Finite Sample Prediction and Recovery Bounds for Ordinal Embedding

Lalit Jain
University of Michigan
Ann Arbor, MI 48109
lalitj@umich.edu

Kevin Jamieson
University of California, Berkeley
Berkeley, CA 94720
kjamieson@berkeley.edu

Robert Nowak
University of Wisconsin
Madison, WI 53706
rdnowak@wisc.edu

Abstract

The goal of ordinal embedding is to represent items as points in a low-dimensional Euclidean space given a set of constraints like "item i is closer to item j than item k". Ordinal constraints like this often come from human judgments. The classic approach to solving this problem is known as *non-metric multidimensional scaling*. To account for errors and variation in judgments, we consider the noisy situation in which the given constraints are independently corrupted by reversing the correct constraint with some probability. The ordinal embedding problem has been studied for decades, but most past work pays little attention to the question of whether accurate embedding is possible, apart from empirical studies. This paper shows that under a generative data model it is possible to learn the correct embedding from noisy distance comparisons. In establishing this fundamental result, the paper makes several new contributions. First, we derive prediction error bounds for embedding from noisy distance comparisons by exploiting the fact that the rank of a distance matrix of points in $\mathbb{R}^d$ is at most $d+2$. These bounds characterize how well a learned embedding predicts new comparative judgments. Second, we show that the underlying embedding can be recovered by solving a simple convex optimization. This result is highly non-trivial since we show that the linear map corresponding to distance comparisons is non-invertible, but there exists a nonlinear map that is invertible. Third, two new algorithms for ordinal embedding are proposed and evaluated in experiments.

1 Ordinal Embedding

Ordinal embedding aims to represent items as points in $\mathbb{R}^d$ so that the distances between items agree as well as possible with a given set of ordinal comparisons such as item i is closer to item j than to item k. In other words, the goal is to find a geometric representation of data that is faithful to comparative similarity judgments. This problem has been studied and applied for more than 50 years, dating back to the classic *non-metric multidimensional scaling* (NMDS) [1, 2] approach, and it is widely used to gauge and visualize how people perceive similarities.

Despite the widespread application of NMDS and recent algorithmic developments [3, 4, 5, 6, 7], the fundamental question of whether an embedding can be learned from noisy distance/similarity comparisons had not been answered. This paper shows that if the data are generated according to a known probabilistic model, then accurate recovery of the underlying embedding is possible by solving a simple convex optimization, settling this long-standing open question. In the process of answering this question, the paper also characterizes how well a learned embedding predicts new distance comparisons and presents two new computationally efficient algorithms for solving the optimization problem.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

1.1 Related Work

The classic approach to ordinal embedding is NMDS [1, 2]. Recently, several authors have proposed new approaches based on more modern techniques. *Generalized NMDS* [3] and *Stochastic Triplet Embedding (STE)* [6] employ hinge or logistic loss measures and convex relaxations of the low-dimensionality (i.e., rank) constraint based on the nuclear norm. These works are most closely related to the theory and methods in this paper. The *Linear partial order embedding (LPOE)* method is similar, but starts with a known Euclidean embedding and learns a kernel/metric in this space based distance comparison data [7]. The *Crowd Kernel* [4] and *t-STE* [6] propose alternative non-convex loss measures based on probabilistic generative models. The main contributions in these papers are new optimization methods and experimental studies, but did not address the fundamental question of whether an embedding can be recovered under an assumed generative model. Other recent work has looked at the asymptotics of ordinal embedding, showing that embeddings can be learned as the number of items grows and the items densely populate the embedding space [8, 9, 10]. In contrast, this paper focuses on the practical setting involving a finite set items. Finally, it is known that at least $2dn \log n$ distance comparisons are necessary to learn an embedding of n points in $\mathbb{R}^d$ [5].

1.2 Ordinal Embedding from Noisy Data

Consider n points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$. Let $\boldsymbol{X} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$. The *Euclidean distance matrix* $\boldsymbol{D}^\star$ is defined to have elements $D^\star_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$. Ordinal embedding is the problem of recovering $\boldsymbol{X}$ given ordinal constraints on distances. This paper focuses on "triplet" constraints of the form $D^\star_{ij} < D^\star_{ik}$, where $1 \leq i \neq j \neq k \leq n$. Furthermore, we only observe noisy indications of these constraints, as follows. Each triplet $t = (i, j, k)$ has an associated probability p_t satisfying

$$p_t > 1/2 \iff \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 < \|\boldsymbol{x}_i - \boldsymbol{x}_k\|^2 \,.$$

Let $\mathcal{S}$ denote a collection of triplets drawn independently and uniformly at random. And for each $t \in \mathcal{S}$ we observe an independent random variable $y_t = -1$ with probability p_t, and $y_t = 1$ otherwise. The goal is to recover the embedding $\boldsymbol{X}$ from these data. Exact recovery of $\boldsymbol{D}^\star$ from such data requires a known *link* between p_t and $\boldsymbol{D}^\star$. To this end, our main focus is the following problem.

Ordinal Embedding from Noisy Data

Consider n points $\boldsymbol{x}_1, \boldsymbol{x}_2 \cdots, \boldsymbol{x}_n$ in d-dimensional Euclidean space. Let $\mathcal{S}$ denote a collection of triplets and for each $t \in \mathcal{S}$ observe an independent random variable

$$y_t \;=\; \begin{cases} -1 & w.p.\ f(D^\star_{ij} - D^\star_{ik}) \\ 1 & w.p.\ 1 - f(D^\star_{ij} - D^\star_{ik}) \end{cases} .$$

where the link function $f : \mathbb{R} \to [0, 1]$ is known. Estimate $\boldsymbol{X}$ from $\mathcal{S}$, $\{y_t\}$, and f.

For example, if f is the logistic function, then for triplet $t = (i, j, k)$

$$p_t = \mathbb{P}(y_t = -1) \;=\; f(D^\star_{ij} - D^\star_{ik}) \;=\; \frac{1}{1 + \exp(D^\star_{ij} - D^\star_{ik})} \,, \tag{1}$$

then $D^\star_{ij} - D^\star_{ik} = \log\left(\frac{1-p_t}{p_t}\right)$. However, we stress that we only require the existence of a link function for exact recovery of $\boldsymbol{D}^\star$. Indeed, if one just wishes to *predict* the answers to unobserved triplets, then the results of Section 2 hold for arbitrary p_t probabilities. Aspects of the statistical analysis are related to one-bit matrix completion and rank aggregation [11, 12, 13]. However, we use novel methods for the recovery of the embedding based on geometric properties of Euclidean distance matrices.

1.3 Organization of Paper

This paper takes the following approach to ordinal embedding.

1. Our samples are assumed to be independently generated according to a probabilistic model based on an underlying low-rank distance matrix. We use relatively standard statistically learning theory

techniques to analyze the minimizer of a bounded, Lipschitz loss with a nuclear norm constraint, and show that an embedding can be learned from the data that predicts nearly as well as the true embedding with $O(dn \log n)$ samples (Theorem 1).

2. Next, assuming the form of the probabilistic generative model is known (e.g., logistic), we show that if the learned embedding is a good predictor of the ordinal comparisons, then it must also be a good estimator of the true differences of distances between the embedding points (Theorem 2). This result hinges on the fact that the (linear) observation model acts approximately like an isometry on differences of distances.

3. While the true differences of distances can be estimated, the observation process is "blind" to the mean distance between embedding points. Despite this, we show that the mean is determined by the differences of distances, due to the special properties of Euclidean distance matrices. Specifically, the second eigenvalue of the "mean-centered" distance matrix (well-estimated by the data from the estimate of the differences of distances, Theorem 3) is proportional to the mean distance (Theorem 4). This allows us to show that the minimizer of the loss with a nuclear norm constraint indeed recovers an accurate estimate of the underlying true distance matrix.

1.4 Notation and Assumptions

We will use $(\boldsymbol{D}^\star, \boldsymbol{G}^\star)$ to denote the distance and Gram matrices of the latent embedding, and $(\boldsymbol{D}, \boldsymbol{G})$ to denote an arbitrary distance matrix and its corresponding Gram matrix. The observations $\{y_t\}$ carry information about $\boldsymbol{D}^\star$, but distance matrices are invariant to rotation and translation, and therefore it may only be possible to recover $\boldsymbol{X}$ up to a rigid transformation. Without loss of generality, we assume assume the points $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n \in \mathbb{R}^d$ are centered at the origin (i.e., $\sum_{i=1}^n \boldsymbol{x}_i = \boldsymbol{0}$).

Define the *centering matrix* $\boldsymbol{V} := \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$. If $\boldsymbol{X}$ is centered, $\boldsymbol{X}\boldsymbol{V} = \boldsymbol{X}$. Note that $\boldsymbol{D}^\star$ is determined by the Gram matrix $\boldsymbol{G}^\star = \boldsymbol{X}^T\boldsymbol{X}$. In addition, $\boldsymbol{X}$ can be determined from $\boldsymbol{G}^\star$ up to a unitary transformation. Note that if $\boldsymbol{X}$ is centered, the Gram matrix is "centered" so that $\boldsymbol{V}\boldsymbol{G}^\star\boldsymbol{V} = \boldsymbol{G}^\star$. It will be convenient in the paper to work with both the distance and Gram matrix representations, and the following identities will be useful to keep in mind. For any distance matrix $\boldsymbol{D}$ and its centered Gram matrix $\boldsymbol{G}$

$$\boldsymbol{G} = -\frac{1}{2}\boldsymbol{V}\boldsymbol{D}\boldsymbol{V} , \tag{2}$$

$$\boldsymbol{D} = \text{diag}(\boldsymbol{G})\boldsymbol{1}^T - 2\boldsymbol{G} + \boldsymbol{1}\text{diag}(\boldsymbol{G})^T , \tag{3}$$

where $\text{diag}(\boldsymbol{G})$ is the column vector composed of the diagonal of $\boldsymbol{G}$. In particular this establishes a bijection between centered Gram matrices and distance matrices. We refer the reader to [14] for an insightful and thorough treatment of the properties of distance matrices. We also define the set of all unique triplets

$$\mathcal{T} := \left\{(i,j,k) : 1 \le i \ne j \ne k \le n, j < k\right\} .$$

Assumption 1. *The observed triplets in $\mathcal{S}$ are drawn independently and unifomly from $\mathcal{T}$.*

2 Prediction Error Bounds

For $t \in \mathcal{T}$ with $t = (i,j,k)$ we define $\mathcal{L}_t$ to be the linear operator satisfying $\mathcal{L}_t(\boldsymbol{X}^T\boldsymbol{X}) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 - \|\boldsymbol{x}_i - \boldsymbol{x}_k\|^2$ for all $t \in \mathcal{T}$. In general, for any Gram matrix $\boldsymbol{G}$

$$\mathcal{L}_t(\boldsymbol{G}) := G_{jj} - 2G_{ij} - G_{kk} + 2G_{ik}.$$

We can naturally view $\mathcal{L}_t$ as a linear operator on $\mathbb{S}_+^n$, the space of $n \times n$ symmetric positive semidefinite matrices. We can also represent $\mathcal{L}_t$ as a symmetric $n \times n$ matrix that is zero everywhere except on the submatrix corresponding to i, j, k which has the form

$$\begin{bmatrix} 0 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

and so we will write

$$\mathcal{L}_t(\boldsymbol{G}) := \langle \mathcal{L}_t, \boldsymbol{G} \rangle$$

where $\langle A, B\rangle = \text{vec}(A)^T \text{vec}(B)$ for any compatible matrices A, B. Ordering the elements of $\mathcal{T}$ lexicographically, we arrange all the $\mathcal{L}_t(\boldsymbol{G})$ together to define the $n\binom{n-1}{2}$-dimensional vector

$$\mathcal{L}(\boldsymbol{G}) = [\mathcal{L}_{123}(\boldsymbol{G}), \mathcal{L}_{124}(\boldsymbol{G}), \cdots, \mathcal{L}_{ijk}(\boldsymbol{G}), \cdots]^T. \tag{4}$$

Let $\ell(y_t\langle \mathcal{L}_t, \boldsymbol{G}\rangle)$ denote a loss function. For example we can consider the $0-1$ loss $\ell(y_t\langle \mathcal{L}_t, \boldsymbol{G}\rangle) = \mathbb{1}_{\{\text{sign}\{y_t\langle \mathcal{L}_t, \boldsymbol{G}\rangle\}\neq 1\}}$, the hinge-loss $\ell(y_t\langle \mathcal{L}_t, \boldsymbol{G}\rangle) = \max\{0, 1 - y_t\langle \mathcal{L}_t, \boldsymbol{G}\rangle\}$, or the logistic loss

$$\ell(y_t\langle \mathcal{L}_t, \boldsymbol{G}\rangle) = \log(1 + \exp(-y_t\langle \mathcal{L}_t, \boldsymbol{G}\rangle)). \tag{5}$$

Let $p_t := \mathbb{P}(y_t = -1)$ and take the expectation of the loss with respect to both the uniformly random selection of the triple t and the observation y_t, we have the risk of $\boldsymbol{G}$

$$R(\boldsymbol{G}) \quad := \quad \mathbb{E}[\ell(y_t\langle \mathcal{L}_t, \boldsymbol{G}\rangle)] = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} p_t\ell(-\langle \mathcal{L}_t, \boldsymbol{G}\rangle) + (1-p_t)\ell(\langle \mathcal{L}_t, \boldsymbol{G}\rangle).$$

Given a set of observations $\mathcal{S}$ under the model defined in the problem statement, the empirical risk is,

$$\widehat{R}_{\mathcal{S}}(\boldsymbol{G}) = \frac{1}{|\mathcal{S}|}\sum_{t\in\mathcal{S}} \ell(y_t\langle \mathcal{L}_t, \boldsymbol{G}\rangle) \tag{6}$$

which is an unbiased estimator of the true risk: $\mathbb{E}[\widehat{R}_{\mathcal{S}}(\boldsymbol{G})] = R(\boldsymbol{G})$. For any $\boldsymbol{G} \in \mathbb{S}^n_+$, let $\|\boldsymbol{G}\|_*$ denote the nuclear norm and $\|\boldsymbol{G}\|_\infty := \max_{ij} |\boldsymbol{G}_{ij}|$. Define the constraint set

$$\mathcal{G}_{\lambda,\gamma} \ := \ \{\boldsymbol{G} \in \mathbb{S}^n_+ : \|\boldsymbol{G}\|_* \leq \lambda, \|\boldsymbol{G}\|_\infty \leq \gamma\} \ . \tag{7}$$

We estimate $\boldsymbol{G}^\star$ by $\widehat{\boldsymbol{G}}$, the solution of the program,

$$\widehat{\boldsymbol{G}} := \underset{\boldsymbol{G}\in\mathcal{G}_{\lambda,\gamma}}{\text{argmin}} \widehat{R}_{\mathcal{S}}(\boldsymbol{G}) \ . \tag{8}$$

Since $\boldsymbol{G}^\star$ is positive semidefinite, we expect the diagonal entries of $\boldsymbol{G}^\star$ to bound the off-diagonal entries. So an infinity norm constraint on the diagonal guarantees that the points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ corresponding to $\boldsymbol{G}^\star$ live inside a bounded ℓ_2 ball. The ℓ_∞ constraint in (7) plays two roles: 1) if our loss function is Lipschitz, large magnitude values of $\langle \mathcal{L}_t, \boldsymbol{G}\rangle$ can lead to large deviations of $\widehat{R}_{\mathcal{S}}(\boldsymbol{G})$ from $R(\boldsymbol{G})$; bounding $||\boldsymbol{G}||_\infty$ bounds $|\langle \mathcal{L}_t, \boldsymbol{G}\rangle|$. 2) Later we will define ℓ in terms of the link function f and as the magnitude of $\langle \mathcal{L}_t, \boldsymbol{G}\rangle$ increases the magnitude of the derivative of the link function f typically becomes very small, making it difficult to "invert"; bounding $||\boldsymbol{G}||_\infty$ tends to keep $\langle \mathcal{L}_t, \boldsymbol{G}\rangle$ within an invertible regime of f.

Theorem 1. *Fix λ, γ and assume $\boldsymbol{G}^\star \in \mathcal{G}_{\lambda,\gamma}$. If the loss function $\ell(\cdot)$ is L-Lipschitz (or $|\sup_y \ell(y)| \leq L\max\{1, 12\gamma\}$) then with probability at least $1-\delta$,*

$$R(\widehat{\boldsymbol{G}}) - R(\boldsymbol{G}^\star) \leq \frac{4L\lambda}{|\mathcal{S}|}\left(\sqrt{\frac{18|\mathcal{S}|\log(n)}{n}} + \frac{\sqrt{3}}{3}\log n\right) + L\gamma\sqrt{\frac{288\log 2/\delta}{|\mathcal{S}|}}$$

Proof. The proof follows from standard statistical learning theory techniques, see for instance [15]. By the bounded difference inequality, with probability $1-\delta$

$$\begin{aligned} R(\widehat{\boldsymbol{G}}) - R(\boldsymbol{G}^\star) &= R(\widehat{\boldsymbol{G}}) - \widehat{R}_{\mathcal{S}}(\widehat{\boldsymbol{G}}) + \widehat{R}_{\mathcal{S}}(\widehat{\boldsymbol{G}}) - \widehat{R}_{\mathcal{S}}(\boldsymbol{G}^\star) + \widehat{R}_{\mathcal{S}}(\boldsymbol{G}^\star) - R(\boldsymbol{G}^\star) \\ &\leq 2\sup_{\boldsymbol{G}\in\mathcal{G}_{\lambda,\gamma}} |\widehat{R}_{\mathcal{S}}(\boldsymbol{G}) - R(\boldsymbol{G})| \leq 2\mathbb{E}[\sup_{\boldsymbol{G}\in\mathcal{G}_{\lambda,\gamma}} |\widehat{R}_{\mathcal{S}}(\boldsymbol{G}) - R(\boldsymbol{G})|] + \sqrt{\frac{2B^2\log 2/\delta}{|S|}} \end{aligned}$$

where $\sup_{\boldsymbol{G}\in\mathcal{G}_{\lambda,\gamma}} \ell(y_t\langle \mathcal{L}_t, \boldsymbol{G}\rangle) - \ell(y_{t'}\langle \mathcal{L}_{t'}, \boldsymbol{G}\rangle) \leq \sup_{\boldsymbol{G}\in\mathcal{G}_{\lambda,\gamma}} L|\langle y_t\mathcal{L}_t - y_{t'}\mathcal{L}_{t'}, \boldsymbol{G}\rangle| \leq 12L\gamma =: B$ using the facts that $\mathcal{L}_t$ has 6 non-zeros of magnitude 1 and $||\boldsymbol{G}||_\infty \leq \gamma$.

Using standard symmetrization and contraction lemmas, we can introduce Rademacher random variables $\epsilon_t \in \{-1, 1\}$ for all $t \in \mathcal{S}$ so that

$$\mathbb{E}\sup_{\boldsymbol{G}\in\mathcal{G}_{\lambda,\gamma}} |\widehat{R}_{\mathcal{S}}(\boldsymbol{G}) - R(\boldsymbol{G})| \leq \mathbb{E}\sup_{\boldsymbol{G}\in\mathcal{G}_{\lambda,\gamma}} \frac{2L}{|\mathcal{S}|}\left|\sum_{t\in\mathcal{S}} \epsilon_t\langle \mathcal{L}_t, \boldsymbol{G}\rangle\right|.$$

The right hand side is just the Rademacher complexity of $\mathcal{G}_{\lambda,\gamma}$. By definition,

$$\{\boldsymbol{G} : \|\boldsymbol{G}\|_* \leq \lambda\} = \lambda \cdot \text{conv}(\{uu^T : |u| = 1\}).$$

where $\text{conv}(U)$ is the convex hull of a set U. Since the Rademacher complexity of a set is the same as the Rademacher complexity of it's closed convex hull,

$$\mathbb{E} \sup_{\boldsymbol{G} \in \mathcal{G}_{\lambda,\gamma}} \left| \sum_{t \in \mathcal{S}} \epsilon_t \langle \mathcal{L}_t, \boldsymbol{G} \rangle \right| \leq \lambda \mathbb{E} \sup_{|u|=1} \left| \sum_{t \in \mathcal{S}} \epsilon_t \langle \mathcal{L}_t, uu^T \rangle \right| = \lambda \mathbb{E} \sup_{|u|=1} \left| u^T \left(\sum_{t \in \mathcal{S}} \epsilon_t \mathcal{L}_t \right) u \right|$$

which we recognize is just $\lambda \mathbb{E} \| \sum_{t \in \mathcal{S}} \epsilon_t \mathcal{L}_t \|$. By [16, 6.6.1] we can bound the operator norm $\| \sum_{t \in S} \epsilon_t \mathcal{L}_t \|$ in terms of the variance of $\sum_{t \in \mathcal{S}} \mathcal{L}_t^2$ and the maximal eigenvalue of $\max_t \mathcal{L}_t$. These are computed in Lemma 1 given in the supplemental materials. Combining these results gives,

$$\frac{2L\lambda}{|\mathcal{S}|} \mathbb{E} \| \sum_{t \in \mathcal{S}} \epsilon_t \mathcal{L}_t \| \leq \frac{2L\lambda}{|\mathcal{S}|} \left(\sqrt{\frac{18|\mathcal{S}| \log(n)}{n}} + \frac{\sqrt{3}}{3} \log n \right).$$

□

We remark that if $\boldsymbol{G}$ is a rank $d < n$ matrix then

$$\|\boldsymbol{G}\|_* \leq \sqrt{d} \|\boldsymbol{G}\|_F \leq \sqrt{d} n \|\boldsymbol{G}\|_\infty$$

so if $\boldsymbol{G}^\star$ is low rank, we really only need a bound on the infinity norm of our constraint set. Under the assumption that $\boldsymbol{G}^\star$ is rank d with $||\boldsymbol{G}^\star||_\infty \leq \gamma$ and we set $\lambda = \sqrt{d} n \gamma$, then Theorem 1 implies that for $|S| > n \log n / 161$

$$R(\widehat{\boldsymbol{G}}) - R(\boldsymbol{G}^\star) \leq 8L\gamma \sqrt{\frac{18dn \log(n)}{|\mathcal{S}|}} + L\gamma \sqrt{\frac{288 \log 2/\delta}{|\mathcal{S}|}}$$

with probability at least $1 - \delta$. The above display says that $|\mathcal{S}|$ must scale like $dn \log(n)$ which is consistent with known finite sample bounds [5].

3 Maximum Likelihood Embedding

We now turn our attention to recovering metric information about $\boldsymbol{G}^\star$. Let $\mathcal{S}$ be a collection of triplets sampled uniformly at random with replacement and let $f : \mathbb{R} \to (0, 1)$ be a known probability function governing the observations. Any link function f induces a natural loss function ℓ_f, namely, the negative log-likelihood of a solution $\boldsymbol{G}$ given an observation y_t defined as

$$\ell_f(y_t \langle \mathcal{L}_t, \boldsymbol{G} \rangle) = \mathbb{1}_{y_t = -1} \log(\tfrac{1}{f(\langle \mathcal{L}_t, \boldsymbol{G} \rangle)}) + \mathbb{1}_{y_t = 1} \log(\tfrac{1}{1 - f(\langle \mathcal{L}_t, \boldsymbol{G} \rangle)})$$

For example, the logistic link function of (1) induces the logistic loss of (5). Recalling that $\mathbb{P}(y_t = -1) = f(\langle \mathcal{L}_t, \boldsymbol{G} \rangle)$ we have

$$\begin{aligned} \mathbb{E}[\ell_f(y_t \langle \mathcal{L}_t, \boldsymbol{G} \rangle)] &= f(\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle) \log(\tfrac{1}{f(\langle \mathcal{L}_t, \boldsymbol{G} \rangle)}) + (1 - f(\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle) \log(\tfrac{1}{1 - f(\langle \mathcal{L}_t, \boldsymbol{G} \rangle)}) \\ &= H(f(\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle)) + KL(f(\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle) | f(\langle \mathcal{L}_t, \boldsymbol{G} \rangle)) \end{aligned}$$

where $H(p) = p \log(\frac{1}{p}) + (1 - p) \log(\frac{1}{1-p})$ and $KL(p, q) = p \log(\frac{p}{q}) + (1 - p) \log(\frac{1-p}{1-q})$ are the entropy and KL divergence of Bernoulli RVs with means p, q. Recall that $||\boldsymbol{G}||_\infty \leq \gamma$ controls the magnitude of $\langle \mathcal{L}_t, \boldsymbol{G} \rangle$ so for the moment, assume this is small. Then by a Taylor series $f(\langle \mathcal{L}_t, \boldsymbol{G} \rangle) \approx \frac{1}{2} + f'(0) \langle \mathcal{L}_t, \boldsymbol{G} \rangle$ using the fact that $f(0) = \frac{1}{2}$, and by another Taylor series we have

$$\begin{aligned} KL(f(\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle) | f(\langle \mathcal{L}_t, \boldsymbol{G} \rangle)) &\approx KL(\tfrac{1}{2} + f'(0) \langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle | \tfrac{1}{2} + f'(0) \langle \mathcal{L}_t, \boldsymbol{G} \rangle) \\ &\approx 2 f'(0)^2 (\langle \mathcal{L}_t, \boldsymbol{G}^\star - \boldsymbol{G} \rangle)^2. \end{aligned}$$

Thus, recalling the definition of $\mathcal{L}(\boldsymbol{G})$ from (4) we conclude that if $\widetilde{\boldsymbol{G}} \in \arg\min_{\boldsymbol{G}} R(\boldsymbol{G})$ with $R(\boldsymbol{G}) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathbb{E}[\ell_f(y_t \langle \mathcal{L}_t, \boldsymbol{G} \rangle)]$ then one would expect $\mathcal{L}(\widetilde{\boldsymbol{G}}) \approx \mathcal{L}(\boldsymbol{G}^\star)$. Moreover, since $\widehat{R}_{\mathcal{S}}(\boldsymbol{G})$ is an unbiased estimator of $R(\boldsymbol{G})$, one expects $\mathcal{L}(\widehat{\boldsymbol{G}})$ to approximate $\mathcal{L}(\boldsymbol{G}^\star)$. The next theorem, combined with Theorem 1, formalizes this observation; its proof is found in the appendix.

Theorem 2. *Let $C_f = \min_{t\in\mathcal{T}} \inf_{\boldsymbol{G}\in\mathcal{G}_{\lambda,\gamma}} |f'\big(\langle \mathcal{L}_t, \boldsymbol{G}\rangle\big)|$ where f' denotes the derivative of f. Then for any $\boldsymbol{G}$*

$$\frac{2C_f^2}{|\mathcal{T}|}\|\mathcal{L}(\boldsymbol{G}) - \mathcal{L}(\boldsymbol{G}^\star)\|_F^2 \;\le\; R(\boldsymbol{G}) - R(\boldsymbol{G}^\star)\;.$$

Note that if f is the logistic link function of (1) then its straightforward to show that $|f'\big(\langle \mathcal{L}_t, \boldsymbol{G}\rangle\big)| \ge \frac{1}{4}\exp(-|\langle \mathcal{L}_t, \boldsymbol{G}\rangle|) \ge \frac{1}{4}\exp(-6||\boldsymbol{G}||_\infty)$ for any t, $\boldsymbol{G}$ so it suffices to take $C_f = \frac{1}{4}\exp(-6\gamma)$.

It remains to see that we can recover $\boldsymbol{G}^\star$ even given $\mathcal{L}(\boldsymbol{G}^\star)$, much less $\mathcal{L}(\widehat{\boldsymbol{G}})$. To do this, it is more convenient to work with distance matrices instead of Gram matrices. Analogous to the operators $\mathcal{L}_t(\boldsymbol{G})$ defined above, we define the operators Δ_t for $t \in \mathcal{T}$ satisfying,

$$\Delta_t(\boldsymbol{D}) := D_{ij} - D_{ik} \equiv \mathcal{L}_t(\boldsymbol{G})\;.$$

We will view the Δ_t as linear operators on the space of symmetric hollow $n \times n$ matrices $\mathbb{S}_h^n$, which includes distance matrices as special cases. As with $\mathcal{L}$, we can arrange all the Δ_t together, ordering the $t \in \mathcal{T}$ lexicographically, to define the $n\binom{n-1}{2}$-dimensional vector

$$\Delta(\boldsymbol{D}) = [D_{12} - D_{13}, \cdots, D_{ij} - D_{ik}, \cdots]^T.$$

We will use the fact that $\mathcal{L}(\boldsymbol{G}) \equiv \Delta(\boldsymbol{D})$ heavily. Because $\Delta(\boldsymbol{D})$ consists of differences of matrix entries, Δ has a non-trivial kernel. However, it is easy to see that $\boldsymbol{D}$ can be recovered given $\Delta(\boldsymbol{D})$ *and* any one off-diagonal element of $\boldsymbol{D}$, so the kernel is 1-dimensional. Also, the kernel is easy to identify by example. Consider the regular simplex in d dimensions. The distances between all $n = d+1$ vertices are equal and the distance matrix can easily be seen to be $\mathbf{1}\mathbf{1}^T - I$. Thus $\Delta(\boldsymbol{D}) = \mathbf{0}$ in this case. This gives us the following simple result.

Lemma 2. *Let $\mathbb{S}_h^n$ denote the space of symmetric hollow matrices, which includes all distance matrices. For any $\boldsymbol{D} \in \mathbb{S}_h^n$, the set of linear functionals $\{\Delta_t(\boldsymbol{D}),\ t \in \mathcal{T}\}$ spans an $\binom{n}{2} - 1$ dimensional subspace of $\mathbb{S}_h^n$, and the 1-dimensional kernel is given by the span of $\mathbf{1}\mathbf{1}^T - \boldsymbol{I}$.*

So we see that the operator Δ is not invertible on $\mathbb{S}_h^n$. Define $\boldsymbol{J} := \mathbf{1}\mathbf{1}^T - \boldsymbol{I}$. For any $\boldsymbol{D}$, let $\boldsymbol{C}$, *the centered distance matrix*, be the component of $\boldsymbol{D}$ orthogonal to the kernel of $\mathcal{L}$ (i.e., $\mathrm{tr}(\boldsymbol{C}\boldsymbol{J}) = 0$). Then we have the orthogonal decomposition

$$\boldsymbol{D} \;=\; \boldsymbol{C} \;+\; \sigma_D\,\boldsymbol{J}\;,$$

where $\sigma_D = \mathrm{trace}(\boldsymbol{D}\boldsymbol{J})/\|\boldsymbol{J}\|_F^2$. Since $\boldsymbol{G}$ is assumed to be centered, the value of σ_D has a simple interpretation:

$$\sigma_D = \frac{1}{2\binom{n}{2}}\sum_{1\le i\le j\le n} D_{ij} = \frac{2}{n-1}\sum_{1\le i\le n}\langle x_i, x_i\rangle = \frac{2\|\boldsymbol{G}\|_*}{n-1}, \tag{9}$$

the average of the squared distances or alternatively a scaled version of the nuclear norm of $\boldsymbol{G}$.

Let $\widehat{\boldsymbol{D}}$ and $\widehat{\boldsymbol{C}}$ be the corresponding distance and centered distance matrices corresponding to $\widehat{\boldsymbol{G}}$ the solution to 8. Though Δ is not invertible on all $\mathbb{S}_h^n$, it is invertible on the subspace orthogonal to the kernel, namely $\boldsymbol{J}^\perp$. So if $\Delta(\widehat{\boldsymbol{D}}) \approx \Delta(\boldsymbol{D}^\star)$, or equivalently $\mathcal{L}(\widehat{\boldsymbol{G}}) \approx \mathcal{L}(\boldsymbol{G}^\star)$, we expect $\widehat{\boldsymbol{C}}$ to be close to $\boldsymbol{C}^\star$. The next theorem quantifies this.

Theorem 3. *Consider the setting of Theorems 1 and 2 and let $\widehat{\boldsymbol{C}}, \boldsymbol{C}^\star$ be defined as above. Then*

$$\frac{1}{2\binom{n}{2}}\|\widehat{\boldsymbol{C}} - \boldsymbol{C}^\star\|_F^2 \;\le \frac{L\lambda}{4C_f^2|\mathcal{S}|}\left(\sqrt{\frac{18|\mathcal{S}|\log(n)}{n}} + \frac{\sqrt{3}}{3}\log n\right) + \frac{L\gamma}{4C_f^2}\sqrt{\frac{288\log 2/\delta}{|\mathcal{S}|}}$$

Proof. By combining Theorem 2 with the prediction error bounds obtainined in 1 we see that

$$\frac{2C_f^2}{n\binom{n-1}{2}}\|\mathcal{L}(\widehat{\boldsymbol{G}}) - \mathcal{L}(\boldsymbol{G}^\star)\|_F^2 \le \frac{4L\lambda}{|\mathcal{S}|}\left(\sqrt{\frac{18|\mathcal{S}|\log(n)}{n}} + \frac{\sqrt{3}}{3}\log n\right) + L\gamma\sqrt{\frac{288\log 2/\delta}{|\mathcal{S}|}}.$$

Next we employ the following *restricted isometry property* of Δ on the subspace $\boldsymbol{J}^\perp$ whose proof is in the supplementary materials.

Lemma 3. *Let $\boldsymbol{D}$ and $\boldsymbol{D}'$ be two different distance matrices of n points in $\mathbb{R}^d$ and $\mathbb{R}^{d'}$. Let $\boldsymbol{C}$ and $\boldsymbol{C}'$ be the components of $\boldsymbol{D}$ and $\boldsymbol{D}'$ orthogonal to $\boldsymbol{J}$. Then*

$$n\|\boldsymbol{C}-\boldsymbol{C}'\|_F^2 \;\le\; \|\Delta(\boldsymbol{C})-\Delta(\boldsymbol{C}')\|^2 = \|\Delta(\boldsymbol{D})-\Delta(\boldsymbol{D}')\|^2 \;\le\; 2(n-1)\|\boldsymbol{C}-\boldsymbol{C}'\|_F^2 \,.$$

The result then follows. □

This implies that by collecting enough samples, we can recover the centered distance matrix. By applying the discussion following Theorem 1 when $\boldsymbol{G}^\star$ is rank d, we can state an upperbound of $\frac{1}{2\binom{n}{2}}\|\widehat{\boldsymbol{C}}-\boldsymbol{C}^\star\|_F^2 \le O\left(\frac{L\gamma}{C_f^2}\sqrt{\frac{dn\log(n)+\log(1/\delta)}{|\mathcal{S}|}}\right)$. However, it is still not clear that this is enough to recover $\boldsymbol{D}^\star$ or $\boldsymbol{G}^\star$. Remarkably, despite this unknown component being in the kernel, we show next that it can be recovered.

Theorem 4. *Let $\boldsymbol{D}$ be a distance matrix of n points in $\mathbb{R}^d$, let $\boldsymbol{C}$ be the component of $\boldsymbol{D}$ orthogonal to the kernel of $\mathcal{L}$, and let $\lambda_2(\boldsymbol{C})$ denote the second largest eigenvalue of $\boldsymbol{C}$. If $n > d+2$, then*

$$\boldsymbol{D} \;=\; \boldsymbol{C} \;+\; \lambda_2(\boldsymbol{C})\,\boldsymbol{J}\,. \tag{10}$$

This shows that $\boldsymbol{D}$ is uniquely determined as a function of $\boldsymbol{C}$. Therefore, since $\Delta(\boldsymbol{D}) = \Delta(\boldsymbol{C})$ and because $\boldsymbol{C}$ is orthogonal to the kernel of Δ, the distance matrix $\boldsymbol{D}$ can be recovered from $\Delta(\boldsymbol{D})$, even though the linear operator Δ is non-invertible.

We now provide a proof of Theorem 4 in the case where $n > d+3$. The result is true in the case when $n > d+2$ but requires a more detailed analysis. This includes the construction of a vector $\boldsymbol{x}$ such that $\boldsymbol{D}\boldsymbol{x} = \mathbf{1}$ and $\mathbf{1}^T\boldsymbol{x} \ge 0$ for any distance matrix a result in [17].

Proof. To prove Theorem 4 we need the following lemma, proved in the supplementary materials.

Lemma 4. *Let $\boldsymbol{D}$ be a Euclidean distance matrix on n points. Then $\boldsymbol{D}$ is negative semidefinite on the subspace*

$$\mathbf{1}^\perp := \{\boldsymbol{x} \in \mathbb{R}^n | \mathbf{1}^T\boldsymbol{x} = 0\}.$$

Furthermore, $\ker(\boldsymbol{D}) \subset \mathbf{1}^\perp$.

For any matrix $\boldsymbol{M}$, let $\lambda_i(\boldsymbol{M})$ denote its ith largest eigenvalue. Under the conditions of the theorem, we show that for $\sigma > 0$, $\lambda_2(\boldsymbol{D}-\sigma\boldsymbol{J}) = \sigma$. Since $\boldsymbol{C} = \boldsymbol{D} - \sigma_D\boldsymbol{J}$, this proves the theorem.

Note that, $\lambda_i(\boldsymbol{D}-\sigma\boldsymbol{J}) = \lambda_i(\boldsymbol{D}-\sigma\mathbf{1}\mathbf{1}^T)+\sigma$ for $1 \le i \le n$ and σ arbitrary. So it suffices to show that $\lambda_2(D-\sigma\mathbf{1}\mathbf{1}^T) = 0$.

By Weyl's Theorem

$$\lambda_2(\boldsymbol{D}-\sigma\mathbf{1}\mathbf{1}^T) \;\le\; \lambda_2(\boldsymbol{D}) + \lambda_1(-\sigma\mathbf{1}\mathbf{1}^T)\,.$$

Since $\lambda_1(-\sigma\mathbf{1}\mathbf{1}^T) = 0$, we have $\lambda_2(\boldsymbol{D}-\sigma\mathbf{1}\mathbf{1}^T) \le \lambda_2(\boldsymbol{D}) = 0$. By the Courant-Fischer Theorem

$$\lambda_2(\boldsymbol{D}) \;=\; \min_{U:\dim(U)=n-1}\;\max_{\boldsymbol{x}\in U,\boldsymbol{x}\neq 0}\frac{\boldsymbol{x}^T\boldsymbol{D}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \;\le\; \min_{U=\mathbf{1}^\perp}\;\max_{\boldsymbol{x}\in U,\boldsymbol{x}\neq 0}\frac{\boldsymbol{x}^T\boldsymbol{D}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \;\le\; 0$$

since $\boldsymbol{D}$ negative semidefinite on $\mathbf{1}^\perp$. Now let $\boldsymbol{v}_i$ denote the ith eigenvector of $\boldsymbol{D}$ with eigenvalue $\lambda_i = 0$. Then

$$(\boldsymbol{D}-\sigma\mathbf{1}\mathbf{1}^T)\boldsymbol{v}_i = \boldsymbol{D}\boldsymbol{v}_i = 0\,,$$

since $\boldsymbol{v}_i^T\mathbf{1} = 0$ by 4. So $\boldsymbol{D}-\sigma\mathbf{1}\mathbf{1}^T$ has at least $n-d-2$ zero eigenvalues, since rank$\boldsymbol{D} \le d+2$. In particular, if $n > d+3$, then $\boldsymbol{D}-\sigma\mathbf{1}\mathbf{1}^T$ must have at least two eigenvalues equal to 0. Therefore, $\lambda_2(\boldsymbol{D}-\sigma\mathbf{1}\mathbf{1}^T) = 0$. □

The previous theorem along with Theorem 3 guarantees that we can recover $\boldsymbol{G}^\star$ as we increase the number of triplets sampled. The final theorem, which follows directly from Theorems 3 and 4, summarizes this.

Theorem 5. *Assume $n > d+2$ and consider the setting of Theorems 1 and 2. As $|\mathcal{S}| \to \infty$, $\widehat{\boldsymbol{D}} \to \boldsymbol{D}^*$ where $\widehat{\boldsymbol{D}}$ is the distance matrix corresponding to $\widehat{\boldsymbol{G}}$ (the solution to 8).*

Proof. Recall $\widehat{\boldsymbol{D}} = \widehat{\boldsymbol{C}} + \lambda_2(\widehat{\boldsymbol{C}})\boldsymbol{J}$, so as $\widehat{\boldsymbol{C}} \to \boldsymbol{C}^*$, $\widehat{\boldsymbol{D}} \to \boldsymbol{D}^*$. □

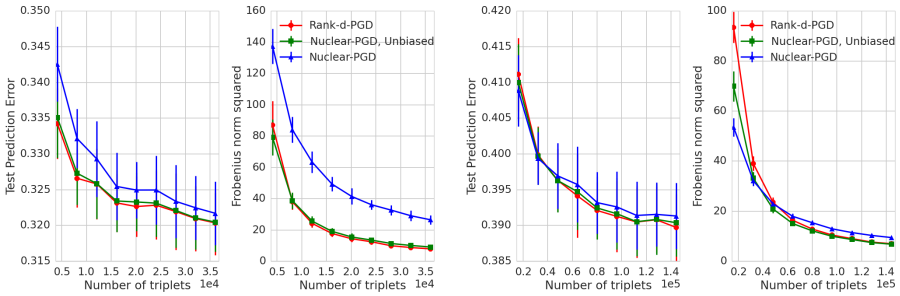

Figure 1: $\boldsymbol{G}^\star$ generated with $n = 64$ points in $d = 2$ and $d = 8$ dimensions on the left and right.

4 Experimental Study

The section empirically studies the properties of estimators suggested by our theory. It is *not* an attempt to perform an exhaustive empirical evaluation of different embedding techniques; for that see [18, 4, 6, 3]. In what follows each of the n points is generated randomly: $\boldsymbol{x}_i \sim \mathcal{N}(0, \frac{1}{2d} I_d) \in \mathbb{R}^d$, $i = 1, \ldots, n$, motivated by the observation that

$$\mathbb{E}[|\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle|] \;=\; \mathbb{E}\big[\big|\, \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 - ||\boldsymbol{x}_i - \boldsymbol{x}_k||_2^2 \,\big|\big] \;\leq\; \mathbb{E}\big[\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2\big] \;=\; 2\mathbb{E}\big[\|\boldsymbol{x}_i\|_2^2\big] \;=\; 1$$

for any triplet $t = (i, j, k)$.We report the prediction error on a holdout set of $10,000$ triplets and the error in Frobenius norm of the estimated Gram matrix over 36 random trials. We minimize the logistic MLE objective $\widehat{R}_{\mathcal{S}}(\boldsymbol{G}) = \frac{1}{|S|} \sum_{t \in S} \log(1 + \exp(-y_t \langle \mathcal{L}_t, \boldsymbol{G} \rangle))$.

For each algorithm considered, the domain of the objective variable $\boldsymbol{G}$ is the space of symmetric positive semi-definite matrices. None of the methods impose the constraint $\max_{ij} |G_{ij}| \leq \gamma$ (as done above), since this was used to simplify the analysis and does not have a large impact in practice. *Rank-d Projected Gradient Descent (PGD)* performs gradient descent on the objective $\widehat{R}_{\mathcal{S}}(\boldsymbol{G})$ with line search, projecting onto the subspace spanned by the top d eigenvalues at each step (i.e. setting the smallest $n - d$ eigenvalues to 0). *Nuclear Norm PGD* performs gradient descent on $\widehat{R}_{\mathcal{S}}(\boldsymbol{G})$ projecting onto the nuclear norm ball with radius $\|\boldsymbol{G}^\star\|_*$, where $\boldsymbol{G}^\star$ is the Gram matrix of the latent embedding. The nuclear norm projection can have the undesirable effect of shrinking the non-zero eigenvalues toward the origin. To compensate for this potential bias, we employ *Nuclear Norm PGD Debiased*, which takes the biased output of *Nuclear Norm PGD*, decomposes it into $\boldsymbol{U}\boldsymbol{E}\boldsymbol{U}^T$ where $\boldsymbol{U} \in \mathbb{R}^{n \times d}$ are the top d eigenvectors, and outputs $\boldsymbol{U}\mathrm{diag}(\widehat{\boldsymbol{s}})\boldsymbol{U}^T$ where $\widehat{s} = \arg\min_{\boldsymbol{s} \in \mathbb{R}^d} \widehat{R}_{\mathcal{S}}(\boldsymbol{U}\mathrm{diag}(\boldsymbol{s})\boldsymbol{U}^T)$. This last algorithm is motivated by the observation that methods for minimizing $\|\cdot\|_1$ or $\|\cdot\|_*$ are good at identifying the true support of a signal, but output biased magnitudes [19]. *Rank-d PGD* and *Nuclear Norm PGD Debiased* are novel ordinal embedding algorithms.

Figure 1 presents how the algorithms behave for $n = 64$ and $d = 2, 8$. We observe that the unbiased nuclear norm solution behaves near-identically to the rank-d solution and remark that this was observed in all of our experiments (see the supplementary materials for other values of n, d, and scalings of $\boldsymbol{G}^\star$). A popular technique for recovering rank d embeddings is to perform (stochastic) gradient descent on $\widehat{R}_{\mathcal{S}}(\boldsymbol{U}^T\boldsymbol{U})$ with objective variable $\boldsymbol{U} \in \mathbb{R}^{n \times d}$ taken as the embedding [18, 4, 6]. In all of our experiments this method produced Gram matrices nearly identical to those produced by our *Rank-d-PGD* method, but *Rank-d-PGD* was an order of magnitude faster in our implementation. Also, in light of our isometry theorem, we can show that the Hessian of $\mathbb{E}[\widehat{R}_{\mathcal{S}}(\boldsymbol{G})]$ is nearly a scaled identity, leading us to hypothesize that a globally optimal linear convergence result for this non-convex optimization may be possible using the techniques of [20, 21]. Finally, we note that previous literature has reported that nuclear norm optimizations like *Nuclear Norm PGD* tend to produce less accurate embeddings than those of non-convex methods [4, 6]. The results imply that *Nuclear Norm PGD Debiased* appears to close the performance gap between the convex and non-convex solutions.

Acknowledgments This work was partially supported by the NSF grants CCF-1218189 and IIS-1447449, the NIH grant 1 U54 AI117924-01, the AFOSR grant FA9550-13-1-0138, and by ONR awards N00014-15-1-2620, and N00014-13-1-0129. We would also like to thank Amazon Web Services for providing the computational resources used for running our simulations.

References

[1] Roger N Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika*, 27(2):125–140, 1962.

[2] Joseph B Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129, 1964.

[3] Sameer Agarwal, Josh Wills, Lawrence Cayton, Gert Lanckriet, David J Kriegman, and Serge Belongie. Generalized non-metric multidimensional scaling. In *International Conference on Artificial Intelligence and Statistics*, pages 11–18, 2007.

[4] Omer Tamuz, Ce Liu, Ohad Shamir, Adam Kalai, and Serge J Belongie. Adaptively learning the crowd kernel. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 673–680, 2011.

[5] Kevin G Jamieson and Robert D Nowak. Low-dimensional embedding using adaptively selected ordinal data. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 1077–1084. IEEE, 2011.

[6] Laurens Van Der Maaten and Kilian Weinberger. Stochastic triplet embedding. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6. IEEE, 2012.

[7] Brian McFee and Gert Lanckriet. Learning multi-modal similarity. *The Journal of Machine Learning Research*, 12:491–523, 2011.

[8] Matthäus Kleindessner and Ulrike von Luxburg. Uniqueness of ordinal embedding. In *COLT*, pages 40–67, 2014.

[9] Yoshikazu Terada and Ulrike V Luxburg. Local ordinal embedding. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 847–855, 2014.

[10] Ery Arias-Castro. Some theory for ordinal embedding. *arXiv preprint arXiv:1501.02861*, 2015.

[11] Mark A Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference*, 3(3), 2014.

[12] Yu Lu and Sahand N Negahban. Individualized rank aggregation using nuclear norm regularization. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1473–1479. IEEE, 2015.

[13] D. Park, J , Neeman, J. Zhang, S. Sanghavi, and I. Dhillon. Preference completion: Large-scale collaborative ranking from pairwise comparisons. *Proc. Int. Conf. Machine Learning (ICML)*, 2015.

[14] Jon Dattorro. *Convex Optimization & Euclidean Distance Geometry*. Meboo Publishing USA, 2011.

[15] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

[16] Joel A. Tropp. An introduction to matrix concentration inequalities, 2015.

[17] Pablo Tarazaga and Juan E. Gallardo. Euclidean distance matrices: new characterization and boundary properties. *Linear and Multilinear Algebra*, 57(7):651–658, 2009.

[18] Kevin G Jamieson, Lalit Jain, Chris Fernandez, Nicholas J Glattard, and Rob Nowak. Next: A system for real-world development, evaluation, and application of active learning. In *Advances in Neural Information Processing Systems*, pages 2638–2646, 2015.

[19] Nikhil Rao, Parikshit Shah, and Stephen Wright. Conditional gradient with enhancement and truncation for atomic norm regularization. In *NIPS workshop on Greedy Algorithms*, 2013.

[20] Samet Oymak, Benjamin Recht, and Mahdi Soltanolkotabi. Sharp time–data tradeoffs for linear inverse problems. *arXiv preprint arXiv:1507.04793*, 2015.

[21] Jie Shen and Ping Li. A tight bound of hard thresholding. *arXiv preprint arXiv:1605.01656*, 2016.

5 Supplementary Materials for "Finite Sample Error Bounds for Ordinal Embedding"

5.1 Proof of Lemma 1

Lemma 1. *For all $t \in \mathcal{T}$,*

$$\lambda_1(\mathcal{L}_t) = \|\mathcal{L}_t\| = \sqrt{3}$$

in addition if $n \geq 3$

$$\|\mathbb{E}_t[\mathcal{L}_t^2]\| = \frac{6}{n-1} \leq \frac{9}{n}$$

Proof. Note that $\mathcal{L}_t^3 - 3\mathcal{L}_t = 0$ for all $t \in \mathcal{T}$. Thus by the Cayley-Hamilton theorem, $\sqrt{3}$ is the largest eigenvalue of $\mathcal{L}_t$. A computation shows that the submatrix of $\mathcal{L}_t^2$ corresponding to i, j, k is

$$\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

and every other element of $\mathcal{L}_t^2$ is zero. Summing over the $t \in \mathcal{T}$ then gives,

$$\mathbb{E}[\mathcal{L}_t^2] = \frac{1}{n\binom{n-1}{2}} \sum_{t \in \mathcal{T}} \mathcal{L}_t^2 = \begin{pmatrix} \frac{6}{n} & \frac{-6}{n(n-1)} & \cdots & \frac{-6}{n(n-1)} \\ \frac{-6}{n(n-1)} & \frac{6}{n} & \cdots & \frac{-6}{n(n-1)} \\ \vdots & \cdots & \cdots & \vdots \\ \frac{-6}{n(n-1)} & \cdots & \frac{-6}{n(n-1)} & \frac{6}{n} \end{pmatrix}$$

This matrix can be rewritten as $\frac{6}{n} I - \frac{6}{n(n-1)} \boldsymbol{J}$. The eigenvalues of $\boldsymbol{J}$ are -1 with multiplicity $n-1$ and $n-1$ with multiplicity 1. Hence the largest eigenvalue of $\mathbb{E}[\mathcal{L}_t^2]$ is $\frac{6}{n-1}$. □

5.2 Proof of Theorem 2

Proof. For $y, z \in (0,1)$ let $g(z) = z \log \frac{z}{y} + (1-z) \log \frac{1-z}{1-y}$. Then $g'(z) = \log \frac{z}{1-z} - \log \frac{y}{1-y}$ and $g''(z) = \frac{1}{z(1-z)}$. By taking a Taylor series around y,

$$g(z) \geq \frac{(z-y)^2/2}{\sup_{x \in [0,1]} x(1-x)} \geq 2(z-y)^2.$$

Now applying this to $z = f(\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle)$ and $y = f(\langle \mathcal{L}_t, \boldsymbol{G} \rangle)$ gives

$$\begin{aligned} f(\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle) \log \frac{f(\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle)}{f(\langle \mathcal{L}_t, \boldsymbol{G} \rangle)} + (1 - f(\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle)) \log \frac{1 - f(\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle)}{1 - f(\langle \mathcal{L}_t, \boldsymbol{G} \rangle)} &\geq 2(f(\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle) - f(\langle \mathcal{L}_t, \boldsymbol{G} \rangle))^2 \\ &\geq 2C_f^2 (\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle - \langle \mathcal{L}_t, \boldsymbol{G} \rangle)^2 \end{aligned}$$

where the last line comes from applying Taylor's theorem to f, $f(x) - f(y) \geq \inf_{z \in [x,y]} f'(z)(x-y)$ for any x, y. Thus

$$\begin{aligned} R(\boldsymbol{G}) - R(\boldsymbol{G}^\star) &= \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} f(\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle) \log \frac{f(\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle)}{f(\langle \mathcal{L}_t, \boldsymbol{G} \rangle)} + (1 - f(\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle)) \log \frac{1 - f(\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle)}{1 - f(\langle \mathcal{L}_t, \boldsymbol{G} \rangle)} \\ &\geq \frac{2C_f^2}{|\mathcal{T}|} \sum_{t \in T} (\langle \mathcal{L}_t, \boldsymbol{G}^\star \rangle - \langle \mathcal{L}_t, \boldsymbol{G} \rangle)^2 \\ &= \frac{2C_f^2}{|\mathcal{T}|} \sum_{t \in T} (\langle \mathcal{L}_t, \boldsymbol{G} - \boldsymbol{G}^\star \rangle)^2 = \frac{2C_f^2}{|\mathcal{T}|} \|\mathcal{L}(\boldsymbol{G}) - \mathcal{L}(\boldsymbol{G}^*)\|_2^2. \end{aligned}$$

□

5.3 Proof of Lemma 3

Lemma 3. *Let $\boldsymbol{D}$ and $\boldsymbol{D}'$ be two different distance matrices of n points in $\mathbb{R}^d$ and $\mathbb{R}^{d'}$ respectively. Let $\boldsymbol{C}$ and $\boldsymbol{C}'$ be the components of $\boldsymbol{D}$ and $\boldsymbol{D}'$ orthogonal to J. Then*

$$n\|\boldsymbol{C}-\boldsymbol{C}'\|_F^2 \;\le\; \|\Delta(\boldsymbol{C})-\Delta(\boldsymbol{C}')\|^2 = \|\Delta(\boldsymbol{D})-\Delta(\boldsymbol{D}')\|^2 \;\le\; 2(n-1)\|\boldsymbol{C}-\boldsymbol{C}'\|_F^2 \,.$$

We can view the operator Δ defined above as acting on the space $\mathbb{R}^{\binom{n}{2}}$ where each symmetric hollow matrix is identified with vectorization of it's upper triangular component. With respect to this basis Δ is an $n\binom{n-1}{2} \times \binom{n}{2}$ matrix, which we will denote by Δ. Since $\boldsymbol{C}$ and $\boldsymbol{C}'$ are orthogonal to the kernel of Δ, the lemma follows immediately from the following characterization of the eigenvalues of $\Delta^T\Delta$.

Lemma 6. *$\Delta^T\Delta : \mathbb{S}_h^n \to \mathbb{S}_h^n$ has the following eigenvalues and eigenspaces,*

- *Eigenvalue 0, with a one dimensional eigenspace.*
- *Eigenvalue n, with a $n-1$ dimensional eigenspace.*
- *Eigenvalue $2(n-1)$, with a $\binom{n}{2}-n$ dimensional eigenspace.*

Proof. The rows of Δ are indexed by triplets $t \in \mathcal{T}$ and columns indexed by pairs i,j with $1 \le i < j \le n$ and vice-versa for Δ^T. The row of Δ^T corresponding to the pair i,j is supported on columns corresponding to triplets $t=(l,m,n)$ where $m<n$ and l and one of m or n form the pair i,j or j,i. Specifically, letting $[\Delta^T]_{(i,j),t}$ denote the entry of Δ^T corresponding to row i,j and column t,

- if $l=i, m=j$ then $[\Delta^T]_{(i,j),t}=1$
- if $l=i, n=j$ then $[\Delta^T]_{(i,j),t}=-1$
- if $l=j, m=i$ then $[\Delta^T]_{(i,j),t}=1$
- if $l=j, n=i$ then $[\Delta^T]_{(i,j),t}=-1$

Using this one can easily check that

$$\begin{aligned}[\Delta^T\Delta\boldsymbol{D}]_{i,j} &= \sum_{(i,j,k)\in\mathcal{T}} D_{ij}-D_{ik} - \sum_{(i,k,j)\in\mathcal{T}} D_{ik}-D_{ij} + \sum_{(j,i,k)\in\mathcal{T}} D_{ji}-D_{jk} - \sum_{(j,k,i)\in\mathcal{T}} D_{jk}-D_{ji} \\ &= 2(n-1)D_{ij} - \sum_{n\neq i} D_{in} - \sum_{n\neq j} D_{jn}. \end{aligned} \tag{11}$$

This representation allows us to find the eigenspaces mentioned above very quickly.

Eigenvalue 0. From the above discussion, we know the kernel is generated by $\boldsymbol{J}=\boldsymbol{1}\boldsymbol{1}^T-\boldsymbol{I}$.

Eigenvalue $2(n-1)$. This eigenspace corresponds to all symmetric hollow matrices such that $\boldsymbol{D}\boldsymbol{1}=\boldsymbol{0}$. For such a matrix each row and column sum is zero and so in particular, the sums in (11) are both zero. Hence for such a $\boldsymbol{D}$,

$$[\Delta^T\Delta\boldsymbol{D}]_{i,j} = 2(n-1)D_{ij}$$

The dimension of this subspace is $\binom{n}{2}-n$, indeed there are $\binom{n}{2}$ degree of freedom to choose the elements of $\boldsymbol{D}$ and $\boldsymbol{D}\boldsymbol{1}=\boldsymbol{0}$ adds n constraints.

Eigenvalue n. This eigenspace corresponds to the span of the matrices $\boldsymbol{D}^{(i)}$ defined as,

$$\boldsymbol{D}^{(i)} = -n(\boldsymbol{e}_i\boldsymbol{1}^T + \boldsymbol{1}\boldsymbol{e}_i^T - 2\boldsymbol{e}_i\boldsymbol{e}_i^T) + 2\boldsymbol{J}$$

where $\boldsymbol{e}_i$ is the standard basis vector with a 1 in the ith row and 0 elsewhere. As an example,

$$\boldsymbol{D}^{(1)} = \begin{pmatrix} 0 & -n+2 & \cdots & -n+2 \\ \vdots & 2 & \cdots & 2 \\ -n+2 & 2 & \cdots & 0 \end{pmatrix}.$$

If $i, j \neq m$, then $D_{ij}^{(m)} := [\boldsymbol{D}^{(m)}]_{ij} = 2$, and we can compute the row and column sums

$$\sum_{n \neq i} D_{in}^{(m)} = \sum_{n \neq j} D_{jn}^{(m)} = 2(n-2) - n + 2 = n - 2.$$

This implies that $D_{ij}^{(m)} = n$, and so by (11)

$$[\Delta^T \Delta \boldsymbol{D}^{(m)}]_{i,j} = 2(n-1) \cdot 2 - (n-2) - (n-2) \ = \ 2n \ = \ nD_{ij}^{(m)} \ .$$

Otherwise, without loss of generality we can assume that $i = m, j \neq m$ in which case, $[\boldsymbol{D}^{(m)}]_{ij} = -n + 2$, the row and columns sums can be computed as

$$\sum_{n \neq i} D_{in}^{(m)} = (n-1)(-n+2)$$

and

$$\sum_{n \neq j} D_{in}^{(m)} = n - 2.$$

Putting it all together,

$$\begin{aligned}
[\Delta^T \Delta \boldsymbol{D}^{(m)}]_{m,j} &= 2(n-1) \cdot (-n+2) - (n-1)(-n+2) - (n-2) \\
&= (n-1)(-n+2) + (-n+2) \\
&= n(-n+2) \\
&= nD_{m,j}^{(m)}
\end{aligned}$$

and $\Delta^T \Delta \boldsymbol{D} = n\boldsymbol{D}$. Note that the dimension of span$\langle \boldsymbol{D}^{(i)} \rangle = n - 1$ since

$$\sum_m \boldsymbol{D}^{(m)} = 0.$$

□

5.4 Proof of Lemmas 4

Lemma 4. *Let $\boldsymbol{D}$ be a Euclidean distance matrix on n points. Then $\boldsymbol{D}$ is negative semidefinite on the subspace*

$$\mathbf{1}^\perp := \{\boldsymbol{x} \in \mathbb{R}^n | \mathbf{1}^T \boldsymbol{x} = 0\}.$$

Furthermore, $\ker(\boldsymbol{D}) \subset \mathbf{1}^\perp$.

Proof. The associated Gram matrix $\boldsymbol{G} = -\frac{1}{2}\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}$ is a positive semidefinite matrix. For $\boldsymbol{x} \in \mathbf{1}^\perp$, $\boldsymbol{J}\boldsymbol{x} = -\boldsymbol{x}$ so

$$\boldsymbol{x}^T \left(-\frac{1}{2}\boldsymbol{V}\boldsymbol{D}\boldsymbol{V} \right) \boldsymbol{x} = -\frac{1}{2}\boldsymbol{x}^T \boldsymbol{D}\boldsymbol{x} \leq 0$$

establishing the first part of the theorem. Now if $\boldsymbol{x} \in \ker \boldsymbol{D}$,

$$0 \leq -\frac{1}{2}\boldsymbol{x}^T \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}\boldsymbol{x} = -\frac{1}{2}\boldsymbol{x}^T \mathbf{1}\mathbf{1}^T \boldsymbol{D}\mathbf{1}\mathbf{1}^T \boldsymbol{x} = -\frac{1}{2}\mathbf{1}^T \boldsymbol{D}\mathbf{1}(\mathbf{1}^T \boldsymbol{x})^2 \leq 0 \ ,$$

where the last inequality follows from the fact that $\mathbf{1}^T \boldsymbol{D}\mathbf{1} > 0$ since $\boldsymbol{D}$ is non-negative. Hence $\mathbf{1}^T \boldsymbol{x} = 0$ and $\ker \boldsymbol{D} \subset \mathbf{1}^\perp$. □

6 Additional Empirical Results

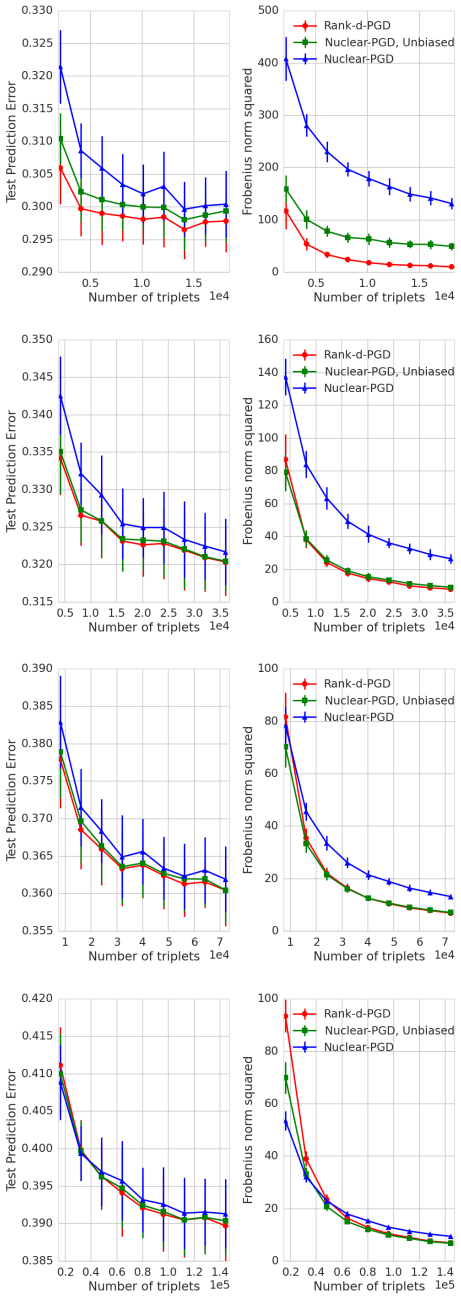

Figure 2: **Varying dimension** $n = 64$, $d = \{1, 2, 4, 8\}$ from top to bottom.

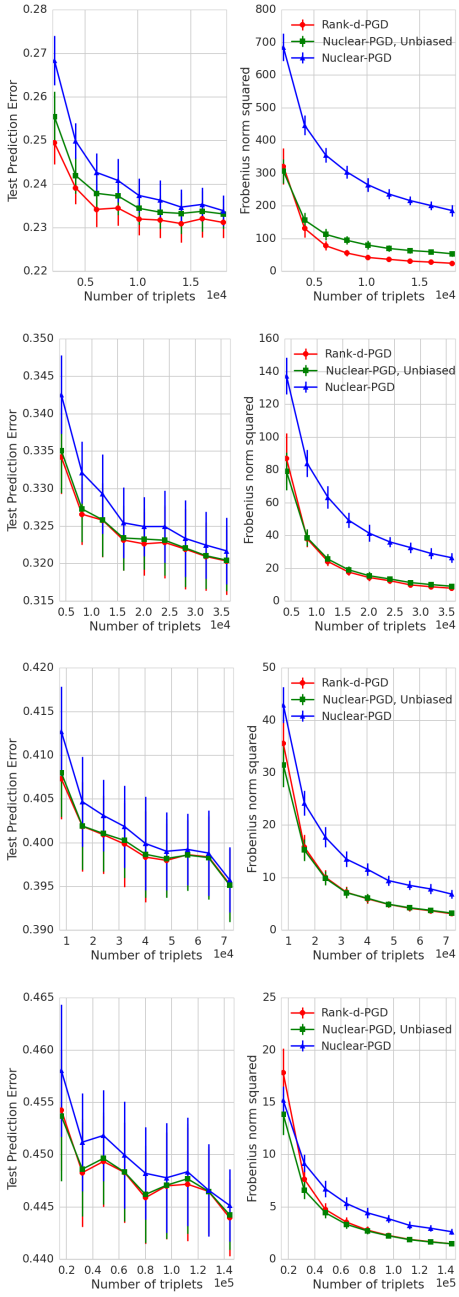

Figure 3: **Varying Noise** $n = 64$, $d = 2$, αM for $\alpha = \{2, 1, \frac{1}{2}, \frac{1}{4}\}$ from top to bottom.

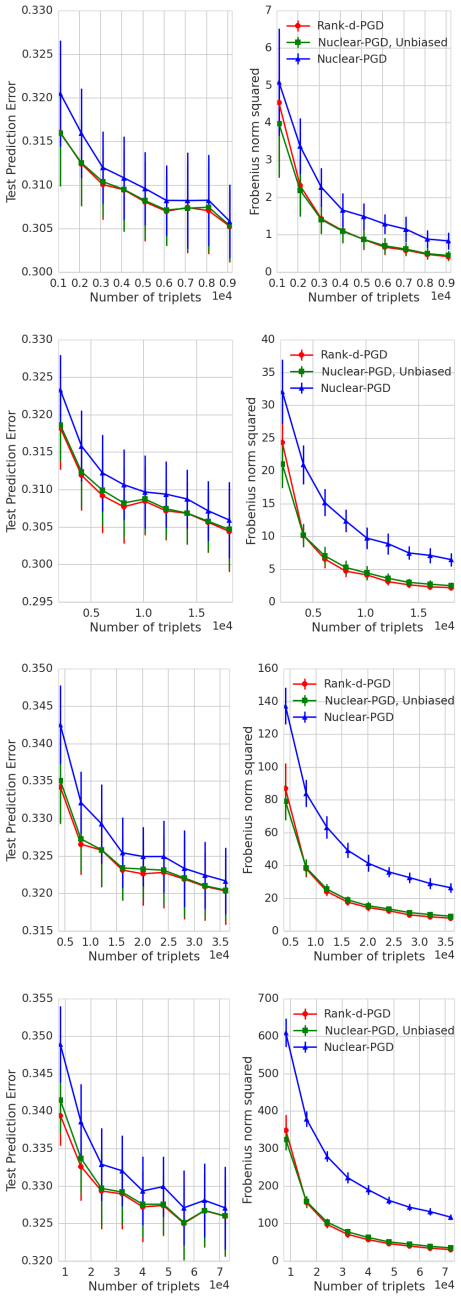

Figure 4: **Varying # items** $n = \{16, 32, 64, 128\}$, $d = 2$ from top to bottom.