Notes on *Complete Dictionary Recovery over the Sphere*

Bora Basa
STAT 525 final project, May 7, 2022

Preliminary remarks This review is based on Refs. [1, 2]. This is a two-part paper introducing an efficient algorithm for (complete) dictionary learning over S^{n-1}. While a practical algorithm is presented in [2], the main focus is on proving that sub-exponential time solutions to the complete dictionary recovery problems are feasible, owing to the geometric peculiarity of the non-convex problem. Since I was only assigned Ref. [1], this is what I will focus on. I will also assume the dictionaries of interest are not just complete but orthogonal matrices. The appropriate relaxation of this constraint is built upon the theory developed for the orthogonal case and is somewhat technical - I opted to omit this. I also will not be reviewing the proofs of the main theorems as they are rather cumbersome and unrelated to the class.

I. THE DICTIONARY LEARNING PROBLEM

The dictionary learning problem seeks a sparse representation of a given input dataset as a linear combination of so called *dictionary entries*. As an optimization problem, it can be formulated as follows:

Let $Y := [y_1, ..., y_p]$ be a dataset of p features where the feature vector, y_i, is valued in some $\mathbb{R}^d$. Let $D \in \mathbb{R}^{d\times n}$ be the dictionary and $R := [r_1, ..., r_p]$ be a collection of n-dimenstional representation vectors. The problem objective is to find a D and R pair such that $\|Y - DR\|_F$ is minimized with a sufficiently sparse R. That is,

$$\arg\min_{D,R} \|Y - DR\|_F^2 + \lambda\|R\|_1 \tag{1}$$

where the dictionary, D, is valued in some suitable space of matrices, $\mathcal{D}$. In general, this problem is not jointly convex in both D and R, even if the two sub-problems are independently convex. This can be seen through the degeneracy of the objective function under

$$(D, R) \sim (D\Pi\Sigma, \Sigma^{-1}\Pi^T R)$$

where Π is any permutation matrix and Σ is any diagonal matrix with entries valued in $\{\pm 1\}$. This is trivial to check:

$$\begin{aligned}\|Y - DR\|_F^2 + \lambda\|R\|_1 &\mapsto \|Y - D(\Pi\Sigma)(\Sigma^{-1}\Pi^T)R\| + \lambda\|(\Sigma^{-1}\Pi^T)R\|_1 \\ &= \|Y - D\underbrace{\Pi\Pi^T}_{\mathbb{I}}R\| + \lambda\underbrace{\|(\Sigma^{-1}\Pi^T)R\|_1}_{\|R\|_1}.\end{aligned}$$

Due to the combinatorial nature of the degeneracy, the global minima are expected to be isolated, making it difficult to choose a suitable convex surrogate. Beyond this degeneracy, the choice of $\mathcal{D}$ is typically non-convex, but this non-convexity is less of a theoretical challenge.

In the remainder we will fix $\mathcal{D}$ to be the group of orthogonal $n \times n$ matrices and consider invertible dictionaries and walk through the arguments of Ref. [1]. The problem tackled there, and reviewed here, is coming up with a solid theoretical framework for obtaining global solutions to the following non-convex problem:

$$\arg\min_{D,R} \frac{1}{2}\|Y - DR\|_F^2 + \lambda\|R\|_1, \quad \text{s.t. } D \in O_n \tag{2}$$

II. HEURISTIC BLOCK COORDINATE DESCENT

While Eq. (2) is seemingly intractable given the severity of the non-convexity, the unreasonable effectiveness of a coordinate descent algorithm indicates that this may not be the case. The idea is simply to exploit the fact that the optimization problem is convex in either D or R (but not both). Hence, one can fix one variable and take a gradient descent step in the other and then alternate to the other variable. These steps end up being

$$R_k = S_\lambda[D_{k-1}^T Y], \quad D_k = UV^T \tag{3}$$

where U and V are obtained via $SVD(YD^T)$. We have little a priori reason to expect this algorithm to converge to a global minimum because it essentially ignores the joint non-convexity of the objective in D and R. The hallmark of an objective function having degnerate, isolated minima is the dependence of the output on initialization. A quick numerical experiment demonstrates that the optimal values obtained by the block coordinate descent algorithm are robust. Following the authors, we determine D^*, R^* for uncompressed image data for a number of different random realizations of D over the uniform orthogonal ensemble. The metric they track is $\|D^T Y\|_1$ after convergence. We plot in Fig. 1 a reproduction of their results. Curiously, in my implementation we see a spurious value of the metric for a single choice of random initial dictionary at realization 39. This is not present in the examples considered by the authors. It is nonetheless remarkable that this algorithm converges to a single optimal value with very high probability. As per the insight of the authors, it is so remarkable, in fact, that there must be more to the story. Before moving on, let us stress that what we are learning is a set of bases for the 8×8 blocks of the image. The two optimizers are valued in equivalence classes of dictionaries, D, and sparse representation vectors, R, with the relation being permutation, scaling and sign flips.

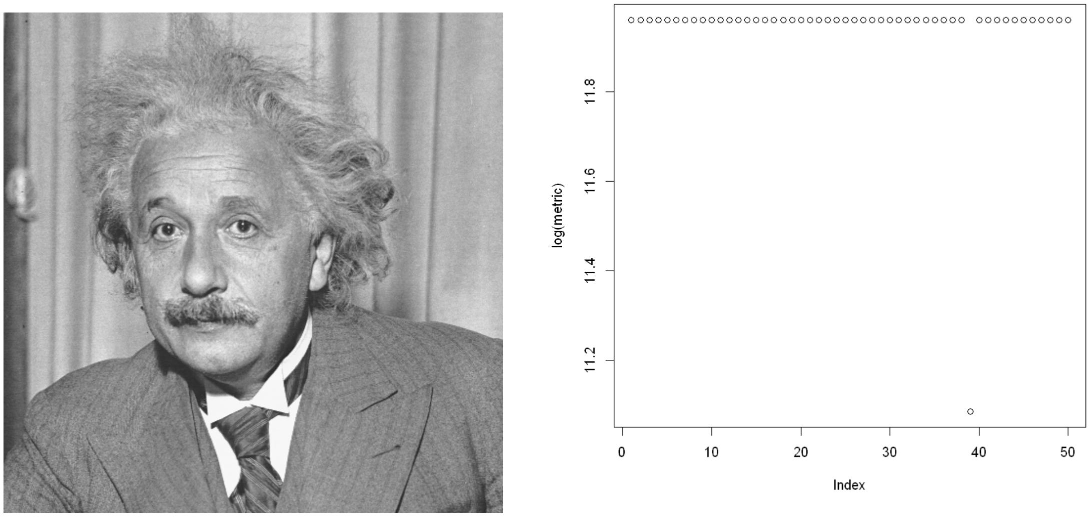

FIG. 1. The image uncompressed image of Einstein is assembled into a Y matrix of dimension (64×4096) by stacking vectorized 8×8 non-overlapping blocks. After the optimizer is ran for 100 steps, the converged value of $\|D^T Y\|_1$ is plotted for 50 different initial realizations of the dictionary, D.

III. TRACTABILITY OF THE DICTIONARY LEARNING PROBLEM

III.1. Reformulation

The dictionary learning problem can be complete, undercomplete or overcomplete. Further, the *sparsity* of the representation vector, R, is somewhat arbitrary in that one has to quantify what it means to be sufficiently sparse. The authors fix a reasonably structured dictionary reconstruction problem by requiring that the dictionary be complete, and R follow a Bernoulli-Gaussian distribution. That is,

$$R_{ij} = \Omega_{ij} V_{ij}, \quad V_{ij} \sim \mathcal{N}(0,1), \; \Omega_{ij} \sim Ber(\theta)$$

Then, θ controls the sparsity of R. The distribution for V is chosen to be normal purely out of the universality of the Gaussian distribution in practical applications.

The fact that the dictionary is complete allows for a manifestly non-convex reformulation of the learning problem. Let $Y = DR$. From completeness, it follows immediately that Y and its dictionary, D, share row spaces. Then, the rows of R, the representation vectors, live in a linear subspace of the row space of Y. Using the fact that R follows a Gaussian Bernoulli distribution, one can conclude that the rows of R correspond to the $n-$sparsest directions in the row space of Y with high probability, if $p \geq n \log n$ [3].

One can parameterize a vector in the row space of Y as $q^T Y$. Then, the sparsest vector in Y may be found via

$$\min_{q \neq 0} \|q^T Y\|_0.$$

Then, one has a recursive way of enumerating the rows of R, from which one can recover the dictionary by solving $Y = DR$. Note that the scale ambiguity allows for the replacement $q \neq 0 \to \|q\|_2 = 1$, which is much more manageable (but is non-convex). The l^0 norm is still a concern, however. The authors opt to solve smooth surrogate problem,

$$\min \underbrace{\frac{1}{p} \sum_{k=1}^{p} h_\mu(q^T \hat{y}_k)}_{f(q)}, \quad \|q\|_2 = 1 \tag{4}$$

where $\hat{y}_k \in$ is a column of the proxy $\hat{Y} \in \mathbb{R}^{n \times p}$ and

$$h_\mu(x) = \mu \log \cosh(x/\mu)$$

is a smoothening of the l^1 surrogate to l^0. As with the motivating example above, this non-convex problem can be solved surprisingly robustly by gradient descent methods.

III.2. Geometry

Most of Ref. [1] is dedicated to motivating the fact that all local minima obtained by solving the surrogate non-convex minimization problem aid us in solving the dictionary reconstruction problem because they all approximate a row in R. Provided we have a good way of obtaining these minimizers, the only real challenge becomes escaping the saddles of the objective function. This is largely the subject of Ref. [2].

The non-convex problem defined by Eq. (4) is geometrized by writing the constraint as $q \in \mathbb{S}^{n-1}$. Let us write $D = \mathbb{I}$ for now. Notice that if $\mathcal{D}$ is the space of orthogonal matrices, there is no loss of generality by this choice: we are simply picking a representative from a family of equivalent problems on the sphere related by rotation. The object of interest, then is $\mathbb{E}_{Y=R}[f(q)]$ on S^{n-1}.

Let

$$q(w) = (w, \sqrt{1 - \|w\|_2^2}), \quad w \in \mathbb{B}^{n-1}$$

be the parameterization of f over a domain of S^{n-1}. Specifically, $f(q(w))$ is a function over an equatorial section, $\text{span}(e_1, ..., e_{n-1}) \cap \mathbb{B}^n$. Here $\{e_i\}$ is a basis for $\mathbb{R}^n$. One can imagine doing this for all equatorial sections defined to be normal to $\{\pm e_i\}_{i=1}^n$. In fact, the authors propose a finer grained subset of w,

$$\Gamma = \left\{ w : \|w\|_2^2 < \frac{4n-1}{4n} \right\}, \tag{5}$$

rather than the projection onto the entire equatorial section. This region can be thought of as a neighborhood of $x \in S^{N-1}$ where the normal vector at x is parallel to one of $\{e_i\}$. The key insight of the authors is that this point is not only a local minimum with high probability $\forall n$, provided sufficient sparsity, but also that these are the only kinds of minima present. Further, the structure of the geometry in an $\varepsilon_{1,2,3}-$neighborhoods of an optimal point have (1) negative curvature, (2) a gradient that increases as $q^\star$ is approached and (3) strong convexity. Here, $\varepsilon_1 > \varepsilon_2 > \varepsilon_3$. An immediate consequence of this structure is that for any $q^\star \neq q \in S^{n-1}$ there exists a ‘correct’ descent direction taking one close to one of these local minima.

The final ingredient is the realization that each of these $2n$[4] minimizers approximately reproduce particular rows of R via $q^\star Y$. This links the sparsity of R with the dimension of the geometric problem. This is a relation the authors make explicit.

Before formalizing the ideas reviewed so far within a probably approximately correct framework, let us reproduce the visualizations for the case of $n = 3$.

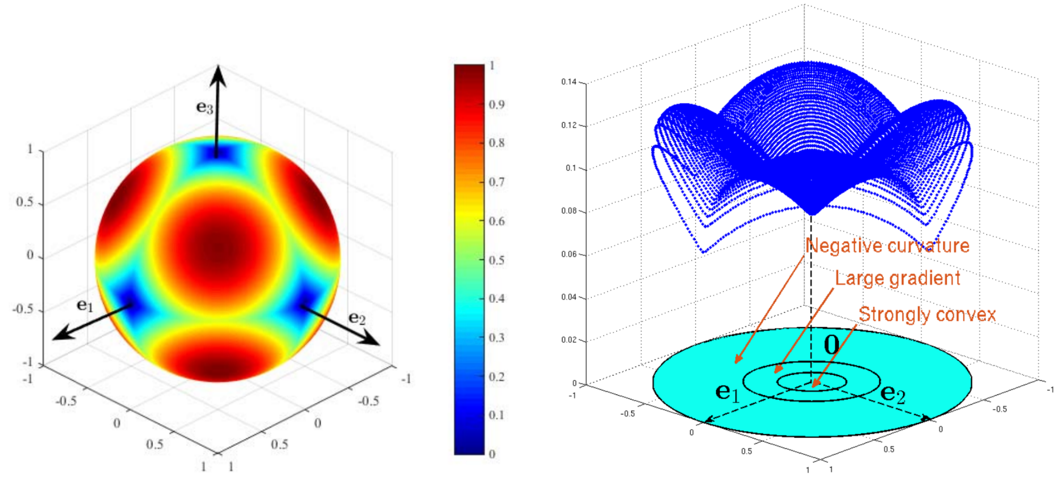

FIG. 2. Reproduced from Ref. [1]. (Left) Heat map of $\mathbb{E}_R[f(q(w))]$ with $q \in S^2$ over 100000 samples. Here $D \in O_3$ and R is a Bernoulli-Gaussian process with $\theta = 0.1$ Also plotted are the three canonical basis vectors for $\mathbb{R}^3$. Notice that the objective achieves its minimum at the intersection of the sphere and the basis rays. There are clearly 6 of these local minima. Notice also the structure of saddles as this will be relevant later. (Right) ε-neighborhood of any $q^\star$ where ε is large enough to capture a region of strong convexity around the critical point, a region of large gradient and a region of negative curvature. The plotted domain is the third equatorial section: $\mathrm{span}(e_1, e_2) \cap \mathbb{B}^3$

Formalizing the low-dimensional geometric intuition

The main result of Ref. [1] is formalizing the geometric intuition depicted in Fig. 2. To this end, define the following three quantities:

$$G_1 = \frac{1}{\|w\|^2}\langle w, \nabla^2 f(q(w))w\rangle, \quad \text{Curvature} \tag{6}$$

$$G_2 := \frac{1}{\|w\|}\langle w, \nabla f(q(w))\rangle, \quad \text{Gradient} \tag{7}$$

$$G_3 := \nabla^2 g(w), \quad \text{Convexity} \tag{8}$$

Fig. 2(right) indicates three regions discussed previously. The $n-1$ dimensional analog of these regions are characterized by

- $G_1(w) < -c_1$ in a ε_1-neighborhood of $q^\star$
- $G_2(w) > c_2$ in a ε_2-neighborhood of $q^\star$
- $G_3(w) > c_3$ in a ε_3-neighborhood of $q^\star$

for positive constants $c_{1,2,3}$. Through a fairly laborious process, one can actually parameterize these constants as well as $\varepsilon_{1,2,3}$ with the relations above understood as concentration inequalities.

Theorem 1 (Sun, Qu and Wright) *Let D be an orthogonal dictionary. Then, $D = \mathbb{I}$ w.l.o.g. and hence $Y = R$. Then, $\exists c_{\star,a,b}, C \in \mathbb{R}_+$ s.t. $\forall \theta \in [0, 1/2]$ and $\mu < c_a \min(\theta/n, n^{-5/4})$, whenever*

$$p \geq \frac{Cn^3}{\mu^2\theta^2} \log \frac{n}{\mu\theta}$$

the following hold with probability at least $1 - c_b p^{-6}$:

- $G_1(w) < -c_\star\theta$ *if* $\frac{1}{20\sqrt{5}} \leq \|w\| \leq \sqrt{\frac{4n-1}{4n}}$
- $G_2(w) \geq c_\star\theta$ *if* $\frac{\mu}{4\sqrt{2}} \leq \|w\| \leq \frac{1}{20\sqrt{5}}$
- $G_3(w) \geq \frac{c_\star\theta}{\mu}\mathbb{I}$ *if* $\|w\| \leq \frac{\mu}{4\sqrt{2}}$

It then follows that, one can find around each patch of S^{N-1}, one can find a favorable local geometry (a generalization of Fig. 2(right)) that has a single descent direction that takes us to $q^\star$. That is, for each $f\big|_{\Gamma_{i\in[1,2n]}}$ centered around a Euclidean basis ray, there exists a unique minimizer. Hence, there are $(2n)$ local minimizers over the whole sphere (with high probability).

The upshot of this result is of course that one can devise a polynomial time algorithm that exploits the geometric structure near a local minimum to descend to it uniquely. The result is a row of R.

$$r_i = q_i^{\star T} Y, \quad r_i \in \mathrm{row}(R).$$

The recovery of all rows will be outlined in the next section. It is worth noting that this is the first proposal of the theoretical feasibility of a non-exponential time algorithm for the DL learning problem.

Saddles and Riemannian trust region algorithm

While the local geometry around $q_i^\star \in S^{N-1}$ is favorable for descent algorithms, there is still the matter of escaping saddle points. The saddles that appear, however, are non-generic: The Hessian, with high probability, has a single negative eigenvalue. Such saddle points (called rideable saddles by the authors) can be escaped with second-order algorithms (or stochastic first-order algorithms). Notice that the challenge of escaping saddles is much less severe than the challenges that are present in the absence of the geometric insight of the authors.

At this point we need to start referring to Ref. [2]. Recall that the quadratic approximation to a function $f : (M, g) \to \mathbb{R}$ in the direction $\delta \in T_q M$ is

$$\hat{f}(q; \delta) \sim f(q) + \langle \nabla f, \delta \rangle + \frac{1}{2}\langle \delta, \nabla^2 \delta \rangle$$

where by ∇ we really mean the connection on (M, g) and $\nabla^2 := \nabla \circ d$ where d is the exterior derivative. We will also need the explicit form of the exponential map, $\exp_q : T_q S^{N-1} \to S^{N-1}$,

$$\exp_q(\delta) = q \cos\|\delta\| + \frac{\delta}{\|\delta\|} \sin\|\delta\|. \tag{9}$$

Given that any non-optimal point has a descent direction that leads to a local minimum, one can escape rideable saddles while approaching $q^\star$ by the descent algorithm

$$q^k = \exp_{q^{k-1}}(\delta^\star) \tag{10}$$

$$\delta^\star = \arg \min_{\delta \in T_q^{k-1} S^{n-1}, \|\delta\| \leq \Delta} \hat{f}(q; \delta) \tag{11}$$

where Δ is the radius of the ball centered around the current iterate. This second order descent method is referred to as the Riemannian trust-region algorithm. Specifically, Eq. (11) is a trust region sub-problem with a variety of approximation schemes available for its efficient solution. Assuming this sub-problem is solved sufficiently accurately, $q^k \to q^\star$ as $k \to \infty$ with high probability.

Theorem 2 (Riemannian trus region method convergence) *Let D be an orthogonal dictionary. Then, $\exists C \in \mathbb{R}_+$ s.t. $\forall \theta \in [0, 1/2]$ and $\mu \leq c_a \min(\theta/n, n^{-5/4})$, whenever*

$$p \geq \frac{Cn^3}{\mu^2\theta^2} \log \frac{n}{\mu\theta},$$

with probability at least $1 - c_b p^{-6}$, the algorithm given in Eq. (10) converge to one $q^\star$ in some fixed number of iterations given a step size satisfying

$$\Delta \leq \frac{c_\star \theta^3 \mu^2}{n^{7/2} \log^{7/2}(np)}.$$

While proving this result in detail is rather non-trivial, it should be clear that the convergence is in essence guaranteed by the three regions (negative curvature, large gradient, strongly convex) outlined in Theorem 1. When the iterate q is in the strongly convex region, the trust algorithm essentially behaves like Newton's algorithm. Away from the strongly convex region, the geometry is either strongly sloped towards the convex region or the curvature is negative. The step in Eq. (10) is guaranteed to reduce $\hat{f}$ in either scenario by construction.

IV. A POLYNOMIAL TIME DL ALGORITHM FOR ORTHOGONAL DICTIONARIES

We now put everything together and state an algorithm that (approximately) solves the orthogonal dictionary recovery problem. As previously stated, the computational backbone is the trust region algorithm. What remains to be fleshed out is how to construct all rows of R given an algorithm that provides one row per successful convergence.

1. Iteratively solve Eq. (10) and (11) to obtain $\hat{q}$ s.t. $\|\hat{q} - q^\star\|_2 \leq \epsilon$ where ϵ is a preset error tolerance.
2. Obtain a row of R by computing $q_\star^T Y$
3. Once l rows of R are obtained, take basis U for $\text{span}(q_\star^1, ..., q_\star^l)$ and solve $\min f(Uq)$ using step 1.
4. Once R is obtained recursively, solve $D = YR^T(RR^T)^{-1}$ to recover the dictionary.

[1] J. Sun, Q. Qu, and J. Wright, IEEE Transactions on Information Theory **63**, 853 (2017).
[2] J. Sun, Q. Qu, and J. Wright, IEEE Transactions on Information Theory **63**, 885 (2017).
[3] D. A. Spielman, H. Wang, and J. Wright, in *Proceedings of the 25th Annual Conference on Learning Theory*, Proceedings of Machine Learning Research, Vol. 23, edited by S. Mannor, N. Srebro, and R. C. Williamson (PMLR, Edinburgh, Scotland, 2012) pp. 37.1–37.18.
[4] To each element of $\{e_i\}_{i=1}^n$ associate a minumum. By the Borsuk-Ulam theorem there are $2n$ points where the objective achieves its local minima.