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 \( Y = A_0 X_0 \), where \( A_0 \in \mathbb{R}^{n \times m} \) and \( X_0 \in \mathbb{R}^{m \times p} \). Existing recovery results pertain only to highly sparse \( X_0 \). For example, [8] showed that a certain linear programming relaxation can recover a complete \((m = n)\) dictionary \( A_0 \) when \( X_0 \) 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 \((m > n)\), based on a combination of initialization and local refinement. These algorithms again succeed when \( X_0 \) has \( O(\sqrt{n}) \) nonzeros per column. [13] gives an efficient algorithm working with \( O(n^{2}) \) nonzeros per column for any \( c < 1 \).

In this work, we consider the problem of recovering a complete dictionary \( A_0 \) from \( Y = A_0 X_0 \). We give the first efficient algorithm that provably recovers \( A_0 \) when \( X_0 \) 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 \( Y = A_0 X_0 \), with \( A_0 \) nonsingular, \( \text{row}(Y) = \text{row}(X_0) \). The rows of \( X_0 \) are sparse vectors in the known subspace \( \text{row}(Y) \). Following [8], we use this fact to first recover the rows of \( X_0 \), and then recover \( A_0 \) by solving a system of linear equations. Under suitable probability models on \( X_0 \), the rows of \( X_0 \) are the \( n \) sparsest vectors (directions) in \( \text{row}(Y) \) [8]. One might attempt to recover them by solving

\[
\min \|q^t Y\|_0 \quad \text{s.t. } q \neq 0.
\]  

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:

\[
\min \ f(q) \doteq \frac{1}{p} \sum_{i=1}^{p} h_\mu(q^t y_i), \quad \text{s.t. } \|q\|_2 = 1,
\]

where \( y_i \) is the \( i \)-th column of \( Y \). Here \( h_\mu(\cdot) \) is a smooth approximation to \( |\cdot| \) and \( \mu \) controls the smoothing level. 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 \( X_0 \) follows the Bernoulli-Gaussian model: \( |X_0|_{ij} = \Omega_{ij} V_{ij} \), with \( \Omega_{ij} \sim \text{Ber}(\theta) \) and \( V_{ij} \sim \mathcal{N}(0, 1) \). For the moment, suppose that \( A_0 \) is orthogonal. Fig. 1 plots the landscape of \( \mathbb{E}_{\Omega_{ij}}[f(q)] \) over \( S^2 \). Remarkably, \( \mathbb{E}_{\Omega_{ij}}[f(q)] \) has no spurious local minima. Every local minimum \( \hat{q} \) produces a row of \( X_0: \hat{q}^t Y = \alpha e_i X_0 \). 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, even when the number of observations \( p \) is large yet finite. Theorem 1 makes this precise, in the special case of \( A_0 = I \). The result characterizes the properties of the reparameterization \( g(w) = f(q(w)) \) obtained by projecting \( S^{n-1} \) onto the equatorial plane \( e_1 \) – see Fig. 1 (center).

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

\[
\nabla^2 g(w) \succeq \frac{1}{\mu^2} c_i I \quad \forall \ w \text{ s.t. } \|w\| \leq \frac{\mu}{\sqrt{c_i}}.
\]

\[
\frac{w^* \nabla^2 g(w) w}{\|w\|^2} \succeq c_i \theta \quad \forall \ w \text{ s.t. } \frac{\mu}{\sqrt{c_i}} < \|w\| \leq \frac{1}{2\sqrt{c_i} \theta}.
\]

\[
\frac{w^* \nabla^2 g(w) w}{\|w\|^2} \preceq -c_i \theta \quad \forall \ w \text{ s.t. } \frac{1}{2\sqrt{c_i} \theta} < \|w\| \leq \frac{1}{2n^{3/2} - 1},
\]

for some constant \( c_i > 0 \), and \( g(w) \) has a unique minimizer \( w^* \), over \( \{w: \|w\| < \frac{2n^{-1/4}}{4n^{3/2} - 1}\} \) and \( w^* \) satisfies

\[
\|w^* - 0\| \leq O\left(\frac{n \log n}{\mu} \right).
\]

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 \( f(q) \) are next to the target solutions in their respective symmetric sections. Here \( \theta \) controls the fraction of nonzeros in \( X_0 \) in our probability model. Where previous results required \( \theta = O(1/\sqrt{n}) \), our algorithm succeeds even when \( \theta = 1/2 - \epsilon \). The geometric characterization in Theorem 1 can be extended to general orthobases \( A_0 \) by a simple rotation, and to general invertible \( A_0 \in \mathbb{R}^{n \times n} \) 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 \( X_0 \), 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.
NOTES

1. The $\tilde{O}$ suppresses logarithmic factors.
2. [13] also guarantees recovery with linear sparsity with super-polynomial (quasipolynomial) running time; see also [18].
3. The notation $\ast$ denotes matrix transposition.
4. To be specific, we choose to work with $h_a(z) = \mu \log \cosh(z/\mu)$, which is infinitely differentiable.
5. Here the probability is with respect to the randomness of $X_0$. 

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


Fig. 1. Why is dictionary learning over $\mathbb{R}^{n \times 1}$ tractable? Assume the target dictionary $A_0$ is orthogonal. Left: Large sample objective function $E_{X_0} [f(q)]$. The only local minima are the columns of $A_0$ and their negatives. Center: the same function, visualized as a height above the plane $a_1^\ast$ ($a_1$ is the first column of $A_0$). 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 $a_1$ is a direction of negative curvature (right).