The Nonconvex Geometry of Low-Rank Matrix Optimizations

Qiuwei Li, Zhihui Zhu, Gongguo Tang
Department of Electrical Engineering and Computer Science, Colorado School of Mines, CO, USA
Email: {qiuli, zzhu}@mymail.mines.edu, gtang@mines.edu

The past few years have seen a surge of interest in nonconvex reformulations of convex optimizations for efficiency and scalability reasons [1]–[8]. Compared with the convex formulations, the nonconvex ones typically involve many fewer variables, allowing them to scale to scenarios with millions of variables. In addition, simple algorithms [8]–[13] applied to the nonconvex formulations have surprisingly good performance in practice. A complete understanding of this phenomenon, particularly the geometrical structures of these nonconvex optimizations, is still an active research area. Unlike the simple geometry of convex optimizations where local minimizers are always global ones, the landscapes of general nonconvex functions can become as complicated as it could be. Fortunately, for a range of convex optimizations, particularly matrix completion and sensing problems, the corresponding nonconvex reformulations have nice geometric structures that allow local search algorithms to converge to global optimality [1]–[3], [6], [14], [15]. We extend this line of investigation by working with a general convex function $f(X)$ and considering the following two optimizations:

$$\operatorname*{minimize}_{X\in\mathbb{R}^{p\times p}} f(X) \text{ subject to } X\succeq 0 \tag{P0}$$

$$\operatorname*{minimize}_{X\in\mathbb{R}^{p\times q}} f(X)+\lambda\|X\|_*, \tag{P1}$$

both of which are assumed to admit a low-rank optimizer $X^\star$ with $\operatorname{rank}(X^\star)=r^\star$ [16]. For these problems, standard first-order convex solvers [17], [18] require performing an eigenvalue (or singularvalue) decomposition in each iteration, severely limiting their efficiency and scalability in applications [4], [19]–[27].

Our Approach: Burer-Monteiro Style Parameterization

Burer and Monteiro [28] proposed to factorize a low-rank variable $X=UU^T$ (for semi-definite matrices) or $X=UV^T$ (for general matrices) where $U\in\mathbb{R}^{p\times r}$ and $V\in\mathbb{R}^{q\times r}$ with $r\ll\{p,q\}$. Moreover, by noting $\|X\|_*=\operatorname{minimize}_{X=UV^T}(\|U\|_F^2+\|V\|_F^2)/2$, we obtain the following nonconvex reparameterizations of (P0)-(P1):

$$\operatorname*{minimize}_{U\in\mathbb{R}^{p\times r}} g(U)=f(UU^T) \tag{F0}$$

$$\operatorname*{minimize}_{U\in\mathbb{R}^{p\times r},V\in\mathbb{R}^{q\times r}} g(U,V)=f(UV^T)+\lambda(\|U\|_F^2+\|V\|_F^2)/2 \tag{F1}$$

Since $r\ll\{p,q\}$, these factored problems (F0)-(F1) involve many fewer variables.

The past two years have seen renewed interest in the Burer-Monteiro factorization for solving trace norm regularized inverse problems [29]–[34]. With technical innovations in analyzing the nonconvex landscape of the factored objective function, several recent works have shown that with exact parameterization (*i.e.*, $r=r^\star$) the factored objective function $g(U)$ (or $g(U,V)$) in (F0)-(F1) has no spurious local minima or degenerate saddle points [1]–[3], [35], [36]. An important implication is that local search algorithms such as gradient descent and its variants are able to converge to the global optimum with even random initialization [2].

We generalize this line of work by assuming a general objective function $f(X)$ in (P0)-(P1), not necessarily coming from a matrix inverse problem. The generality allows us to view the factored problems (F0)-(F1) as a way to solve the convex optimizations (P0)-(P1) to the global optimum, rather than a new modeling method. This perspective, also taken by Burer and Monteiro in their original work [28], frees us from rederiving the statistical performances of the factored optimizations (F0)-(F1). Instead, its performance inherits from that of the convex optimizations (P0)-(P1), whose performance can be developed using a suite of powerful convex analysis techniques accumulated from several decades of research. In addition, our general analysis technique also sheds light on the connection between the geometries of the convex programs (P0)-(P1) and its nonconvex counterparts (F0)-(F1).

Our Main Result

Our governing assumption on the objective function $f(X)$ is $2r$-restricted well-conditionedness:

$$m\mathbf{I}\preceq\nabla^2 f(X)\preceq M\mathbf{I} \text{ with } M/m\le 1.5 \text{ if } \operatorname{rank}(X)\le 2r \tag{1}$$

This assumption is standard in matrix inverse problem [37], [38]. We show that under this assumption combined with a small condition number M/m, we have the following theorem:

Theorem 1. *Suppose the objective function $f(X)$ is convex and satisfies (1). Assume $X^\star$ is an optimal solution of the minimization (P0) or (P1) with $\operatorname{rank}(X^\star)=r^\star$. Set $r\ge r^\star$ in (F0)-(F1). Then any critical point U (or (U,V)) of g in (F0)-(F1) either corresponds to the global optimizer $X^\star$ where $X^\star=UU^T$ (or $X^\star=UV^T$) or is a strict saddle point (or a local maximum) of the factored problems (F0)-(F1), where the Hessian $\nabla^2 g$ has a strictly negative eigenvalue,* i.e., *$\lambda_{\min}(\nabla^2 g(U))<0$ or $\lambda_{\min}(\nabla^2 g(U,V))<0$.*

Note that our result covers both over-parameterization where $r>r^\star$ and exact parameterization where $r=r^\star$. The geometric property established in the theorem ensures that many iterative algorithms [8]–[11] converge to a square-root factor (or a factorization) of $X^\star$, even with random initialization. Therefore, we can recover the rank-$r^\star$ global minimizer $X^\star$ of (P0)-(P1) by running local search algorithms on the factored function $g(U)$ (or $g(U,V)$) if we know an upper bound on the rank $r^\star$. Furthermore, our main result only relies on the restricted well-conditionedness of $f(X)$. Therefore, in addition to low-rank matrix recovery problems [5], [15], [39], it is also applicable to many other low-rank matrix optimization problems with non-quadratic objective functions, including 1-bit matrix completion [40], [41], robust PCA [42]–[44], Poisson PCA [45], and other low-rank models with generalized loss functions [46]. For problems with additional linear constraints, as those studied in [28], [47], one can combine the original objective function with a least-squares term that penalizes the deviation from the linear constraints. As long as the penalization parameter is large enough, the solution is equivalent to that of the constrained minimizations and hence is also covered by our main theorem.

Acknowledgment

This work is supported by the NSF via Grant CCF-1464205.

REFERENCES

[1] C. Jin, S. M. Kakade, and P. Netrapalli, "Provable efficient online matrix completion via non-convex stochastic gradient descent," *arXiv preprint arXiv:1605.08370*, 2016.

[2] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," *arXiv preprint arXiv:1605.07221*, 2016.

[3] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi, "Dropping convexity for faster semi-definite optimization," *arXiv preprint*, 2015.

[4] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM review*, vol. 57, no. 2, pp. 225–251, 2015.

[5] P. Jain, C. Jin, S. M. Kakade, and P. Netrapalli, "Computing matrix squareroot via non convex local search," *arXiv preprint arXiv:1507.05854*, 2015.

[6] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via nonconvex factorization," in *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pp. 270–289, IEEE, 2015.

[7] Q. Li, A. Prater, L. Shen, and G. Tang, "Overcomplete tensor decomposition via convex optimization," *arXiv preprint arXiv:1602.08614*, 2016.

[8] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere ii: Recovery by riemannian trust-region method," *arXiv preprint arXiv:1511.04777*, 2015.

[9] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent converges to minimizers," *University of California, Berkeley*, vol. 1050, p. 16, 2016.

[10] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle pointsłonline stochastic gradient for tensor decomposition," in *Proceedings of The 28th Conference on Learning Theory*, pp. 797–842, 2015.

[11] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *arXiv preprint arXiv:1602.06664*, 2016.

[12] A. Anandkumar, R. Ge, and M. Janzamin, "Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates," *arXiv preprint arXiv:1402.5180*, 2014.

[13] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.

[14] J. Sun, Q. Qu, and J. Wright, "When are nonconvex problems not scary?," *arXiv preprint arXiv:1510.06096*, 2015.

[15] T. Zhao, Z. Wang, and H. Liu, "A nonconvex optimization framework for low rank matrix estimation," in *Advances in Neural Information Processing Systems*, pp. 559–567, 2015.

[16] Q. Li and G. Tang, "The nonconvex geometry of low-rank matrix optimizations with general objective functions," *arXiv:1611.03060*, 2016.

[17] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[18] Y. Chen and M. J. Wainwright, "Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees," *arXiv preprint arXiv:1509.03025*, 2015.

[19] I. Waldspurger, A. dAspremont, and S. Mallat, "Phase recovery, maxcut and complex semidefinite programming," *Mathematical Programming*, vol. 149, no. 1-2, pp. 47–81, 2015.

[20] S. Aaronson, "The learnability of quantum states," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 463, pp. 3089–3114, The Royal Society, 2007.

[21] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, "Quantum state tomography via compressed sensing," *Physical review letters*, vol. 105, no. 15, p. 150401, 2010.

[22] D. DeCoste, "Collaborative prediction using ensembles of maximum margin matrix factorizations," in *Proceedings of the 23rd international conference on Machine learning*, pp. 249–256, ACM, 2006.

[23] J. D. Rennie and N. Srebro, "Fast maximum margin matrix factorization for collaborative prediction," in *Proceedings of the 22nd international conference on Machine learning*, pp. 713–719, ACM, 2005.

[24] N. Srebro, J. Rennie, and T. S. Jaakkola, "Maximum-margin matrix factorization," in *Advances in neural information processing systems*, pp. 1329–1336, 2004.

[25] M. Weimer, A. Karatzoglou, Q. V. Le, and A. Smola, "Maximum margin matrix factorization for collaborative ranking," *Advances in neural information processing systems*, pp. 1–8, 2007.

[26] P. Biswas, T.-C. Liang, K.-C. Toh, Y. Ye, and T.-C. Wang, "Semidefinite programming approaches for sensor network localization with noisy distance measurements," *IEEE transactions on automation science and engineering*, vol. 3, no. 4, p. 360, 2006.

[27] P. Biswas and Y. Ye, "Semidefinite programming for ad hoc wireless sensor network localization," in *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pp. 46–54, ACM, 2004.

[28] S. Burer and R. D. Monteiro, "A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization," *Mathematical Programming*, vol. 95, no. 2, pp. 329–357, 2003.

[29] Q. Zheng and J. Lafferty, "Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent," *arXiv preprint arXiv:1605.07051*, 2016.

[30] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi, "Non-square matrix sensing without spurious local minima via the burer-monteiro approach," *arXiv preprint arXiv:1609.03240*, 2016.

[31] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi, "Finding low-rank solutions to matrix problems, efficiently and provably," *arXiv preprint arXiv:1606.03168*, 2016.

[32] Q. Li and G. Tang, "The nonconvex geometry of low-rank matrix optimizations with general objective functions," *arXiv preprint arXiv:1611.03060*, 2016.

[33] L. Wang, X. Zhang, and Q. Gu, "A unified computational and statistical framework for nonconvex low-rank matrix estimation," *arXiv preprint arXiv:1610.05275*, 2016.

[34] Q. Zheng and J. Lafferty, "A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements," in *Advances in Neural Information Processing Systems*, pp. 109–117, 2015.

[35] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," *arXiv preprint arXiv:1605.07272*, 2016.

[36] S. Tu, R. Boczar, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," *arXiv preprint arXiv:1507.03566*, 2015.

[37] A. Agarwal, S. Negahban, and M. J. Wainwright, "Fast global convergence rates of gradient methods for high-dimensional statistical recovery," in *Advances in Neural Information Processing Systems*, pp. 37–45, 2010.

[38] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *Journal of Machine Learning Research*, vol. 13, no. May, pp. 1665–1697, 2012.

[39] E. J. Candes and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.

[40] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters, "1-bit matrix completion," *Information and Inference*, vol. 3, no. 3, pp. 189–223, 2014.

[41] L. Kozma, A. Ilin, and T. Raiko, "Binary principal component analysis in the netflix collaborative filtering task," in *2009 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6, IEEE, 2009.

[42] Q. Li, G. Tang, and A. Nehorai, "Robust principal component analysis based on low-rank and block-sparse matrix decomposition," *Handbook of Robust Low-Rank and Sparse Matrix Decomposition: Applications in Image and Video Processing*, 2016.

[43] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, "Non-convex robust pca," in *Advances in Neural Information Processing Systems*, pp. 1107–1115, 2014.

[44] Q. Zhao, D. Meng, Z. Xu, W. Zuo, and L. Zhang, "Robust principal component analysis with complex noise," in *Proceedings of The 31st International Conference on Machine Learning*, pp. 55–63, 2014.

[45] J. Salmon, Z. Harmany, C.-A. Deledalle, and R. Willett, "Poisson noise reduction with non-local pca," *Journal of mathematical imaging and vision*, vol. 48, no. 2, pp. 279–294, 2014.

[46] M. Udell, C. Horn, S. Boyd, and R. Zadeh, "Generalized low rank models," *Foundations and Trends(r) in Machine Learning*, vol. 9, no. 1, pp. 1–118, 2016.

[47] N. Boumal, V. Voroninski, and A. S. Bandeira, "The non-convex burer-monteiro approach works on smooth semidefinite programs," *arXiv preprint arXiv:1606.04970*, 2016.