Complete Dictionary Recovery over the Sphere

Ju Sun, Qing Qu, John Wright
Department of Electrical Engineering, Columbia University, New York, USA
Email: {js4038, qq2105, jw2966}@columbia.edu

Dictionary learning (DL) is the problem of finding a sparse representation for a collection of input signals. Its applications span classical image processing, visual recognition, compressive signal acquisition, as well as recent deep architectures for signal classification [1, 2]. Despite many empirical successes, relatively little is known about the theoretical properties of DL algorithms. Typical formulations are nonconvex. Even proving that the target solution is a *local* minimum requires nontrivial analysis [3–7]. Obtaining *global* solutions with *efficient* algorithms is a standing challenge. Suppose that the data matrix $\boldsymbol{Y}$ is generated as $\boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$, where $\boldsymbol{A}_0 \in \mathbb{R}^{n\times m}$ and $\boldsymbol{X}_0 \in \mathbb{R}^{m\times p}$. Existing recovery results pertain only to highly sparse $\boldsymbol{X}_0$. For example, [8] showed that a certain linear programming relaxation can recover a complete ($m = n$) dictionary $\boldsymbol{A}_0$ from $\boldsymbol{Y}$, when $\boldsymbol{X}_0$ is a sparse random matrix with $O(\sqrt{n})$ nonzeros per column. [9, 10] and [11, 12] have subsequently given efficient algorithms for the overcomplete setting ($m \geq n$), based on a combination of initialization and local refinement. These algorithms again succeed when $\boldsymbol{X}_0$ has $\widetilde{O}(\sqrt{n})$ nonzeros[1] per column. [13] gives an efficient algorithm working with $O(n^c)$ nonzeros per column for any $c < 1$. [2]

In this work, we consider the problem of recovering a complete dictionary $\boldsymbol{A}_0$, from $\boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$. We give the first efficient algorithm that provably recovers $\boldsymbol{A}_0$ when $\boldsymbol{X}_0$ has $O(n)$ nonzeros per column. This algorithm is based on nonconvex optimization. Our proofs give a geometric characterization of the high-dimensional objective landscape, which shows that w.h.p. there are no "spurious" local minima. This abstract is based on our recent work [14].

OUR WORK: A GLIMPSE INTO HIGH-DIMENSIONAL GEOMETRY

Since $\boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$, with $\boldsymbol{A}_0$ nonsingular, $\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X}_0)$. The rows of $\boldsymbol{X}_0$ are sparse vectors in the known subspace $\mathrm{row}(\boldsymbol{Y})$. Following [8], we use this fact to first recover the rows of $\boldsymbol{X}_0$, and then recover $\boldsymbol{A}_0$ by solving a system of linear equations. Under suitable probability models on $\boldsymbol{X}_0$, the rows of $\boldsymbol{X}_0$ are the n *sparsest* vectors (directions) in $\mathrm{row}(\boldsymbol{Y})$ [8]. One might attempt to recover them by solving[3]

$$\min \ \|\boldsymbol{q}^*\boldsymbol{Y}\|_0 \ \text{ s.t. } \ \boldsymbol{q} \neq \boldsymbol{0}. \tag{1}$$

This objective is discontinuous, and the domain is an open set. Known convex relaxations [8, 15] break down beyond the aforementioned $\sqrt{n}$ barrier. Instead, we work with a *nonconvex* alternative:

$$\min \ f(\boldsymbol{q}) \doteq \frac{1}{p}\sum_{i=1}^{p} h_\mu(\boldsymbol{q}^*\boldsymbol{y}_i), \ \text{ s.t. } \ \|\boldsymbol{q}\|_2 = 1, \tag{2}$$

where $\boldsymbol{y}_i$ is the i-th column of $\boldsymbol{Y}$. Here $h_\mu(\cdot)$ is a smooth approximation to $|\cdot|$ and μ controls the smoothing level. [4] The spherical constraint is nonconvex.

Despite this nonconvexity, simple descent algorithms for (2) exhibit very striking behavior: on many practical numerical examples, they appear to produce global solutions. To attempt to shed some light on this phenomenon, we analyze their behavior when $\boldsymbol{X}_0$ follows the Bernoulli-Gaussian model: $[\boldsymbol{X}_0]_{ij} = \Omega_{ij}V_{ij}$, with $\Omega_{ij} \sim \mathrm{Ber}(\theta)$ and $V_{ij} \sim \mathcal{N}(0, 1)$. For the moment, suppose that $\boldsymbol{A}_0$ is orthogonal. Fig. 1 plots the landscape of $\mathbb{E}_{\boldsymbol{X}_0}[f(\boldsymbol{q})]$ over $\mathbb{S}^2$. Remarkably, $\mathbb{E}_{\boldsymbol{X}_0}[f(\boldsymbol{q})]$ has no spurious local minima. Every local minimum $\widehat{\boldsymbol{q}}$ produces a row of $\boldsymbol{X}_0$: $\widehat{\boldsymbol{q}}^*\boldsymbol{Y} = \alpha\boldsymbol{e}_i^*\boldsymbol{X}_0$. Moreover, the geometry implies that at any nonoptimal point, there is always at least one direction of descent.

Probabilistic arguments show that this structure persists in high dimensions ($n \geq 3$), with high probability,[5] even when the number of observations p is large *yet finite*. Theorem 1 makes this precise, in the special case of $\boldsymbol{A}_0 = \boldsymbol{I}$. The result characterizes the properties of the reparameterization $g(\boldsymbol{w}) = f(\boldsymbol{q}(\boldsymbol{w}))$ obtained by projecting $\mathbb{S}^{n-1}$ onto the equatorial plane $\boldsymbol{e}_n^\perp$ – see Fig. 1 (center).

Theorem 1: For any $\theta \in (0, 1/2)$ and $\mu < O(\theta n^{-1}, n^{-5/4})$, when $p \geq Cn^3\log(n/\mu\theta)/(\mu^2\theta^2)$ the following hold w.h.p.:

$$\nabla^2 g(\boldsymbol{w}) \succeq \tfrac{1}{\mu}c_\star\theta\boldsymbol{I} \quad \forall\, \boldsymbol{w} \text{ s.t. } \|\boldsymbol{w}\| \leq \tfrac{\mu}{4\sqrt{2}}, \tag{3}$$

$$\frac{\boldsymbol{w}^*\nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} \geq c_\star\theta \quad \forall\, \boldsymbol{w} \text{ s.t. } \tfrac{\mu}{4\sqrt{2}} < \|\boldsymbol{w}\| \leq \tfrac{1}{20\sqrt{5}}, \tag{4}$$

$$\frac{\boldsymbol{w}^*\nabla^2 g(\boldsymbol{w})\boldsymbol{w}}{\|\boldsymbol{w}\|^2} \leq -c_\star\theta \quad \forall\, \boldsymbol{w} \text{ s.t. } \tfrac{1}{20\sqrt{5}} < \|\boldsymbol{w}\| \leq \sqrt{\tfrac{4n-1}{4n}}, \tag{5}$$

for some constant $c_\star > 0$, and $g(\boldsymbol{w})$ has a unique minimizer $\boldsymbol{w}_\star$ over $\{\boldsymbol{w} : \|\boldsymbol{w}\| < \sqrt{\frac{4n-1}{4n}}\}$ and $\boldsymbol{w}_\star$ satisfies

$$\|\boldsymbol{w}_\star - \boldsymbol{0}\| \leq O\left(\frac{\mu}{\theta}\sqrt{\frac{n\log p}{p}}\right). \tag{6}$$

In words, one sees the strongly convex, nonzero gradient, and negative curvature regions successively when moving away from each target solution, and the local (also global) minimizers of $f(\boldsymbol{q})$ are next to the target solutions in their respective symmetric sections. Here θ controls the fraction of nonzeros in $\boldsymbol{X}_0$ in our probability model. Where previous results required $\theta = O(1/\sqrt{n})$, our algorithm succeeds even when $\theta = 1/2 - \varepsilon$. The geometric characterization in Theorem 1 can be extended to general orthobases $\boldsymbol{A}_0$ by a simple rotation, and to general invertible $\boldsymbol{A}_0 \in \mathbb{R}^{n\times n}$ by preconditioning, in conjunction with a perturbation argument.

Although the problem has no spurious local minima, it does have many saddle points (Fig. 1). We describe a Riemannian trust region method (TRM) [16, 17] over the sphere which can escape these saddle points. Using the geometric characterization in Theorem 1, we prove that from any initialization, it converges to a close approximation to the target solution in a polynomial number of steps. Using this algorithm, together with rounding and deflation techniques to obtain all n rows of $\boldsymbol{X}_0$, we obtain a *polynomial-time* algorithm for *complete DL*, working in *linear sparsity* regime. This can be compared to previous analyses, which either demanded much more stringent (sublinear) sparsity assumptions [8–11], or did not provide efficient algorithms [13, 18].

The particular geometry of this problem does not demand any clever initialization, in contrast with most recent approaches to analyzing nonconvex recovery of structured signals [9–11, 18–31]. The geometric approach taken here may apply to these problems as well. Finally, for dictionary learning, the geometry appears to be stable to small noise, allowing almost plug-and-play stability analysis.

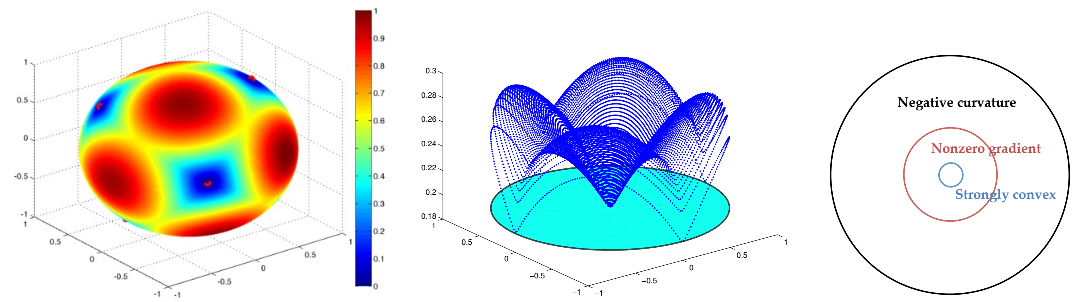

Fig. 1. **Why is dictionary learning over $\mathbb{S}^{n-1}$ tractable?** Assume the target dictionary $\boldsymbol{A}_0$ is orthogonal. Left: Large sample objective function $\mathbb{E}_{\boldsymbol{X}_0}[f(\boldsymbol{q})]$. The only local minima are the columns of $\boldsymbol{A}_0$ and their negatives. Center: the same function, visualized as a height above the plane $\boldsymbol{a}_1^{\perp}$ ($\boldsymbol{a}_1$ is the first column of $\boldsymbol{A}_0$). Around the optimum, the function exhibits a small region of positive curvature, a region of large gradient, and finally a region in which the direction away from $\boldsymbol{a}_1$ is a direction of negative curvature (right).

NOTES

[1]The $\widetilde{O}$ suppresses logarithmic factors.

[2][13] also guarantees recovery with linear sparsity with super-polynomial (quasipolynomial) running time; see also [18].

[3]The notation $*$ denotes matrix transposition.

[4]To be specific, we choose to work with $h_\mu(z) = \mu \log\cosh(z/\mu)$, which is infinitely differentiable.

[5]Here the probability is with respect to the randomness of $\boldsymbol{X}_0$.

ACKNOWLEDGEMENT

This work was partially supported by grants ONR N00014-13-1-0492, NSF 1343282, and funding from the Moore and Sloan Foundations, the Wei Family Private Foundation.

REFERENCES

[1] M. Elad, *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer, 2010.

[2] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *Foundations and Trends in Computer Graphics and Vision*, vol. 8, no. 2-3, 2014.

[3] R. Gribonval and K. Schnass, "Dictionary identification - sparse matrix-factorization via ℓ^1-minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3523–3539, 2010.

[4] Q. Geng and J. Wright, "On the local correctness of ℓ^1-minimization for dictionary learning," Submitted to *IEEE Transactions on Information Theory*, 2011, preprint: http://www.columbia.edu/~jw2966.

[5] K. Schnass, "Local identification of overcomplete dictionaries," *arXiv preprint arXiv:1401.6354*, 2014.

[6] ——, "On the identifiability of overcomplete dictionaries via the minimisation principle underlying k-svd," *Applied and Computational Harmonic Analysis*, vol. 37, no. 3, pp. 464–491, 2014.

[7] ——, "Convergence radius and sample complexity of itkm algorithms for dictionary learning," *arXiv preprint arXiv:1503.07027*, 2015.

[8] D. A. Spielman, H. Wang, and J. Wright, "Exact recovery of sparsely-used dictionaries," in *COLT*, 2012.

[9] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon, "Learning sparsely used overcomplete dictionaries via alternating minimization," *arXiv preprint arXiv:1310.7991*, 2013.

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

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

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

[13] B. Barak, J. A. Kelner, and D. Steurer, "Dictionary learning and tensor decomposition via the sum-of-squares method," *arXiv preprint arXiv:1407.1543*, 2014.

[14] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere," *arXiv preprint arXiv:1504.06785*, 2015.

[15] L. Demanet and P. Hand, "Scaling law for recovering the sparsest element in a subspace," *Information and Inference*, vol. 3, no. 4, pp. 295–309, 2014.

[16] P.-A. Absil, C. G. Baker, and K. A. Gallivan, "Trust-region methods on riemannian manifolds," *Foundations of Computational Mathematics*, vol. 7, no. 3, pp. 303–330, 2007.

[17] P.-A. Absil, R. Mahoney, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

[18] S. Arora, A. Bhaskara, R. Ge, and T. Ma, "More algorithms for provable dictionary learning," *arXiv preprint arXiv:1401.0579*, 2014.

[19] 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.

[20] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *STOC*. ACM, 2013, pp. 665–674.

[21] M. Hardt, "Understanding alternating minimization for matrix completion," in *FOCS*. IEEE, 2014, pp. 651–660.

[22] M. Hardt and M. Wootters, "Fast matrix completion without the condition number," in *COLT*, 2014, pp. 638–678.

[23] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, "Non-convex robust pca," in *NIPS*, 2014, pp. 1107–1115.

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

[25] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *NIPS*, 2013, pp. 2796–2804.

[26] E. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *arXiv preprint arXiv:1407.1065*, 2014.

[27] P. Jain and S. Oh, "Provable tensor factorization with missing data," in *NIPS*, 2014, pp. 1431–1439.

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

[29] Q. Qu, J. Sun, and J. Wright, "Finding a sparse vector in a subspace: Linear sparsity using alternating directions," in *NIPS*, 2014, pp. 3401–3409.

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

[31] K. Lee, Y. Wu, and Y. Bresler, "Near optimal compressed sensing of sparse rank-one matrices via sparse power factorization," *arXiv preprint arXiv:1312.0525*, 2013.