Ju Sun | Research Statement

sunju@stanford.edu • www.sunju.org

SYNOPSIS. *Exploiting low-complexity structures to solve foundational & applicational problems in modern data analysis with practical algorithms & computational guarantees.*

Modern society is awash with data. In scientific domains, innovative sensory and measurement technologies are spawning gigantic data collections, ranging from high-resolution brain and cell maps [1, 2], century-long climate records [3], to full-angle sky survey [4]. In civil life, consumer electronic devices interconnected by social media are creating enormous amounts of texts, images, videos, and web searches, nonstop. Modern data are huge in size, complex in form, and often high-dimensional in raw representation. All these aspects pose stark challenges to effective and efficient data analysis. Taming these challenges will undoubtedly bear multiple scientific and societal impacts, enabling improved healthcare, intuitive web search, vibrant social interactions, and more.

Representation learning. High dimensionality is a hallmark of modern data and causes severe conceptual and technical difficulties for learning—exponential (in dimension) number of samples are needed for effective estimation and inference in general [5, 6]. Nonetheless, practical high-dimensional data ($\boldsymbol{y}$) are invariably generated by processes (f) with few deciding factors ($\boldsymbol{x}$) and hence have small degrees of freedom: $\boldsymbol{y} = f(\boldsymbol{x})$. Deriving *low-complexity representation* that reflects the intrinsic data complexity (i.e., recover $\boldsymbol{x}$ from $\boldsymbol{y}$) cures high dimensionality and dominates all data processing stages, from data compression to data-driven decision making. For the case f is linear and known, remarkable theoretical and practical advances on recovery of $\boldsymbol{x}$ have been made over the past decade [7]. However, in practice f is generally unknown, motivating the fundamental problem of *representation learning*: f (or an "inverse" map $g : \boldsymbol{y} \mapsto \boldsymbol{x}$) and $\{\boldsymbol{x}_i\}_{i=1,\dots,m}$ are simultaneously learned from data samples $\{\boldsymbol{y}_i = f(\boldsymbol{x}_i)\}_{i=1,\dots,m}$. This learning-based paradigm is sensible due to the vast availability of data, and it is attractive because the sheer volume and form of new data preclude deriving f analytically. Many successful heuristics in this line have come with little performance guarantee [8, 9, 10, 11]. Our results [12, 13] show that when f is linear but *unknown*, representation learning can be performed provably and efficiently with practical algorithms, in regimes inaccessible to previous theory [14].

Nonconvex optimization. Most problems in modern data analysis can be naturally cast as optimization problems. Convex analysis and optimization has matured and become an indispensable modeling language and problem-solving tool across diverse disciplines [15, 16, 17]. However, natural formulations for practical problems, including representation learning, are predominantly nonconvex. In theory, nonconvex problems are difficult to optimize globally [18]. In practice, heuristic methods are surprisingly effective and efficient in finding good solutions—best exemplified by the recent breakthroughs of training deep neural networks [11]. The puzzling gap between theory and practice was unexplained for decades. We have recently discovered surprising *low-complexity nonconvex problems* arising in dictionary learning [12, 13] and generalized phase retrieval [19]: *when data are numerous and generic enough, all local minimizers of the nonconvex problems are global*. The benign structure has enabled us to derive novel computational guarantees for both problems [12, 13, 19]. Analogous structures have subsequently been confirmed on other nonconvex problems of practical interest [20, 21]. This line of development signifies the first step toward bridging the curious theory-practice gap for nonconvex optimization.

Next, I will describe our successful examples in more detail, sketch our ongoing efforts of addressing new problems out of the confluence of representation learning and provable nonconvex optimization, and hint on application to computer vision, scientific imaging, and beyond.

REPRESENTATION – WHEN CAN EFFICIENT REPRESENTATIONS BE LEARNED?

As alluded to above, efficient and informative data representations form the backbone of modern data analysis. Representation learning is a timely and necessary paradigm shift that capitalizes on the defining characteristics of modern data sets. In this section, we focus on learning representation of data with linear f (represented by matrix $\boldsymbol{A}(f)$), as a

first nontrivial step:

$$\underbrace{\boldsymbol{Y}}_{\text{data}} = \underbrace{\boldsymbol{A}\,(f)}_{\text{unknown linear basis}} \quad \underbrace{\boldsymbol{X}}_{\text{unknown coefficients (representation)}} . \tag{1}$$

Here both $\boldsymbol{A}$ and $\boldsymbol{X}$ are unknown. The bilinear structure tends to resist effective convex relaxation. Different structural assumptions put on $\boldsymbol{A}$ and $\boldsymbol{X}$, often toward a "simple" $\boldsymbol{X}$ and hence efficient representation, will lead to different learning problems that fit different applicational needs.

Learning multiple subspaces. Principal component analysis (PCA) is a basic yet powerful data modeling tool that approximates a collection of high-dimensional data points with a low-dimensional subspace. When the data points stem from multiple sources, say facial images of multiple persons, it makes sense to fit the data with multiple subspaces, each corresponding to an individual source. This corresponds to assuming in (1): $\boldsymbol{A} = [\boldsymbol{A}_1, \ldots, \boldsymbol{A}_k]$ and $\boldsymbol{X} = \operatorname{diag}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)$. In words, given the data $\boldsymbol{Y}$, the task is to segment the data into their originating subspaces and provide the subspace bases $\boldsymbol{A}_i$ and the representations $\boldsymbol{X}_i$. This *subspace clustering* problem finds novel applications in image and video processing and optimal control [22] and is the first tool of choice when modeling hybrid data. We showed [23, 24] that when the data $\boldsymbol{Y}$ are clean, under mild conditions the underlying subspaces can be provably recovered from the solution of the convex program:

$$\text{minimize}_{\boldsymbol{Z},\boldsymbol{E}} \ \|\boldsymbol{Z}\|_* + \lambda \|\boldsymbol{E}\|_\ell \quad \text{subject to } \boldsymbol{Y} = \boldsymbol{Y}\boldsymbol{Z} + \boldsymbol{E}.$$

In practice, $\boldsymbol{Y}$ may contain outlying data points or be corrupted by gross errors, and the above formulation is capable of robustly recovering the subspaces [23]. This work has spurred substantial further developments and applications[1]; it is one of the principal methods for subspace clustering [22].

Learning sparse representation. *Sparsifying dictionary learning* (SDL) concerns the case when $\boldsymbol{A}$ is generic and $\boldsymbol{X}$ is sparse, i.e., containing maximal possible zeros. $\boldsymbol{A}$ is naturally called the sparsifying dictionary. Initially proposed in neuroscience [8], SDL has seen intensive algorithmic developments across several communities with numerous successes and extensions in applications, spanning low-level image processing to high-level visual recognition [25].

Despite the abundant empirical successes, until recently there has been little understanding of when and how SDL can be performed provably. The recent breakthrough [14] showed that when $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is square and invertible and $\boldsymbol{X}$ obeys a certain random model, $\boldsymbol{A}$ and $\boldsymbol{X}$ can be efficiently and provably recovered from $\boldsymbol{Y}$. However, both [14] and subsequent works only succeed in the super-sparse regime, i.e., when each column of $\boldsymbol{X}$ contains no more than $O(\sqrt{n})$ nonzeros, which may be stringent for applications. Breaking this $O(\sqrt{n})$ barrier was a standing challenge for years.

In [12, 13, 26], we gave the first method that *efficiently and correctly recovers* $\boldsymbol{A}$ *and* $\boldsymbol{X}$ *when* $\boldsymbol{X}$ *has* $O(n)$ *nonzeros per column*[2]. Similar to [14], we rephrase the recovery problem as the geometric problem of finding sparse vectors in a given linear subspace. To break the $O(\sqrt{n})$ barrier, we moved off the convex path set in [14] and analyzed a natural nonconvex heuristic. We showed that despite the nonconvexity, the resulting optimization problem has a benign global structure that allows efficient global optimization—simple iterative methods can provably find a global optimizer from arbitrary initializations! This stands in contrast to a later approach based on sum-of-squares [27, 28]: although it works in the same regime, the need for solving huge-scale semidefinite programs renders it impractical.

Optimization – When Do Nonconvex Heuristics Succeed in Practice?

Nonconvexity plays a crucial role in our breaking the $O(\sqrt{n})$ barrier; in [29], we broke a similar barrier for the related problem of recovering a planted sparse vector from a generic linear subspace, again based on nonconvex optimization. Our quest into provable nonconvex heuristics was ignited by an intriguing numerical discovery: certain nonconvex heuristics of learning orthobases for natural images seem to always return solutions of equivalent quality, irrespective of initializations! This appears to contradict the prevailing intuition that nonconvex problems may possess many spurious local minimizers of vastly distinct qualities.

Our analysis [12, 13, 26] gave a clear geometric explanation of this curious phenomenon for SDL. The explanation uncovers an emerging "blessing of massive data" effect in optimization with modern data: *when the data are numerous and*

[1] According to Google Scholar, this work has been cited over $1,000$ times to date.
[2] This work received the Best Student Paper Prize from SPARS 2015.

generic enough, certain nonconvex problems become highly structured geometrically such that they can be solved to global optimality with simple iterative methods, independent of initializations. The geometric structure is simple to describe, but presently difficult to demonstrate on specific problems: (1) all local minimizers are global, and (2) around all other critical points there are directions of negative curvature. We call it the $\mathcal{X}$ structure.

In [19], we showed the $\mathcal{X}$ structure obtains for the *generalized phase retrieval* (GPR) problem of recovering a complex signal $\boldsymbol{x} \in \mathbb{C}^n$ from magnitude measurements $|\boldsymbol{A}\boldsymbol{x}|$, where $\boldsymbol{A} : \mathbb{C}^n \mapsto \mathbb{C}^m$ represents a linear map. This problem arises in areas such as crystallography, electron microscopy, acoustics, and more [30, 31]. Our geometric analysis indicates that when the measurements are generic ($\boldsymbol{A}$ is i.i.d. Gaussian) and numerous enough ($m \sim n \log^3 n$), a natural least-squares formulation for GPR has the $\mathcal{X}$ structure, which allows efficient iterative methods to find the global solutions and recover the complex signal, starting from arbitrary initializations.

Over a mere three years, the $\mathcal{X}$ structure has been demonstrated on a number of nonconvex problems spanning signal processing, machine learning, and scientific imaging. These include SDL [12, 13], GPR [19], orthogonal tensor decomposition [32], low-rank matrix recovery [33, 34], linear/shallow neural networks [35, 36]; we anticipate more examples across diverse fields.

Solving nonconvex problems based on global geometric characterization contrasts with the popular alternative which relies on problem-specific initialization followed by local refinement [37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61][3]. For practical problems, deriving an effective initialization might be technically demanding or infeasible, while the global geometric knowledge allows "initialization-free" optimization. For most of the above problems, guaranteed results based on semidefinite programming (SDP) are also available, e.g., SDL [27, 28], GPR [62, 63], and low-rank matrix recovery [64, 65, 66]. While these results are theoretically rigorous and beautiful, the entailed SDP optimization is computationally expensive. In contrast, the provable nonconvex methods often enjoy lower computational complexity and are more practical for applications.

Current analyses in verifying the $\mathcal{X}$ structure require study of the first- and second-order derivatives [21, 67]. This is problematic for nonsmooth functions, which are frequently chosen in optimization to promote low-complexity structures [7]. Adopting smooth surrogates could lead to success (e.g.,[12, 13]) but meanwhile induce cumbersome analytic problems. In our ongoing work [68], we show how one can directly work with nonsmooth functions using only first-order subdifferentials [69, 70] [4] and demonstrate the global geometry. This removes serious technical hurdles in solving practical problems.

Once the $\mathcal{X}$ structure is confirmed on a specific problem, a number of first- and second-order iterative methods can provably find a global minimizer from arbitrary initializations [13, 20, 21, 32, 71, 72, 73, 74]. Besides the derivative-based methods, derivative-free methods such as the *alternating direction method of multipliers* (ADMM) and *block coordinate descent* (BCD) method are also observed to often work well on practical nonconvex problems: for certain structured nonconvex, nonsmooth problems, they return the solutions with considerably low cost. In our ongoing work [75], we are making progress in understanding the local and the global convergence behaviors of ADMM on GPR[5], as a first step toward establishing provable derivative-free methods for structured nonconvex problems.

Outlook and Future Directions

Toward disciplined nonconvex modeling and optimization. The discovery of structured nonconvex optimization has led to novel and practical computational guarantees for a number of fundamental problems. However, currently, demonstrating the $\mathcal{X}$ structure on specific problems is highly, if not overwhelmingly, technical. The technicality should not stop us from identifying the $\mathcal{X}$ structure on new problems and thereof producing original results. Nonetheless, toward impacting practical modeling and problem-solving, the relevant analytic tools and optimization methods need to be reasonably modular and disciplined. In this regard, identifying operations that preserve the $\mathcal{X}$ structure[6] and understanding when and why the structure arises from practical problems are two intertwined and pressing problems to solve [21]. Calculus of variations might be a pertinent mathematical viewpoint to take, based on the intuition that minimizing appropriate functional objective could produce functions without local ripples, or spurious local minimizers.

[3] I maintain an updated list of similar work at my webpage `http://sunju.org/research/nonconvex/`.
[4] Working with generalized second-order derivatives is possible [70] but over-technical for analysis and impractical for computation.
[5] The only known workable methods for practical 2D Fourier phase retrieval are all variants of ADMM on a feasibility formulation of the problem [76].
[6] Summation which is the simplest operation preserving convexity (unfortunately) does not preserve the $\mathcal{X}$ structure.

Toward understanding of deep learning. Empirical success of deep learning is an unambiguous demonstration of the power and versatility of representation learning and nonconvex optimization. However, there is limited theoretical backup. The lack of correctness guarantees might be tolerable for applications such as online image search or speech-to-text input, but is detrimental for use in scientific discovery [77, 78] and mission-critical applications including autopilot driving and security [79, 80] systems.

There exists considerable understanding of the capacity of deep neural networks: certain neural network architectures are shown to be capable of approximating any target function with reasonable regularity [81]. But two critical aspects are still missing: (1) optimization—when and why deep networks can be successfully trained given samples from the target function, such that a close approximation is found; and (2) generalization—why in practice limited amounts of samples, much less than typical theory would predict, suffice for successful prediction. Toward a complete theory, understanding how deep learning solves inverse problems may be a reasonable first cut: in solving inverse problems, the target functions are often mathematically precise, whereas in representation learning for general recognition tasks the "right" functions sought are mostly difficult to delineate.

Applications in recognition and scientific inverse problems. The theoretical developments outlined above have been motivated by and have provided provable and efficient solutions to practical problems in computational vision and imaging, including segmenting multiple moving objects in video [23, 24], learning-based image denoising and super-resolution, and recovering phaseless imaging data [19]. The theoretical insights and technical tools have been inspiring new progress beyond vision and imaging. In collaborative work [82], we show how nonconvex optimization techniques inspired by the described developments help produce much more reliable results that the state-of-the-art in analyzing electron microscopy data.

Looking ahead, development of disciplined nonconvex optimization will undoubtedly impact all applicational fields that use optimization. Particularly, this may provide a viable way of penetrating the mystery of deep learning, on which the current analyses appear to be hard to generalize [35, 36]. A solid theory of deep learning on inverse problems will not only allow a systematic way of solving inverse problems of practical interest but also inform how deep neural networks used in recognition can be improved in design and training, which will elevate the recognition performance with reduced computing resources.

Funding & acknowledgement. Because of the broad scope and impact on both theory and applications, my research has received strong and diverse support. I have been personally supported by the Wei Family Private Foundation Fellowship. I have also been partially supported by grants from the National Science Foundation (NSF 1343282, NSF CCF 1527809, NSF IIS 1546411) and Office of Naval Research (ONR N00014-13-1-0492) and funding from the Moore and Sloan Foundations. As a postdoc scholar at Stanford, I am funded under the Math+X fellowship from the Simons Foundation which encourages interdisciplinary collaboration.

References

[1] M. J. Hawrylycz, S. Lein, A. L. Guillozet-Bongaarts, E. H. Shen, L. Ng, J. A. Miller, L. N. Van De Lagemaat, K. A. Smith, A. Ebbert, Z. L. Riley *et al.*, "An anatomically comprehensive atlas of the adult human brain transcriptome," *Nature*, vol. 489, no. 7416, p. 391, 2012.

[2] B. Roberts, A. Haupt, A. Tucker, T. Grancharova, J. Arakaki, M. A. Fuqua, A. Nelson, C. Hookway, S. A. Ludmann, I. M. Mueller *et al.*, "Systematic gene tagging using crispr/cas9 in human stem cells to illuminate cell organization," *bioRxiv*, p. 123042, 2017.

[3] NOAA Earth System Research Laboratory. PSD gridded climate datasets. [Online]. Available: https://www.esrl.noaa.gov/psd/data/gridded/

[4] B. Abolfathi, D. Aguado, G. Aguilar, C. A. Prieto, A. Almeida, T. T. Ananna, F. Anders, S. F. Anderson, B. H. Andrews, B. Anguiano *et al.*, "The fourteenth data release of the sloan digital sky survey: First spectroscopic data from the extended baryon oscillation sky survey and from the second phase of the apache point observatory galactic evolution experiment," *arXiv preprint arXiv:1707.09322*, 2017.

[5] R. E. Bellman, *Adaptive control processes: a guided tour*. Princeton university press, 2015.

[6] D. L. Donoho, "Aide-memoire. high-dimensional data analysis: The curses and blessings of dimensionality," *American Math. Society Lecture-Math Challenges of the 21st Century*, 2000.

[7] E. J. Candès, "Mathematics of sparsity (and a few other things)," in *Proceedings of the International Congress of Mathematicians, Seoul, South Korea*, vol. 123, 2014.

[8] B. A. Olshausen *et al.*, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*,

vol. 381, no. 6583, pp. 607–609, 1996.

[9] J. Mairal, F. Bach, J. Ponce *et al.*, "Sparse modeling for image and vision processing," *Foundations and Trends® in Computer Graphics and Vision*, vol. 8, no. 2-3, pp. 85–283, 2014.

[10] R. Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52–68, 2011.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere I: Overview and the geometric picture," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 853–884, 2017.

[13] ——, "Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 885–914, 2017.

[14] D. A. Spielman, H. Wang, and J. Wright, "Exact recovery of sparsely-used dictionaries," in *Conference on Learning Theory*, 2012, pp. 37–1.

[15] R. T. Rockafellar, *Convex analysis*. Princeton university press, 2015.

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

[17] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[18] K. G. Murty and S. N. Kabadi, "Some NP-complete problems in quadratic and nonlinear programming," *Mathematical programming*, vol. 39, no. 2, pp. 117–129, 1987.

[19] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *Foundations of Computational Mathematics*, Aug 2017. [Online]. Available: https://doi.org/10.1007/s10208-017-9365-9

[20] ——, "When are nonconvex problems not scary?" *NIPS Workshop on Non-convex Optimization for Machine Learning: Theory and Practice*, 2015.

[21] J. Sun and J. Wright, "Nonconvex optimization meets symmetry: Examples, algorithms, and open problems," 2017, in preparation for SIAM Review.

[22] R. Vidal, Y. Ma, and S. S. Sastry, *Generalized Principal Component Analysis*. Springer, New York, NY, 2016.

[23] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2013.

[24] Y. Ni, J. Sun, X. Yuan, S. Yan, and L.-F. Cheong, "Robust low-rank subspace segmentation with semidefinite guarantees," in *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*. IEEE, 2010, pp. 1179–1188.

[25] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM review*, vol. 51, no. 1, pp. 34–81, 2009.

[26] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery using nonconvex optimization," in *International Conference on Machine Learning*, 2015, pp. 2351–2360.

[27] B. Barak, J. A. Kelner, and D. Steurer, "Dictionary learning and tensor decomposition via the sum-of-squares method," in *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. ACM, 2015, pp. 143–151.

[28] T. Ma, J. Shi, and D. Steurer, "Polynomial-time tensor decompositions with sum-of-squares," in *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*. IEEE, 2016, pp. 438–446.

[29] Q. Qu, J. Sun, and J. Wright, "Finding a sparse vector in a subspace: Linear sparsity using alternating directions," *IEEE Transactions on Information Theory*, vol. 62, no. 10, pp. 5855–5880, 2016.

[30] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, "Phase retrieval with application to optical imaging: a contemporary overview," *IEEE signal processing magazine*, vol. 32, no. 3, pp. 87–109, 2015.

[31] K. Jaganathan, Y. C. Eldar, and B. Hassibi, "Phase retrieval: An overview of recent developments," *arXiv preprint arXiv:1510.07713*, 2015.

[32] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—online stochastic gradient for tensor decomposition," in *Conference on Learning Theory*, 2015, pp. 797–842.

[33] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Advances in Neural Information Processing Systems*, 2016, pp. 2973–2981.

[34] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," *arXiv preprint arXiv:1704.00708*, 2017.

[35] K. Kawaguchi, "Deep learning without poor local minima," in *Advances in Neural Information Processing Systems*, 2016, pp. 586–594.

[36] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks," *arXiv preprint arXiv:1707.04926*, 2017.

[37] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *Information Theory, IEEE Transactions on*, vol. 56, no. 6, pp. 2980–2998, 2010.

[38] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proceedings of the forty-fifth annual ACM symposium on Theory of Computing*. ACM, 2013, pp. 665–674.

[39] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via non-convex factorization," *arXiv preprint arXiv:1411.8003*, 2014.

[40] 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.

[41] S. Arora, R. Ge, and A. Moitra, "New algorithms for learning incoherent and overcomplete dictionaries," *arXiv preprint arXiv:1308.6273*, 2013.

[42] A. Agarwal, A. Anandkumar, and P. Netrapalli, "Exact recovery of sparsely used overcomplete dictionaries," *arXiv preprint arXiv:1309.1952*, 2013.

[43] S. Arora, R. Ge, T. Ma, and A. Moitra, "Simple, efficient, and neural algorithms for sparse coding," *arXiv preprint arXiv:1503.00778*, 2015.

[44] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Advances in Neural Information Processing Systems*, 2013, pp. 2796–2804.

[45] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *Information Theory, IEEE Transactions on*, vol. 61, no. 4, pp. 1985–2007, April 2015.

[46] Y. Chen and E. J. Candès, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," *arXiv preprint arXiv:1505.05114*, 2015.

[47] G. Wang, G. B. Giannakis, and Y. C. Eldar, "Solving systems of random quadratic equations via truncated amplitude flow," *arXiv preprint*, 2016.

[48] T. Bendory and Y. C. Eldar, "Non-convex phase retrieval from STFT measurements," *arXiv preprint arXiv:1607.08218*, 2016.

[49] P. Jain and S. Oh, "Provable tensor factorization with missing data," in *Advances in Neural Information Processing Systems*, 2014, pp. 1431–1439.

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

[51] Q. Qu, J. Sun, and J. Wright, "Finding a sparse vector in a subspace: Linear sparsity using alternating directions," in *Advances in Neural Information Processing Systems*, 2014, pp. 3401–3409.

[52] S. B. Hopkins, T. Schramm, J. Shi, and D. Steurer, "Speeding up sum-of-squares for tensor decomposition and planted sparse vectors," *arXiv preprint arXiv:1512.02337*, 2015.

[53] X. Yi, C. Caramanis, and S. Sanghavi, "Alternating minimization for mixed linear regression," *arXiv preprint arXiv:1310.3745*, 2013.

[54] H. Sedghi and A. Anandkumar, "Provable tensor methods for learning mixtures of classifiers," *arXiv preprint arXiv:1412.3046*, 2014.

[55] K. Lee and M. Junge, "RIP-like properties in subsampled blind deconvolution," *arXiv preprint arXiv:1511.06146*, 2015.

[56] K. Lee, Y. Li, M. Junge, and Y. Bresler, "Blind recovery of sparse signals from subsampled convolution," *arXiv preprint arXiv:1511.06149*, 2015.

[57] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," *arXiv preprint arXiv:1606.04933*, 2016.

[58] S. Ling and T. Strohmer, "Regularized gradient descent: A nonconvex recipe for fast joint blind deconvolution and demixing," *arXiv preprint arXiv:1703.08642*, 2017.

[59] A. Eftekhari and M. B. Wakin, "Greed is super: A fast algorithm for super-resolution," *arXiv preprint arXiv:1511.03385*, 2015.

[60] Y. Chen and E. Candès, "The projected power method: An efficient algorithm for joint alignment from pairwise differences," *arXiv preprint arXiv:1609.05820*, 2016.

[61] M. Soltanolkotabi, "Learning relus via gradient descent," *arXiv preprint arXiv:1705.04591*, 2017.

[62] E. J. Candès, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1241–1274, 2013.

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

[64] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, p. 717, 2009.

[65] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[66] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[67] J. Sun, "When are nonconvex optimization problems not scary?" Ph.D. dissertation, Columbia University, 2016.

[68] Y. Bai, J. Sun, and E. Candès, "Nonsmooth least-sqaures for Gaussian phase retrieval," 2017, in preparation.

[69] F. H. Clarke, *Optimization and nonsmooth analysis*. SIAM, 1990.

[70] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer Science & Business Media, 2009, vol. 317.

[71] N. Boumal, P.-A. Absil, and C. Cartis, "Global rates of convergence for nonconvex optimization on manifolds," *arXiv preprint*

arXiv:1605.08101, 2016.

[72] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," *arXiv preprint arXiv:1703.00887*, 2017.

[73] D. Goldfarb, "Curvilinear path steplength algorithms for minimization which use directions of negative curvature," *Mathematical programming*, vol. 18, no. 1, pp. 31–40, 1980.

[74] D. Goldfarb, C. Mu, J. Wright, and C. Zhou, "Using negative curvature in solving nonlinear programs," *arXiv preprint arXiv:1706.00896*, 2017.

[75] D. Barmherzig, J. Sun, and E. Candès, "On convergence of ADMM for Gaussian phase retrieval," 2017, in preparation.

[76] Z. Wen, C. Yang, X. Liu, and S. Marchesini, "Alternating direction methods for classical and ptychographic phase retrieval," *Inverse Problems*, vol. 28, no. 11, p. 115010, 2012.

[77] Y. Hezaveh, L. Levasseur, and P. Marshall, "Fast automated analysis of strong gravitational lenses with convolutional neural networks," *Nature*, vol. 548, no. 7669, p. 555, 2017.

[78] G. Carleo and M. Troyer, "Solving the quantum many-body problem with artificial neural networks," *Science*, vol. 355, no. 6325, pp. 602–606, 2017.

[79] L. Fridman. MIT 6.S094: Deep learning for self-driving cars. [Online]. Available: http://selfdrivingcars.mit.edu/

[80] D. Faggella. Facial recognition applications—security, retail, and beyond. [Online]. Available: https://www.techemergence.com/facial-recognition-applications/

[81] T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda, and Q. Liao, "Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review," *International Journal of Automation and Computing*, pp. 1–17, 2017.

[82] S. Cheung, Y. Lau, Z. Chen, J. Sun, Y. Zhang, J. Wright, and A. Pasupathy, "Beyond the Fourier transform: A machine learning approach to microscopy analysis," 2016, submitted to Nature Communications.