When Nonconvexity Meets Nonsmoothness

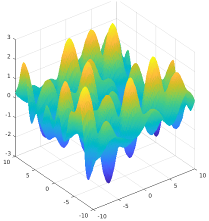
Ju Sun
Department of Mathematics
Stanford University

Nonconvex Statistical Estimation at Allerton Conference 2018

October 4, 2018

Nonscary nonconvex optimization

Many problems in modern **signal processing**, **machine learning**, **statistics**, **imaging**, ..., are most naturally formulated as **nonconvex** optimization problems.

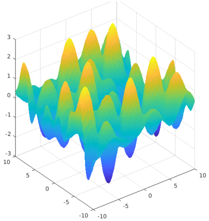


In theory: Even computing a local minimizer is NP-hard!

In practice: Heuristic algorithms are often surprisingly successful.

Nonscary nonconvex optimization

Many problems in modern **signal processing**, **machine learning**, **statistics**, **imaging**, ..., are most naturally formulated as **nonconvex** optimization problems.



In theory: Even computing a local minimizer is NP-hard!

In practice: Heuristic algorithms are often surprisingly successful.

Which nonconvex optimization problems are easy?

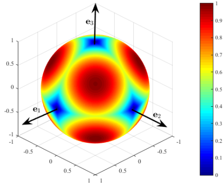
Problems with nice global landscapes

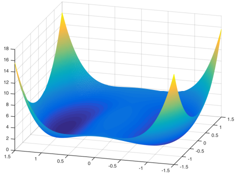
All local mins are global, all saddles are strict

Problems with nice global landscapes

All local mins are global, all saddles are strict

Eigenvalue problems (folklore!)
Sparsifying dictionary learning [Sun et al., 2015]
Generalized phase retrieval [Sun et al., 2016]
Orthogonal tensor decomposition [Ge et al., 2015]
Low-rank matrix recovery and completion
[Ge et al., 2016, Bhojanapalli et al., 2016]
Phase synchronization [Boumal, 2016]
Community detection [Bandeira et al., 2016]
Deep/shallow networks [Kawaguchi, 2016,
Lu and Kawaguchi, 2017, Soltanolkotabi et al., 2017]
Sparse blind deconvolution [Zhang et al., 2017]

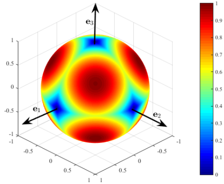


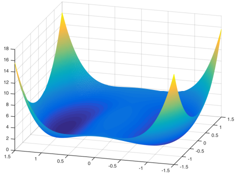


Problems with nice global landscapes

All local mins are global, all saddles are strict

Eigenvalue problems (folklore!)
Sparsifying dictionary learning [Sun et al., 2015]
Generalized phase retrieval [Sun et al., 2016]
Orthogonal tensor decomposition [Ge et al., 2015]
Low-rank matrix recovery and completion
[Ge et al., 2016, Bhojanapalli et al., 2016]
Phase synchronization [Boumal, 2016]
Community detection [Bandeira et al., 2016]
Deep/shallow networks [Kawaguchi, 2016,
Lu and Kawaguchi, 2017, Soltanolkotabi et al., 2017]
Sparse blind deconvolution [Zhang et al., 2017]

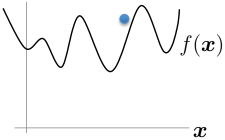




Algorithms: virtually everything reasonable works!

[Conn et al., 2000, Nesterov and Polyak, 2006, Goldfarb, 1980, Jin et al., 2017]

Problems with nice local landscapes



Use problem structure to find a
clever (sometimes random) initial guess.

Analyze iteration-by-iteration
in the vicinity of the optimum.

- **Matrix completion/recovery**: [Keshavan et al., 2010], [Jain et al., 2013], [Hardt, 2014], [Hardt and Wootters, 2014], [Netrapalli et al., 2014], [Jain and Netrapalli, 2014], [Sun and Luo, 2014], [Zheng and Lafferty, 2015], [Tu et al., 2015], [Chen and Wainwright, 2015], [Sa et al., 2015], [Wei et al., 2015]. Also [Jain et al., 2010]
- **Dictionary learning**: [Agarwal et al., 2013a], [Arora et al., 2013], [Agarwal et al., 2013b], [Arora et al., 2015], [Chatterji and Bartlett, 2017], [Gilboa et al., 2018]
- **Tensor recovery**: [Jain and Oh, 2014], [Anandkumar et al., 2014b], [Anandkumar et al., 2014a], [Anandkumar et al., 2015]
- **Phase retrieval**: [Netrapalli et al., 2013], [Candès et al., 2015], [Chen and Candès, 2015], [White et al., 2015], [Wang et al., 2016], [Chen et al., 2018]

Nonscary nonconvex problems

Problems with nice global/local landscapes

- My webpage: `http://sunju.org/research/nonconvex/`
- Jain, Prateek, and Purushottam Kar. **Non-convex optimization for machine learning**. Foundations and Trends® in Machine Learning 10.3–4 (2017): 142–336.
- Chen, Yudong, and Yuejie Chi. **Harnessing structures in big data via guaranteed low-rank matrix estimation**. arXiv preprint arXiv:1802.08397 (2018).
- Chi, Yuejie, Yue M. Lu, and Yuxin Chen. **Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview**. arXiv preprint arXiv:1809.09573 (2018).

Common ingredients in analysis

1st order geometry: ∇f or $\boldsymbol{v}^\top \nabla f$ (directional derivatives)

2nd order geometry: $\nabla^2 f$ or $\boldsymbol{v}^\top \nabla^2 f \boldsymbol{v}$ (directional curvatures)

This talk: **What about nonsmooth, nonconvex problems?**

nonsmooth: may be non-differentiable

Nonsmooth problems are everywhere

Optimization: exact penalty functions

$$\min \ f(\boldsymbol{x}) \ \text{s.t.} \ g_i(x) \leq 0, h_j(x) = 0$$

$$\longrightarrow P(\boldsymbol{x}, c) = f(\boldsymbol{x}) + c\left(\sum_i g_i(x)_+ + \sum_j |h_j(\boldsymbol{x})|\right)$$

Nonsmooth problems are everywhere

Optimization: exact penalty functions

$$\min \ f(\boldsymbol{x}) \ \text{s.t.} \ g_i(x) \leq 0, h_j(x) = 0$$

$$\longrightarrow P(\boldsymbol{x}, c) = f(\boldsymbol{x}) + c \left(\sum_i g_i(x)_+ + \sum_j |h_j(\boldsymbol{x})| \right)$$

Robust estimation:

$$\min \ \|f(\boldsymbol{x}) - \boldsymbol{y}\|_p \qquad f \text{ nonlinear}$$

Nonsmooth problems are everywhere

Optimization: exact penalty functions

$$\min\ f(\boldsymbol{x})\ \text{s.t.}\ g_i(x) \le 0, h_j(x) = 0$$

$$\longrightarrow P(\boldsymbol{x}, c) = f(\boldsymbol{x}) + c\left(\sum_i g_i(x)_+ + \sum_j |h_j(\boldsymbol{x})|\right)$$

Robust estimation:

$$\min\ \|f(\boldsymbol{x}) - \boldsymbol{y}\|_p \qquad f \text{ nonlinear}$$

Promoting structures:

Sparse phase retrieval
Sparse principal component analysis (SPCA)
Sparse blind deconvolution
Neural networks with nonsmooth activations (e.g., ReLU)

Nonsmooth problems are everywhere

Optimization: exact penalty functions

$$\min \ f(\boldsymbol{x}) \ \text{s. t.} \ g_i(x) \le 0, h_j(x) = 0$$
$$\longrightarrow P(\boldsymbol{x}, c) = f(\boldsymbol{x}) + c \left(\sum_i g_i(x)_+ + \sum_j |h_j(\boldsymbol{x})| \right)$$

Robust estimation:

$$\min \ \|f(\boldsymbol{x}) - \boldsymbol{y}\|_p \qquad f \text{ nonlinear}$$

Promoting structures:

- Sparse phase retrieval
- Sparse principal component analysis (SPCA)
- Sparse blind deconvolution
- Neural networks with nonsmooth activations (e.g., ReLU)

Others [Bagirov et al., 2014, Absil and Hosseini, 2017]

Language for nonsmooth functions?

This talk: **What about nonsmooth, nonconvex problems?**

1st order geometry: ∇f or $v^\top \nabla f$ (directional derivatives)

$$\Longrightarrow \textbf{?}$$

2nd order geometry: $\nabla^2 f$ or $v^\top \nabla^2 f v$ (directional curvatures)

$$\Longrightarrow \textbf{?}$$

Locally Lipschitz functions

... functions that are Lipschitz **locally**:

- Continuous convex and concave functions
- Continuously differentiable functions
- Distance function to a set
- Sum of two locally Lipschitz functions: e.g., weakly convex functions ($f\left(\boldsymbol{x}\right)$ so that $f(\boldsymbol{x})+\rho\left\|\boldsymbol{x}\right\|_2^2$ is convex)
- Components of two locally Lipschitz functions: e.g., $h\left(g\left(\boldsymbol{x}\right)\right)$ with h convex and $g \in \mathcal{C}^1$
- Products/Quotients of two locally Lipschitz functions
- ...

Clarke subdifferentials

We restrict to **finite-dimensional** functions, i.e., $f : X \mapsto \mathbb{R}$ with $X \subset \mathbb{R}^n$.

Clarke subdifferentials

We restrict to **finite-dimensional** functions, i.e., $f : X \mapsto \mathbb{R}$ with $X \subset \mathbb{R}^n$.

Rademacher's theorem: If f is locally Lipschitz, f is differentiable almost everywhere.

Definition (Clarke subdifferential [Clarke, 1990])

$$\partial f(\boldsymbol{x}) \doteq \operatorname{conv}\left\{\lim \nabla f(\boldsymbol{x}_k) : \boldsymbol{x}_k \to \boldsymbol{x}, f \text{ diff. at } \boldsymbol{x}_k\right\}$$

Clarke's subdifferential

Definition (Clarke subdifferential [Clarke, 1990])

$$\partial f\left(\boldsymbol{x}\right) \doteq \operatorname{conv}\left\{\lim \nabla f\left(\boldsymbol{x}_k\right) : \boldsymbol{x}_k \to \boldsymbol{x}, f \text{ diff. at } \boldsymbol{x}_k\right\}$$

- $f \in \mathcal{C}^1$, $\partial f\left(\boldsymbol{x}\right) = \left\{\nabla f\left(\boldsymbol{x}\right)\right\}$
- f convex: the usual subdifferential in convex analysis
- Most natural calculus rules hold (under regularity conditions, Chap 2 of [Clarke, 1990])
- Optimality: $\boldsymbol{x}_0$ is local min when $\mathbf{0} \in \partial f\left(\boldsymbol{x}_0\right)$

Language for nonsmooth functions?

This talk: **What about nonsmooth, nonconvex problems?**

Language for nonsmooth functions?

This talk: **What about nonsmooth, nonconvex problems?**

1st order geometry: ∇f or $\boldsymbol{v}^\top \nabla f$ (directional derivatives)
$\implies \partial f$ or $\boldsymbol{v}^\top \partial f$

Language for nonsmooth functions?

This talk: **What about nonsmooth, nonconvex problems?**

1st order geometry: ∇f or $\boldsymbol{v}^\top \nabla f$ (directional derivatives) $\implies \partial f$ or $\boldsymbol{v}^\top \partial f$

2nd order geometry: $\nabla^2 f$ or $\boldsymbol{v}^\top \nabla^2 f \boldsymbol{v}$ (directional curvatures)
$\implies$ motononicity of ∂f: f is convex **iff**

$$\langle \boldsymbol{u_x} - \boldsymbol{u_y}, \boldsymbol{x} - \boldsymbol{y} \rangle \geq 0 \quad \forall\, \boldsymbol{x}, \boldsymbol{y} \text{ and } \boldsymbol{u_x} \in \partial f\left(\boldsymbol{x}\right), \boldsymbol{u_y} \in \partial f\left(\boldsymbol{y}\right).$$

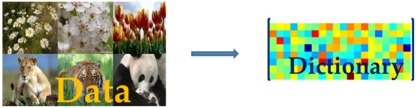
Nonsmoothness in action

Learning sparsifying transformation

Given $\boldsymbol{Y}$, learn $\boldsymbol{Q}$ so that $\boldsymbol{Q}^* \boldsymbol{Y}$ is sparse, i.e., $\|\boldsymbol{Q}^* \boldsymbol{Y}\|_0$ is small.

Nonsmoothness in action

Learning sparsifying transformation



Given $\boldsymbol{Y}$, learn $\boldsymbol{Q}$ so that $\boldsymbol{Q}^*\boldsymbol{Y}$ is sparse, i.e., $\|\boldsymbol{Q}^*\boldsymbol{Y}\|_0$ is small.

To study possibility of **recovery**, given $\boldsymbol{Q}_0$ orthogonal and $\boldsymbol{X}_0$ sparse,

$$\boldsymbol{Y} \quad = \quad \boldsymbol{Q}_0 \times \boldsymbol{X}_0,$$

recover $\boldsymbol{Q}_0$ and $\boldsymbol{X}_0$ (**up to signed permutation & scaling**) .

One element each time

Assume $\boldsymbol{Y} = \boldsymbol{Q}_0 \boldsymbol{X}_0$, $\boldsymbol{Q}_0$ orthogonal

A naive formulation:

$$\min \quad \|\boldsymbol{q}^* \boldsymbol{Y}\|_0 \quad \text{s.t.} \quad \boldsymbol{q} \neq \boldsymbol{0}.$$

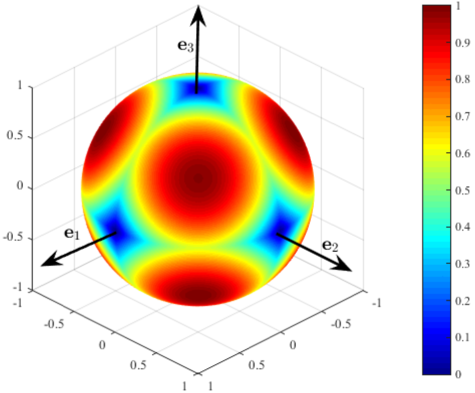
Nonconvex "relaxation":

$$\boxed{\min \quad f(\boldsymbol{q}) \doteq \frac{1}{m} \|\boldsymbol{q}^* \boldsymbol{Y}\|_1 \quad \text{s.t.} \ \|\boldsymbol{q}\|_2^2 = 1.}$$

Many precedents, e.g., [Zibulevsky and Pearlmutter, 2001] in blind source separation. Here, inspired by [Spielman et al., 2012, Sun et al., 2015]

Toward geometric intuition

A low-dimensional example ($n = 3$) of the landscape when the target dictionary $\boldsymbol{Q}_0$ is $\boldsymbol{I}$ and $m \to \infty$



The landscape

$$\boxed{\min \quad f(\boldsymbol{q}) \doteq \frac{1}{m}\|\boldsymbol{q}^*\boldsymbol{Y}\|_1 = \frac{1}{m}\sum_i |\boldsymbol{q}^*\boldsymbol{y}_i| \quad \text{s.t. } \|\boldsymbol{q}\|_2^2 = 1.}$$

The landscape

$$\min \quad f(\boldsymbol{q}) \doteq \frac{1}{m}\|\boldsymbol{q}^*\boldsymbol{Y}\|_1 = \frac{1}{m}\sum_i |\boldsymbol{q}^*\boldsymbol{y}_i| \quad \text{s.t. } \|\boldsymbol{q}\|_2^2 = 1.$$

Riemannian language: $\partial_R f(\boldsymbol{q}) = (\boldsymbol{I} - \boldsymbol{q}\boldsymbol{q}^*)\,\partial f(\boldsymbol{q})$

[Hosseini and Uschmajew, 2017]

The landscape

$$\min \quad f(\boldsymbol{q}) \doteq \frac{1}{m}\|\boldsymbol{q}^*\boldsymbol{Y}\|_1 = \frac{1}{m}\sum_i |\boldsymbol{q}^*\boldsymbol{y}_i| \qquad \text{s.t. } \|\boldsymbol{q}\|_2^2 = 1.$$

Riemannian language: $\partial_R f(\boldsymbol{q}) = (\boldsymbol{I} - \boldsymbol{q}\boldsymbol{q}^*)\,\partial f(\boldsymbol{q})$

[Hosseini and Uschmajew, 2017]

For analysis: Bernoulli-Gaussian model $\boldsymbol{X}_0 = \boldsymbol{\Omega}_0 \circ \boldsymbol{V}_0$, $\boldsymbol{\Omega}_0 \sim_{\text{iid}} \mathrm{Ber}(\theta)$, $\boldsymbol{V}_0 \sim_{\text{iid}} \mathcal{N}(0,1)$. Sparsity parameter θ

The landscape

$$\min \quad f(\boldsymbol{q}) \doteq \frac{1}{m}\|\boldsymbol{q}^*\boldsymbol{Y}\|_1 = \frac{1}{m}\sum_i |\boldsymbol{q}^*\boldsymbol{y}_i| \qquad \text{s.t. } \|\boldsymbol{q}\|_2^2 = 1.$$

Riemannian language: $\partial_R f(\boldsymbol{q}) = (\boldsymbol{I} - \boldsymbol{q}\boldsymbol{q}^*)\,\partial f(\boldsymbol{q})$

[Hosseini and Uschmajew, 2017]

For analysis: Bernoulli-Gaussian model $\boldsymbol{X}_0 = \boldsymbol{\Omega}_0 \circ \boldsymbol{V}_0$, $\boldsymbol{\Omega}_0 \sim_{\text{iid}} \text{Ber}(\theta)$, $\boldsymbol{V}_0 \sim_{\text{iid}} \mathcal{N}(0,1)$. Sparsity parameter θ

When m is large, w.h.p., in a "reasonably large" region of $\boldsymbol{e}_n$:

$$\inf \langle \partial_R f(\boldsymbol{q}), \boldsymbol{q} - \boldsymbol{e}_n \rangle \geq \gamma \|\boldsymbol{q} - \boldsymbol{e}_n\|$$

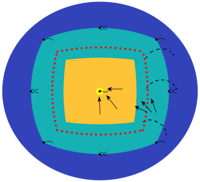


image credit: [Gilboa et al., 2018]

Subgradient descent learns orthogonal dictionaries!

Starting from a $\boldsymbol{q}^{(0)}$ **uniformly random** on $\mathbb{S}^{n-1}$, for $k = 0, 1, 2, \dots$:

$$\boldsymbol{q}^{(k+1)} = \frac{\boldsymbol{q}^{(k)} - \eta^{(k)} \boldsymbol{v}^{(k)}}{\left\| \boldsymbol{q}^{(k)} - \eta^{(k)} \boldsymbol{v}^{(k)} \right\|} \quad \text{for any } \boldsymbol{v} \in \partial_R f\left(\boldsymbol{q}^{(k)}\right)$$

Subgradient descent learns orthogonal dictionaries!

Starting from a $\boldsymbol{q}^{(0)}$ **uniformly random** on $\mathbb{S}^{n-1}$, for $k = 0, 1, 2, \ldots$:

$$\boldsymbol{q}^{(k+1)} = \frac{\boldsymbol{q}^{(k)} - \eta^{(k)} \boldsymbol{v}^{(k)}}{\left\| \boldsymbol{q}^{(k)} - \eta^{(k)} \boldsymbol{v}^{(k)} \right\|} \quad \text{for any } \boldsymbol{v} \in \partial_R f\left(\boldsymbol{q}^{(k)}\right)$$

Ideas:

- Each runs finds an $\boldsymbol{e}_i$ with constant probability
- All basis vectors found in $O(n \log n)$ independent runs

Subgradient descent learns orthogonal dictionaries!

Starting from a $\boldsymbol{q}^{(0)}$ **uniformly random** on $\mathbb{S}^{n-1}$, for $k = 0, 1, 2, \ldots$:

$$\boldsymbol{q}^{(k+1)} = \frac{\boldsymbol{q}^{(k)} - \eta^{(k)}\boldsymbol{v}^{(k)}}{\left\|\boldsymbol{q}^{(k)} - \eta^{(k)}\boldsymbol{v}^{(k)}\right\|} \quad \text{for any } \boldsymbol{v} \in \partial_R f\left(\boldsymbol{q}^{(k)}\right)$$

Ideas:

- Each runs finds an $\boldsymbol{e}_i$ with constant probability
- All basis vectors found in $O(n \log n)$ independent runs

Theorem (Informal, Bai, Jiang, S.'18)

Assume $\theta \in [1/n, 1/2]$. When $m \geq \Omega\left(\theta^{-2} n^4 \log^4 n\right)$, whp, the proposed algorithm pipeline recovers all basis vectors in polynomial time.

Comparison with the DL literature

Algorithms working in the constant sparsity regime, i.e., $\theta \in \Theta(1)$

– **Convex relaxation based on Sum-of-Squares (SOS)**:

[Barak et al., 2015, Ma et al., 2016, Schramm and Steurer, 2017]

solving huge SDP's or tensor decompositions

Comparison with the DL literature

Algorithms working in the constant sparsity regime, i.e., $\theta \in \Theta(1)$

- **Convex relaxation based on Sum-of-Squares (SOS)**: [Barak et al., 2015, Ma et al., 2016, Schramm and Steurer, 2017] solving huge SDP's or tensor decompositions
- **Nonconvex relaxation based on smoothed ℓ_1**: 2nd order method [Sun et al., 2015] or 1st order method [Gilboa et al., 2018], still expensive in computation and involved for analysis

Comparison with the DL literature

Algorithms working in the constant sparsity regime, i.e., $\theta \in \Theta(1)$

- **Convex relaxation based on Sum-of-Squares (SOS)**: [Barak et al., 2015, Ma et al., 2016, Schramm and Steurer, 2017] solving huge SDP's or tensor decompositions
- **Nonconvex relaxation based on smoothed ℓ_1**: 2nd order method [Sun et al., 2015] or 1st order method [Gilboa et al., 2018], still expensive in computation and involved for analysis
- **This work: nonconvex relaxation based directly on ℓ_1**: lightweight computation and neater analysis — compress the smoothed ℓ_1 analysis by $2/3$!

A word on technicalities

Subdifferentials are (convex) sets in general, and randomness in data leads to **random sets**.

- Measure set difference: Hausdorff distance
- Expectation of random sets: selection integrals and support functions [Aubin and Frankowska, 2009, Molchanov, 2013]
- Concentration of Minkowski sum of random sets: support functions and concentration of empirical processes [Molchanov, 2017, Molchanov, 2013]

A word on technicalities

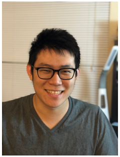
Subdifferentials are (convex) sets in general, and randomness in data leads to **random sets**.

- Measure set difference: Hausdorff distance
- Expectation of random sets: selection integrals and support functions [Aubin and Frankowska, 2009, Molchanov, 2013]
- Concentration of Minkowski sum of random sets: support functions and concentration of empirical processes [Molchanov, 2017, Molchanov, 2013]

The $\mathrm{sign}(\cdot)$ function is not Lipschitz in the usual sense

- Careful construction of the ε-net for covering in showing uniform convergence of the subdifferential

Thanks to ...



Yu Bai

Stanford

Qijia Jiang

Stanford

Emmaneul Candès

Stanford

Thank you!

References i

[Absil and Hosseini, 2017] Absil, P. and Hosseini, S. (2017). **A collection of nonsmooth riemannian optimization problems.**

[Agarwal et al., 2013a] Agarwal, A., Anandkumar, A., Jain, P., Netrapalli, P., and Tandon, R. (2013a). **Learning sparsely used overcomplete dictionaries via alternating minimization.** *arXiv preprint arXiv:1310.7991*.

[Agarwal et al., 2013b] Agarwal, A., Anandkumar, A., and Netrapalli, P. (2013b). **Exact recovery of sparsely used overcomplete dictionaries.** *arXiv preprint arXiv:1309.1952*.

[Anandkumar et al., 2014a] Anandkumar, A., Ge, R., and Janzamin, M. (2014a). **Analyzing tensor power method dynamics: Applications to learning overcomplete latent variable models.** *arXiv preprint arXiv:1411.1488*.

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

[Anandkumar et al., 2015] Anandkumar, A., Jain, P., Shi, Y., and Niranjan, U. N. (2015). **Tensor vs matrix methods: Robust tensor decomposition under block sparse perturbations.** *arXiv preprint arXiv:1510.04747*.

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

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

[Aubin and Frankowska, 2009] Aubin, J.-P. and Frankowska, H. (2009). **Set-valued analysis.** Springer Science & Business Media.

[Bagirov et al., 2014] Bagirov, A., Karmitsa, N., and Mäkelä, M. M. (2014). **Introduction to Nonsmooth Optimization: theory, practice and software.** Springer.

[Bandeira et al., 2016] Bandeira, A. S., Boumal, N., and Voroninski, V. (2016). **On the low-rank approach for semidefinite programs arising in synchronization and community detection.** *arXiv preprint arXiv:1602.04426*.

[Barak et al., 2015] Barak, B., Kelner, J. A., and Steurer, D. (2015). **Dictionary learning and tensor decomposition via the sum-of-squares method.** In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 143–151. ACM.

References iii

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

[Boumal, 2016] Boumal, N. (2016). **Nonconvex phase synchronization.** *arXiv preprint arXiv:1601.06114*.

[Candès et al., 2015] Candès, E. J., Li, X., and Soltanolkotabi, M. (2015). **Phase retrieval via wirtinger flow: Theory and algorithms.** *Information Theory, IEEE Transactions on*, 61(4):1985–2007.

[Chatterji and Bartlett, 2017] Chatterji, N. S. and Bartlett, P. L. (2017). **Alternating minimization for dictionary learning with random initialization.** *arxiv:1711.03634*.

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

[Chen et al., 2018] Chen, Y., Chi, Y., Fan, J., and Ma, C. (2018). **Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval.** *arXiv:1803.07726*.

References iv

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

[Clarke, 1990] Clarke, F. H. (1990). **Optimization and nonsmooth analysis, volume 5.** Siam.

[Conn et al., 2000] Conn, A. R., Gould, N. I. M., and Toint, P. L. (2000). **Trust-region Methods.** Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

[Ge et al., 2015] Ge, R., Huang, F., Jin, C., and Yuan, Y. (2015). **Escaping from saddle points—online stochastic gradient for tensor decomposition.** In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842.

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

[Gilboa et al., 2018] Gilboa, D., Buchanan, S., and Wright, J. (2018). **Efficient dictionary learning with gradient descent.** *arXiv:1809.10313*.

[Goldfarb, 1980] Goldfarb, D. (1980). **Curvilinear path steplength algorithms for minimization which use directions of negative curvature.** *Mathematical programming*, 18(1):31–40.

References v

[Hardt, 2014] Hardt, M. (2014). **Understanding alternating minimization for matrix completion.** In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 651–660. IEEE.

[Hardt and Wootters, 2014] Hardt, M. and Wootters, M. (2014). **Fast matrix completion without the condition number.** In *Proceedings of The 27th Conference on Learning Theory*, pages 638–678.

[Hosseini and Uschmajew, 2017] Hosseini, S. and Uschmajew, A. (2017). **A Riemannian gradient sampling algorithm for nonsmooth optimization on manifolds.** *SIAM Journal on Optimization*, 27(1):173–189.

[Jain et al., 2010] Jain, P., Meka, R., and Dhillon, I. S. (2010). **Guaranteed rank minimization via singular value projection.** In *Advances in Neural Information Processing Systems*, pages 937–945.

[Jain and Netrapalli, 2014] Jain, P. and Netrapalli, P. (2014). **Fast exact matrix completion with finite samples.** *arXiv preprint arXiv:1411.1087*.

[Jain et al., 2013] Jain, P., Netrapalli, P., and Sanghavi, S. (2013). **Low-rank matrix completion using alternating minimization.** In *Proceedings of the forty-fifth annual ACM symposium on Theory of Computing*, pages 665–674. ACM.

References vi

[Jain and Oh, 2014] Jain, P. and Oh, S. (2014). **Provable tensor factorization with missing data.** In *Advances in Neural Information Processing Systems*, pages 1431–1439.

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

[Kawaguchi, 2016] Kawaguchi, K. (2016). **Deep learning without poor local minima.** *arXiv preprint arXiv:1605.07110*.

[Keshavan et al., 2010] Keshavan, R. H., Montanari, A., and Oh, S. (2010). **Matrix completion from a few entries.** *Information Theory, IEEE Transactions on*, 56(6):2980–2998.

[Lu and Kawaguchi, 2017] Lu, H. and Kawaguchi, K. (2017). **Depth creates no bad local minima.** *arXiv preprint arXiv:1702.08580*.

[Ma et al., 2016] Ma, T., Shi, J., and Steurer, D. (2016). **Polynomial-time tensor decompositions with sum-of-squares.**

[Molchanov, 2013] Molchanov, I. (2013). **Foundations of stochastic geometry and theory of random sets.** In *Stochastic Geometry, Spatial Statistics and Random Fields*, pages 1–20. Springer.

[Molchanov, 2017] Molchanov, I. (2017). **Theory of random sets.** Springer-Verlag London, 2 edition.

[Nesterov and Polyak, 2006] Nesterov, Y. and Polyak, B. T. (2006). **Cubic regularization of newton method and its global performance.** *Mathematical Programming*, 108(1):177–205.

[Netrapalli et al., 2013] Netrapalli, P., Jain, P., and Sanghavi, S. (2013). **Phase retrieval using alternating minimization.** In *Advances in Neural Information Processing Systems*, pages 2796–2804.

[Netrapalli et al., 2014] Netrapalli, P., Niranjan, U. N., Sanghavi, S., Anandkumar, A., and Jain, P. (2014). **Non-convex robust PCA.** In *Advances in Neural Information Processing Systems*, pages 1107–1115.

[Sa et al., 2015] Sa, C. D., Re, C., and Olukotun, K. (2015). **Global convergence of stochastic gradient descent for some non-convex matrix problems.** In *The 32nd International Conference on Machine Learning*, volume 37, pages 2332–2341.

[Schramm and Steurer, 2017] Schramm, T. and Steurer, D. (2017). **Fast and robust tensor decomposition with applications to dictionary learning.** *arXiv preprint arXiv:1706.08672*.

References viii

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

[Spielman et al., 2012] Spielman, D. A., Wang, H., and Wright, J. (2012). **Exact recovery of sparsely-used dictionaries.** In *Proceedings of the 25th Annual Conference on Learning Theory*.

[Sun et al., 2015] Sun, J., Qu, Q., and Wright, J. (2015). **Complete dictionary recovery over the sphere.** *arXiv preprint arXiv:1504.06785*.

[Sun et al., 2016] Sun, J., Qu, Q., and Wright, J. (2016). **A geometric analysis of phase retreival.** *arXiv preprint arXiv:1602.06664*.

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

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

[Wang et al., 2016] Wang, G., Giannakis, G. B., and Eldar, Y. C. (2016). **Solving systems of random quadratic equations via truncated amplitude flow.** *arxiv:1605.08285*.

[Wei et al., 2015] Wei, K., Cai, J.-F., Chan, T. F., and Leung, S. (2015). **Guarantees of Riemannian optimization for low rank matrix recovery.** *arXiv preprint arXiv:1511.01562*.

[White et al., 2015] White, C. D., Ward, R., and Sanghavi, S. (2015). **The local convexity of solving quadratic equations.** *arXiv preprint arXiv:1506.07868*.

[Zhang et al., 2017] Zhang, Y., Lau, Y., Kuo, H.-w., Cheung, S., Pasupathy, A., and Wright, J. (2017). **On the global geometry of sphere-constrained sparse blind deconvolution.** In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Zheng and Lafferty, 2015] Zheng, Q. and Lafferty, J. (2015). **A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements.** *arXiv preprint arXiv:1506.06081*.

[Zibulevsky and Pearlmutter, 2001] Zibulevsky, M. and Pearlmutter, B. (2001). **Blind source separation by sparse decomposition in a signal dictionary.** *Neural computation*, 13(4):863–882.