Efficient Dictionary Learning with Gradient Descent

Dar Gilboa [1 2] **Sam Buchanan** [3 2] **John Wright** [3 2]

Abstract

Randomly initialized first-order optimization algorithms are the method of choice for solving many high-dimensional nonconvex problems in machine learning, yet general theoretical guarantees cannot rule out convergence to critical points of poor objective value. For some highly structured nonconvex problems however, the success of gradient descent can be understood by studying the geometry of the objective. We study one such problem – complete orthogonal dictionary learning, and provide converge guarantees for randomly initialized gradient descent to the neighborhood of a global optimum. The resulting rates scale as low order polynomials in the dimension even though the objective possesses an exponential number of saddle points. This efficient convergence can be viewed as a consequence of negative curvature normal to the stable manifolds associated with saddle points, and we provide evidence that this feature is shared by other nonconvex problems of importance as well.

1. Introduction

Many central problems in machine learning and signal processing are most naturally formulated as optimization problems. These problems are often both nonconvex and high-dimensional. High dimensionality makes the evaluation of second-order information prohibitively expensive, and thus randomly initialized first-order methods are usually employed instead. This has prompted great interest in recent years in understanding the behavior of gradient descent on nonconvex objectives (Hardt et al., 2015; Ge et al., 2015; Hardt et al., 2016; Dauphin et al., 2014). General analysis of first- and second-order methods on such problems can provide guarantees for convergence to critical points but these may be highly suboptimal, since nonconvex optimization is in general an NP-hard probem (Bertsekas, 1999). Outside of a convex setting (Nesterov, 2013) one must assume additional structure in order to make statements about convergence to optimal or high quality solutions. It is a curious fact that for certain classes of problems such as ones that involve sparsification (Lee et al., 2013; Bronstein et al., 2005) or matrix/tensor recovery (Keshavan et al., 2010; Jain et al., 2013; Anandkumar et al., 2014) first-order methods can be used effectively. Even for some highly nonconvex problems where there is no ground truth available such as the training of neural networks first-order methods converge to high-quality solutions (Zhang et al., 2016).

Dictionary learning is a problem of inferring a sparse representation of data that was originally developed in the neuroscience literature (Olshausen & Field, 1996), and has since seen a number of important applications including image denoising, compressive signal acquisition and signal classification (Elad & Aharon, 2006; Mairal et al., 2014). In this work we study a formulation of the dictionary learning problem that can be solved efficiently using randomly initialized gradient descent despite possessing a number of saddle points exponential in the dimension. A feature that appears to enable efficient optimization is the existence of sufficient negative curvature in the directions normal to the stable manifolds of all critical points that are not global minima [1]. This property ensures that the regions of the space that feed into small gradient regions under gradient flow do not dominate the parameter space. Figure 1 illustrates the value of this property: negative curvature prevents measure from concentrating about the stable manifold. As a consequence randomly initialized gradient methods avoid the "slow region" of around the saddle point.

The main results of this work is a convergence rate for randomly initialized gradient descent for complete orthogonal dictionary learning to the neighborhood of a global minimum of the objective. Our results are probabilistic since they rely on initialization in certain regions of the parameter space, yet they allow one to flexibly trade off between the maximal number of iterations in the bound and the probabil-

[1]Department of Neuroscience, Columbia University [2]Data Science Institute, Columbia University [3]Department of Electrical Engineering, Columbia University. Correspondence to: Dar Gilboa <dargilboa@gmail.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

[1] As well as a lack of spurious local minimizers, and the existence of large gradients or strong convexity in the remaining parts of the space

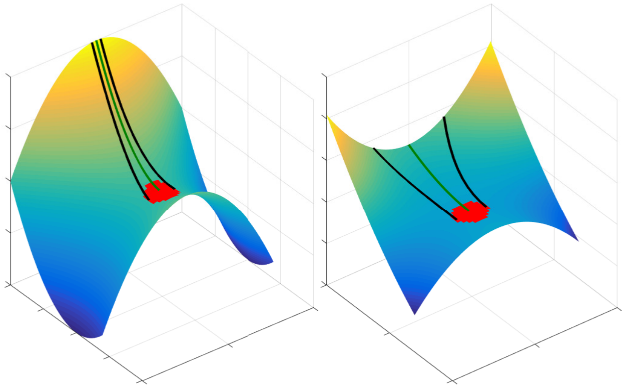

Figure 1. **Negative curvature helps gradient descent.** Red: "slow region" of small gradient around a saddle point. Green: stable manifold associated with the saddle point. Black: points that flow to the slow region. Left: global negative curvature normal to the stable manifold. Right: positive curvature normal to the stable manifold – randomly initialized gradient descent is more likely to encounter the slow region.

ity of the bound holding. The sample complexity required for the concentration results in the paper to hold with high probability is $p \in \mathcal{O}(n^4)$ up to polylog(n) factors, where p is the number of samples and n is the dimensionality of the space.

While our focus is on dictionary learning, it has been recently shown that for other important nonconvex problems such as phase retrieval (Chen et al., 2018) performance guarantees for randomly initialized gradient descent can be obtained as well. In fact, in Appendix C we show that negative curvature normal to the stable manifolds of saddle points (illustrated in Figure 1) is also a feature of the population objective of generalized phase retrieval, and can be used to obtain an efficient convergence rate.

2. Related Work

Easy nonconvex problems. There are two basic impediments to solving nonconvex problems globally: **(i) spurious local minimizers**, and **(ii) flat saddle points**, which can cause methods to stagnate in the vicinity of critical points that are not minimizers. The latter difficulty has motivated the study of *strict saddle functions* (Sun et al., 2015b; Ge et al., 2015), which have the property that at every point in the domain of optimization, there is a large gradient, a direction of strict negative curvature, or the function is strongly convex. By leveraging this curvature information, it is possible to escape saddle points and obtain a local minimizer in polynomial time.[2] Perhaps more surprisingly, many known strict saddle functions also have the property that every local minimizer is global; for these problems, this implies that efficient methods find global solutions. Examples of problems with this property include variants of sparse dictionary learning (Sun et al., 2017), phase retrieval (Sun et al., 2016), tensor decomposition (Ge et al., 2015), community detection (Bandeira et al., 2016) and phase synchronization (Boumal, 2016).

Minimizing strict saddle functions. Strict saddle functions have the property that at every saddle point there is a direction of strict negative curvature. A natural approach to escape such saddle points is to use second order methods (e.g., trust region (Conn et al., 2000) or curvilinear search (Goldfarb, 1980)) that explicitly leverage curvature information. Alternatively, one can attempt to escape saddle points using first order information only. However, some care is needed: canonical first order methods such as gradient descent will not obtain minimizers if initialized at a saddle point (or at a point that flows to one) – at any critical point, gradient descent simply stops. A natural remedy is to randomly perturb the iterate whenever needed. A line of recent works shows that noisy gradient methods of this form efficiently optimize strict saddle functions (Lee et al., 2016; Du et al., 2017; Jin et al., 2017). For example, (Jin et al., 2017) obtains rates on strict saddle functions that match the optimal rates for smooth convex programs up to a polylogarithmic dependence on dimension.[3]

Randomly initialized gradient descent? The aforementioned results are broad, and nearly optimal. Nevertheless, important questions about the behavior of first order methods for nonconvex optimization remain unanswered. For example: *in every one of the aforemented benign nonconvex optimization problems, randomly initialized gradient descent rapidly obtains a minimizer.* This may seem unsurprising: general considerations indicate that the stable manifolds associated with non-minimizing critical points have measure zero (Nicolaescu, 2011), this implies that a variety of small-stepping first order methods converge to minimizers in the large-time limit (Lee et al., 2017). However, it is not difficult to construct strict saddle problems that *are not* amenable to efficient optimization by randomly initialized gradient descent – see (Du et al., 2017) for an example. This contrast between the excellent empirical performance of randomly initialized first order methods and worst case examples suggests that there are important geometric and/or topological properties of "easy nonconvex problems" that are not captured by the strict saddle hypothesis. Hence, the motivation of this paper is twofold: (i) to provide theoretical corroboration (in certain specific situa-

[2] This statement is nontrivial: finding a local minimum of a smooth function is NP-hard.

[3] This work also proves convergence to a second-order stationary point under more general smoothness assumptions.

tions) for what is arguably the simplest, most natural, and most widely used first order method, and (ii) to contribute to the ongoing effort to identify conditions which make nonconvex problems amenable to efficient optimization.

3. Dictionary Learning over the Sphere

Suppose we are given data matrix $\boldsymbol{Y} = [\boldsymbol{y}_1, \dots \boldsymbol{y}_p] \in \mathbb{R}^{n \times p}$. The *dictionary learning* problem asks us to find a concise representation of the data (Elad & Aharon, 2006), of the form $\boldsymbol{Y} \approx \boldsymbol{A}\boldsymbol{X}$, where $\boldsymbol{X}$ is a sparse matrix. In the *complete, orthogonal dictionary learning* problem, we restrict the matrix $\boldsymbol{A}$ to have orthonormal columns ($\boldsymbol{A} \in O(n)$). This variation of dictionary learning is useful for finding concise representations of small datasets (e.g., patches from a single image, in MRI (Ravishankar & Bresler, 2011)).

To analyze the behavior of dictionary learning algorithms theoretically, it useful to posit that $\boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$ for some true dictionary $\boldsymbol{A}_0 \in O(n)$ and sparse coefficient matrix $\boldsymbol{X}_0 \in \mathbb{R}^{n \times p}$, and ask whether a given algorithm recovers the pair $(\boldsymbol{A}_0, \boldsymbol{X}_0)$.[4] In this work, we further assume that the sparse matrix $\boldsymbol{X}_0$ is random, with entries i.i.d. Bernoulli-Gaussian[5]. For simplicity, we will let $\boldsymbol{A}_0 = \boldsymbol{I}$; our arguments extend directly to general $\boldsymbol{A}_0$ via the simple change of variables $\boldsymbol{q} \mapsto \boldsymbol{A}_0^*\boldsymbol{q}$.

(Spielman et al., 2012) showed that under mild conditions, the complete dictionary recovery problem can be reduced to the geometric problem of finding a sparse vector in a linear subspace (Qu et al., 2014). Notice that because $\boldsymbol{A}_0$ is orthogonal, $\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X}_0)$. Because $\boldsymbol{X}_0$ is a sparse random matrix, the rows of $\boldsymbol{X}_0$ are sparse vectors. Under mild conditions (Spielman et al., 2012), they are the *sparsest* vectors in the row space of $\boldsymbol{Y}$, and hence can be recovered by solving the conceptual optimization problem

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

This is not a well-structured optimization problem: the objective is discontinuous, and the constraint set is open. A natural remedy is to replace the ℓ^0 norm with a smooth sparsity surrogate, and to break the scale ambiguity by constraining $\boldsymbol{q}$ to the sphere, giving

$$\min \; f_{\mathrm{DL}}(\boldsymbol{q}) \equiv \frac{1}{p}\sum_{k=1}^{p} h_\mu(\boldsymbol{q}^*\boldsymbol{y}_k) \quad \text{s.t.} \quad \boldsymbol{q} \in \mathbb{S}^{n-1}. \qquad (1)$$

Here, we choose $h_\mu(t) = \mu \log(\cosh(t/\mu))$ as a smooth sparsity surrogate. This choice is motivated by convenience of analysis and analogous performance guarantees should be obtainable for other choices. This objective was analyzed in (Sun et al., 2015a), which showed that (i) although this optimization problem is nonconvex, when the data are sufficiently large, with high probability every local optimizer is near a signed column of the true dictionary $\boldsymbol{A}_0$, (ii) every other critical point has a direction of strict negative curvature, and (iii) as a consequence, a second-order Riemannian trust region method efficiently recovers a column of $\boldsymbol{A}_0$.[6] The Riemannian trust region method is of mostly theoretical interest: it solves complicated (albeit polynomial time) subproblems that involve the Hessian of f_{DL}.

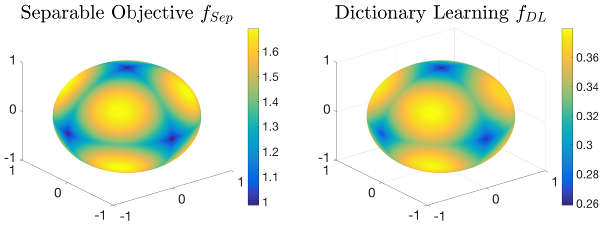

Figure 2. Left: The separable objective for $n = 3$. Note the similarity to the dictionary learning objective. *Right:* The objective for complete orthogonal dictionary learning (discussed in section 6) for $n = 3$.

In practice, simple iterative methods, including randomly initialized gradient descent are also observed to rapidly obtain high-quality solutions. In the sequel, we will give a geometric explanation for this phenomenon, and bound the rate of convergence of randomly initialized gradient descent to the neighborhood of a column of $\boldsymbol{A}_0$. Our analysis of f_{DL} is probabilistic in nature: it argues that with high probability in the sparse matrix $\boldsymbol{X}_0$, randomly initialized gradient descent rapidly produces a minimizer.

To isolate more clearly the key intuitions behind this analysis, we first analyze the simpler *separable objective*

$$\min \; f_{\mathrm{Sep}}(\boldsymbol{q}) \equiv \sum_{i=1}^{n} h_\mu(\boldsymbol{q}_i) \quad \text{s.t.} \quad \boldsymbol{q} \in \mathbb{S}^{n-1}. \qquad (2)$$

Figure 2 plots both f_{Sep} and f_{DL} as functions over the sphere. Notice that many of the key geometric features in f_{DL} are present in f_{Sep}; indeed, f_{Sep} can be seen as an "ultrasparse" version of f_{DL} in which the columns of the true sparse matrix $\boldsymbol{X}_0$ are taken to have only one nonzero entry. A virtue of this model function is that its critical points and their stable manifolds have simple closed form expressions (see Lemma 1).

[4] This problem exhibits a sign permutation symmetry: $\boldsymbol{A}_0\boldsymbol{X}_0 = (\boldsymbol{A}_0\boldsymbol{\Gamma})(\boldsymbol{\Gamma}^*\boldsymbol{X}_0)$ for any signed permutation matrix $\boldsymbol{\Gamma}$. Hence, we only ask for recovery up to a signed permutation.

[5] $[\boldsymbol{X}_0]_{ij} = \boldsymbol{V}_{ij}\boldsymbol{\Omega}_{ij}$, with $\boldsymbol{V}_{ij} \sim \mathcal{N}(0,1)$, $\boldsymbol{\Omega}_{ij} \sim \mathrm{Bern}(\theta)$ independent.

[6] Combining with a deflation strategy, one can then efficiently recover the entire dictionary $\boldsymbol{A}_0$.

4. Outline of Important Geometric Features

Our problems of interest have the form

$$\min\ f(\boldsymbol{q}) \quad \text{s.t.} \quad \boldsymbol{q} \in \mathbb{S}^{n-1},$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function. We let $\nabla f(\boldsymbol{q})$ and $\nabla^2 f(\boldsymbol{q})$ denote the Euclidean gradient and hessian (over $\mathbb{R}^n$), and let $\operatorname{grad}[f](\boldsymbol{q})$ and $\operatorname{Hess}[f](\boldsymbol{q})$ denote their Riemannian counterparts (over $\mathbb{S}^{n-1}$). The projection operator onto $\mathbb{S}^{n-1}$ is denoted by $\mathcal{P}_{\mathbb{S}^{n-1}}$.We will obtain results for Riemannian gradient descent defined by the update

$$\boldsymbol{q} \to \exp_{\boldsymbol{q}}(-\eta \operatorname{grad}[f](\boldsymbol{q}))$$

for some step size $\eta > 0$, where $\exp_{\boldsymbol{q}} : T_{\boldsymbol{q}}\mathbb{S}^{n-1} \to \mathbb{S}^{n-1}$ is the exponential map (Absil et al., 2009). The Riemannian gradient on the sphere is given by $\operatorname{grad}[f](\boldsymbol{q}) = (\boldsymbol{I} - \boldsymbol{q}\boldsymbol{q}^*)\nabla f(\boldsymbol{q})$.

We let A denote the set of critical points of f over $\mathbb{S}^{n-1}$ – these are the points $\bar{\boldsymbol{q}}$ s.t. $\operatorname{grad}[f](\bar{\boldsymbol{q}}) = \mathbf{0}$. We let $\breve{A}$ denote the set of local minimizers, and $\hat{A}$ its complement. Both f_{Sep} and f_{DL} are *Morse functions* on $\mathbb{S}^{n-1}$,[7] we can assign an index α to every $\bar{\boldsymbol{q}} \in A$, which is the number of negative eigenvalues of $\operatorname{Hess}[f](\bar{\boldsymbol{q}})$.

Our goal is to understand when gradient descent efficiently converges to a local minimizer. In the small-step limit, gradient descent follows gradient flow lines $\boldsymbol{\gamma} : \mathbb{R} \to \mathcal{M}$, which are solution curves of the ordinary differential equation

$$\dot{\boldsymbol{\gamma}}(t) = -\operatorname{grad}[f](\boldsymbol{\gamma}(t))$$

To each critical point $\boldsymbol{\alpha} \in A$ of index λ, there is an associated *stable manifold* of dimension $\dim(\mathcal{M}) - \lambda$, which is roughly speaking, the set of points that flow to α under gradient flow:

$$W^s(\boldsymbol{\alpha}) \equiv \left\{ \boldsymbol{q} \in \mathcal{M} \;\middle|\; \begin{array}{l} \lim\limits_{t\to\infty} \boldsymbol{\gamma}(t) = \boldsymbol{\alpha} \\ \boldsymbol{\gamma} \text{ a gradient flow line s.t. } \boldsymbol{\gamma}(0) = \boldsymbol{q} \end{array} \right\}.$$

Our analysis uses the following convenient coordinate chart

$$\varphi(\boldsymbol{w}) = \left(\boldsymbol{w}, \sqrt{1 - \|\boldsymbol{w}\|^2}\right) \equiv \boldsymbol{q}(\boldsymbol{w}) \tag{3}$$

where $\boldsymbol{w} \in B_1(0)$. We also define two useful sets:

$$\mathcal{C} \equiv \{\boldsymbol{q} \in \mathbb{S}^{n-1} | q_n \geq \|\boldsymbol{w}\|_\infty\}$$

$$\mathcal{C}_\zeta \equiv \left\{ \boldsymbol{q} \in \mathbb{S}^{n-1} \;\middle|\; \frac{q_n}{\|\boldsymbol{w}\|_\infty} \geq 1 + \zeta \right\}. \tag{4}$$

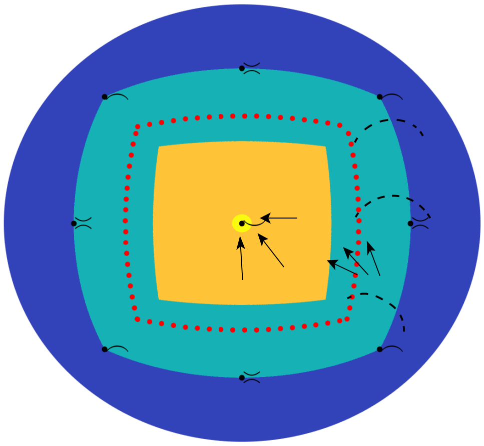

Figure 3. **Negative curvature and efficient gradient descent.** The union of the light blue, orange and yellow sets is the set $\mathcal{C}$. In the light blue region, there is negative curvature normal to $\partial\mathcal{C}$, while in the orange region the gradient norm is large, as illustrated by the arrows. There is a single global minimizer in the yellow region. For the separable objective, the stable manifolds of the saddles and maximizers all lie on $\partial\mathcal{C}$ (the black circles denote the critical points, which are either maximizers "⌢", saddles "≍", or minimizers "⌣"). The red dots denote $\partial\mathcal{C}_\zeta$ with $\zeta = 0.2$.

The set $\mathcal{C}$ is simply the subset of $\mathbb{S}^{n-1}$ where the n-th coordinate is largest in magnitude and positive. Since the problems considered here are symmetric with respect to a signed permutation of the coordinates we can consider a certain $\mathcal{C}$ and the results will hold for the other symmetric sections as well. We will show that at every point in $\mathcal{C}$ aside from a neighborhood of a global minimizer for the separable objective (or a solution to the dictionary problem that may only be a local minimizer), there is either a large gradient component in the direction of the minimizer or negative curvature in a direction normal to $\partial\mathcal{C}$. For the case of the separable objective, one can show that the stable manifolds of the saddles lie on this boundary, and hence this curvature is normal to the stable manifolds of the saddles and allows rapid progress away from small gradient regions and towards a global minimizer [8]. These regions are depicted in Figure 3.

In the sequel, we will make the above ideas precise for the two specific nonconvex optimization problems discussed in Section 3 and use this to obtain a convergence rate to a neighborhood of a global minimizer. Our analysis are specific to these problems. However, as we will describe in more detail later, they hinge on important geometric characteristics of these problems which make them amenable to

[7] Strictly speaking, f_{DL} is Morse with high probability, due to results of (Sun et al., 2017).

[8] The direction of this negative curvature is important here, and it is this feature that distinguishes these problems from other problems in the strict-saddle class where this direction may be arbitrary

efficient optimization, which may obtain in much broader classes of problems.

5. Separable Function Convergence Rate

In this section, we study the behavior of randomly initialized gradient descent on the separable function f_{Sep}. We begin by characterizing the critical points:

Lemma 1 (Critical points of f_{Sep})**.** *The critical points of the separable problem (2) are*

$$A = \left\{ \mathcal{P}_{\mathbb{S}^{n-1}}[\boldsymbol{a}] \,\middle|\, \boldsymbol{a} \in \{-1,0,1\}^{\otimes n}, \|\boldsymbol{a}\| > 0 \right\}. \tag{5}$$

For every $\boldsymbol{\alpha} \in A$ and corresponding $\boldsymbol{a}(\boldsymbol{\alpha})$, for $\mu < \frac{c}{\sqrt{n}\log n}$ the stable manifold of $\boldsymbol{\alpha}$ takes the form

$$W^s(\boldsymbol{\alpha}) = \left\{ \begin{array}{c} \mathcal{P}_{\mathbb{S}^{n-1}}\left[\boldsymbol{a}(\boldsymbol{\alpha}) + \boldsymbol{b}\right] \mid \\ \operatorname{supp}(\boldsymbol{a}(\boldsymbol{\alpha})) \cap \operatorname{supp}(\boldsymbol{b}) = \varnothing, \\ \|\boldsymbol{b}\|_\infty < 1 \end{array} \right\} \tag{6}$$

where $c > 0$ is a numerical constant.

Proof. Please see Appendix A □

By inspecting the dimension of the stable manifolds, it is easy to verify that that there are $2n$ global minimizers at the 1-sparse vectors on the sphere $\pm\widehat{\boldsymbol{e}}_i$, 2^n maximizers at the least sparse vectors and an exponential number of saddle points of intermediate sparsity. This is because the dimension of $W^s(\alpha)$ is simply the dimension of b in 6, and it follows directly from the stable manifold theorem that only minimizers will have a stable manifold of dimension $n-1$. The objective thus possesses no spurious local minimizers.

When referring to critical points and stable manifolds from now on we refer only to those that are contained in $\mathcal{C}$ or on its boundary. It is evident from Lemma 1 that the critical points in $\hat{A}$ all lie on $\partial\mathcal{C}$ and that $\bigcup_{\boldsymbol{\alpha}\in\hat{A}} W^s(\boldsymbol{\alpha}) = \partial\mathcal{C}$, and there is a minimizer at its center given by $\boldsymbol{q}(\boldsymbol{0}) = \widehat{\boldsymbol{e}}_n$.

5.1. The effect of negative curvature on the gradient

We now turn to making precise the notion that negative curvature normal to stable manifolds of saddle points enables gradient descent to rapidly exit small gradient regions. We do this by defining vector fields $\boldsymbol{u}^{(i)}(\boldsymbol{q}), i \in [n-1]$ such that each field is normal to a continuous piece of $\partial\mathcal{C}_\zeta$ and points outwards relative to $\mathcal{C}_\zeta$ defined in 4. By showing that the Riemannian gradient projected in this direction is positive and proportional to ζ, we are then able to show that gradient descent acts to increase $\zeta(\boldsymbol{q}(\boldsymbol{w})) = \frac{q_n}{\|\boldsymbol{w}\|_\infty} - 1$ geometrically. This corresponds to the behavior illustrated in the light blue region in Figure 3.

Lemma 2 (Separable objective gradient projection)**.** *For any $\boldsymbol{w}$ such that $\boldsymbol{q}(\boldsymbol{w}) \in \mathcal{C}_\zeta, i \in [n-1]$, we define a vector $\boldsymbol{u}^{(i)} \in T_{\boldsymbol{q}(\boldsymbol{w})}\mathbb{S}^{n-1}$ by*

$$u_j^{(i)} = \begin{cases} 0 & j \notin \{i, n\}, \\ \operatorname{sign}(w_i) & j = i, \\ -\frac{|w_i|}{q_n} & j = n. \end{cases} \tag{7}$$

If $\mu \log\left(\frac{1}{\mu}\right) \leq w_i$ and $\mu < \frac{1}{16}$, then

$$\boldsymbol{u}^{(i)*}\operatorname{grad}[f_{\text{Sep}}](\boldsymbol{q}(\boldsymbol{w})) \geq c\,\|\boldsymbol{w}\|_\infty \zeta,$$

where $c > 0$ is a numerical constant.

Proof. Please see Appendix A. □

Since we will use this property of the gradient in $\mathcal{C}_\zeta$ to derive a convergence rate, we will be interested in bounding the probability that gradient descent initialized randomly with respect to a uniform measure on the sphere is initialized in $\mathcal{C}_\zeta$. This will require bounding the volume of this set, which is done in the following lemma:

Lemma 3 (Volume of $\mathcal{C}_\zeta$)**.** *For $\mathcal{C}_\zeta$ defined as in (4) we have*

$$\frac{\operatorname{Vol}(\mathcal{C}_\zeta)}{\operatorname{Vol}(\mathbb{S}^{n-1})} \geq \frac{1}{2n} - \frac{\log(n)}{n}\zeta$$

Proof. Please see Appendix D.3. □

5.2. Convergence rate

Using the results above, one can obtain the following convergence rate:

Theorem 1 (Gradient descent convergence rate for separable function)**.** *For any $0 < \zeta_0 < 1$, $r > \mu\log\left(\frac{1}{\mu}\right)$, Riemannian gradient descent with step size $\eta < \min\left\{\frac{c_1}{n}, \frac{\mu}{2}\right\}$ on the separable objective (2) with $\mu < \frac{c_2}{\sqrt{n}\log n}$, enters an L^∞ ball of radius r around a global minimizer in*

$$T < \frac{C}{\eta}\left(\frac{\sqrt{n}}{r^2} + \log\left(\frac{1}{\zeta_0}\right)\right)$$

iterations with probability

$$\mathbb{P} \geq 1 - 2\log(n)\zeta_0,$$

where $c_i, C > 0$ are numerical constants.

Proof. Please see Appendix A. □

We have thus obtained a convergence rate for gradient descent that relies on the negative curvature around the stable manifolds of the saddles to rapidly move from these regions of the space towards the vicinity of a global minimizer. This is evinced by the logarithmic dependence of the rate on ζ. As was shown for orthogonal dictionary learning in (Sun et al., 2017), we also expect a linear convergence rate due to strong convexity in the neighborhood of a minimizer, but do not take this into account in the current analysis.

6. Dictionary Learning Convergence Rate

The proofs in this section will be along the same lines as those of Section 5. While we will not describe the positions of the critical points explicitly, the similarity between this objective and the separable function motivates a similar argument. It will be shown that initialization in some $\mathcal{C}_\zeta$ will guarantee that Riemannian gradient descent makes uniform progress in function value until reaching the neighborhood of a global minimizer. We will first consider the population objective which corresponds to the infinite data limit

$$f_{\mathrm{DL}}^{\mathrm{pop}}(\boldsymbol{q}) \equiv \mathop{\mathbb{E}}_{\boldsymbol{X}_0} f_{\mathrm{DL}}(\boldsymbol{q}) = \mathbb{E}_{\boldsymbol{x}\sim_{\mathrm{i.i.d.}}\mathrm{BG}(\theta)}\left[h_\mu(\boldsymbol{q}^*\boldsymbol{x}) \right]. \quad (8)$$

and then bounding the finite sample size fluctuations of the relevant quantities. We begin with a lemma analogous to Lemma 2:

Lemma 4 (Dictionary learning population gradient)**.** *For any $\boldsymbol{w}$ such that $\boldsymbol{q}(\boldsymbol{w}) \in \mathcal{C}_\zeta, r < |w_i|, \mu < c_1 r^{5/2}\sqrt{\zeta}$ the dictionary learning population objective 8 obeys*

$$\boldsymbol{u}^{(i)*}\mathrm{grad}[f_{\mathrm{DL}}^{\mathrm{pop}}](\boldsymbol{q}(\boldsymbol{w})) \geq c_\theta r^3 \zeta$$

where c_θ depends only on θ, c_1 is a positive numerical constant and $\boldsymbol{u}^{(i)}$ is defined in 7.

Proof. Please see Appendix B □

Using this result, we obtain the desired convergence rate for the population objective, presented in Lemma 11 in Appendix B. After accounting for finite sample size fluctuations in the gradient, one obtains a rate of convergence to the neighborhood of a solution (which is some signed basis vector due to our choice $\boldsymbol{A}_0 = \boldsymbol{I}$)

Theorem 2 (Gradient descent convergence rate for dictionary learning)**.** *For any $1 > \zeta_0 > 0, s > \frac{\mu}{4\sqrt{2}}$, Riemannian gradient descent with step size $\eta < \frac{c_5\theta s}{n\log np}$ on the dictionary learning objective 1 with $\mu < \frac{c_6\sqrt{\zeta_0}}{n^{5/4}}, \theta \in (0, \frac{1}{2})$, enters a ball of radius $c_3 s$ from a target solution in*

$$T < \frac{C_2}{\eta\theta}\left(\frac{1}{s} + n\log\frac{1}{\zeta_0}\right)$$

iterations with probability

$$\mathbb{P} \geq 1 - 2\log(n)\zeta_0 - \mathbb{P}_y - c_8 p^{-6}$$

where $\mathbb{P}_y$ is given in Lemma 10 with $y = \frac{c_7\theta(1-\theta)\zeta_0}{n^{3/2}}$ and c_i, C_i are positive constants. Additionally, $\mathbb{P}_y \leq \exp\left(-\frac{\tilde{c}(\theta,\zeta_0)p}{n^3} + n polylog(n, \frac{1}{\mu}, \frac{1}{\zeta_0}, \theta) + \log n\right)$ for some $\tilde{c}(\theta, \zeta_0)$ > 0.

Proof. Please see Appendix B □

The two terms in the rate correspond to an initial geometric increase in the distance from the set containing the small gradient regions around saddle points, followed by convergence to the vicinity of a minimizer in a region where the gradient norm is large. The latter is based on results on the geometry of this objective provided in (Sun et al., 2017).

7. Discussion

The above analysis suggests that second-order properties - namely negative curvature normal to the stable manifolds of saddle points - play an important role in the success of randomly initialized gradient descent in the solution of complete orthogonal dictionary learning. This was done by furnishing a convergence rate guarantee that holds when the random initialization is not in regions that feed into small gradient regions around saddle points, and bounding the probability of such an initialization. In Appendix C we provide an additional example of a nonconvex problem for which an efficient rate can be obtained based on an analysis that relies on negative curvature normal to stable manifolds of saddles - generalized phase retrieval. An interesting direction of further work is to more precisely characterize the class of functions that share this feature.

The effect of curvature can be seen in the dependence of the maximal number of iterations T on the parameter ζ_0. This parameter controlled the volume of regions where initialization would lead to slow progress and the failure probability of the bound $1 - \mathbb{P}$ was linear in ζ_0, while T depended logarithmically on ζ_0. This logarithmic dependence is due to a geometric increase in the distance from the stable manifolds of the saddles during gradient descent, which is a consequence of negative curvature. Note that the choice of ζ_0 allows one to flexibly trade off between T and $1 - \mathbb{P}$. By decreasing ζ_0, the bound holds with higher probability, at the price of an increase in T. This is because the volume of acceptable initializations now contains regions of smaller minimal gradient norm. In a sense, the result is an extrapolation of works such as (Lee et al., 2017) that analyze the $\zeta_0 = 0$ case to finite ζ_0.

Our analysis uses precise knowledge of the location of the stable manifolds of saddle points. For less symmetric problems, including variants of sparse blind deconvolution (Zhang et al., 2017) and overcomplete tensor decomposition, there is no closed form expression for the stable manifolds. However, it is still possible to coarsely localize them in regions containing negative curvature. Understanding the implications of this geometric structure for randomly initialized first-order methods is an important direction for future work. One may also hope that studying simple model problems and identifying structures (here, negative curvature orthogonal to the stable manifold) that enable efficient optimization will inspire approaches to broader classes of problems.

Acknowledgements

The authors gratefully acknowledge support from NSF 1740833, 1733857, 1527809, and an NDSEG fellowship.

References

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

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

Balan, R., Casazza, P., and Edidin, D. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345–356, 2006.

Bandeira, A. S., Boumal, N., and Voroninski, V. On the low-rank approach for semidefinite programs arising in synchronization and community detection. In *Conference on Learning Theory*, pp. 361–382, 2016.

Bertsekas, D. P. *Nonlinear programming*. Athena scientific Belmont, 1999.

Boumal, N. Nonconvex phase synchronization. *SIAM Journal on Optimization*, 26(4):2355–2377, 2016.

Bronstein, M. M., Bronstein, A. M., Zibulevsky, M., and Zeevi, Y. Y. Blind deconvolution of images using optimal sparse representations. *IEEE Transactions on Image Processing*, 14(6):726–736, 2005.

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

Chen, Y., Chi, Y., Fan, J., and Ma, C. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *arXiv preprint arXiv:1803.07726*, 2018.

Conn, A. R., Gould, N. I., and Toint, P. L. *Trust region methods*, volume 1. Siam, 2000.

Corbett, J. V. The pauli problem, state reconstruction and quantum-real numbers. *Reports on Mathematical Physics*, 57:53–68, 2006.

Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pp. 2933–2941, 2014.

Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Poczos, B., and Singh, A. Gradient descent can take exponential time to escape saddle points. *arXiv preprint arXiv:1705.10412*, 2017.

Elad, M. and Aharon, M. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points?online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pp. 797–842, 2015.

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

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

Hardt, M., Ma, T., and Recht, B. Gradient descent learns linear dynamical systems. *arXiv preprint arXiv:1609.05191*, 2016.

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

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

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

Kreutz-Delgado, K. The complex gradient operator and the cr-calculus. *arXiv preprint arXiv:0906.4835*, 2009.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pp. 1246–1257, 2016.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

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

Mairal, J., Bach, F., Ponce, J., et al. Sparse modeling for image and vision processing. *Foundations and Trends® in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

Miao, J., Ishikawa, T., Johnson, B., Anderson, E. H., Lai, B., and Hodgson, K. O. High resolution 3d x-ray diffraction microscopy. *Physical review letters*, 89(8):088303, 2002.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Nicolaescu, L. *An invitation to Morse theory*. Springer Science & Business Media, 2011.

Olshausen, B. A. and Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607, 1996.

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

Ravishankar, S. and Bresler, Y. Mr image reconstruction from highly undersampled k-space data by dictionary learning. *IEEE transactions on medical imaging*, 30(5): 1028–1041, 2011.

Shechtman, Y., Eldar, Y. C., Cohen, O., Chapman, H. N., Miao, J., and Segev, M. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE signal processing magazine*, 32(3):87–109, 2015.

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

Sun, J., Qu, Q., and Wright, J. Complete dictionary recovery over the sphere. In *Sampling Theory and Applications (SampTA), 2015 International Conference on*, pp. 407–410. IEEE, 2015a.

Sun, J., Qu, Q., and Wright, J. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015b.

Sun, J., Qu, Q., and Wright, J. A geometric analysis of phase retrieval. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pp. 2379–2383. IEEE, 2016.

Sun, J., Qu, Q., and Wright, J. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Zhang, Y., Lau, Y., Kuo, H.-w., Cheung, S., Pasupathy, A., and Wright, J. On the global geometry of sphere-constrained sparse blind deconvolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4894–4902, 2017.