THE LANDSCAPE OF NON-CONVEX QUADRATIC FEASIBILITY

Amanda Bower ⋆*, Lalit Jain*⋆*, Laura Balzano* †

⋆ Department of Mathematics, † Department of Electrical Engineering and Computer Science
University of Michigan
amandarg, lalitj, girasole@umich.edu

ABSTRACT

Motivated by applications such as ordinal embedding and collaborative ranking, we formulate homogeneous quadratic feasibility as an unconstrained, non-convex minimization problem. Our work aims to understand the landscape (local minimizers and global minimizers) of the non-convex objective, which corresponds to hinge losses arising from quadratic constraints. Under certain assumptions, we give necessary conditions for non-global, local minimizers of our objective and additionally show that in two dimensions, every local minimizer is a global minimizer. Empirically, we demonstrate that finding feasible points by solving the unconstrained optimization problem with stochastic gradient descent works reliably by utilizing large initializations.

Index Terms: Non-Convex Optimization, Preference Learning

1. INTRODUCTION

In this paper, we consider quadratic feasibility problems and present theory and experimental results utilizing first order methods for recovering a feasible point. To motivate this approach, we consider a natural set of quadratic feasibility problems that arise when using ordinal comparisons to find a Euclidean embedding for a set of items. Such embeddings are useful for downstream machine learning applications such as rank aggregation, visualization, or recommender systems. We give two concrete examples of this now. The first type of embedding method we consider is *ordinal embedding* (also known as non-metric multidimensional scaling [1, 2]) and is based on Euclidean distance comparisons. Given ordinal constraints on distances of the form $T = \{(i, j, k) : \text{item } i \text{ is closer to item } j \text{ than item } k\}$, the goal is to find n points, $X = \{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{R}^d$, that satisfy these Euclidean distance constraints. Let $L_{i,j,k}$ be the matrix that captures the quadratic constraint (i, j, k), i.e., $X^T L_{i,j,k} X = \|x_i - x_k\|^2 - \|x_i - x_j\|^2 > 0$. We can formulate the ordinal embedding feasibility problem as follows: Find $X \in \mathbb{R}^{nd}$ such that $X^T L_{i,j,k} X > 0$ for all $(i, j, k) \in T$. The second type of embedding method is *collaborative ranking*. We assume there are m users, corresponding to columns of a matrix $W \in \mathbb{R}^{d \times m}$, and n items, corresponding to rows of a matrix $U \in \mathbb{R}^{n \times d}$, and each user gives a ranking σ_i on the set of items. Then the i-th column of $X = UW$ contains the scores used by user i to rank all the items. Finding U, W given the rankings σ_i corresponds to the feasibility problem: $\langle W_i, U^T_{\sigma_i(j)} \rangle - \langle W_i, U^T_{\sigma_i(k)} \rangle > 0$.

Crucially, both ordinal embedding and collaborative ranking can be cast as homogeneous quadratic feasibility problems. Both problems have the form:

$$\begin{aligned} \text{find} \quad & x \\ \text{subject to} \quad & x^T P_i x > 0, \quad i = 1, \ldots, m \,, \end{aligned} \tag{1}$$

where P_i is a symmetric matrix corresponding to the i-th constraint. Quadratic feasibility is a special case of quadratically constrained quadratic programming, which has been extensively studied (see the excellent survey [3]). In general, quadratic feasibility with indefinite P_i matrices is NP-hard.

We propose to solve (1) by solving the following optimization problem that penalizes a candidate point when it does not satisfy a quadratic constraint:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \sum_{i \in [m]} \max\{0, 1 - x^T P_i x\}. \tag{2}$$

Similar to support vector machines, the hinge loss in (2) captures a margin, which quantifies the amount a

The authors were supported in part by DARPA grant 16-43-D3M-FP-03 and the University of Michigan MCubed program. A. Bower was also partially supported by the NSF Graduate Research Fellowship under Grant No. DGE 1256260.

constraint is violated. Furthermore, the 1 in the objective of (2) avoids convergence to the infeasible point $\hat{x} = \mathbf{0}$ and can be replaced with any positive constant. In both examples above, the constraint matrices can be shown to be indefinite and thus (2) is non-convex. We focus our attention on this case.

Assuming (1) is feasible, there is a tight connection between feasible points and global minimizers of (2). Indeed, a feasible point can be scaled to have an objective value of 0, a global minimum. Furthermore, any global minimizer corresponds to a feasible point. Thus our goal is to find a global minimum of the objective in (2).

We propose to solve (2) with a first order method (FOM), like stochastic gradient descent (SGD). FOMs are essentially the only option in big data scenarios due to low memory and computation requirements. Although FOMs are computationally advantageous, they can converge to non-global, local minimizers for non-convex problems. In general the landscape of local and global minimizers of non-convex functions can be very complex, but a heightened interest in machine learning has lead to a flurry of activity showing several non-convex problems for which all local minima are global. Examples include matrix completion [4] and Burer-Monteiro factorization for semidefinite programming [5]. In these cases, a FOM can successfully avoid saddle points and so converges to global minima [6].

To the best of our knowledge, the local minimizers of the objective in (2) have not been studied extensively making it unclear whether a FOM applied to (2) finds a solution to (1). We hope to close that gap theoretically and experimentally. We point out that [7] recently also proposed a similar method applying SGD to a smoothed version of (2) that shows promising empirical results. However, they do not prove any results about the existence of non-global, local minima nor provide any assumptions regarding the success of recovering a feasible point of (1) by applying a FOM to (2).

Our Contributions:

- Assuming all P_i are trace 0 and share a feasible point, we give necessary conditions for a point to be a local minimum of the objective of (2); see Theorem 1.
- In $\mathbb{R}^2$ under suitable assumptions, we show the objective of (2) has no local minima; see Theorem 4.
- Finally we provide experiments showing the success of a FOM applied to (2) for solving (1).

Remark: We point out that the formulation of (2) has been used in the specific case of ordinal embedding and collaborative ranking. For example, see [8, 9, 10, 11]. In both of these applications, extensive work has been done on bounding the sample complexity and determining the uniqueness of an embedding [12, 13, 11, 14, 15], but little work has been done on theoretically understanding the proposed non-convex optimization problems and methods used to solve them.

2. THEORY

2.1. Necessary Conditions for Minimizers

The following theorem gives necessary conditions for a point to be a non-global, local minimizer of optimization problem (2). The matrices arising in collaborative ranking are trace 0 because each constraint depends only on inner products between columns of W and rows of U, so each entry of the diagonal of the matrix that corresponds to each constraint is 0. Similarly, the ordinal embedding constraints result in trace 0 matrices [12]. Hence we restrict to trace 0 matrices. We say a set of matrices $\{P_i\}$ have a feasible point or share a feasible point if there is an x so that $x^T P_i x > 0$ for all i.

Theorem 1. *Let $\{P_i \in \mathbb{R}^{n\times n}\}$ be a set of real, symmetric trace 0 matrices that share a feasible point. Assume x is not a global minimizer of* (2)*. If $x \in \mathbb{R}^n$ is a non-global, local minimizer of* (2)*, x must satisfy the following two equations:*

P1) $\sum_{\{i:x^T P_i x<1\}} x^T P_i x < 0$

P2) $\sum_{\{i:x^T P_i x<1\}} x^T P_i x + \sum_{\{i:x^T P_i x=1\}} x^T P_i x \geq 0.$

In particular, $\{i : x^T P_i x = 1\} \neq \varnothing$.

Proof (Sketch of Proof). First we set some notation. Let $L(x)$ be the objective of optimization problem (2). Consider the partition of the constraints at x given by $I_x^{=1} := \{i : x^T P_i x = 1\}$ with $I_x^{>1}$ and $I_x^{<1}$ defined similarly. Therefore, $L(x) = |I_x^{<1}| - x^T P_x^{<1} x$ where $P_x^{<1} := \sum_{i\in I_x^{<1}} P_i$ and $P_x^{=1} := \sum_{i\in I_x^{=1}} P_i$.

If P1 or P2 is not true at some x' that is not a global minimizer, we claim x' cannot be a local minimizer by finding x arbitrarily close to x' with $L(x) < L(x')$.

First, assume P1 is not true. Take u to be a unit eigenvector of $P_{x'}^{<1}$ with positive eigenvalue λ, guaranteed since x' is not a global minimizer, the trace 0 condition, and feasibility. We can also assume $x'^T u \geq 0$.

Let $v_{\delta,\epsilon} = \epsilon x' + \delta u$ and $x = x' + v_{\delta,\epsilon}$. If $j \in I_{x'}^{=1}$, by choosing ϵ, δ sufficiently small, $x^T P_j x = (1+\epsilon)^2 + (1+\epsilon)\delta u^T P_j x' + \delta^2 u^T P_j u > 1$, and x is sufficiently close to x'. This implies $I_{x'}^{>1} \cup I_{x'}^{=1} = I_{x'+v_{\delta,\epsilon}}^{>1}$ and as a result $I_x^{<1} = I_{x'}^{<1}$. Thus since P_1 is not true and since $x'^T u \geq 0$,

$$\begin{aligned} &L(x) \\ &= |I_x^{<1}| - x^T P_x^{<1} x \\ &= |I_{x'}^{<1}| - x^T P_{x'}^{<1} x \\ &= |I_{x'}^{<1}| - (1+\epsilon)^2 x'^T P_{x'}^{<1} x' - \delta\lambda(2(1+\epsilon)x'^T u + \delta) \\ &< L(x'). \end{aligned}$$

In the second case, we assume P2 is not true. Consider $x = (1-\epsilon)x'$ for $\epsilon > 0$. For ϵ sufficiently small $I_{x'}^{<1} \subseteq I_x^{<1}$ and $I_{x'}^{>1} \subseteq I_x^{>1}$. If $j \in I_{x'}^{=1}$, then $x^T P_j x = (1-\epsilon)^2 x'^T P_j x' < 1$, so $I_{x'}^{=1} \subseteq I_x^{<1}$ and as a result $I_x^{<1} = I_{x'}^{<1} \cup I_{x'}^{=1}$. Then

$$\begin{aligned} L(x) &= |I_x^{<1}| - x^T P_x^{<1} x \\ &= |I_{x'}^{<1}| + |I_{x'}^{=1}| - (1-\epsilon)^2 x'^T \left(P_{x'}^{<1} + P_{x'}^{=1}\right) x' \\ &< |I_{x'}^{<1}| + |I_{x'}^{=1}| - \left(x'^T P_{x'}^{<1} x' + x'^T P_{x'}^{=1} x'\right) \\ &= L(x'), \end{aligned}$$

where P2 not being true implies the second to last line. □

2.2. Two Dimensions

In this section, for trace 0 constraint matrices in $\mathbb{R}^2$ sharing a feasible point, we show the objective of (2) has no local minima. In the case of homogeneous quadratic equations in $\mathbb{R}^2$, there is a simple algorithm for finding a feasible point. Since each quadratic equation $x^T P_i x$ is homogeneous, there is an interval $I_i \subset \mathbb{R}$ so that the line $y = \alpha x, \alpha \in I_i$, is contained in the cone $x^T P_i x > 0$. Thus the quadratic equations share a feasible point if $\cap_i I_i \neq \varnothing$. However, this algorithm does not generalize to higher dimensions unlike solving (2). We hope that our results in $\mathbb{R}^2$ generalize to higher dimensions.

Lemma 2. *Assume $A, B \in \mathbb{R}^{2\times 2}$ are linearly independent, trace 0 matrices. At any point x' on the curve $x^T B x = 1$, there is a tangent direction of $x^T B x = 1$ at x' which is a descent direction for $a - x^T A x$ at x' where $a \in \mathbb{R}$ is a constant.*

Proof. By the method of Lagrange multipliers, if $x' \in \mathbb{R}^2$ is a local minimizer of $a - x^T A x$ subject to $x^T B x = 1$, there exists $\lambda \in \mathbb{R}$ such that $Ax' = \lambda B x'$. However, since $\text{tr}(A - \lambda B) = 0$ and they are independent, $A - \lambda B$ is invertible so no such x' or λ can exist. □

Lemma 3. *Assume $P_1, P_2, P_3 \in \mathbb{R}^{2\times 2}$ are trace 0, pairwise independent matrices sharing a feasible point. Assume for some x', $x'^T P_1 x' = x'^T P_2 x' = 1$ and $x'^T P_3 x' < 0$. Then at x', there is a tangent direction of $x^T P_1 x$ (respectively $x^T P_2 x$) which is an ascent direction of $x^T P_2 x = 1$ (respectively $x^T P_1 x$) and a descent direction for $A - x^T P_3 x$, where $A \in \mathbb{R}$ is a constant.*

Proof (Sketch of Proof.). By an orthogonal rotation and assuming the trace 0 condition, $P_3 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$. Similarly let $P_1 = \left(\begin{smallmatrix} a & c \\ c & -a \end{smallmatrix}\right)$ and $P_2 = \left(\begin{smallmatrix} b & d \\ d & -b \end{smallmatrix}\right)$. For $i \in [1,2]$, let $f_i(x) = x^T P_i x$ and $f_3(x) = A - x^T P_3 x$. Since P_1, P_2 are independent of P_3, $c, d \neq 0$. In addition a computation shows if $cd < 0$, there cannot be a feasible point. Thus, $cd > 0$.

Let $U = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$. Then a tangent vector to the curve $x^T P_i x = 1$ at x' is $UP_i x'$. A computation shows

$$\langle \nabla f_2(x'), UP_1 x' \rangle = -2(ad - bc)\|x'\|_2^2, \tag{3}$$

$$\langle \nabla f_1(x'), UP_2 x' \rangle = 2(ad - bc)\|x'\|_2^2. \tag{4}$$

Likewise

$$\langle \nabla f_3(x'), UP_1 x' \rangle = -2c\|x'\|_2^2, \tag{5}$$

$$\langle \nabla f_3(x'), UP_2 x' \rangle = -2d\|x'\|_2^2. \tag{6}$$

WLOG, assume $c, d > 0$. Since $\langle \nabla f_2(x'), UP_1 x' \rangle$ and $\langle \nabla f_1(x'), UP_2 x' \rangle$ have opposite signs, one is nonnegative. WLOG say $\langle \nabla f_1(x'), UP_2 x' \rangle \geq 0$, so $UP_2 x'$ is an ascent direction of f_1 restricted to $x^T P_2 x = 1$.

Because $c, d > 0$, (6) is negative, so $UP_2 x'$ is also a descent direction for f_3. Therefore, at x', as we move along the curve $x^T P_2 x = 1$, in the tangent direction $UP_2 x'$, $x^T P_1 x$ increases by (4) and $A - x^T P_3 x$ decreases by (6). If $c, d < 0$, then the same argument applies but with the tangent vector $-UP_i x'$. □

Theorem 4 (Arbitrary Number of Constraints)**.** *Let $\{P_i \in \mathbb{R}^{2\times 2}\}$ be real, symmetric, trace zero matrices satisfying the conditions of Theorem 1. Additionally assume no three of the curves $x^T P_i x = 1$ intersect at a point. Then every local minimizer of the objective of* (2) *is a global minimizer.*

Proof. For $x \in \mathbb{R}^2$, let $I_x^{=1}, I_x^{<1}$, $I_x^{>1}$, and $P_x^{<1}$ be as defined in the proof of Theorem 1. By contradiction, suppose $\hat{z} \in \mathbb{R}^2$ is a non-global, local minimizer of objective (2). By Theorem 1, $\hat{z}^T P_{\hat{z}}^{<1} \hat{z} < 0$ and $1 \leq |I_{\hat{z}}^{=1}| \leq 2$, where the upper bound follows since at most two of the $x^T P_i x = 1$ intersect. We will now break into cases depending on the size of $I_{\hat{z}}^{=1}$. Recall that $L(x) = |I_x^{<1}| - x^T P_x^{<1} x$.

First, assume $|I_{\hat{z}}^{=1}| = 1$, so WLOG, $I_{\hat{z}}^{=1} = \{1\}$. Assume P_1 and $P_{\hat{z}}^{<1}$ are linearly independent. In this case,

Lemma 2 shows that there is a direction to move along the curve $x^T P_1 x = 1$ from $\hat{z}$ such that $L(x)$ decreases. If P_1 and $P_{\hat{z}}^{<1}$ are linearly dependent, then $\lambda P_1 = P_{\hat{z}}^{<1}$ for some λ; a feasible point for all the P_i imply $\lambda > 0$. However, $\lambda = \lambda \hat{z}^T P_1 \hat{z} = \hat{z}^T P_{\hat{z}}^{<1} \hat{z} < 0$, a contradiction. Thus, P_1 and $P_{\hat{z}}^{<1}$ must be linearly independent.

Second, assume $|I_{\hat{z}}^{=1}| = 2$ and WLOG, $I_{\hat{z}}^{=1} = \{1, 2\}$. If P_1, P_2 and $P_{\hat{z}}^{<1}$ are pairwise linearly independent, an identical argument as above now follows from Lemma 3. Now assume P_1, P_2 and $P_{\hat{z}}^{<1}$ are not pairwise independent. Since $|I_{\hat{z}}^{=1}| = 2$, $P_1 \neq \lambda P_2$ for any $\lambda \in \mathbb{R}$. Now, if $P_1 = \lambda P_{\hat{z}}^{<1}$ or $P_2 = \lambda P_{\hat{z}}^{<1}$, we repeat the argument from the case when $|I_{\hat{z}}^{=1}| = 1$. Therefore, P_1, P_2, and $P^{<1}$ are pairwise independent. □

2.3. Importance of Assumptions

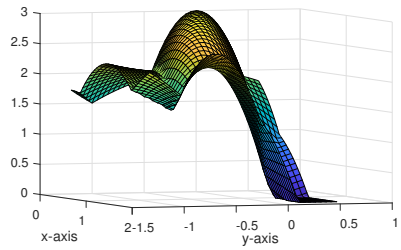

Fig. 1. Existence of non-global, local minimum of objective of (2) when trace 0 assumptions are not satisfied.

The trace 0 assumption of Theorem 4 is necessary. Otherwise, consider $P_1 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -.5 \end{smallmatrix}\right), P_2 = \left(\begin{smallmatrix} .5 & 1 \\ 1 & 1 \end{smallmatrix}\right), P_3 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 5 \end{smallmatrix}\right)$, which share a feasible point: $[1, 1]^T$. Figure 1 shows that $x \approx [1.1, -.7]$ is a non-global, local minimizer of the objective of (2) since the global minimum is 0. Therefore, proper initialization of FOMs and appropriate assumptions on the constraint matrices need to be more thoroughly studied to guarantee the success of solving (2) with FOMs.

3. EXPERIMENTS

In our experiments, we focus on validating SGD on (2) for finding feasible points. Due to the non-convexity of the problem, it seems to be challenging to determine how step size and initialization affect the success of a FOM like SGD. Therefore, we experiment with different step sizes and initializations at different scales. We remark that [7] contains an extensive set of experiments that validate using FOMs on a smoothed version of (2) to find feasible points. However, they did not consider different initializations.

The first experiment is in the case of ordinal embedding. To construct our constraints, we sampled a set of 50 points from $\mathcal{N}(0, I)$ in $\mathbb{R}^2$ and used all ordinal constraints arising from these points. To find a feasible embedding, we used SGD on objective (2). We varied the initial step size $(.001, .01, .1, .5)$ and the scale of the initialization, i.e., the initialization was sampled from $\mathcal{N}(0, \alpha I)$ for $\alpha = 1, 10, 100, \ldots, 10^6$. The step sizes decayed exponentially as $\frac{1}{2^t}$ where t is the number of epochs. Figure 2 shows the proportion of success over 20 experiments per choice of step size and initial scale, where a new set of points was sampled each time. SGD was given a budget of 8000 epochs.

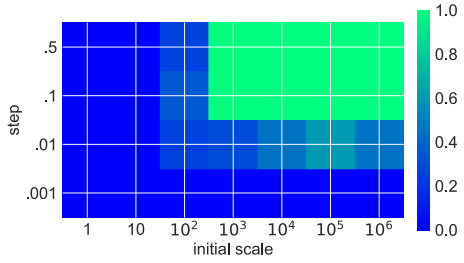

Fig. 2. Success of recovering a feasible embedding.

For the next experiment, we sampled 2000 symmetric matrices $\{P_i\}_{i \in [2000]} \subset \mathbb{R}^{20 \times 20}$ from $\mathcal{N}(0, I)$ and then projected them onto the subspace of trace 0 matrices. We picked a vector x and negated the P_i as needed so that $x^T P_i x > 0$ for all i ensuring feasibility. Initial step sizes and scalings were varied as in the previous experiment and exponentially decaying weights were used. SGD was given a budget of 4000 epochs. See Figure 3.

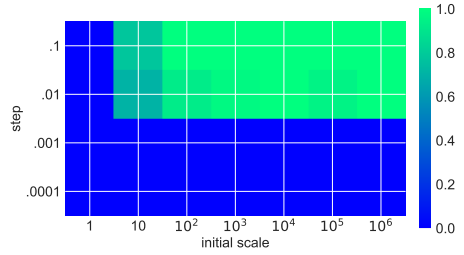

Fig. 3. Success of general quadratic feasibility in $\mathbb{R}^{20}$.

In both experiments, for a large enough initial step size and initialization, SGD reliably recovers a feasible point. Although not illustrated, SGD with small, constant step sizes produced similar results. Interestingly, initialization seems to play a large role in success of SGD in both of the above experiments.

4. REFERENCES

[1] Roger N Shepard, "The analysis of proximities: multidimensional scaling with an unknown distance function. i.," *Psychometrika*, vol. 27, no. 2, pp. 125–140, 1962.

[2] Joseph B Kruskal, "Nonmetric multidimensional scaling: a numerical method," *Psychometrika*, vol. 29, no. 2, pp. 115–129, 1964.

[3] Jaehyun Park and Stephen Boyd, "General heuristics for nonconvex quadratically constrained quadratic programming," *arXiv preprint*, 2017.

[4] Rong Ge, Jason D Lee, and Tengyu Ma, "Matrix completion has no spurious local minimum," in *Advances in Neural Information Processing Systems*, 2016, pp. 2973–2981.

[5] Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira, "The non-convex Burer-Monteiro approach works on smooth semidefinite programs," in *Advances in Neural Information Processing Systems*, 2016, pp. 2757–2765.

[6] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht, "Gradient descent only converges to minimizers," in *Conference on Learning Theory*, 2016, pp. 1246–1257.

[7] Aritra Konar and Nicholas D. Sidiropoulos, "First-order methods for fast feasibility pursuit of non-convex QCQPs," *IEEE Transactions on Signal Processing*, vol. 65, no. 22, pp. 5927–5941, 11 2017.

[8] Yoshikazu Terada and Ulrike Von Luxburg, "Local ordinal embedding," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. 2014, ICML'14, pp. II–847–II–855, JMLR.org.

[9] 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*. IEEE, 2011, pp. 1077–1084.

[10] Sameer Agarwal, Josh Wills, Lawrence Cayton, Gert Lanckriet, David Kriegman, and Serge Belongie, "Generalized non-metric multidimensional scaling," in *Artificial Intelligence and Statistics*, 2007, pp. 11–18.

[11] Dohyung Park, Joe Neeman, Jin Zhang, Sujay Sanghavi, and Inderjit Dhillon, "Preference completion: Large-scale collaborative ranking from pairwise comparisons," in *Proceedings of the 32nd International Conference on Machine Learning*, Francis Bach and David Blei, Eds., Lille, France, 07–09 Jul 2015, vol. 37 of *Proceedings of Machine Learning Research*, pp. 1907–1916, PMLR.

[12] Lalit Jain, Kevin G Jamieson, and Rob Nowak, "Finite sample prediction and recovery bounds for ordinal embedding," in *Advances in Neural Information Processing Systems*, 2016, pp. 2711–2719.

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

[14] Sewoong Oh, Kiran K Thekumparampil, and Jiaming Xu, "Collaboratively learning preferences from ordinal data," in *Advances in Neural Information Processing Systems*, 2015, pp. 1909–1917.

[15] Ery Arias-Castro et al., "Some theory for ordinal embedding," *Bernoulli*, vol. 23, no. 3, pp. 1663–1693, 2017.