OUTLIER-ROBUST ORTHOGONAL REGRESSION ON MANIFOLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Motivated by machine learning and computer vision applications, we formulate the problem of Outlier-Robust Orthogonal Regression to find a point in a manifold that satisfies as many linear equations as possible. Existing approaches addressing special cases of our formulation either lack theoretical support, are computationally costly, or somewhat ignore the manifold constraint; the latter two limit them from many applications. In this paper, we propose a unified approach based on solving a non-convex and non-smooth ℓ^1 optimization problem over the manifold. We give conditions on the geometry of the input data, the manifold, and their interplay, under which the minimizers recover the ground truth; notably the conditions can hold even when the inliers are skewed within the true hyperplane. We provide a Riemannian subgradient method and an iteratively reweighted least squares method, suiting different computational oracles, and prove their linear/sub-linear convergence to minimizers/critical points. Experiments demonstrate that respecting the manifold constraints increases robustness against outliers in robust essential matrix estimation and robust rotation search.

1 INTRODUCTION

Given a dataset $\mathcal{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_L] \in \mathbb{R}^{D\times L}$ of L points and a manifold $\mathcal{M} \subset \mathbb{R}^D$, we propose *Outlier-Robust Orthogonal Regression* (OR^2), formulated using the ℓ^0 semi-norm as

$$\min_{\boldsymbol{b}} \quad \left\|\mathcal{Y}^\top \boldsymbol{b}\right\|_0 \quad \text{s.t.} \quad \boldsymbol{b} \in \mathcal{M}, \quad \boldsymbol{b} \neq \mathbf{0}. \qquad (\text{OR}^2\text{-}\ell^0)$$

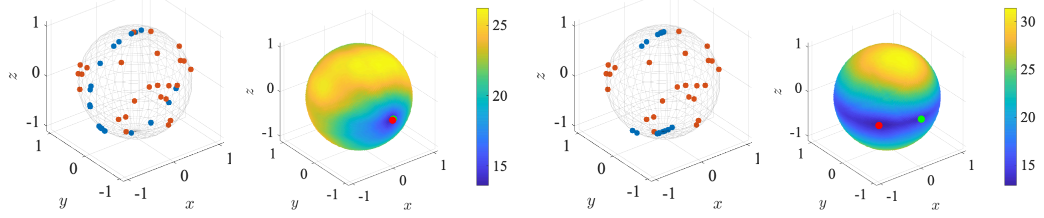
In words, (OR^2-ℓ^0) is an ℓ^0 *regression* problem of finding some vector $\boldsymbol{b}^*$ from $\mathcal{M}$ that is *orthogonal* to as many points in $\mathcal{Y}$ as possible. We assume such $\boldsymbol{b}^*$ is orthogonal to N points in $\mathcal{Y}$ ($N \leq L$). Points of $\mathcal{Y}$ orthogonal to $\boldsymbol{b}^*$ are called *inliers*, or otherwise *outliers*. By solving (OR^2-ℓ^0) one not only finds $\boldsymbol{b}^*$, but also identifies inliers and outliers, thus (OR^2-ℓ^0) is an *outlier-robust* formulation.

OR^2 has massive applications in *machine learning* such as robust subspace recovery ($\mathcal{M}$ is the unit sphere), hyperplane clustering (the Chow variety), fixed-rank matrix sensing (fixed-rank matrices), and in *geometric vision* including robust rotation search ($\mathbb{SO}(3)$), essential matrix estimation (the essential manifold), and trifocal tensor estimation (the calibrated trifocal tensor variety), among many others. To further motivate the paper, we begin with a tour of two important instances of OR^2.

OR^2 over the Sphere $\mathbb{S}^{D-1}$. In the late 1980s, Spath & Watson (1987) first considered the problem

$$\min_{\boldsymbol{b}} \quad \left\|\mathcal{Y}^\top \boldsymbol{b}\right\|_1 \quad \text{s.t.} \quad \boldsymbol{b} \in \mathcal{M} = \mathbb{S}^{D-1}, \qquad (1)$$

and called it ℓ^1 *orthogonal regression*, which motivates the name of this paper. (1) can be viewed as the ℓ^1 relaxation of (OR^2-ℓ^0) with a choice of $\mathcal{M} = \mathbb{S}^{D-1}$ to fix $\boldsymbol{b}$'s scale, since (OR^2-ℓ^0) is invariant to scalings. It was not until the past decade that a fruitful line of research considered (1) (sometimes with other surrogate objectives) in robust subspace recovery (Tsakiris & Vidal, 2015; 2017; 2018b; Zhu et al., 2018; Ding et al., 2019a; 2021), dictionary learning (Qu et al., 2014; Sun et al., 2017b;a; Bai et al., 2018; Gilboa et al., 2018), and sparse blind deconvolution (Qu et al., 2019); see Qu et al. (2020) for a summary. Despite being non-convex non-smooth, (1) enjoys strong guarantees of successful recovery. For example, Zhu et al. (2018) shows that (1) is robust to as many outliers as the square of the number of inliers, and Zhu et al. (2018) furthermore provides linearly convergent algorithms to find $\boldsymbol{b}^*$.



(a) Uniform inliers and outliers

(b) Non-uniform inliers and outliers

Figure 1: Inliers (blue) and outliers (orange) over $\mathbb{S}^2$ (left); objective of (1) over $\mathbb{S}^2$ (right), with the true $\boldsymbol{b}^*$ (green) and the minimizer (red).

OR^2 over the Essential Manifold. Given two images of a static scene taken from a camera that undergoes a rotation $\boldsymbol{R}^* \in \mathbb{SO}(3)$ and translation $\boldsymbol{t}^*$, how can we estimate $\boldsymbol{R}^*, \boldsymbol{t}^*$? This is a classic textbook question (Ma et al., 2003; Hartley & Zisserman, 2004) in geometric vision and has been a building block of modern autonomous driving systems. To solve it, some matching algorithm is typically run first on the two images to produce *correspondences* $\{(\boldsymbol{z}_j, \boldsymbol{z}'_j)\}_{j=1}^{L} \subset \mathbb{S}^2 \times \mathbb{S}^2$. *Inlier correspondences* $(\boldsymbol{z}_j, \boldsymbol{z}'_j)$ satisfy

$$\boldsymbol{z}_j^\top \boldsymbol{E}^* \boldsymbol{z}'_j = 0 \quad \Leftrightarrow \quad \operatorname{tr}(\boldsymbol{z}'_j \boldsymbol{z}_j^\top \boldsymbol{E}^*) = 0, \tag{2}$$

where $\boldsymbol{E}^* = [\boldsymbol{t}^*]_\times \boldsymbol{R}^*$ is the so-called *essential matrix*[1], lying in the *essential manifold*

$$\mathcal{M}_E := \{\boldsymbol{E} \in \mathbb{R}^{3\times 3} : \quad \boldsymbol{E} = [\boldsymbol{t}]_\times \boldsymbol{R}/\sqrt{2}, \quad \boldsymbol{t} \in \mathbb{S}^2, \quad \boldsymbol{R} \in \mathbb{SO}(3)\}, \tag{3}$$

in which we restricted $\boldsymbol{t} \in \mathbb{S}^2$ as (2) is invariant to the scale of $\boldsymbol{t}$. Note then that (2) is equivalent to

$$\langle \boldsymbol{y}_j, \boldsymbol{b}^* \rangle = 0, \quad \text{where} \quad \boldsymbol{y}_j = \operatorname{vec}(\boldsymbol{z}'_j \boldsymbol{z}_j^\top), \quad \boldsymbol{b}^* = \operatorname{vec}(\boldsymbol{E}^*), \tag{4}$$

which naturally leads to two different approaches for estimating $\boldsymbol{E}^*$ (and therefore $\boldsymbol{R}^*$ and $\boldsymbol{t}^*$):

- (*Essential Relaxation*?) Observe that $\boldsymbol{E}^*$ satisfies $\operatorname{tr}(\boldsymbol{E}^*(\boldsymbol{E}^*)^\top) = 1$, so $\boldsymbol{b}^* = \operatorname{vec}(\boldsymbol{E}^*)$ lies in a *proper* subset of $\mathbb{S}^8$. Therefore a first approach is to relax the constraint $\boldsymbol{E}^* \in \mathcal{M}_E$ and simply assume that $\boldsymbol{b}^*$ in (4) lies in the unit sphere $\mathbb{S}^8$. This idea is classic in geometric vision and has received significant popularity (Hartley & Zisserman, 2004). In particular, with this relaxation, robustly estimating an essential matrix reduces to (1), which can be solved via diverse sets of methods mentioned above and notably (Ding et al., 2020b), with powerful theoretical guarantees. However, these guarantees all build on the assumption that inliers and outliers are somewhat uniformly distributed, which is violated here. Specifically, the entries of each $\boldsymbol{y}_j = \operatorname{vec}(\boldsymbol{z}'_j \boldsymbol{z}_j^\top)$ in (4) are correlated; it has been further proved that often $\boldsymbol{y}_j$'s are *clustered* in some proper subspace (Ding et al., 2020b) under certain *degeneracies*, rather than uniformly distributed. Figure 1 presents a visual analogy in $\mathbb{R}^3$. On the left of Figure 1a (*resp.* Figure 1b), the blue points are uniform (*resp.* non-uniform) inliers and orange points are outliers. Then we see that, for uniform inliers, (1) has a unique up-to-sign minimizer in red coinciding with $\boldsymbol{b}^*$ in green (Figure 1a, right), while non-uniform inliers, in contrast, lead to an arbitrary minimizer far from $\boldsymbol{b}^*$ (Figure 1b, right).
- (*OR^2 on $\mathcal{M}_E$*?) To address the issue, one can consider ℓ^1 orthogonal regression directly on $\mathcal{M}_E$:

$$\min_{\boldsymbol{b}} \quad \left\| \mathcal{Y}^\top \boldsymbol{b} \right\|_1 \quad \text{s.t.} \quad \boldsymbol{b} \in \mathcal{M} = \mathcal{M}_E. \tag{5}$$

That being said, we are not aware of any existing theoretical guarantees for (5) regarding its minimizers or recovery algorithms. More broadly, to the best of our knowledge, such guarantees have rarely if ever been derived for many other similar outlier-robust estimation problems in geometric vision that admit an orthogonal regression formulation and are associated naturally with manifold constraints (e.g., rotations). This thus invokes a strong call for carrying out a theoretical and principled investigation into orthogonal regression in a general, manifold optimization context.

[1] For any $\boldsymbol{a} \in \mathbb{R}^3$, $[\boldsymbol{a}]_\times := \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix}$ is a matrix representing cross product, i.e., $[\boldsymbol{a}]_\times \boldsymbol{b} = \boldsymbol{a} \times \boldsymbol{b}$.

Contributions. Going beyond unit spheres and essential manifolds, we propose to study

$$\min_{\boldsymbol{b}} \quad \left\|\mathcal{Y}^\top \boldsymbol{b}\right\|_1 \quad \text{s.t.} \quad \boldsymbol{b} \in \mathcal{M} \tag{OR^2-ℓ^1}$$

where $\mathcal{M}$ is assumed to be a smooth and compact submanifold of the unit sphere $\mathbb{S}^{D-1}$ (Assumption 1). This is a non-convex and non-smooth optimization problem. We consider two efficient algorithms for (OR^2-ℓ^1): a Riemannian subgradient method (Algorithm 1) when one can compute a retraction map of $\mathcal{M}$ and a projection map onto $T_{\boldsymbol{b}}\mathcal{M}$, or otherwise an iteratively reweighted least squares method (Algorithm 2) when a weighted least squares problem over $\mathcal{M}$ can be easily solved.

- *Landscape*: We provide conditions on the geometry of the inliers, outliers, $\mathcal{M}$, and their interplay, under which the minimizer of (OR^2-ℓ^1) recovers (Theorems 1 and 2) or is close to (Theorem 3) the ground truth. Notably, (OR^2-ℓ^1) admits a successful recovery even if the inliers are skewed within the true hyperplane, as long as $\mathcal{M}$ is in some sense posed close to the mode of inliers (Figure 2) and does not curve too much near the ground truth (Figure 3); a novel insight to our knowledge.
- *Convergence*: We prove that the Riemannian subgradient method converges linearly to the ground truth with suitable initialization and step sizes (Theorem 4). Moreover, we show that the iteratively reweighted least squares converges globally sub-linearly to a critical point of a surrogate to (Theorem 5); this has not been shown in the literature even for particular instances of (OR^2-ℓ^1).
- *Experiments*: We show that respecting the manifold constraints (as opposed to relaxing them to the unit sphere ones) increases robustness against outliers in tasks of robust estimation of essential matrix (§5) and rotation (A.2) with different parameters (e.g., number of points, outlier ratios).

2 Prior Art: Outlier-Robust Estimation

Here we review two types of popular approaches for outlier-robust estimation in geometric vision.

RANSAC. One of the most widely applied approach for (OR^2-ℓ^0) is RANdom SAmple Consensus (Fischler & Bolles, 1981) that arised in the early 1980s. The most basic version operates by iteratively i) sampling say $d < D$ points from $\mathcal{Y}$, ii) fitting to them a model from $\mathcal{M}$, iii) computing a score that measures how many points in $\mathcal{Y}$ are consensus with this model. It terminates and returns the model with the highest score, when the probability of having sampled an outlier-free model is above a prescribed threshold.

A well-known drawback of RANSAC is, however, that the iteration complexity increases dramatically with the percent of outliers in $\mathcal{Y}$ as well as d. Efforts mitigating the latter have unfortunately led to a dilemma: While it is possible to reduce d to the theoretical lower bound, the dimension of $\mathcal{M}$, by utilizing as many algebraic constraints from $\mathcal{M}$ as possible (Nistér, 2004), yet decreasing d undermines the accuracy of step ii) against noise in $\mathcal{Y}$.

We conclude by remarking that numerous heuristics have been proposed to improve RANSAC in aspects of sampling (Chum & Matas, 2005; Torr et al., 2002; Barath et al., 2019a), score function (Torr & Zisserman, 2000; Stewart, 1995; Moisan et al., 2016), degeneracy handling (Chum et al., 2005; Ivashechkin et al., 2021), refitting (Chum et al., 2003; Lebeda et al., 2012; Barath & Matas, 2018; Barath et al., 2020; 2019b; Yang et al., 2021), and combinations thereof (Raguram et al., 2013; Ivashechkin et al.). Despite their empirical success on specific problem and dataset instances, little has been done on the theoretical understanding of when and why these variants offer accurate or efficient recovery. Indeed, the only theoretical result we are aware of is on the *basic* RANSAC algorithm in the problem of robust subspace recovery (Maunu & Lerman, 2019, §4).

Continuous Robust Surrogates. Since the ℓ^0 objective of (OR^2-ℓ^0) is discrete, one naturally considers its continuous surrogates to faciliate continuous optimization. Indeed, people have replaced the objective in (OR^2-ℓ^0) with ℓ^1, Huber, Logcosh or truncated quadratic, albeit these works are restricted to particular instances of $\mathcal{M}$, to be reviewed in the sequel. Observe that the relaxed problem is still *non-convex* since the feasible set $\mathcal{M}$ is so.

When $\mathcal{M}$ is the set of rotations $\mathbb{SO}(3)$, Lajoie et al. (2018); Yang & Carlone (2019; 2022); Peng et al. (2022a) considers a surrogate of (OR^2-ℓ^0) using the truncated quadratic objective. Notably, they established its equivalence to a quadratically constrainted quadratic programming (QCQP) , which is in turn relaxed to a semi-definite programming (SDP), a *convex* problem. This allows computing and analyzing the minimizer of QCQP and SDP problems under certain conditions; it has been shown to be rather robust to noise and outliers. However, this approach scales poorly with the number of samples L: indeed, the SDP has $O(L^2)$ variables, hindering real-time computation for even $L = 50$.

A fruitful series of research consider instead optimizing directly some surrogates of (OR^2-ℓ^0), allowing much faster computation. When $\mathcal{M}$ is $\mathbb{S}^{D-1}$, the minimizer of (OR^2-ℓ^1) recovers the ground truth (Zhu et al., 2018), and (Projected) Riemannian sub-gradient method with suitable initialization and step sizes converges (piece-wise (Zhu et al., 2018)) linearly (Zhu et al., 2019; Li et al., 2019) to the minimizer. Nevertheless, these results requires, among other conditions, that the inliers to be uniformly distributed in the hyperplane orthogonal to the true $\boldsymbol{b}^*$, a condition that can often be violated, e.g., in geometric vision (see Example 1 and experiments). On the other hand, when $\mathcal{M}$ is the normalize essential variety, a family of surrogates *empirically* enjoyes accurate and efficient recovery (Zhao, 2022); this in principle applies to a rich family of $\mathcal{M}$ as long as weighted least squares problems can be solved (§3.2). Be that as it may, there is no theoretical guarantee whatsoever telling when the minimizer is the ground truth or why algorithms should converge to it.

3 Formulation

In this section, we first formulate the problem in a precise manner (§3.1), and then discuss two algorithms useful when different mappings can be efficiently computed with respect to the manifold $\mathcal{M}$ (§3.2). The theoretical analysis of the objective and algorithms are deferred to §4.

3.1 Set up: Data Model, Assumptions, Notations

To begin with, we define the manifold and the data model, making clear the associated assumptions. The first assumption we make is the following.

Assumption 1 (Regularity on manifold)**.** $\mathcal{M}$ is a smooth and compact submanifold of $\mathbb{S}^{D-1}$.

Let $\mathcal{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_L] \in \mathbb{R}^{D \times L}$ be a given *normalized* dataset, i.e., $\boldsymbol{y}_i \in \mathbb{S}^{D-1}$. When the context is clear we also treat $\mathcal{Y}$ as the set $\{\boldsymbol{y}_i\}_{i=1}^{L}$ with some abuse of notation. Let $\mathcal{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N]$ be the subset of $\mathcal{Y}$ containing points from $\mathcal{Y}$ that are orthogonal to a ground truth vector $\boldsymbol{b}^* \in \mathcal{M}$, which we refer to as *inliers*. It turns out to be handy to define the ground truth hyperplane $\mathcal{H} := \operatorname{span}(\boldsymbol{b}^*)^\perp$. On the other hand, denote the points from $\mathcal{Y} \backslash \mathcal{X}$ by $\mathcal{O} = [\boldsymbol{o}_1, \dots, \boldsymbol{o}_M]$, termed *outliers*; observe that $L = N + M$. Finally, we make a mild assumption on the outliers.

Assumption 2 (Outliers are in general position)**.** Any size-D subset of $\mathcal{O}$ are linearly independent.

Remark 1. The requirements that points $\boldsymbol{y}_i$'s from $\mathcal{Y}$ lie on the unit sphere $\mathbb{S}^{D-1}$ and $\mathcal{M}$ is a submanifold of $\mathbb{S}^{D-1}$ do not compromise generality. Note that the objective of (OR^2-ℓ^0) is invariant to the scale of $\boldsymbol{b}$ and points from $\mathcal{Y}$, i.e., (OR^2-ℓ^0) concerns $\boldsymbol{b}$ and $\boldsymbol{y}_i$'s only in *directions* but not *scaling*. As such, for an arbitrary dataset we can project[2] its points onto $\mathbb{S}^{D-1}$ to form a qualified $\mathcal{Y}$; for an *arbitrary* smooth manifold $\overline{\mathcal{M}}$, taking $\boldsymbol{\Pi}_{\mathbb{S}^{D-1}}\left(\overline{\mathcal{M}} \backslash \{\boldsymbol{0}\}\right)$ gives a $\mathcal{M}$ obeying Assumption 1.

Remark 2. Assumption 2 is not a strong assumption, since for outliers from a uniform distribution over $\mathbb{S}^{D-1}$ it holds with probability 1.

Notations. We denote the objective function of (OR^2-ℓ^1) as $f(\cdot)$, i.e., $f(\boldsymbol{b}) = \left\|\mathcal{Y}^\top \boldsymbol{b}\right\|_1$. For a set $\mathcal{S} \subset \mathbb{R}^D$ and a vector $\boldsymbol{x} \in \mathbb{R}^D$, $\boldsymbol{\Pi}_{\mathcal{S}}(\boldsymbol{x}) := \operatorname{argmin}_{\boldsymbol{y}}\{\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 : \boldsymbol{y} \in \mathcal{S}\}$ is the projection of $\boldsymbol{x}$ onto $\mathcal{S}$ provided it exists. For any $\boldsymbol{v} \in \mathbb{R}^D$, $\langle \boldsymbol{v} \rangle$ denotes the linear subspace spanned by $\boldsymbol{v}$. For any linear subspace $\mathcal{S}$ of $\mathbb{R}^D$, $\mathcal{S}^\perp$ is its orthogonal complement. Let $\operatorname{sign}(\cdot)$ be such that $\operatorname{sign}(x)$ is 1 if $x > 0$, -1 if $x < 0$ and 0 otherwise. Denote by $T_{\boldsymbol{b}}\mathcal{M}$ the tangent space of $\mathcal{M}$ at $\boldsymbol{b} \in \mathcal{M}$. Given $\boldsymbol{b} \in \mathcal{M}$ we denote by $\operatorname{retr}_{\boldsymbol{b}} : T_{\boldsymbol{b}}\mathcal{M} \to \mathcal{M}$ the retraction map that sends a vector $\boldsymbol{v}$ in the tangent space $T_{\boldsymbol{b}}\mathcal{M}$ back to the manifold $\mathcal{M}$. We use rgrad for the Riemannian gradient.

[2] A minor detail here is that there is not a unique projection of the zero vector $\mathbf{0}$ onto $\mathbb{S}^{D-1}$, a point that we can remove from $\mathcal{Y}$ regardless without affecting minimizers of (OR^2-ℓ^0).

Algorithm 1: RSGM for (OR^2-ℓ^1)

```
Initialization $\boldsymbol{b}_0 \in \mathcal{M}$,   step sizes: $\{\alpha_k > 0\}_{k=1,2,\ldots}$
For $k \leftarrow 1, 2, \ldots, K$:
    $\boldsymbol{g}_k \in \partial f(\boldsymbol{b}_k)$    (6)
    $\boldsymbol{r}_k \leftarrow \boldsymbol{\Pi}_{T_{\boldsymbol{b}_k}\mathcal{M}}\, \boldsymbol{g}_k$    (7)
    $\boldsymbol{b}_{k+1} \leftarrow \operatorname{retr}_{\boldsymbol{b}_k}(-\alpha_k \boldsymbol{r}_k)$    (8)
```

Algorithm 2: IRLS for (OR^2-ℓ^1)

```
Initialization $\boldsymbol{b}_0 \in \mathcal{M}$,   parameter $\delta > 0$
For $k \leftarrow 0, \ldots, K$:
    $w_j^{k+1} \leftarrow \frac{1}{\max\{|\boldsymbol{y}_j^\top \boldsymbol{b}_k|, \delta\}}, \quad \forall j = 1, \ldots, L$    (9)
    $\boldsymbol{b}_{k+1} \in \operatorname*{argmin}_{\boldsymbol{b} \in \mathcal{M}} \sum_{j=1}^{L} w_j^{k+1} \cdot \left(\boldsymbol{y}_j^\top \boldsymbol{b}\right)^2$    (10)
```

3.2 Algorithms

To solve (OR^2-ℓ^1), we describe two algorithms: a Riemannian subgradient method when a retraction map of $\mathcal{M}$ and a projection map onto $T_{\boldsymbol{b}}\mathcal{M}$ are accessible, or otherwise an iteratively reweighted least squares method when a weighted least squares problem over $\mathcal{M}$ can be easily solved.

Riemannian Sub-gradient Method (RSGM). Algorithm 1 gives the classic RSGM algorithm for (OR^2-ℓ^1), which iteratively computes a subgradient (6), projects it onto the tangent space (7), and takes a retraction step (8). To give a concrete example, suppose for the moment that $\mathcal{M}$ is the unit sphere. In this case the computation of the subgradient in step (6) can be contextualized as $\boldsymbol{g}_k \leftarrow \sum_{j=1}^{L} \operatorname{sign}(\boldsymbol{y}_j^\top \boldsymbol{b}_k)\boldsymbol{y}_j$, the projection step (7) is given as $\boldsymbol{r}_k \leftarrow \boldsymbol{g}_k - \boldsymbol{b}_k \boldsymbol{b}_k^\top \boldsymbol{g}_k$, and the retraction step (8) can be implemented simply as a projection onto the sphere, $\boldsymbol{b}_{k+1} \leftarrow (\boldsymbol{b}_k - \alpha_{k+1}\boldsymbol{r}_k)/\|\boldsymbol{b}_k - \alpha_{k+1}\boldsymbol{r}_k\|_2$. In general, the subgradient step (6) can be computed for typical objective functions and the two Riemannian optimization steps (7) and (8) can be computed efficiently for a wide range of manifolds $\mathcal{M}$'s. Therefore, we assume that Algorithm 1 can be implemented efficiently, and will analyze its convergence rate in §4.3 and performance in applications in §5.

Iteratively Reweighted Least Squares (IRLS). Algorithm 2 presents the classic IRLS framework to solve (OR^2-ℓ^1). IRLS alternates between updating the weights (9) and solving a weighted least-squares problem (10). The choice of weight w_j^{k+1} in (9) is a classic one and has been used in prior works (Tsakiris & Vidal, 2018a; Ding et al., 2019b; 2020a; Iwata et al., 2020; Kümmerle et al., 2021; Peng et al., 2022b). While we have not seen the appearance of (10) in its exact form in the literature, several papers have considered (10) with particualr $\mathcal{M}$'s for the problems of robust subspace recovery (Tsakiris & Vidal, 2018a; Ding et al., 2019b; 2020a) and essential matrix estimation (Zhao et al., 2020; Zhao, 2022). In the former case, (10) is a weighted eigenvalue problem, easily solvable; in the latter case, (10) is a QCQP, which can be solved via its SDP relaxations. Since (10) does not involve binary decision variables (different from Lajoie et al. (2018); Yang & Carlone (2019; 2022); Peng et al. (2022a)), the SDP solver of Zhao et al. (2020); Zhao (2022) is actually efficient. Given these, we implicitly assume that (10) is efficiently solvable, and later we will analyze the convergence rate of IRLS in generality and evaluate its performance in action.

4 Theoretical Analysis

In this section, we first characterize the geometry of inliers, outliers, and their interplay with $\mathcal{M}$ via some quantities (§4.1). They turn out to be useful in the analysis of the minimizers (§4.2) of and algorithms (§4.3) for (OR^2-ℓ^1).

4.1 Geometric Quantities

We will define quantities that depend on the inliers $\mathcal{X}$, outliers $\mathcal{O}$, the true vector $\boldsymbol{b}^*$ (equivalently, the true hyperplane $\mathcal{H}$), and the manifold $\mathcal{M}$. It should be clear that these quantities are for theoretical understanding and not for running the algorithms. To begin with, define

$$c_{\text{in}} := \inf \left\{ \frac{1}{N} \sum_{j=1}^{N} |\boldsymbol{x}_j^\top \boldsymbol{s}| : \boldsymbol{b} \in \mathcal{M},\ \boldsymbol{b} \neq \boldsymbol{b}^*,\ \boldsymbol{s} = \boldsymbol{\Pi}_{\mathbb{S}^{D-1}}(\boldsymbol{\Pi}_{\mathcal{H}}(\boldsymbol{b})) \right\}. \tag{11}$$

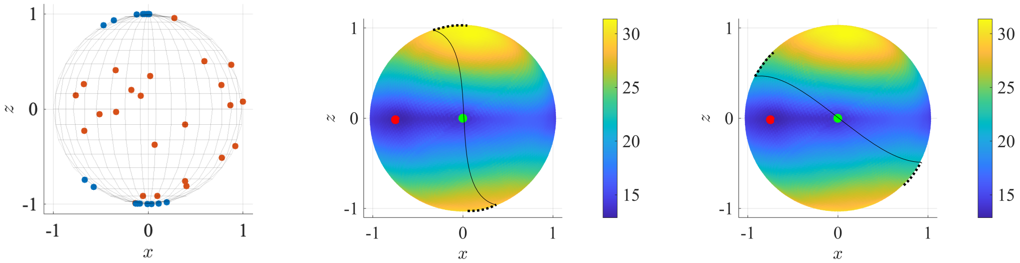


Figure 2: The non-uniform data from Figure 1 (left), and $f(\cdot)$ over $\mathbb{S}^2$ using the data overlaid by two $\mathcal{M}$'s (dark solid lines) yielding large c_{in} (middle) and small c_{in} (right) respectively. Dotted lines are locations of $\boldsymbol{s}$'s in c_{in} (11). For clarity, the viewpoints are set at the true $\boldsymbol{b}^*$ facing towards the origin.

Remark 3. To understand c_{in}, note that one can expect it to be large when for any $\boldsymbol{b} \in \mathcal{M}, \boldsymbol{b} \neq \boldsymbol{b}^*$, its projection $\boldsymbol{s}$ onto $\mathcal{H}$ is *not* orthogonal to most inliers $\boldsymbol{x}_j$'s. This is visualized in Figure 2, where c_{in} is large as the dotted lines are far from the equator $z = 0$ (middle), and vice versa (right). To better appreciate the role of $\mathcal{M}$ in c_{in}, we make a contrast with a similar quantity in Zhu et al. (2018) for guaranteeing (1):

$$\bar{c}_{\text{in}} = \inf \left\{ \frac{1}{N} \sum_{j=1}^{N} |\boldsymbol{x}_j^\top \boldsymbol{s}| : \boldsymbol{s} \in \mathcal{H} \cap \mathbb{S}^{D-1} \right\}. \tag{12}$$

Observe that $\bar{c}_{\text{in}}$ is large only when *all directions* in $\mathcal{H}$ are not orthogonal to most inliers, essentially requiring inliers to be uniformly distributed in $\mathcal{H} \cap \mathbb{S}^{D-1}$. Such an assumption can fail in applications as illustrated in Example 1. In contrast, one can have large c_{in} albeit some vectors in $\mathcal{H}$ are orthogonal to a vast number of inliers, as long as these directions are ruled out by $\mathcal{M}$.

Example 1. In geometric vision applications, inliers tend to lie close to a *proper* subspace of the true hyperplane. For instance, in *essential matrix* or *trifocal tensor estimation*, data resulting from 3D points on an affine plane provably lies in a subspace of lower dimension (Chum et al., 2005; Ding et al., 2020b); this is more generally known as critical surfaces (Hartley & Zisserman, 2003, §22).

With the above said, we can further define similar quantities for outliers:

$$c_{\text{out,min}} := \min_{\boldsymbol{b} \in \mathbb{S}^{D-1}} \frac{1}{M} \sum_{j=1}^{M} |\boldsymbol{o}_j^\top \boldsymbol{b}|, \quad c_{\text{out,max}} := \max_{\boldsymbol{b} \in \mathbb{S}^{D-1}} \frac{1}{M} \sum_{j=1}^{M} |\boldsymbol{o}_j^\top \boldsymbol{b}|. \tag{13}$$

Remark 4. It is now not hard to see that $c_{\text{out,min}}$ and $c_{\text{out,max}}$ are close if the outliers are uniformly distributed in $\mathbb{S}^{D-1}$. While in principle one can consider the $\mathcal{M}$-constrained version of these outlier-related quantities, we find doing so complicates the picture without shedding useful insights.

Finally, we define

$$\gamma_{\text{in},\theta} := \max \left\{ \frac{1}{N} \left\| \boldsymbol{\Pi}_{T_{\boldsymbol{b}}\mathcal{M}^\perp \cap \langle b \rangle^\perp} \sum_{j=1}^{N} \operatorname{sgn}(\boldsymbol{x}_j^\top \boldsymbol{b}) \boldsymbol{x}_j \right\|_2 : \boldsymbol{b} \in \mathcal{B}_{\mathcal{M}}(\boldsymbol{b}^*, \theta) \right\}, \tag{14}$$

$$\bar{\eta}_{\text{out}} := \max_{\boldsymbol{b} \in \mathbb{S}^{D-1}} \frac{1}{M} \left(\left\| \boldsymbol{\Pi}_{T_{\boldsymbol{b}}\mathcal{M}} \sum_{j=1}^{M} \operatorname{sign}(\boldsymbol{o}_j^\top \boldsymbol{b}) \boldsymbol{o}_j \right\|_2 + D \right), \tag{15}$$

where $\mathcal{B}_{\mathcal{M}}(\boldsymbol{b}^*, \theta)$ is a ball of angular radius θ centered at $\boldsymbol{b}^*$ over $\mathcal{M}$

$$\mathcal{B}_{\mathcal{M}}(\boldsymbol{b}^*, \theta) := \{\boldsymbol{b} \in \mathcal{M} : \angle(\boldsymbol{b}, \boldsymbol{b}^*) \leq \theta\}. \tag{16}$$

Remark 5. Roughly speaking, $T_{\boldsymbol{b}}\mathcal{M}^\perp \cap \langle b \rangle^\perp$ contains directions *tangent* to $\mathbb{S}^{D-1}$ and *normal* to $\mathcal{M}$ at $\boldsymbol{b}$, as seen by the arrows in Figure 3. As such, $\gamma_{\text{in},\theta}$ is small if within $\mathcal{B}_{\mathcal{M}}(\boldsymbol{b}^*, \theta)$, all such directions are orthogonal to most inliers $\boldsymbol{x}_j$'s; this is, e.g., the case for the left and middle plots in Figure 3 with $\theta = \frac{\pi}{2}$ and $\theta = \frac{\pi}{3}$ respectively, but *not* the case for the right plot with $\theta = \frac{\pi}{3}$. Observe also that $\gamma_{\text{in},\theta}$ is increasing in θ. Last but not least, $\bar{\eta}_{\text{out}}$ is small if outliers are uniformly distributed in $\mathbb{S}^{D-1}$.

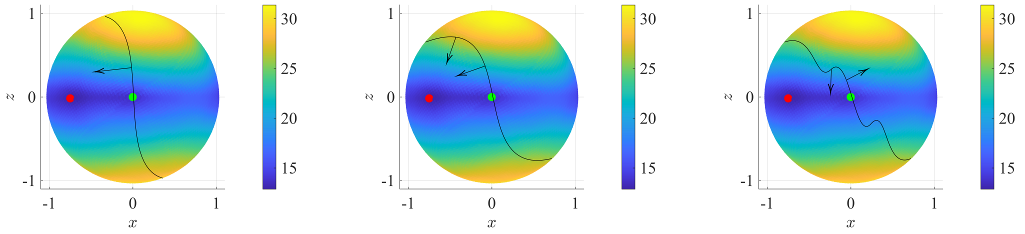


Figure 3: $f(\cdot)$ over $\mathbb{S}^2$ using the non-uniform data from Figure 1 overlaid by three $\mathcal{M}$'s (dark solid lines) yielding $\gamma_{\text{in},\theta}$ globally small (left), locally small (middle) and not small (right) respectively. Arrows are examples of vectors from $T_{\boldsymbol{b}}\mathcal{M}^\perp \cap \langle \boldsymbol{b} \rangle^\perp$ in $\gamma_{\text{in},\theta}$ (14).

4.2 Guarantees on Global Minimizers

Using the quantities developed above, we are now ready to localize global minimizers of (OR^2-ℓ^1). To begin with, when $\gamma_{\text{in},\theta}$ is small for all $\theta \in (0, \frac{\pi}{2}]$ (e.g., Figure 3 left), we have the theorem below:

Theorem 1. *Let* $\gamma_{\text{in}} := \sup_{\theta \in (0,\frac{\pi}{2}]} \gamma_{\text{in},\theta}$. *Every global minimizer of* (OR^2-ℓ^1) *must be* $\pm\boldsymbol{b}^*$ *if*

$$\frac{\frac{M^2}{N^2} \cdot \left[\left(c_{\text{out,max}} - c_{\text{out,min}} \right)^2 + \overline{\eta}_{\text{out}}^2 \right] + \gamma_{\text{in}}^2}{c_{\text{in}}^2} < 1. \tag{17}$$

Intuitively, Theorem 1 says the minimizers of (OR^2-ℓ^1) recover the ground truth $\pm\boldsymbol{b}^*$ as long as $N \gg M$, inliers are well distributed *with respect to* $\mathcal{M}$ (c_{in} large), outliers are well distributed ($c_{\text{out,max}} - c_{\text{out,min}}$ and $\overline{\eta}_{\text{out}}$ small), and $\mathcal{M}$ does not curve too much against the inliers (γ_{in} small).

On the other hand, $\gamma_{\text{in},\theta}$ could be small only for small θ and increases drastically for large θ (e.g., Figure 3 middle). In this case, we provide the following guarantee:

Theorem 2. *Every global minimizer of* (OR^2-ℓ^1) *must be* $\pm\boldsymbol{b}^*$ *if*

$$\sin(\theta) > \frac{M}{N} \cdot \frac{c_{\text{out,max}} - c_{\text{out,min}}}{c_{\text{in}}}, \tag{18}$$

$$N^2 c_{\text{in}}^2 - N^2 \gamma_{\text{in},0}^2 \geq M^2 \overline{\eta}_{\text{out}}^2, \tag{19}$$

where θ *is the unique solution to*

$$\cos^2(\theta) N^2 c_{\text{in}}^2 - N^2 \gamma_{\text{in},\theta}^2 = M^2 \overline{\eta}_{\text{out}}^2. \tag{20}$$

Note that the LHS of (20) is strictly decreasing in θ, and non-positive with $\theta = \frac{\pi}{2}$; the RHS of (20) is non-negative. Thus, with condition (19) in place, (20) is guaranteed to have a unique solution. Theorem 2 is tighter than Theorem 1: with some inspection one can see that the former entails the latter; if for some $\theta = \overline{\theta}$ (18) holds and in (20) the LHS is greater than or equal to the RHS, then the minimizers of (OR^2-ℓ^1) are $\pm\boldsymbol{b}^*$ *regardless of* the behavior of $\gamma_{\text{in},\theta}$ for $\theta > \overline{\theta}$.

Finally, if $\gamma_{\text{in},\theta}$ is large even for small θ (e.g., Figure 3 right), we can only characterize the approximate location of the minimizer with the following theorem that does not depend on $\gamma_{\text{in},\theta}$.

Theorem 3. *Every global minimizer of* (OR^2-ℓ^1) *has an angle away from* $\pm\boldsymbol{b}^*$ *of at most*

$$\arcsin\left(\frac{M}{N} \cdot \frac{c_{\text{out,max}} - c_{\text{out,min}}}{c_{\text{in}}} \right). \tag{21}$$

4.3 Convergence Rates of RSGM and IRLS

We now consider convergence analysis of the proposed algorithms. We begin with Algorithm 1:

Theorem 4. *With a suitable initialization and geometrically diminishing step size, the sequence* $\{\boldsymbol{b}_k\}_{k=0}^K$ *given by Algorithm 1 converges at a linear rate to the true vector* $\pm\boldsymbol{b}^*$. *More precisely, with the step size* $\alpha_k = \beta^k \cdot \alpha_0$, *the principal angle* θ_k *between* $\mathbf{b}_k$ *and* $\mathbf{b}^*$ *is upper bounded as per*

$$\sin(\theta_k/2) \leq \beta^k \sin(\theta_0/2) \tag{22}$$

if the following conditions are satisfied:

(i) Initialization not too far from $\pm\boldsymbol{b}^*$*:* $\sin(\theta_0/2) < \frac{s}{2c_2}$,

(ii) Small initial step size: $0 < \alpha_0 < \min\left(\frac{\sin(\theta_0/2)}{s - 2c_2\sin(\theta_0/2)}, \frac{4(s-2c_2\sin(\theta_0/2))}{(c_1+1)\kappa^2/\sin(\theta_0/2)}\right)$,

(iii) Suitable step size decay: $1 > \beta \geq \sqrt{1 + \alpha_0\left(\alpha_0\frac{(c_1+1)\kappa^2}{4\sin(\theta_0/2)^2} + 2c_2 - \frac{s}{\sin(\theta_0/2)}\right)}$.

In the above, $s := \frac{1}{2}Nc_{\text{in}} - Mc_{\text{out,max}}$, κ *is a Lipschitz constant of the objective* $f(\cdot)$, $c_2 = \kappa \cdot c_3$, *and* c_1 *and* c_3 *are constants that depend only on* $\mathcal{M}$.

Assuming condition (i) holds, one can verify that conditions (ii) and (iii) of Theorem 4 also holds true if the initial step size α_0 is small enough as $\beta \to 1$. Therefore, with α_0 and β suitably chosen, Algorithm 1 converges linearly if initialized in a neighborhood of $\boldsymbol{b}^*$ of size s/c_2 (Theorem 4).

We then proceed to Algorithm 2, an IRLS method of an entirely different style than Algorithm 1. The theoretical convergence rate guarantee, stated below, is also in a different style than Theorem 4:

Theorem 5. *The sequence* $\{\boldsymbol{b}_k\}_{k=0}^K$ *produced by Algorithm 2 converges at a sublinear rate to a critical point of the following problem:*

$$\min_{\boldsymbol{b}\in\mathcal{M}} \sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}), \qquad h(r) := \begin{cases} |r| & |r| > \delta \\ \frac{1}{2}\left(r^2/\delta + \delta\right) & |r| \leq \delta \end{cases} \tag{OR^2-Huber}$$

More precisely, the Riemannian gradient of the objective of (OR^2-Huber) *is bounded above:*

$$\min_{k=0,\dots,K} \left\| \operatorname{rgrad} \sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}_k) \right\|_2 \leq \sqrt{\frac{2\sigma \cdot \sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}_0)}{K+1}} \tag{23}$$

In (23), σ *is a finite positive constant depending on the manifold* $\mathcal{M}$*, data* $\boldsymbol{y}_j$*'s, and* δ.

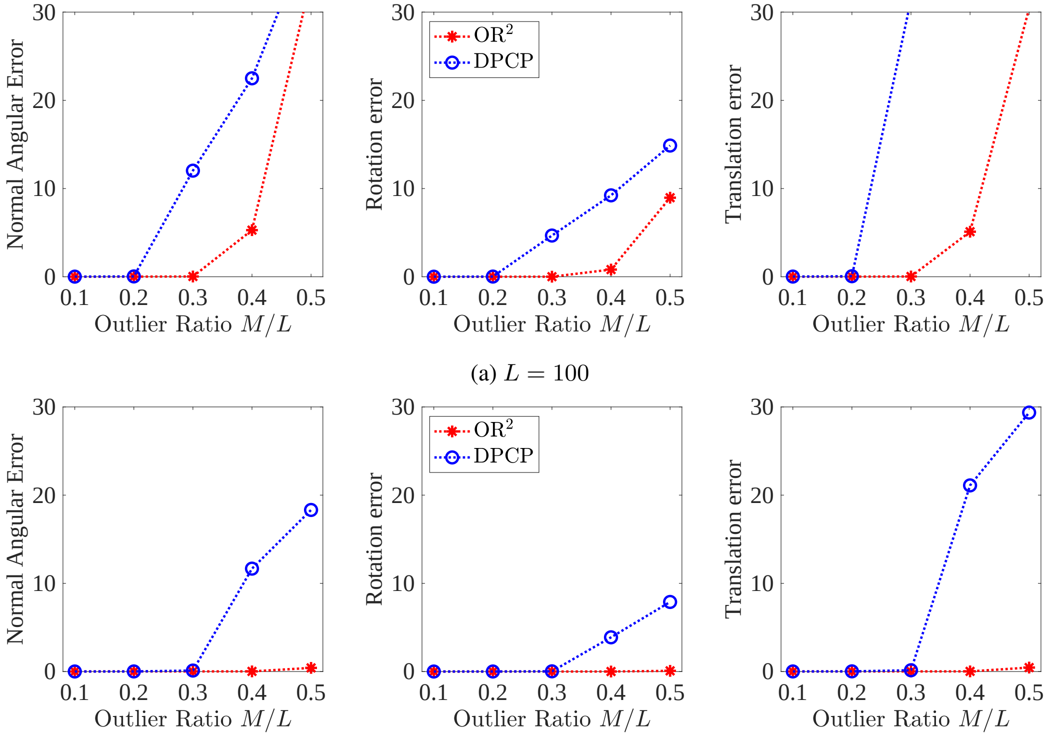
Algorithm 2 converges to a critical point of a different Huber-tyle objective in (OR^2-Huber) than the ℓ^1 objective of (OR^2-ℓ^1). The key is that the (rescaled and translated) Huber loss $h(\cdot)$ in (OR^2-Huber) tends to the ℓ^1 loss as $\delta \to 0$, and (OR^2-Huber) can be thought of as a smoothed version of (OR^2-ℓ^1), easier to minimize. Theorem 5 reveals Algorithm 2 performs such smoothing implicitly with a sublinear convergence rate. Our proof of Theorem 5 extends the idea of Hong et al. (2017), which is specifically for the convex case, into the case with non-convex manifolds $\mathcal{M}$.

5 Experiments: Robust Essential Matrix Estimation

Recall that our major claim is that enforcing manifold constraints would lead to theoretically more meaningful and practically more robust estimation than relaxing, if not ignoring, the manifold constraints. While previous sections focus on providing theoretical insights, here we are interested in consolidating our claim using numerical experiments. In particular, this section presents experimental results on robust essential matrix estimation, while in the interest of space we refer the reader to the Appendix for more experiments and details.

In order to validate our claim, we consider the following setup. We follow Kneip & Furgale (2014) to generate random rotations $\boldsymbol{R}^* \in \mathbb{SO}(3)$, translations $\boldsymbol{t}^* \in \mathbb{R}^3$, and L correspondences, of which N are inliers and M are outliers. The correspondences are used to produce $\mathcal{Y} \in \mathbb{R}^{9\times L}$ following (4) and a normalization of columns. We test the IRLS algorithm for solving (1) and (OR^2-ℓ^1). For (1), the weighted least-squares step (10) can be implemented via SVD; the resulting IRLS algorithm was proposed in Tsakiris & Vidal (2018b) and here we call it DPCP[3] (*dual principal component pursuit*). For (OR^2-ℓ^1), with $\mathcal{M}$ being the essential manifold $\mathcal{M}_E$, the weighted least-squares step (10) can be implemented via an efficient SDP solver (Zhao, 2022); here we call it OR^2. Each method is run on $\mathcal{Y}$ to obtain an estimate $\hat{\boldsymbol{b}}$, while we need to project the estimate of DPCP onto $\mathcal{M}_E$ as it does not enforce the essential manifold constraint. We compute the angular error between $\hat{\boldsymbol{b}}$ and the ground truth $\boldsymbol{b}^*$. We further decompose each estimated essential matrix into rotations and translations, and calculate angular rotation and translation error with $\boldsymbol{R}^*, \boldsymbol{t}^*$.

[3] We refer the reader to Tsakiris & Vidal (2018b); Ding et al. (2020b); Zhao (2022), where DPCP or IRLS really stood out with excellent performance in comparison to other competitive methods including RANSAC.



(a) $L = 100$

(b) $L = 200$

Figure 4: Robust essential matrix estimation over 50 randomly generated scene data.

In Figure 4 we report the median errors for varying outlier ratio $M/L \in \{0.1 : 0.1 : 0.5\}$ and total number of points $L \in \{100, 200\}$. As expected, the two methods, DPCP and OR^2, give almost 0 error with a small (e.g., $\leq 20\%$) outlier ratio. Notably, also can be seen is that (OR^2-ℓ^1) is much more robust than (1): when $L = 100$, (OR^2-ℓ^1) has a translation error of 5.2° with 40% outliers, while (1) yields that of more than 30° with only 30% outliers, a sharp contrast. Moreover, with more points ($L = 200$), (OR^2-ℓ^1) achieves less than 1° for any error metric even with 50% outliers, while in that case (1) has, e.g., 7.8° rotation error already. The experiments therefore confirm our claim that directly tackling the manifold constraints (e.g., essential manifolds) yield visible computational benefits than relaxing the constraints (e.g., into a unit sphere).

6 Conclusion and Discussion

In this paper, we presented a meaningful generalization of the seminal work of Späth & Watson (1987) on ℓ^1 orthogonal regression (on the sphere) to outlier-robust orthogonal regression on smooth manifolds, thereby bringing together robustness with the proper manifold constraints that are inherent in geometric vision (e.g., essential manifolds). Our experiments suggested that taking manifold constraints into consideration does improve the performance of the estimation algorithms. Even though handling manifold constraints directly is arguably more challenging, our core contribution consists of performing theoretical analysis for the proposed approach and derived several results regarding the optimization landscape and convergence of associated algorithms; all these are ultimately useful guarantees for outlier-robust orthogonal regression on manifolds.

Our analysis is not without limitations. First of all, to obtain conditions for global minimizers (§4.2), we defined several quantities in §4.1 that are not directly computable. However, for future work it is possible to derive probabilistic bounds for them, thereby obtaining more informative conditions. The second, related, question to consider is on specializing our conditions for manifolds of interest (e.g., essential manifolds). Finally, we focused on orthogonal regression in the paper which is associated with the ℓ^1 (and Huber) loss. It is an important future direction to generalize our analysis for other common loss functions (e.g., Logcosh, Geman-McClure, truncated quadratic).

References

Yu Bai, Qijia Jiang, and Ju Sun. Subgradient descent learns orthogonal dictionaries. October 2018. 1

Daniel Barath and Jiří Matas. Graph-cut RANSAC. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6733–6741, 2018. 3

Daniel Barath, Maksym Ivashechkin, and Jiri Matas. Progressive NAPSAC: sampling from gradually growing neighborhoods. June 2019a. 3

Daniel Barath, Jiri Matas, and Jana Noskova. MAGSAC: Marginalizing sample consensus. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019b. 3

Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiri Matas. MAGSAC++, a fast, reliable and accurate robust estimator. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020. 3

Nicolas Boumal, Bamdev Mishra, P-A Absil, and Rodolphe Sepulchre. Manopt, a matlab toolbox for optimization on manifolds. August 2013. 15

Nicolas Boumal, Pierre-Antoine Absil, and Coralia Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019. 20, 22

O Chum and J Matas. Matching with PROSAC — progressive sample consensus. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE, 2005. 3

Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized RANSAC. In *Pattern Recognition*, pp. 236–243. Springer Berlin Heidelberg, 2003. 3

Ondřej Chum, Tomáš Werner, and Jiří Matas. Two-view geometry estimation unaffected by a dominant plane. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, volume I, pp. 772–779, 2005. 3, 6

Tianjiao Ding, Yunchen Yang, Zhihui Zhu, Daniel P Robinson, René Vidal, Laurent Kneip, and Manolis C. Tsakiris. Robust homography estimation via dual principal component pursuit. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020a. 5

Tianjiao Ding, Yunchen Yang, Zhihui Zhu, Daniel P Robinson, René Vidal, Laurent Kneip, and Manolis C Tsakiris. Robust homography estimation via dual principal component pursuit. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6080–6089, 2020b. 2, 6, 8, 15

Tianyu Ding, Zhihui Zhu, Tianjiao Ding, Yunchen Yang, Daniel P Robinson, Manolis C Tsakiris, and René Vidal. Noisy dual principal component pursuit. In *International Conference on Machine Learning*, pp. 1617–1625, 2019a. 1

Tianyu Ding, Zhihui Zhu, Tianjiao Ding, Yunchen Yang, René Vidal, Manolis C. Tsakiris, and Daniel Robinson. Noisy dual principal component pursuit. In *International Conference on Machine Learning*, 2019b. 5

Tianyu Ding, Zhihui Zhu, Manolis Tsakiris, René Vidal, and Daniel Robinson. Dual principal component pursuit for learning a union of hyperplanes: Theory and algorithms. In *Artificial Intelligence and Statistics*, 2021. 1

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24: 381–395, 1981. 3

Dar Gilboa, Sam Buchanan, and John Wright. Efficient dictionary learning with gradient descent. September 2018. 1

Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003. 6

Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. 2

Mingyi Hong, Xiangfeng Wang, Meisam Razaviyayn, and Zhi-Quan Luo. Iteration complexity analysis of block coordinate descent methods. *Mathematical Programming*, 163:85–114, 2017. 8

Maksym Ivashechkin, Daniel Barath, and Jiri Matas. USACv20: robust essential, fundamental and homography matrix estimation. *arXiv*. 3

Maksym Ivashechkin, Daniel Barath, and Jiri Matas. VSAC: Efficient and accurate estimator for H and F. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2021. 3

Daichi Iwata, Michael Waechter, Wen-Yan Lin, and Yasuyuki Matsushita. An analysis of sketched irls for accelerated sparse residual regression. In *European Conference on Computer Vision*, 2020. 5

Laurent Kneip and Paul Furgale. OpenGV: A unified and generalized approach to real-time calibrated geometric vision. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2014. 8

Christian Kümmerle, Claudio Mayrink Verdun, and Dominik Stöger. Iteratively reweighted least squares for basis pursuit with global linear convergence rate. *Advances in Neural Information Processing Systems*, 2021. 5

Pierre-Yves Lajoie, Siyi Hu, Giovanni Beltrame, and Luca Carlone. Modeling perceptual aliasing in SLAM via Discrete-Continuous graphical models. October 2018. 3, 5

Karel Lebeda, Jirí Matas, and Ondrej Chum. Fixing the locally optimized RANSAC. In *Procedings of the British Machine Vision Conference 2012*, number 95. British Machine Vision Association, 2012. 3

Xiao Li, Shixiang Chen, Zengde Deng, Qing Qu, Zhihui Zhu, and Anthony Man Cho So. Weakly convex optimization over stiefel manifold using riemannian Subgradient-Type methods. November 2019. 4, 20

Yi Ma, Stefano Soatto, Jana Košecká, and S Shankar Sastry. *An Invitation to 3-D Vision*. Springer New York, 2003. 2

Tyler Maunu and Gilad Lerman. Robust subspace recovery with adversarial outliers. April 2019. 3

Lionel Moisan, Pierre Moulon, and Pascal Monasse. Fundamental matrix of a stereo pair, with a contrario elimination of outliers. *Image Processing On Line*, 5:89–113, 2016. 3

David Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):756–770, 2004. 3

Liangzu Peng, Mahyar Fazlyab, and René Vidal. Semidefinite relaxations of truncated Least-Squares in robust rotation search: Tight or not. In *Computer Vision – ECCV 2022*, pp. 673–691. Springer Nature Switzerland, 2022a. 3, 5, 15

Liangzu Peng, Christian Kümmerle, and René Vidal. Global linear and local superlinear convergence of IRLS for non-smooth robust regression. In *Advances in Neural Information Processing Systems*, 2022b. 5

Liangzu Peng, Manolis C. Tsakiris, and René Vidal. ARCS: Accurate rotation and correspondences search. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022c. 15

Qing Qu, Ju Sun, and John Wright. *Finding a sparse vector in a subspace: Linear sparsity using alternating directions*, pp. 3401–3409. 2014. 1

Qing Qu, Xiao Li, and Zhihui Zhu. A nonconvex approach for exact and efficient multichannel sparse blind deconvolution. August 2019. 1

Qing Qu, Zhihui Zhu, Xiao Li, Manolis C Tsakiris, John Wright, and René Vidal. Finding the sparsest vectors in a subspace: Theory, algorithms, and applications. January 2020. 1

Rahul Raguram, Ondrej Chum, Marc Pollefeys, Jiri Matas, and Jan Michael Frahm. USAC: A universal framework for random sample consensus. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35 (8):2022–2038, 2013. 3

H Spath and G A Watson. On orthogonal linear l1 approximation. *Numerische Mathematik*, 51: 531–543, 1987. 1

H Späth and GA Watson. On orthogonal linear ℓ_1 approximation. *Numerische Mathematik*, 51: 531–543, 1987. 9

C V Stewart. MINPRAN: a new robust estimator for computer vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(10):925–938, October 1995. 3

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Trans. Inf. Theory*, 63(2):853–884, 2017a. 1

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere II: Recovery by riemannian Trust-Region method. *IEEE Trans. Inf. Theory*, 63(2):885–914, 2017b. 1

P H S Torr and A Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Comput. Vis. Image Underst.*, 78(1):138–156, April 2000. 3

Philip Hilaire Torr, Slawomir J Nasuto, and John Mark Bishop. Napsac: High noise, high dimensional robust estimation-it's in the bag. In *British Machine Vision Conference (BMVC)*, 2002. 3

Manolis C Tsakiris and Rene Vidal. Dual principal component pursuit. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. IEEE, December 2015. 1

Manolis C Tsakiris and René Vidal. Hyperplane clustering via dual principal component pursuit. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3472–3481. PMLR, 2017. 1

Manolis C. Tsakiris and René Vidal. Dual principal component pursuit. *Journal of Machine Learning Research*, 19(18):1–50, 2018a. 5

Manolis C Tsakiris and Rene Vidal. Dual principal component pursuit. *Journal of Machine Learning Research*, 19:1–49, 2018b. 1, 8

Heng Yang and Luca Carlone. A quaternion-based certifiably optimal solution to the wahba problem with outliers. May 2019. 3, 5, 15, 16

Heng Yang and Luca Carlone. Certifiably optimal Outlier-Robust geometric perception: Semidefinite relaxations and scalable global optimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP, May 2022. 3, 5

Yunchen Yang, Xinyue Zhang, Tianjiao Ding, Daniel P Robinson, Rene Vidal, and Manolis C Tsakiris. Boosting RANSAC via dual principal component pursuit. October 2021. 3

Ji Zhao. An efficient solution to Non-Minimal case essential matrix estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(4):1777–1792, April 2022. 4, 5, 8

Ji Zhao, Wanting Xu, and Laurent Kneip. A certifiably globally optimal solution to generalized essential matrix estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020. 5

Zhihui Zhu, Yifan Wang, Daniel Robinson, Daniel Naiman, Rene Vidal, and Manolis Tsakiris. Dual principal component pursuit: Improved analysis and efficient algorithms. In *Advances in Neural Information Processing Systems*, 2018. 1, 4, 6, 15, 19

Zhihui Zhu, Tianyu Ding, Daniel Robinson, Manolis Tsakiris, and René Vidal. A linearly convergent method for non-smooth non-convex optimization on the grassmannian with applications to robust subspace and dictionary learning. In *Adv. Neural Inf. Process. Syst.*, volume 32. proceedings.neurips.cc, 2019. 4

A More on Experiments

A.1 Data Generation for Essential Matrix Estimation (§5)

We first generate the ground truth rotation $\boldsymbol{R}^*$ and translation $\boldsymbol{t}^*$. $\boldsymbol{t}^*$ is sampled uniformly at random from $[-2,2]^3$. $\boldsymbol{R}^*$ is generated corresponding to rotating about the x-axis for θ_x radians, the y-axis for θ_y radians, and the z-axis for θ_z radians; $\theta_x, \theta_y, \theta_z$ are sampled independently and uniformly at random from $[-0.5, 0.5]$. A total of L points in $\mathbb{R}^3$ are randomly sampled using the following procedure:

Algorithm 3: 3D Point Cloud Generation

```
Number of points: $L$
$d_{\min} \leftarrow 4, d_{\max} \leftarrow 8$
For $j \leftarrow 1, \dots, L$:
    $\boldsymbol{p}_j \sim \operatorname{unif}([-1,1]^3)$    (24)
    $\boldsymbol{d}_j \leftarrow \boldsymbol{\Pi}_{\mathbb{R}^2}(\boldsymbol{p}_j)$    (25)
    $\boldsymbol{q}_j \leftarrow (d_{\max} - d_{\min})\boldsymbol{p}_j + d_{\min}\boldsymbol{d}_j$    (26)
return Generated 3D points $\boldsymbol{Q}$
```

Roughly speaking, each point is randomly sampled to have a distance $4-8$ around the origin. These points in $\mathbb{R}^3$ are projected onto the cameras defined by $[\boldsymbol{I}, \boldsymbol{0}]$ and $[\boldsymbol{R}^*, \boldsymbol{t}^*]$ to give L correspondences. We take M of them and randomly permute the correspondence relations to produce outliers. Then, as said in §5, the correspondences are used to produce $\mathcal{Y} \in \mathbb{R}^{9\times L}$ following (4) and a projection onto $\mathbb{S}^8$ for normalization.

A.2 Robust Rotation Estimation

Problem. Given a point cloud in $\mathbb{R}^3$ and its copy that undergoes a rotation $\boldsymbol{R}^* \in \mathbb{SO}(3)$, how can we recover $\boldsymbol{R}^*$? Again, one can first run some matching algorithm on the point clouds to obtain correspondences $\{(\boldsymbol{z}_j, \boldsymbol{z}'_j)\}_{j=1}^L \subset \mathbb{R}^3 \times \mathbb{R}^3$. Inlier correspondences $(\boldsymbol{z}_j, \boldsymbol{z}'_j)$ now satisfy

$$\boldsymbol{z}'_j = \boldsymbol{R}^* \boldsymbol{z}_j, \tag{27}$$

while outlier correspondences exist due to false matches. Our focus is to robustly estimate $\boldsymbol{R}^*$ given $\{(\boldsymbol{z}_j, \boldsymbol{z}'_j)\}_{j=1}^L$. Being able to estimate $\boldsymbol{R}^*$ accurately is crucial to a more challenging task of *point cloud registration*, where inlier correspondences $(\boldsymbol{z}_j, \boldsymbol{z}'_j)$ further differ by a translation $\boldsymbol{t}^*$, that is $\boldsymbol{z}'_j = \boldsymbol{R}^* \boldsymbol{z}_j + \boldsymbol{t}^*$.

OR^2 over Rotation Matrices. To see that this problem can be posed as an instance of OR^2, observe that (27) implies

$$[\boldsymbol{z}'_j]_\times \boldsymbol{R}^* \boldsymbol{z}_j = \boldsymbol{0} \in \mathbb{R}^3, \tag{28}$$

by noting $\boldsymbol{a} \times \boldsymbol{a} = \boldsymbol{0}$ for any $\boldsymbol{a} \in \mathbb{R}^3$ and [1]. These are three equations linear in $\boldsymbol{R}^*$, only two of which are linearly independent. Thus one can write (28) as

$$\boldsymbol{Y}_j^\top \boldsymbol{b}^* = \boldsymbol{0} \in \mathbb{R}^2, \tag{29}$$

where $\boldsymbol{Y}_j \in \mathbb{R}^{9\times 2}$ is some embedding of $(\boldsymbol{z}_j, \boldsymbol{z}'_j)$, and $\boldsymbol{b}^* = \operatorname{vec}(\boldsymbol{R}^*)/\sqrt{3}$ lies in the manifold

$$\mathcal{M}_R := \{\boldsymbol{b} \in \mathbb{R}^9 : \quad \boldsymbol{b} = \operatorname{vec}(\boldsymbol{R})/\sqrt{3}, \quad \boldsymbol{R} \in \mathbb{SO}(3)\}, \tag{30}$$

Since $\boldsymbol{Y}_j$ is an inlier when *both* of its columns are orthogonal to $\boldsymbol{b}^*$, we consider a natural extention[4] of (OR^2-ℓ^1)

$$\min_{\boldsymbol{b}} \sum_{j=1}^{L} \left\| \boldsymbol{Y}_j^\top \boldsymbol{b} \right\|_2 \quad \text{s.t.} \quad \boldsymbol{b} \in \mathcal{M} = \mathcal{M}_R. \tag{OR^2-ℓ^1-group}$$

[4] Indeed, it is easy to verify that (OR^2-ℓ^1-group) reduces to (OR^2-ℓ^1) if each $\boldsymbol{Y}_j$ only has one column.

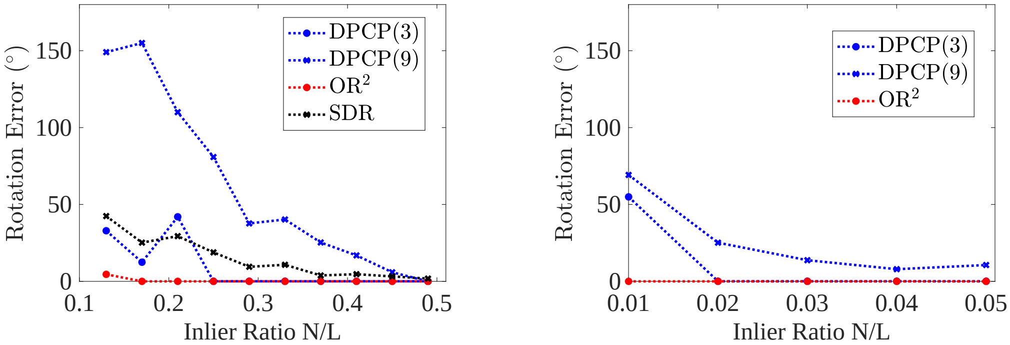


Figure 5: Robust rotation search with varying number of inliers and total number of points $L = 100$ (left) and $L = 10000$ (right). The SDR method fail to run in the latter case due to insufficient memory.

To normalize the data, each column of $\boldsymbol{Y}_j$ is scaled to have a unit ℓ^2 norm. We consider an RSGM algorithm for (OR^2-ℓ^1-group), in which the projection map onto $T_{\boldsymbol{b}}\mathcal{M}$ and the retraction map are implemented by the manopt toolbox (Boumal et al., 2013).

Relaxation to $\mathbb{S}^8$. Similar to the case of robustly estimating an essential matrix (§1), since $\mathcal{M}_R \subset \mathbb{S}^8$ one can relax (OR^2-ℓ^1-group) to

$$\min_{\boldsymbol{b}} \quad \sum_{j=1}^{L} \left\| \boldsymbol{Y}_j^\top \boldsymbol{b} \right\|_2 \quad \text{s.t.} \quad \boldsymbol{b} \in \mathcal{M} = \mathbb{S}^8. \tag{31}$$

This special instance has been considered in the related problem of homography estimation (Ding et al., 2020b). We use an RSGM algorithm for solving (31), and refer to such a method as DPCP(9) (Zhu et al., 2018). The number 9 represents the coordinate dimension of $\boldsymbol{b}^*$. As one might expect, we also have a baseline that operates in a different dimension.

Relaxation to $\mathbb{S}^2$. Let $\overline{\boldsymbol{b}} \in \mathbb{S}^2$ be the rotation axis of $\boldsymbol{R}^*$. Applying dot product on (27) with $\overline{\boldsymbol{b}}$ yields

$$\langle \overline{\boldsymbol{b}}, \boldsymbol{z}_j' \rangle = \langle \overline{\boldsymbol{b}}, \boldsymbol{R}^* \boldsymbol{z}_j \rangle = \langle \overline{\boldsymbol{b}}, \boldsymbol{z}_j \rangle \tag{32}$$

$$\Leftrightarrow \langle \overline{\boldsymbol{b}}, \boldsymbol{z}_j' - \boldsymbol{z}_j \rangle = 0. \tag{33}$$

Therefore, we can estimate $\overline{\boldsymbol{b}}$ by solving (1) with $\boldsymbol{y}_j$ taken as $(\boldsymbol{z}_j' - \boldsymbol{z}_j)/\left\|\boldsymbol{z}_j' - \boldsymbol{z}_j\right\|_2$ and $D = 3$. Once we have a solution, say $\hat{\boldsymbol{b}}$, we identify a set of inliers by thresholding the error $|\langle \hat{\boldsymbol{b}}, \boldsymbol{y}_j \rangle|$ by 10^{-3}. Given the estimated rotation axis $\hat{\boldsymbol{b}}$ and estimated inliers, one can compute the rotation angle via interval stabbing as in Peng et al. (2022c), after which a rotation estimation can be formed by composing the estimated axis and angle. We call such an approach DPCP(3) in the experiments.

Semidefinite Programming. Finally, we compare with a truncated least squares surrogate of (OR^2-ℓ^0) with $\mathcal{M} = \mathcal{M}_R$, which is solved via a *Shor* semidefinite relaxation (Yang & Carlone, 2019; Peng et al., 2022a); see also §2. In our experiments this is referred to as SDR.

Data Generation. To generate the ground-truth rotation $\boldsymbol{R}^*$, we sample uniformly at random an axis from $\mathbb{S}^2$ and an angle from $[0, 2\pi]$, and form $\boldsymbol{R}^*$ by composing the axis and angle. We sample a point cloud of N points from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_3)$, and rotate them by $\boldsymbol{R}^*$ to give the inlier correspondences. To generate outlier correspondences, we sample *two* point clouds of M points from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_3)$, and scale the points from the second point cloud such that $\|\boldsymbol{z}_j'\|_2 = \|\boldsymbol{z}_j\|_2$; otherwise the outliers can be easily removed by comparing the norms.

Result. Figure 5 reports the median rotation error in degrees over 10 random seeds with varying number of points and outlier ratio. Firstly, in all cases OR^2 has a lower error than DPCP (the red curve stays below all other curves), as also observed in estimating essential matrix §5. Notably, when $L = 100$, OR^2 is more accurate than SDR, the latter being a state-of-the-art approach (Yang & Carlone, 2019). Moving on to the regime of large $L = 10,000$, the SDR method fails due to the prohibitive memory required to handle $O(L^2)$ variables. In contrast, OR^2 continue to give the most accurate estimate: even with 1% inliers it gives close to $0°$ median error.

B Proofs

B.1 Proofs of Minimizers

We begin this section by giving a lemma on the critical points which stands as a building block in proving Theorems 1 and 2.

Lemma 6. *Any critical point of* (OR^2-ℓ^1) *is either* $\pm\boldsymbol{b}^*$ *or has an angle of at least* θ *away from* $\pm\boldsymbol{b}^*$*, where* θ *is the solution to*

$$\cos^2(\theta)N^2c_{\text{in}}^2 - N^2\gamma_{\text{in},\theta}^2 = M^2\overline{\eta}_{\text{out}}^2. \tag{34}$$

Proof. Let $\boldsymbol{b}$ be a critical point of (OR^2-ℓ^1). We are done if it is $\pm\boldsymbol{b}^*$, so suppose the otherwise. Note that $\boldsymbol{b}$ by definition satisfies the following subgradient condition

$$\boldsymbol{0} = (\boldsymbol{I} - \boldsymbol{b}\boldsymbol{b}^\top - \boldsymbol{Q}\boldsymbol{Q}^\top)\left(\sum_{j=1}^{N}\operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j + \sum_{j=1}^{M}\operatorname{sgn}(\boldsymbol{o}_j^\top\boldsymbol{b})\boldsymbol{o}_j\right), \tag{35}$$

where $\boldsymbol{I} - \boldsymbol{b}\boldsymbol{b}^\top - \boldsymbol{Q}\boldsymbol{Q}^\top$ is the orthogonal projection onto $T_{\boldsymbol{b}}\mathcal{M}$ such that $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}$ and $\boldsymbol{Q}^\top\boldsymbol{b} = \boldsymbol{0}$. We will now bound the 'size' of the subgradients for inliers and outliers respectively.

Lower bounding the size of subgradient for inliers. Define θ to be the angle between $\boldsymbol{b}$ and $\boldsymbol{b}^*$. Since $\boldsymbol{b} \subset \mathcal{M} \subset \mathbb{S}^{D-1}$ and $b \not\perp \mathcal{H}$, one write $\boldsymbol{b}$ as

$$\boldsymbol{b} = \sin(\theta)\boldsymbol{s} + \cos(\theta)\boldsymbol{n}, \tag{36}$$

where $\boldsymbol{s} := \boldsymbol{\Pi}_{\mathcal{H}}(\boldsymbol{b})/\|\,\boldsymbol{\Pi}_{\mathcal{H}}(\boldsymbol{b})\|_2$, $\boldsymbol{n} := \boldsymbol{\Pi}_{\mathcal{H}^\perp}(\boldsymbol{b})/\|\,\boldsymbol{\Pi}_{\mathcal{H}^\perp}(\boldsymbol{b})\|_2$, and . Now,

$$
\begin{aligned}
&\left\|(\boldsymbol{I}-\boldsymbol{b}\boldsymbol{b}^\top-\boldsymbol{Q}\boldsymbol{Q}^\top)\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j\right\|_2^2 && (37)\\
&\overset{(36)}{=}\left\|(\boldsymbol{I}-\boldsymbol{Q}\boldsymbol{Q}^\top)\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j-\sin(\theta)\boldsymbol{b}\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{s})(\boldsymbol{x}_j^\top\boldsymbol{s})\right\|_2^2 && (38)\\
&\overset{\|\boldsymbol{b}\|=1}{=}\left\|(\boldsymbol{I}-\boldsymbol{Q}\boldsymbol{Q}^\top)\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j\right\|_2^2+\sin^2(\theta)\left\|\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{s})(\boldsymbol{x}_j^\top\boldsymbol{s})\right\|_2^2\\
&\qquad-2\left\langle(\boldsymbol{I}-\boldsymbol{Q}\boldsymbol{Q}^\top)\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j,\boldsymbol{b}\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{s})(\boldsymbol{x}_j^\top\boldsymbol{s})\right\rangle && (39)\\
&\overset{\boldsymbol{Q}^\top\boldsymbol{b}=\boldsymbol{0}}{=}\left\|(\boldsymbol{I}-\boldsymbol{Q}\boldsymbol{Q}^\top)\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j\right\|_2^2+\sin^2(\theta)\left\|\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{s})(\boldsymbol{x}_j^\top\boldsymbol{s})\right\|_2^2\\
&\qquad-2\left\langle\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j,\boldsymbol{b}\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{s})(\boldsymbol{x}_j^\top\boldsymbol{s})\right\rangle && (40)\\
&\overset{(36)}{=}\left\|(\boldsymbol{I}-\boldsymbol{Q}\boldsymbol{Q}^\top)\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j\right\|_2^2+\sin^2(\theta)\left\|\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{s})(\boldsymbol{x}_j^\top\boldsymbol{s})\right\|_2^2\\
&\qquad-2\sin^2(\theta)\left\|\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{s})(\boldsymbol{x}_j^\top\boldsymbol{s})\right\|_2^2 && (41)\\
&=\left\|(\boldsymbol{I}-\boldsymbol{Q}\boldsymbol{Q}^\top)\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j\right\|_2^2-\sin^2(\theta)\left(\sum_{j=1}^N |\boldsymbol{x}_j^\top\boldsymbol{s}|\right)^2 && (42)
\end{aligned}
$$

Note that the above can be further written as

$$
\left\|\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j\right\|_2^2-\left\|\boldsymbol{Q}\boldsymbol{Q}^\top\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j\right\|_2^2-\sin^2(\theta)\left(\sum_{j=1}^N |\boldsymbol{x}_j^\top\boldsymbol{s}|\right)^2 \tag{43}
$$

Now, the first term can be lower bounded by

$$
\begin{aligned}
\left\|\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j\right\|_2^2 &\overset{\|\boldsymbol{s}\|=1}{=}\left\|\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j\right\|_2^2\|\boldsymbol{s}\|_2^2 && (44)\\
&\overset{\text{Cauchy-Schwartz}}{\geq}\left(\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j^\top\boldsymbol{s}\right)^2 && (45)\\
&\overset{(36)}{=}\left(\sum_{j=1}^N \operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{s})\boldsymbol{x}_j^\top\boldsymbol{s}\right)^2=\left(\sum_{j=1}^N |\boldsymbol{x}_j^\top\boldsymbol{s}|\right)^2. && (46)
\end{aligned}
$$

Substituting this back to (43), we have

$$\left\|(\boldsymbol{I}-\boldsymbol{b}\boldsymbol{b}^\top-\boldsymbol{Q}\boldsymbol{Q}^\top)\sum_{j=1}^{N}\operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j\right\|_2^2 \tag{47}$$

$$\geq \cos^2(\theta)\left(\sum_{j=1}^{N}|\boldsymbol{x}_j^\top\boldsymbol{s}|\right)^2 - \left\|\boldsymbol{Q}\boldsymbol{Q}^\top\sum_{j=1}^{N}\operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j\right\|_2^2. \tag{48}$$

This leads to final lower bound

$$\left\|(\boldsymbol{I}-\boldsymbol{b}\boldsymbol{b}^\top-\boldsymbol{Q}\boldsymbol{Q}^\top)\sum_{j=1}^{N}\operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j\right\|_2^2 \tag{49}$$

$$\geq \cos^2(\theta)N^2c_{\text{in}}^2 - N^2\gamma_{\text{in},\theta}^2 \tag{50}$$

Note that the lower bound is a decreasing function of θ.

Upper bounding the size of subgradient for outliers. From Assumption 2, we have that $\boldsymbol{b}$ is orthogonal to at most $D-1$ outliers. Therefore, we have

$$(\boldsymbol{I}-\boldsymbol{b}\boldsymbol{b}^\top-\boldsymbol{Q}\boldsymbol{Q}^\top)\left(\sum_{j=1}^{M}\operatorname{sgn}(\boldsymbol{o}_j^\top\boldsymbol{b})\boldsymbol{o}_j\right) \tag{51}$$

$$= (\boldsymbol{I}-\boldsymbol{b}\boldsymbol{b}^\top-\boldsymbol{Q}\boldsymbol{Q}^\top)\left(\sum_{j=1}^{M}\operatorname{sign}(\boldsymbol{o}_j^\top\boldsymbol{b})\boldsymbol{o}_j + \sum_{j=1}^{D-1}\lambda_j\boldsymbol{o}_{l_j}\right). \tag{52}$$

Here sign is the usual sign function, $\{\boldsymbol{o}_{l_j}\}_{j=1}^{D-1}$ are $D-1$ outliers and $\lambda_j \in [-1,1]$ for $j=1,\dots,D-1$. Note further that

$$\left\|(\boldsymbol{I}-\boldsymbol{b}\boldsymbol{b}^\top-\boldsymbol{Q}\boldsymbol{Q}^\top)\sum_{j=1}^{D-1}\lambda_j\boldsymbol{o}_{l_j}\right\|_2 \leq \left\|\sum_{j=1}^{D-1}\lambda_j\boldsymbol{o}_{l_j}\right\|_2 \leq \sum_{j=1}^{D-1}\left\|\lambda_j\boldsymbol{o}_{l_j}\right\|_2 \leq D. \tag{53}$$

The last inequality makes use of the fact that $\boldsymbol{o}_j \in \mathbb{S}^{D-1}$. Therefore, we have

$$\left\|(\boldsymbol{I}-\boldsymbol{b}\boldsymbol{b}^\top-\boldsymbol{Q}\boldsymbol{Q}^\top)\left(\sum_{j=1}^{M}\operatorname{sign}(\boldsymbol{o}_j^\top\boldsymbol{b})\boldsymbol{o}_j + \sum_{j=1}^{D-1}\lambda_j\boldsymbol{o}_{l_j}\right)\right\|_2 \tag{54}$$

$$\overset{\text{Triangle Ineq.}}{\leq} \left\|(\boldsymbol{I}-\boldsymbol{b}\boldsymbol{b}^\top-\boldsymbol{Q}\boldsymbol{Q}^\top)\left(\sum_{j=1}^{M}\operatorname{sign}(\boldsymbol{o}_j^\top\boldsymbol{b})\boldsymbol{o}_j\right)\right\|_2 + D, \tag{55}$$

$$\leq M\overline{\eta}_{\text{out}}. \tag{56}$$

Putting everything together. From (35) it is necessary that the size of subgradients for inliers and outliers must be equal

$$\left\|(\boldsymbol{I}-\boldsymbol{b}\boldsymbol{b}^\top-\boldsymbol{Q}\boldsymbol{Q}^\top)\left(\sum_{j=1}^{N}\operatorname{sgn}(\boldsymbol{x}_j^\top\boldsymbol{b})\boldsymbol{x}_j\right)\right\| = \left\|(\boldsymbol{I}-\boldsymbol{b}\boldsymbol{b}^\top-\boldsymbol{Q}\boldsymbol{Q}^\top)\left(\sum_{j=1}^{M}\operatorname{sgn}(\boldsymbol{o}_j^\top\boldsymbol{b})\boldsymbol{o}_j\right)\right\|. \tag{57}$$

This together with (50) and (56) yields

$$\cos^2(\theta)N^2c_{\text{in}}^2 - N^2\gamma_{\text{in},\theta}^2 \leq M^2\overline{\eta}_{\text{out}}^2. \tag{58}$$

Since $\cos^2(\theta)N^2c_{\text{in}}^2 - N^2\gamma_{\text{in},\theta}^2 - M^2\overline{\eta}_{\text{out}}^2$ is strictly decreasing in θ, and (58) holds for $\theta=\frac{\pi}{2}$, there exists a unique $\theta_1 \geq 0$ such that (58) is equivalent to $\theta \geq \theta_1$. □

Theorem 3. *Every global minimizer of* (OR^2-ℓ^1) *has an angle away from* $\pm \boldsymbol{b}^*$ *of at most*

$$\arcsin\left(\frac{M}{N} \cdot \frac{c_{\text{out,max}} - c_{\text{out,min}}}{c_{\text{in}}}\right). \tag{21}$$

Proof. Suppose $\boldsymbol{b}$ is a global minimizer. If $\boldsymbol{b} = \pm\boldsymbol{b}^*$ then we are done, so suppose the otherwise. We can again decompose $\boldsymbol{b}$ as in (36). Now, on the one hand,

$$\begin{aligned} f(\boldsymbol{b}) &= \sum_{j=1}^{N} |\boldsymbol{x}_j^\top \boldsymbol{b}| + \sum_{j=1}^{M} |\boldsymbol{o}_j^\top \boldsymbol{b}| && (59)\\ &= \sin(\theta) \sum_{j=1}^{N} |\boldsymbol{x}_j^\top \boldsymbol{s}| + \sum_{j=1}^{M} |\boldsymbol{o}_j^\top \boldsymbol{b}| && (60)\\ &\geq \sin(\theta) N c_{\text{in}} + M c_{\text{out,min}}. && (61) \end{aligned}$$

On the other hand,

$$f(\boldsymbol{b}) \leq f(\boldsymbol{b}^*) = \sum_{j=1}^{M} |\boldsymbol{o}_j^\top \boldsymbol{b}^*| \leq M c_{\text{out,max}}. \tag{62}$$

Combining the above, it must hold that

$$M c_{\text{out,max}} \geq \sin(\theta) N c_{\text{in}} + M c_{\text{out,min}}, \tag{63}$$

and we conclude the proof. □

Theorem 2 then follows from Lemma 6 and Theorem 3, and Theorem 1 is a result of Theorem 2. We conclude the section with the following remark. While the high-level proof strategy is motivated by Zhu et al. (2018), the problem we study (OR^2-ℓ^1) handles *general* manifolds $\mathcal{M}$'s satisfying Assumption 1, while Zhu et al. (2018) considers only $\mathcal{M} = \mathbb{S}^{D-1}$. Further, our analysis induces novel quantities c_{in}, $\gamma_{\text{in},\theta}$ that are critical to tolerating non-uniform inliers, while Zhu et al. (2018) requires uniform inliers, as explained in §4.1.

B.2 Convergence Proof of PSGM

It turns out to be useful to define for any $\boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{S}^{D-1}$

$$\operatorname{dist}(\boldsymbol{b}_1, \boldsymbol{b}_2) := \min_{\tau \in \{\pm 1\}} \|\boldsymbol{b}_1 - \tau \boldsymbol{b}_2\|_2 = \sqrt{2 - 2|\langle \boldsymbol{b}_1, \boldsymbol{b}_2 \rangle|} = 2\sin(\angle(\boldsymbol{b}_1, \boldsymbol{b}_2)/2). \tag{64}$$

Lemma 7 (Sharpness of (OR^2-ℓ^1))**.** *The objective $f(\cdot)$ of* (OR^2-ℓ^1) *satisfies*

$$f(\boldsymbol{b}) - f(\boldsymbol{b}^*) \geq s \operatorname{dist}(\boldsymbol{b}, \boldsymbol{b}^*), \quad \forall \boldsymbol{b} \in \mathcal{M}, \tag{65}$$

where $s := \frac{1}{2} N c_{\text{in}} - M c_{\text{out,max}}$.

Proof. It is clear that when $\boldsymbol{b} = \pm\boldsymbol{b}^*$, (65) holds for any $s > 0$ so we are done. Suppose $\boldsymbol{b} \neq \pm\boldsymbol{b}^*$. Again taking a decomposition of $\boldsymbol{b}$ as in (36), we have

$$\begin{aligned} f(\boldsymbol{b}) - f(\boldsymbol{b}^*) &= \sin(\theta) \sum_{j=1}^{N} |\boldsymbol{x}_j^\top \boldsymbol{s}| + \sum_{j=1}^{M} |\boldsymbol{o}_j^\top \boldsymbol{b}| - \sum_{j=1}^{M} |\boldsymbol{o}_j^\top \boldsymbol{b}^*| && (66)\\ &\geq \sin(\theta) N c_{\text{in}} + \sum_{j=1}^{M} \big(|\boldsymbol{o}_j^\top \boldsymbol{b}| - |\boldsymbol{o}_j^\top \boldsymbol{b}^*|\big). && (67) \end{aligned}$$

To lower bound the first term, we observe that

$$\operatorname{dist}(\boldsymbol{b}, \boldsymbol{b}^*) = 2\sin(\theta/2) \leq 2\sin(\theta), \tag{68}$$

where the equality follows from (64), and the inequality is due to $\sin(\cdot)$ is increasing on $[0, \pi/2]$. Now we lower bound the last two terms in (67). For any $\tau \in \{\pm 1\}$, the following holds

$$\begin{aligned}
\sum_{j=1}^{M} \big(|\boldsymbol{o}_j^\top \boldsymbol{b}| - |\boldsymbol{o}_j^\top \boldsymbol{b}^*|\big) &= \sum_{j=1}^{M} \big(|\boldsymbol{o}_j^\top \tau \boldsymbol{b}| - |\boldsymbol{o}_j^\top \boldsymbol{b}^*|\big) && (69)\\
&= \sum_{j=1}^{M} \big(|\boldsymbol{o}_j^\top (\tau \boldsymbol{b} - \boldsymbol{b}^* + \boldsymbol{b}^*)| - |\boldsymbol{o}_j^\top \boldsymbol{b}^*|\big) && (70)\\
&\geq -\sum_{j=1}^{M} |\boldsymbol{o}_j^\top (\tau \boldsymbol{b} - \boldsymbol{b}^*)| && (71)\\
&= -\|\tau \boldsymbol{b} - \boldsymbol{b}^*\|_2 \sum_{j=1}^{M} \left| \boldsymbol{o}_j^\top \frac{\tau \boldsymbol{b} - \boldsymbol{b}^*}{\|\tau \boldsymbol{b} - \boldsymbol{b}^*\|_2} \right| && (72)\\
&\geq -\|\tau \boldsymbol{b} - \boldsymbol{b}^*\|_2 M c_{\text{out,max}} && (73)
\end{aligned}$$

where (71) is from the triangle inequality. Therefore, it is true in particular that

$$\sum_{j=1}^{M} \big(|\boldsymbol{o}_j^\top \boldsymbol{b}| - |\boldsymbol{o}_j^\top \boldsymbol{b}^*|\big) \geq -\operatorname{dist}(\boldsymbol{b}, \boldsymbol{b}^*) M c_{\text{out,max}}. \tag{74}$$

Combining (67), (68) and (74) yields

$$f(\boldsymbol{b}) - f(\boldsymbol{b}^*) \geq \left(\frac{1}{2} N c_{\text{in}} - M c_{\text{out,max}}\right) \operatorname{dist}(\boldsymbol{b}, \boldsymbol{b}^*). \tag{75}$$

□

Lemma 8 (Distance Lemma)**.** *Let $\{\mathbf{b}_k\}_{k=0}^{K}$ be the sequence generated by Algorithm 1. Let κ be a Lipschitz constant for $f(\cdot)$ restricted to $\mathcal{M}$. Then, it holds for any $\mathbf{b} \in \mathcal{M}$ that*

$$\|\boldsymbol{b}_k - \boldsymbol{b}\|_2^2 \leq \|\boldsymbol{b}_{k-1} - \boldsymbol{b}\|_2^2 + (c_1 + 1)\alpha_{k-1}^2 \kappa^2 + 2\alpha_{k-1}(f(\boldsymbol{b}) - f(\boldsymbol{b}_{k-1})) + 2\alpha_{k-1} c_2 \|\boldsymbol{b}_{k-1} - \boldsymbol{b}\|_2^2, \tag{76}$$

where $c_2 = \kappa \cdot c_3$, and c_1, c_3 are positive constants that depend only on $\mathcal{M}$.

Proof. Since $f(\cdot)$ is convex, it follows from (Li et al., 2019, Corollary 1) that

$$f(\boldsymbol{b}) - f(\boldsymbol{b}_{k-1}) + c_2 \|\boldsymbol{b}_{k-1} - \boldsymbol{b}\|_2^2 \geq \langle \boldsymbol{b} - \boldsymbol{b}_{k-1}, \boldsymbol{r}_{k-1} \rangle. \tag{77}$$

Therefore, we have

$$\begin{aligned}
\|\boldsymbol{b}_k - \boldsymbol{b}\|_2^2 &= \big\|\operatorname{retr}_{\boldsymbol{b}_{k-1}}(-\alpha_{k-1}\boldsymbol{r}_{k-1}) - \boldsymbol{b}\big\|_2^2 && (78)\\
&\leq \|\boldsymbol{b}_{k-1} - \alpha_{k-1}\boldsymbol{r}_{k-1} - \boldsymbol{b}\|_2^2 + c_1 \|\alpha_{k-1}\boldsymbol{r}_{k-1}\|_2^2 && (79)\\
&= \|\boldsymbol{b}_{k-1} - \boldsymbol{b}\|_2^2 + (c_1 + 1)\|\alpha_{k-1}\boldsymbol{r}_{k-1}\|_2^2 + 2\langle \boldsymbol{b} - \boldsymbol{b}_{k-1}, \alpha_{k-1}\boldsymbol{r}_{k-1} \rangle && (80)\\
&\leq \|\boldsymbol{b}_{k-1} - \boldsymbol{b}\|_2^2 + (c_1 + 1)\alpha_{k-1}^2 \kappa^2 + 2\alpha_{k-1}\langle \boldsymbol{b} - \boldsymbol{b}_{k-1}, \boldsymbol{r}_{k-1} \rangle && (81)\\
&\leq \|\boldsymbol{b}_{k-1} - \boldsymbol{b}\|_2^2 + (c_1 + 1)\alpha_{k-1}^2 \kappa^2 + 2\alpha_{k-1}(f(\boldsymbol{b}) - f(\boldsymbol{b}_{k-1})) + 2\alpha_{k-1} c_2 \|\boldsymbol{b}_{k-1} - \boldsymbol{b}\|_2^2,
\end{aligned}$$

where the first inequality follows from Assumption 1 and (Boumal et al., 2019, Appendix B) , c_1 is a constant that depends only on $\mathcal{M}$; the second inequality follows from the Lipschitzness of $f(\cdot)$, and the last inequality follows from the argument above. □

Theorem 4. *With a suitable initialization and geometrically diminishing step size, the sequence $\{\mathbf{b}_k\}_{k=0}^{K}$ given by Algorithm 1 converges at a linear rate to the true vector $\pm\mathbf{b}^*$. More precisely, with the step size $\alpha_k = \beta^k \cdot \alpha_0$, the principal angle θ_k between $\mathbf{b}_k$ and $\mathbf{b}^*$ is upper bounded as per*

$$\sin(\theta_k/2) \leq \beta^k \sin(\theta_0/2) \tag{22}$$

if the following conditions are satisfied:

(i) *Initialization not too far from* $\pm \boldsymbol{b}^*$*:* $\sin(\theta_0/2) < \frac{s}{2c_2}$,

(ii) *Small initial step size:* $0 < \alpha_0 < \min\left(\frac{\sin(\theta_0/2)}{s-2c_2\sin(\theta_0/2)}, \frac{4(s-2c_2\sin(\theta_0/2))}{(c_1+1)\kappa^2/\sin(\theta_0/2)}\right)$,

(iii) *Suitable step size decay:* $1 > \beta \geq \sqrt{1+\alpha_0\left(\alpha_0\frac{(c_1+1)\kappa^2}{4\sin(\theta_0/2)^2}+2c_2-\frac{s}{\sin(\theta_0/2)}\right)}$.

In the above, $s := \frac{1}{2}Nc_{\text{in}} - Mc_{\text{out,max}}$*,* κ *is a Lipschitz constant of the objective* $f(\cdot)$*,* $c_2 = \kappa \cdot c_3$*, and* c_1 *and* c_3 *are constants that depend only on* $\mathcal{M}$*.*

Proof. Due to (64), it suffices to prove that

$$\operatorname{dist}(\boldsymbol{b}_k, \boldsymbol{b}^*) \leq \beta^k \operatorname{dist}(\boldsymbol{b}_0, \boldsymbol{b}^*). \tag{82}$$

We prove the theorem by induction on k. When $k = 0$, (22) holds trivially. Suppose $k > 0$, and let $\tau^* := \operatorname{argmin}_{\tau\in\{\pm1\}}\|\boldsymbol{b}_{k-1} - \tau\boldsymbol{b}^*\|_2^2$. Then,

$$\begin{aligned}
\operatorname{dist}^2(\boldsymbol{b}_k, \boldsymbol{b}^*) &= \min_{\tau\in\{\pm1\}}\|\boldsymbol{b}_k - \tau\boldsymbol{b}^*\|_2^2 \leq \|\boldsymbol{b}_k - \tau^*\boldsymbol{b}^*\|_2^2 && (83)\\
&\leq (1+2\alpha_{k-1}c_2)\|\boldsymbol{b}_{k-1} - \tau^*\boldsymbol{b}^*\|_2^2 + (c_1+1)\alpha_{k-1}^2\kappa^2 + 2\alpha_{k-1}(f(\boldsymbol{b}^*) - f(\boldsymbol{b}_{k-1})) && (84)\\
&= (1+2\alpha_{k-1}c_2)\operatorname{dist}^2(\boldsymbol{b}_{k-1}, \boldsymbol{b}^*) + (c_1+1)\alpha_{k-1}^2\kappa^2 + 2\alpha_{k-1}(f(\boldsymbol{b}^*) - f(\boldsymbol{b}_{k-1})) && (85)\\
&\leq (1+2\alpha_0c_2)\operatorname{dist}^2(\boldsymbol{b}_{k-1}, \boldsymbol{b}^*) - 2\alpha_{k-1}s\operatorname{dist}(\boldsymbol{b}_{k-1}, \boldsymbol{b}^*) + (c_1+1)\alpha_{k-1}^2\kappa^2. && (86)
\end{aligned}$$

To upper bound it Consider the quadratic function $l(t) := (1+2\alpha_0c_2)t^2 - 2\alpha_{k-1}st + (c_1+1)\alpha_{k-1}^2\kappa^2$. It has an axis of symmetry given by the vertical line $t = t_0$, where $t_0 := \frac{\alpha_{k-1}s}{1+2\alpha_0c_2} > 0$; so $l(0) = l(2t_0)$. On the other hand, since $\alpha_0 < \frac{\sin(\theta_0/2)}{s-2c_2\sin(\theta_0/2)}$ and $s - 2c_2\sin(\theta_0/2) > 0$ due to conditions (i) and (ii), we have

$$\alpha_0(s - 2c_2\sin(\theta_0/2)) < \sin(\theta_0/2) \tag{87}$$

which by (64) is equivalent to

$$\begin{aligned}
&2\alpha_0(s - c_2\operatorname{dist}(\boldsymbol{b}_0, \boldsymbol{b}^*)) < \operatorname{dist}(\boldsymbol{b}_0, \boldsymbol{b}^*) && (88)\\
\Leftrightarrow\ &\frac{2\alpha_0 s}{1+2\alpha_0c_2} < \operatorname{dist}(\boldsymbol{b}_0, \boldsymbol{b}^*) && (89)\\
\Leftrightarrow\ &\frac{2\alpha_0\beta^{k-1}s}{1+2\alpha_0c_2} < \beta^{k-1}\operatorname{dist}(\boldsymbol{b}_0, \boldsymbol{b}^*). && (90)
\end{aligned}$$

This is to say that $2t_0 < \beta^{k-1}\operatorname{dist}(\boldsymbol{b}_0, \boldsymbol{b}^*)$. Since the induction hypothesis is

$$\operatorname{dist}(\boldsymbol{b}_{k-1}, \boldsymbol{b}^*) \leq \beta^{k-1}\operatorname{dist}(\boldsymbol{b}_0, \boldsymbol{b}^*), \tag{91}$$

we have $l(\operatorname{dist}(\boldsymbol{b}_{k-1}, \boldsymbol{b}^*)) < l(\beta^{k-1}\operatorname{dist}(\boldsymbol{b}_0, \boldsymbol{b}^*))$. Combining this with (86) yields

$$\begin{aligned}
\operatorname{dist}^2(\boldsymbol{b}_k, \boldsymbol{b}^*) &\leq (1+2\alpha_0c_2)\beta^{2k-2}\operatorname{dist}^2(\boldsymbol{b}_0, \boldsymbol{b}^*) - 2\beta^{2k-2}\alpha_0 s\operatorname{dist}(\boldsymbol{b}_0, \boldsymbol{b}^*) + (c_1+1)\beta^{2k-2}\alpha_0^2\kappa^2\\
&= \beta^{2k-2}\operatorname{dist}^2(\boldsymbol{b}_0, \boldsymbol{b}^*)\left(1 + 2\alpha_0c_2 - 2\frac{\alpha_0 s}{\operatorname{dist}(\boldsymbol{b}_0, \boldsymbol{b}^*)} + \frac{(c_1+1)\alpha_0^2\kappa^2}{\operatorname{dist}^2(\boldsymbol{b}_0, \boldsymbol{b}^*)}\right) && (92)\\
&\leq \beta^{2k}\operatorname{dist}^2(\boldsymbol{b}_0, \boldsymbol{b}^*), && (93)
\end{aligned}$$

where the last inequality follows from condition (iii). This completes the induction step. Finally, note that with condition (i) in place, (ii) is not vacuous since both terms in $\min$ are positive; thus (iii) is also not vacuous, since (ii) implies $\alpha_0\frac{(c_1+1)\kappa^2}{4\sin(\theta_0/2)^2} + 2c_2 - \frac{s}{\sin(\theta_0/2)} < 0$. □

B.3 Convergence Proof of IRLS

Lemma 9. *The weighted eigenvalue subproblem* (10) *in Algorithm* 2 *is equivalent to*

$$\boldsymbol{b}_{k+1} \in \operatorname*{argmin}_{\boldsymbol{b}\in\mathcal{M}} \sum_{j=1}^{L} g\big(\boldsymbol{y}_j^\top\boldsymbol{b}, \boldsymbol{y}_j^\top\boldsymbol{b}_k\big), \qquad g(r,v) := h(v) + \frac{r^2 - v^2}{2\cdot\max\{|v|,\delta\}}. \tag{94}$$

Furthermore, g is a tight upper bound of h of (OR^2-Huber): For every $r \in \mathbb{R}$ and every $v \in \mathbb{R}$ we have

$$h(r) \leq g(r, v)$$
$$h'(r) = \frac{\partial g(r, v)}{\partial r}\Big|_{r=v}$$

In the first inequality, equality is attained if and only if $r = v$.

Proof. This is an elementary result, and the proof is omitted. □

Lemma 10 (Lipschitz Smoothness)**.** *Recall the definition of $g(\cdot, \cdot)$ in (94). For every $\boldsymbol{b} \in \mathcal{M}$, every $\boldsymbol{b}' \in \mathcal{M}$, and every $\boldsymbol{z} \in \mathcal{M}$ we have*

$$\left\|\nabla g\big(\boldsymbol{y}_j^\top \boldsymbol{b}, \boldsymbol{y}_j^\top \boldsymbol{z}\big) - \nabla g\big(\boldsymbol{y}_j^\top \boldsymbol{b}', \boldsymbol{y}_j^\top \boldsymbol{z}\big)\right\|_2 \leq \frac{1}{\delta}\|\boldsymbol{y}_j\|_2^2 \cdot \left\|\boldsymbol{b} - \boldsymbol{b}'\right\|_2$$

In the above, $\nabla g(\boldsymbol{y}_j^\top \boldsymbol{b}, \boldsymbol{y}_j^\top \boldsymbol{z})$ denotes the gradient of g with respect to $\boldsymbol{b}$.

Proof. First note that for each $\boldsymbol{b} \in \mathcal{M}$ and $\boldsymbol{z} \in \mathcal{M}$ we have

$$\nabla g(\boldsymbol{y}_j^\top \boldsymbol{b}, \boldsymbol{y}_j^\top \boldsymbol{z}) = \frac{\boldsymbol{y}_j \boldsymbol{y}_j^\top \boldsymbol{b}}{\max\{|\boldsymbol{y}_j^\top \boldsymbol{z}|, \delta\}}$$

Therefore

$$\begin{aligned}\left\|\nabla g\big(\boldsymbol{y}_j^\top \boldsymbol{b}, \boldsymbol{y}_j^\top \boldsymbol{z}\big) - \nabla g\big(\boldsymbol{y}_j^\top \boldsymbol{b}', \boldsymbol{y}_j^\top \boldsymbol{z}\big)\right\|_2 &= \left\|\frac{\boldsymbol{y}_j \boldsymbol{y}_j^\top \boldsymbol{b}}{\max\{|\boldsymbol{y}_j^\top \boldsymbol{z}|, \delta\}} - \frac{\boldsymbol{y}_j \boldsymbol{y}_j^\top \boldsymbol{b}'}{\max\{|\boldsymbol{y}_j^\top \boldsymbol{z}|, \delta\}}\right\|_2 \\ &\leq \left\|\frac{\boldsymbol{y}_j \boldsymbol{y}_j^\top}{\max\{|\boldsymbol{y}_j^\top \boldsymbol{z}|, \delta\}}\right\|_2 \cdot \left\|\boldsymbol{b} - \boldsymbol{b}'\right\|_2 \\ &\leq \frac{1}{\delta}\|\boldsymbol{y}_j\|_2^2 \cdot \left\|\boldsymbol{b} - \boldsymbol{b}'\right\|_2\end{aligned}$$

The proof is finished. □

Lemma 11 (Sufficient Decrease)**.** *The sequence $\{\boldsymbol{b}_k\}_{k=0}^K$ produced by Algorithm 2 satisfies*

$$\sum_{j=1}^L g\big(\boldsymbol{y}_j^\top \boldsymbol{b}_{k+1}, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big) - \sum_{j=1}^L g\big(\boldsymbol{y}_j^\top \boldsymbol{b}_k, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big) \leq -\frac{1}{2\sigma}\left\|\operatorname{rgrad}\sum_{j=1}^L g\big(\boldsymbol{y}_j^\top \boldsymbol{b}_k, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big)\right\|_2^2.$$

Here, σ is a finite positive constant depending on the manifold $\mathcal{M}$ and data $\boldsymbol{y}_j$'s.

Proof. By Lemma 10 we know that $\sum_{j=1}^L g\big(\boldsymbol{y}_j^\top \boldsymbol{b}, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big)$ is Lipschitz smooth in $\boldsymbol{b}$ (with constant $\frac{1}{\delta}\sum_{j=1}^L \|\boldsymbol{y}_j\|_2^2$). Note also that $\mathcal{M}$ is a compact smooth submanifold of $\mathbb{S}^{D-1}$ by Assumption 1. Therefore, it follows from Appendix B of Boumal et al. (2019) that there is some finite constant $\sigma > 0$ satisfying

$$\sum_{j=1}^L g\big(\boldsymbol{y}_j^\top \operatorname{retr}_{\boldsymbol{b}_k}(\boldsymbol{r}), \boldsymbol{y}_j^\top \boldsymbol{b}_k\big) \leq \sum_{j=1}^L g\big(\boldsymbol{y}_j^\top \boldsymbol{b}_k, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big) + \boldsymbol{r}^\top\Big(\operatorname{rgrad}\sum_{j=1}^L g\big(\boldsymbol{y}_j^\top \boldsymbol{b}_k, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big)\Big) + \frac{\sigma}{2}\|\boldsymbol{r}\|_2^2$$

for every $\boldsymbol{r} \in T_{\boldsymbol{b}_k}\mathcal{M}$; σ depends on the manifold $\mathcal{M}$ and the Lipschitz smoothness $\frac{1}{\delta}\sum_{j=1}^L \|\boldsymbol{y}_j\|_2^2$ of $\sum_{j=1}^L g\big(\boldsymbol{y}_j^\top \boldsymbol{b}, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big)$, i.e., σ depends on $\mathcal{M}$, data $\boldsymbol{y}_j$'s, and δ. In particular, the above holds with $\boldsymbol{r} = -\frac{1}{\sigma}\operatorname{rgrad}\sum_{j=1}^L g\big(\boldsymbol{y}_j^\top \boldsymbol{b}_k, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big)$, meaning that

$$\sum_{j=1}^L g\big(\boldsymbol{y}_j^\top \operatorname{retr}_{\boldsymbol{b}_k}(\boldsymbol{r}), \boldsymbol{y}_j^\top \boldsymbol{b}_k\big) \leq \sum_{j=1}^L g\big(\boldsymbol{y}_j^\top \boldsymbol{b}_k, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big) - \frac{1}{2\sigma}\left\|\operatorname{rgrad}\sum_{j=1}^L g\big(\boldsymbol{y}_j^\top \boldsymbol{b}_k, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big)\right\|_2^2.$$

Since $\boldsymbol{b}_{k+1}$ is a global minimizer of $\sum_{j=1}^L g\big(\boldsymbol{y}_j^\top \boldsymbol{b}, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big)$ in $\boldsymbol{b} \in \mathcal{M}$, $\sum_{j=1}^L g\big(\boldsymbol{y}_j^\top \boldsymbol{b}_{k+1}, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big)$ is smaller than or equal to $\sum_{j=1}^L g\big(\boldsymbol{y}_j^\top \operatorname{retr}_{\boldsymbol{b}_k}(\boldsymbol{r}), \boldsymbol{y}_j^\top \boldsymbol{b}_k\big)$, and the proof is complete. □

Theorem 5. *The sequence* $\{\boldsymbol{b}_k\}_{k=0}^{K}$ *produced by Algorithm 2 converges at a sublinear rate to a critical point of the following problem:*

$$\min_{\boldsymbol{b}\in\mathcal{M}} \sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}), \qquad h(r) := \begin{cases} |r| & |r| > \delta \\ \frac{1}{2}\left(r^2/\delta + \delta\right) & |r| \le \delta \end{cases} \tag{OR^2-Huber}$$

More precisely, the Riemannian gradient of the objective of (OR^2-Huber) *is bounded above:*

$$\min_{k=0,\dots,K} \left\| \operatorname{rgrad} \sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}_k) \right\|_2 \le \sqrt{\frac{2\sigma \cdot \sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}_0)}{K+1}} \tag{23}$$

In (23)*,* σ *is a finite positive constant depending on the manifold* $\mathcal{M}$*, data* $\boldsymbol{y}_j$*'s, and* δ*.*

Proof. For every $k = 0, \dots, K$ we have

$$\begin{aligned} \sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}_{k+1}) - \sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}_k) &\le \sum_{j=1}^{L} g\big(\boldsymbol{y}_j^\top \boldsymbol{b}_{k+1}, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big) - \sum_{j=1}^{L} g\big(\boldsymbol{y}_j^\top \boldsymbol{b}_k, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big) \\ &\le -\frac{1}{2\sigma} \left\| \operatorname{rgrad} \sum_{j=1}^{L} g\big(\boldsymbol{y}_j^\top \boldsymbol{b}_k, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big) \right\|_2^2 \end{aligned}$$

where the first inequality follows from Lemma 9 and the second from Lemma 11. Rearrange the above inequality and sum over $k = 0, \dots, K$ and we obtain

$$\sum_{k=0}^{K} \left\| \operatorname{rgrad} \sum_{j=1}^{L} g\big(\boldsymbol{y}_j^\top \boldsymbol{b}_k, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big) \right\|_2^2 \le 2\sigma \Big(\sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}_0) - \sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}_{K+1}) \Big)$$

Finally, by Lemma 9 we have $\nabla \sum_{j=1}^{L} g\big(\boldsymbol{y}_j^\top \boldsymbol{b}_k, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big) = \nabla \sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b})$, and therefore

$$\operatorname{rgrad} \sum_{j=1}^{L} g\big(\boldsymbol{y}_j^\top \boldsymbol{b}_k, \boldsymbol{y}_j^\top \boldsymbol{b}_k\big) = \operatorname{rgrad} \sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}).$$

Combining the above yields

$$\sum_{k=0}^{K} \left\| \operatorname{rgrad} \sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}) \right\|_2^2 \le 2\sigma \Big(\sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}_0) - \sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}_{K+1}) \Big),$$

which means

$$\begin{aligned} (K+1) \cdot \min_{k=0,\dots,K} \left\| \operatorname{rgrad} \sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}) \right\|_2^2 &\le 2\sigma \cdot \Big(\sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}_0) - \sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}_{K+1}) \Big), \\ &\le 2\sigma \cdot \sum_{j=1}^{L} h(\boldsymbol{y}_j^\top \boldsymbol{b}_0) \end{aligned}$$

concluding the proof. □