A Well-Tempered Landscape for Non-convex Robust Subspace Recovery

Tyler Maunu   MAUNUT@MIT.EDU
Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139

Teng Zhang   TENG.ZHANG@UCF.EDU
Department of Mathematics
University of Central Florida
Orlando, FL 32816

Gilad Lerman   LERMAN@UMN.EDU
School of Mathematics
University of Minnesota
Minneapolis, MN 55455

Editor: Moritz Hardt

Abstract

We present a mathematical analysis of a non-convex energy landscape for robust subspace recovery. We prove that an underlying subspace is the only stationary point and local minimizer in a specified neighborhood under a deterministic condition on a dataset. If the deterministic condition is satisfied, we further show that a geodesic gradient descent method over the Grassmannian manifold can exactly recover the underlying subspace when the method is properly initialized. Proper initialization by principal component analysis is guaranteed with a simple deterministic condition. Under slightly stronger assumptions, the gradient descent method with a piecewise constant step-size scheme achieves linear convergence. The practicality of the deterministic condition is demonstrated on some statistical models of data, and the method achieves almost state-of-the-art recovery guarantees on the Haystack Model for different regimes of sample size and ambient dimension. In particular, when the ambient dimension is fixed and the sample size is large enough, we show that our gradient method can exactly recover the underlying subspace for any fixed fraction of outliers (less than 1).

Keywords: robust subspace recovery, non-convex optimization, dimension reduction, optimization on the Grassmannian

1. Introduction

Robust subspace recovery (RSR) involves estimating a low-dimensional linear subspace in a corrupted dataset. It assumes that a portion of the given dataset lies close to or on a subspace, which we will refer to as the "underlying subspace". The other portion of the dataset is assumed to be corrupted and may lie far from the underlying subspace. In this regime, noise and corruption are separate entities: corruption refers to large and potentially arbitrary changes to a data point, while noise is a small perturbation of a data point.

A basic method for modeling data by a low-dimensional subspace is Principal Component Analysis (PCA) (Jolliffe, 2002). PCA is popular for both reducing noise and capturing low-dimensional structure within data. However, PCA is notoriously sensitive to corrupted data and does not perform well in many regimes of the RSR problem.

Many strategies have been proposed for the RSR problem, which are reviewed in Lerman and Maunu (2018b). However, despite nice progress, many of the proposed methods have inherent issues. Perhaps the largest flaw in many existing methods is computational time: many methods require calculation of a full covariance matrix or matrix inversion. If we have a dataset of N points in $\mathbb{R}^D$, these calculations typically have complexity $O(ND^2)$ or $O(D^3)$. On the other hand, the PCA d-dimensional subspace can be calculated with complexity $O(NDd)$. Some recent proposals for RSR have complexity that scales like $O(NDd)$, but they either do not have satisfying theoretical guarantees or have extra user specified parameters. Ideally, we would like algorithms that run in $O(NDd)$ because the set of d-dimensional subspaces has dimension $O(Dd)$.

The key point of our work is the development of a computationally efficient and provably accurate method for RSR. We desire a method that has complexity $O(NDd)$ and that does not sacrifice theoretical guarantees. The method we propose involves minimization of the robust least absolute deviations energy function. Minimizing this function involves solving a non-convex optimization problem that is NP-hard.

Even though the problem is NP-hard in general, we derive conditions that ensure the energy landscape is well-behaved in a substantial neighborhood of an underlying subspace. These conditions also ensure that a geodesic gradient descent method can locally recover an underlying subspace, and the convergence of this method is linear under some slightly stronger assumptions. This linear convergence implies that this $O(NDd)$ algorithm is very efficient. Furthermore, we give the most complete discussion of recovery under various regimes of corruption and relatively broad statistical models. In particular, we show that our method is robust to very high percentages of corruption under special generative models. To our knowledge, we give the strongest guarantees on a non-convex method for RSR to date and even obtain stronger results than some convex methods.

In the rest of this section, we give some necessary background for our method and an overview of this work. First, in Section 1.1, we give some essential background to understand our approach to RSR. Then, Section 1.2 outlines the main contributions of our work. Finally, Section 1.3 summarizes the structure of the paper, and Section 1.4 discusses necessary notation.

1.1. Essential Background

Here, we briefly summarize the necessary background to understand the primary contribution of this work. First, for the rest of this paper, we assume a linear subspace setting, which means that we only consider underlying subspaces that are linear. We leave the case of affine subspaces to future work. For simplicity of discussion, we advocate centering by the geometric median for real data when the center is not known.

The essential problem of RSR is an optimization problem over the Grassmannian, which is the set of linear subspaces of a fixed dimension. The Grassmannian is a non-convex set, which makes optimization over it hard. Frequently, this leads to NP-hard or SSE-hard

formulations (Hardt and Moitra, 2013; Clarkson and Woodruff, 2015). In this paper, we denote by $G(D, d)$ the Grassmannian of linear d-dimensional subspaces in $\mathbb{R}^D$ and refer to such subspaces as d-subspaces.

It is illuminating to first outline the PCA subspace problem, since it has a similar form to our methodology. The basic formulation for the PCA subspace problem can be cast as an optimization over $G(D, d)$. For a dataset $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subset \mathbb{R}^D$ centered at the origin, the PCA d-subspace is the solution of the least squares problem

$$\min_{L \in G(D,d)} \sum_{i=1}^{N} \|\boldsymbol{x}_i - \boldsymbol{P}_L \boldsymbol{x}_i\|_2^2, \tag{1}$$

where $\boldsymbol{P}_L$ denotes the orthogonal projection matrix onto the subspace L. As has been noted in many past works, the least squares formulation is sensitive to corrupted data.

Some have attempted to make the PCA formulation robust by considering least absolute deviations:

$$\min_{L \in G(D,d)} \sum_{i=1}^{N} \|\boldsymbol{x}_i - \boldsymbol{P}_L \boldsymbol{x}_i\|_2. \tag{2}$$

This optimization can be thought of as estimating a geometric median subspace. Some have tried to directly optimize this problem (Ding et al., 2006; Lerman and Maunu, 2018a), while others have tried to solve convex relaxations of it (McCoy and Tropp, 2011; Xu et al., 2012; Zhang and Lerman, 2014; Lerman et al., 2015).

The goal of this work is to directly analyze the energy landscape of (2) and guarantee that a non-convex gradient descent method for this energy minimization can recover an underlying subspace. This gradient descent method leads to huge gains in speed over previous convex methods but does not sacrifice accuracy in subspace recovery.

In the RSR problem setup, it is common to refer to the uncorrupted portion of the dataset as inliers, which lie on or near the underlying subspace. The case where the inliers lie on the underlying subspace is referred to as the noiseless RSR setting, while the case where the inliers lie near the underlying subspace is referred to as the noisy RSR setting. The corrupted points in the dataset are referred to as outliers, which are assumed to lie somewhere in the ambient space. Exact recovery in the noiseless setting refers to when a method outputs the underlying subspace exactly. Near recovery in the noisy setting refers to when a method outputs a good approximation of the underlying subspace, where the goodness of approximation depends on the noise level.

1.2. Contribution of This Work

As mentioned, the goal of this work is to recover an underlying subspace by directly optimizing the non-convex function in (2). To motivate why such a procedure might work, Figure 1 demonstrates the landscape of the energy function in (2) for two simulated datasets. The novelty of this paper consists of the following observation in these and certain other datasets: despite non-convexity, the energy landscape appears to exhibit basins of attraction around the underlying subspaces. In other words, the energy function decreases over $G(D, d)$ in the direction of L_* within some neighborhood. Indeed, it appears that direct minimization of the energy in a local neighborhood would yield exact recovery of the underlying subspace.
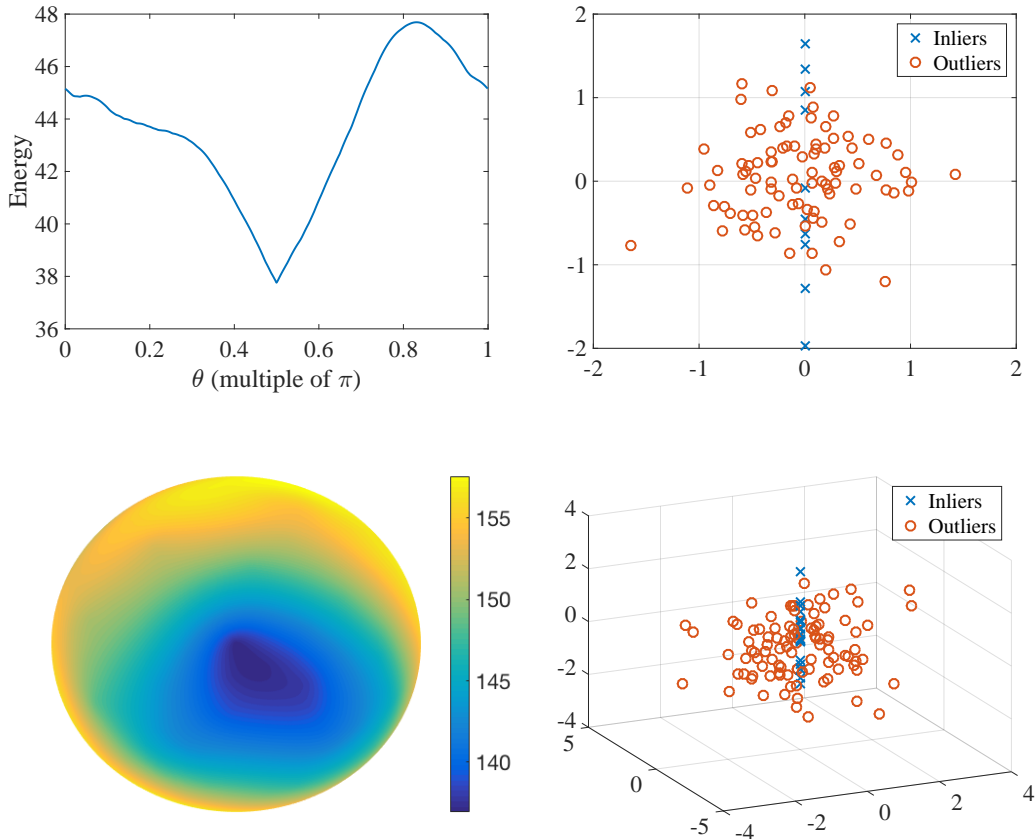
Figure 1: Demonstration of the energy landscape of (2) over $G(2,1)$ and $G(3,1)$ with simulated Gaussian data. The simulated datasets are demonstrated on the right and the corresponding energy landscape is depicted on the left. Top: In $\mathbb{R}^2$, 90 outliers are i.i.d. $\mathcal{N}(\mathbf{0}, \boldsymbol{I}/2)$ and 10 inliers are i.i.d. $\mathcal{N}(0,1)$ along the y-axis. $G(2,1)$ is identified with a semicircle and parameterized by angle. The energy is depicted as a function of the parametrizing angle. Its global minimum is at $\pi/2$, which corresponds to the underlying line at the y-axis of $\mathbb{R}^2$. Bottom: In $\mathbb{R}^3$, 100 outliers are i.i.d. $\mathcal{N}(\mathbf{0}, 4\boldsymbol{I}_3/3)$ and 20 inliers are i.i.d. $\mathcal{N}(0,1)$ along the z-axis. $G(3,1)$ is identified with the top hemisphere which is flattened to a circle. The energy function is depicted by a heat map on that circle. The global minimum is at the center of the circle, which corresponds to the underlying line at the z-axis of $\mathbb{R}^3$.

It is important to emphasize that this phenomenon is inherently local. Looking at the energy plots in Figure 1, it appears that there may be other local minimizers far from the underlying subspace, and so proper initialization is quite important.

Our key contributions follow:

- We show that, under deterministic conditions, the robust energy landscape of (2) exhibits basins of attraction around an underlying subspace as seen in Figure 1. Theorems 1 and 2 formulate this result for the energy landscape of the noiseless and noisy RSR settings respectively.

- We propose a geodesic gradient descent algorithm in Section 4. Theorems 3 and 4 show that this algorithm exactly recovers the underlying subspace under the conditions of Theorem 1 with proper initialization. Theorem 3 guarantees sublinear convergence of this algorithm to the underlying subspace. With some additional, slightly stronger assumptions, Theorem 4 guarantees linear convergence to the underlying subspace using a piecewise constant step-size scheme. These results can be generalized to near recovery in the noisy RSR setting as well (see Remark 5).

- Lemma 9 guarantees that we can initialize in the correct local neighborhood using PCA under a similar deterministic condition. This yields a complete guarantee for geodesic gradient descent with PCA initialization.

- The deterministic guarantees are shown to hold for a variety of statistical models of data. In particular, we achieve competitive guarantees for recovery under the special Haystack Model. More specifically, we consider three different regimes of sample size. The first regime, $N = O(D)$, describes the scenario of a relatively small sample. The smallest ratio of inliers to outliers where exact recovery is still possible for this model among all algorithms is $d/(D-d)$, although this has only been established for algorithms of complexity $O(ND^2)$ at best (Hardt and Moitra, 2013; Zhang, 2016). We guarantee instead exact recovery for our $O(NDd)$ algorithm in this regime under a larger ratio of inliers to outliers, namely of order $O(d/\sqrt{D})$ (see Corollary 16). In the regime of larger samples, when $N = O(D^2)$, exact recovery with the even smaller ratio $O(d/\sqrt{D(D-d)})$ was previously obtained for a convex algorithm (Zhang and Lerman, 2014). Theorem 17 implies recovery with this same small fraction but in the regime $N = O(d(D-d)^2 \log(D))$. Except for this convex algorithm and our proposed algorithm, we are unaware of any other method that is guaranteed to obtain exact recovery under such a small fraction in the regime $N = O(d(D-d)^2 \log(D))$. Beyond this, in the regime of very large samples, we show that our method can exactly recover an underlying subspace with any fixed fraction of outliers, where N is at least some polynomial order of D and also depends on the ratio of inliers to outliers. This is the only efficient RSR method with such a guarantee.

We will close this section by briefly commenting on how to read the theoretical results of this paper. The first three bullets constitute the primary theoretical interest of this work, since they guarantee the usefulness of our non-convex RSR method under some fairly general conditions. In particular, the stability statistic developed in Section 3.1.1 lies at the core of much of our analysis, and many of the more complicated results are extensions of this statistic's analysis in Theorem 1. The discussion of statistical models as outlined in the fourth bullet above is merely included for interpretability of the general conditions we offer. With such models, we develop a heuristic understanding of these conditions, and, in particular, we gain insight into trade-offs between inliers and outliers. These models also allow for easier comparison of the theoretical guarantees of all RSR methods.

1.3. Paper Organization

First, in Section 2, we review previous work on the RSR problem. In Section 3, we describe deterministic conditions that ensure that the energy landscape of (2) behaves nicely around an underlying subspace. Then, in Section 4, we outline a geodesic gradient descent method on the set of subspaces. We show that this method can locally recover an underlying subspace for datasets satisfying the deterministic conditions and that the convergence rate is linear under some slightly stronger assumptions. We obtain exact recovery in the noiseless RSR setting and near recovery in the noisy RSR setting. Then, Section 5 shows that the conditions hold for certain statistical models of data. In particular, we obtain almost state-of-the-art results on recovery under the Haystack Model and also consider a range of other models. Next, Section 6 gives simulations that agree with the theoretical results of this paper. Finally, Section 7 concludes this work and discusses possible future directions.

1.4. Notation

Before presenting the results of this paper, we explain our commonly used notation. The letter L is used to refer to d-subspaces. We use bold upper case letters for matrices and bold lower case letters for column vectors. For a matrix $\boldsymbol{A}$, $\mathrm{Sp}(\boldsymbol{A})$ is the subspace spanned by the columns of $\boldsymbol{A}$, $\sigma_j(\boldsymbol{A})$ denotes the jth singular value of $\boldsymbol{A}$, and, if $\boldsymbol{A}$ is square, then $\lambda_j(\boldsymbol{A})$ denotes its jth eigenvalue. The spectral norm of $\boldsymbol{A}$ is $\|\boldsymbol{A}\|_2$, and the Euclidean 2-norm for vectors is denoted by $\|\cdot\|$. The notation $\widetilde{\boldsymbol{A}}$ denotes projection of the columns of $\boldsymbol{A}$ to the unit sphere. For $d \leq D$, the set of semi-orthogonal $D \times d$ matrices is denoted by $O(D, d) = \{\boldsymbol{V} \in \mathbb{R}^{D \times d} : \boldsymbol{V}^T \boldsymbol{V} = \boldsymbol{I}_d\}$. For $\boldsymbol{V} \in O(D, d)$, we denote its columns by $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$. We recall that $G(D, d)$ denotes the Grassmannian, that is, the set of d-dimensional linear subspaces of $\mathbb{R}^D$. The orthogonal projection matrix onto the subspace $L = \mathrm{Sp}(\boldsymbol{V})$ is denoted by $\boldsymbol{P}_L$, and we interchangeably use $\boldsymbol{P}_{\boldsymbol{V}} = \boldsymbol{P}_L$. The projection onto the orthogonal complement of L is $\boldsymbol{Q}_L = \boldsymbol{I} - \boldsymbol{P}_L$. We denote the largest principal angle between two subspaces L_1 and L_2 by $\theta_1(L_1, L_2)$, that is, $\theta_1(L_1, L_2) = \arccos\left(\sigma_d(\boldsymbol{P}_{L_1}\boldsymbol{P}_{L_2})\right)$. We say that an event occurs with high probability (w.h.p.) if the probability is bounded below by $1 - O(N^a)$, for some absolute constant $a > 0$. We say that an event occurs with overwhelming probability (w.o.p.) if the probability is bounded below by $1 - O(e^{-aN^b})$, for absolute constants $a, b > 0$. The notation $f(x) \lesssim g(x)$ is used to denote that $f(x) < Cg(x)$ for some absolute constant C (and the notation $\gtrsim$ is used in the same way).

2. Background and Review of Previous Work

In this section, we will review past work on the RSR problem and necessary background concepts for this work. First, Section 2.1 discusses past attempts to solve the RSR problem. Then, Section 2.2 gives the background concepts that are necessary to understand our later results.

2.1. Review of Previous Work

The most ubiquitous subspace modeling framework uses principal component analysis (Jolliffe, 2002). Its optimization problem, which is formulated in (1), is non-convex since $G(D, d)$

is non-convex. However, despite its non-convexity, this problem has a direct solution, which is calculated from the singular value decomposition of the data matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$. This problem also has a nice energy landscape. Indeed, if the dth singular value of $\boldsymbol{X}$ is larger than the $(d+1)$st singular value, then the global minimum is unique, and there are no other local minima; otherwise, if the dth singular value is equal to the $(d+1)$st, then all local minima are globally optimal. Saddle points are also guaranteed to be sufficiently far from the global minimum, and they have explicit expressions. We discuss the PCA energy landscape further in Appendix A. These nice properties of the PCA subspace optimization are not shared by the algorithms for RSR that we discuss next.

Examples of works on RSR include Maronna (2005); Maronna et al. (2006); Ding et al. (2006); Zhang et al. (2009); Lerman and Zhang (2011); McCoy and Tropp (2011); Xu et al. (2012); Coudron and Lerman (2012); Hardt and Moitra (2013); Zhang and Lerman (2014); Goes et al. (2014); Lerman and Zhang (2014); Lerman et al. (2015); Zhang (2016); Lerman and Maunu (2018a); Cherapanamjeri et al. (2017). A comprehensive overview of this topic is given in Lerman and Maunu (2018b).

We note that this problem is distinct from what is typically called robust PCA (RPCA) (Candès et al., 2011; Zhou et al., 2010), where the corruptions occur element-wise throughout the whole data matrix rather than some samples being wholly corrupted. Algorithms for RPCA typically do not perform well in the RSR setting, and algorithms for RSR do not perform well in the RPCA setting.

RSR is inherently non-convex due to the non-convexity of $G(D, d)$. Robust versions of the PCA energy may have more complicated landscapes in general. One way of making PCA robust is to simply project the data to the unit sphere, S^{D-1}, before running PCA (Locantore et al., 1999; Maronna, 2005; Maronna et al., 2006). This deals with PCA's sensitivity to the scaling of the data and makes it easier to find directions that robustly capture variance. However, it is still not able to deal with correlated outlier directions and does not have good asymptotic guarantees even for simple models (Lerman and Maunu, 2018a).

As mentioned, another way to make PCA robust is to consider least absolute deviations (Ding et al., 2006; McCoy and Tropp, 2011; Xu et al., 2012; Zhang and Lerman, 2014; Lerman and Zhang, 2014; Lerman et al., 2015; Lerman and Maunu, 2018a). The first use of least absolute deviations in subspace modeling was the work on orthogonal regression by Osborne and Watson (1985). This was not extended to general subspace modeling until much later (Watson, 2001; Ding et al., 2006). Previous works have considered convex relaxation of this energy (McCoy and Tropp, 2011; Xu et al., 2012; Zhang and Lerman, 2014; Lerman et al., 2015). However, such convex relaxations are generally slow and may not approximate the underlying problem well. Indeed, most either have complexity $O(ND^2)$ or $O(D^3)$.

The works of Lerman and Zhang (2011, 2014) established under a certain model that an underlying subspace is recoverable by the minimizer of (2). However, they did not provide a guaranteed algorithm for minimizing this energy. The estimates of these works do not hold for small sample sizes: they only hold for large N. Lerman and Maunu (2018a) developed the FMS algorithm, which employs iteratively reweighted least squares to optimize (2). However, the FMS algorithm does not have deterministic guarantees of fast convergence or deterministic results on recovery of the underlying subspace. The FMS algorithm does

have theoretical guarantees of approximate recovery for a very special model of data, with relatively large samples. In contrast, we directly minimize (2) by gradient descent, and we provide deterministic guarantees of fast convergence and subspace recovery.

Another recent work on RSR was given by Cherapanamjeri et al. (2017), where they propose Thresholding based Outlier Robust PCA (TORP). TORP has analysis for arbitrary outliers and noise, as long as the percentage of outliers is known in advance. While the tolerance to very low percentages of arbitrary outliers is not that impressive, the noise analysis is somewhat novel. Under Gaussian noise, the authors are able to show similar sample complexity as that of PCA. One downside of this algorithm is that one must know the percentage of outliers as an input. Further, since the guarantees are only for adversarial models of outliers, there is no discussion of improved estimates when the outliers are not adversarial but instead obey a specific statistical model.

In the existing literature, only a few methods achieve the complexity bound of $O(TNDd)$, where T is iteration count. These include SPCA (Maronna, 2005), RANSAC and RandomizedFind (Hardt and Moitra, 2013; Arias-Castro and Wang, 2017), FMS (Lerman and Maunu, 2018a), and TORP (Cherapanamjeri et al., 2017). Among these, the algorithms either do not have sufficiently satisfying guarantees for recovery, or they do not have a good bound on T, or they require additional parameters. SPCA is the fastest out of these algorithms since it has $T = 1$. Here, we are slightly abusing the complexities and assuming the cost of running PCA is $O(NDd)$, despite the fact that PCA is also an iterative algorithm. We choose this convention due to the fact that many methods use PCA as a sub-routine. While SPCA is somewhat robust to arbitrary outliers, it cannot exactly recover subspaces in the presence of outliers. However, SPCA is nice since it is quite general and lacks the specialized assumptions of many methods. RANSAC requires a user to input specialized parameters, such as the consensus number and a consensus threshold. RANSAC can also only bound T in probability under certain conditions. This also goes for the analysis of RandomizedFind given by Arias-Castro and Wang (2017), along with their updated algorithm that has complexity $O(TDd)$. In many cases, though, this T can be very large. For both the RANSAC and RandomizedFind methods, recovery guarantees exist in the noiseless RSR setting under specialized assumptions, but there are no satisfying extensions of either method to noise. On the other hand, TORP (Cherapanamjeri et al., 2017) requires a user to input the percentage of outliers that is not known in general. TORP has a guarantee of linear convergence under certain conditions, but, as we mentioned earlier, it does not have satisfying guarantees for subspace recovery. FMS (Lerman and Maunu, 2018a) only has guarantees for rate of convergence and recovery for very special models of data.

One way to compare the theoretical guarantees of various methods is to assume a statistical model of data and then determine which algorithm performs best in this model. For example, one common choice of model in past works was the Haystack Model, which can be seen in Lerman et al. (2015). Another model was to assume spherically symmetric outliers, and inliers spherically symmetric on an underlying subspace (Lerman and Zhang, 2014; Lerman and Maunu, 2018a). Others have examined models with arbitrary outliers (Xu et al., 2012; Cherapanamjeri et al., 2017). In this work, after giving our general theoretical guarantees, we will show how they can be applied to a variety of statistical models of data.

This paper also fits in to the surge of recent work that has focused on non-convex optimization for many structured data problems (Dauphin et al., 2014; Hardt, 2014; Jain

et al., 2014; Ge et al., 2015; Lee et al., 2016; Arora et al., 2015; Mei et al., 2018; Ge et al., 2016; Boumal, 2016; Sun et al., 2015b,a; Ma et al., 2018). Some work has focused on non-convex optimization for robust PCA (Netrapalli et al., 2014; Yi et al., 2016; Zhang and Yang, 2017), which is a related but different problem than RSR. Others have attempted to solve non-convex versions of the RSR problem (Lerman and Maunu, 2018a; Cherapanamjeri et al., 2017), but, as we have discussed, these methods each have their own shortcomings.

This work is partially built on optimization on manifolds, and in particular there are important results on optimization over the Grassmannian manifold (Edelman et al., 1999; Absil et al., 2004). Edelman et al. (1999) develop gradient descent on the Grassmannian and give formulations for Newton's method and conjugate gradient for the Grassmannian. We discuss optimization on the Grassmannian in more detail in the next section.

Many other recent works have also focused on using optimization on the Grassmannian to solve various problems (Zhang et al., 2009; Goes et al., 2014; St. Thomas et al., 2014; Zhang and Balzano, 2016; Ye and Lim, 2016; Lim et al., 2016). The work of Zhang and Balzano (2016) examines a rank one geodesic gradient scheme for solving online PCA. Their setting is distinctly different from ours since they attempt to solve the PCA problem rather than RSR. They also only prove recovery of the PCA solution for a specific model of Gaussian noise, and no deterministic condition for global recovery is given. Further, while we assume centered data in this paper, St. Thomas et al. (2014) and Lim et al. (2016) consider estimation on the affine Grassmannian.

2.2. Review of Optimization over $G(D, d)$

The minimization in (2) involves optimization over the Grassmannian manifold. To understand the energy landscape, one must have a basic understanding of the geometry of $G(D, d)$ and how to calculate derivatives over it.

We can write the energy function in (2) in two equivalent ways. First, as a function over $G(D, d)$, we write

$$F(L; \mathcal{X}) = \sum_{\mathcal{X}} \|\boldsymbol{Q}_L \boldsymbol{x}_i\|. \tag{3}$$

On the other hand, we can represent points in $G(D, d)$ by equivalence classes of points in $O(D, d)$. For any $\boldsymbol{V} \in O(D, d)$, the subspace $\mathrm{Sp}(\boldsymbol{V})$ can be represented by the equivalence class $[\boldsymbol{V}] = \{\boldsymbol{V}\boldsymbol{R} : \boldsymbol{R} \in O(d, d)\}$. For $\boldsymbol{V} \in O(D, d)$, the energy (3) is equivalent to

$$F(\boldsymbol{V}; \mathcal{X}) = \sum_{i=1}^{N} \|(\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^T)\boldsymbol{x}_i\|. \tag{4}$$

While both formulations are equivalent, we use (3) to formulate geodesic derivatives over $G(D, d)$ and the coordinate representation in (4) to calculate gradients. In the following, the geodesic derivative of (3) will be used to characterize the local landscape, and the gradient of (4) will be used to analyze the performance of the gradient descent algorithm we discuss later in Section 4.

One can measure the distance between subspaces in $G(D, d)$ using the principal angles. For a discussion of principal angles between subspaces, see Appendix B. Denoting the largest principal angle between L_0 and L_1 by $\theta_1(L_0, L_1)$, we can define a metric on $G(D, d)$

by $\text{dist}(L_0, L_1) = \theta_1(L_0, L_1)$. We then define a ball on the metric space $G(D, d)$ by

$$B(L, \gamma) = \{L' \in G(D, d) : \theta_1(L', L) < \gamma\}.$$

We say that an element of $O(D, d)$ lies in the ball $B(L, \gamma)$ if the subspace spanned by its columns lies in $B(L, \gamma)$.

In the following, we frequently use a construction for geodesics on the Grassmannian. For a review of this construction and necessary terminology, see Appendix B or §3.2.1 of Lerman and Zhang (2014). Suppose that the interaction dimension between L_1 and L_2 is k, that is, $k = d - \dim(L_0 \cap L_1)$. Let $\theta_1, \ldots, \theta_k$ be the nonzero principal angles for L_0 and L_1 (in decreasing order), and let the respective principal vectors for L_0 and L_1 be $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$. Finally, let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ be a complementary orthogonal basis for L_1 with respect to L_0. We can use these to parameterize a geodesic $L(t)$ with $L(0) = L_0$ and $L(1) = L_1$, where the formula is given in (62) of Appendix B. Then, following Lerman and Zhang (2014) and Lerman and Maunu (2018a), we can calculate the directional geodesic subderivative of (3) at L_0 in the direction of L_1:

$$\frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X}) \Big|_{t=0} = - \sum_{\|\boldsymbol{Q}_0 \boldsymbol{x}_i\| > 0} \frac{\sum_{j=1}^{k} \theta_j (\boldsymbol{v}_j^T \boldsymbol{x}_i)(\boldsymbol{u}_j^T \boldsymbol{x}_i)}{\|\boldsymbol{Q}_{L_0} \boldsymbol{x}_i\|}. \tag{5}$$

A subderivative of (4) with respect to $\boldsymbol{V}$ is

$$\frac{\partial}{\partial \boldsymbol{V}} F(\boldsymbol{V}; \mathcal{X}) = - \sum_{\|\boldsymbol{Q}_{\boldsymbol{V}} \boldsymbol{x}_i\| > 0} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{V}}{\|\boldsymbol{Q}_{\boldsymbol{V}} \boldsymbol{x}_i\|}. \tag{6}$$

The definition of subderivative and subdifferential as we use them are given next. For a more in depth discussion of these concepts, see, for example, Clarke (1990) and Ledyaev and Zhu (2007). In both of the derivatives (5) and (6), the sum is taken over all points that do not lie in $\text{Sp}(\boldsymbol{V})$. This restriction is what makes them both subderivatives. For any general function $f(x)$, a subderivative of f at x_0 is any number in the subdifferential $\partial f(x_0)$. In turn, the subdifferential of f at x_0 is the set of all numbers between the one-sided derivatives of f at x_0. For (5), the subdifferential is defined to be the set of all numbers between

$$a = \lim_{t \to 0^-} \frac{F(L(t); \mathcal{X}) - F(L(0); \mathcal{X})}{t} \text{ and } b = \lim_{t \to 0^+} \frac{F(L(t); \mathcal{X}) - F(L(0); \mathcal{X})}{t}.$$

In other words, the subdifferential is $[\min(a, b), \max(a, b)]$, which is the set of all instantaneous tangent slopes at $L(0)$. For the other case of (6), for any entry of $\boldsymbol{V}$, $\boldsymbol{V}_{ij}$, let $\boldsymbol{\Delta}$ be the matrix of all zeros except $\boldsymbol{\Delta}_{ij} = 1$. Then, the subdifferential of $F(\boldsymbol{V}; \mathcal{X})$ for $\boldsymbol{V}_{ij}$ is all numbers between

$$a_{ij} = \lim_{t \to 0^-} \frac{F(\boldsymbol{V} + t\boldsymbol{\Delta}; \mathcal{X}) - F(\boldsymbol{V}; \mathcal{X})}{t} \text{ and } b_{ij} = \lim_{t \to 0^+} \frac{F(\boldsymbol{V} + t\boldsymbol{\Delta}; \mathcal{X}) - F(\boldsymbol{V}; \mathcal{X})}{t}.$$

This can be generalized to any direction $\boldsymbol{\Delta}$ with $\|\boldsymbol{\Delta}\|_F = 1$, where the subdifferential is the convex hull of the one sided derivatives. We say that the subdifferential is less than a number if all of its elements are bounded above by that number, that is,

$$\partial F(L(t); \mathcal{X})|_{t=0} < M \iff a < M \ \forall \ a \in \partial F(L(t); \mathcal{X})|_{t=0}. \tag{7}$$
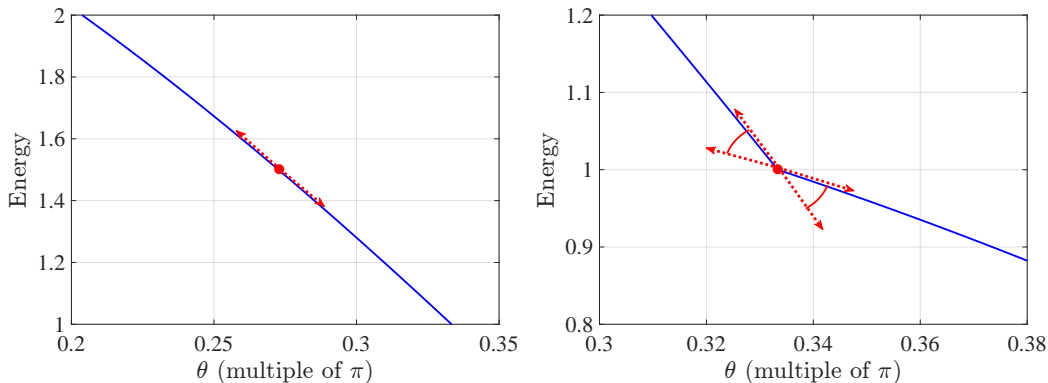
10

Figure 2: Demonstration of the derivative and subdifferential of the energy in (2). We assume that $d = 1$, $D = 2$ and identify $G(2, 1)$ with the unit circle. The energy function in (2) is thus parameterized by angle and its graph is similar to the one in the top left sub-figure of Figure 1. In both images, the slope of the red two-sided arrow represents the magnitude of the directional geodesic derivative or subderivative over $G(2, 1)$, which is the same as derivative or subderivative with respect to the representing angle. The first image shows a differentiable energy function on the given domain. In the second image, there is a value where the energy function is not differentiable. In this case, we use the subdifferential, which is the set of slopes of all lines between the acute angles formed by the two red two-sided arrows. Note that this subdifferential is bounded above by a negative number. We later prove that this property generally holds for the energy function of (2) under certain conditions.

To gain an intuition for these concepts, we display a visualization of the derivative and subdifferential for a simulated energy landscape in Figure 2. The derivative follows the standard definition from calculus on manifolds and is just the slope of the tangent line displayed on the left in Figure 2. On the other hand, at points where the function $F(L(t); \mathcal{X})$ is non-smooth at $t = 0$, we use the subdifferential instead. The extreme slopes for the subdifferential are displayed on the right in Figure 2.

In future sections, to save space, we will write the sums in (5) and (6) as $\sum_{\mathcal{X}}$ and leave the condition $\|\boldsymbol{Q_V x_i}\| > 0$ as implied. Following Section 2.5.3 of Edelman et al. (1999), to respect the geometry of the Grassmannian, the (sub)gradient of (4) is defined as

$$\nabla F(\boldsymbol{V}; \mathcal{X}) = \boldsymbol{Q_V} \frac{\partial}{\partial \boldsymbol{V}} F(\boldsymbol{V}; \mathcal{X}). \tag{8}$$

3. A Well-Tempered Landscape for Least Absolute Deviations

We assume a dataset $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subset \mathbb{R}^D$ that can be partitioned into corrupted (outlier) and uncorrupted (inlier) parts. We refer to $\mathcal{X}$ as an inlier-outlier dataset, where in the coming sections, we will more rigorously define this notion in the noiseless and noisy RSR settings. We denote the subsets of inliers and outliers in $\mathcal{X}$ as $\mathcal{X}_{\text{in}}$ and $\mathcal{X}_{\text{out}}$,

11

respectively. The corresponding data matrices for $\mathcal{X}_{\text{in}}$ and $\mathcal{X}_{\text{out}}$ are $\boldsymbol{X}_{\text{in}}$ and $\boldsymbol{X}_{\text{out}}$, where columns represent data points.

As stated previously, the basic problem of RSR is to recover the subspace L_* from an inlier-outlier dataset. In the noiseless setting one can try to exactly recover this subspace, and in the noisy setting one may try to approximately recover it. In the latter case, this means that we wish to estimate it up to a specified approximation error, which may depend on the level of noise in the inliers. In order for this problem to be well-defined, basic assumptions must be made. Indeed, if all inliers lie at the origin, then any subspace would be a solution to the RSR problem. This issue, among others, was extensively discussed in §III-A of Lerman and Maunu (2018b). Our theoretical results for recovery will depend on a condition formulated later in this section that ensures the problem is well-defined.

First, Section 3.1 discusses some statistics that play a fundamental role in our analysis. Next, Section 3.2 uses these statistics to develop the deterministic conditions that ensure the energy landscape of (2) behaves nicely around an underlying subspace in both the noiseless and noisy RSR settings.

3.1. Landscape Statistics

Equipped with the notions laid out in Section 2.2, we are ready to define some important statistics for the landscape of (2). These statistics are inspired by those originally discussed in Lerman et al. (2015), and they are later used for our stability results in Theorems 1 and 2. We first discuss the noiseless RSR setting in Section 3.1.1 and then the noisy RSR setting in Section 3.1.2.

3.1.1. The Noiseless RSR Setting

For the noiseless setting, we assume that the inliers, $\mathcal{X}_{\text{in}} \subset \mathcal{X}$, lie on a low-dimensional linear subspace $L^* \in G(D, d)$, and the rest of the points, $\mathcal{X}_{\text{out}} = \mathcal{X} \setminus \mathcal{X}_{\text{in}}$, are in $\mathbb{R}^D \setminus \{L^*\}$. We call $\mathcal{X}$ defined in this way a noiseless inlier-outlier dataset.

The permeance of the inliers in a noiseless inlier-outlier dataset is defined as

$$\mathcal{P}(\mathcal{X}_{\text{in}}) = \lambda_d \left(\sum_{\boldsymbol{x} \in \mathcal{X}_{\text{in}}} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\|\boldsymbol{x}_i\|} \right). \tag{9}$$

Here, $\lambda_d(\cdot)$ denotes the dth eigenvalue of a matrix. Notice that large values of $\mathcal{P}$ ensure that the inliers are well-distributed. In other words, they permeate throughout L_*.

We also define an alignment statistic for the noiseless inlier-outlier dataset $\mathcal{X}$. With some abuse of notation, we write $\nabla F(L; \mathcal{X})$ to refer to the gradient with respect to some basis of L, where the choice of basis does not matter. The alignment statistic of a set of outliers with respect to a subspace is

$$\mathcal{A}(\mathcal{X}_{\text{out}}, L) = \|\nabla F(L; \mathcal{X}_{\text{out}})\|_2. \tag{10}$$

It is not hard to show that (10) is invariant with respect to choice of basis for L. In effect, if this term is always small, then the outliers are not concentrated in any low-dimensional space. In other words, they are not aligned. In our later analysis, we use a simple and

illuminating bound for $\mathcal{A}$:

$$\mathcal{A}(\mathcal{X}_{\text{out}}, L) \leq \sqrt{N_{\text{out}}} \|\boldsymbol{X}_{\text{out}}\|_2. \tag{11}$$

The derivation for this bound is left to Appendix C. We note that this bound may be tight, but in most cases it is not.

We will define a stability statistic for a neighborhood of L_*, $B(L_*, \gamma)$. This neighborhood depends on a parameter γ, which fixes the maximum principal angle of subspaces in this neighborhood with L_*. Using the permeance and alignment defined in (9) and (10), the stability statistic of a noiseless inlier-outlier dataset is

$$\mathcal{S}(\mathcal{X}, L_*, \gamma) = \cos(\gamma)\mathcal{P}(\mathcal{X}_{\text{in}}) - \sup_{L \in B(L_*, \gamma)} \mathcal{A}(\mathcal{X}_{\text{out}}, L). \tag{12}$$

The simple condition required in most of our theoretical analysis is $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$. This essentially means that the amount that the inliers permeate the underlying subspace must be able to beat the alignment of the outliers with respect to any subspace.

Note that $\mathcal{S}(\mathcal{X}, L_*, 0) = \mathcal{P}(\mathcal{X}_{\text{in}}) - \mathcal{A}(\mathcal{X}_{\text{out}}, L_*)$ is a tighter stability condition than the one in (2.4) of Lerman et al. (2015). Indeed, the stability expression of Lerman et al. (2015), takes the form

$$\mathcal{S}_{\text{REAP}}(\mathcal{X}, L_*) = \frac{1}{4\sqrt{d}}\mathcal{P}_{\text{REAP}}(\mathcal{X}_{\text{in}}) - \mathcal{A}_{\text{REAP}}(\mathcal{X}_{\text{out}}, L_*). \tag{13}$$

Here, $\mathcal{P}_{\text{REAP}}$ and $\mathcal{A}_{\text{REAP}}$ are actually lower and upper bounds on the permeance and alignment defined in (9) and (10), respectively. This, together with the extra factor of $1/(4\sqrt{d})$, means $\mathcal{S}_{\text{REAP}}(\mathcal{X}, L_*)$ is not as tight as $\mathcal{S}(\mathcal{X}, L_*, \gamma)$ for $\gamma = 0$. The upside is that the REAPER alignment only needs to be examined at a single point L_*, whereas $\mathcal{S}(\mathcal{X}, L_*, \gamma)$ becomes hard to estimate as γ increases. It is not clear in general which statistic is tighter when $\gamma > 0$.

3.1.2. THE NOISY RSR SETTING

The noisy setting occurs when the inliers lie near the low-dimensional subspace rather than exactly on it. In this case, we need to be more careful with the statistics of our inlier points. For each inlier point, we write $\boldsymbol{x}_i = \boldsymbol{P}_{L_*}\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$, where $\boldsymbol{P}_{L_*}\boldsymbol{x}_i \in L_*$ and $\boldsymbol{\epsilon}_i \in L_*^{\perp}$ is added noise. Then, $\mathcal{X}^{\text{dns}} = (\boldsymbol{P}_{L_*}\mathcal{X}_{\text{in}}) \cup \mathcal{X}_{\text{out}}$ is the corresponding noiseless inlier-outlier dataset (here, the $\cdot^{\text{dns}}$ superscript stands for "de-noised"). We assume that the noise in our data is uniformly bounded by ϵ, that is, $\|\boldsymbol{x}_i - \boldsymbol{P}_{L_*}\boldsymbol{x}_i\| < \epsilon$ for all $\boldsymbol{x}_i \in \mathcal{X}_{\text{in}}$.

Some small technical issues come up with noisy RSR datasets that make the conditions harder to interpret. However, the following discussion is just a generalization of the previous section on the noiseless case after dealing with these technicalities.

To write the stability statistic in the noisy RSR setting, we must define the following set-valued functions of $\mathcal{X}_{\text{in}}$. These are defined for a unit vector $\boldsymbol{w} \in L_* \cap S^{D-1}$ and small-projection cutoff δ. They are meant to distinguish between inliers who have a projection onto $\boldsymbol{w}$ with length bigger than δ, and inliers who have a projection onto $\boldsymbol{w}$ with length less than or equal to δ. These functions are defined as

$$\mathcal{F}_0(\mathcal{X}_{\text{in}}, \boldsymbol{w}, \delta) = \{\boldsymbol{x} \in \mathcal{X}_{\text{in}} : |\boldsymbol{w}^T\boldsymbol{x}| \leq \delta\}, \tag{14}$$

$$\mathcal{F}_1(\mathcal{X}_{\text{in}}, \boldsymbol{w}, \delta) = \{\boldsymbol{x} \in \mathcal{X}_{\text{in}} : |\boldsymbol{w}^T\boldsymbol{x}| > \delta\}. \tag{15}$$

Inliers in the first set are coined "small-projection inliers", and inliers in the latter set are coined "large-projection inliers".

With these sets, our noisy inlier-outlier stability statistic is

$$
\mathcal{S}_n(\mathcal{X}, L_*, \epsilon, \delta, \gamma) = \frac{\cos(\gamma - 2\arctan(\epsilon/\delta))}{2} \min_{\boldsymbol{w} \in L_* \cap S^{D-1}} \left(\sum_{\boldsymbol{x}_i \in \mathcal{F}_1(\mathcal{X}_{\text{in}}, \boldsymbol{w}, \delta)} \frac{\boldsymbol{w}^T \boldsymbol{P}_{L_*} \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{P}_{L_*} \boldsymbol{w}}{\|\boldsymbol{P}_{L_*} \boldsymbol{x}\| + \sqrt{\epsilon^2 + \delta^2}} \right)
$$
$$
- \sqrt{\delta^2 + \epsilon^2} \max_{\boldsymbol{w} \in L_* \cap S^{D-1}} \# \left(\mathcal{F}_0(\mathcal{X}_{\text{in}}, \boldsymbol{w}, \delta) \right) - \sup_{B(L_*, \gamma)} \mathcal{A}(\mathcal{X}_{\text{out}}, L). \tag{16}
$$

This statistic is somewhat similar to what we had in the noiseless RSR setting, although now we have separated our inlier terms into two parts. The first term behaves like the permeance from the noiseless RSR setting, with the addition that small-projection inliers are trimmed. The last term is again the alignment of the outliers. The middle term is quite technical, and it is meant to capture cases when inliers may have large angle with a fixed direction of L_*. If we take $\delta \to 0$ and $\epsilon/\delta \to 0$, then the stability almost becomes our original stability, with the added factor of $1/2$ on the permeance term. If $\mathcal{S}_n(\mathcal{X}, L_*, \epsilon, \delta, \gamma) > 0$, then we will demonstrate later that recovery is possible up to accuracy $\eta = 2\arctan(\epsilon/\delta)$.

An illustration of the small-projection cutoff δ, the noise bound ϵ, and the accuracy η for noisy inliers is given in Figure 3 to help ease understanding of our statistic. For simplicity, we show the case of a one-dimensional subspace in $\mathbb{R}^2$. Here, the vectors $\boldsymbol{w} \in L_* \cap S^{D-1}$ are $(1,0)^T$ or $(-1,0)^T$. Since these two vectors are equivalent for the two functions in (14) and (15), the large-projection inliers and small-projection inliers are only determined by the magnitude of δ. Thus, the cutoff defined by a certain choice of δ in (14) and (15) corresponds to separating small and large inliers by their x-value.

We note that the statistic is by no means tight and future work should analyze how accurate these methods can be with noise. As we will discuss in Section 7, one could also study RSR in settings with high noise, such as heavy tailed noise or under the spiked model. The main point of the noisy statistic is to show that our results yield ϵ/δ-approximate recovery when the noise is uniformly bounded by ϵ. Here, δ is constrained in that the stability condition, $\mathcal{S}_n(\mathcal{X}, L_*, \epsilon, \delta, \gamma) > 0$, must hold.

3.2. The Local Landscape under Stability

In this section, we will give results that prove the local stability of the energy landscape of (2). We begin with the theorem for the noiseless RSR setting in Section 3.2.1, and then prove an analogous result for the noisy RSR setting in Section 3.2.2.

3.2.1. Stability in the Noiseless Case

We show that positivity of the stability statistic given in (12) with $0 < \gamma < \pi/2$ implies stability of L_* as a minimizer in the neighborhood $B(L_*, \gamma)$. Stability of L_* means that it is the only critical point and minimizer in $B(L_*, \gamma)$, and, at all other points in this neighborhood, there exists a direction in $G(D, d)$ such that the energy landscape looks like one of the two cases displayed in Figure 2; in other words, there is a direction of decrease.

Theorem 1 (Stability of L_*) *Suppose that a noiseless inlier-outlier dataset with an underlying subspace L_* satisfies $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$, for some $0 < \gamma < \pi/2$. Then, all points in*
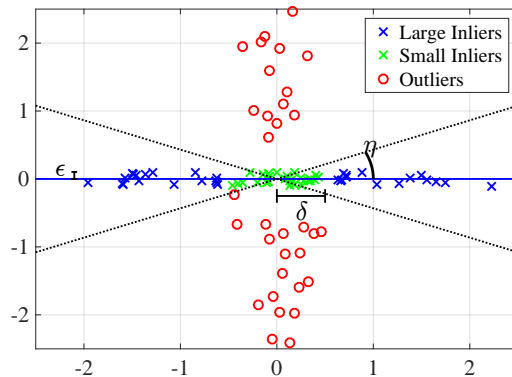
Figure 3: Demonstration of a noisy inlier-outlier dataset, where $d = 1$ and $D = 2$. The following parameters used in our analysis are also displayed: the small-projection cutoff δ, the noise bound ϵ, and the accuracy η. Note that, in this example, inliers within a δ-neighborhood of the origin are removed from the permeance calculation, and ϵ the maximum distance of the inliers to the X-axis. Our analysis guarantees recovery up to the accuracy η, which means that we would recover the x-axis within the acute angle formed by the dotted lines.

$B(L_*, \gamma) \setminus \{L_*\}$ have a subdifferential along a geodesic strictly less than $-\mathcal{S}(\mathcal{X}, L_*, \gamma)$, that is, it is a direction of decreasing energy. This implies that L_* is the only critical point and local minimizer in $B(L_*, \gamma)$.

Proof [Proof of Theorem 1] The main point behind the proof is the following statement. We show that, for any $L \in B(L_*, \gamma) \setminus \{L_*\}$, there is a geodesic $L(t)$ with $L(0) = L$, and an open interval around 0, $\mathcal{I} = (\theta_1(L, L_*) - \gamma, \delta(L))$, for some $\delta(L) > 0$, such that

$$\text{For } t \in \mathcal{I}, \ \theta_1(L(t), L_*) \text{ is a strictly decreasing function;} \qquad (17)$$

$$\text{For } t \in \mathcal{I}, \ F(L(t); \mathcal{X}) \text{ is a strictly decreasing function.} \qquad (18)$$

In simple words, the function $F(L(t); \mathcal{X})$ is decreasing as $L(t)$ moves closer to L_*. This implies that L_* is the only critical point and minimizer in $B(L_*, \gamma)$ by a perturbation argument, which we will explicitly state at the end of the proof.

Fix a subspace $L \in B(L_*, \gamma) \setminus \{L_*\}$, and let the principal angles between L and L_* be $\theta_1, \ldots, \theta_d$. Also, choose a set of corresponding principal vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_d$ for L and L_*, respectively, and let $l \geq 1$ be the maximum index such that $\theta_1 = \cdots = \theta_l$. We let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l$ be complementary orthogonal vectors for $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_l$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_l$. For $t \in [0, 1]$, we form the geodesic

$$L(t) = \text{Sp}(\boldsymbol{v}_1 \cos(t) + \boldsymbol{u}_1 \sin(t), \ldots, \boldsymbol{v}_l \cos(t) + \boldsymbol{u}_l \sin(t), \boldsymbol{v}_{l+1}, \ldots, \boldsymbol{v}_d). \qquad (19)$$

Notice that this geodesic is parameterized by arclength in terms of the metric defined in Section 2.2. This geodesic moves only the l furthest directions of $L(0)$ towards L_* and fixes those directions that are closer than θ_1. This geodesic certainly satisfies (17). We

have also removed dependence on θ_1, since this unnecessarily impacts the magnitude of the geodesic subderivative (5). Following (5) with no dependence on $\theta_1, \ldots, \theta_l$ and then using (6) and (8), we have

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X})\Big|_{t=0} &= -\sum_{j=1}^{l} \sum_{\mathcal{X}} \frac{\boldsymbol{v}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{u}_j}{\|\boldsymbol{Q}_L \boldsymbol{x}_i\|} \\
&= -\sum_{j=1}^{l} \left(\sum_{\mathcal{X}_{\text{in}}} \frac{\boldsymbol{v}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{u}_j}{\|\boldsymbol{Q}_L \boldsymbol{x}_i\|} + \sum_{\mathcal{X}_{\text{out}}} \frac{\boldsymbol{v}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{u}_j}{\|\boldsymbol{Q}_L \boldsymbol{x}_i\|} \right) \\
&\leq -\sum_{j=1}^{l} \left(\frac{\cos(\theta_1) \sin(\theta_1)}{\sin(\theta_1)} \sum_{\mathcal{X}_{\text{in}}} \frac{\boldsymbol{y}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{y}_j}{\|\boldsymbol{x}_i\|} - \sup_{\boldsymbol{V} \in B(L_*, \gamma)} \|\nabla F(\boldsymbol{V}; \mathcal{X}_{\text{out}})\|_2 \right) \\
&\leq -l\mathcal{S}(\mathcal{X}, L_*, \gamma) \leq -\mathcal{S}(\mathcal{X}, L_*, \gamma) < 0.
\end{aligned}
\tag{20}
$$

The third line is obtained by noting that the inliers and the vectors $\boldsymbol{y}_j$, $1 \leq j \leq l$, are contained L_*, the vectors $\boldsymbol{v}_j$, $1 \leq j \leq l$, are contained in L, and $\theta_1 = \ldots = \theta_l$. Therefore, for all $1 \leq j \leq l$ and all inliers $\boldsymbol{x}_i$, $\boldsymbol{v}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{u}_j \geq \cos(\theta_1) \sin(\theta_1)(\boldsymbol{y}_j^T \boldsymbol{x}_i)^2$. In the third line we also used (6) and (8) for the outlier term and maximized it over $B(L_*, \gamma)$. Thus, (20) implies that every subspace in $\overline{B(L_*, \gamma)} \setminus \{L_*\}$ has a direction with negative local subderivative.

The argument above was for a special subderivative at $L(0)$. We now extend this argument to show that this implies that the subdifferential at every point is bounded above by $-\mathcal{S}(\mathcal{X}, L_*, \gamma)$. On the one hand, if $L(0)$ contains no points of $\mathcal{X}$, then the subderivative corresponds to the actual derivative, and we therefore have that $F(L(t); \mathcal{X})$ is decreasing, as desired.

On the other hand, assume that $L(0) \cap \mathcal{X}$ is non-empty. Since $\mathcal{X}$ is finite, there exists a neighborhood of 0, $(-\omega, \omega)$, for $0 < \omega < \pi/2$, such that $L(t)$ has empty intersection with $\mathcal{X} \setminus L(0)$, that is,

$$
\{\boldsymbol{x} \in \mathcal{X} : \boldsymbol{x} \in L(t), \text{ for some } t \in (-\omega, \omega)\} = L(0) \cap \mathcal{X}.
\tag{21}
$$

Notice that the changing directions in (19) form a geodesic along $G(D, l)$ while the other directions are fixed (i.e., $L(t) \supset \text{Sp}(\boldsymbol{v}_{l+1}, \ldots, \boldsymbol{v}_d)$ for all t). Thus, derivatives of $F(L(t); \mathcal{X})$ over $G(D, d)$ are equivalent to derivatives of $F(\tilde{L}(t); \boldsymbol{Q}_{\boldsymbol{v}_{l+1}, \ldots, \boldsymbol{v}_d} \mathcal{X})$ over $G(D, l)$, where $\tilde{L}(t) = \text{Sp}(\boldsymbol{v}_1 \cos(t) + \boldsymbol{u}_1 \sin(t), \ldots, \boldsymbol{v}_l \cos(t) + \boldsymbol{u}_l \sin(t)))$. Here, $\boldsymbol{Q}_{\boldsymbol{v}_{l+1}, \ldots, \boldsymbol{v}_d}$ is the projection onto the orthogonal complement of $\text{Sp}(\boldsymbol{v}_{l+1}, \ldots, \boldsymbol{v}_d)$, which is applied to all points of $\mathcal{X}$. By (21), this geodesic over $G(D, l)$ has the property

$$
\tilde{L}(t) \cap \mathcal{X} = \emptyset, \ \forall \, t \in (-\omega, 0) \cup (0, \omega).
$$

Consequently, $F(\tilde{L}(t); \boldsymbol{Q}_{\boldsymbol{v}_{l+1}, \ldots, \boldsymbol{v}_d} \mathcal{X})$ is continuously differentiable on $(-\omega, 0)$ and $(0, \omega)$ as a function over $G(D, l)$, and it is apparent that the bounds in (20) also hold for all of these derivatives. Putting these facts together, by the continuity of the derivatives of $F(\tilde{L}(t); \boldsymbol{Q}_{\boldsymbol{v}_{l+1}, \ldots, \boldsymbol{v}_d} \mathcal{X})$ for $t \in (-\omega, 0) \cup (0, \omega)$, the subdifferential of $F(L(t); \mathcal{X})$ at $t = 0$ is bounded above by $-\mathcal{S}(\mathcal{X}, L_*, \gamma)$, which in turn implies that (18) holds.

Finally, there are clearly no other critical points than L_* in $B(L_*, \gamma)$, because every point has a direction of decrease. To show that (17) and (18) imply that L_* is a local minimizer, consider a one-dimensional perturbation of L_*, L'. In other words, $\theta_1(L_*, L') > 0$

and $\theta_2(L_*, L') = 0$. Then, (17) and (18) imply that if $L(t)$ is the geodesic between L' and L_*, then $F(L(t); \mathcal{X})$ is decreasing for $t \in (\theta_1(L', L_*) - \gamma, \theta_1(L', L_*))$. The more general perturbation case is just an extension of this argument. Indeed, a d-dimensional perturbation may be written as a sequence of one-dimensional perturbations. ∎

3.2.2. Stability with Small Noise

The following theorem generalizes Theorem 1 for the case of a noisy inlier-outlier dataset. The proof of this theorem is left to Appendix D.1, and it essentially follows that of the noiseless RSR setting with the altered stability statistic. Here, we only guarantee that there is a large region with no critical points up to a precision of $\eta = 2\arctan(\epsilon/\delta)$, which is determined by the noise level and inlier permeance.

Theorem 2 (Stability of $B(L_*, \eta)$ with Noise) *Assume a noisy inlier-outlier dataset, with an underlying subspace L_* and noise parameter $\epsilon > 0$, that satisfies for some $\delta > \epsilon > 0$ the stability condition $\mathcal{S}_n(\mathcal{X}, L_*, \epsilon, \delta, \gamma) > 0$. Let $\eta = 2\arctan(\epsilon/\delta)$ and assume further that $\eta < \gamma$. Then, all points in $B(L_*, \gamma) \setminus B(L_*, \eta)$ have a subdifferential along a geodesic strictly less than $-\mathcal{S}_n(\mathcal{X}, L_*, \epsilon, \delta, \gamma)$, that is, it is a direction of decreasing energy. This implies that the only local minimizers and saddle points in $B(L_*, \gamma)$ are in $B(L_*, \eta)$.*

4. A Geodesic Gradient Method for RSR and its Guarantees

In this section, we discuss a geodesic gradient descent method for minimizing (3). First, Section 4.1 gives the details of our algorithm. After laying out the algorithm, we discuss convergence to a local minimizer under deterministic conditions in Section 4.2. Then, Section 4.3 shows that the PCA d-subspace is a good initializer under a separate deterministic condition.

4.1. Minimization by Geodesic Gradient Descent

We use gradient descent to minimize (4). Denoting the singular value decomposition of the negative gradient by $-\nabla F(\boldsymbol{V}; \mathcal{X}) = \boldsymbol{U\Sigma W}^T$, Theorem 2.3 of Edelman et al. (1999) states that the geodesic starting at $\boldsymbol{V}(0) = \boldsymbol{V}$ with $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{V}(t)|_{t=0} = -\nabla F(\boldsymbol{V}; \mathcal{X})$ is

$$\boldsymbol{V}(t) = \boldsymbol{VW}\cos(\boldsymbol{\Sigma}t)\boldsymbol{W}^T + \boldsymbol{U}\sin(\boldsymbol{\Sigma}t)\boldsymbol{W}^T. \tag{22}$$

Here, sin and cos are the typical matrix sin and cos, defined by their corresponding power series.

We develop a geodesic gradient descent method using the construction in (22). At a point $\boldsymbol{V}^k$, we may choose a value of t and move along the geodesic to the next iterate. For a sequence of step-sizes $(t^k)_{k \in \mathbb{N}}$, the sequence of subspaces is defined recursively by $\boldsymbol{V}^{k+1} = \boldsymbol{V}^k(t^k)$. The full algorithm with a specific choice of step-size is given in Algorithm 1 and is referred to as GGD. The complexity of this algorithm is $O(TNDd)$, where T is the number of iterations. Note that we use a piecewise constant scheme for the step-sizes, which starts with step-size $t^1 = s$, and every K iterations the step-size is shrunk by a factor of $1/2$.

Algorithm 1 RSR by Geodesic Gradient Descent (GGD)

1: **Input:** dataset $\mathcal{X}$, subspace dimension d, initial step-size s, tolerance τ, constant step interval length K
2: **Output:** $\boldsymbol{V}^* \in O(D, d)$, whose columns span the robust subspace
3: $\boldsymbol{V}^1 = PCA(\boldsymbol{X}, d)$
4: $k = 1$, $s = 1$
5: **while** $\theta_1(\boldsymbol{V}^k, \boldsymbol{V}^{k-1}) > \tau$ or $k = 1$ **do**
6:     Compute $\nabla F(\boldsymbol{V}^k; \mathcal{X})$ by (6) and (8)
7:     Compute the SVD $\boldsymbol{U}^k \boldsymbol{\Sigma}^k \boldsymbol{W}^k = -\nabla F(\boldsymbol{V}^k; \mathcal{X})$
8:     $s^k = s/2^{\lfloor k/K \rfloor}$
9:     $\boldsymbol{V}^{k+1} = \boldsymbol{V}^k \boldsymbol{W}^k \cos(\boldsymbol{\Sigma}^k s^k) \boldsymbol{W}^{kT} + \boldsymbol{U}^k \sin(\boldsymbol{\Sigma}^k s^k) \boldsymbol{W}^{kT}$ ▷ (22)
10:     $k = k + 1$
11: **end while**

The next section provides convergence guarantees for GGD with both the piecewise constant step-size of Algorithm 1 and with step-size $s/\sqrt{k}$. We later compare these two examples of step-size on a simulated dataset in Figure 6. The factor of $1/2$ in line 8 is called the shrink-factor and could be chosen to be any fraction. Later, in our experiments, we display the convergence properties of GGD with two choices of this factor.

4.2. Local Convergence of GGD Under Stability

In this section, we give convergence guarantees for GGD. Theorem 3 shows that the convergence is sublinear under the deterministic conditions of Theorem 1 for step-size $t^k = s/\sqrt{k}$. Then, Theorem 4 shows that the convergence of Algorithm 1 is linear under a slightly stronger assumption. These results are all for the noiseless RSR setting. However, all of these results can be extended to the noisy RSR setting in a simple fashion using the notions described in Section 3.1.2.

As a reminder, Theorem 1 implies that if $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$ in the noiseless RSR setting then L_* is the only limit point in $B(L_*, \gamma)$ for GGD. In other words, there is no need to worry about saddle points or non-optimal critical points in this neighborhood of L_*. In the following theorem, we give a general sublinear convergence bound for GGD in the local neighborhood considered in Theorem 1. This implies that the algorithm can exactly recover the underlying subspace in the noiseless RSR problem. The step-size used in this theorem is $t^k = s/\sqrt{k}$ at iteration k instead of the piecewise constant scheme seen in Algorithm 1. The proof of this theorem is left to Appendix D.2.

Theorem 3 (Noiseless Sublinear Convergence) *Suppose that $\mathcal{X}$ is an inlier-outlier dataset with an underlying subspace L_*. Suppose also that there exists $0 < \gamma < \pi/2$ such that $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$ and that the initial GGD iterate is $\boldsymbol{V}^1 \in B(L_*, \gamma)$. Then, for sufficiently small s as input (which may depend on $\mathcal{S}(\mathcal{X}, L_*, \gamma)$, d, D, N_{in}, and N_{out}), modified GGD with $t^k = s/\sqrt{k}$ converges to L_* with rate $\theta_1(L_k, L_*) < C/\sqrt{k}$, for some constant C.*

While the rate $O(1/\sqrt{k})$ matches typical results in non-smooth optimization, faster convergence is desirable. With the piecewise constant step-size given in Algorithm 1 and a further deterministic condition, GGD linearly converges to the underlying subspace L_*.

Theorem 4 (Noiseless Linear Convergence) *Suppose that $\mathcal{X}$ is an inlier-outlier dataset with an underlying subspace L_*. Suppose also that there exists $0 < \gamma < \pi/2$ such that $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$ and that the initial GGD iterate is $\boldsymbol{V}^1 \in B(L_*, \gamma)$. Assume further that*

$$\inf_{L \in B(L_*,\gamma)\setminus\{L_*\}} \frac{1}{4} \left| \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X})|_{t=0} \right| > \sup_{L \in B(L_*,\gamma)\setminus\{L_*\}} \sum_{\mathcal{X}\cap L} 2\|\boldsymbol{x}_i\|, \tag{23}$$

where $L(t)$ is a geodesic parameterized by arclength from L through L_*. Then, for sufficiently large K and sufficiently small s as input (which may depend on $\mathcal{S}(\mathcal{X}, L_*, \gamma)$, d, D, N_{in}, and N_{out}), the sequence generated by Algorithm 1 converges linearly to L_*.

Remark 5 *As mentioned earlier, Theorems 3 and 4 can also be extended to the noisy RSR setting with more complicated statements. Indeed, these extensions follow the same ideas as Theorem 2 (which extends Theorem 1 to the noisy RSR setting).*

We remark that the restriction in (23) can be weakened, although it results in a more complicated theorem statement: for clarity, we show the simpler version in the theorem statement. We refer to (23) as the strong gradient condition. In general, the sum in the right hand side of (23) only contains a few points when the inliers and outliers are not too linearly dependent. For example, consider the case of inliers lying in general position within the subspace and outliers lying in general position. This is the case if all D-subsets of $\mathcal{X}_{\mathrm{out}}$ are linearly independent and all d-subsets of $\mathcal{X}_{\mathrm{in}}$ are linearly independent. Under this assumption, the right hand side contains less than $2(d-1)$ points. On the other hand, the left hand side may be sufficiently large for a wide range of statistical models of data. For more interpretation of this condition for a specific model of data, we point the reader to Lemma 20, where we demonstrate a simple way in which the condition might hold.

We are not able to give an explicit bound on the rate of convergence factor, which depends on certain statistical characteristics of the data. While we do not currently have estimates of this factor and it may depend on N, D, and d, we expect this dependence to not be too bad, especially based on our numerical experiments.

Proof [Proof of Theorem 4] The proof of this theorem is a consequence of the following lemmas. The proof of these lemmas is deferred to Appendix D.3.

The first lemma locally bounds above the increase in cost around L_*.

Lemma 6 *If $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$, then*

$$F(L; \mathcal{X}) - F(L_*; \mathcal{X}) < 2\theta_1(L, L_*) \sum_{\mathcal{X}_{\mathrm{in}}} \|\boldsymbol{x}_i\|, \ \ \forall L \in B(L_*, \gamma). \tag{24}$$

Notice that $\theta_1(L, L_*)$ is a measure of distance between L and L_*. The next lemma bounds the magnitude of the gradient in $B(L_*, \gamma) \setminus L_*$.

Lemma 7 *If $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$, then, for some $C_1 > 0$ (that depends on the data),*

$$\left| \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X})|_{t=0} \right| > C_1 \sum_{\mathcal{X}_{\mathrm{in}}} \|\boldsymbol{x}_i\|, \ \ \forall L \in B(L_*, \gamma) \setminus L_*, \tag{25}$$

where $L(t)$ is a geodesic parameterized by arclength from L through L_*.

19

Finally, the third lemma bounds the decrease in cost between consecutive iterates.

Lemma 8 *If $L_k \in B(L_*, \gamma)$ and (23) holds, then there exists $c_0 > 0$ such that for each step-size choice $t^k = c\theta_1(L_k, L_*)$ with $c < c_0$,*

$$F(L_k) - F(L_{k+1})) \geq \frac{c\theta_1(L_k, L_*)}{2} \left| \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X})|_{t=0} \right|, \tag{26}$$

where $L(t)$ is a geodesic parameterized by arclength from L_k through L_*.

Choosing the step-size $t^k = c\theta_1(L_k, L_*)$, with c coming from Lemma 8, and combining the results of Lemmas 6, 7 and 8, we find that

$$F(L_{k+1}; \mathcal{X}) - F(L_*; \mathcal{X}) < F(L_k; \mathcal{X}) - F(L_*; \mathcal{X}) - \frac{c\theta_1(L_k, L_*)}{2} \left| \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X})|_{t=0} \right| \tag{27}$$

$$\leq F(L_k; \mathcal{X}) - F(L_*; \mathcal{X}) - \frac{C_1}{2} c\theta_1(L_k, L_*) \left(\sum_{\mathcal{X}_{\mathrm{in}}} \|\boldsymbol{x}_i\| \right)$$

$$\leq F(L_k; \mathcal{X}) - F(L_*; \mathcal{X}) - \frac{C_1 c}{4} (F(L_k; \mathcal{X}) - F(L_*; \mathcal{X}))$$

$$\leq (1 - C_2)(F(L_k; \mathcal{X}) - F(L_*; \mathcal{X})),$$

where $0 < C_2 < 1$. Here, $(1 - C_2)$ is the rate of convergence factor, which depends on d. Thus, if one could choose the step-size $t^k = c\theta_1(L_k, L_*)$, then the sequence of costs $F(L_k; \mathcal{X})$ would converge linearly to $F(L_*; \mathcal{X})$.

For all $L \in B(L_*, \gamma) \setminus \{L_*\}$, letting $a = \theta_1(L, L_*)$, we find that

$$F(L; \mathcal{X}) - F(L_*; \mathcal{X}) = \int_{t=0}^{a} \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X}) \mathrm{d}t \geq a \inf_{L' \in B(L_*, \gamma)} \left| \frac{\mathrm{d}}{\mathrm{d}t} F(L'(t); \mathcal{X}) \right| \tag{28}$$

$$\geq a\mathcal{S}(\mathcal{X}, L_*, \gamma),$$

where $L(t)$ is the geodesic parameterized by arclength from L_* through L and $L'(t)$ is the geodesic parameterized by arclength from L_* through L'. The last inequality follows from the argument in Theorem 1, and in particular from (20). Thus, for $C_3 = 1/\mathcal{S}(\mathcal{X}, L_*, \gamma)$, where $0 < C_3 < \infty$, (28) implies that

$$\theta_1(L_{k+1}, L_*) \leq C_3(F(L_{k+1}; \mathcal{X}) - F(L_*; \mathcal{X})). \tag{29}$$

This means that linear convergence of the energy sequence, $(F(L_k; \mathcal{X}))_{k \in \mathbb{N}}$, gives linear convergence of the iterates, $(L_k)_{k \in \mathbb{N}}$.

So far, we have shown that there exists a c such that choosing step-size $t^k = c\theta_1(L_k, L_*)$ leads to linear convergence of L_k to L_*. However, this choice of step-size is purely theoretical, since in practice we would not know $\theta_1(L_k, L_*)$ at each iteration. We must now rectify this choice of step-size with that used in Algorithm 1.

Suppose a constant step-size s and a constant c satisfying the above argument. Then, the sequence $(L_k)_{k \in \mathbb{N}}$ at least converges linearly to an element of the set $B(L_*, s/c)$ because (27) holds as long as $t^k = s < c\theta_1(L_k, L_*)$. If instead the constant step-size is $s/2$, we will get

linear convergence to an element of the set $B(L_*, s/2c)$, albeit at a slower rate. Notice that, at each shrinking step (i.e., switching to step-size $s/2$ from s), if the sequence has already reached $B(L_*, s/c)$, we can bound the rate of convergence factor $(1 - C_2)$ by

$$\left(1 - \frac{s/2}{c\theta_1(L_k, L_*)}\frac{C_1 c}{4}\right) \leq \left(1 - \frac{C_1 c}{8}\right). \tag{30}$$

Further, for all subsequent steps where $s/2 < c\theta_1(L_k, L_*)$, we have that the rate of convergence factor is strictly less than $1 - C_1 c/(8\sqrt{d}(d+1))$. Thus, as long as the time between shrinking is large enough, we are guaranteed to have linear convergence. If the first step-size is allowed to run long enough, we find that the number of steps m between shrinking needs to be at most

$$\left(1 - \frac{C_1 c}{8}\right)^m < \frac{1}{2}.$$

$\blacksquare$

4.3. Complete Guarantee with PCA Initialization

Notice that the previous section does not give a complete guarantee because the results are local. In other words, they assume that we first initialize in $B(L_*, \gamma)$ and then run GGD. To make these results practical, we will show that it is possible to initialize in this neighborhood under a simple deterministic condition. As before with Theorems 3 and 4, we consider only the noiseless RSR setting here. Extensions to the noisy case do not require much more effort.

The result of this section shows that PCA initializes in $B(L_*, \gamma)$ under a similar deterministic condition to $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$. This is quantified in the following lemma.

Lemma 9 *Suppose that, for a noiseless inlier-outlier dataset,*

$$\sqrt{2}\sin(\gamma)\lambda_d(\boldsymbol{X}_{\text{in}}\boldsymbol{X}_{\text{in}}^T) - \|\boldsymbol{X}_{\text{out}}\|_2^2 > 0. \tag{31}$$

Then, $L_{PCA} \in B(L_*, \gamma)$.

Proof The proof of this lemma is a direct consequence of the Davis-Kahan $\sin\theta$ Theorem (Davis and Kahan, 1970), which has a nice formulation in Vu et al. (2013). Let $L_* \in G(D, d)$ span the principal d-subspace of $\boldsymbol{X}_{\text{in}}\boldsymbol{X}_{\text{in}}^T$ and $L_{PCA} \in G(D, d)$ span the principal d-subspace of $\boldsymbol{X}_{\text{out}}\boldsymbol{X}_{\text{out}}^T + \boldsymbol{X}_{\text{in}}\boldsymbol{X}_{\text{in}}^T$. Then, applying Corollary 3.1 of Vu et al. (2013) to these matrices yields

$$|\sin(\theta_1(L_*, L_{PCA}))| \leq \sqrt{2}\frac{\|\boldsymbol{X}_{\text{out}}\|^2}{\lambda_d(\boldsymbol{X}_{\text{in}}\boldsymbol{X}_{\text{in}}^T)}. \tag{32}$$

Thus, if (31) holds, then we are guaranteed that

$$|\sin(\theta_1(L_*, L_{PCA}))| < |\sin(\gamma)|. \tag{33}$$

$\blacksquare$

The condition required in (31) bears some nice similarity to the earlier condition of Theorems 1 and 3, $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$. Here the first term in (31) is like the inlier permeance and the second term is like the outlier alignment.

In summary, if both of the conditions $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$ and (31) hold, then GGD with step-size $s/\sqrt{k}$ exactly recovers L_* with convergence rate $O(1/\sqrt{k})$. If we additionally have that (23) holds, then GGD with the step-size in Algorithm 1 linearly converges to L_*.

5. Guarantees for Specific Statistical Models

In this section, we discuss some statistical models of data and determine when they satisfy the assumptions of our theorems. These models are meant to illustrate that our conditions are satisfied in a wide range of RSR examples, and they begin to explore the recovery limits of our algorithm. These results should provide some useful context and lead to easier interpretation of the general conditions given in Section 3 and Section 4.

The main idea behind the study of statistical models of data is to compare the theoretical guarantees of various algorithms with a given choice of metric. A natural metric for this purpose is the signal to noise ratio (SNR). The SNR is the ratio of inliers to outliers, $N_{\text{in}}/N_{\text{out}}$, and we are interested in the minimal SNR that allows exact (or sufficiently near) recovery of an RSR algorithm under a given set of assumptions on the data. In other words, under further assumptions on the data generating model, we want to derive a more interpretable condition than $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$.

Due to the properties of these models, different SNR bounds may arise for different regimes of sample size. A first common small sample regime assumes that $N = O(D)$, where the SNR bounds are usually higher to account for the increased variation in the data. Another regime of slightly larger samples is obtained when $N = O(D^p)$ and $p > 1$ is sufficiently small. Under some special statistical models, the SNR may decrease as p increases. That is, a larger fraction of outliers can be tolerated with larger orders of sample size. A third regime uses very large, and possibly arbitrarily large, N. We refer to the SNR bound of this third case as the very large N regime. Under some very special models of data, the SNR bound can go to zero in this regime as the sample size increases. However, in this case, the sample size must depend on the SNR itself.

In the following, not only will we show almost state-of-the-art results for GGD on the Haystack Model of Lerman et al. (2015), but we will also demonstrate how our convergence theorem holds for other more general models of data. This is an important step towards understanding how RSR algorithms perform outside of the simple Haystack Model. The general statistical models considered below assume that the inliers and outliers are sampled from probability distributions that obey certain assumptions. These models are the following:

- Assumptions on the outliers:
  - Bounded support distributions
  - Sub-Gaussian distributions
- Assumptions on the inliers:

- Continuous, bounded support distributions
- Sub-Gaussian distributions

In each of these cases, the inliers and outliers will be assumed to be i.i.d. samples from the given distributions. The precise definition of each of these distributions will be given later.

In Section 5.1, we bound the alignment of outliers under the above outlier models. Then, in Section 5.2, we bound the permeance of inliers under the above inlier models. The goal of these first two subsections is to understand how each part of $\mathcal{S}(\mathcal{X}, L_*, \gamma)$ behaves on its own. After this, in Section 5.3 we prove that $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$ under certain conditions on these models of inliers and outliers. Next, we show in Section 5.4 that PCA can initialize in $B(L_*, \gamma)$ in a wide range of cases. Then, Section 5.5 gives an in depth discussion of the Haystack Model, where we show that GGD with step-size $s/\sqrt{k}$ has almost state-of-the-art guarantees. The discussion considers the previously mentioned three regimes of sample size. Finally, Section 5.6 gives an idea of how statistical models can also ensure that the strong gradient condition in Theorem 4 holds, which gives more evidence that the method may converge linearly in practice.

5.1. Outlier Distributions with Restricted Alignment

We explain the two assumptions on outliers listed above, which lead to bounds on the alignment. We first discuss bounded distributions in Section 5.1.1 and then discuss sub-Gaussian distributions in Section 5.1.2.

5.1.1. Bounded Support Distributions

We consider the case of outliers drawn from a distribution with bounded support. This assumption is needed because our bound on the alignment scales like the spectral norm of $\boldsymbol{X}_{\text{out}}$, which can be very large for even a single large outlier. An outlier distribution of this type has the form

$$\mathcal{X}_{\text{out}} \sim \mu, \ \mu(\mathbb{R}^D \setminus B(\mathbf{0}, M)) = 0, \tag{34}$$

where μ represents the probability measure and M is a uniform bound on the magnitude of the outliers. In this case, we have the worst-case bound

$$\|\boldsymbol{X}_{\text{out}}\|_2 < M\sqrt{N_{\text{out}}}. \tag{35}$$

In the special case where $\mathcal{X}_{\text{out}} \sim \text{Unif}(B(\mathbf{0}, M))$, the following bound was provided in Lemma 8.4 of Lerman et al. (2015):

$$\|\boldsymbol{X}_{\text{out}}\|_2 \leq M\left(\sqrt{\frac{N_{\text{out}}}{D - 0.5}} + \sqrt{2} + \frac{t}{\sqrt{D - 0.5}}\right), \text{ w.p. at least } 1 - 1.5e^{-t^2}. \tag{36}$$

We remark that (35) holds under any sampling from a bounded distribution and (36) holds under i.i.d. sampling of a special distribution.

From these bounds and (11), we get a sense of how the alignment scales for different types of outliers. When outliers are more adversarial but still bounded, the alignment scales like $O(N_{\text{out}})$. On the other hand, when outliers have the special distribution $\text{Unif}(B(\mathbf{0}, M))$, we can bound the alignment by $N_{\text{out}}/\sqrt{D}$. Later, in Theorem 17, we show how to improve this to $O(N_{\text{out}}/\sqrt{D(D - d)})$, due to the fact that the bound in (11) is not tight.

5.1.2. SUB-GAUSSIAN DISTRIBUTIONS

Rather than assume that the outliers are bounded, we can instead assume they come from a sub-Gaussian distribution. In this case, we have the following lemma.

Lemma 10 *Suppose that the outliers follow a sub-Gaussian distribution with covariance* $\Sigma_{\mathrm{out}}/D_{\mathrm{out}}$, *which has rank* D_{out}. *Then,*

$$\|\boldsymbol{X}_{\mathrm{out}}\|_2 \le \left\|\boldsymbol{\Sigma}_{\mathrm{out}}^{1/2}\right\|_2 \left(2\frac{\sqrt{N_{\mathrm{out}}}}{\sqrt{D_{\mathrm{out}}}} + C\right), \tag{37}$$

with probability at least $1 - e^{-cN_{\mathrm{out}}}$. *Here,* c *and* C *depend on the sub-Gaussian norm of* $\boldsymbol{\Sigma}_{\mathrm{out}}^{-1/2}\boldsymbol{x}$, *where* $\boldsymbol{x}$ *follows the outlier distribution.*

Proof Note that the transformed data, $\boldsymbol{\Sigma}_{\mathrm{out}}^{-1/2}\boldsymbol{X}_{\mathrm{out}}$, is isometric. Therefore, we bound the spectral norm of the outliers by

$$\|\boldsymbol{X}_{\mathrm{out}}\|_2 = \|\boldsymbol{\Sigma}_{\mathrm{out}}^{1/2}\boldsymbol{\Sigma}_{\mathrm{out}}^{-1/2}\boldsymbol{X}_{\mathrm{out}}\|_2 \le \|\boldsymbol{\Sigma}_{\mathrm{out}}^{1/2}\|_2\|\boldsymbol{\Sigma}_{\mathrm{out}}^{-1/2}\boldsymbol{X}_{\mathrm{out}}\|_2.$$

We can apply Theorem 5.39 of Vershynin (2012b) to the last term in this inequality. This yields the bound in (37). ∎

5.2. Permeating Inlier Distributions

We will look at two assumptions that yield permeance of inliers. First, Section 5.2.1 examines continuous distribution inliers, and then Section 5.2.2 looks at sub-Gaussian inliers.

5.2.1. BOUNDED CONTINUOUS INLIER DISTRIBUTIONS

An i.i.d. sample from a distribution with a continuous density lies in general position with probability 1. In the case of a continuous distribution on a subspace L_*, this means that no d-subset of them lies on a $d-1$-dimensional subspace. Some slightly stronger assumptions are also needed to easily prove that the inlier permeance has a nontrivial lower bound, which includes the distribution having bounded support, although other assumptions could be used. We will refer to these distributions as bounded continuous distributions. The proof of this proposition is given in Appendix D.4.1.

Proposition 11 *Suppose that the inliers are sampled from a distribution that has a continuous density with respect to the uniform measure on* L_*. *Suppose further that this distribution has mean zero and has support contained in* $B(\boldsymbol{0}, M) \cap L_*$, *for some constant* M. *Then,*

$$\mathcal{P}(\mathcal{X}_{\mathrm{in}}) \gtrsim \frac{N_{\mathrm{in}}}{M} \min_{\boldsymbol{v} \in L_* \cap S^{D-1}} \mathrm{Var}(\boldsymbol{v}^T\boldsymbol{x}), \ w.h.p. \tag{38}$$

Here, the probability goes to 1 and the permeance goes to ∞ as $N_{\mathrm{in}} \to \infty$.

5.2.2. Sub-Gaussian Inlier Distributions

We show how the assumption of a sub-Gaussian distribution provides a lower bound for the permeance of inliers. The proof of this theorem is given in Appendix D.4.2.

Proposition 12 *Suppose that inliers are sampled i.i.d. from a sub-Gaussian distribution with covariance $\mathbf{\Sigma}_{\text{in}}/d$, which has rank d. Then, for $0 < a < 1$ satisfying $(1-a)^2 N_{\text{in}} > C_1^2 d$,*

$$\mathcal{P}(\mathcal{X}_{\text{in}}) \geq \frac{\lambda_d(\mathbf{\Sigma}_{\text{in}})}{\lambda_1(\mathbf{\Sigma}_{\text{in}})^{1/2}} (1-a)^2 \frac{N_{\text{in}}}{d} + O\left(\sqrt{\frac{N_{\text{in}}}{d}}\right), \quad \text{w.p. at least } 1 - 4e^{-c_1 a^2 N_{\text{in}}}. \tag{39}$$

Here, c_1 and C_1 are constants that depend on the sub-Gaussian norms of $\mathbf{\Sigma}_{\text{in}}^{-1/2}\boldsymbol{x}$ and $\widetilde{\mathbf{\Sigma}_{\text{in}}^{-1/2}\boldsymbol{x}}$, where $\boldsymbol{x}$ is a random vector that follows the inlier distribution.

Notice that the choice of a here affects both the bound and the probability. One could, in principal, choose $a = N^{-1/2+\epsilon}$ for some constant $\epsilon > 0$ and still achieve an overwhelming probability bound.

5.3. Combining Statistical Models to Enforce $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$

In this section, we explicitly compare the permeance and alignment bounds for these statistical models of data to see when we can expect to have $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$, which is the essential assumption in Theorems 1 and 3. Together with the result of the next section on PCA initialization, this implies that GGD exactly recovers L_* provided that the SNR is appropriately bounded from below in these models. First, Section 5.3.1 will look at the case of bounded outliers and bounded continuous inliers. Then, Section 5.3.2 will discuss the case of sub-Gaussian inliers and outliers in what we call the Generalized Haystack Model.

5.3.1. Bounded Outliers and Bounded Continuous Inliers

Under the assumption of bounded outliers and bounded continuous inliers, we can guarantee $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$ for large enough SNR and sample sizes. This results in the following proposition.

Proposition 13 (Stability with Bounded Outliers and Continuous Bounded Inliers)
Suppose that the outliers follow a bounded distribution and the inliers follow a mean zero, bounded distribution with continuous density on L_*. Then, for a fixed parameter $0 < \gamma < \pi/2$, $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$ w.h.p. for sufficiently large SNR and N.

Proof First, the result of Proposition 11 bounds the permeance of inliers. On the other hand, the outliers follow a bounded distribution. This implies that

$$\max_{L \in G(D,d)} \mathcal{A}(\mathcal{X}, L) \leq MN_{\text{out}}. \tag{40}$$

Thus, comparing (38) and (40), for both $N_{\text{in}}/N_{\text{out}}$ and N sufficiently large, $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$ w.h.p. ∎
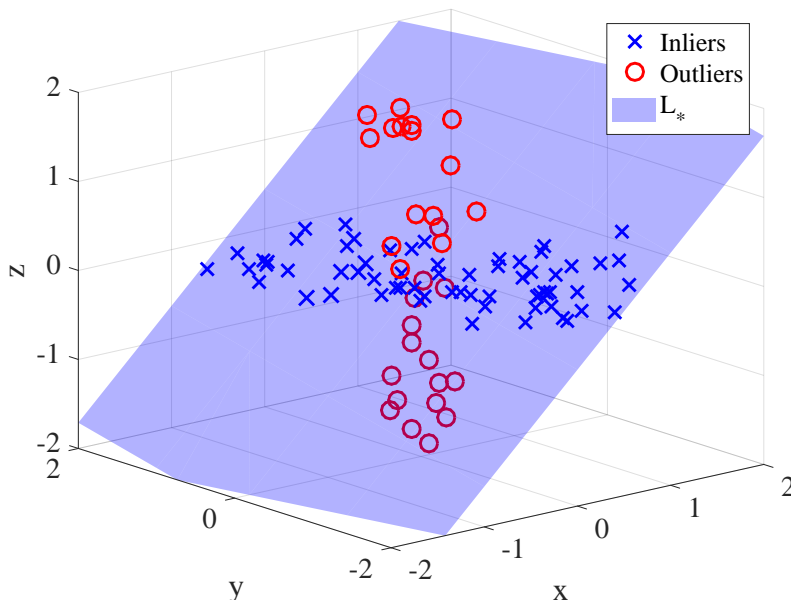
Figure 4: Example dataset drawn from a Generalized Haystack Model, where $d = 2$, $D = 3$, $N_{\text{in}} = 100$, and $N_{\text{out}} = 40$. Here, L_* is a random 2-dimensional subspace. Inliers are sampled i.i.d. from a normal distribution supported on L_* that has variance 4 and 0.09 in its principal directions. The outliers are sampled i.i.d. from a normal distribution with covariance $\mathbf{\Sigma}_{\text{out}} = \text{diag}(.04, .04, 2.25)$.

5.3.2. THE GENERALIZED HAYSTACK MODEL: SUB-GAUSSIAN INLIERS AND OUTLIERS

Next, we propose the Generalized Haystack Model as a special case of sub-Gaussian inliers and outliers. Fix a positive diagonal matrix $\mathbf{\Lambda}_{\text{in}} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{V}^* \in O(D, d)$, which spans $L_* \in G(D, d)$. Letting $\mathbf{\Sigma}_{\text{in}} = \boldsymbol{V}^* \mathbf{\Lambda}_{\text{in}} \boldsymbol{V}^{*T}$, we assume that N_{in} inliers are i.i.d. sampled from a sub-Gaussian distribution with covariance $\mathbf{\Sigma}_{\text{in}}/d$. Fix a symmetric positive semi-definite matrix $\mathbf{\Sigma}_{\text{out}} \in \mathbb{R}^{D \times D}$ and assume N_{out} outliers are i.i.d. sampled from a sub-Gaussian distribution with covariance $\mathbf{\Sigma}_{\text{out}}/D_{\text{out}}$, where D_{out} is the rank of $\mathbf{\Sigma}_{\text{out}}$. This specifies a Generalized Haystack Model with parameters N_{in}, $\mathbf{\Sigma}_{\text{in}}$, N_{out}, $\mathbf{\Sigma}_{\text{out}}$, D_{out}, and d. An example dataset drawn from a Generalized Haystack Model is given in Figure 4. This model generalizes the Haystack Model, which was proposed by Lerman et al. (2015) as a simple model with spherically symmetric Gaussian distributions of inliers and outliers. In the latter model, inliers are distributed i.i.d. $\mathcal{N}(\mathbf{0}, \sigma_{\text{in}}^2 \boldsymbol{P}_{L_*}/d)$, and outliers are distributed i.i.d. $\mathcal{N}(\mathbf{0}, \sigma_{\text{out}}^2 \boldsymbol{I}/D)$. This defines the Haystack Model with parameters N_{in}, σ_{in}, N_{out}, σ_{out}, D and d.

We can combine the previous results on sub-Gaussian inliers and outliers to yield a theoretical guarantee. We show that, under certain conditions on the Generalized Haystack Model, $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$ with overwhelming probability.

Theorem 14 (Stability of the Generalized Haystack Model) *Suppose that the dataset* $\mathcal{X}$ *follows the Generalized Haystack Model with parameters* N_{in}, $\mathbf{\Sigma}_{\text{in}}$, N_{out}, $\mathbf{\Sigma}_{\text{out}}$, D_{out}, *and*

d. Suppose also that $0 < \gamma < \pi/2$ and

$$\text{SNR} > \frac{1}{\cos(\gamma)} \frac{\lambda_1(\mathbf{\Sigma}_{\text{in}})^{1/2}}{\lambda_d(\mathbf{\Sigma}_{\text{in}})} \lambda_1(\mathbf{\Sigma}_{\text{out}}^{1/2}) \frac{2}{(1-a)^2} \frac{d}{\sqrt{D_{\text{out}}}} + o(1). \tag{41}$$

Then $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$ with probability at least $1 - 4e^{-c_1 a^2 N_{\text{in}}} - 2e^{-c_2 N_{\text{out}}}$, provided that $(1-a)^2 N_{\text{in}} > C_1^2 d$, where c_1, c_2 and C_1 are constants depending on the sub-Gaussian norms of the inliers and outliers.

Proof It is left to compare the bounds derived earlier in Lemma 10 and Proposition 12. Notice that (41) can be obtained by requiring the right hand side of (37) to be less than $\cos(\gamma)$ times the right hand side of (39). This results in precisely the statement in the theorem. ∎

Scaling the inlier and outlier covariance matrices by d and D_{out}, respectively, ensures that in the spherically symmetric case (where $\mathbf{\Sigma}_{\text{in}}$ and $\mathbf{\Sigma}_{\text{out}}$ are orthogonal projections onto subspaces of $\mathbb{R}^D$ of dimensions d and D_{out} respectively) the inliers and outliers have the same typical length. Thus, with this normalization, differences in the traces of $\mathbf{\Sigma}_{\text{in}}$ and $\mathbf{\Sigma}_{\text{out}}$ translate into differences in typical scale between inliers and outliers. We emphasize that it is important to prove results for general sub-Gaussian distributions rather than just spherically symmetric Gaussians. This is due to the fact that simpler strategies, like running PCA and then filtering points far from the PCA subspace, can be applied to the symmetric case with great success. The Generalized Haystack Model allows for certain adversarial outliers: for example, outliers can be contained in a low-dimensional subspace as well. Nevertheless, since the Haystack Model has been addressed by several previous works and since it is easy to improve our estimates for it, we address it in Section 5.5.

5.4. PCA Initialization

The discussed models can also guarantee good initialization by PCA. This is an essential ingredient to actually have a practical algorithm. We demonstrate this on the specific case of the Generalized Haystack Model. However, this sort of argument can be extended to the other types of models discussed above as well (such as bounded distributions of outliers and bounded continuous inliers).

We must have a lower bound on the SNR that depends on the parameters of the sub-Gaussian distributions in order for the following proposition to hold. A short proof for this proposition is given in Appendix D.5. It essentially states that Lemma 9 holds with high probability under certain conditions on the Generalized Haystack Model.

Proposition 15 (PCA Initialization with Sub-Gaussian Models) *Suppose that the dataset $\mathcal{X}$ follows the Generalized Haystack Model with parameters N_{in}, $\mathbf{\Sigma}_{\text{in}}$, N_{out}, $\mathbf{\Sigma}_{\text{out}}$, D_{out}, and d. Suppose also that, for some $0 < \gamma < \pi/2$,*

$$\text{SNR} \geq \frac{\sqrt{2}}{\sin(\gamma)} \frac{d}{D_{\text{out}}} \frac{\lambda_1(\mathbf{\Sigma}_{\text{out}})}{\lambda_d(\mathbf{\Sigma}_{\text{in}})} + o(1). \tag{42}$$

Then, for large enough $N = N_{\text{out}} + N_{\text{in}}$, the PCA d-subspace is contained in $B(L_*, \gamma)$ w.h.p.

5.5. Performance of GGD Under the Haystack Model

We assume here the simpler Haystack Model and show that GGD performs almost as well as state-of-the-art methods on datasets drawn from this model. We compute results for three different regimes of sample size. These are the small sample regime $N = O(D)$, the larger sample regime $N = O(D^p)$, for $p > 1$ sufficiently small, and the very large N regime, where N must depend on the SNR as well. In the larger sample regime, GGD requires at least $N = O(d(D-d)^2 \log(D))$, which is not more than $N = O(D^{3+\epsilon})$. In the very large N regime, in addition to dependence of N on a power of D, it also depends on a negative power of the SNR and thus is very large for small SNR. In our case, the very large N regime considers sample sizes of the order $N_{\text{out}} = O(\max(d^3 D^3 \log^3(N_{\text{out}}), (dN_{\text{out}}/N_{\text{in}})^6))$. The big O notation is slightly abused here, as we are really indicating results for finite N and D: the order is meant to illustrate the relation between these finite values. We compare all of these results together in Table 1.

5.5.1. Bounds for Sample Size $N = O(D)$

We first translate the bounds obtained previously in Theorem 14 to this special model. We choose $a = 1/2$ and thus obtain the following corollary:

Corollary 16 *Suppose that the dataset $\mathcal{X}$ follows the Haystack Model with parameters N_{in}, σ_{in}, N_{out}, σ_{out}, D, and d. Then, if $N_{\text{in}} > 4C_1^2 d$ and*

$$\text{SNR} \geq 8 \frac{1}{\cos(\gamma)} \frac{\sigma_{\text{out}}}{\sigma_{\text{in}}} \frac{d}{\sqrt{D}} + o(1), \tag{43}$$

$\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$ with probability at least $1 - 4e^{-c_1 N_{\text{in}}/4} - 2e^{-c_2 N_{\text{out}}}$, where c_1, c_2 and C_1 are absolute constants.

Notice that in this case, we obtain strong probabilistic estimates for even small sample sizes of $N = O(D)$. For the full theoretical guarantee, we also need to consider (42), and we must choose a value for γ. To balance between the $\sin(\gamma)$ and $\cos(\gamma)$ in (42) and (43), respectively, we fix $\gamma = \pi/4$. From these equations, for this fixed γ, we conclude that our theoretical SNR for the Haystack Model in the small sample regime is

$$\text{SNR} \geq \max \left(8\sqrt{2} \frac{\sigma_{\text{out}}}{\sigma_{\text{in}}} \frac{d}{\sqrt{D}}, 2 \frac{\sigma_{\text{out}}^2}{\sigma_{\text{in}}^2} \frac{d}{D} \right). \tag{44}$$

On the other hand, previous works (Hardt and Moitra, 2013; Lerman et al., 2015; Zhang, 2016) obtained optimal bounds for this model when $N = O(D)$ and the SNR is on the order of

$$\text{SNR} \gtrsim \frac{\sigma_{\text{out}}}{\sigma_{\text{in}}} \frac{d}{(D-d)}. \tag{45}$$

We remark that the bound of Lerman et al. (2015) for the REAPER algorithm requires the assumption $d < (D-1)/2$ and its constant is relatively large. This is in contrast to Hardt and Moitra (2013) and Zhang (2016), who do not have restrictions on d and do not have dependence on $\sigma_{\text{out}}/\sigma_{\text{in}}$. In this regime, we are unable to establish sharp results like the ones of REAPER, Tyler's M-estimator, or RandomizedFind. These estimates are

better by a factor of $\sqrt{D}/(D-d)$ than our current estimate. Nevertheless, in this regime the complexity of our algorithm is $O(NDd)$, whereas the complexity of the mentioned algorithms is $O(ND^2)$ or $O(D^3)$.

5.5.2. Bounds for Sample Size $N = O(d(D-d)^2 \log(D))$

Zhang and Lerman (2014) obtained the following sharper bound for the GMS algorithm under the Haystack Model and the larger sample regime of $N = O(D^2)$:

$$\text{SNR} \gtrsim \frac{\sigma_{\text{out}}}{\sigma_{\text{in}}} \frac{d}{\sqrt{D(D-d)}}. \tag{46}$$

We remark that this is the sharpest bound for any similar sample regime under this model when $\sigma_{\text{out}} \approx \sigma_{\text{in}}$. While the bounds mentioned above (Hardt and Moitra, 2013; Lerman et al., 2015; Zhang, 2016) hold for any regime of sample size, they are worse by a factor of $\sqrt{D/(D-d)}$. We show that a similar bound in a similar regime holds for the GGD algorithm. Indeed, the primary deficiency in Corollary 16 is that we use the loose bound on the alignment, $\sqrt{N_{\text{out}}}\|\boldsymbol{X}_{\text{out}}\|_2$. However, one could instead operate using the precursor to this bound,

$$\mathcal{A}(\mathcal{X}_{\text{out}}, L) \leq \|\widetilde{\boldsymbol{Q}_L \boldsymbol{X}_{\text{out}}}\|_2 \|\boldsymbol{X}_{\text{out}}\|_2. \tag{47}$$

Using this bound instead, we have the following theorem, which shows that GGD achieves the optimal SNR bound under the Haystack Model in the region $N = O(d\,(D-d)^2\,\log(D))$, which is at worst $O(D^3 \log D)$.

Theorem 17 (Stability of the Haystack Model) *Suppose that the dataset $\mathcal{X}$ follows the Haystack Model with parameters N_{in}, σ_{in}, N_{out}, σ_{out}, D, and d. If*

$$\text{SNR} \geq \frac{\sigma_{\text{in}}}{\sigma_{\text{out}}} \frac{1}{\cos(\gamma)} \frac{5}{(1-a)^2} \frac{d}{\sqrt{D(D-d)}} + o(1), \tag{48}$$

then $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$ with probability at least

$$1 - 2e^{-N_{\text{out}}/16} - e^{-N_{\text{out}}/4} - C_1 \exp\left(-\frac{N_{\text{out}}}{4(D-d)} + \frac{d(D-d)}{2} \log\left(\frac{D}{D-d}\right)\right) - 4e^{-c_1 a^2 N_{\text{in}}}. \tag{49}$$

Here, we can take a close to 0 (e.g., we can take $a = o(N_{\text{in}}^{-1/2})$), and so we ignore the factor of $1/(1-a)^2$ in (48). The full theoretical guarantee, which also considers (42), is established similarly to (44). We conclude that our theoretical SNR bound for the Haystack Model in this regime is

$$\text{SNR} \geq \max\left(5\sqrt{2}\frac{\sigma_{\text{out}}}{\sigma_{\text{in}}} \frac{d}{\sqrt{D(D-d)}}, 2\frac{\sigma_{\text{out}}^2}{\sigma_{\text{in}}^2} \frac{d}{D}\right). \tag{50}$$

The final results for the two different regimes, that is, (44) and (50), are for $\gamma = \pi/4$ and initialization by PCA. As a side note, if the SNR grows, we see that larger values of γ may be tolerated for GGD. In particular, for large sample sizes and sufficiently large SNRs, γ can be sufficiently close to $\pi/2$. In this case, random initialization of GGD are expected

to work as well as the PCA initialization. We quantify this claim more rigorously in the special case where $d < D/2$ and $d, D \to \infty$. Based on the analysis of extreme singular values of random Gaussian matrices (Rudelson and Vershynin, 2008), it can be shown that with high probability, a random initialization lies in $B(L_*, \gamma) \setminus \{L_*\}$, where $\cos(\gamma) = O(1/\sqrt{Dd})$. Therefore, GGD with random initialization succeeds with high probability under the given assumptions on d and D when $N_{\text{in}}/N_{\text{out}} \geq O(d\sigma_{\text{out}}/\sigma_{\text{in}})$.

5.5.3. Bounds for Very Large N

In the large sample regime, one can prove something much stronger. Indeed, for any fraction of outliers, it is obvious that PCA asymptotically recovers the underlying subspace L_* (Lerman and Maunu, 2018a). Further, it was shown that FMS can asymptotically recover L_* for any fraction of outliers with better dependence on the sample size than PCA (Lerman and Maunu, 2018a) (while the result given for FMS is for the spherized Haystack Model, the result can be extended to the non-spherized version as well). It would be very surprising if GGD could not do something similar.

In fact, this type of result can be extended to GGD as well, which means that, as $N \to \infty$, we can take SNR$\to 0$. However, while the PCA and FMS subspace estimators converge to the underlying subspace as $N \to \infty$, the probability of either exactly recovering the underlying subspace for any fixed N is zero. In contrast, GGD can exactly recover the underlying subspace with overwhelming probability for finite, yet large, N. This result is stated in the following theorem, which is proved in Appendix D.7.

Theorem 18 *Suppose that $\mathcal{X}$ follows the Haystack Model with parameters N_{in}, σ_{in}, N_{out}, σ_{out}, D, and d. For any SNR lower bound $\alpha > 0$ (i.e., SNR $> \alpha$) and for any N_{out} at least $O(\max(d^3 D^3 \log^3(N_{\text{out}}), (dN_{\text{out}}/N_{\text{in}})^6))$, GGD recovers L_* w.o.p. In particular, N_{out} is at least $O(d/\alpha^6)$,*

For this theorem to hold, we see that N needs to be quite large, especially for low SNRs. However, note that we still obtain strong probabilistic bounds for large enough finite N. On the other hand, taking $\alpha \to 0$ in this theorem requires $N_{\text{out}} \to \infty$, which implies asymptotic recovery for GGD for any SNR in this model.

5.5.4. Comparison of all Haystack Model Results

We compare the lowest SNR guarantees for a variety of RSR algorithms under the Haystack Model. Table 1 replicates Table 1 of Lerman and Maunu (2018b) with updated estimates. It both compares lower bounds on SNR under the Haystack Model and also briefly describes the actual data model that each algorithm has guarantees for. The algorithms include geodesic gradient descent (GGD), FMS (Lerman and Maunu, 2018a), REAPER (Lerman et al., 2015), GMS (Zhang and Lerman, 2014), OP (Xu et al., 2012) and LLD (McCoy and Tropp, 2011), HR-PCA (Xu et al., 2013), Tyler's M-estimator (TME) (Zhang, 2016), TORP (Cherapanamjeri et al., 2017), CP (Rahmani and Atia, 2016), SSC (Soltanolkotabi and Candès, 2012), HOSC (Arias-Castro et al., 2011), RANSAC (Arias-Castro and Wang, 2017), and RF (Hardt and Moitra, 2013). For each SNR bound, we also give the associated sample size, N, for each result to begin holding with high probability. Again, we note that the big O notation is slightly abused here since the results are really for non-asymptotic N and D.

PCA	$N_{\text{in}}/N_{\text{out}} \, \text{``} \gtrsim \text{''} 0$, when $N \to \infty$, D fixed
	No exact recovery and poor estimates for finite N.
GGD	$N_{\text{in}}/N_{\text{out}} \geq \max\left(8\sqrt{2}\frac{\sigma_{\text{out}}}{\sigma_{\text{in}}}\frac{d}{\sqrt{D}}, 2\frac{\sigma_{\text{out}}^2}{\sigma_{\text{in}}^2}\frac{d}{D}\right)$, when $N = O(D)$
	$N_{\text{in}}/N_{\text{out}} \geq \max\left(5\sqrt{2}\frac{\sigma_{\text{out}}}{\sigma_{\text{in}}}\frac{d}{\sqrt{D(D-d)}}, 2\frac{\sigma_{\text{out}}^2}{\sigma_{\text{in}}^2}\frac{d}{D}\right)$, when $N = O(d(D-d)^2 \log(D))$
	Any $N_{\text{in}}/N_{\text{out}} > 0$, when $N_{\text{out}} \gtrsim \max(d^3 D^3 \log^3(N), (dN_{\text{out}}/N_{\text{in}})^6)$
	$\implies N_{\text{in}}/N_{\text{out}} \gtrsim 0$, when $N \to \infty$, D fixed
	Deterministic condition, results for a variety of data models.
FMS	$N_{\text{in}}/N_{\text{out}} \, \text{``} \gtrsim \text{''} 0$, when $N \to \infty$, D fixed
	Approximate recovery for large samples from spherized Haystack or from two one-dimensional subspaces on the sphere. Much better estimates for finite N than PCA.
REAPER	$N_{\text{in}}/N_{\text{out}} \geq 16\frac{\sigma_{\text{out}}}{\sigma_{\text{in}}}\frac{d}{D}$, when $N = O(D)$, $1 \leq d \leq (D-1)/2$
	Deterministic condition, results for Haystack where $d < (D-1)/2$.
GMS	$N_{\text{in}}/N_{\text{out}} \geq 4\frac{\sigma_{\text{out}}}{\sigma_{\text{in}}}\frac{d}{\sqrt{(D-d)D}}$, when $N = O(D^2)$
	Deterministic condition, results for Haystack that extends to elliptical outliers.
OP	$N_{\text{in}}/N_{\text{out}} \geq \frac{121d}{9}O\left(\max(1, \log(N)/d)\right)$, when $N = O(D)$
	Deterministic condition (formulated for arbitrary outliers) with last term in above formula replaced by an inlier incoherence parameter μ.
HR-PCA	$N_{\text{in}}/N_{\text{out}} \to \infty$, when $N \to \infty$, D fixed
	Weak lower bound on the expressed variance, requires fraction of outliers as input.
TME/(D)RF	$N_{\text{in}}/N_{\text{out}} > \frac{d}{D-d}$, when $N = O(D)$
	Result for "general-position" data, but does not extend to noise.
TORP	$N_{\text{in}}/N_{\text{out}} \geq 128d \max(1, \log(N)/d)^2$, when $N = O(D)$
	Deterministic condition (formulated for arbitrary outliers) with last term replaced by an inlier incoherence parameter μ, requires fraction of outliers as input.
CP	$N_{\text{in}}/N_{\text{out}} \geq d/(D-d^2)$ $(N = O(D), d < \sqrt{D})$
	$N_{\text{in}}/N_{\text{out}} \gtrsim 0$, when $N \to \infty$, $d < \sqrt{D}$, D fixed
	Exact recovery for the spherized Haystack model with a random inlier subspace and $d < \sqrt{D}$, and also for a special model of outliers around a line.
SSC	$N_{\text{in}}/N_{\text{out}} \geq d/D \cdot ((\frac{N_{\text{in}}-1}{d})^{\frac{cD}{d}-1} - 1)^{-1}$, when $N < e^{c\sqrt{D}}/D$, w.h.p. in D
	Exact recovery for the spherized Haystack model with a random inlier subspace.
HOSC	$N_{\text{in}}/N_{\text{out}} \geq \log(N)N^{-2(D-d)/(2D-d)}$
	$\implies N_{\text{in}}/N_{\text{out}} \gtrsim 0$, where $N \to \infty$, D fixed
	Result for outliers uniformly sampled from $[0,1]^D$ and inliers uniformly sampled from the intersection of a d-subspace with $(0,1)^d$. Also extends to a union of manifolds and to settings with small noise.

Table 1: Updated table from Lerman and Maunu (2018b), which compares the lower bounds on SNR for many RSR methods under the Haystack Model. Included here are theoretical guarantees along with the corresponding sample size requirements for the result to hold with probability close to 1.

Here, the symbol $\gtrsim 0$ is used for the SNR lower bound for recovery when a method can exactly recover an underlying subspace for any fixed SNR and large enough N. Similarly, the symbol " $\gtrsim$ "0 is used when a method can *approximate* L_* to any accuracy for any fixed

SNR and large enough N. However, methods with " $\gtrapprox$ "0 instead of $\gtrapprox 0$ cannot exactly recover the underlying subspace.

Among all algorithms, only PCA, GGD, TORP, RANSAC, and FMS run in $O(NDd)$ time. Since the strong gradient condition can be shown to hold for the Haystack model (see Corollary 21 in Section 5.6), GGD also achieves linear convergence. This means that it is theoretically guaranteed to be the fastest out of these $O(NDd)$ algorithms (TORP has a guarantee for linear convergence, but under very restrictive assumptions on the SNR). Furthermore, among all algorithms, GGD is the only one with guarantees close to state-of-the-art for the small sample size regime. Among all algorithms it has the state-of-the-art result for the regime $N = O(d\,(D - d)^2 \log(D))$, although GMS obtains such a result for the smaller regime $N = O(D^2)$. GGD is also the fastest algorithm with a result for the large sample regime as well. We note that algorithms with worse complexity, such as CP, SSC, and HOSC, have guarantees in this setting as well. We have found it too complicated to compare the exact theoretical results of these various methods, and so we have instead opted to just show that they are guaranteed to have exact recovery for any percentage of outliers for large enough N.

5.6. A Note on the Strong Gradient Condition in Theorem 4

While it may be hard to interpret the strong gradient condition in Theorem 4, it is possible to show that it holds for a variety of the data models in this section. The brief discussion here is just meant to illustrate that this condition is, in fact, practical. It is not hard to extend the arguments below to other cases discussed earlier as well. This lemma relies on the following definition.

Definition 19 *Sets of inliers and outliers, $\mathcal{X}_{\mathrm{in}}$ and $\mathcal{X}_{\mathrm{out}}$, within an inlier-outlier dataset are said to lie in general position with respect to each other if*

$$\max_{G(D,d)\backslash\{L_*\}} \#(\mathcal{X} \cap L) \le d. \tag{51}$$

With this definition, we prove the following lemma.

Lemma 20 *Suppose that the inliers and outliers follow distributions that satisfy (34) and lie in general position with respect to each other. Assume further that the inliers are drawn from a distribution such that $\mathcal{P}(\mathcal{X}_{\mathrm{in}}) \to \infty$ as $N \to \infty$, and that*

$$\max_{L \in G(D,d)} \mathcal{A}(\mathcal{X}_{\mathrm{out}}, L) < c\cos(\gamma)\mathcal{P}(\mathcal{X}_{\mathrm{in}}), \tag{52}$$

for some $0 < c < 1$ for all N. Then, (23) holds for N sufficiently large.

The general position assumption implies that the amount of data points in any d-subspace that is not L_* is at most d. This general position-type assumption is also satisfied with probability 1 by any sample drawn from the Haystack Model, or by inliers drawn uniformly from $L_* \cap B(\mathbf{0}, M)$ and outliers drawn uniformly from $B(\mathbf{0}, M)$.

Proof Under the assumptions of the lemma, (23) reduces to

$$\inf_{L \in B(L_*,\gamma)\backslash\{L_*\}} \left| \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X})|_{t=0} \right| > 2dM, \tag{53}$$

where $L(t)$ is a geodesic parameterized by arclength from L_k through L_*. Since we know that $\left|\frac{\mathrm{d}}{\mathrm{d}t}F(L(t);\mathcal{X})|_{t=0}\right| \geq \mathcal{S}(\mathcal{X},L_*,\gamma)$ (from the proof of Theorem 1), a sufficient condition for the strong gradient condition to hold is

$$\mathcal{S}(\mathcal{X},L_*,\gamma) > 8dM. \tag{54}$$

The assumptions also imply that

$$\mathcal{S}(\mathcal{X},L_*,\gamma) > \cos(\gamma)(1-c)\mathcal{P}(\mathcal{X}_{\mathrm{in}}). \tag{55}$$

Therefore, the reduction of the strong gradient condition in (54) is satisfied for N sufficiently large, since $\mathcal{P}(\mathcal{X}_{\mathrm{in}}) \to \infty$ as $N \to \infty$. ∎

One can easily modify the proof of Theorem 18 and the proof of Lemma 20 to obtain linear convergence for the Haystack model. We state this result without proof in the following corollary.

Corollary 21 *Suppose that the inliers and outliers are distributed according to the Haystack Model with parameters N_{in}, σ_{in}, N_{out}, σ_{out}, D, and d. Then, (23) holds w.h.p. for N sufficiently large.*

6. Simulations

In this section, we run two simulations to verify the theory we proved earlier. All experiments are run using the Haystack Model. Some more comprehensive comparisons of the various RSR algorithms on synthetic and stylized datasets are contained in the review paper of Lerman and Maunu (2018b).

First, we attempt to demonstrate the stability condition $\mathcal{S}$ from (12). While we cannot explicitly evaluate the maximum within this expression, we can instead simulate the values achieved by

$$\cos(\gamma)\mathcal{P}(\mathcal{X}_{\mathrm{in}}) - \mathcal{A}(\mathcal{X}_{\mathrm{out}},L), \tag{56}$$

for L in a small neighborhood of L_*.

The values achieved by (56) are simulated in Figure 5 for the noiseless RSR settings with the fixed value of $\gamma = \pi/4$. The dataset for this figure is generated according to the Haystack Model outlined in Section 5.5, with parameters $N_{\mathrm{in}} = 200$, $\sigma_{\mathrm{in}} = 1$, $N_{\mathrm{out}} = 200$, $\sigma_{\mathrm{out}} = 1$, $D = 200$, and $d = 10$. For this plot, the y-axis represents a distance from the underlying subspace L_* in terms of the maximum principal angle up to γ. The x-axis represents randomly generated subspaces at that distance from L_*. The color represents the value of (56). As we can see, the value of (56) is indeed positive within this large neighborhood of L_*.

We also simulate the convergence properties of the GGD method in Figure 6. The data was generated according to the Haystack Model with parameters $N_{\mathrm{in}} = 200$, $\sigma_{\mathrm{in}} = 1$, $N_{\mathrm{out}} = 200$, $\sigma_{\mathrm{out}} = 1$, $D = 100$, and $d = 5$. We compare different choices of step-size in accordance with the statements of Theorems 3 and 4. The green line denotes the convergence of the step-size $t^k = s/\sqrt{k}$ in GGD with $s = 1/D$, which is the modified GGD from Theorem 3. We also display three different types of piecewise constant step-size schemes
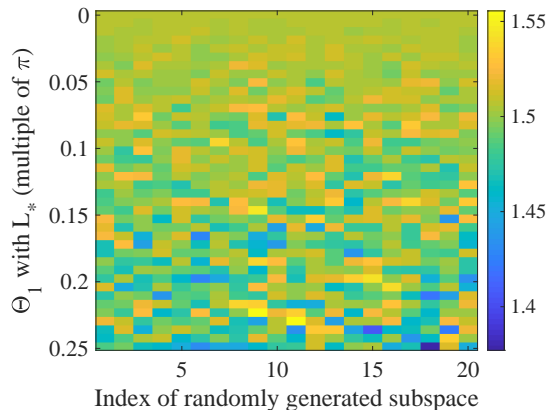
Figure 5: Simulation of the stability statistic. Data was generated for the Haystack Model with parameters $N_{\text{out}} = 200$, $\sigma_{\text{out}} = 1$, $D = 200$, and $d = 10$. The y-axis of each figure represents the maximum principal angle of a randomly generated subspace with L_*. For each value of the maximum principal angle, the x-axis represents the index in the set of 20 randomly generated subspaces with the specified maximum principal angle.

with initial step-size $1/D$. The blue and red lines represent Algorithm 1 with different choices of K and shrink-factors. In one case, we use $K = 20$ and a shrink-factor of $1/2$. In the other case, we use $K = 50$ and a shrink-factor of $1/10$. The pink line is a modification of this piecewise constant step-size scheme that decreases the cost monotonically. In this scheme, the algorithm only shrinks the step-size when the current constant step-size does not decrease the energy function. When this is the case, the step-size is shrunk by the largest factor $1/2^n$, for $n \in \mathbb{N}$, such that the energy decreases. While this set-up performs the best, we do not yet have a theoretical justification for this scheme. We see that all of the choices of piecewise constant step-size converge linearly to L_* (although the convergence is not monotonic). The small fluctuations around an error of 10^{-7} occur due to the machine precision limit in MATLAB.

7. Conclusion

We have presented a deterministic condition that ensures the landscape of (2) behaves nicely around an underlying subspace: if $\mathcal{S}(\gamma, 0, L_*) > 0$, the underlying subspace is the only minimizer and stationary point in $B(L_*, \gamma)$. The deterministic condition also ensures the convergence of a non-convex gradient method for RSR. The convergence of this method is linear under some slightly stronger assumptions on the data. We have shown that the condition $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$ and the strong gradient condition in (23) hold for certain statistical models of data and have used these examples to understand the limits of recovery in various regimes. These models indicate the flexibility of the conditions to deal with a wide range of RSR datasets. GGD is even shown to have almost state-of-the-art SNR guarantees under the Haystack Model.
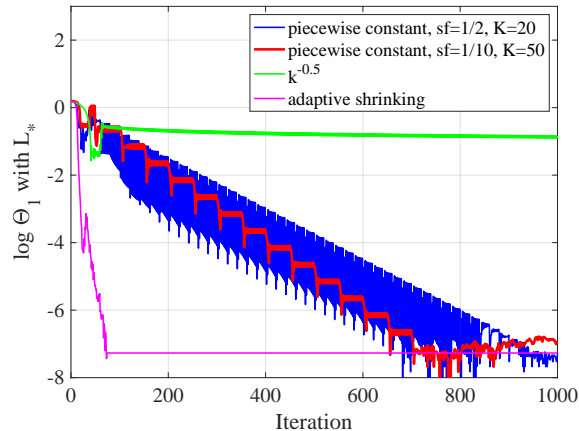
Figure 6: Convergence characteristics of the GGD algorithm with different step-size choices. Data was generated according to the Haystack Model with parameters $N_{\text{in}} = 200$, $\sigma_{\text{in}} = 1$, $N_{\text{out}} = 200$, $\sigma_{\text{out}} = 1$, $D = 100$, and $d = 5$.. The y-axis is the logarithm of the top principal angle with the underlying subspace L_*. For all algorithms, the initial step-size is $s = 1/D$. For the piecewise constant step-size scheme of Algorithm 1, we use two difference combinations of shrink-factor and time between shrinking, K. We also include an adaptive shrinking method, that only shrinks the constant step-size when the current step-size does not decrease the energy.

A main point of this work is to obtain exact recovery guarantees that are similar in spirit to the REAPER algorithm of Lerman et al. (2015) for a non-convex algorithm. We even manage to provide better estimates in large sample regimes. Indeed, the stability analysis of this work is inspired by the stability analysis of the REAPER algorithm, which is done with respect to a convex relaxation of the energy function considered here. Since we do not relax the non-convex problem, our stability is tighter than the REAPER stability when considering sufficiently small neighborhoods. However, for large enough neighborhoods, the REAPER stability might be tighter than ours. For example, a difference shows up in small sample regimes of the Haystack Model: the dependence on a neighborhood makes our result weaker in terms of the SNR regimes we can tolerate. Nevertheless, as far as we know, there is no non-convex RSR competitor for the types of estimates we have developed in this paper and no other competitor with computational complexity of order $O(NDd)$. Furthermore, in larger sample regimes we obtain a stronger result than REAPER's.

While there are many directions that future work can take, we only specify a couple here. One avenue for future work is to extend this result to other data models. For example, one may consider more adversarial models of corruption. This is pursued in a forthcoming work (Maunu and Lerman).

Another direction for future work involves study of subspace recovery in the presence of heavier noise. The only works that consider RSR in really noisy settings are Coudron and Lerman (2012), Minsker (2015), and Cherapanamjeri et al. (2017), although the area remains largely unexplored. The noise considered in our work involves small perturbations

from the underlying subspace. However, in general, more modern work in data science has focused on settings with heavier noise. For example, some have considered noise drawn from distributions with heavy tails (Minsker, 2015), while others have considered PCA in the spiked model (Johnstone, 2001; Baik et al., 2005). In the latter case, when the dimension is very large, the noisy inliers would most likely be very far from the underlying subspace. It is an interesting direction for future work to study RSR in both of these settings.

Future work may also consider proof of convergence for other methods, such as Newton's method and conjugate gradient (Edelman et al., 1999), or IRLS (Lerman and Maunu, 2018a), using the guarantees on the energy landscape of this paper. One could also consider using different frameworks for optimization over $G(D, d)$, such as a retraction based method like that specified in Absil et al. (2009). A quick heuristic argument indicates that the retraction formulation should agree with GGD up to first order, but we leave rigorous examination of this to future work.

Finally, one may also directly follow the ideas of Lim et al. (2016, 2018) and consider extensions to the affine Grassmannian, since in practice we cannot assume that the data is properly centered.

A supplementary web page with code will be made available at `https://twmaunu.github.io/WTL/`.

Acknowledgments

Appendix A. Landscape of the PCA energy

In this appendix, we discuss the landscape of the energy function described in (1). Let $\boldsymbol{X} \in \mathbb{R}^{D \times N}$ be a matrix with columns given by $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, and assume that its singular values and left singular vectors are $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_D$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_D \in \mathbb{R}^D$, respectively. Under a generic scenario when $\sigma_d > \sigma_{d+1}$ and $\sigma_{D-d} > \sigma_{D-d+1}$, the landscape of the energy in (1) has the following properties:

- There exists a unique local minimum given by $L = \text{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d)$, which is also the global minimizer; and a unique local maximum given by $L = \text{span}(\boldsymbol{v}_{D-d+1}, \ldots, \boldsymbol{v}_D)$, which is also the global maximizer.

- The set of saddle points are given by the set of L such that $L = \text{span}(\boldsymbol{v}_{i_1}, \ldots, \boldsymbol{v}_{i_d})$ for any d distinct integers $(i_1, \ldots, i_d)$ between 1 and D, and $(i_1, \ldots, i_d)$ can not be $(1, \ldots, d)$ or $(D-d+1, \ldots, D)$ (which correspond to the minimizer and the maximizer).

These properties can be derived by casting the PCA problem as a constrained optimization problem and then examining the corresponding Lagrangian. Since the number of saddle points is finite and all saddle points are orthogonal to the minimizer, there exists a local neighborhood around the minimizer such that the minimizer is the unique critical point

inside this neighborhood, which is a property similar to Theorem 1. Using a strategy similar to this work, if an initialization is chosen appropriately such that the initialization lies in a neighborhood around the global minimizer, a gradient descent algorithm would converge to the global minimizer of (1).

One other case of interest involves $\sigma_d = \sigma_{d+1} = \cdots = \sigma_k$, for $d < k \leq D$. In this case, the singular vectors corresponding to $\sigma_d, \ldots, \sigma_k$ would span a subspace, and any orthonormal basis for $\mathrm{Sp}(\boldsymbol{v}_d, \ldots, \boldsymbol{v}_k)$ would result in a valid set of singular vectors corresponding to the singular values $\sigma_d, \ldots, \sigma_k$. As an example of what happens in this case, if $\sigma_1 > \sigma_2 > \cdots > \sigma_d = \cdots = \sigma_k$ then the d-subspace spanned by $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{d-1}$, $\boldsymbol{v}_i \in \mathrm{Sp}(\boldsymbol{v}_d, \ldots, \boldsymbol{v}_k) \cap S^{D-1}$ are minimizers. One can construct other sorts of continuums by also considering equality of singular values between σ_1 and σ_d. All cases constructed in these ways would yield a continuum of minimizers for the PCA d-subspace energy, which all achieve the same energy. Analogously, one can consider the cases where $\sigma_k = \ldots = \sigma_{D-d+1}$ for $1 \leq k < D-d+1$, and show a continuum of maximizers. Or, one can also consider the cases where $\sigma_k = \ldots = \sigma_l$ where $d < k < l < D - d + 1$, and show a continuum of saddle points.

In addition, while it is unrelated to the focus of this work, we remark that the landscape of PCA also has the "strict saddle point property" discussed in Ge et al. (2015): the Hessian of every saddle point has a negative eigenvalue. Ge et al. (2015) propose algorithms with theoretical guarantees to minimize such energy functions. This property can be seen by again examining the Lagrangian formulation of the PCA problem.

Appendix B. Grassmannian Geodesics

In this appendix, we describe some basic geometric notions on the Grassmannian manifold, $G(D, d)$. Given two subspaces $L_1, L_2 \in G(D, d)$, the principal angles between the two subspaces are defined sequentially. The smallest angle, θ_d, is given by

$$\theta_d = \min_{\substack{\boldsymbol{v} \in L_1, \boldsymbol{y} \in L_2 \\ \|\boldsymbol{v}\| = \|\boldsymbol{y}\| = 1}} \arccos(|\boldsymbol{v}^T \boldsymbol{y}|). \tag{57}$$

The vectors $\boldsymbol{v}_d$ and $\boldsymbol{y}_d$ which achieve the minimum are the principal vectors corresponding to θ_d. The d principal angles are defined sequentially by

$$\theta_k = \min_{\substack{\boldsymbol{v} \in L_1, \|\boldsymbol{v}\| = 1, \boldsymbol{v} \perp \boldsymbol{v}_{k+1}, \ldots, \boldsymbol{v}_d \\ \boldsymbol{y} \in L_2, \|\boldsymbol{y}\| = 1, \boldsymbol{y} \perp \boldsymbol{y}_{d+1}, \ldots, \boldsymbol{y}_d}} \arccos(|\boldsymbol{v}^T \boldsymbol{y}|), \tag{58}$$

and the corresponding principal vectors are found in the same way. The ordering defined in (57) and (58) is the reverse of what is usually used for principal angles: here, θ_1 is the largest principal angle, while most other works denote the smallest principal angle with θ_1. Notice that if two principal angles are equal, the choice of principal vectors is not unique. Principal angles and vectors can be efficiently calculated: if $\boldsymbol{W}_1 \in O(D, d)$ spans L_1 and $\boldsymbol{W}_2 \in O(D, d)$ spans L_2, then we write the singular value decomposition

$$\boldsymbol{W}_1^T \boldsymbol{W}_2 = \boldsymbol{V}_{12} \boldsymbol{\Sigma}_{12} \boldsymbol{Y}_{12}^T. \tag{59}$$

The principal angles are given in reverse order by $\arccos(\mathrm{diag}(\boldsymbol{\Sigma}_{12}))$, and the corresponding principal vectors are given by the columns of $\boldsymbol{V}_{12}$ and $\boldsymbol{Y}_{12}$ in reverse order. Now let two

subspaces L_1 and L_2 have principal angles $\theta_1, \ldots, \theta_d$ and principal vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_d$, respectively. Let k be the largest index such that $\theta_k > 0$, which is also known as the interaction dimension (Lerman and Zhang, 2014), where $d - k$ is the dimension of $L_1 \cap L_2$. Then, we can define a complementary orthogonal basis $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ for L_2 with respect to L_1 as

$$\boldsymbol{u}_j \in \mathrm{Sp}(\boldsymbol{v}_j, \boldsymbol{y}_j), \ \boldsymbol{u}_j \perp \boldsymbol{v}_j, \ \boldsymbol{u}_j^T \boldsymbol{y}_j > 0. \tag{60}$$

As explained in Lerman and Zhang (2014), for any two subspaces L_1 and L_2 such that $\theta_1 = \arccos(|\boldsymbol{v}_1^T \boldsymbol{y}_1|) < \pi/2$, a unique geodesic on $G(D, d)$ with $L(0) = L_1$ and $L(1) = L_2$ can be parameterized by

$$L(t) = \mathrm{Sp}(\boldsymbol{v}_1 \cos(\theta_1 t) + \boldsymbol{u}_1 \sin(\theta_1 t), \ldots, \boldsymbol{v}_k \cos(\theta_k t) + \boldsymbol{u}_k \sin(\theta_k t), \boldsymbol{v}_{k+1}, \ldots, \boldsymbol{v}_d). \tag{61}$$

In the paper, we will frequently use a reparametrization of geodesics to prevent the angles from needlessly affecting the magnitude of the derivatives. In these cases, we reparametrize the geodesic in (62) by arclength (in terms of metric defined Section 2.2) by writing

$$L(t) = \mathrm{Sp}\left(\boldsymbol{v}_1 \cos(t) + \boldsymbol{u}_1 \sin(t), \boldsymbol{v}_2 \cos\left(\frac{\theta_2}{\theta_1} t\right) + \boldsymbol{u}_2 \sin\left(\frac{\theta_2}{\theta_1} t\right), \ldots, \right. \tag{62}$$
$$\left. \boldsymbol{v}_k \cos\left(\frac{\theta_k}{\theta_1} t\right) + \boldsymbol{u}_k \sin\left(\frac{\theta_k}{\theta_1} t\right), \boldsymbol{v}_{k+1}, \ldots, \boldsymbol{v}_d\right).$$

Appendix C. Bound on the Alignment

Here we derive the simple bound on the alignment statistic seen in (11). Here, the $\widetilde{}$ notation denotes projection of the data points (columns of a data matrix) to the unit sphere, S^{D-1}. This bound is used to prove Theorem 17. The first step follows from (6), (8), and (10).

$$\mathcal{A}(\mathcal{X}_{\mathrm{out}}, \boldsymbol{V}) = \left\| \sum_{\boldsymbol{x}_i \in \mathcal{X}_{\mathrm{out}}} \frac{\boldsymbol{Q}_{\boldsymbol{V}} \boldsymbol{x}_i}{\|\boldsymbol{Q}_{\boldsymbol{V}} \boldsymbol{x}_i\|} \boldsymbol{x}_i^T \boldsymbol{V} \right\|_2 = \left\| \widetilde{\boldsymbol{Q}_{\boldsymbol{V}} \boldsymbol{X}}_{\mathrm{out}} \boldsymbol{X}_{\mathrm{out}}^T \boldsymbol{V} \right\|_2 \tag{63}$$
$$\leq \left\| \widetilde{\boldsymbol{Q}_{\boldsymbol{V}} \boldsymbol{X}}_{\mathrm{out}} \right\|_2 \left\| \boldsymbol{X}_{\mathrm{out}}^T \boldsymbol{V} \right\|_2 \leq \left\| \widetilde{\boldsymbol{Q}_{\boldsymbol{V}} \boldsymbol{X}}_{\mathrm{out}} \right\|_F \|\boldsymbol{X}_{\mathrm{out}}\|_2$$
$$\leq \sqrt{N_{\mathrm{out}}} \|\boldsymbol{X}_{\mathrm{out}}\|_2.$$

Appendix D. Supplementary Proofs

In this appendix, we give the supplementary proofs for the various theorems, propositions, and lemmas given in the paper.

D.1. Proof of Theorem 2

The proof of this theorem proceeds much in the same way as the proof of Theorem 1.

We first define the following geodesic on $G(D, d)$: Fix a subspace $L \in B(L_*, \gamma) \setminus B(L_*, \eta)$, and let the principal angles between L and L_* be $\theta_1, \ldots, \theta_d$. Here, we define $\eta = 2 \arctan(\epsilon/\delta)$. Also, choose a set of corresponding principal vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$

and $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d$ for L and L_*, respectively, and let l be the maximum index such that $\theta_1 = \cdots = \theta_l$. We let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l$ be complementary orthogonal vectors for $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_l$ and $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_l$, which exist since $\theta_1(L, L_*) > 0$. For $t \in [0, 1]$, we form the geodesic

$$L(t) = \mathrm{Sp}(\boldsymbol{v}_1 \cos(t) + \boldsymbol{u}_1 \sin(t), \ldots, \boldsymbol{v}_l \cos(t) + \boldsymbol{u}_l \sin(t), \boldsymbol{v}_{l+1}, \ldots, \boldsymbol{v}_d).$$

This geodesic moves only the furthest directions of $L(0)$ towards L_*, and we have removed dependence on θ_1, since this unnecessarily impacts the magnitude of the geodesic derivative (5).

We will first prove the inequality $\partial F(L(t); \mathcal{X}) < -\mathcal{S}_n(\mathcal{X}, L_*, \epsilon, \delta, \gamma)$. Using the derivative formula in (5), a subderivative of $F(L(t); \mathcal{X})$ at $t = 0$ is given by

$$\frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X}) \Big|_{t=0} = - \sum_{\substack{\boldsymbol{x}_i \in \mathcal{X} \\ \|\boldsymbol{Q}_L \boldsymbol{x}_i\| > 0}} \sum_{j=1}^{l} \frac{\boldsymbol{v}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{u}_j}{\|\boldsymbol{Q}_L \boldsymbol{x}_i\|} \tag{64}$$

$$= - \sum_{j=1}^{l} \left(\sum_{\substack{\boldsymbol{x}_i \in \mathcal{F}_1(\mathcal{X}_{\mathrm{in}}, \boldsymbol{w}_j, \delta) \\ \|\boldsymbol{Q}_L \boldsymbol{x}_i\| > 0}} \frac{\boldsymbol{v}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{u}_j}{\|\boldsymbol{Q}_L \boldsymbol{x}_i\|} + \sum_{\substack{\boldsymbol{x}_i \in \mathcal{F}_0(\mathcal{X}_{\mathrm{in}}, \boldsymbol{w}_j, \delta) \\ \|\boldsymbol{Q}_L \boldsymbol{x}_i\| > 0}} \frac{\boldsymbol{v}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{u}_j}{\|\boldsymbol{Q}_L \boldsymbol{x}_i\|} + \sum_{\substack{\boldsymbol{x}_i \in \mathcal{X}_{\mathrm{out}} \\ \|\boldsymbol{Q}_L \boldsymbol{x}_i\| > 0}} \frac{\boldsymbol{v}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{u}_j}{\|\boldsymbol{Q}_L \boldsymbol{x}_i\|} \right).$$

We examine the terms in (64) one by one. Using (6) and (8), we can bound the outlier term

$$- \sum_{\substack{\boldsymbol{x}_i \in \mathcal{X}_{\mathrm{out}} \\ \|\boldsymbol{Q}_L \boldsymbol{x}_i\| > 0}} \frac{\boldsymbol{v}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{u}_j}{\|\boldsymbol{Q}_L \boldsymbol{x}_i\|} \leq \|\nabla F(L; \mathcal{X}_{\mathrm{out}})\|_2. \tag{65}$$

We know that $|\boldsymbol{x}_i \cdot \boldsymbol{u}_j| \leq \|\boldsymbol{Q}_L \boldsymbol{x}_i\|$ for all i since $\boldsymbol{u}_j \in \mathrm{Sp}(\boldsymbol{Q}_L)$. We also know that, since $\theta_j > \eta$, $|\boldsymbol{v}_j^T \boldsymbol{x}_i| \leq |\boldsymbol{w}_j^T \boldsymbol{x}_i| \leq \sqrt{\delta^2 + \epsilon^2}$ for all $\boldsymbol{x}_i \in \mathcal{F}_0(\mathcal{X}_{\mathrm{in}}, \boldsymbol{w}_j, \delta)$. These two observations imply that

$$- \sum_{\substack{\boldsymbol{x}_i \in \mathcal{F}_0(\mathcal{X}_{\mathrm{in}}, \boldsymbol{w}_j, \delta) \\ \|\boldsymbol{Q}_L \boldsymbol{x}_i\| > 0}} \frac{\boldsymbol{v}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{u}_j}{\|\boldsymbol{Q}_L \boldsymbol{x}_i\|} \leq \sqrt{\delta^2 + \epsilon^2} \max_{\boldsymbol{w} \in L_* \cap S^{D-1}} \#(\mathcal{F}_0(\mathcal{X}_{\mathrm{in}}, \boldsymbol{w}, \delta)). \tag{66}$$

Thus, we finally must deal with the inlier term. We have the inequalities

$$\boldsymbol{v}_j \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{u}_j \geq \cos(\theta_1 - \eta) \sin(\theta_1 - \eta), \tag{67}$$

$$\|\boldsymbol{Q}_L \boldsymbol{x}_i\| \leq \sin(\theta_1) \|\boldsymbol{P}_{L_*} \boldsymbol{x}_i\| + \epsilon. \tag{68}$$

Applying these to the inlier term, as long as $\theta_1 > \eta$, we find

$$-\sum_{\substack{\boldsymbol{x}_i \in \mathcal{F}_1(\mathcal{X}_{\mathrm{in}}, \boldsymbol{w}_j, \delta) \\ \|\boldsymbol{Q}_L \boldsymbol{x}_i\| > 0}} \frac{\boldsymbol{v}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{u}_j}{\|\boldsymbol{Q}_L \boldsymbol{x}_i\|} \tag{69}$$

$$\leq -\frac{\cos(\theta_1 - \eta) \sin(\theta_1 - \eta)}{\sin(\theta_1)} \sum_{\substack{\boldsymbol{x}_i \in \mathcal{F}_1(\mathcal{X}_{\mathrm{in}}, \boldsymbol{w}_j, \delta) \\ \|\boldsymbol{Q}_L \boldsymbol{x}_i\| > 0}} \frac{\boldsymbol{w}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{w}_j}{\|\boldsymbol{P}_{L_*} \boldsymbol{x}_i\| + \epsilon/\sin(\theta_1)}$$

$$\leq -\frac{\cos(\gamma - \eta)}{2} \lambda_d \left(\sum_i \frac{\boldsymbol{P}_{L_*} \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{P}_{L_*}}{\|\boldsymbol{P}_{L_*} \boldsymbol{x}_i\| + \sqrt{\delta^2 + \epsilon^2}} \right).$$

Putting together (65), (66), and (69), we find that

$$\partial F(L(t); \mathcal{X}) \Big|_{t=0} \leq -l \mathcal{S}_n(\mathcal{X}, L_*, \epsilon, \delta, \gamma) < 0. \tag{70}$$

Thus, (70) implies that every subspace in $B(L_*, \gamma) \backslash B(L_*, \eta)$ has a direction with negative local subderivative. From here, the proof is the same as that of Theorem 1.

D.2. Proof of Theorem 3

We will first show that

$$\theta_1(\boldsymbol{V}^{k+1}, L_*) < \theta_1(\boldsymbol{V}^k, L_*), \tag{71}$$

for sufficiently small t^k. Let $\boldsymbol{V}^* \in O(D, d)$ span L_*. We will establish (71) by showing

$$\sigma_d \left(\boldsymbol{V}^{*T} \boldsymbol{V}^{k+1} \right) > \sigma_d \left(\boldsymbol{V}^{*T} \boldsymbol{V}^k \right).$$

Using (22) and the fact that

$$\cos(\boldsymbol{\Sigma}^k t^k) = \boldsymbol{I} - O((t^k)^2), \qquad \sin(\boldsymbol{\Sigma}^k t^k) = \boldsymbol{\Sigma}^k t^k - O((t^k)^3),$$

we can write

$$\sigma_d \left(\boldsymbol{V}^{*T} \boldsymbol{V}^{k+1} \right) = \sigma_d \left(\boldsymbol{V}^{*T} \left(\boldsymbol{V}^k \boldsymbol{W}^k \cos(\boldsymbol{\Sigma}^k t^k) \boldsymbol{W}^{kT} + \boldsymbol{U}^k \sin(\boldsymbol{\Sigma}^k t^k) \boldsymbol{W}^{kT} \right) \right) \tag{72}$$

$$= \sigma_d \left(\boldsymbol{V}^{*T} \left(\boldsymbol{V}^k - t^k \nabla F(\boldsymbol{V}^k; \mathcal{X}) + O((t^k)^2) \right) \right).$$

Let $\boldsymbol{v}_1^k \in \mathrm{Sp}(\boldsymbol{V}^k)$ be a unit vector corresponding to the maximum principal angle with L_*, and let $\boldsymbol{u}_1^k$ be its complementary orthogonal vector. Define the unit vector $\boldsymbol{y}_1^k \in L_* \cap \mathrm{Sp}(\boldsymbol{v}_1^k, \boldsymbol{u}_1^k)$, and write $\theta_1^k = \theta_1(\mathrm{Sp}(\boldsymbol{V}^k), L_*)$. Suppose that $\sigma_d^k = \sigma_d(\boldsymbol{V}^{*T} \boldsymbol{V}^k)$ has multiplicity r, and let $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in O(d, r)$ be such that

$$\boldsymbol{\beta}_1^T \boldsymbol{V}^{*T} \boldsymbol{V}^k \boldsymbol{\beta}_2 = \mathrm{diag}(\sigma_d^k, \dots, \sigma_d^k). \tag{73}$$

We now apply Result 4.1 in Söderström (1999), which states the following. Suppose a matrix $\boldsymbol{A}$ has a singular value σ with multiplicity r, with corresponding left and right

singular vectors $\boldsymbol{U}$ and $\boldsymbol{V}$. Suppose that we perturb $\boldsymbol{A}$ by $\epsilon\boldsymbol{B}$. Then, $\boldsymbol{A}+\epsilon\boldsymbol{B}$ has r singular values $\sigma_1(\boldsymbol{A}+\epsilon\boldsymbol{B}),\ldots,\sigma_r(\boldsymbol{A}+\epsilon\boldsymbol{B})$ which satisfy

$$\sigma_j(\boldsymbol{A}+\epsilon\boldsymbol{B}) = \sigma_j(\boldsymbol{A}) + \frac{\epsilon}{2}\lambda_j\left(\boldsymbol{V}^T\boldsymbol{B}^T\boldsymbol{U} + \boldsymbol{U}^T\boldsymbol{B}\boldsymbol{V}\right) + O(\epsilon^2).$$

Applying this to (72) yields

$$\sigma_d\left(\boldsymbol{V}^{*T}\boldsymbol{V}^{k+1}\right) \geq \sigma_d\left(\boldsymbol{V}^{*T}\boldsymbol{V}^k\right) + \frac{t^k}{2}\lambda_d\left(\boldsymbol{\beta}_1^T\boldsymbol{V}^{*T}\sum_{\mathcal{X}}\frac{\boldsymbol{Q}_{\boldsymbol{V}^k}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{V}^k}{\|\boldsymbol{Q}_{\boldsymbol{V}^k}\boldsymbol{x}_i\|}\boldsymbol{\beta}_2\right) + O((t^k)^2) \quad (74)$$

$$= \sigma_d\left(\boldsymbol{V}^{*T}\boldsymbol{V}^k\right) + t^k\left(\boldsymbol{y}_1^{kT}\sum_{\mathcal{X}}\frac{\boldsymbol{Q}_{\boldsymbol{V}^k}\boldsymbol{x}_i\boldsymbol{x}_i^T}{\|\boldsymbol{Q}_{\boldsymbol{V}^k}\boldsymbol{x}_i\|}\boldsymbol{v}_1^k\right) + O((t^k)^2)$$

$$= \sigma_d\left(\boldsymbol{V}^{*T}\boldsymbol{V}^k\right) + t^k\sin(\theta_1^k)\left(\boldsymbol{u}_1^{kT}\sum_{\mathcal{X}_{\text{in}}}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{\|\boldsymbol{Q}_{\boldsymbol{V}^k}\boldsymbol{x}_i\|}\boldsymbol{v}_1^k + \boldsymbol{u}_1^{kT}\sum_{\mathcal{X}_{\text{out}}}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{\|\boldsymbol{Q}_{\boldsymbol{V}^k}\boldsymbol{x}_i\|}\boldsymbol{v}_1^k\right) + O((t^k)^2).$$

Here, the $O((t^k)^2)$ term is bounded below by $-C_2(t^k)^2$, where C_2 does not depend on $\boldsymbol{V}^k$, which follows from compactness of $O(D,d)$.

Notice that the inlier term in (74) is positive and bounded below

$$\boldsymbol{u}_1^{kT}\sum_{\mathcal{X}_{\text{in}}}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{\|\boldsymbol{Q}_{\boldsymbol{V}^k}\boldsymbol{x}_i\|}\boldsymbol{v}_1^k \geq \frac{1}{\sin(\theta_1(\text{Sp}(\boldsymbol{V}^k),L_*))}\sum_{\mathcal{X}_{\text{in}}}\frac{\boldsymbol{u}_1^{kT}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{v}_1^k}{\|\boldsymbol{x}_i\|} \quad (75)$$

$$\geq \cos(\theta_1(\text{Sp}(\boldsymbol{V}^k),L_*)))\sum_{\mathcal{X}_{\text{in}}}\frac{\boldsymbol{y}_1^{kT}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{y}_1^k}{\|\boldsymbol{x}_i\|} \geq \cos(\gamma)\lambda_d\left(\sum_{\mathcal{X}_{\text{in}}}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{\|\boldsymbol{x}_i\|}\right).$$

Using the fact that $\boldsymbol{u}_1^k \in \text{Sp}(\boldsymbol{Q}_{\boldsymbol{V}^k})$ and $\boldsymbol{v}_1^k \in \text{Sp}(\boldsymbol{V}^k)$, we can bound the outlier term in (74)

$$\left|\boldsymbol{u}_1^{kT}\sum_{\mathcal{X}_{\text{out}}}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{\|\boldsymbol{Q}_{\boldsymbol{V}^k}\boldsymbol{x}_i\|}\boldsymbol{v}_1^k\right| \leq \sigma_1\left(\sum_{\mathcal{X}_{\text{out}}}\frac{\boldsymbol{Q}_{\boldsymbol{V}^k}\boldsymbol{x}_i\boldsymbol{x}_i^T}{\|\boldsymbol{Q}_{\boldsymbol{V}^k}\boldsymbol{x}_i\|}\boldsymbol{V}^k\right) = \sigma_1\left(\nabla F(\boldsymbol{V}^k;\mathcal{X}_{\text{out}})\right). \quad (76)$$

Thus, from (75) and (76) we conclude

$$\sigma_d\left(\boldsymbol{V}^{*T}\boldsymbol{V}^{k+1}\right) - \sigma_d\left(\boldsymbol{V}^{*T}\boldsymbol{V}^k\right) \quad (77)$$

$$\geq t^k\sin(\theta_1^k)\left(\cos(\gamma)\lambda_d\left(\sum_{\mathcal{X}_{\text{in}}}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{\|\boldsymbol{x}_i\|}\right) - \sigma_1\left(\nabla F(\boldsymbol{V}^k;\mathcal{X}_{\text{out}})\right)\right) - C_2(t^k)^2$$

$$\geq t^k\sin(\theta_1^k)\left(\cos(\gamma)\lambda_d\left(\sum_{\mathcal{X}_{\text{in}}}\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{\|\boldsymbol{x}_i\|}\right) - \sup_{\boldsymbol{V}\in B(L_*,\gamma)}\sigma_1\left(\nabla F(\boldsymbol{V};\mathcal{X}_{\text{out}})\right)\right) - C_2(t^k)^2$$

$$\geq t^k\sin(\theta_1^k)C_1 - C_2(t^k)^2 = t^k\left(\sin(\theta_1^k)C_1 - C_2t^k\right),$$

for positive constants C_1 and C_2 which do not depend on $\boldsymbol{V}^k$. Hence, for small enough t^k, we have that (71) holds.

It remains to show that the sequence with step-size $s/\sqrt{k}$ converges to L_* for sufficiently small s. Suppose that s satisfies

$$s < \min\left(\frac{C_1 \sin(\gamma)}{2C_2}, \frac{1}{4\sqrt{C_2}}\right). \tag{78}$$

Then, for any $\boldsymbol{V}^k$ with $\frac{\gamma}{2\sqrt{k}} \le \theta_1(\boldsymbol{V}^k, L_*) \le \gamma$, looking at (77) and the first term in (78), $s/\sqrt{k}$ decreases the principal angle by at least C_3/k, for some constant C_3. On the other hand, for any $\boldsymbol{V}^k$ such that $\theta_1(\boldsymbol{V}^k, L_*) < \frac{\gamma}{2\sqrt{k}}$, we have the bound

$$\sigma_d\left(\boldsymbol{V}^{*T}\boldsymbol{V}^{k+1}\right) - \sigma_d\left(\boldsymbol{V}^{*T}\boldsymbol{V}^k\right) > -C_2(t^k)^2. \tag{79}$$

Note that (79) gives the inequality

$$\sigma_d\left(\boldsymbol{V}^{*T}\boldsymbol{V}^{k+1}\right) > \sigma_d\left(\boldsymbol{V}^{*T}\boldsymbol{V}^k\right) - C_2(t^k)^2 \ge \cos\left(\frac{\gamma}{2\sqrt{k}}\right) - C_2(t^k)^2. \tag{80}$$

It is straightforward to show that if

$$t^k < \frac{1}{4\sqrt{C_2}\sqrt{k}},$$

then the right hand side of (80) is greater than $\cos(\gamma/\sqrt{k})$. Thus, the second term in the minimum of (78) implies that if $\theta_1(\boldsymbol{V}^k, L_*) < \frac{\gamma}{2\sqrt{k}}$, then $\theta_1(\boldsymbol{V}^{k+1}, L_*) < \gamma/\sqrt{k}$.

We summarize this in the following way. For any k, either $\frac{\gamma}{2\sqrt{k}} \le \theta_1(\boldsymbol{V}^k, L_*) \le \gamma$ or $\theta_1(\boldsymbol{V}^k, L_*) < \frac{\gamma}{2\sqrt{k}}$. If the former holds, then $\theta_1(\boldsymbol{V}^{k+1}, L_*) < \theta_1(\boldsymbol{V}^k, L_*) - C_3/k$. If the latter holds, then we have the bound $\theta_1(\boldsymbol{V}^{k+1}, L_*) < \gamma/\sqrt{k}$. Thus, the maximum principal angle with L_* either decreases by C_3/k or the distance is bounded by $\gamma/\sqrt{k}$. Put together, these imply that $\boldsymbol{V}^k$ converges to L_* with $O(1/\sqrt{k})$ rate of convergence.

D.3. Proof of Lemmas in Theorem 4

This appendix contains proofs for the three lemmas in Theorem 4.

D.3.1. Proof of Lemma 6

For any fixed subspace $L \in B(L_*, \gamma)$, let the principle vectors and angles for L with respect to L_* be given by $\{\boldsymbol{v}_j\}_{j=1}^d$ and $\{\theta_j\}_{j=1}^d$. Then, writing $\boldsymbol{\Theta} = \mathrm{diag}(\theta_1, \cdots, \theta_d)$, we know that $F(L_*; \mathcal{X}_{\mathrm{in}}) = 0$ and

$$F(L; \mathcal{X}_{\mathrm{in}}) = \sum_{\boldsymbol{x} \in \mathcal{X}_{\mathrm{in}}} \|\sin(\theta)\boldsymbol{V}^T\boldsymbol{x}\| < \sum_{\boldsymbol{x} \in \mathcal{X}_{\mathrm{in}}} \|\boldsymbol{\Theta}\boldsymbol{V}^T\boldsymbol{x}\| \le \|\boldsymbol{\Theta}\|_2 \max_{\boldsymbol{V} \in O(D,d)} \sum_{\mathcal{X}_{\mathrm{in}}} \|\boldsymbol{V}^T\boldsymbol{x}_i\| \tag{81}$$

$$\le \theta_1 \sum_{\mathcal{X}_{\mathrm{in}}} \|\boldsymbol{x}_i\|.$$

Let $L(t)$ be the geodesic from L_* to L parameterized by arclength, and let $s = \theta_1$. The difference in energies for the outliers is bounded by

$$|F(L; \mathcal{X}_{\text{out}}) - F(L_*; \mathcal{X}_{\text{out}})| = \left| \int_{t=0}^{s} \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X}_{\text{out}}) \mathrm{d}t \right| \tag{82}$$

$$\leq \theta_1 \sup_{L' \in B(L_*, \sin(\gamma))} (\sigma_1(\nabla F(L'; \mathcal{X}_{\text{out}}))).$$

By the assumption $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$, we have by the definition in (12) that

$$\sup_{L' \in B(L_*, \sin(\gamma))} (\sigma_1(\nabla F(L'; \mathcal{X}_{\text{out}}))) < \cos(\gamma) \lambda_d \left(\sum_{\mathcal{X}_{\text{in}}} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\|\boldsymbol{x}_i\|} \right) \leq \max_{\boldsymbol{v} \in S^{D-1}} \left(\sum_{\mathcal{X}_{\text{in}}} \boldsymbol{v}^T \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\|\boldsymbol{x}_i\|} \boldsymbol{v} \right) \tag{83}$$

$$< \max_{\boldsymbol{v} \in S^{D-1}} \sum_{\mathcal{X}_{\text{in}}} |\boldsymbol{v}^T \boldsymbol{x}| \leq \sum_{\mathcal{X}_{\text{in}}} \|\boldsymbol{x}_i\|_2.$$

Thus, combining (81)-(83) yields

$$|F(L; \mathcal{X}) - F(L_*; \mathcal{X})| = |F(L; \mathcal{X}_{\text{in}}) + F(L; \mathcal{X}_{\text{out}}) - (F(L_*; \mathcal{X}_{\text{in}}) + F(L; \mathcal{X}_{\text{out}}))| \tag{84}$$

$$\leq |F(L; \mathcal{X}_{\text{in}})| + |F(L; \mathcal{X}_{\text{out}}) - F(L; \mathcal{X}_{\text{out}})|$$

$$\leq 2\theta_1 \sum_{\mathcal{X}_{\text{in}}} \|\boldsymbol{x}_i\|.$$

As a result, the lemma is proved.

D.3.2. PROOF OF LEMMA 7

For any $L \in B(L_*, \gamma) \setminus \{L_*\}$ and geodesic $L(t)$ parameterized by arclength from L through L_*,

$$\left| \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X})|_{t=0} \right| = \left| \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X}_{\text{in}})|_{t=0} + \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X}_{\text{out}})|_{t=0} \right| \tag{85}$$

$$\geq \left| \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X}_{\text{in}})|_{t=0} \right| - \left| \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X}_{\text{out}})|_{t=0} \right|$$

$$\geq \cos(\gamma) \lambda_1 \left(\sum_{\mathcal{X}_{\text{in}}} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\|\boldsymbol{x}_i\|} \right) - \sup_{L \in B(L_*, \gamma)} \sigma_1(\nabla F(L; \mathcal{X}_{\text{out}}))$$

$$\geq \mathcal{S}(\mathcal{X}, L_*, \gamma) \geq C_1 \sum_{\mathcal{X}_{\text{in}}} \|\boldsymbol{x}_i\|,$$

where, $C_1 = \mathcal{S}(\mathcal{X}, L_*, \gamma) / \sum_{\mathcal{X}_{\text{in}}} \|\boldsymbol{x}_i\|$.

D.3.3. PROOF OF LEMMA 8

First, assume that c_0 is small enough such that

$$\inf_{L(0) \in B(L_*, \gamma) \setminus \{L_*\}} \frac{1}{4} \left| \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X})|_{t=0} \right| > \sum_{\mathcal{X}} 3\sqrt{c_0} \|\boldsymbol{x}_i\|, \tag{86}$$

43

where $L(t)$ is a geodesic parameterized by arclength between $L(0)$ and L_*. We are guaranteed that such a c_0 exists by Lemma 6.

Fix $c < c_0$ and define a geodesic line $L(t)$ on $G(D, d)$ such that $L(0) = L_k$ and $L(c\theta_1(L_k, L_*)) = L_{k+1}$. We first investigate the derivatives of the function $\mathrm{dist}(\boldsymbol{x}, L(t))$. We will then show, under the given assumptions on the data, that $\sum_i \mathrm{dist}(\boldsymbol{x}_i, L(t))$ is close to being Lipschitz.

Applying Lerman and Zhang (2014, (23)-(24)), and assuming that $\sum_{j=1}^d \theta_j^2 = 1$, we have

$$
\frac{\mathrm{d}^2}{\mathrm{d}t^2} \mathrm{dist}(\boldsymbol{x}, L(t)) \tag{87}
$$
$$
= \frac{\mathrm{d}}{\mathrm{d}t} \left[-\frac{\sum_{j=1}^d \theta_j((\cos(t\theta_j)\boldsymbol{v}_j + \sin(t\theta_j)\boldsymbol{u}_j) \cdot \boldsymbol{x})((-\sin(t\theta_j)\boldsymbol{v}_j + \cos(t\theta_j)\boldsymbol{u}_j) \cdot \boldsymbol{x})}{\mathrm{dist}(\boldsymbol{x}, L(t))} \right]
$$
$$
= -\frac{\left(\sum_{j=1}^d \theta_j((\cos(t\theta_j)\boldsymbol{v}_j + \sin(t\theta_j)\boldsymbol{u}_j) \cdot \boldsymbol{x})((-\sin(t\theta_j)\boldsymbol{v}_j + \cos(t\theta_j)\boldsymbol{u}_j) \cdot \boldsymbol{x}) \right)^2}{\mathrm{dist}(\boldsymbol{x}, L(t))^3}
$$
$$
- \frac{\sum_{j=1}^d \theta_j^2((-\sin(t\theta_j)\boldsymbol{v}_j + \cos(t\theta_j)\boldsymbol{u}_j) \cdot \boldsymbol{x})((-\sin(t\theta_j)\boldsymbol{v}_j + \cos(t\theta_j)\boldsymbol{u}_j) \cdot \boldsymbol{x})}{\mathrm{dist}(\boldsymbol{x}, L(t))}
$$
$$
- \frac{\sum_{j=1}^d \theta_j^2((\cos(t\theta_j)\boldsymbol{v}_j + \sin(t\theta_j)\boldsymbol{u}_j) \cdot \boldsymbol{x})((-\cos(t\theta_j)\boldsymbol{v}_j - \sin(t\theta_j)\boldsymbol{u}_j) \cdot \boldsymbol{x})}{\mathrm{dist}(\boldsymbol{x}, L(t))}.
$$

Using $|(\cos(t\theta_j)\boldsymbol{v}_j + \sin(t\theta_j)\boldsymbol{u}_j) \cdot \boldsymbol{x}| \le \|\boldsymbol{x}\|$ and the Cauchy-Schwarz inequality in this equation gives

$$
\left(\sum_{j=1}^d \theta_j((\cos(t\theta_j)\boldsymbol{v}_j + \sin(t\theta_j)\boldsymbol{u}_j) \cdot \boldsymbol{x})((-\sin(t\theta_j)\boldsymbol{v}_j + \cos(t\theta_j)\boldsymbol{u}_j) \cdot \boldsymbol{x}) \right)^2 \tag{88}
$$
$$
\le \left(\sum_{j=1}^d \theta_j^2((\cos(t\theta_j)\boldsymbol{v}_j + \sin(t\theta_j)\boldsymbol{u}_j) \cdot \boldsymbol{x})^2 \right) \left(\sum_{j=1}^d ((-\sin(t\theta_j)\boldsymbol{v}_j + \cos(t\theta_j)\boldsymbol{u}_j) \cdot \boldsymbol{x})^2 \right)
$$
$$
\le \sum_{j=1}^d \theta_j^2 \|\boldsymbol{x}\|^2 \mathrm{dist}(\boldsymbol{x}, L(t))^2 = \|\boldsymbol{x}\|^2 \mathrm{dist}(\boldsymbol{x}, L(t))^2.
$$

Putting (87) and (88) together yields

$$
\left| \frac{\mathrm{d}^2}{\mathrm{d}t^2} \mathrm{dist}(\boldsymbol{x}, L(t)) \right| \le \frac{\|\boldsymbol{x}\|^2 \mathrm{dist}(\boldsymbol{x}, L(t))^2}{\mathrm{dist}(\boldsymbol{x}, L(t))^3} + \frac{2\|\boldsymbol{x}\|^2 \sum_{j=1}^d \theta_j^2}{\mathrm{dist}(\boldsymbol{x}, L(t))} = \frac{3\|\boldsymbol{x}\|^2}{\mathrm{dist}(\boldsymbol{x}, L(t))}. \tag{89}
$$

On the other hand, Lerman and Zhang (2014, Lemma 3.2) implies that

$$
\left| \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{dist}(\boldsymbol{x}, L(t)) \right| \le \|\boldsymbol{x}\|. \tag{90}
$$

Then, define the set

$$
\mathcal{G}(\mathcal{X}, L_*, L(t), c) = \left\{ \boldsymbol{x} \in \mathcal{X} : \min_{t \in [0, c\theta_1(L, L_*)]} \frac{\mathrm{dist}(\boldsymbol{x}, L(t))}{\|\boldsymbol{x}\|} \le \sqrt{c}\theta_1(L_k, L_*) \right\}. \tag{91}
$$

Then, for all $0 \leq s \leq c\theta_1(L_k, L_*)$,

$$\left| \sum_{\boldsymbol{x} \in \mathcal{X}} \left(\frac{\mathrm{d}}{\mathrm{d}t} \operatorname{dist}(x, L(t)) \Big|_{t=s} - \frac{\mathrm{d}}{\mathrm{d}t} \operatorname{dist}(x, L(t)) \Big|_{t=0} \right) \right| \tag{92}$$

$$\leq \sum_{\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{G}(\mathcal{X}, L_*, L(t), c)} \int_{t=0}^{s} \left| \frac{\mathrm{d}^2}{\mathrm{d}t^2} \operatorname{dist}(\boldsymbol{x}, L(t)) \right| \mathrm{d}t + \sum_{\boldsymbol{x} \in \mathcal{G}(\mathcal{X}, L_*, L(t), c)} 2\|\boldsymbol{x}\|$$

$$\leq \sum_{\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{G}(\mathcal{X}, L_*, L(t), c)} s \frac{3\|\boldsymbol{x}\|}{\sqrt{c}\theta_1(L_k, L_*)} + \sum_{\boldsymbol{x} \in \mathcal{G}(\mathcal{X}, L_*, L(t), c)} 2\|\boldsymbol{x}\|$$

$$\leq \sum_{\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{G}(\mathcal{X}, L_*, L(t), c))} 3\sqrt{c}\|\boldsymbol{x}\| + \sum_{\boldsymbol{x} \in \mathcal{G}(\mathcal{X}, L_*, L(t), c)} 2\|\boldsymbol{x}\|.$$

In the last line of (92), the first term can be made as small as one would like by taking c_0 small enough. On the other hand, for c small enough, all points in $\mathcal{G}(\mathcal{X}, L_*, L(t), c)$ must be contained in a subspace $L \in B(L_*, \gamma) \setminus \{L_*\}$. Thus, for small enough c, the function $F(L(t); \mathcal{X})$ is approximately Lipschitz for $t \in [0, c\theta_1(L_k, L_*)]$.

As a result,

$$F(L_k; \mathcal{X}) - F(L_{k+1}; \mathcal{X})) = F(L(0); \mathcal{X}) - F(L(c\theta_1(L_k, L_*)); \mathcal{X})$$

$$= -\int_{t=0}^{c\theta_1(L_k, L_*)} \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X}) \mathrm{d}t$$

$$\geq -c\theta_1(L_k, L_*) \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X})|_{t=0}$$

$$- c\theta_1(L_k, L_*) \left(\sum_{\boldsymbol{x} \in \mathcal{X} \setminus \mathcal{G}(L_k, L_{k+1})} 3\sqrt{c}\|\boldsymbol{x}\| + \sum_{\boldsymbol{x} \in \mathcal{G}(L_k, L_{k+1})} 2\|\boldsymbol{x}\| \right).$$

Finally, noting that $\frac{\mathrm{d}}{\mathrm{d}t} F(L(t))|_{t=0} < 0$, we use (86) and the assumption in (23) to find

$$F(L_k) - F(L_{k+1})) \geq \left| c\theta_1(L_k, L_*) \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X})|_{t=0} \right| - \frac{c\theta_1(L_k, L_*)}{2} \left| \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X})|_{t=0} \right|$$

$$\geq \frac{c\theta_1(L_k, L_*)}{2} \left| \frac{\mathrm{d}}{\mathrm{d}t} F(L(t); \mathcal{X})|_{t=0} \right|.$$

D.4. Proof of Inlier Permeance Bounds

D.4.1. Proof of Proposition 11

The continuous distribution assumption implies that all directions in $S^{D-1} \cap L_*$ have nonzero probability. Let $\boldsymbol{x}$ be a random variable following the inlier distribution. By the central limit theorem, we have that

$$\frac{1}{N_{\mathrm{in}}} \sum_{\mathcal{X}_{\mathrm{in}}} \left(\frac{|\boldsymbol{v}^T \boldsymbol{x}_i|^2}{\|\boldsymbol{x}_i\|} - \operatorname{Var}\left(\frac{\boldsymbol{v}^T \boldsymbol{x}}{\|\boldsymbol{x}\|^{1/2}} \right) \right) \xrightarrow{d} N(0, 1).$$

We also have that $\max_i \|\boldsymbol{x}_i\| < M$ by the bounded-support assumption assumption. Thus, for large enough N_{in} and a covering argument, we have for some absolute constant C that

$$\min_{\boldsymbol{v} \in L_* \cap S^{D-1}} \sum_{\mathcal{X}_{\text{in}}} \frac{|\boldsymbol{v}^T \boldsymbol{x}_i|^2}{\|\boldsymbol{x}_i\|} \geq C \frac{N_{\text{in}}}{M} \min_{\boldsymbol{v} \in L_* \cap S^{D-1}} \text{Var} \left(\frac{\boldsymbol{v}^T \boldsymbol{x}}{\|\boldsymbol{x}\|^{1/2}} \right), \text{ w.h.p.}$$

This can be done by using, e.g., Hoeffding's inequality on the bounded random variable $\boldsymbol{v}^T \boldsymbol{x} / \|\boldsymbol{x}\|^{1/2}$.

D.4.2. PROOF OF PROPOSITION 12

For the inliers, we need to bound $\mathcal{P}(\mathcal{X}_{\text{in}})$. Using Theorem 3.1 in Lu and Pearce (2000) and letting $\tilde{\cdot}$ denote the spherization operator,

$$\lambda_d \left(\sum_{\mathcal{X}_{\text{in}}} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\|\boldsymbol{x}_i\|} \right) = \sigma_d(\widetilde{\boldsymbol{X}_{\text{in}}} \boldsymbol{X}_{\text{in}}^T) \geq \sigma_d(\widetilde{\boldsymbol{X}_{\text{in}}}) \sigma_d(\boldsymbol{X}_{\text{in}}). \tag{93}$$

We proceed by bounding the last term in (93). Notice that $\boldsymbol{\Sigma}_{\text{in}}^{-1/2} \boldsymbol{X}_{\text{in}}$ has directional variance $1/d$ in all directions of L_* (i.e., these transformed inliers are isotropic with variance $1/d$). We again apply Theorem 3.1 in Lu and Pearce (2000) and then apply Theorem 5.39 in Vershynin (2012b) to $\sigma_d \left(\boldsymbol{\Sigma}_{\text{in}}^{-1/2} \boldsymbol{X}_{\text{in}} \right)$, where we scale by the standard deviation $1/\sqrt{d}$ and choose $t = a\sqrt{N_{\text{in}}}$:

$$\sigma_d \left(\boldsymbol{X}_{\text{in}} \right) = \sigma_d \left(\boldsymbol{\Sigma}_{\text{in}}^{1/2} \boldsymbol{\Sigma}_{\text{in}}^{-1/2} \boldsymbol{X}_{\text{in}} \right) \geq \sigma_d \left(\boldsymbol{\Sigma}_{\text{in}}^{1/2} \right) \sigma_d \left(\boldsymbol{\Sigma}_{\text{in}}^{-1/2} \boldsymbol{X}_{\text{in}} \right) \tag{94}$$

$$> \lambda_d \left(\boldsymbol{\Sigma}_{\text{in}}^{1/2} \right) \left((1-a)\sqrt{\frac{N_{\text{in}}}{d}} - C_1 \right), \text{ w.p. at least } 1 - 2e^{-c_1 a^2 N_{\text{in}}}.$$

Here, c_1 and C_1 are constants that depend on the sub-Gaussian norm of $\boldsymbol{\Sigma}_{\text{in}}^{-1/2} \boldsymbol{x}$, where $\boldsymbol{x}$ follows the inlier distribution, and a must be chosen such that $(1-a)^2 N_{\text{in}} > C_1^2 d$. On the other hand, $\widetilde{\boldsymbol{X}_{\text{in}}}$ is still sub-Gaussian and $\widetilde{\boldsymbol{\Sigma}_{\text{in}}^{-1/2} \boldsymbol{X}_{\text{in}}}$ is isotropic sub-Gaussian. Therefore, by using (94) and applying Theorem 3.1 in Lu and Pearce (2000), we have

$$\sigma_d \left(\widetilde{\boldsymbol{X}_{\text{in}}} \right) = \sigma_d \left(\left[\frac{\boldsymbol{x}_1}{\|\boldsymbol{x}_1\|}, \dots, \frac{\boldsymbol{x}_{N_{\text{in}}}}{\|\boldsymbol{x}_{N_{\text{in}}}\|} \right] \right) \tag{95}$$

$$= \sigma_d \left(\boldsymbol{\Sigma}_{\text{in}}^{1/2} \left[\frac{\boldsymbol{\Sigma}_{\text{in}}^{-1/2} \boldsymbol{x}_1}{\|\boldsymbol{\Sigma}_{\text{in}}^{1/2} \boldsymbol{\Sigma}_{\text{in}}^{-1/2} \boldsymbol{x}_1\|}, \dots, \frac{\boldsymbol{\Sigma}_{\text{in}}^{-1/2} \boldsymbol{x}_{N_{\text{in}}}}{\|\boldsymbol{\Sigma}_{\text{in}}^{1/2} \boldsymbol{\Sigma}_{\text{in}}^{-1/2} \boldsymbol{x}_{N_{\text{in}}}\|} \right] \right)$$

$$\geq \sigma_d \left(\frac{\boldsymbol{\Sigma}_{\text{in}}^{1/2}}{\|\boldsymbol{\Sigma}_{\text{in}}^{1/2}\|_2} \left[\frac{\boldsymbol{\Sigma}_{\text{in}}^{-1/2} \boldsymbol{x}_1}{\|\boldsymbol{\Sigma}_{\text{in}}^{-1/2} \boldsymbol{x}_1\|}, \dots, \frac{\boldsymbol{\Sigma}_{\text{in}}^{-1/2} \boldsymbol{x}_{N_{\text{in}}}}{\|\boldsymbol{\Sigma}_{\text{in}}^{-1/2} \boldsymbol{x}_{N_{\text{in}}}\|} \right] \right)$$

$$\geq \frac{\lambda_d(\boldsymbol{\Sigma}_{\text{in}}^{1/2})}{\lambda_1(\boldsymbol{\Sigma}_{\text{in}}^{1/2})} \sigma_d \left(\widetilde{\boldsymbol{\Sigma}_{\text{in}}^{-1/2} \boldsymbol{X}_{\text{in}}} \right)$$

$$> \frac{\lambda_d(\boldsymbol{\Sigma}_{\text{in}}^{1/2})}{\lambda_1(\boldsymbol{\Sigma}_{\text{in}}^{1/2})} \left((1-a)\sqrt{\frac{N_{\text{in}}}{d}} - C_1' \right), \text{ w.p. at least } 1 - 2e^{-c_1' a^2 N_{\text{in}}}.$$

Here, c_1' and C_1' are constants that depend on the sub-Gaussian norm of $\widetilde{\boldsymbol{\Sigma}_{\text{in}}^{-1/2}\boldsymbol{x}}$, where $\boldsymbol{x}$ is a random vector that follows the inlier distribution, and a must chosen such that $(1-a)^2 N_{\text{in}} > C_1'^2 d$. Therefore, for a to satisfy both requirements, we need $(1-a)^2 N_{\text{in}} > \max(C_1, C_1')^2 d$. Abusing notation, we let c_1 be the minimum of c_1 and c_1', and we let C_1 be the maximum of C_1 and C_1'. Putting (94) and (95) together, we find

$$\lambda_d \left(\sum_{\mathcal{X}_{\text{in}}} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\|\boldsymbol{x}_i\|} \right) > \frac{\lambda_d(\boldsymbol{\Sigma}_{\text{in}})}{\lambda_1(\boldsymbol{\Sigma}_{\text{in}})^{1/2}} \left((1-a)\sqrt{\frac{N_{\text{in}}}{d}} - C_1 \right)^2, \text{ w.p. at least } 1 - 4e^{-c_1 a^2 N_{\text{in}}}. \quad (96)$$

D.5. Proof of Proposition 15

To prove this proposition, we only need to show that PCA initializes in $B(L_*, \gamma)$. Note that the covariance matrix for the data $\mathcal{X}$ is $\boldsymbol{\Sigma} = N_{\text{out}}\boldsymbol{\Sigma}_{\text{out}}/(N D_{\text{out}}) + N_{\text{in}}\boldsymbol{\Sigma}_{\text{in}}/(Nd)$. We want to see how close the sample covariance approximates $\boldsymbol{\Sigma}_{\text{in}}$, since the PCA subspace is spanned by the top d eigenvectors of the sample covariance. Denoting the sample covariance by $\boldsymbol{\Sigma}_N$, let $\boldsymbol{V}_{PCA}$ be its top d eigenvectors. Also, let $\boldsymbol{V}^*$ be the top d eigenvectors of $\boldsymbol{\Sigma}_{\text{in}}$ and let $\boldsymbol{V}$ be the top d eigenvectors of $\boldsymbol{\Sigma}$. We note that

$$|\sin(\theta_1(\boldsymbol{V}_{PCA}, \boldsymbol{V}^*))| \leq |\sin(\theta_1(\boldsymbol{V}_{PCA}, \boldsymbol{V}))| + |\sin(\theta_1(\boldsymbol{V}, \boldsymbol{V}^*))|. \quad (97)$$

To deal with the last term in (97), the Davis-Kahan $\sin\theta$ Theorem (Davis and Kahan, 1970), or more precisely Corollary 3.1 of Vu et al. (2013), gives

$$|\sin(\theta_1(\boldsymbol{V}, \boldsymbol{V}^*)| \leq \sqrt{2} \frac{\lambda_1(N_{\text{out}}\boldsymbol{\Sigma}_{\text{out}}/(N D_{\text{out}}))}{\lambda_d(N_{\text{in}}\boldsymbol{\Sigma}_{\text{in}}/(Nd))} = \sqrt{2} \frac{N_{\text{out}}}{N_{\text{in}}} \frac{d}{D_{\text{out}}} \frac{\lambda_1(\boldsymbol{\Sigma}_{\text{out}})}{\lambda_d(\boldsymbol{\Sigma}_{\text{in}})}.$$

Thus,

$$\sin(\gamma) > \sqrt{2} \frac{N_{\text{out}}}{N_{\text{in}}} \frac{d}{D_{\text{out}}} \frac{\lambda_1(\boldsymbol{\Sigma}_{\text{out}})}{\lambda_d(\boldsymbol{\Sigma}_{\text{in}})} \quad (98)$$

is a sufficient condition for $|\sin(\theta_1(\boldsymbol{V}, \boldsymbol{V}^*)| < \sin(\gamma)$.

On the other hand, similar to the derivation of (94), for the first term in the right hand side of (97), we must bound how close the sample covariance is to the true covariance. Proposition 2.1 of Vershynin (2012a) implies that, for every $\delta > 0$,

$$\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_N\|_2 \leq \epsilon(\delta, \boldsymbol{\Sigma}_{\text{in}}, \boldsymbol{\Sigma}_{\text{out}}) \left(\frac{D_{\text{out}}}{N} \right)^{\frac{1}{2}}, \quad (99)$$

with probability at least $1 - \delta$, where $\epsilon(\delta, \boldsymbol{\Sigma}_{\text{in}}, \boldsymbol{\Sigma}_{\text{out}})$ is a constant depending on δ, $\boldsymbol{\Sigma}_{\text{in}}$, and $\boldsymbol{\Sigma}_{\text{out}}$. By assumption, $\boldsymbol{\Sigma}$ has a positive dth eigengap. Thus, another application of the Davis-Kahan $\sin\theta$ Theorem yields

$$|\sin(\theta_1(\boldsymbol{V}_{PCA}, \boldsymbol{V}))| \leq \frac{\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_N\|_2}{\lambda_d(\boldsymbol{\Sigma}) - \lambda_{d+1}(\boldsymbol{\Sigma})} \leq \epsilon_2(\delta, \boldsymbol{\Sigma}_{\text{in}}, \boldsymbol{\Sigma}_{\text{out}}) \left(\frac{dD}{N} \right)^{\frac{1}{2}},$$

with probability $1 - \delta$. For large enough N, we have

$$|\sin(\theta_1(\boldsymbol{V}_{PCA}, \boldsymbol{V}))| \leq \sin(\gamma) - |\sin(\theta_1(\boldsymbol{V}, \boldsymbol{V}^*))|. \quad (100)$$

Rearranging (100) and using the triangle inequality in (97) yields

$$|\sin(\theta_1(\boldsymbol{V}_{PCA}, \boldsymbol{V}^*))| \leq \sin(\gamma),$$

with probability $1 - \delta$.

D.6. Proof of Theorem 17

We begin by bounding the outlier term. The last term on the right hand side of (47) is optimally bounded in the proof of Theorem 14. On the other hand, in the following proposition, we obtain tighter bounds for the first term on the right hand side of this inequality.

Proposition 22 *Suppose that $\mathcal{X}$ is drawn from the Haystack Model with parameters N_{in}, σ_{in}, N_{out}, σ_{out}, D, and d. Then, for $D - d \geq 3$*

$$\max_{L \in G(D,d)} \left\| \widetilde{\boldsymbol{Q}_L \boldsymbol{X}_{\mathrm{out}}} \right\|_2 \leq \frac{7}{2} \sqrt{\frac{N_{\mathrm{out}}}{D - d}} + \sqrt{2}, \tag{101}$$

$$w.p. \ at \ least \ 1 - \exp\left(-\frac{N_{\mathrm{out}}}{4}\right) - C_1 \exp\left(-\frac{N_{\mathrm{out}}}{4(D - d)} + \frac{d(D - d)\log(D)}{2}\right),$$

where C_1 is an absolute constant.

Proof [Proof of Proposition 22] First, for a single L, Lemma 8.4 in Lerman et al. (2015) yields

$$\left\| \widetilde{\boldsymbol{Q}_L \boldsymbol{X}_{\mathrm{out}}} \right\|_2 \leq \left(\sqrt{\frac{N_{\mathrm{out}}}{D - d - 0.5}} + \sqrt{2} + \frac{t}{\sqrt{D - d - 0.5}} \right), \ \text{w.p. at least } 1 - 1.5e^{-t^2}. \tag{102}$$

We set $t = \sqrt{N_{\mathrm{out}}}/2$ to obtain the desired bound for a single subspace L. However, in order to cover all of $G(D, d)$, we much use a more complicated argument.

Assume that we have two subspaces L_0 and L_1 such that $\theta_1(L_0, L_1) < 1/(2\sqrt{D})$. First, due to the triangle inequality,

$$\left\| \widetilde{\boldsymbol{Q}_{L_0} \boldsymbol{X}_{\mathrm{out}}} - \widetilde{\boldsymbol{Q}_{L_1} \boldsymbol{X}_{\mathrm{out}}} \right\|_2 < \left\| \widetilde{\boldsymbol{Q}_{L_0} \boldsymbol{X}_{\mathrm{out}}^0} - \widetilde{\boldsymbol{Q}_{L_1} \boldsymbol{X}_{\mathrm{out}}^0} \right\|_2 + \left\| \widetilde{\boldsymbol{Q}_{L_0} \boldsymbol{X}_{\mathrm{out}}^1} - \widetilde{\boldsymbol{Q}_{L_1} \boldsymbol{X}_{\mathrm{out}}^1} \right\|_2. \tag{103}$$

Here, we have partitioned the outliers into two parts. The following set-valued functions define these parts. Let $L(t)$ be the geodesic between L_0 and L_1 such that $L(0) = L_0$ and $L(1) = L_1$. Then, we define

$$\mathcal{X}_{\mathrm{out}}^0 = \left\{ \boldsymbol{x} \in \mathcal{X}_{\mathrm{out}} : \min_{t \in [0,1]} \angle(\boldsymbol{x}, L(t)) < \frac{1}{2\sqrt{D}} \right\}, \tag{104}$$

$$\mathcal{X}_{\mathrm{out}}^1 = \left\{ \boldsymbol{x} \in \mathcal{X}_{\mathrm{out}} : \min_{t \in [0,1]} \angle(\boldsymbol{x}, L(t)) \geq \frac{1}{2\sqrt{D}} \right\}. \tag{105}$$

Notice that

$$\mathcal{X}_{\mathrm{out}}^0 \subset \left\{ \boldsymbol{x} \in \mathcal{X}_{\mathrm{out}} : \angle(\boldsymbol{x}, L_0) < \frac{1}{\sqrt{D}} \right\}. \tag{106}$$

With these datasets, their data matrices are $\boldsymbol{X}^0_{\text{out}}$ and $\boldsymbol{X}^1_{\text{out}}$.

For the last term in (103), since $\theta_1(L_0, L_1) < 1/(2\sqrt{D})$, we have

$$\left\| \widetilde{\boldsymbol{Q}_{L_0}\boldsymbol{X}^1_{\text{out}}} - \widetilde{\boldsymbol{Q}_{L_1}\boldsymbol{X}^1_{\text{out}}} \right\|_2 < \frac{1}{2}\sqrt{\frac{N_{\text{out}}}{D}}. \tag{107}$$

On the other hand, we must look at the concentration of points around subspaces for the second to last term in (103). For any given L, we have the following concentration lemma.

Lemma 23 *If $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, then*

$$\Pr\left(\angle(\boldsymbol{x}, L) < \frac{1}{2}\sqrt{\frac{D-d}{D}} \right) < \exp\left(-\frac{D-d}{2} \right).$$

Proof This is a direct consequence of Lemma 2.2 in Dasgupta and Gupta (2003). ∎

For $D - d \geq 3$, we have

$$\Pr\left(\angle(\boldsymbol{x}, L) < \frac{1}{2}\sqrt{\frac{D-d}{D}} \right) \leq \exp\left(-\frac{D-d}{2} \right) \leq \frac{1}{D-d}. \tag{108}$$

Using (108) together with a loose Chernoff bound for the concentration of binomial random variables (Mitzenmacher and Upfal, 2005), we know that

$$\Pr\left(\#\left(\left\{ \boldsymbol{x} \in \mathcal{X}_{\text{out}} : \angle(\boldsymbol{x}, L) < \frac{1}{2}\sqrt{\frac{D-d}{D}} \right\} \right) > \frac{3}{2}\frac{N_{\text{out}}}{D-d} \right) < \exp\left(-\frac{N_{\text{out}}}{4(D-d)} \right). \tag{109}$$

Thus, we use (109) on the second to last term in (103) to find

$$\left\| \widetilde{\boldsymbol{Q}_{L_0}\boldsymbol{X}^0_{\text{out}}} - \widetilde{\boldsymbol{Q}_{L_1}\boldsymbol{X}^0_{\text{out}}} \right\|_2 < \sqrt{\#(\mathcal{X}^0_{\text{out}})} \leq \sqrt{\frac{3}{2}\frac{N_{\text{out}}}{D-d}}, \tag{110}$$

$$\text{w.p. at least } 1 - \exp\left(-\frac{N_{\text{out}}}{4(D-d)} \right).$$

Notice that also, by construction, this is true for all $L'_0, L'_1 \in B(L(t), \frac{1}{4}\sqrt{(D-d)/D})$.

We finish by completing the covering argument on $G(D, d)$. By remark 8.4 of (Szarek, 1983), $G(D, d)$ can be covered by $(C)^{d(D-d)}/(\gamma_1)^{d(D-d)}$ balls of radius γ_1. This means that, using a union bound and taking (102), (103), (107), and (110) together,

$$\Pr\left(\max_{L \in G(D,d)} \|\widetilde{\boldsymbol{Q}_L\boldsymbol{X}_{\text{out}}}\| < \frac{7}{2}\sqrt{\frac{N_{\text{out}}}{D-d-0.5}} + \sqrt{2} \right) > \tag{111}$$

$$1 - C_1 \exp\left(-\frac{N_{\text{out}}}{4(D-d)} + C_2 d(D-d)\log\left(\frac{D}{D-d} \right) \right) - \exp\left(-\frac{N_{\text{out}}}{4} \right),$$

where C_1 and C_2 are absolute constants. Thus, we see that N must be on the order $N = O(d(D-d)^2 \log(D))$, and we have have the desired statement. ∎

We now continue with the proof of Theorem 17. In Proposition 22, We see that N needs to be on the order of $d(D-d)^2 \log(D)$, which is not the case of the small sample regime. On the other hand, Theorem 5.39 of Vershynin (2012b) states that

$$\|\boldsymbol{X}_{\text{out}}\|_2 \le \sigma_{\text{out}} \left[\frac{5}{4}\sqrt{\frac{N_{\text{out}}}{D}} + 1\right], \text{ w.p. at least } 1 - 2e^{-N_{\text{out}}/16}. \tag{112}$$

We combine the result of Proposition 22 with (112) to find that

$$\|\widetilde{\boldsymbol{Q}_L \boldsymbol{X}_{\text{out}}}\|_2 \|\boldsymbol{X}_{\text{out}}\|_2 \le \frac{7}{2}\frac{N_{\text{out}}}{\sqrt{D(D-d)}} + O\left(\sigma_{\text{out}}\sqrt{\frac{N_{\text{out}}}{D}}\right), \tag{113}$$

w.p. at least $1 - 2e^{-N_{\text{out}}/16} - e^{-N_{\text{out}}/4} - C_1 \exp\left(-\frac{N_{\text{out}}}{4(D-d)} + C_2 d(D-d)\log\left(\frac{D}{D-d}\right)\right)$.

We can also improve the results for the inlier permeance bound since they are already isometric. In this case, we have that (94) still holds. On the other hand, $\widetilde{\boldsymbol{X}}_{\text{in}}$ is isotropic, which implies that

$$\sigma_d\left(\widetilde{\boldsymbol{X}}_{\text{in}}\right) > \left((1-a)\sqrt{\frac{N_{\text{in}}}{d}} - C_1'\right), \text{ w.p. at least } 1 - 2e^{-c_1' a^2 N_{\text{in}}}, \tag{114}$$

where as before we must have $(1-a)^2 N_{\text{in}}^2 > C_1'^2 d$. Abusing notation again (as we did in the proof of Proposition 12), we let c_1 be the minimum of c_1 and c_1' and C_1 be the max of C_1 and C_1' and we find that

$$\lambda_d\left(\sum_{\mathcal{X}_{\text{in}}} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\|\boldsymbol{x}_i\|}\right) > \sigma_{\text{in}}\left((1-a)\sqrt{\frac{N_{\text{in}}}{d}} - C_1\right)^2, \text{ w.p. at least } 1 - 4e^{-c_1 a^2 N_{\text{in}}}. \tag{115}$$

This results in the following exact statement for the Haystack Model. We combine (113) with (115) to find that if (48) holds, then $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$ with probability at least (49).

D.7. Proof of Theorem 18

Here, we will show that GGD can recover the underlying subspace for any fixed fraction of outliers, provided that N_{out} is large enough. We denote the percentage of inliers by α_1 and the percentage of outliers by α_0. Then, the SNR is given by α_1/α_0.

Assume for simplicity that $\sigma_{\text{out}} = 1$, since the general case follows from the same logic. First, we will bound the maximum norm of the set of outlier points. Each $\boldsymbol{x}_i$ is i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}/D)$, we can bound its norm by

$$\Pr\left(\|\boldsymbol{x}_i\| \le 1 + \frac{t}{\sqrt{D}}\right) \ge 1 - \exp\left(-ct^2\right), \tag{116}$$

where c is just some universal constant. Thus, applying a union bound to (116) yields

$$\Pr\left(\max_i \|\boldsymbol{x}_i\| \le 1 + \frac{t}{\sqrt{D}}\right) \ge 1 - N_{\text{out}}\exp\left(-ct^2\right). \tag{117}$$

We can use a value of $t = \sqrt{D} N_{\text{out}}^{1/6}$ to find

$$\Pr\left(\max_i \|\boldsymbol{x}_i\| \leq 1 + N^{1/6}\right) \geq 1 - N_{\text{out}} \exp\left(-cD N_{\text{out}}^{1/3}\right). \tag{118}$$

Next, consider the alignment for a single fixed subspace $L \neq L_*$:

$$\mathcal{A}(\mathcal{X}_{\text{out}}, L) = \left\| \sum_{\mathcal{X}_{\text{out}}} \frac{\boldsymbol{Q}_L \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{P}_L}{\|\boldsymbol{Q}_L \boldsymbol{x}_i\|} \right\|_2. \tag{119}$$

Taking an expectation within the norm in (119) yields

$$\mathbb{E} \sum_{\mathcal{X}_{\text{out}}} \frac{\boldsymbol{Q}_L \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{P}_L}{\|\boldsymbol{Q}_L \boldsymbol{x}_i\|} = \boldsymbol{0}. \tag{120}$$

This follows by the symmetry of the outlier distribution, $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}/D)$. Fix arbitrary vectors $\boldsymbol{u} \in \text{Sp}(\boldsymbol{Q}_L) \cap S^{D-1}$ and $\boldsymbol{v} \in \text{Sp}(\boldsymbol{P}_L) \cap S^{D-1}$. Then, (120) implies that

$$\mathbb{E} \sum_{\mathcal{X}_{\text{out}}} \frac{\boldsymbol{u}^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{v}}{\|\boldsymbol{Q}_L \boldsymbol{x}_i\|} = 0. \tag{121}$$

In the following, we will continue to use $\boldsymbol{u}$ and $\boldsymbol{v}$ defined in this way.

To this end, define the following random variable

$$J(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = \frac{\boldsymbol{u}^T \boldsymbol{x} \boldsymbol{x}^T \boldsymbol{v}}{\|\boldsymbol{Q}_L \boldsymbol{x}\|}, \tag{122}$$

where $\boldsymbol{x}$ is distributed $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}/D)$ (i.e., it is an outlier in the Haystack model). We will first give a concentration bound for $\sum_{\mathcal{X}_{\text{out}}} J(\boldsymbol{x}_i, \boldsymbol{u}, \boldsymbol{v})$, which appears in (121). Notice that $J(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})$ is a mean zero random variable and is bounded by

$$|J(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})| \leq |\boldsymbol{x}^T \boldsymbol{v}|. \tag{123}$$

Here, $\boldsymbol{x}^T \boldsymbol{v}$ is Gaussian with variance $1/D$. Thus, $J(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})$ is sub-Gaussian with variance proxy $1/D$. This implies that

$$\Pr\left(\left|\sum_{\mathcal{X}_{\text{out}}} J(\boldsymbol{x}_i, \boldsymbol{u}, \boldsymbol{v})\right| > N_{\text{out}} t\right) \leq 2 \exp\left(-\frac{t^2 D N_{\text{out}}}{2}\right). \tag{124}$$

Letting $t = N_{\text{out}}^{-1/3}$ we find that

$$\Pr\left(\left|\sum_{\mathcal{X}_{\text{out}}} J(\boldsymbol{x}_i, \boldsymbol{u}, \boldsymbol{v})\right| > N_{\text{out}}^{2/3}\right) \leq 2 \exp\left(-\frac{D N_{\text{out}}^{1/3}}{2}\right). \tag{125}$$

Notice that $J(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})$ is continuous as a function of $\boldsymbol{u}$ and $\boldsymbol{v}$. Using (123), between two points $\boldsymbol{v}_1, \boldsymbol{v}_2 \in S^{D-1}$, we can bound the deviation in $\sum_{\mathcal{X}_{\text{out}}} J(\boldsymbol{x}_i, \boldsymbol{u}, \cdot)$ by

$$\left|\sum_{\mathcal{X}_{\text{out}}} J(\boldsymbol{x}_i, \boldsymbol{u}, \boldsymbol{v}_1) - \sum_{\mathcal{X}_{\text{out}}} J(\boldsymbol{x}_i, \boldsymbol{u}, \boldsymbol{v}_2)\right| \leq \sum_{\mathcal{X}_{\text{out}}} |(\boldsymbol{v}_1 - \boldsymbol{v}_2)^T \boldsymbol{x}_i| \leq \|\boldsymbol{v}_1 - \boldsymbol{v}_2\| N_{\text{out}} \max_i \|\boldsymbol{x}_i\|. \tag{126}$$

Combining (118), (125) and (126) with $\|\boldsymbol{v}_1 - \boldsymbol{v}_2\| < N_{\text{out}}^{-1/3}$ yields

$$\Pr\left(\max_{\boldsymbol{v}_2:\|\boldsymbol{v}_2-\boldsymbol{v}_1\|\leq N_{\text{out}}^{-1/3}}\left|\sum_i J(\boldsymbol{x}_i,\boldsymbol{u},\boldsymbol{v}_2)\right| \leq 3N_{\text{out}}^{5/6}\right) \geq 1 - N_{\text{out}}\exp\left(-cDN_{\text{out}}^{1/3}\right) \quad (127)$$

$$- 2\exp\left(-\frac{DN_{\text{out}}^{1/3}}{2}\right).$$

Since $L \cap S^{D-1}$ can be covered by $\left(2N_{\text{out}}^{1/3}+1\right)^d$ balls of radius $N_{\text{out}}^{-1/3}$, we use a covering argument with (127) to find

$$\Pr\left(\max_{\boldsymbol{v}\in L\cap S^{D-1}}\left|\sum_{\mathcal{X}_{\text{out}}} J(\boldsymbol{x}_i,\boldsymbol{u},\boldsymbol{v})\right| \leq 3N_{\text{out}}^{5/6}\right) \geq 1 - N_{\text{out}}\exp\left(-cDN_{\text{out}}^{1/3}\right) \quad (128)$$

$$- 2\left(2N_{\text{out}}^{1/3}+1\right)^d\exp\left(-\frac{DN_{\text{out}}^{1/3}}{2}\right).$$

Notice that we do not include the probability from (118) in the covering, since this probability holds independently of the choice of $\boldsymbol{v} \in L \cap S^{D-1}$.

The previous argument was actually independent of choice of $\boldsymbol{u} \in L^\perp \cap S^{D-1}$. Thus, we finally have

$$\Pr\left(\left\|\sum_{\mathcal{X}_{\text{out}}}\frac{\boldsymbol{Q}_L\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{P}_L}{\|\boldsymbol{Q}_L\boldsymbol{x}_i\|}\right\|_2 \leq 3N_{\text{out}}^{5/6}\right) \geq \quad (129)$$

$$1 - N_{\text{out}}\exp\left(-cDN_{\text{out}}^{1/3}\right) - 2\left(2N_{\text{out}}^{1/3}+1\right)^d\exp\left(-\frac{DN_{\text{out}}^{1/3}}{2}\right).$$

Now that we have covered a single $L \in G(D,d)$, we must extend this to all of $G(D,d)$ by another covering argument. From Lemma 2.2 of Dasgupta and Gupta (2003), we have, for each $\boldsymbol{x}_i \in \mathcal{X}_{\text{out}}$

$$\Pr\left(\angle(\boldsymbol{x}_i,L) < \frac{\pi}{2}\sqrt{\frac{\beta(D-d)}{D}}\right) \leq \exp\left(\frac{D-d}{2}(1+\log(\beta))\right). \quad (130)$$

If we choose $\beta < N_{\text{out}}^{-2/3}$, then it is not hard to show that

$$\Pr\left(\angle(\boldsymbol{x}_i,L) < \frac{\pi}{2}\sqrt{\frac{\beta(D-d)}{D}}\right) \leq N_{\text{out}}^{-1/3}. \quad (131)$$

Define the cone around a subspace L as

$$\mathcal{C}(L,\xi) = \{\boldsymbol{x}\in\mathbb{R}^D : \angle(\boldsymbol{x},L) < \xi\}. \quad (132)$$

Using a loose Chernoff bound for the concentration of binomial random variables (Mitzenmacher and Upfal, 2005), we have

$$\Pr\left(\left|\#\left(\mathcal{X}_{\text{out}} \cap \mathcal{C}\left(L, N_{\text{out}}^{-1/3}\frac{\pi}{2}\sqrt{\frac{(D-d)}{D}}\right)\right) - \delta N_{\text{out}}^{2/3}\right| \geq N_{\text{out}}^{2/3}\right) \leq \exp\left(-\frac{N_{\text{out}}^{2/3}\delta^2}{3}\right).$$
(133)

Choosing $\delta = N_{\text{out}}^{-1/6}$ yields

$$\Pr\left(\left|\#\left(\mathcal{X}_{\text{out}} \cap \mathcal{C}\left(L, N_{\text{out}}^{-1/3}\frac{\pi}{2}\sqrt{\frac{(D-d)}{D}}\right)\right) - N_{\text{out}}^{1/2}\right| \geq N_{\text{out}}^{2/3}\right) \leq \exp\left(-\frac{N_{\text{out}}^{1/3}}{3}\right). \quad (134)$$

For any $L_0 \in G(D, d)$, and $L_1 \in B(L_0, \xi)$, we can separate the alignment term into two parts: those $\boldsymbol{x}_i$ that are close to L_0 and L_1 and those that lie further away. The idea behind this is that points that are far away from L_0 and L_1 will contribute similar amounts to the alignment. On the other hand, those that are very close can contribute at most 2 times their norm to the alignment.

Rigorously, we write

$$\left\|\sum_{\mathcal{X}_{\text{out}}} \frac{\boldsymbol{Q}_{L_0}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{P}_{L_0}}{\|\boldsymbol{Q}_{L_0}\boldsymbol{x}_i\|_2} - \frac{\boldsymbol{Q}_{L_1}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{P}_{L_1}}{\|\boldsymbol{Q}_{L_1}\boldsymbol{x}_i\|_2}\right\|_2 \tag{135}$$

$$\leq \left\|\sum_{\boldsymbol{x}_i \in \mathcal{X}_{\text{out}} \cap \mathcal{C}(L_0, \xi)} \frac{\boldsymbol{Q}_{L_0}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{P}_{L_0}}{\|\boldsymbol{Q}_{L_0}\boldsymbol{x}_i\|_2} - \frac{\boldsymbol{Q}_{L_1}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{P}_{L_1}}{\|\boldsymbol{Q}_{L_1}\boldsymbol{x}_i\|_2}\right\|_2 +$$

$$\left\|\sum_{\boldsymbol{x}_i \in \mathcal{X}_{\text{out}} \cap \mathcal{C}(L_0, \xi)^C} \frac{\boldsymbol{Q}_{L_0}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{P}_{L_0}}{\|\boldsymbol{Q}_{L_0}\boldsymbol{x}_i\|_2} - \frac{\boldsymbol{Q}_{L_1}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{P}_{L_1}}{\|\boldsymbol{Q}_{L_1}\boldsymbol{x}_i\|_2}\right\|_2$$

$$\leq \sum_{\boldsymbol{x}_i \in \mathcal{X}_{\text{out}} \cap \mathcal{C}(L_0, \xi)} 2\|\boldsymbol{x}_i\| + 2\xi \sum_{\boldsymbol{x}_i \in \mathcal{X}_{\text{out}} \cap \mathcal{C}(L_0, \xi)^C} \|\boldsymbol{x}_i\|_2.$$

We will examining (135) term by term. For the first term, if we choose $\xi = N_{\text{out}}^{-1/3}$ and combine (118) and (134), we find that

$$\sum_{\boldsymbol{x}_i \in \mathcal{X}_{\text{out}} \cap \mathcal{C}(L_0, \xi)} 2\|\boldsymbol{x}_i\| \leq 2\#(\mathcal{X}_{\text{out}} \cap \mathcal{C}(L_0, \xi)) \max_{\mathcal{X}_{\text{out}}} \|\boldsymbol{x}_i\| \leq 2\left(1 + N_{\text{out}}^{1/6}\right)N_{\text{out}}^{2/3}, \tag{136}$$

$$\text{w.p. at least } 1 - N_{\text{out}}\exp\left(-cDN_{\text{out}}^{1/3}\right) - \exp\left(-\frac{N_{\text{out}}^{1/3}}{3}\right).$$

For the second term, choosing $\xi = N_{\text{out}}^{-1/3}$ and again using (118) yields

$$2\xi \sum_{\boldsymbol{x}_i \in \mathcal{X}_{\text{out}} \cap \mathcal{C}(L_0, \xi)^C} \|\boldsymbol{x}_i\|_2 \leq 2\left(1 + N_{\text{out}}^{1/6}\right)N_{\text{out}}^{2/3} + o(1), \tag{137}$$

$$\text{w.p. at least } 1 - N_{\text{out}}\exp\left(-cDN_{\text{out}}^{1/3}\right).$$

Putting (136) and (137) yields

$$\left\| \sum_{\mathcal{X}_{\text{out}}} \frac{\boldsymbol{Q}_{L_0} \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{P}_{L_0}}{\|\boldsymbol{Q}_{L_0} \boldsymbol{x}_i\|_2} - \frac{\boldsymbol{Q}_{L_1} \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{P}_{L_1}}{\|\boldsymbol{Q}_{L_1} \boldsymbol{x}_i\|_2} \right\|_2 \leq 8 N_{\text{out}}^{5/6}, \tag{138}$$

w.p. at least $1 - N_{\text{out}} \exp\left(-cD N_{\text{out}}^{1/3}\right) - \exp\left(-\frac{N_{\text{out}}^{1/3}}{3}\right).$

As already mentioned in Appendix D.6 (following Szarek (1983)), $G(D, d)$ can be covered by $(C_1)^{d(D-d)}/(\gamma_1)^{d(D-d)}$ balls of radius γ_1. Thus, by a union bound with (129) and (138),

$$\left\| \sum_{\mathcal{X}_{\text{out}}} \frac{\boldsymbol{Q}_L \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{P}_L}{\|\boldsymbol{Q}_L \boldsymbol{x}_i\|} \right\|_2 \leq 11 N_{\text{out}}^{5/6}, \ \forall \ L \in G(D, d), \tag{139}$$

w.p. at least $1 - N_{\text{out}} \exp\left(-cD N_{\text{out}}^{1/3}\right) - 2\left((C_1)^{d(D-d)} (N_{\text{out}}^{1/3})^{d(D-d)}\right).$

$$\left(\left(2 N_{\text{out}}^{1/3} + 1\right)^d \exp\left(-\frac{D N_{\text{out}}^{1/3}}{2}\right) + \exp\left(-\frac{N_{\text{out}}^{1/3}}{3}\right) \right).$$

As $N_{\text{out}} \to \infty$, we see that this probability goes to 1.

On the other hand, we recall the bound on the permeance of inliers from (115). Combining this with (139) and adding back in the scale factor σ_{out}, we have that if

$$\cos(\gamma) \sigma_{\text{in}} \left((1-a) \sqrt{\frac{N_{\text{in}}}{d}} - C_1 \right)^2 \geq 11 \sigma_{\text{out}} N_{\text{out}}^{5/6}, \tag{140}$$

then $\mathcal{S}(\mathcal{X}, L_*, \gamma) > 0$ w.o.p. This equates to

$$\text{SNR} \geq \frac{11 d \sigma_{\text{out}}}{\cos(\gamma) N_{\text{out}}^{1/6} \sigma_{\text{in}} (1-a)^2} + O\left(\frac{\sqrt{N_{\text{in}}}}{N_{\text{out}}}\right), \tag{141}$$

which goes to 0 as $N \to \infty$ for any fixed fraction of outliers. We see, in terms of dependence on parameters in (139) and (141), N_{out} must be at least on the order of $O(\max(d^3 D^3 \log^3(N_{\text{out}}), (d N_{\text{out}}/N_{\text{in}})^6))$. This is due to the fact that (141) must hold and the limiting probability in (139) is

$$2\left((C_1)^{d(D-d)} (N_{\text{out}}^{1/3})^{d(D-d)}\right) \exp\left(-\frac{N_{\text{out}}^{1/3}}{3}\right),$$

which is only close to zero when the number of outliers is more than the specified regime.

References

P.-A. Absil, R. Mahony, and R. Sepulchre. Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematica*, 80(2):199–220, 2004.

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

E. Arias-Castro and J. Wang. RANSAC algorithms for subspace recovery and subspace clustering. *ArXiv e-prints*, November 2017.

E. Arias-Castro, G. Chen, and G. Lerman. Spectral clustering based on local linear approximations. *Electron. J. Statist.*, 5:1537–1587, 2011.

S. Arora, R. Ge, T. Ma, and A. Moitra. Simple, efficient, and neural algorithms for sparse coding. In *COLT*, 2015.

J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

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

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

Y. Cherapanamjeri, P. Jain, and P. Netrapalli. Thresholding based outlier robust PCA. In *COLT*, pages 593–628, 2017.

F. H. Clarke. *Optimization and nonsmooth analysis*, volume 5. SIAM, 1990.

K. L. Clarkson and D. P. Woodruff. Input sparsity and hardness for robust subspace approximation. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 310–329. IEEE, 2015.

M. Coudron and G. Lerman. On the sample complexity of robust PCA. In *NIPS*, pages 3221–3229. 2012.

S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

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

C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM J. on Numerical Analysis*, 7:1–46, 1970.

C. Ding, D. Zhou, X. He, and H. Zha. R1-PCA: rotational invariant L_1-norm principal component analysis for robust subspace factorization. In *ICML*, pages 281–288. ACM, 2006.

A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353 (electronic), 1999. ISSN 0895-4798.

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *COLT*, pages 797–842, 2015.

R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *NIPS*, pages 2973–2981, 2016.

J. Goes, T. Zhang, R. Arora, and G. Lerman. Robust stochastic principal component analysis. *JMLR W&CP*, page 266274, 2014.

M. Hardt. Understanding alternating minimization for matrix completion. In *FOCS*, pages 651–660. IEEE, 2014.

M. Hardt and A. Moitra. Algorithms and hardness for robust subspace recovery. In *COLT*, pages 354–375, 2013.

P. Jain, A. Tewari, and P. Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *NIPS*, pages 685–693, 2014.

I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.

I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2nd edition, 2002.

Y. Ledyaev and Q. Zhu. Nonsmooth analysis on smooth manifolds. *Transactions of the American Mathematical Society*, 359(8):3687–3732, 2007.

J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In *COLT*, pages 1246–1257, 2016.

G. Lerman and T. Maunu. Fast, robust and non-convex subspace recovery. *Information and Inference: A Journal of the IMA*, 7(2):277–336, 2018a.

G. Lerman and T. Maunu. An overview of robust subspace recovery. *Proceedings of the IEEE*, 106(8):1380–1410, Aug 2018b. ISSN 0018-9219. doi: 10.1109/JPROC.2018. 2853141.

G. Lerman and T. Zhang. Robust recovery of multiple subspaces by geometric l_p minimization. *Ann. Statist.*, 39(5):2686–2715, 2011.

G. Lerman and T. Zhang. l_p-recovery of the most significant subspace among multiple subspaces with outliers. *Constructive Approximation*, 40(3):329–385, 2014.

G. Lerman, M. B. McCoy, J. A. Tropp, and T. Zhang. Robust computation of linear models by convex relaxation. *Foundations of Computational Mathematics*, 15(2):363–410, 2015.

L. Lim, K. S. Wong, and K. Ye. Statistical estimation and the affine Grassmannian. *arXiv preprint arXiv:1607.01833*, 2016.

L. Lim, K. S. Wong, and K. Ye. The Grassmannian of affine subspaces. *arXiv preprint arXiv:1807.10883*, 2018.

N. Locantore, J. S. Marron, D. G. Simpson, N. Tripoli, J. T. Zhang, and K. L. Cohen. Robust principal component analysis for functional data. *Test*, 8(1):1–73, 1999.

L.-Z. Lu and C. E. M. Pearce. Some new bounds for singular values and eigenvalues of matrix products. *Annals of Operations Research*, 98(1-4):141–148, 2000.

C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *PMLR*, volume 80, pages 3345–3354, 10–15 Jul 2018.

R. A. Maronna. Principal components and orthogonal regression based on robust scales. *Technometrics*, 47:264–273, 2005. ISSN 1537-2723.

R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust statistics: Theory and methods*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2006. ISBN 978-0-470-01092-1; 0-470-01092-4.

T. Maunu and G. Lerman. Robust subspace recovery with adverserial outliers. In preparation.

M. McCoy and J. A Tropp. Two proposals for robust PCA using semidefinite programming. *Electronic Journal of Statistics*, 5:1123–1160, 2011.

S. Mei, Y. Bai, and A. Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

S. Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4): 2308–2335, 2015.

M. Mitzenmacher and E. Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge university press, 2005.

P. Netrapalli, U. N. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain. Non-convex robust PCA. In *NIPS*, pages 1107–1115, 2014.

M. R. Osborne and G. A. Watson. An analysis of the total approximation problem in separable norms, and an algorithm for the total l_1 problem. *SIAM journal on scientific and statistical computing*, 6(2):410–424, 1985.

M. Rahmani and G. K. Atia. Coherence pursuit: Fast, simple, and robust principal component analysis. *IEEE Transactions on Signal Processing*, 65(23):6260–6275, 2016.

M. Rudelson and R. Vershynin. The Littlewood–Offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600 – 633, 2008. ISSN 0001-8708.

T. Söderström. *Perturbation results for singular values*. Institutionen för informationsteknologi, Uppsala universitet, 1999.

M. Soltanolkotabi and E. J. Candès. A geometric analysis of subspace clustering with outliers. *Ann. Stat.*, 40(4):2195–2238, 2012. doi: 10.1214/12-AOS1034.

B. St. Thomas, L. Lin, L. Lim, and S. Mukherjee. Learning subspaces of different dimension. *arXiv preprint arXiv:1404.6841*, 2014.

J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere. In *SAMPTA*, pages 407–410, May 2015a.

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

S. J. Szarek. The finite-dimensional basis problem with an appendix on nets of Grassmann manifolds. *Acta Math.*, 151(3-4):153–179, 1983. ISSN 0001-5962.

R. Vershynin. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012a.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012b.

V. Q. Vu, J. Cho, J. Lei, and K. Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In *NIPS*, pages 2670–2678, 2013.

G. A. Watson. *Some Problems in Orthogonal Distance and Non-Orthogonal Distance Regression*. Defense Technical Information Center, 2001. URL `http://books.google.com/books?id=WKKWGwAACAAJ`.

H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. *IEEE Trans. Information Theory*, 58(5):3047–3064, 2012.

H. Xu, C. Caramanis, and S. Mannor. Outlier-robust PCA: the high-dimensional case. *IEEE Trans. Information Theory*, 59(1):546–572, 2013.

K. Ye and L. Lim. Schubert varieties and distances between subspaces of different dimensions. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1176–1197, 2016.

X. Yi, D. Park, Y. Chen, and C. Caramanis. Fast algorithms for robust PCA via gradient descent. In *NIPS*, pages 4152–4160, 2016.

D. Zhang and L. Balzano. Global convergence of a Grassmannian gradient descent algorithm for subspace estimation. In *AISTATS*, pages 1460–1468, 2016.

T. Zhang. Robust subspace recovery by Tyler's M-estimator. *Information and Inference*, 5 (1):1–21, 2016.

T. Zhang and G. Lerman. A novel M-estimator for robust PCA. *Journal of Machine Learning Research*, 15(1):749–808, 2014.

T. Zhang and Y. Yang. Robust principal component analysis by manifold optimization. *arXiv preprint arXiv:1708.00257*, 2017.

T. Zhang, A. Szlam, and G. Lerman. Median K-flats for hybrid linear modeling with many outliers. In *International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 234–241, Kyoto, Japan, 2009.

Z. Zhou, X. Li, J. Wright, E. Candès, and Y. Ma. Stable principal component pursuit. In *International Symposium on Information Theory Proceedings (ISIT)*, pages 1518–1522. IEEE, 2010.