "" convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions," *arXiv preprint arXiv:1705.02766*, 2017.

[19] Z. Allen-Zhu, "Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter," in *International Conference on Machine Learning*, 2017, pp. 89–97.

[20] ——, "Natasha 2: Faster non-convex optimization than sgd," *arXiv preprint arXiv:1708.08694*, 2017.

[21] G. Lan and Y. Yang, "Accelerated stochastic algorithms for nonconvex finite-sum and multi-block optimization," *arXiv preprint arXiv:1805.05411*, 2018.

[22] H. Sun and M. Hong, "Distributed non-convex first-order optimization and information processing: Lower complexity bounds and rate optimal algorithms," *arXiv preprint arXiv:1804.02729*, 2018.

[23] I. Notarnicola, Y. Sun, G. Scutari, and G. Notarstefano, "Distributed big-data optimization via block-iterative convexification and averaging," *arXiv preprint arXiv:1805.00658*, 2018.

[24] Z. Allen-Zhu, "Katyusha x: Practical momentum method for stochastic sum-of-nonconvex optimization," *arXiv preprint arXiv:1802.03866*, 2018.

[25] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2016, pp. 795–811.

[26] M. Hardt, T. Ma, and B. Recht, "Gradient descent learns linear dynamical systems," *arXiv preprint arXiv:1609.05191*, 2016.

[27] A. Tyurin, "Mirror version of similar triangles method for constrained optimization problems," *arXiv preprint arXiv:1705.09809*, 2017.

[28] S. Guminov, A. Gasnikov, A. Anikin, and A. Gornov, "A universal modification of the linear coupling method," *arXiv preprint arXiv:1711.01850*, 2017.

[29] P. Ochs, J. Fadili, and T. Brox, "Non-smooth non-convex bregman minimization: Unification and new algorithms," *arXiv preprint arXiv:1707.02278*, 2017.

[30] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[31] S. Ghadimi, H. Liu, and T. Zhang, "Second-order methods with cubic regularization under inexact information," *arXiv preprint arXiv:1710.05782*, 2017.

[32] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[33] C. A. Uribe, S. Lee, A. Gasnikov, and A. Nedić, "Optimal algorithms for distributed optimization," *arXiv:1712.00232*, 2017.

[34] D. P. Bertsekas, *Nonlinear programming*. Athena scientific Belmont, 1999.

[35] A. Agarwal and J. C. Duchi, "Distributed delayed stochastic optimization," in *Advances in Neural Information Processing Systems*, 2011, pp. 873–881.

[36] P. Dvurechensky and A. Gasnikov, "Stochastic intermediate gradient method for convex problems with stochastic inexact oracle," *Journal of Optimization Theory and Applications*, vol. 171, no. 1, pp. 121–145, 2016.

[37] B. Recht, C. Re, S. Wright, and F. Niu, "Hogwild: A lock-free approach to parallelizing stochastic gradient descent," in *Advances in neural information processing systems*, 2011, pp. 693–701.

[38] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ: Princeton University Press, 2008.

[39] A. Griewank and A. Walther, *Evaluating derivatives: principles and techniques of algorithmic differentiation*. Siam, 2008, vol. 105.

[40] K. R. Aida-zade and Y. G. Evtushenko, "Fast automatic differentiation on computers," *Matematicheskoe Modelirovanie*, vol. 1, no. 1, pp. 120–131, 1989.

[41] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization," *ACM Transactions on Mathematical Software (TOMS)*, vol. 23, no. 4, pp. 550–560, 1997.

[42] D. J. Wales and J. P. Doye, "Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms," *The Journal of Physical Chemistry A*, vol. 101, no. 28, pp. 5111–5116, 1997.

[43] M. Wetter and J. Wright, "Comparison of a generalized pattern search and a genetic algorithm optimization method," in *Proceedings of the 8th International IBPSA Conference, Eindhoven, Netherlands*, 2003, pp. 1401–1408.

[44] S. J. Wright, "Coordinate descent algorithms," *Mathematical Programming*, vol. 151, no. 1, pp. 3–34, 2015.