Dealing with Intransitivity, Non-Convexity, and Algorithmic Bias in Preference Learning

by

Amanda Bower

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Applied and Interdisciplinary Mathematics)
in the University of Michigan
2020

Doctoral Committee:

Associate Professor Laura Balzano, Co-Chair
Professor Martin Strauss, Co-Chair
Professor Alexander Barvinok
Assistant Professor Yuekai Sun

Amanda Bower
amandarg@umich.edu
ORCID iD: 0000-0002-4497-3088

© Amanda Bower 2020

Acknowledgments

I believe that the outcomes of my life are mostly a function of randomness–starting from where I was born. Although hard work and perseverance plays a necessary factor, I arguably would never have started and completed this thesis without the support of the people that I have fortunately crossed paths with.

First and foremost, I thank my co-advisors Laura Balzano and Martin Strauss. Laura is an incredible researcher and supportive mentor who has helped me navigate the technical and non-technical aspects of doing research. I could go on and on, but in short, she showed me how to approach the boundaries of research with courage and optimism, believed in me even when I did not, afforded me unwavering patience, and gave me intellectual freedom. Martin has also been a supportive mentor to me and always available to give me feedback. He is the reason that I got interested in algorithmic fairness back when all the researchers in this field could fit into a small room.

I also thank all my collaborators, especially Yuekai Sun, Mikhail Yurochkin, and Lalit Jain. Yuekai has been a quasi-advisor to me. I admire his ability to see through research problems and the passion with which he tackles problems. I have learned a lot from Mikhail about how to approach research problems and about how to do solid empirical work. I am amazed by the ease with which he solves these problems. Lalit has always been just a phone call away, which is much appreciated. He helped me immensely early on by both shaping how I view research and taking the time to set me up for success with the practical aspects of machine learning research.

I have been fortunate to have awesome friends and peers in grad school. I especially thank Shelby Heinecke and Farrah Yhee for always being there for me. Shelby has always been a few steps ahead of me, which means that she warns me about what to watch out for next and gives me invaluable practical advice. She has shaped the way that I approach work, pushing me to be more effective, efficient, and happy. I am so grateful that we reconnected at that MSRI workshop. If COVID-19 ever goes away, we plan to hike Machu Picchu, and hopefully she and other hikers won't have to carry me up. Farrah has truly believed in me when I needed it the most. I admire her passion for social justice and her actions she has taken towards this goal, doing more work and exhibiting more courage than anyone I know. This passion pushes me to be a better algorithmic fairness researcher. I also appreciate our mutual love for good food even though she got me addicted to Soma raspberry bars. We plan make it to Noma in Copenhagen one day. I also thank Laura Niss and Alex Vargo for their friendship and collaborations. We literally navigated grad school together side-by-side from nearly the beginning, and I value all our long conversations about life and research. I thank Dejiao Zhang for her support, analytical eye, and always pushing me to be a better version of myself. I have always and continue to look up to her as a research role model, and I hope one day we can go to China. I thank my other close math friends, Umang Varma, Feng Zhu, and Anthony Della Pella, and the SPADA lab: Alex Ritchie, Yutong Wang, Kyle Gilman, Dejiao Zhang, Haroon Raja, Zhu Du, David Hong, Greg Ongie, John Lipor, and Ali Soltani-Tehrani. I especially thank any of my aforementioned friends who have let me crash on their bed or couch. I thank Takumi Murayama for this thesis template.

I interned at Netflix, and I thank everyone who made that summer great. In particular, I thank my mentors Dave Kale and Melody Dye and my manager Ashish Rastogi who all gave me the freedom to explore algorithmic fairness at Netflix. I learned a lot from them about Hollywood, working in industry, and how to apply machine learning in practice. I also admire their passion for social justice. I thank Kelly Uphoff for being an incredible role model and for making the internship program possible. I thank the other interns, Jeeheh Oh, Gayatri Bhat, Ying-Chun Li, and Claudia Roberts, for their friendship and support. Claudia, we will go to Berlin one day and dance all night.

Finally, I thank my family and closest friends for their love and support. My parents have always believed in and trusted my abilities more than I ever have. I thank my mom for being brave and working hard to move from the Philippines to the US, which was a huge jump from poverty to give me a better life than she had. I thank my dad for always respecting my views, valuing an education, and encouraging me to always pursue what makes me happy. I thank Adam's parents for supporting me throughout school and the countless drives to Ann Arbor. To Sam, Megan, Adam, and Eirik: I do not even need to say anything since you all know how much you and your support mean to me. Finally, I thank my cats for being my biggest stress relievers. Their cuddles are much appreciated.

Table of Contents

List of Figures

List of Tables

Abstract

Rankings are ubiquitous since they are a natural way to present information to people who are making decisions. There are seemingly countless scenarios where rankings arise, such as deciding whom to hire at a company, determining what movies to watch, purchasing products, understanding human perception, judging science fair projects, voting for political candidates, and so on. In many of these scenarios, the number of items in consideration is prohibitively large, such that asking someone to rank all of the choices is essentially impossible. On the other hand, collecting preference data on a small subset of the items is feasible, e.g., collecting answers to “Do you prefer item A or item B?” or “Is item A closer to item B or item C?”. Therefore, an important machine learning task is to learn a ranking of the items based on this preference data. This thesis theoretically and empirically addresses three key challenges of preference learning: intransitivity in preference data, non-convex optimization, and algorithmic bias.

Chapter 2 addresses the challenge of learning a ranking given pairwise comparison data that violate rational choice assumptions such as transitivity. Our key observation is that two items compared in isolation from other items may be compared based only on a salient subset of features. Formalizing this framework, we propose the *salient feature preference model* and prove a sample complexity result for learning the parameters of our model and the underlying ranking with maximum likelihood estimation.

Chapter 3 addresses the non-convexity of a class of optimization problems that find feasible points to a set of quadratic inequalities. This class contains the ordinal embedding problem, which is a preference learning task. We aim to understand the local minimizers and global minimizers of the non-convex objective, which corresponds to penalizing each violated quadratic inequality with the hinge loss. Under certain assumptions, we give necessary conditions for non-global, local minimizers of the objective and additionally show that in two dimensions, every local minimizer is a global minimizer.

Chapters 4 and 5 address the challenge of algorithmic bias. We consider training

machine learning models that are fair in the sense that their performance is invariant under certain sensitive perturbations to the inputs. For example, the performance of a résumé screening system should be invariant under changes to the gender and ethnicity of the applicant. We formalize this notion of algorithmic fairness as a variant of individual fairness. In Chapter 4, we consider classification and develop a distributionally robust optimization approach, *SenSR*, that enforces this notion of individual fairness during training and provably learns individually fair classifiers.

Chapter 5 builds upon Chapter 4. We develop a related algorithm, *SenSTIR*, to train provably individually fair learning-to-rank (LTR) models. The proposed approach ensures items from minority groups appear alongside similar items from majority groups. This notion of fair ranking is based on the individual fairness definition considered in Chapter 4 for the classification context and is more nuanced than prior fair LTR approaches that simply provide underrepresented items with a basic level of exposure. The crux of our method is an optimal transport-based regularizer that enforces individual fairness, and we provide an efficient algorithm for optimizing the regularizer.

Chapter 1

Introduction

Faced with a set of choices, people typically use rankings to facilitate decision-making. This scenario is ubiquitous: search engines rank webpages given a query, recommender systems rank products for purchase or movies for entertainment, universities and companies rank applicants, and in some jurisdictions in the United States and other countries, voters rank political candidates and vote via these rankings, which is known as "ranked choice voting." In many of these applications, the number of choices is prohibitively large for a person to manually rank, so data-driven techniques are employed to rank the choices. However, there are several challenges encountered when utilizing these data-driven approaches. This thesis in particular addresses three key challenges in preference learning: intransitivity in preference data, non-convexity of preference models, and algorithmic biases of ranking models and data.

1.1 Intransitivity

The first challenge of preference learning this thesis considers is intransitivity in preference data. We specifically consider the scenario where there is a set of n items and one unknown ranking of these items. Although collecting full ranking data from humans is prohibitive, it is relatively easy for people to answer pairwise comparison questions, i.e., questions of the form "Is item A better than item B?". Then, the underlying ranking can be efficiently estimated by aggregating this pairwise comparison data. For example, if the pairwise comparison data is noiseless, ranking the items is equivalent to sorting a list. Some sorting algorithms like merge sort can sort a list with $O(n \log n)$ pairwise comparisons [CLRS09], which is significantly fewer than all $O(n^2)$ unique pairwise comparisons.

There is one implicit but crucial assumption in this discussion: we assumed that pairwise comparison data is consistent with the underlying ranking of all of the items. In other words, we assumed the decision-making processes people use to rank all of the items at once are the same decision-making processes people use to answer pairwise comparison questions. In fact, many ranking models and algorithms, like the Bradley-Terry-Luce model [BT52, Luc59] and ranking SVM [Joa02], make this assumption. If pairwise comparison data is consistent with the underlying ranking, then the pairwise comparison data cannot contain intransitivity. That is, there cannot be three items A, B, C such that on average item A is preferred to item B, item B is preferred to item C, but item C is preferred to item A. This example contradicts transitivity, since transitivity of the choices implies item A should instead be preferred to item C.

However, we argue–just as researchers in social science [She64, Tor65, Tve77, Tve72, BGS13] and recently in machine learning have [SPU19, ROS20, HSR+19, PGH19, KMU17, SW17, RU16, NR17, BKT16, CJ16b, CJ16a, RGLA15, YBW15, Agr12]–that intransitivity is a prevalent characteristic of real preference data. For illustration, see Table 1.1. Let $P_{A,B}$ be the empirical probability that item A beats item B in a pairwise comparison. The first column "Valid Triplets" refers to $|\{(A,B,C) : P_{A,B} \geq \frac{1}{2} \text{ and } P_{B,C} \geq \frac{1}{2}\}|$, and the number in parenthesis in the last three columns is the fraction of "Valid Triples" that violate one of three stochastic transitivity properties. The last three columns of the table correspond to different types of transitivity violations. Three items A, B, and C such that $P_{A,B} \geq \frac{1}{2}$ and $P_{B,C} \geq \frac{1}{2}$ violate *weak stochastic transitivity* if $P_{A,C} < \frac{1}{2}$, violate *moderate stochastic transitivity* if $P_{A,C} < \min\{P_{A,B}, P_{B,C}\}$, and violate *strong stochastic transitivity* if $P_{A,C} < \max\{P_{A,B}, P_{B,C}\}$. Clearly, each data set contains a significant amount of stochastic transitivity violations. This is problematic for the Bradley-Terry-Luce model [BT52, Luc59]–arguably one of the most popular and widely used ranking models–since it assumes the pairwise comparison data does not violate even strong stochastic transitivity.

Although there are several different reasons for why intransitivity can arise [RBM06] including heterogeneous preferences, in Chapter 2, we attribute intransitivity to pairwise contextual effects since each pairwise comparison asks for a human judgement about two items in isolation of all the other items. Our model is inspired by theories in social science [Tve72, TS93, RBM06, BP09, She64, Tor65, Tve77, BGS13, KKK17]. For

Data Set	Valid Triplets	Strong Violations	Moderate Violations	Weak Violations
NBA 2015 [Kel20]	2654	1439 (54%)	1185 (45%)	272 (10%)
Tennis 2014 [Gob20]	4793	1092 (23%)	1080 (23%)	651 (14%)
Nascar [GS09]	65003	26354 (41%)	17128 (26%)	4171 (6%)
Jester [GRGP01]	161700	14560 (9%)	327 (.2%)	78 (.05%)
Sushi-A [KA09]	120	28 (23%)	0 (0%)	0 (0%)
Sushi-B [KA09]	139992	66013 (47%)	26366 (19%)	4939 (4%)
District [KKK17]	48	25 (52%)	8 (16%)	0 (0%)
Car [ASBP13]	120	46 (38%)	7 (6%)	0 (0%)
Sonancia [LLY17]	874	175 (20%)	175 (20%)	108 (13%)
New Yorker [Sie20]	3990	1823 (51%)	606 (17%)	199 (6%)

Table 1.1: Intransitivity is a prevalent characteristic of real preference data.

example, the authors in [KKK17] wanted to obtain a ranking of legislative districts in the United States from most compact to least compact in order to better understand human perception of compactness. They attempted to use a pairwise comparison approach but deemed the pairwise comparison data unreliable potentially because a pairwise "approach enables respondents to make each paired comparison independently of the others, and may even encourage, them to use different dimensions for different comparisons" [KKK17].

Inspired by this idea for why intransitivity arises, we propose the *salient feature preference model*, which reconciles intransitive pairwise preferences with a global ranking of the items. Specifically, we posit that for each pair of items, there is a potentially different subset of salient features that stand out to people. For each pair, these salient features are the only features taken into consideration when answering the corresponding pairwise comparison question. On the other hand, our model assumes that if a person could view all of the items at once, they would consider all the features when deciding how to rank the items. Therefore, in our model, pairwise contextual effects due to which two items are being compared prevent the underlying ranking from being perfectly reflected in pairwise comparison data. In Chapter 2, we study the statistical properties of the maximum likelihood estimator of our model, and we demonstrate strong performance of our model and algorithm on real preference data that contain intransitive preferences.

1.2 Non-Convexity

The second challenge of preference learning this thesis addresses is non-convex optimization. In order to turn data into an actionable model, we need to find a model that fits the data as well as possible. Typically, this requires solving an optimization problem. When the resulting optimization problem is convex, first order methods like stochastic gradient descent are guaranteed to find an optimal model. On the other hand, stochastic gradient descent can get stuck in a local optimum instead of finding a global optimum when the optimization problem is non-convex.

The aforementioned salient feature preference model considered in Chapter 2 turns out to be convex. However, other preference models, like the ordinal embedding model, result in non-convex optimization problems. In the ordinal embedding model, we assume there is a set of items such that each item has an unknown low-dimensional representation, which we would like to estimate. We collect answers to questions of the form "Is item A closer to item B or item C?" and assume the answers to these questions are governed by the Euclidean distances between the low-dimensional representations of the items. Downstream applications of ordinal embedding include visualization and rankings, e.g., items can be ranked in order of their distances to a fixed item. Estimating the low-dimensional representation of each point can be written as a non-convex optimization problem.

To illustrate ordinal embedding, consider Figure 1.1, which shows the state capitals of the continental United States. Using data of the form "capital A is closer to capital B than capital C," we attempt to estimate the locations of the state capitals to fit this data as well as possible. Figure 1.2 shows the estimated location of each state capital where the estimates were obtained by solving a non-convex problem with stochastic gradient descent.

Interestingly, the estimated locations of the state capitals are perfectly consistent with the observed data, i.e., if capital A is truly closer to capital B than capital C in the observed data, then the estimated location of capital A is closer to the estimated location of capital B than the estimated location of capital C. In other words, despite solving a non-convex optimization problem, stochastic gradient descent finds a global optimum. This observation suggests that every local optima of this problem is a global optima since gradient descent does not get stuck in saddle points [LSJR16]. In fact,

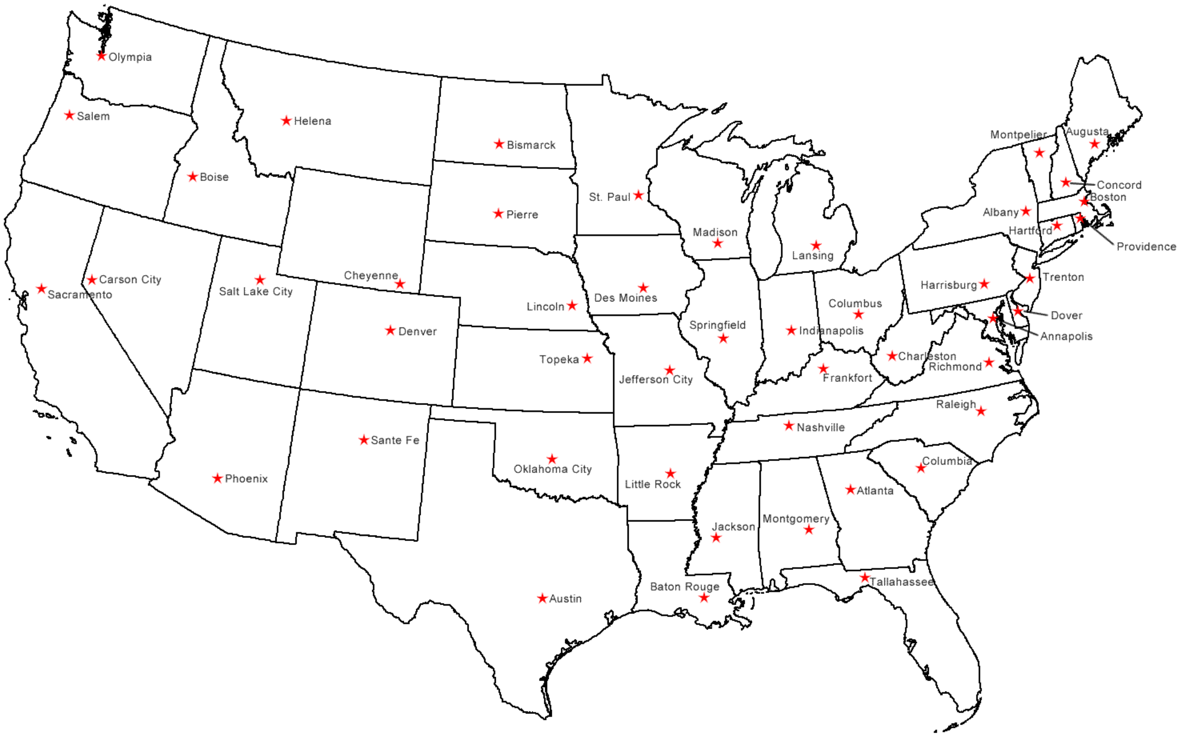


Figure 1.1: This map shows the true locations of the state capitals of the continental United States [Wat20].

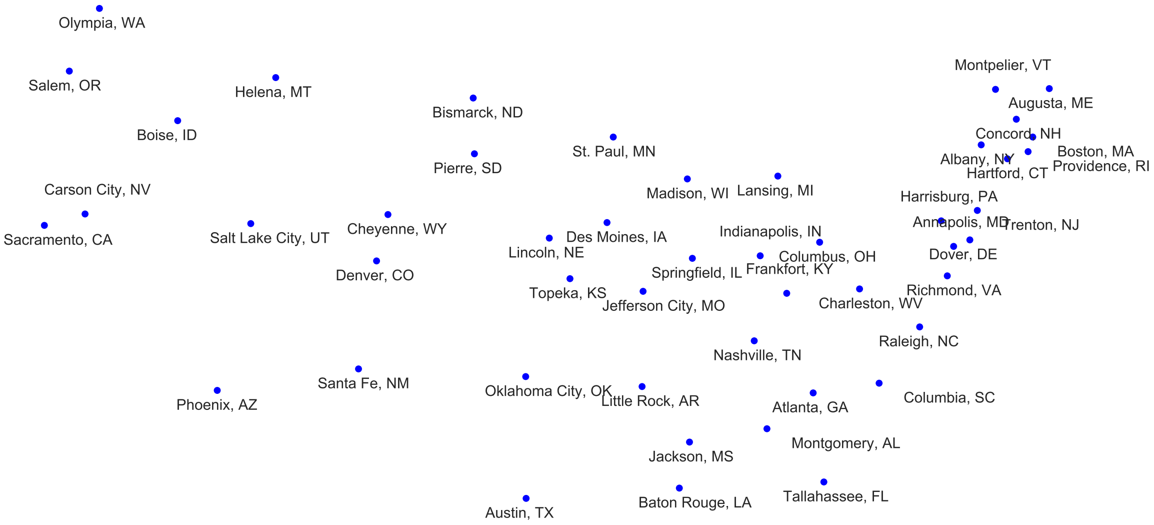


Figure 1.2: Using data of the form "capital A is closer to capital B than capital C," this figure shows the estimated locations of the state capitals of the continental United States, which are obtained by solving a non-convex problem with stochastic gradient descent.

there has been a flurry of recent work showing that all optima are global optima in many non-convex problems [GJZ17, BVB16]. Motivated by these findings and the empirical success of solving ordinal embedding problems, like we just have illustrated with the state capitals, in Chapter 3, we study the the local and global optima of the non-convex quadratic feasibility problem theoretically and empirically. This class of optimization problems includes the ordinal embedding problem and other preference learning problems as special cases.

1.3 Algorithmic Bias

The third challenge of preference learning this thesis addresses is bias in ranking models and ranking data. Algorithms touch several facets of our daily lives ranging from seemingly inconspicuous tasks like web search to high-stakes scenarios like access to employment. Alarmingly, it has been well-established that algorithms are not neutral since they can perpetuate or exacerbate existing biases due to reasons like historical discrimination and racism, underrepresentation of certain demographic groups, or poor data collection and algorithm design choices. In high-stakes domains such as access to financial services, access to employment, policing, and criminal justice, algorithms can have serious and grave consequences.

For example, algorithms in high-stakes settings can have gender biases. In the financial services domain, Apple credit card is under investigation by New York State regulators due to potential algorithmic gender biases. Several pairs of heterosexual married couples claim that although both partners have essentially the same data, e.g., same bank account, similar credit scores, etc., men were given substantially higher credit limits than women [Vig19]. In the employment domain, Amazon stopped using an internal résumé screening tool that was biased against women: "It penalized résumés that included the word 'women's,' as in 'women's chess club captain.' And it downgraded graduates of two all-women's colleges, according to people familiar with the matter" [Das18].

Furthermore, algorithms in high-stakes settings can have racial biases. In the policing domain, a Black man was recently wrongfully handcuffed and arrested at his home in front of his family, was held overnight in a detention center, took a mugshot, gave his DNA and fingerprints, and used a vacation day to appear in court for an arraignment all because a facial recognition system incorrectly identified him as a shoplifter [Hil20].

Although dependent on the specific tool, facial recognition tools are known to typically have significantly higher false positive rates for Black people than for Caucasian people [GNH19]. In the criminal justice domain, ProPublica showed that the false positive rate (respectively, false negative rates) of the COMPAS algorithm–used to predict whether or not someone will recommit a crime and taken into account by judges when considering sentencing or parole–are significantly higher (respectively lower) for Black people than for white people [ALMK16].

Biases in algorithms extend far beyond these striking examples. Therefore, as machine learning researchers, we have a responsibility to understand the ethical implications of our algorithms and understand how and the extent to which we can mitigate these biases. Although defining what algorithmic "fairness" means is still an active area of research especially in areas outside of classification like rankings, most definitions fall into either the "group fairness" category or "individual fairness" category.

Group fair definitions typically assume the data can be partitioned into demographic groups, e.g., women and men, and require a statistical quantity, like false positive rates or proportion of positive labels, to be equal over these demographic groups [HPPS16]. For example, ProPublica showed that the COMPAS algorithm violated group fairness: the false positive rate for Blacks is much higher than the false positive rate for whites [ALMK16]. However, group fairness has several deficiencies: although group fairness guarantees that individuals from different groups are treated the same on average, group fair algorithms provide no guarantees to individuals themselves, and algorithms can even be "Gerrymandered" to make an arguably unfair algorithm appear group fair [KNRW18]. To illustrate, we use the following example from [KNRW18]. Consider a classification task such that each person is a woman or a man and Black or white, and assume that among the four resulting demographic groups, the number of people in each group is equal. A classifier that always gives a negative label to Black women and white men and a positive label to Black men and white women is both gender group fair and race group fair, but it is clearly unfair to Black women and white men since they always receive a negative label.

In contrast to group fairness, individual fairness requires similar individuals be treated similarly by an algorithm [DHP$^+$12]. For instance, the algorithm used to determine credit limits in the Apple credit example purportedly violates individual fairness: each husband and wife pair claim that despite having similar features, e.g., filing joint tax returns, owning the same exact assets, and having joint bank accounts, the credit limit given to the husband is substantially higher than the credit limit given to the wife. Individual fairness can be advantageous over group fairness. For example, individually fair algorithms are not susceptible to being "Gerrymandered" like group fair algorithms can be. However, despite being introduced nearly a decade ago, individual fairness has largely not been operationalized since defining the similarity between individuals, i.e., the fair metric, is non-trivial. In Chapter 4, we view individual fairness through the lens of distributional robustness, propose a model and algorithm to learn the fair metric from data, and propose an algorithm to learn an individually fair classifier. We study the statistical properties of our model as well as empirically demonstrate the efficacy of our model to mitigate biases on real data.

Until recently, the fairness of ranking systems has been given relatively little attention in comparison to classification, and furthermore, most of the work in fair rankings has focused on group fairness notions. In Chapter 5, we apply similar ideas as in Chapter 4 to propose an individually fair based definition for fair ranking systems and to propose an algorithm to learn individually fair ranking systems. We study the statistical properties of our model and illustrate that our model can mitigate biases on real data. Our proposed notion for individual fairness in ranking systems requires rankings to be stable with respect to certain perturbations of the features. To illustrate, see Figure 1.3. Suppose a job recruiter is searching for software engineers, and they are presented a ranking of potential job candidates, each of which is a man or a woman. Women are represented by relatively longer hair than men. Consider a counterfactual set of job candidates where, for sake of simplicity, the gender of each candidate is flipped. We require an individually fair ranking system to rank the original set of candidates and counterfactual set of candidates the same as illustrated on the right hand side of Figure 1.3. In contrast, the left hand side of Figure 1.3 shows an unfair ranking system that is biased against women. The man who was originally ranked first is now ranked significantly lower in third under the counterfactual ranking when he is regarded as a woman. Similarly, the women in the second and third positions are boosted up in the counterfactual ranking when they are regarded as men.

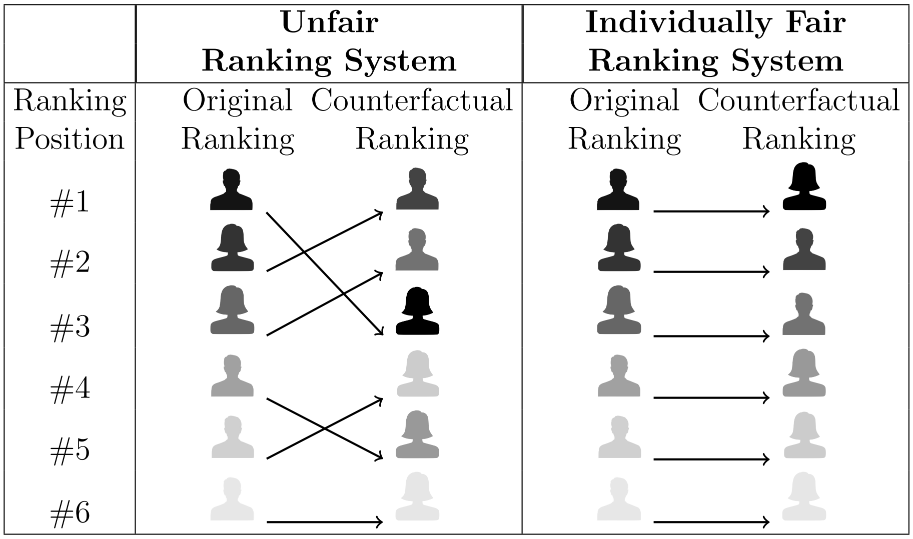


Figure 1.3: In this example, a job recruiter is searching for software engineers. Given the original set of job candidates, consider a counterfactual set of job candidates where each person's gender is flipped. The ranking system on the left hand side is biased against women since the counterfactual ranking changes substantially in favor of men, whereas the stable system on the right hand side is considered fair.

1.4 Publications

The following are my publications where $*$ indicates equal contribution.

- Chapter 2: Amanda Bower and Laura Balzano. “Preference Modeling with Context-Dependent Salient Features.” In ICML 2020.
- Chapter 3: Amanda Bower, Lalit Jain, Laura Balzano. “The Landscape of Non-Convex Quadratic Feasibility.” In ICASSP 2018.
- Chapter 4: Mikhail Yurochkin*, Amanda Bower*, and Yuekai Sun. “Training individually fair ML models with sensitive subspace robustness.” In ICLR 2020.
- Chapter 5: Amanda Bower, Hamid Eftekhari, Mikhail Yurochkin, and Yuekai Sun. “Individually Fair Rankings.” In submission.
- Other:
  - Amanda Bower*, Laura Niss*, Yuekai Sun*, and Alex Vargo*. “Debiasing Representations by Removing Unwanted Variation Due to Protected Attributes.” In FAT-ML workshop at ICML 2018.
  - Amanda Bower, Sarah Kitchen*, Laura Niss*, Martin Strauss*, Alex Vargo*, and Suresh Venkatasubramanian*. “Fair Pipelines.” In FAT-ML workshop at KDD 2017.

Chapter 2

Preference Modeling with Context-Specific Salient Features

The work in this chapter is joint with Laura Balzano. This work is published as *Preference Modeling with Context-Specific Salient Features* at ICML 2020.

2.1 Introduction

The problem of estimating a ranking is ubiquitous and has applications in a wide variety of areas such as recommender systems, review of scientific articles or proposals, search results, sports tournaments, and understanding human perception. Collecting full rankings of n items from human users is infeasible if the number of items n is large. Therefore, k-wise comparisons, $k < n$, are typically collected and aggregated instead. Pairwise comparisons ($k = 2$) are popular since it is believed that humans can easily and quickly answer these types of comparisons. However, it has been observed that data from k-wise comparisons for small k often exhibit what looks like irrational choice, such as systematic intransitivity among comparisons. Common models address this issue with modeling noise, ignoring its systematic nature. We observe, as others have before us [SPU19, ROS20, PGH19, KMU17, BKT16, CJ16b, CJ16a], that these systematic irrational behaviors can likely be better modeled as *rational behaviors made in context*, meaning that the particular k items used in a k-wise comparison will affect the comparison outcome.

Consider the most common model for learning a single ranking from pairwise comparisons, the Bradley-Terry-Luce (BTL) model. In this model, there exists a judgment vector $w^* \in \mathbb{R}^d$ that indicates the favorability of each of the d features of an item (e.g. for shoes: cost, width, material quality, etc), and each item has an embedding $U_i \in \mathbb{R}^d$, $i = 1, \ldots, n$, indicating the value of each feature for that given item. Subsequently, the outcome of a comparison is made with probability related to the inner product $\langle U_i, w^* \rangle$; the larger this inner product, the more likely item i will be ranked above other items to which it is compared. A key implicit assumption is that the features used to rank all n items are the same features used to rank just k items in the absence of the other $n-k$ items. However, we argue that the context of that particular pairwise comparison is also relevant; it is likely that when a pairwise comparison is collected, if there are a small number of features that "stand out," a person will use only these features and ignore the rest when he or she makes a comparison judgment. Otherwise, if there are no salient features between a pair of items, a person will take all features into consideration. This theory has been hypothesized by the social science community to explain violations of rational choice [Tve72, TS93, RBM06, BP09, She64, Tor65, Tve77, BGS13]. For example, [KKK17] collected preference data to understand human perception of the compactness of legislative districts. They hypothesized that the features respondents use in a pairwise comparison task to judge district compactness vary from pair to pair, which explains why their data are more reliable for larger k. To illustrate this point, we highlight a concrete example from their experiments. Given two images of districts, they asked respondents to pick which district is more compact. When comparing district A with district B or district C in Figure 2.1, one of the most salient features is the degree of nonconvexity. However, when comparing district B and district C, the degree of nonconvexity is no longer a salient feature. These districts look similar on many dimensions, forcing a person to really think and consider all the features before making a judgment. Let P_{ij} be the empirical probability that district i beats district j with respect to compactness. Then, from the experiments of [KKK17], we have $P_{AB} = 100\%$, $P_{BC} = 67\%$, and $P_{AC} = 70\%$. These three districts violate strong stochastic transitivity, the requirement that if $P_{AB} \geq 50\%$ and $P_{BC} \geq 50\%$, then $P_{AC} \geq \max\{P_{AB}, P_{BC}\}$.

We propose a novel probabilistic model called the *salient feature preference model* for pairwise comparisons such that the features used to compare two items are dependent on the context in which two items are being compared. The salient feature preference model is a variation of the standard Bradley-Terry-Luce model. At a high level, given a pair

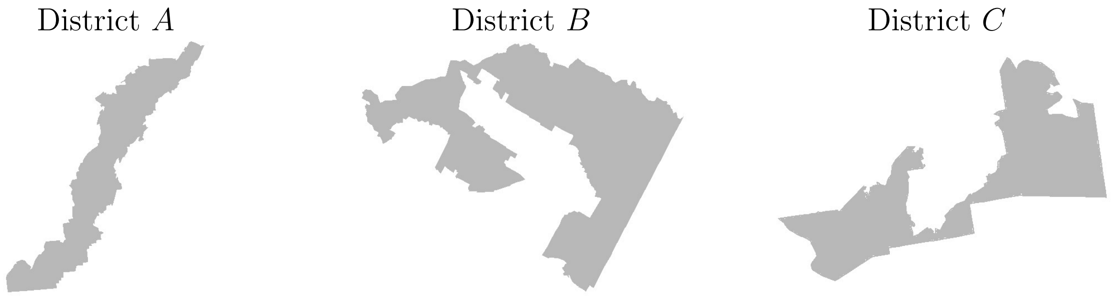


Figure 2.1: Three districts used in pairwise comparison tasks in [KKK17]

of items in $\mathbb{R}^d$, we posit that humans perform the pairwise comparison in a coordinate subspace of $\mathbb{R}^d$. The particular subspace depends on the salience of each feature of the pairs being compared. Crucially, if any human were able to rank all the items at once, he or she would compare the items in the ambient space without projection onto a smaller subspace. This single ranking in the ambient space is the ranking that we would like to estimate. Our contributions are threefold. First, we precisely formulate this model and derive the associated maximum likelihood estimator (MLE) where the log-likelihood is convex. Our model can result in intransitive preferences, despite the fact that comparisons are based off a single universal ranking. In addition, our model generalizes to unseen items and unseen pairs. Second, we then prove a necessary and sufficient identifiability condition for our model and finite sample complexity bounds for the MLE. Our result specializes to the sample complexity of the MLE for the BTL model with features, which to the best of our knowledge has not been provided in the literature. Third, we provide synthetic experiments that support our theoretical results and also illustrate scenarios where our salient feature preference model results in systematic intransitives. We also demonstrate the efficacy of our model and maximum likelihood estimation on real preference data about legislative district compactness and the `UT Zappos50K` data set.

2.1.1 Related Work

The Bradley-Terry-Luce Model One popular probabilistic model for pairwise comparisons is the Bradley-Terry-Luce (BTL) model [BT52, Luc59]. In this model, there are n items each with an unknown utility u_i for $i \in [n]$, and the items are ranked by sorting the utilities. The BTL model defines

$$\mathbb{P}(\text{item } i \text{ beats item } j) = \frac{e^{u_i}}{e^{u_i} + e^{u_j}}. \tag{2.1}$$

Although the BTL model makes strong parametric assumptions, it has been analyzed extensively by both the machine learning and social science community and has been applied in practice. For instance, the World Chess Federation has used a variation of the BTL model in the past for ranking chess players [MM08]. The sample complexity of learning the utilities or the ranking of the items with maximum likelihood estimation (MLE) has been studied recently in [RA14, NOS16]. Moreover, there is a recent line of work that analyzes the sample complexity of learning the utilities with MLE and other algorithms under several variations of the BTL model, including when the items have features that may or may not be known [LCF$^+$18, OTX15a, LN15a, PNZ$^+$15a, SR18, NR17]. Our model is also a variation of the BTL model where the utility of each item is dependent on the items it is being compared to.

Violations of Rational Choice The social science community has long recognized and hypothesized about irrational choice [She64, Tor65, Tve77, Tve72, BGS13]. See [RBM06] for an excellent survey of this area including references to social science experiments that demonstrate scenarios where humans make choices that can violate a variety of rational choice axioms such as transitivity. There has been recent progress in modeling and providing evidence for violations of rational choice axioms in the machine learning community [SPU19, ROS20, HSR$^+$19, PGH19, KMU17, SW17, RU16, NR17, BKT16, CJ16b, CJ16a, RGLA15, YBW15, Agr12]. In contrast to our work, none of these works model preference data that both violates rational choice and admits a universal ranking of the items with the exception of [SW17, HSR$^+$19]. Assuming there is a true ranking of the items, our model makes a direct connection between pairwise comparison data that violates rational choice and the underlying ranking. Violations of rational choice, including intransitivty, occur in our model because of contextual effects due to which pairs of items are being compared. These contextual effects distort the true ranking, whereas in the work of [SW17, HSR$^+$19] the intransitive choices define the ranking. Specifically, the items are ranked by sorting the items by the probability that an item beats any other item.

We now focus on the works most similar to ours. The work in [SPU19], which generalizes [CJ16b, CJ16a] from pairwise comparisons to k-wise comparisons, considers a model for context dependent comparisons. However, because they do not assume access to features, their model cannot predict choices based on new items, which is a key task for very large modern data sets. In contrast, our model can predict pairwise outcomes and rankings of new items. Both [ROS20] and [PGH19] assume access to features of items and propose learning contextual utilities with neural networks. In contrast, we propose a linear approach with typically far fewer parameters to estimate. Furthermore, the latter work does not contain any theory, whereas we prove a sample complexity result on estimating the parameters of our model. In all of the aforementioned works in this paragraph, the resulting optimization problems are non-convex with the exception of a special case in [SPU19] that requires sampling every pairwise comparison. In contrast, the negative log likelihood of our model is convex. Interestingly, the work in [MU19] shows that for a class of parametric models for pairwise preference probabilities, if intransitives exist, then the negative log likelihood cannot be convex. Our model does not belong to the class of parametric models they consider.

Notation For an integer $d > 0$, $[d] := \{1, \ldots, d\}$. For $x, y \in \mathbb{R}^d$, $\langle x, y \rangle := \sum_{i=1}^{d} x_i y_i$. For $x \in \mathbb{R}^d$ and $\Omega \subset [d]$, let $x^{\Omega} \in \mathbb{R}^d$ where $(x^{\Omega})_i = x_i$ if $i \in \Omega$ and 0 otherwise. For $i, j \in [n]$, "$i >_B j$" means "item i beats item j." Let $\mathcal{P}(X)$ be the power set of a set X. Given a set of vectors $S = \{x_i \in \mathbb{R}^d\}_{i=1}^{q}$, $\text{span}(S) = \{\sum_{i=1}^{q} \alpha_i x_i : \alpha_i \in \mathbb{R}\}$.

2.2 Model and Algorithm

Salient Feature Preference Model Suppose there are n items, and each item $j \in [n]$ has a known feature vector $U_j \in \mathbb{R}^d$. Let $U := \begin{bmatrix} U_1 U_2 \cdots U_n \end{bmatrix} \in \mathbb{R}^{d \times n}$. Let $w^* \in \mathbb{R}^d$ be the unknown *judgment weights*, which signify the importance of each feature when comparing items. Let $\tau : [n] \times [n] \to \mathcal{P}([d])$ be the known *selection function* that determines which features are used in each pairwise comparison. Let $P := \{(i, j) \in [n] \times [n] : i < j\}$ be the set of all pairs of items. Let $S_m = \{(i_\ell, j_\ell, y_\ell)\}_{\ell=1}^{m}$ be a set of m independent pairwise comparison samples where $(i_\ell, j_\ell) \in P$ are chosen uniformly at random from P with replacement, and $y_\ell \in \{0, 1\}$ indicates the outcome of the pairwise comparison where 1 indicates item i_ℓ beat item j_ℓ and 0 indicates item j_ℓ beat item i_ℓ. We model

$y_\ell \sim \text{Bern}(\mathbb{P}(i_\ell >_B j_\ell))$ where

$$\mathbb{P}(i_\ell >_B j_\ell) = \frac{\exp\left(\langle U_{i_\ell}^{\tau(i_\ell, j_\ell)}, w^* \rangle\right)}{\exp\left(\langle U_{j_\ell}^{\tau(i_\ell, j_\ell)}, w^* \rangle\right) + \exp\left(\langle U_{i_\ell}^{\tau(i_\ell, j_\ell)}, w^* \rangle\right)}. \tag{2.2}$$

To understand the probability model given by Equation (2.2), note that $\langle U_i^{\tau(i,j)}, w^* \rangle$ is the inner product of U_i and w^* after U_i is projected to the coordinate subspace given by $\tau(i,j)$. Therefore, Equation (2.2) is simply the utility model of Equation (2.1) where the utilities are inner products computed in the subspace defined by the selection function τ. If the selection function returns all the coordinates, i.e. $\tau(i,j) = [d]$, then Equation (2.2) becomes the standard BTL model where the utility of item i is $\langle U_i, w^* \rangle$ and fixed regardless of context, i.e., regardless of which pair is being compared. This model is typically called “BTL with features,” and we will refer to it as FBTL. See Section 2.6.1 in the Supplement for a natural extension of Equation (2.2) to k-wise comparisons for $k > 2$. Furthermore, we assume that the true ranking of all the items depends on all the features and is given by sorting the items by $\langle U_i, w^* \rangle$ for $i \in [n]$.

Selection Function We propose a selection function τ inspired by the social science literature, which posits that violations of rational choice axioms arise in certain scenarios because people make comparison judgments on a set of items based on the features that differentiate them the most [RBM06, BP09, BGS13].

For two variables $w, z \in \mathbb{R}$, let $\mu := (w+z)/2$ be their mean and $\bar{s} := ((w-\mu)^2 + (z-\mu)^2)/2$ be their sample variance. Given $t \in [d]$ and items $i, j \in [n]$, the *top-t selection function* selects the t coordinates with the t largest sample variances in the entries of the feature vectors U_i, U_j.

Algorithm: Maximum Likelihood Estimation Given observations $S_m = \{(i_\ell, j_\ell, y_\ell)\}_{\ell=1}^m$, item features $U \in \mathbb{R}^{d \times n}$, and a selection function τ, the negative log-likelihood of $w \in \mathbb{R}^d$ is

$$\mathcal{L}_m(w; U, S_m, \tau) = \sum_{\ell=1}^{m} \log\left(1 + \exp\left(u_{i_\ell, j_\ell}\right)\right) - y_\ell u_{i_\ell, j_\ell}, \tag{2.3}$$

where $u_{i_\ell,j_\ell} = \left\langle w, U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)} \right\rangle$.

Equation 2.3 is equivalent to logistic regression with features $x_\ell = U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}$. See Section 2.6.2 of the Supplement for the derivation. We estimate w^* with the maximum likelihood estimator $\hat{w}$, which requires minimizing a convex function: $\hat{w} := \text{argmin}_w \mathcal{L}_m(w; U, S_m, \tau)$.

2.3 Theory

In this section, we analyze the sample complexity of estimating the judgment weights with the MLE given by minimizing $\mathcal{L}_m$ of Equation (2.3). We first consider the sample complexity under an arbitrary selection function, and then specialize to two concrete selection functions: one that selects all features per pair and another that selects just one feature per pair. Throughout this section, we assume the set-up and notation presented in the beginning of Section 2.2.

First, the following proposition completely characterizes the identifiability of w^*. Identifiability means that with infinite samples, it is possible to learn w^*. Precisely, the salient feature preference model is identifiable if for all $(i,j) \in P$ and for $w_1, w_2 \in \mathbb{R}^d$, if $\mathbb{P}(i >_B j; w_1) = \mathbb{P}(i >_B j; w_2)$, then $w_1 = w_2$ where $\mathbb{P}(i >_B j; w)$ refers to Equation (2.2) where w is the judgement vector. The proof is in Section 2.6.3 of the Supplement.

Proposition 2.3.1 (Identifiability). *Given item features $U \in \mathbb{R}^{n \times d}$, the salient feature preference model with selection function τ is identifiable if and only if span$\{U_i^{\tau(i,j)} - U_j^{\tau(i,j)} : (i,j) \in P\} = \mathbb{R}^d$.*

Now we present our main theorem on the sample complexity of estimating w^*. Let

$$b^* := \max_{(i,j)\in P} |\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle|,$$

which is the maximum absolute difference between two items' utilities when comparing them in context, i.e. based on the features given by the selection function τ. Let

$$\mathcal{W}(b^*) := \{w \in \mathbb{R}^d : \max_{(i,j)\in P} |\langle w, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle| \leq b^*\}.$$

We constrain the MLE to $\mathcal{W}(b^*)$ so that we can bound the entries of the Hessian of $\mathcal{L}_m$ in our theoretical analysis. We do not enforce this constraint in our synthetic experiments.

Theorem 2.3.2 (Sample complexity of learning w^*)**.** *Let $U \in \mathbb{R}^{d\times n}$, $w^* \in \mathbb{R}^d$, τ, and S_m be defined as in the beginning of Section 2.2. Let $\hat{w}$ be the maximum likelihood estimator, i.e. the minimum of $\mathcal{L}_m$ in Equation (2.3), restricted to the set $\mathcal{W}(b^*)$. The following expectations are taken with respect to a uniformly chosen random pair of items from P. For $(i,j) \in P$, let*

$$\begin{aligned} Z_{(i,j)} &:= (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T \\ \lambda &:= \lambda_{\min}(\mathbb{E}Z_{(i,j)}), \\ \eta &:= \sigma_{\max}(\mathbb{E}((Z_{(i,j)} - \mathbb{E}Z_{(i,j)})^2)), \\ \zeta &:= \max_{(k,\ell)\in P} \lambda_{\max}(\mathbb{E}Z_{(i,j)} - Z_{(k,\ell)}), \end{aligned}$$

where for a positive semidefinite matrix X, $\lambda_{\min}(X)$ and $\lambda_{\max}(X)$ are the smallest/largest eigenvalues of X, and where for any matrix X, $\sigma_{\max}(X)$ is the largest singular value of X. Let

$$\beta := \max_{(i,j)\in P} \|U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\|_\infty. \tag{2.4}$$

Let $\delta > 0$. If $\lambda > 0$ and

$$\begin{aligned} m \geq \max \Big\{ & C_1(\beta^2 d + \beta\sqrt{d}) \log(4d/\delta) \ , \\ & C_2(\eta + \lambda\zeta)\frac{\log(2d/\delta)}{\lambda^2} \Big\} \ , \end{aligned}$$

then with probability at least $1 - \delta$,

$$\|w^* - \hat{w}\|_2 = O\left(\frac{\exp(b^*)}{\lambda} \sqrt{\frac{(\beta^2 d + \beta\sqrt{d}) \log(4d/\delta)}{m}} \right)$$

where C_1, C_2 are constants given in the proof and the randomness is from the randomly chosen pairs and the outcomes of the pairwise comparisons.

We utilize the proof technique of Theorem 4 in [NOS16], which proves a similar result for the standard BTL model of Equation (2.1), i.e. when $U = I_{n\times n}$, the $n \times n$ identity matrix, $d = n$, and $\tau(i,j) = [d]$ for all $(i,j) \in P$. We modify the proofs for arbitrary U

and d. See Section 2.6.5 in the Supplement for the proof.

We now discuss the terms that appear in Theorem 2.3.2. First, the $d\log(d/\delta)$ terms are natural since we are estimating d parameters. Second, estimating w^* well essentially requires inverting the logistic function. When b^* is large, we need to invert the logistic function for pairwise probabilities that are close to 0 and 1. This is precisely the challenging regime, since a small change in probabilities results in a large change in the estimate of w^*, and thus we expect to require many samples to estimate w^* when b^* is large. The exponential dependence on b^* is standard for this type of analysis and arises from the Hessian of $\mathcal{L}_m$. Third, η and ζ arise from a matrix concentration bound applied to the Hessian of $\mathcal{L}_m$. Fourth, λ arises from the minimum eigenvalue of the Hessian of $\mathcal{L}_m$ in a neighborhood of w^*, which controls the convexity of $\mathcal{L}_m$. This type of dependence also appears in other state of the art finite sample complexity analyses [NRW$^+$12]. In addition, to better understand the role of λ, we present the following proposition whose proof is in Section 2.6.4 in the Supplement. Proposition 2.3.3 shows that the requirement $\lambda > 0$ in Theorem 2.3.2 is fundamental, because we would otherwise be unable to bound the estimation error for the non-identifiable part of w^*, i.e., the projection of w^* onto the orthogonal complement of $\text{span}\{U_i^{\tau(i,j)} - U_j^{\tau(i,j)} : (i,j) \in P\} = \mathbb{R}^d$.

Proposition 2.3.3. $\lambda > 0$ *if and only if the salient feature preference model is identifiable.*

Finally, if one assumes $\lambda, \eta, \zeta, \beta, \exp(b^*)$ are $O(1)$, then $\Omega(d\log(d/\delta))$ samples are enough to guarantee the error is $O(1)$. However, as we will show in the corollaries, these parameters are not always $O(1)$, increasing the complexity. We point out that the combination of the features U and the selection function τ is what dictates the parameters of Theorem 2.3.2. For the top-t selection function in particular, we plot $\lambda, \zeta, \eta, b^*, \beta$, the number of samples required by Theorem 2.3.2, and the bound on the estimation error as a function of intransitivity rates in the Supplement in Section 2.6.8, to provide further insight into these parameters. Since we envision practical selection functions will be dependent on the features themselves, further analysis is a challenging but exciting subject of future work.

For deterministic U, we now specialize our results to FBTL as well as to the case where a single feature is used in each comparison. The following corollaries provide insight into how a particular selection function τ impacts λ, η, and ζ and thus the sample complexity.

First, we consider FBTL. In this case, the selection function selects all the features in each pairwise comparison, so there cannot be intransitivities in the preference data. The following Corollary of Theorem 2.3.2 gives a simplified form for λ and upper bounds ζ and η. The terms involving the conditioning of UU^T are natural; since we make no assumption on w^*, if the feature vectors are concentrated in a lower dimensional subspace, estimation of w^* will be more difficult. See Section 2.6.6 of the Supplement for the proof.

Corollary 2.3.4 (Sample complexity for FBTL)**.** *For the selection function τ, suppose $|\tau(i,j)| = d$ for any $(i,j) \in P$. In other words, all the features are used in each pairwise comparison. Let $\nu := \max\{\max_{(i,j)\in P} \|U_i - U_j\|_2^2, 1\}$. Assume $n > d$. Without loss of generality, assume the columns of U sum to zero: $\sum_{i=1}^n U_i = 0$. Let $\delta > 0$. Then,*

$$\begin{aligned}\lambda &= \frac{n\lambda_{\min}(UU^T)}{\binom{n}{2}},\\ \zeta &\leq \nu + \frac{n\lambda_{\max}(UU^T)}{\binom{n}{2}}, \ \textit{and}\\ \eta &\leq \frac{\nu n\lambda_{\max}(UU^T)}{\binom{n}{2}} + \frac{n^2\lambda_{\max}(UU^T)^2}{\binom{n}{2}^2}.\end{aligned}$$

Hence, if

$$\begin{aligned}m \geq \max\Big\{&C_1(\beta^2 d + \beta\sqrt{d})\log(4d/\delta),\\ &C_3\log(2d/\delta)\nu n\bar{\lambda}\Big\}\end{aligned}$$

where

$$\bar{\lambda} = \left(\frac{\lambda_{\max}(UU^T) + \lambda_{\max}(UU^T)^2 + \lambda_{\min}(UU^T)}{\lambda_{\min}(UU^T)^2}\right)$$

then with probability at least $1-\delta$,

$$\|w^* - \hat{w}\|_2 = O\left(\frac{\exp(b^*)n}{\lambda_{\min}(UU^T)}\sqrt{\frac{(\beta^2 d + \beta\sqrt{d})\log(\frac{4d}{\delta})}{m}}\right)$$

where C_1 and C_3 are constants given in the proof.

To the best of our knowledge, this is the first analysis of the sample complexity for the

MLE of FBTL parameters. There are related results in [SR18, NRW$^+$12, HSR$^+$19, SW17] to which our bound compares favorably, and we discuss this in Section 2.6.6 of the Supplement.

Second, suppose the selection function is very aggressive and selects only one coordinate for each pair, i.e. $|\tau(i,j)| = 1$ for all $(i,j) \in P$. For instance, the top-1 selection function has this property. This type of selection function can cause intransitivities in the preference data as we show in the synthetic experiments of Section 2.4.1.

Corollary 2.3.5. *Assume that for any $(i,j) \in P$, $|\tau(i,j)| = 1$. Partition $P = \sqcup_{k=1}^{d} P_k$ into d sets where $(i,j) \in P_k$ if $\tau(i,j) = \{k\}$ for $k \in [d]$. Let β be defined as in Theorem 2.3.2 and*

$$\epsilon := \min_{(i,j)\in P} \|U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\|_\infty.$$

Let $\delta > 0$. Then

$$\begin{aligned}
\lambda &\geq \frac{\epsilon^2}{\binom{n}{2}} \min_{k\in[d]} |P_k|,\\
\zeta &\leq \beta^2 + \frac{\beta^2}{\binom{n}{2}} \max_{k\in[d]} |P_k|, \text{ and}\\
\eta &\leq \frac{\beta^4}{\binom{n}{2}} \max_{k\in[d]} \left(|P_k| + \frac{|P_k|^2}{\binom{n}{2}} \right).
\end{aligned}$$

Hence, if

$$m \geq \max\left\{ C_1(\beta^2 d + \beta\sqrt{d}) \log(4d/\delta), C_4(Q_1 + Q_2) \right\},$$

where

$$\begin{aligned}
Q_1 &= \left(\frac{\beta^4}{\epsilon^4}\right) \frac{\binom{n}{2} \max_{k\in[d]} |P_k| + \max_{k\in[d]} |P_k|^2}{\min_{k\in[d]} |P_k|^2},\\
Q_2 &= \left(\frac{\beta^2}{\epsilon^2}\right) \frac{\binom{n}{2} + \max_{k\in[d]} |P_k|}{\min_{k\in[d]} |P_k|},
\end{aligned}$$

then with probability at least $1-\delta$*,*

$$\|w^* - \hat{w}\|_2 = O\left(\frac{\exp(b^*)\binom{n}{2}}{\epsilon^2 \min\limits_{k\in[d]}|P_k|}\sqrt{\frac{(\beta^2 d + \beta\sqrt{d})\log(\frac{4d}{\delta})}{m}}\right)$$

where C_1 *and* C_4 *are constants given in the proof.*

There are two main implications of Corollary 2.3.5 if we consider β and ϵ constant. First, suppose there is a coordinate $k \in [d]$ such that $|P_k| := |\{(i,j) \in P : \tau(i,j) = k\}|$ is small. Intuitively it will take many samples to estimate w^* well, since the chance of sampling a pairwise comparison that uses the k-th coordinate of w^* is $|P_k|/\binom{n}{2}$. Corollary 2.3.5 formalizes this intuition. In particular, $\lambda = O(|P_k|/\binom{n}{2})$, and since λ comes into the bounds of Theorem 2.3.2 in the denominator of both the lower bound on samples and the upper bound on error, a small λ makes estimation more difficult.

Second, on the other hand, if ϵ is fixed, the maximum lower bound on λ given by Corollary 2.3.5 is $\max\min_{i\in[d]}|P_i| = \binom{n}{2}/d$ where the maximum is with respect to any partition of P. In this case, $|P_i| \approx |P_j|$ for all $i, j \in [d]$, so the chance of sampling a pairwise comparison that uses any coordinate is approximately equal. Therefore, $\lambda, \eta, \zeta = O(1/d)$, and by tightening a bound used in the proof of Theorem 2.3.2, $\Omega(d^2 \log(d/\delta))$ samples ensures the estimation error is $O(1)$. See Section 2.6.6 in the Supplement for an explanation.

Ultimately, we seek to estimate the underlying ranking of the items. The following corollary of Theorem 2.3.2 says that by controlling the estimation error of w^*, the underlying ranking can be estimated approximately. The sample complexity depends inversely on the square of the differences of full feature item utilities. Intuitively, if the absolute difference between the utilities of two items is small, then many samples are required in order to rank these items correctly relative to each other. See Section 2.6.7 in the Supplement for the proof.

Corollary 2.3.6 (Sample complexity of estimating the ranking)**.** *Assume the set-up of Theorem 2.3.2. Pick* $k \in [\binom{n}{2}]$*. Let* α_k *be the k-th smallest number in* $\{|\langle w^*, U_i - U_j\rangle| : (i,j) \in P\}$*. Let* $M := \max_{i\in[n]} \|U_i\|_2$*. Let* $\gamma^* : [n] \to [n]$ *be the ranking obtained from* w^* *by sorting the items by their full-feature utilities* $\langle w^*, U_i\rangle$ *where* $\gamma^*(i)$ *is the position of item* i *in the ranking. Define* $\hat{\gamma}$ *similarly but for the estimated ranking obtained from the*

MLE estimate $\hat{w}$. Let $\delta > 0$. If

$$m \geq \max\Bigg\{C_1(\beta^2 d + \beta\sqrt{d})\log(4d/\delta), \\ C_2(\eta + \lambda\zeta)\frac{\log(2d/\delta)}{\lambda^2}, \\ \frac{C_5 M^2 e^{2b^*}(\beta^2 d + \beta\sqrt{d})\log(4d/\delta)}{\alpha_k^2\lambda^2}\Bigg\},$$

then with probability $1-\delta$,

$$K(\gamma^*, \hat{\gamma}) \leq k - 1,$$

where $K(\gamma^*, \hat{\gamma}) = |\{(i,j) \in P : (\gamma^*(i) - \gamma^*(j))(\hat{\gamma}(i) - \hat{\gamma}(j)) < 0\}|$ is the Kendall tau distance between two rankings and C_1, C_2, and C_5 are constants given in the proof.

2.4 Experiments

See Section 2.6.9 of the Supplement for additional details about the algorithm implementation, data, preprocessing, hyperparameter selection, and training and validation error for both synthetic and real data experiments.

2.4.1 Synthetic Data

We investigate violations of rational choice arising from the salient feature preference model and illustrate Theorem 2.3.2 while highlighting the differences between the salient feature preference model and the FBTL model throughout. Given the very reasonable simulation setup we use, these experiments suggest that the salient feature preference model may sometimes be better suited to real data than FBTL.

For these experiments, the ambient dimension $d = 10$, the number of items $n = 100$, and comparisons are sampled from the salient feature preference model with top-t selection function. The coordinates of U, respectively w^*, are drawn from $\mathcal{N}(0, 1/\sqrt{d})$, respectively $\mathcal{N}(0, 4/\sqrt{d})$, so that $\mathbb{P}(i >_B j)$ is bounded away from 0 and 1 for $i, j \in [n]$. This set-up ensures b^* does not become too large.

First, the salient feature preference model can produce preferences that systematically violate rational choice. In contrast, the FBTL model cannot. Let $P_{ij} = \mathbb{P}(i >_B j)$

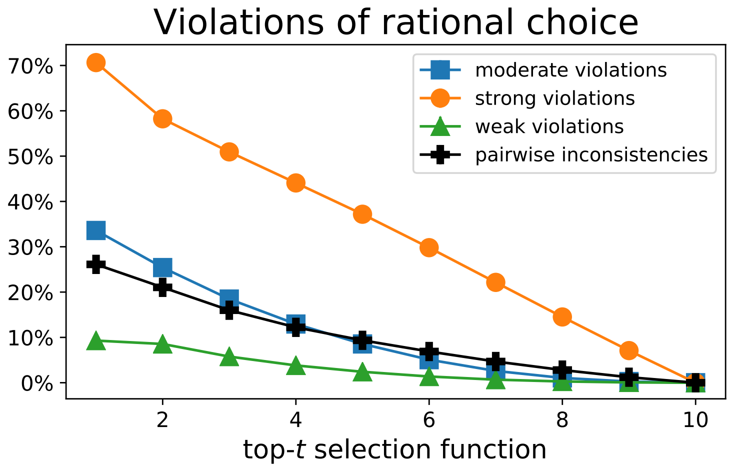


Figure 2.2: The salient feature preference model with the top-t selection function produces systematic intransitives and pairwise comparisons that are inconsistent with the underlying ranking. When $t = 10$, the salient feature preference model with the top-t selection function is the FBTL model, and hence there are no intransitives or pairwise inconsistencies.

and $T = \{(i, j, k) \in [n]^3 : P_{ij} > .5, P_{jk} > .5\}$. Then $(i, j, k) \in T$ satisfies strong stochastic transitivity if $P_{ik} \geq \max\{P_{ij}, P_{jk}\}$, moderate stochastic transitivity if $P_{ik} \geq \min\{P_{ij}, P_{jk}\}$, and weak stochastic transitivity if $P_{ik} \geq .5$ [Cat12]. We sample U and w^* 10 times as described in the beginning of the section and allow t to vary in $[d]$. Figure 2.2 shows the average ratio of the number of weak, moderate, and strong stochastic transitivity violations to $|T|$ as a function of $t \in [d]$. There is very little deviation from the average. The standard error bars over the 10 experiments were plotted but they are so small that the markers covered them. All $\binom{n}{2}$ probabilities given by Equation (2.2) are used to calculate the intransitivity rates. In the same figure we also show the percentage of pairwise comparisons that are inconsistent with the true ranking under the same experimental set-up. These are the pairs i, j such that $\langle U_i - U_j, w^* \rangle < 0$, meaning item i is ranked lower than item j in the true ranking, but $\langle U_i^{\tau_t(i,j)} - U_j^{\tau_t(i,j)}, w^* \rangle > 0$ meaning item i beats item j by at least 50% when compared in isolation from the other items. Notice that when $t = 10$, the salient feature preference model is the FBTL model, so there are no pairwise inconsistencies or intransitives. Although this example is synthetic, real data exhibits intransitivity and even inconsistent pairs with the underlying ranking as discussed in the real data experiments in Section 2.4.2.

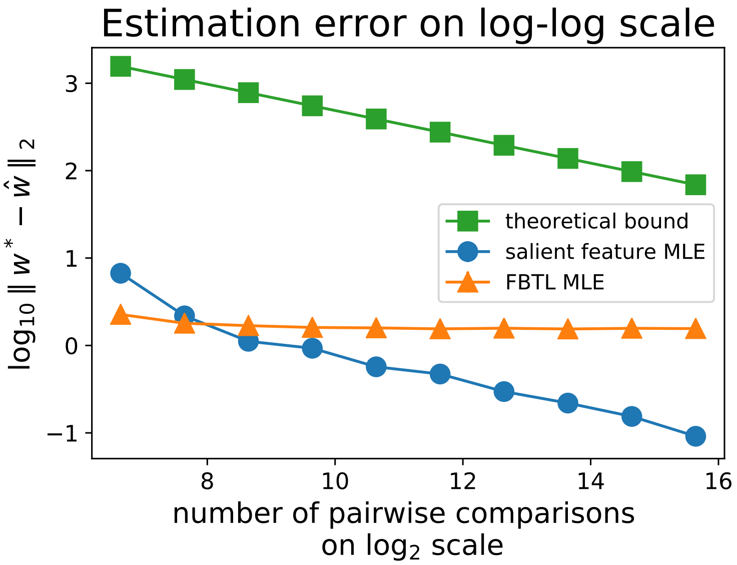


Figure 2.3: Illustration of Theorem 2.3.2 with the exact theoretical upper bound for the salient feature preference model with the top-1 selection function. Although there is a gap between the bound and the observed estimation error, they decrease at the same rate eventually. Excluding the first two points, the salient feature MLE error's slope on the log-log scale is -0.154, whereas the theoretical bound's slope is -0.151.

Second, we illustrate Theorem 2.3.2 with the top-1 selection function, and where U and w are sampled once as described in the beginning of this section. We sample m pairwise comparisons for $m \in \{(100)2^{i-1} : i \in [10]\}$, fit the MLEs of both the salient preference model with the top-1 selection function and FBTL, and repeat 10 times. Figure 2.3 shows the average estimation error of w^* on a logarithmic scale as a function of the number of pairwise comparison samples also on a logarithmic scale. Figure 2.3 also shows the exact theoretical upper bound where $\delta = \frac{1}{d} = \frac{1}{10}$ of Theorem 2.3.2 without constants C_1 and C_2 as stated in Section 2.6.5 of the Supplement. Again, there is very little deviation from the average. The standard error bars over the 10 experiments were plotted but they are so small that the markers covered them. There is a gap between the observed error and the theoretical bound, though the error decreases at the same rate. The error of the MLE of FBTL does not improve with more samples, since the pairwise comparisons are generated according to the salient feature preference model with the top-1 selection function. See Section 2.6.8 in the Supplement for investigating

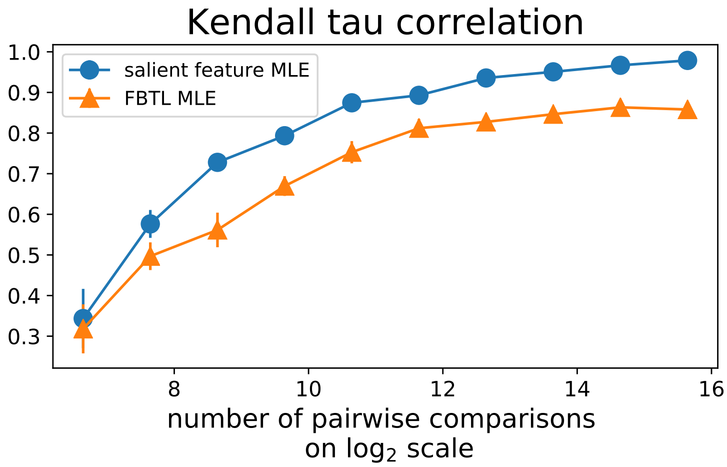


Figure 2.4: Kendall tau correlation between the true ranking and the estimated ranking where pairwise comparisons are sampled from the salient feature preference model with the top-1 selection function. Estimating w^* well implies being able to estimate the underlying ranking well as stated in Corollary 2.3.6.

model misspecification, i.e. fitting the MLE of the top-t selection function for $t \neq 1$ with the same experimental set-up.

By estimating w^* well, we can estimate the underlying ranking well by Corollary 2.3.6. Under the same experimental set up, Figure 2.4 shows the Kendall tau correlation (definition given in Supplement 2.6.8) between the true ranking (obtained by ranking the items according to $\langle U_i, w^* \rangle$) and the estimated ranking (according to $\langle U_i, \hat{w} \rangle$) but on a new set of 100 items drawn from the same distribution. The maximum Kendall tau correlation between two rankings is 1 and occurs when both rankings are equal. Also, estimating w^* well allows us to predict the outcome of unseen pairwise comparisons well, as shown in the Supplement in Section 2.6.8.

2.4.2 Real Data

For the following experiments, we use the top-t selection function for the salient feature preference model, where t is treated as a hyperparameter and tuned on a validation set. We compare to FBTL, RankNet [BSR$^+$05] with one hidden layer, and Ranking SVM [Joa02]. We append an ℓ_2 penalty to $\mathcal{L}_m$ for the salient feature preference model and the

Table 2.1: Average Kendall tau correlation over individual rankings on test sets for district compactness. The number in parenthesis is the standard deviation.

Model:	`Shiny1`	`Shiny2`	`UG1-j1`	`UG1-j2`	`UG1-j3`	`UG1-j4`	`UG1-j5`
Salient features	**.14** (.26)	**.26** (.2)	**.48** (.21)	**.41** (.09)	**.6** (.1)	.14 (.14)	**.42** (.09)
FBTL	.09 (.22)	.18 (.17)	.2 (.12)	.26 (.07)	.45 (.15)	.2 (.13)	.06 (.14)
Ranking SVM	.09 (.22)	.18 (.17)	.22 (.12)	.26 (.07)	.45 (.15)	.2 (.13)	.06 (.14)
RankNet	.12 (.24)	.24 (.18)	.28 (.14)	.37 (.08)	.53 (.11)	**.28** (.08)	.15 (.15)

FBTL model, that is, for regularization parameter μ, we solve $\min_{w \in \mathbb{R}^d} \mathcal{L}_m(w) + \mu \|w\|_2^2$. For RankNet, we add to the objective function an ℓ_2 penalty on the weights. As explained in more detail in subsection 2.6.9 in the Supplement, the hyperparameters for the salient feature preference model are t for the top-t selection function and μ, the hyperparameter for FBTL is μ, the hyperparameter for Ranking SVM is the coefficient corresponding to the norm of the learned hyperplane, and the hyperparameters for RankNet are the number of nodes in the single hidden layer and the coefficient for the ℓ_2 regularization of the weights.

District Compactness [KKK17] collected preference data to understand human perception of compactness of legislative districts in the United States. Their data include both pairwise comparisons and k-wise ranking data for $k > 2$ as well as 27 continuous features for each district, including geometric features and compactness metrics. Although difficult to define precisely, the United States law suggests compactness is universally understood [KKK17]. In fact, the authors provide evidence that most people agree on a universal ranking, but they found the pairwise comparison data was extremely noisy. They hypothesize that pairwise comparisons may not directly capture the full ranking, since all features may not be used when comparing two districts in isolation from the other districts. Hence, this problem is applicable to our salient feature preference model and its motivation.

The goal as set forth by [KKK17] is to learn a ranking of districts. We train on 5,150 pairwise comparisons collected from 94 unique pairs of districts to learn $\hat{w}$, an estimate of the judgment vector w^*, then estimate a ranking by sorting the districts by $\langle \hat{w}, U_i \rangle$. The k-wise ranking data sets are used for validation and testing. Since there is no ground truth for the universal ranking, we measure how close the estimated ranking is to each individual ranking. In this scenario, we care about the accuracy of the full ranking, and so we consider Kendall tau correlation. Given a k-wise comparison data set, Table 2.1 shows the average Kendall tau correlation between the estimated ranking and each individual ranking where the number in parenthesis is the standard deviation. The standard deviation on `shiny1` and `shiny2` is relatively high because the Kendall tau correlation between pairs of rankings in these data sets has high variability, shown in Figure 2.10 in the Supplement.

The MLE of the salient feature preference model under the top-t selection function outperforms both the MLE of FBTL and Ranking SVM by a significant amount on 6 out 7 test sets, suggesting that pairwise comparison decisions may be better modeled by incorporating context. The MLE of the salient feature preference model, which is linear, is competitive with RankNet, which models pairwise comparisons as in Equation (2.1) except where the utility of each item uses a function f defined by a neural network, i.e. $u_i = f(U_i)$.

The salient feature preference model may be outperforming FBTL and Ranking SVM since this data exhibits significant violations of rational choice. First, on the training set of pairwise comparisons, there are 48 triplets of districts (i, j, k) where both (1) all three distinct pairwise comparisons were collected and (2) $P_{ij} > .5$ and $P_{jk} > .5$. Seventeen violate strong transitivity, 3 violate moderate transitivity, but none violate weak transitivity. Second, given a set of k-wise ranking data, let $\hat{P}_{ij}$ be the proportion of rankings in which item i is ranked higher than item j. There are 20 pairs of districts that appear in both the k-wise ranking data and the pairwise comparison training data. Four of these pairs of items i, j have the property that $(.5 - P_{ij})(.5 - \hat{P}_{ij}) < 0$, meaning item i is typically ranked higher than item j in the ranking data, but j typically beats i in the pairwise comparisons.

`UT Zappos50k` The `UT Zappos50K` data set consists of pairwise comparisons on images of shoes and 960 extracted vision features for each shoe [YG14, YG17]. Given images of two shoes and an attribute from {"open," "pointy," "sporty," "comfort"}, respondents picked which shoe exhibited the attribute more. The data consists of easier, coarse questions, i.e. based on comfort, pick between a slipper or high-heel, and harder, fine grained questions i.e. based on comfort, pick between two slippers.

Table 2.2: Average pairwise prediction accuracy over 10 train/validation/test splits on the test sets by attribute for `UT Zappos50k`. *C* stands for coarse and *F* stands for fine grained. *O* stands for open, *P* stands for pointy, *S* stands for sporty, and *Co* stands for comfort. The number in parenthesis is the standard deviation.

Model:	*O-C*	*P-C*	*S-C*	*Co-C*	*O-F*	*P-F*	*S-F*	*Co-F*
Salient features	.73(.02)	.78(.02)	.78(.03)	.77(.03)	.6(.04)	.59(.04)	.59(.03)	.56(.03)
FBTL	.73(.02)	.77(.03)	.8(.03)	.78(.03)	.6(.03)	.6(.03)	.59(.03)	.58(.05)
Ranking SVM	.74(.02)	.78(.03)	.79(.03)	.78(.03)	.6(.03)	.6(.04)	.6(.04)	.58(.03)
RankNet	.73(.01)	.79(.01)	.78(.03)	.8(.02)	.61(.02)	.59(.02)	.59(.04)	.59(.05)

We now consider predicting pairwise comparisons instead of estimating a ranking since there is no ranking data available. We train four models, one for each attribute. See Table 2.2 for the average pairwise comparison accuracy over ten train (70%), validation (15%), and test splits (15%) of the data. The pairwise comparison accuracy is defined as the percentage of items (i, j) where i beats j a majority of the time and the model estimates the probability that i beats j exceeds 50%.

In this case, the MLE of the FBTL model and the salient feature preference model under the top t selection function perform similarly. Nevertheless, while the FBTL model utilizes all 990 features, the best t's on each validation set and split of the data do not use all features, so our model is different from yet competitive to FBTL. See Table 2.3 in the Supplement. This suggests that the salient feature preference model under the top-t selection function for relatively small t is still a reasonable model for real data.

2.5 Conclusion

We focused on the problem of learning a ranking from pairwise comparison data with irrational choice behaviors, and we formulated the salient feature preference model where one uses projections onto salient coordinates in order to perform comparisons. We proved sample complexity results for MLE on this model and demonstrated the efficacy of our model on both synthetic and real data. Going forward, we would like to develop techniques to learn both the selection function τ and feature embeddings simultaneously. Finally, it will be useful to consider how to incorporate context into models more sophisticated than BTL, and also consider contextual effects in other tasks that use human judgements such as ordinal embedding [TL14].

2.6 Supplement

2.6.1 k-wise Comparisons Extension

We describe how to extend the salient feature preference model of Equation (2.2) from pairwise comparisons to k-wise comparisons when $k > 2$. We base our generalization on the Placket-Luce model [Pla75, Luc59], which is a generalization of the BTL model from pairwise comparisons to k-wise comparisons.

Let the domain of the selection function τ be $[n]^k$ instead of $[n] \times [n]$, i.e. $\tau : [n]^k \to \mathcal{P}([d])$. Then for $T_\ell = (t_1, \ldots, t_k)$ where $t_i \in [n]$ are items, the probability of picking the ranking $t_1 >_B \cdots >_B t_k$ is

$$\mathbb{P}(t_1 >_B \cdots >_B t_k) = \prod_{\ell=1}^{k} \frac{\exp\left(\langle U_{t_\ell}^{\tau(T_\ell)}, w^* \rangle\right)}{\sum_{j \in [k] \setminus [\ell-1]} \exp\left(\langle U_{t_j}^{\tau(T_\ell)}, w^* \rangle\right)}, \tag{2.5}$$

where "$t_1 >_B \cdots >_B t_k$" means item t_1 is preferred to item t_2 and so on and so forth.

We explain Equation (2.5): Given items $T_\ell = (t_1, \ldots, t_k)$, first project each item's features U_{t_i} onto the coordinate subspace spanned by the coordinates given by $\tau(T_\ell)$. Then the utility of item t_i in the presence of the other items in T is given by the inner product of its projected features with w^*: $\langle (U_{t_i})^{\tau(T_\ell)}, w^* \rangle$. The higher the utility an item has, the more likely the item will be ranked higher among the items in T_ℓ. Now imagine a bag of balls where each ball corresponds to one of the items in T_ℓ. We select balls from this bag without replacement where the probability of picking a ball is the ratio of its utility to the sum of the utilities of all the remaining balls. The order in which we select balls results in a ranking of the k items. This process is what Equation (2.5) represents.

In the pairwise comparison case ($k = 2$) for two items $T_\ell = (i, j)$, Equation (2.5) reduces to Equation (2.2), which is the salient preference model. We can also extend the top-t selection function naturally to accommodate k-wise comparisons.

2.6.2 Negative Log-Likelihood Derivation

Lemma 2.6.1. *Under the set-up of Section 2.2, the negative log-likelihood of $w \in \mathbb{R}^d$ is*

$$\mathcal{L}_m(w; U, S_m, \tau) = \sum_{\ell=1}^{m} \log\left(1 + \exp\left(\langle U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}, w\rangle\right)\right) - y_\ell \langle U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}, w\rangle. \tag{2.6}$$

Proof. Let $P_w(S_m)$ be the joint distribution of the m samples S_m with respect to the judgement vector w. Then

$$\begin{aligned}
&\mathcal{L}_m(w; U, S_m, \tau) && (2.7)\\
&= -\log P_w(S_m) && (2.8)\\
&= -\log\left(\prod_{\ell=1}^{m} (\mathbb{P}(y_\ell = 1)^{y_\ell} \mathbb{P}(y_\ell = 0)^{1-y_\ell})\right) \text{ by independence and since } y_\ell \in \{0, 1\} && (2.9)\\
&= -\sum_{i=1}^{m} y_\ell \log(\mathbb{P}(y_\ell = 1)) + (1 - y_\ell)\log(1 - \mathbb{P}(y_\ell = 1)) && (2.10)\\
&= -\sum_{i=1}^{m} y_\ell \log\left(\frac{\exp\left(\langle U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}, w\rangle\right)}{1 + \exp\left(\langle U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}, w\rangle\right)}\right) && (2.11)\\
&+ (1 - y_\ell)\log\left(\frac{1}{1 + \exp\left(\langle U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}, w\rangle\right)}\right)\\
&= \sum_{i=1}^{m} \log\left(1 + \exp\left(\langle U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}, w\rangle\right)\right) - y_\ell \langle U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}, w\rangle && (2.12)
\end{aligned}$$

□

2.6.3 Proof of Proposition 2.3.1

Proposition 2.6.2 (Restatement of Proposition 2.3.1)**.** *Given item features $U \in \mathbb{R}^{d \times n}$, the salient feature preference model with selection function τ is identifiable if and only if span$\{U_i^{\tau(i,j)} - U_j^{\tau(i,j)} : (i, j) \in P\} = \mathbb{R}^d$.*

Proof. Let $w \in \mathbb{R}^d$. Then for any $(i,j) \in P$,

$$\begin{align}
&\mathbb{P}(i >_B j; w) = \mathbb{P}(i >_B j; w^*) \tag{2.13}\\
\iff &\frac{\exp\left(\langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w\rangle\right)}{1+\exp\left(\langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w\rangle\right)} = \frac{\exp\left(\langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w^*\rangle\right)}{1+\exp\left(\langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w^*\rangle\right)} \tag{2.14}\\
\iff &\exp\left(\langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w\rangle\right) = \exp\left(\langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w^*\rangle\right) \tag{2.15}\\
\iff &\langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w\rangle = \langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w^*\rangle \tag{2.16}\\
\iff &\langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w^* - w\rangle = 0. \tag{2.17}
\end{align}$$

$\Rightarrow$ Assume identifiability. By contradiction, if $\text{span}\{U_i^{\tau(i,j)} - U_j^{\tau(i,j)} : (i,j) \in P\} \neq \mathbb{R}^d$, then there is some vector $x \neq 0$ that is orthogonal to $\text{span}\{U_i^{\tau(i,j)} - U_j^{\tau(i,j)} : (i,j) \in P\}$. Consider $w^* - x$. Then, for any $(i,j) \in P$

$$\begin{align}
\langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w^* - (w^* - x)\rangle &= \langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, x\rangle \tag{2.18}\\
&= 0. \tag{2.19}
\end{align}$$

Therefore, with $w = w^* - x$, Equation (2.17) is true and implies Equation (2.13) meaning

$$\mathbb{P}(i > j; w^* - x) = \mathbb{P}(i > j; w^*),$$

contradicting identifiability since $w^* - x \neq w^*$ because $x \neq 0$.

$\Leftarrow$ Now assume $\text{span}\{U_i^{\tau(i,j)} - U_j^{\tau(i,j)} : (i,j) \in P\} = \mathbb{R}^d$. We want to prove identifiability so suppose there exists w such that Equation (2.13) holds. We will show $w = w^*$. Let $x \in \mathbb{R}^d$ where $x = \sum_{(i,j)\in P} \alpha_{i,j}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})$ for $\alpha_{i,j} \in \mathbb{R}$. Then by Equation (2.17),

$$\left\langle \sum_{(i,j)\in P} \alpha_{i,j}\left(U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\right), w^* - w \right\rangle = 0.$$

Since this is true for any $x \in \mathbb{R}^d$, $w^* - w = 0$, which means $w = w^*$. □

2.6.4 Proof of Proposition 2.3.3

Proposition 2.6.3 (Restatement of Proposition 2.3.3)**.** *Under the set-up of Section 2.2,* $\lambda := \lambda_{\min}(\mathbb{E}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T) > 0$ *if and only if the salient feature preference model with selection function* τ *is identifiable.*

Proof. For both directions, we prove the contrapositive.

⇒ Assume $\lambda_{\min}(\mathbb{E}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T) = 0$. Recall the expectation is with respect to a uniformly at random chosen pair of items. Let $\mathbf{0} \in \mathbb{R}^d$ be the all 0 vector. Then there exists $y \neq \mathbf{0} \in \mathbb{R}^d$ that has unit norm such that

$$
\begin{aligned}
&(\mathbb{E}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T)y = \mathbf{0} && (2.20)\\
\implies\ & y^T(\mathbb{E}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T)y = 0 && (2.21)\\
\implies\ & \frac{1}{\binom{n}{2}} \sum_{(i,j)\in P} y^T(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T)y = 0 && (2.22)\\
\implies\ & \frac{1}{\binom{n}{2}} \sum_{(i,j)\in P} \|(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T y\|_2^2 = 0 && (2.23)\\
\implies\ & \|(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T y\|_2^2 = 0 \ \forall (i,j) \in P && (2.24)\\
\implies\ & (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T y = \mathbf{0} \ \forall (i,j) \in P, && (2.25)
\end{aligned}
$$

where Equation (2.22) is because $(i,j) \in P$ is chosen uniformly at random.

We now show $y \notin \operatorname{span}\{U_i^{\tau(i,j)} - U_j^{\tau(i,j)} : (i,j) \in P\}$, which establishes the salient feature preference model is not identifiable by Proposition 2.3.1. By contradiction, suppose there exist $\alpha_{i,j} \in \mathbb{R}$ such that

$$
y = \sum_{(i,j)\in P} \alpha_{i,j}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)}).
$$

Then

$$
\begin{aligned}
1 &= \langle y, y \rangle && (2.26)\\
&= \left\langle \sum_{(i,j)\in P} \alpha_{i,j} \left(U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\right), y \right\rangle && (2.27)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{(i,j)\in P} \alpha_{i,j} \left\langle \left(U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \right), y \right\rangle \qquad &(2.28)\\
&= 0, &(2.29)
\end{aligned}
$$

a contradiction.

$\Leftarrow$ Now suppose that the preference model is not identifiable. By Proposition 2.3.1, $\text{span}\{U_i^{\tau(i,j)} - U_j^{\tau(i,j)} : (i,j) \in P\} \neq \mathbb{R}^d$. In particular, there exists $y \in \mathbb{R}^d$ such that $y \neq \mathbf{0}$ and $\langle y, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle = 0$ for all $(i,j) \in P$, i.e. y is in the orthogonal complement of $\text{span}\{U_i^{\tau(i,j)} - U_j^{\tau(i,j)} : (i,j) \in P\}$. Furthermore,

$$
\begin{aligned}
\frac{1}{\binom{n}{2}} \sum_{(i,j)\in P} (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T y = \mathbf{0} \qquad &(2.30)\\
\implies (\mathbb{E}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T) y = \mathbf{0}, \qquad &(2.31)\\
&(2.32)
\end{aligned}
$$

since the expectation is with respect to a uniformly at random chosen pair of items. Therefore, $\lambda_{\min}(\mathbb{E}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T) = 0$ since all the eigenvalues of $\mathbb{E}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T$ are non-negative since it is a sum of positive semidefinite matrices, and 0 is an eigenvalue. □

2.6.5 Proof of Theorem 2.3.2

Recall the set-up from the beginning of Section 2.2. There are n items where the features of the items are given by the columns of $U \in \mathbb{R}^{d\times n}$ and let $w^* \in \mathbb{R}^d$ be the judgment vector. Let τ be the selection function. Let $S_m = \{(i_\ell, j_\ell, y_\ell)\}_{\ell=1}^m$ be the m samples of independent pairwise comparisons where each pair of items (i_ℓ, j_ℓ) is chosen uniformly at random from all the pairs of items $P := \{(i,j) \in [n] \times [n] : i < j\}$. Furthermore, y_ℓ is 1 if the i_ℓ-th item beats the j_ℓ-th item and 0 otherwise where $y_\ell \sim \text{Bernoulli}\left(\frac{\exp\left(\langle U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}, w^* \rangle\right)}{1+\exp\left(\langle U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}, w^* \rangle\right)} \right)$. We will not repeat these assumptions in the following lemmas.

In this section, we present the exact lower bounds on the number of samples and upper bound on the estimation error. The exact values of the constants that appear in the main text, i.e. C_1 and C_2, appear at the end of the proof.

Theorem 2.6.4 (restatement of Theorem 2.3.2: sample complexity of estimating w^*)**.**

Let U, w^*, τ, and S_m be defined as above. Let $\hat{w}$ be the maximum likelihood estimator, i.e. the minimum of $\mathcal{L}_m$ in Equation (2.3), restricted to the set $\mathcal{W}(b^*)$. The following expectations are taken with respect to a uniformly chosen random pair of items from P. For $(i,j) \in P$, let

$$\begin{aligned} Z_{(i,j)} &:= (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T \\ \lambda &:= \lambda_{\min}(\mathbb{E}Z_{(i,j)}), \\ \eta &:= \sigma_{\max}(\mathbb{E}((Z_{(i,j)} - \mathbb{E}Z_{(i,j)})^2)), \\ \zeta &:= \max_{(k,\ell)\in P} \lambda_{\max}(\mathbb{E}Z_{(i,j)} - Z_{(k,\ell)}), \end{aligned}$$

where for a positive semidefinite matrix X, $\lambda_{\min}(X)$ and $\lambda_{\max}(X)$ are the smallest/largest eigenvalues of X, and where for any matrix X, $\sigma_{\max}(X)$ is the largest singular value of X. Let

$$\beta := \max_{(i,j)\in P} \|U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\|_\infty. \tag{2.33}$$

Let $\delta > 0$. If $\lambda > 0$ and if

$$m \geq \max\left\{ \frac{3\beta^2 \log(4d/\delta) d + 4\sqrt{d}\beta \log(4d/\delta)}{6}, \frac{8\log(2d/\delta)(6\eta + \lambda\zeta)}{3\lambda^2} \right\},$$

then with probability at least $1 - \delta$,

$$\|w^* - \hat{w}\|_2 \leq \frac{4(1+\exp(b^*))^2}{\exp(b^*)\lambda} \sqrt{\frac{3\beta^2 \log(4d/\delta) d + 4\sqrt{d}\beta \log(4d/\delta)}{6m}}$$

where the randomness is from the randomly chosen pairs and the outcomes of the pairwise comparisons.

Proof. We use the proof technique of Theorem 4 in [NOS16]. We use the notation $\mathcal{L}_m(w)$ instead of $\mathcal{L}_m(w; U, S_m, \tau)$ throughout the proof since it is clear from context.

By definition $\mathcal{L}_m(\hat{w}) \leq \mathcal{L}_m(w^*)$. Let $\Delta := \hat{w} - w^*$. Then

$$\mathcal{L}_m(w^* + \Delta) - \mathcal{L}_m(w^*) - \langle \nabla \mathcal{L}_m(w^*), \Delta \rangle \tag{2.34}$$

$$\leq -\langle \nabla \mathcal{L}_m(w^*), \Delta \rangle \tag{2.35}$$

$$\leq \|\nabla \mathcal{L}_m(w^*)\|_2 \|\Delta\|_2, \tag{2.36}$$

by the Cauchy-Schwarz inequality.

Recall Taylor's theorem:

Theorem 2.6.5 (Taylor's Theorem)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$. If the Hessian H_f of f exists everywhere on its domain, then for any $x, \Delta \in \mathbb{R}^n$, there exists $\lambda \in [0,1]$ such that $f(x+\Delta) = f(x) + \langle \nabla f(x), \Delta \rangle + \frac{1}{2}\Delta^T H_f(x+\lambda\Delta)\Delta$.*

Now, we lower bound Equation (2.34). Let $H_{\mathcal{L}_m}$ be the Hessian of $\mathcal{L}_m$. Then by Taylor's theorem, there exists $\lambda \in [0,1]$ such that

$$\frac{1}{m}\left(\mathcal{L}_m(w^* + \Delta) - \mathcal{L}_m(w^*) - \langle \nabla \mathcal{L}_m(w^*), \Delta \rangle\right) \tag{2.37}$$

$$= \frac{1}{2m}\Delta^T H_{\mathcal{L}_m}(w^* + \lambda\Delta)\Delta \tag{2.38}$$

$$= \frac{1}{2m}\sum_{\ell=1}^{m} h(\langle w^* + \lambda\Delta, U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\rangle)\Delta^T(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})^T\Delta \tag{2.39}$$

where the Hessian $H_{\mathcal{L}_m}$ is computed in Lemma 2.6.10 and $h(x) := \frac{e^x}{(1+e^x)^2}$.

Note

$$|\langle w^* + \lambda\Delta, U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\rangle| \tag{2.40}$$

$$= |(1-\lambda)\langle w^*, U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\rangle + \lambda\langle \hat{w}, U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\rangle| \tag{2.41}$$

$$\leq (1-\lambda)b^* + \lambda b^* \tag{2.42}$$

$$= b^* \tag{2.43}$$

where the second to last inequality is by definition of b^* and since $\hat{w} \in \mathcal{W}(b^*)$. Because $h(x) = \frac{e^x}{(1+e^x)^2}$ is symmetric and decreases on $[0,\infty)$ by Lemma 2.6.11, for any $i, j \in [n]$,

$$h(\langle w^* + \lambda\Delta, U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\rangle) \geq h(b^*) = \frac{\exp(b^*)}{(1+\exp(b^*))^2}.$$

Therefore,

$$\frac{1}{m}\left(\mathcal{L}_m(w^* + \Delta) - \mathcal{L}_m(w^*) - \langle \nabla \mathcal{L}_m(w^*), \Delta \rangle\right) \tag{2.44}$$

$$\geq \frac{\exp(b^*)}{2m(1+\exp(b^*))^2} \sum_{\ell=1}^{m} \Delta^T (U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})^T \Delta. \tag{2.45}$$

By Lemma 2.6.6 and 2.6.8 and combining Equation (2.36) and Equation (2.45), with probability at least $1-\delta$ if

$$m \geq \max\left\{ \frac{3\beta^2 \log(4d/\delta)d + 4\sqrt{d}\beta \log(4d/\delta)}{6}, \frac{8\log(2d/\delta)(6\eta+\lambda\zeta)}{3\lambda^2} \right\},$$

$$\left(\frac{\exp(b^*)}{2(1+\exp(b^*))^2}\right) \frac{\lambda}{2} \|\Delta\|_2^2 \leq \frac{1}{m} \left(\mathcal{L}_m(w^*+\Delta) - \mathcal{L}_m(w^*) - \langle \nabla \mathcal{L}_m(w^*), \Delta \rangle\right) \tag{2.46}$$

$$\leq \sqrt{\frac{3\beta^2 \log(4d/\delta)d + 4\sqrt{d}\beta \log(4d/\delta)}{6m}} \|\Delta\|_2 \tag{2.47}$$

$$\implies \|\Delta\|_2 \leq \frac{4(1+\exp(b^*))^2}{\exp(b^*)\lambda} \sqrt{\frac{3\beta^2 \log(4d/\delta)d + 4\sqrt{d}\beta \log(4d/\delta)}{6m}}. \tag{2.48}$$

In the main text of this chapter with order terms, it is easy to see the $O(\cdot)$ bound on the upper bound on the estimation error. Furthermore, it is easy to see that for the constants C_1 and C_2 given in the main text of this chapter, we have $C_1 = 4/6$ and $C_2 = 48/3$. □

We now present the lemmas used in the prior proof.

Lemma 2.6.6. *Let $\delta > 0$. Under the model assumptions in this section, if*

$$m \geq \frac{3\beta^2 \log(4d/\delta)d + 4\sqrt{d}\beta \log(4d/\delta)}{6},$$

then with probability at least $1 - \frac{\delta}{2}$,

$$\left\| \frac{1}{m} \nabla \mathcal{L}_m(w^*) \right\|_2 \leq \sqrt{\frac{3\beta^2 \log(4d/\delta)d + 4\sqrt{d}\beta \log(4d/\delta)}{6m}}$$

where $\beta := \max_{(i,j)\in P} \left\| U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \right\|_\infty$.

Proof. For $\ell \in [m]$, let

$$X_\ell = \frac{1}{m}\left(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\right)\left(\frac{\exp(\langle w^*, U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\rangle)}{1+\exp(\langle w^*, U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\rangle)} - y_\ell\right),$$

so $\frac{1}{m}\nabla\mathcal{L}_m(w^*) = \sum_{\ell=1}^m X_\ell$ by Lemma 2.6.10.

We now show (1) $\mathbb{E}(X_\ell) = 0$ where the expectation is taken with respect to a uniformly chosen pair of items, (2) the coordinates of X_ℓ are bounded, and (3) the coordinates of X_ℓ have bounded second moments.

First $\mathbb{E}(X_\ell) = 0$. By conditioning on each pair of items, each of which have the same probability of being chosen,

$$\mathbb{E}(X_\ell) = \frac{1}{\binom{n}{2}}\sum_{(i,j)\in P}\mathbb{E}(X_\ell|\text{items } i,j \text{ are chosen}) \tag{2.49}$$

$$= \frac{1}{\binom{n}{2}}\sum_{(i,j)\in P}\frac{1}{m}\left(U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\right)\left(\frac{\exp(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\rangle)}{1+\exp(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\rangle)} - \mathbb{E}(y_{(i,j)})\right) \tag{2.50}$$

$$= \frac{1}{\binom{n}{2}}\sum_{(i,j)\in P}\frac{1}{m}\left(U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\right) \tag{2.51}$$

$$\left(\frac{\exp(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\rangle)}{1+\exp(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\rangle)} - \frac{\exp(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\rangle)}{1+\exp(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\rangle)}\right)$$

$$= 0, \tag{2.52}$$

where the expectation is with respect to the random pair that is drawn and the outcome of the pairwise comparison.

Second, $|X_\ell^{(k)}| \leq \frac{\beta}{m}$ where $X_\ell^{(k)}$ is the k-th coordinate of X_ℓ. Then for $k \in [d]$

$$|X_\ell^{(k)}| \tag{2.53}$$

$$= \left|\frac{1}{m}\left((U_{i_\ell}^{\tau(i_\ell,j_\ell)})^{(k)} - (U_{j_\ell}^{\tau(i_\ell,j_\ell)})^{(k)}\right)\left(\frac{\exp(\langle w^*, U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\rangle)}{1+\exp(\langle w^*, U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\rangle)} - y_\ell\right)\right| \tag{2.54}$$

$$\leq \frac{1}{m}\left|\left((U_{i_\ell}^{\tau(i_\ell,j_\ell)})^{(k)} - (U_{j_\ell}^{\tau(i_\ell,j_\ell)})^{(k)}\right)\right| \text{ since } \frac{\exp(\langle w^*, U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\rangle)}{1+\exp(\langle w^*, U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\rangle)}, y_\ell \in [0,1] \tag{2.55}$$

$$\leq \frac{1}{m}\max_{(i,j)\in P} \|U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\|_\infty \tag{2.56}$$

$$= \frac{\beta}{m}, \tag{2.57}$$

by definition of β.

Third, $\mathbb{E}((X_\ell^{(k)})^2) \leq \frac{\beta^2}{m^2}$. Let $p(x) = \frac{e^x}{1+e^x}$. For $k \in [d]$,

$$\mathbb{E}((X_\ell^{(k)})^2) \tag{2.58}$$

$$= \frac{1}{\binom{n}{2}}\sum_{(i,j)\in P} \mathbb{E}((X_\ell^{(k)})^2 | \text{items } i,j \text{ are chosen}) \tag{2.59}$$

$$= \frac{1}{\binom{n}{2}}\sum_{(i,j)\in P} \frac{1}{m^2}\left((U_i^{\tau(i,j)})^{(k)} - (U_j^{\tau(i,j)})^{(k)}\right)^2 \mathbb{E}\left(\left(p(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\rangle) - y_{(i,j)}\right)^2\right) \tag{2.60}$$

$$= \frac{1}{m^2\binom{n}{2}}\sum_{(i,j)\in P}\left((U_i^{\tau(i,j)})^{(k)} - (U_j^{\tau(i,j)})^{(k)}\right)^2 \tag{2.61}$$

$$\left(p(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\rangle)^2 - 2\mathbb{E}(y_{(i,j)})p(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\rangle) + \mathbb{E}((y_{(i,j)})^2)\right) \tag{2.62}$$

$$= \frac{1}{m^2\binom{n}{2}}\sum_{(i,j)\in P}\left((U_i^{\tau(i,j)})^{(k)} - (U_j^{\tau(i,j)})^{(k)}\right)^2\left(-p(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\rangle)^2 + \mathbb{E}((y_{(i,j)})^2)\right) \tag{2.63}$$

$$= \frac{1}{m^2\binom{n}{2}}\sum_{(i,j)\in P}\left((U_i^{\tau(i,j)})^{(k)} - (U_j^{\tau(i,j)})^{(k)}\right)^2\left(-p(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\rangle)^2 + \mathbb{E}(y_{(i,j)})\right) \tag{2.64}$$

$$= \frac{1}{m^2\binom{n}{2}}\sum_{(i,j)\in P}\left((U_i^{\tau(i,j)})^{(k)} - (U_j^{\tau(i,j)})^{(k)}\right)^2 \tag{2.65}$$

$$\left(p(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\rangle) - p(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\rangle)^2\right) \tag{2.66}$$

$$\leq \frac{\beta^2}{4m^2} \tag{2.67}$$

where Equation (2.64) is because $y_{(i,j)} \in \{0,1\}$ and where the last line is by definition of

β and since $p(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\rangle) \in [0,1]$ and $x - x^2 \leq \frac{1}{4}$ for $x \in [0,1]$.

Therefore, $\frac{1}{m}\nabla\mathcal{L}_m(w^*) = \sum_{\ell=1}^m X_\ell$ is a sum of i.i.d. mean zero random variables. Hence, each coordinate is also a sum of i.i.d. random variables with mean zero, so Bernstein's inequality applies. Recall Bernstein's inequality:

Theorem 2.6.7 (Bernstein's inequality)**.** *Let X_i be i.i.d. random variables such that $\mathbb{E}(X_i) = 0$ and $|X_i| \leq M$. Then for any $t > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^m X_i > t\right) \leq \exp\left(-\frac{\frac{1}{2}t^2}{\sum \mathbb{E}X_i^2 + \frac{1}{3}Mt}\right).$$

We apply Bernstein's inequality to the k-th coordinate of $\frac{1}{m}\nabla\mathcal{L}_m(w^*)$:

$$\mathbb{P}\left(\left|\frac{1}{m}\nabla\mathcal{L}_m(w^*)^{(k)}\right| > t\right) \leq 2\exp\left(-\frac{\frac{1}{2}t^2}{\frac{\beta^2}{4m} + \frac{\beta t}{3m}}\right) \tag{2.68}$$

since $\sum_{\ell=1}^m \mathbb{E}((X_\ell^{(k)})^2) \leq \frac{\beta^2}{4m}$ and $|X_\ell^{(k)}| \leq \frac{\beta}{m}$.
Since $\|x\|_2 \leq \sqrt{d}\|x\|_\infty$ for any $x \in \mathbb{R}^d$,

$$\begin{aligned}
&\mathbb{P}\left(\left\|\frac{1}{m}\nabla\mathcal{L}_m(w^*)\right\|_2 > t\right) && (2.69)\\
&\leq \mathbb{P}\left(\frac{\sqrt{d}}{m}\left\|\nabla\mathcal{L}_m(w^*)\right\|_\infty > t\right) && (2.70)\\
&= \mathbb{P}\left(\left\|\frac{1}{m}\nabla\mathcal{L}_m(w^*)\right\|_\infty > \frac{t}{\sqrt{d}}\right) && (2.71)\\
&\leq 2d\exp\left(-\frac{\frac{1}{2}\frac{t^2}{d}}{\frac{\beta^2}{4m} + \frac{\beta\frac{t}{\sqrt{d}}}{3m}}\right) \text{ by union bound and inequality Equation (2.68)} && (2.72)\\
&= 2d\exp\left(-\frac{t^2}{\frac{d\beta^2}{2m} + \frac{2\beta t\sqrt{d}}{3m}}\right) && (2.73)\\
&= 2d\exp\left(-\frac{6mt^2}{3d\beta^2 + 4\beta t\sqrt{d}}\right). && (2.74)
\end{aligned}$$

In other words, for $t > 0$, with probability at least $1 - 2d\exp\left(-\frac{6mt^2}{3d\beta^2+4\beta t\sqrt{d}}\right)$,

$$\|\frac{1}{m}\nabla\mathcal{L}_m(w^*)\|_2 \leq t.$$

Let

$$\alpha := 3\beta^2\log(4d/\delta)d + 4\sqrt{d}\beta\log(4d/\delta).$$

Set

$$t = \sqrt{\frac{\alpha}{6m}}.$$

If

$$m \geq \frac{3\beta^2\log(4d/\delta)d + 4\sqrt{d}\beta\log(4d/\delta)}{6} = \frac{\alpha}{6},$$

then

$$2d\exp\left(-\frac{6mt^2}{3d\beta^2+4\beta t\sqrt{d}}\right) \leq \frac{\delta}{2},$$

which we establish below.

If

$$\begin{aligned}
&m \geq \frac{\alpha}{6} && (2.75)\\
\implies\ & m \geq \frac{\alpha(4\beta\log(4d/\delta))^2 d}{6(4\beta\log(4d/\delta))^2 d} && (2.76)\\
\implies\ & m \geq \frac{\alpha(4\beta\log(4d/\delta))^2 d}{6(\alpha - 3\beta^2\log(4d/\delta)d)^2} && (2.77)\\
\implies\ & m \geq \frac{\alpha d}{6\left(\frac{\alpha-3\beta^2\log(4d/\delta)d}{4\beta\log(4d/\delta)}\right)^2} && (2.78)\\
\implies\ & \left(\frac{\alpha - 3\beta^2\log(4d/\delta)d}{4\beta\log(4d/\delta)}\right)^2 \geq \frac{\alpha d}{6m} && (2.79)\\
\implies\ & \frac{\frac{\alpha}{\log(4d/\delta)} - 3\beta^2 d}{4\beta} \geq \sqrt{\frac{\alpha d}{6m}} && (2.80)\\
\implies\ & \frac{\alpha}{\log(4d/\delta)} \geq 4\beta\sqrt{\frac{\alpha d}{6m}} + 3\beta^2 d && (2.81)\\
\implies\ & \frac{\alpha}{4\beta\sqrt{\frac{\alpha d}{6m}} + 3\beta^2 d} \geq \log(4d/\delta) && (2.82)
\end{aligned}$$

$$\implies \frac{t^2 6m}{4\beta t\sqrt{d} + 3\beta^2 d} \geq \log\left(4d/\delta\right) \tag{2.83}$$

$$\implies 2d\exp\left(-\frac{6mt^2}{4\beta t\sqrt{d} + 3\beta^2 d}\right) \leq \frac{\delta}{2} \tag{2.84}$$

$$\tag{2.85}$$

Therefore, if

$$m \geq \frac{3\beta^2 \log\left(4d/\delta\right)d + 4\sqrt{d}\beta\log\left(4d/\delta\right)}{6}$$

with probability at least $1 - \frac{\delta}{2}$,

$$\left\|\frac{1}{m}\nabla\mathcal{L}_m(w^*)\right\|_2 < \sqrt{\frac{3\beta^2 \log\left(4d/\delta\right)d + 4\sqrt{d}\beta\log\left(4d/\delta\right)}{6m}}.$$

□

Lemma 2.6.8. *For $(i,j) \in P$, let $Z_{(i,j)} = (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T$. Let*

$$\lambda := \lambda_{\min}(\mathbb{E}Z_{(i,j)})$$

where for a square matrix U, $\lambda_{\min}(U)$ is the smallest eigenvalue of U. Let

$$\eta := \sigma_{\max}(\mathbb{E}((Z_{(i,j)} - \mathbb{E}Z_{(i,j)})^2))$$

where $\sigma_{\max}(X)$ is the largest singular value of a matrix X. Let

$$\zeta := \max_{(i,j)\in P} \lambda_{\max}(\mathbb{E}Z_{(i,j)} - Z_{(i,j)}),$$

where $\lambda_{\max}(X)$ is the largest eigenvalue of X. The expectation in λ, η, and ζ is taken with respect to a uniformly chosen random pair of items.

Let $\delta > 0$. Under the model assumptions in this section, if $\lambda > 0$ and if

$$m \geq \frac{8\log(2/\delta)(6\eta + \lambda\zeta)}{3\lambda^2},$$

then with probability at least $1-\frac{\delta}{2}$,

$$\frac{1}{m}\sum_{\ell=1}^{m}\Delta^T(U_{i_\ell}^{\tau(i_\ell,j_\ell)}-U_{j_\ell}^{\tau(i_\ell,j_\ell)})(U_{i_\ell}^{\tau(i_\ell,j_\ell)}-U_{j_\ell}^{\tau(i_\ell,j_\ell)})^T\Delta \geq \|\Delta\|_2^2\frac{\lambda}{2}$$

where

$$\Delta = \hat{w}-w^*.$$

Proof. Let

$$X_\ell = \frac{1}{m}(U_{i_\ell}^{\tau(i_\ell,j_\ell)}-U_{j_\ell}^{\tau(i_\ell,j_\ell)})(U_{i_\ell}^{\tau(i_\ell,j_\ell)}-U_{j_\ell}^{\tau(i_\ell,j_\ell)})^T-\frac{1}{m}\mathbb{E}((U_i^{\tau(i,j)}-U_j^{\tau(i,j)})(U_i^{\tau(i,j)}-U_j^{\tau(i,j)})^T)).$$

Notice that $\frac{1}{m}\sum_{\ell=1}^{m}(U_{i_\ell}^{\tau(i_\ell,j_\ell)}-U_{j_\ell}^{\tau(i_\ell,j_\ell)})(U_{i_\ell}^{\tau(i_\ell,j_\ell)}-U_{j_\ell}^{\tau(i_\ell,j_\ell)})^T$ is a sum of random matrices where the randomness is from the random pairs of items that are chosen in the samples. Therefore, bounding the smallest eigenvalue of this random matrix is sufficient to get the desired lower bound as we show.

Since $\mathbb{E}X_\ell = 0$ by construction and X_ℓ is self-adjoint since it is symmetric and real, we apply the following concentration bound to $\sum_{\ell=1}^{m} X_\ell$:

Theorem 2.6.9 (Theorem 1.4 in [Tro12])**.** *Consider a finite sequence* $\{X_k\}$ *of independent, random, self-adjoint matrices with dimension* d*. Assume that each random matrix satisfies* $\mathbb{E}X_k = 0$ *and* $\lambda_{\max}(X_k) \leq R$ *almost surely. Then for all* $t \geq 0$

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_k X_k\right)\geq t\right) \leq d\exp\left(\frac{-t^2/2}{\sigma^2+Rt/3}\right), \tag{2.86}$$

where

$$\sigma^2 := \sigma_{\max}\left(\sum_k \mathbb{E}\left(X_k^2\right)\right).$$

Notice

$$\begin{aligned}\sigma_{\max}\left(\sum_{\ell=1}^{m}\mathbb{E}\left(X_\ell^2\right)\right) &= m\sigma_{\max}(\mathbb{E}\left(X_1^2\right)) \text{ since each } X_\ell \text{ is distributed the same} && (2.87)\\ &= \frac{m}{m^2}\eta && (2.88)\\ &= \frac{1}{m}\eta. && (2.89)\end{aligned}$$

Then applying the above theorem, for $t \geq 0$,

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{\ell=1}^{m} -X_\ell\right) \geq t\right) \leq d\exp\left(\frac{-t^2/2}{\eta/m + \zeta t/(3m)}\right) \tag{2.90}$$

$$\leq d\exp\left(\frac{-3mt^2}{6\eta + 2\zeta t}\right). \tag{2.91}$$

In other words, for all $t \geq 0$, with probability at least $1 - d\exp\left(\frac{-3mt^2}{6\eta+2\zeta t}\right)$,

$$\lambda_{\max}\left(\sum_{\ell=1}^{m} -X_\ell\right) \leq t \tag{2.92}$$

$$\implies \frac{\Delta^T}{\|\Delta\|_2}\left(\sum_{\ell=1}^{m} -X_\ell\right)\frac{\Delta}{\|\Delta\|_2} \leq t \tag{2.93}$$

$$\implies \Delta^T(\mathbb{E}((U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T))- \\ \frac{1}{m}\sum_{\ell=1}^{m}(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})^T)\Delta \leq t\|\Delta\|_2^2 \tag{2.94}$$

$$\implies \Delta^T\left(\mathbb{E}((U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T))\right)\Delta - t\|\Delta\|_2^2 \\ \leq \Delta^T\left(\frac{1}{m}\sum_{\ell=1}^{m}(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})^T\right)\Delta \tag{2.95}$$

$$\implies \|\Delta\|_2^2\frac{\Delta^T}{\|\Delta\|_2}\left(\mathbb{E}((U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T))\right)\frac{\Delta}{\|\Delta\|_2} - t\|\Delta\|_2^2 \\ \leq \Delta^T\left(\frac{1}{m}\sum_{\ell=1}^{m}(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})^T\right)\Delta \tag{2.96}$$

$$\implies (\lambda - t)\|\Delta\|_2^2 \leq \Delta^T\left(\frac{1}{m}\sum_{\ell=1}^{m}(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})^T\right)\Delta \tag{2.97}$$

since $\lambda := \lambda_{\min}(\mathbb{E}((U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T))$.

Set $t = \frac{\lambda}{2}$. Since $\lambda > 0$ by assumption, Equation (2.97) becomes

$$\frac{\lambda}{2}\|\Delta\|_2^2 \leq \Delta^T\left(\frac{1}{m}\sum_{\ell=1}^{m}(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})^T\right)\Delta$$

and holds with probability at least $1 - \frac{\delta}{2}$ if

$$m \geq \frac{8\log(2d/\delta)(6\eta + \lambda\zeta)}{3\lambda^2}$$

since

$$\begin{aligned} d\exp\left(\frac{-3m\frac{\lambda^2}{4}}{6\eta + 2\frac{\lambda}{2}\zeta t}\right) &\leq \frac{\delta}{2} && (2.98)\\ \implies \frac{-3m\frac{\lambda^2}{4}}{6\eta + 2\frac{\lambda}{2}\zeta t} &\leq -\log(2d/\delta) && (2.99)\\ \implies \frac{3m\frac{\lambda^2}{4}}{6\eta + 2\frac{\lambda}{2}\zeta t} &\geq 2\log(2d/\delta) && (2.100)\\ \implies m &\geq \frac{8\log(2d/\delta)(6\eta + \lambda\zeta)}{3\lambda^2}. && (2.101)\\ & && (2.102) \end{aligned}$$

□

Lemma 2.6.10 (Gradient and Hessian of Equation (2.3)). *Given samples S_m, features of the n items $U \in \mathbb{R}^{d\times n}$, and $w \in \mathbb{R}^d$,*

$$\begin{aligned} &\frac{1}{m}\nabla\mathcal{L}_m(w; U, S_m, \tau) && (2.103)\\ &= \frac{1}{m}\sum_{\ell=1}^{m} \frac{\exp(\langle w, U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\rangle)}{1 + \exp(\langle w, U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\rangle)}\left(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\right) - y_\ell\left(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\right)\\ & && (2.104) \end{aligned}$$

and

$$\begin{aligned} &\frac{1}{m}H_{\mathcal{L}_m}(w; U, S_m, \tau) && (2.105)\\ &= \frac{1}{m}\sum_{\ell=1}^{m} \frac{\exp(\langle w, U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\rangle)}{(1 + \exp(\langle w, U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\rangle))^2}(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})^T\\ & && (2.106) \end{aligned}$$

Proof. **Gradient:** Let $f(x) := \log(1+e^x)$ for $x \in \mathbb{R}$ and $g(w;y) := \langle w, y\rangle$ for $w, y \in \mathbb{R}^d$, so

$$\frac{1}{m}\mathcal{L}_m(w;U,S_m,\tau) = \frac{1}{m}\sum_{\ell=1}^{m}(f \circ g)(w; U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}) + y_\ell g(w; U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}).$$

Note

$$f'(x) = \frac{e^x}{1+e^x}$$

and $\nabla_w g(w;y) = y$.

We arrive at the desired result by the chain rule:

$$\begin{aligned}
&\frac{1}{m}\mathcal{L}_m(w;U,S_m,\tau) = &(2.107)\\
&\frac{1}{m}\sum_{\ell=1}^{m} f'(g(w; U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})) &(2.108)\\
&\nabla_w g(w; U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}) - y_\ell \nabla_w g(w; U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}).
\end{aligned}$$

Hessian: Note

$$f''(x) = \frac{e^x(1+e^x) - e^{2x}}{(1+e^x)^2} = \frac{e^x}{(1+e^x)^2}.$$

Let $[H_{\mathcal{L}_m}(w;U,S_m)]_k$ be the kth row of the Hessian and $\nabla\mathcal{L}_m(w;U,S_m)^{(k)}$ be the kth entry of the gradient. Then by the chain rule again,

$$\begin{aligned}
&[H_{\mathcal{L}_m}(w;U,S_m)]_k^T\\
&= \nabla_w(\nabla\mathcal{L}_m(w;U,S_m)^{(k)})\\
&= \sum_{\ell=1}^{m}((U_{i_\ell}^{\tau(i_\ell,j_\ell)})^{(k)} - (U_{j_\ell}^{\tau(i_\ell,j_\ell)})^{(k)}) f''(g(w; U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}))\nabla_w g(w; U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)})\\
&= \sum_{\ell=1}^{m}\frac{\exp(\langle w, U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\rangle)}{(1+\exp(\langle w, U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}\rangle))^2}((U_{i_\ell}^{\tau(i_\ell,j_\ell)})^{(k)} - (U_{j_\ell}^{\tau(i_\ell,j_\ell)})^{(k)})(U_{i_\ell}^{\tau(i_\ell,j_\ell)} - U_{j_\ell}^{\tau(i_\ell,j_\ell)}),
\end{aligned}$$

which proves the claim.

□

Lemma 2.6.11. *Let $h(x) = \frac{e^x}{(1+e^x)^2}$. Then $h(x)$ is symmetric and decreases on $[0,\infty)$.*

Proof. Symmetry:

$$h(-x) = \frac{e^{-x}}{(1+e^{-x})^2} \tag{2.109}$$

$$= \frac{e^{-x}}{e^{-2x}(e^x+1)^2} \tag{2.110}$$

$$= \frac{e^x}{(e^x+1)^2} \tag{2.111}$$

$$= h(x). \tag{2.112}$$

Decreasing on $[0, \infty)$:

Note

$$h'(x) = \frac{e^x(1+e^x)^2 - e^{2x}2(1+e^x)}{(1+e^x)^4} \tag{2.113}$$

$$= \frac{e^x(1+e^x) - e^{2x}2}{(1+e^x)^3} \tag{2.114}$$

$$= \frac{e^x(1-e^x)}{(1+e^x)^3} \tag{2.115}$$

$$\leq 0 \tag{2.116}$$

for $x \in [0, \infty)$ since on this interval, $1 - e^x \leq 0$ but $e^x, (1+e^x)^3 \geq 0$. Thus $h(x)$ is decreasing on $[0, \infty)$. □

2.6.6 Specific Selection Functions: Proofs of Corollaries 2.3.4 and 2.3.5

In this section, we present the full lower bounds on the number of samples and upper bound on the estimation error. The definitions of the constants that appear in the main text, i.e. C_3 and C_4, appear at the end of the applicable proofs.

Proof of Corollary 2.3.4

The following lemma is a straight forward generalization from [NOS16], but we include the proof for completeness. We need this lemma to prove Corollary 2.3.4.

Lemma 2.6.12. *Let $U \in \mathbb{R}^{d\times n}$. Assume that the columns of U sum to 0: $\sum_{i=1}^{n} U_i = 0$. Then*

$$\mathbb{E}((U_i - U_j)(U_i - U_j)^T) = \frac{n}{\binom{n}{2}} UU^T$$

where the expectation is with respect to a uniformly at randomly chosen pair of items.

Proof. Let $e_i \in \mathbb{R}^n$ denote the i-th standard basis vector, $I_{n\times n}$ denote the $n \times n$ identity matrix, and $\mathbb{1} \in \mathbb{R}^n$ be the vector of all ones. Since the expectation is over a uniformly chosen pair of items $(i,j) \in P$,

$$\mathbb{E}((U_i - U_j)(U_i - U_j)^T) \tag{2.117}$$

$$= \mathbb{E}(U(e_i - e_j)(e_i - e_j)^T U^T) \tag{2.118}$$

$$= \frac{1}{\binom{n}{2}} U \left(\sum_{(i,j)\in P} e_i e_i^T - e_i e_j^T - e_j e_i^T + e_j e_j^T \right) U^T \tag{2.119}$$

$$= \frac{1}{\binom{n}{2}} U \left((n-1) \sum_{i=1}^{n} e_i e_i^T - \sum_{(i,j)\in P} e_i e_j^T + e_j e_i^T \right) U^T, \text{ each item is in } n-1 \text{ comparisons} \tag{2.120}$$

$$= \frac{1}{\binom{n}{2}} U \left((n-1) I_{n\times n} - \sum_{(i,j)\in P} e_i e_j^T + e_j e_i^T \right) U^T \tag{2.121}$$

$$= \frac{1}{\binom{n}{2}} U \left((n-1) I_{n\times n} - \left(\mathbb{1}\mathbb{1}^T - I_{n\times n} \right) \right) U^T \text{ explained below} \tag{2.122}$$

$$= \frac{1}{\binom{n}{2}} U \left(n I_{n\times n} - \mathbb{1}\mathbb{1}^T \right) U^T \tag{2.123}$$

$$= \frac{1}{\binom{n}{2}} (nUU^T - U\mathbb{1}\mathbb{1}^T U^T) \tag{2.124}$$

$$= \frac{n}{\binom{n}{2}} UU^T \text{since } U\mathbb{1} = \sum_{i=1}^{n} U_i = \mathbf{0} \text{ by assumption.} \tag{2.125}$$

Equation (2.122) is because $e_i e_j^T$ is the matrix with a 1 in the i-th row and j-th column and 0 elsewhere and we are summing over all $(i,j) \in [n] \times [n]$ where $i < j$. Thus, the sum equals $\mathbb{1}\mathbb{1}^T - I_{n\times n}$, which is the matrix with ones everywhere except for the diagonal. □

Corollary 2.6.13 (Restatement of Corollary 2.3.4)**.** *Assume the set-up stated in the beginning of Section 2.2. For the selection function τ, suppose $\tau(i,j) = [d]$ for any $(i,j) \in P$. In other words, all the features are used in each pairwise comparison. Assume $n > d$. Let $\nu := \max\{\max_{(i,j)\in P} \|U_i - U_j\|_2^2, 1\}$. Without loss of generality, assume the columns of U sum to zero: $\sum_{i=1}^n U_i = 0$. Then,*

$$\lambda = \frac{n\lambda_{\min}(UU^T)}{\binom{n}{2}},$$

$$\zeta \leq \nu + \frac{n\lambda_{\max}(UU^T)}{\binom{n}{2}},$$

and

$$\eta \leq \frac{\nu n\lambda_{\max}(UU^T)}{\binom{n}{2}} + \frac{n^2\lambda_{\max}(UU^T)^2}{\binom{n}{2}^2}.$$

Let

$$m_1 = \frac{3\beta^2 \log{(4d/\delta)}d + 4\sqrt{d}\beta\log{(4d/\delta)}}{6}$$

and

$$m_2 = \frac{48\log(2d/\delta)\binom{n}{2}^2}{3n^2\lambda_{\min}(UU^T)^2}\left(\frac{\nu n\lambda_{\max}(UU^T)}{\binom{n}{2}} + \frac{n^2\lambda_{\max}(UU^T)^2}{\binom{n}{2}^2}\right) + \frac{8\log(2d/\delta)\binom{n}{2}}{3n\lambda_{\min}(UU^T)}\left(\nu + \frac{n\lambda_{\max}(UU^T)}{\binom{n}{2}}\right).$$

Let $\delta > 0$. Hence, if

$$m \geq \max\{m_1, m_2\},$$

then with probability at least $1-\delta$,

$$\|w^* - \hat{w}\|_2 \leq \frac{4(1+\exp(b^*))^2\binom{n}{2}}{\exp(b^*)n\lambda_{\min}(UU^T)}\sqrt{\frac{3\beta^2\log{(4d/\delta)}d + 4\sqrt{d}\beta\log{(4d/\delta)}}{6m}}. \tag{2.126}$$

Proof. Throughout this proof, we use U_i instead of $U_i^{\tau(i,j)}$ for any items i,j since $\tau(i,j)$ selects all coordinates.

If $\sum_{i=1}^n U_i \neq 0$, simply subtract the column mean, $\bar{U} := \frac{1}{n}\sum_{i=1}^n U_i$, from each column. This operation does not affect the underlying pairwise probabilities since

$$\mathbb{P}(\text{item } i \text{ beats item } j) = \frac{1}{1+\exp(-\langle w^*, U_i - U_j\rangle)} \tag{2.127}$$

$$= \frac{1}{1+\exp(-\langle w^*, (U_i - \bar{U}) - (U_j - \bar{U})\rangle)}. \tag{2.128}$$

Let $\widetilde{U} = U(I - \frac{1}{n}\mathbb{1}\mathbb{1}^T)$ be the centered version of U, i.e. where we subtract $\bar{U}$ from each column of U. Since $n > d$ and by Proposition 2.6.14, if $\lambda_{\min}(U) > 0$, then $\lambda_{\min}(\widetilde{U}) > 0$ generically. Therefore, WLOG, we may assume $\sum_{i=1}^n U_i = 0$.

First, we simplify λ. By Lemma 2.6.12,

$$\lambda = \lambda_{\min}(\mathbb{E}((U_i - U_j)(U_i - U_j)^T)) = \frac{n\lambda_{\min}(UU^T)}{\binom{n}{2}}.$$

Second, we upper bound ζ. Let $(k, \ell) \in P$, then

$$\lambda_{\max}\left(\mathbb{E}(U_i - U_j)(U_i - U_j)^T - (U_k - U_\ell)(U_k - U_\ell)^T\right) \tag{2.129}$$

$$= \lambda_{\max}\left(\frac{n}{\binom{n}{2}}UU^T - (U_k - U_\ell)(U_k - U_\ell)^T\right) \text{ by Lemma 2.6.12} \tag{2.130}$$

$$\leq \lambda_{\max}\left(\frac{n}{\binom{n}{2}}UU^T\right) + \lambda_{\max}\left((U_k - U_\ell)(U_k - U_\ell)^T\right) \tag{2.131}$$

$$= \lambda_{\max}\left(\frac{n}{\binom{n}{2}}UU^T\right) + \|(U_k - U_\ell)\|_2^2 \tag{2.132}$$

$$\leq \lambda_{\max}\left(\frac{n}{\binom{n}{2}}UU^T\right) + \nu, \tag{2.133}$$

$$\tag{2.134}$$

where the second to last line is since the largest eigenvalue of a rank one matrix xx^T is $\|x\|_2^2$ and the last line is by definition of ν.

Third, we upper bound η. Let $e_i \in \mathbb{R}^n$ denote the i-th standard basis vector. For any random variable X, we have

$$\mathbb{E}(X-\mathbb{E}(X))^2=\mathbb{E}(X^2)-\mathbb{E}(X)^2. \tag{2.135}$$

Furthermore, since η is the largest singular value of a symmetric matrix squared, the largest eigenvalue of that matrix is also equal to η. Therefore,

$$\eta=\lambda_{\max}\left(\mathbb{E}((U_i-U_j)(U_i-U_j)^T(U_i-U_j)(U_i-U_j)^T)-\mathbb{E}((U_i-U_j)(U_i-U_j)^T)^2\right).$$

Most steps are explained below after the equations. Because the expectation is with respect to a uniformly at random pair of items $(i,j)\in P$ and by Lemma 2.6.12,

$$\lambda_{\max}\left(\mathbb{E}((U_i-U_j)(U_i-U_j)^T(U_i-U_j)(U_i-U_j)^T)-\mathbb{E}((U_i-U_j)(U_i-U_j)^T)^2\right) \tag{2.136}$$

$$=\lambda_{\max}\left(\frac{1}{\binom{n}{2}}\sum_{(i,j)\in P}(U_i-U_j)(U_i-U_j)^T(U_i-U_j)(U_i-U_j)^T-\frac{n^2}{\binom{n}{2}^2}UU^TUU^T\right) \tag{2.137}$$

$$=\lambda_{\max}\left(\frac{1}{\binom{n}{2}}\sum_{(i,j)\in P}\left((U_i-U_j)^T(U_i-U_j)\right)(U_i-U_j)(U_i-U_j)^T-\frac{n^2}{\binom{n}{2}^2}UU^TUU^T\right) \tag{2.138}$$

$$=\lambda_{\max}\left(\frac{1}{\binom{n}{2}}\sum_{(i,j)\in P}\left((U_i-U_j)^T(U_i-U_j)\right)U(e_i-e_j)(e_i-e_j)^TU^T-\frac{n^2}{\binom{n}{2}^2}UU^TUU^T\right) \tag{2.139}$$

$$\le\lambda_{\max}\left(\frac{1}{\binom{n}{2}}\sum_{(i,j)\in P}\left((U_i-U_j)^T(U_i-U_j)\right)U(e_i-e_j)(e_i-e_j)^TU^T\right) \tag{2.140}$$
$$+\lambda_{\max}\left(\frac{n^2}{\binom{n}{2}^2}UU^TUU^T\right)$$

$$=\max_x\frac{x^T}{\|x\|}\left(\frac{1}{\binom{n}{2}}\sum_{(i,j)\in P}\left((U_i-U_j)^T(U_i-U_j)\right)U(e_i-e_j)(e_i-e_j)^TU^T\right)\frac{x}{\|x\|} \tag{2.141}$$
$$+\lambda_{\max}\left(\frac{n^2}{\binom{n}{2}^2}UU^TUU^T\right)$$

$$= \max_x \left(\frac{1}{\binom{n}{2}} \sum_{(i,j)\in P} \left((U_i - U_j)^T(U_i - U_j)\right) \frac{x^T}{\|x\|} U(e_i - e_j)(e_i - e_j)^T U^T \frac{x}{\|x\|} \right)$$

$$+ \lambda_{\max}\left(\frac{n^2}{\binom{n}{2}^2} UU^TUU^T \right) \tag{2.142}$$

$$\leq \max_x \left(\frac{\nu}{\binom{n}{2}} \sum_{(i,j)\in P} \frac{x^T}{\|x\|} U(e_i - e_j)(e_i - e_j)^T U^T \frac{x}{\|x\|} \right) + \lambda_{\max}\left(\frac{n^2}{\binom{n}{2}^2} UU^TUU^T \right) \tag{2.143}$$

$$= \lambda_{\max}\left(\frac{\nu}{\binom{n}{2}} \sum_{(i,j)\in P} U(e_i - e_j)(e_i - e_j)^T U^T \right) + \lambda_{\max}\left(\frac{n^2}{\binom{n}{2}^2} UU^TUU^T \right) \tag{2.144}$$

$$= \frac{\nu n}{\binom{n}{2}} \lambda_{\max}\left(UU^T\right) + \frac{n^2}{\binom{n}{2}^2} \lambda_{\max}\left(UU^T\right)^2 \text{ by Lemma 2.6.12.} \tag{2.145}$$

$$\tag{2.146}$$

Equation (2.138) is because $(U_i - U_j)^T(U_i - U_j) \in \mathbb{R}$. Equation (2.143) is because $(U_i - U_j)^T(U_i - U_j) \geq 0$ and $\frac{x^T}{\|x\|} U(e_i - e_j)(e_i - e_j)^T U^T \frac{x}{\|x\|} \geq 0$.

Now that we have bounds on η and ζ and a simplified form for λ, we apply Theorem 2.3.2, completing the proof.

Now we explain how to get from these results to those in the main text of this chapter with the order terms. The $O(\cdot)$ upper bound on the estimation error is easy to see. The value of C_1 is given at the end of the proof of Theorem 2.3.2. The only remaining term to explain from the main text of this chapter is the upper bound of $\frac{8\log(2d/\delta)(6\eta+\lambda\zeta)}{3\lambda^2}$, which gives us a lower bound on the number of samples required.

In particular,

$$\frac{8\log(2d/\delta)(6\eta + \lambda\zeta)}{3\lambda^2} \tag{2.147}$$

$$= \frac{48\log(2d/\delta)\eta}{3\lambda^2} + \frac{8\log(2d/\delta)\zeta}{3\lambda} \tag{2.148}$$

$$= \frac{48\log(2d/\delta)\binom{n}{2}^2}{3n^2\lambda_{\min}(UU^T)^2} \left(\frac{\nu n \lambda_{\max}(UU^T)}{\binom{n}{2}} + \frac{n^2\lambda_{\max}(UU^T)^2}{\binom{n}{2}^2} \right) \tag{2.149}$$

$$+ \frac{8\log(2d/\delta)\binom{n}{2}}{3n\lambda_{\min}(UU^T)} \left(\nu + \frac{n\lambda_{\max}(UU^T)}{\binom{n}{2}} \right)$$

$$= \frac{48\log(2d/\delta)}{3\lambda_{\min}(UU^T)^2}\left(\frac{\binom{n}{2}\nu\lambda_{\max}(UU^T)}{n} + \lambda_{\max}(UU^T)^2\right) \tag{2.150}$$

$$+ \frac{8\log(2d/\delta)}{3\lambda_{\min}(UU^T)}\left(\frac{\binom{n}{2}\nu}{n} + \lambda_{\max}(UU^T)\right)$$

$$\leq \frac{48\log(2d/\delta)}{3\lambda_{\min}(UU^T)^2}\left(\frac{\binom{n}{2}\nu\lambda_{\max}(UU^T)}{n} + n\lambda_{\max}(UU^T)^2\right) \tag{2.151}$$

$$+ \frac{8\log(2d/\delta)}{3\lambda_{\min}(UU^T)}\left(\frac{\binom{n}{2}\nu}{n} + n\lambda_{\max}(UU^T)\right)$$

$$\leq \frac{48\log(2d/\delta)}{3\lambda_{\min}(UU^T)^2}\left(n\nu\lambda_{\max}(UU^T) + n\lambda_{\max}(UU^T)^2\right) + \frac{48\log(2d/\delta)}{3\lambda_{\min}(UU^T)}\left(n\nu + n\lambda_{\max}(UU^T)\right) \tag{2.152}$$

$$\leq \frac{48\log(\frac{2d}{\delta})n\nu}{3}\left(\frac{\lambda_{\max}(UU^T)}{\lambda_{\min}(UU^T)^2} + \frac{\lambda_{\max}(UU^T)^2}{\lambda_{\min}(UU^T)^2} + \frac{1}{\lambda_{\min}(UU^T)} + \frac{\lambda_{\max}(UU^T)}{\lambda_{\min}(UU^T)}\right) \tag{2.153}$$

$$\leq \frac{2 * 48\log(2d/\delta)n\nu}{3}\left(\frac{\lambda_{\max}(UU^T)}{\lambda_{\min}(UU^T)^2} + \frac{\lambda_{\max}(UU^T)^2}{\lambda_{\min}(UU^T)^2} + \frac{1}{\lambda_{\min}(UU^T)}\right) \tag{2.154}$$

$$= C_3\log(2d/\delta)n\nu\left(\frac{\lambda_{\max}(UU^T)}{\lambda_{\min}(UU^T)^2} + \frac{\lambda_{\max}(UU^T)^2}{\lambda_{\min}(UU^T)^2} + \frac{1}{\lambda_{\min}(UU^T)}\right) \tag{2.155}$$

where Equation (2.153) is because $\nu \geq 1$, Equation (2.154) is because $\frac{\lambda_{\max}(UU^T)}{\lambda_{\min}(UU^T)} \geq 1$, and $C_3 = 2 * 48/3$. We remark that the assumption that $\nu \geq 1$ was made to simplify the upper bound and is not required. □

As we mentioned, we can assume U is centered without loss of generality, because we can subtract the mean column from all columns if they are not centered. However one may wonder then what happens to $\lambda_{\min}(UU^T) = \sqrt{\sigma_{\min}(U)}$ once U is centered. Since we assume $n > d$, it will generically be non-zero, as we make precise in the following proposition.

Proposition 2.6.14. *Given an arbitrary rank-d, $d \times n$ matrix $\widetilde{U}$, let U be its centered version, i.e. $U = \widetilde{U}(I - \frac{1}{n}\mathbb{1}\mathbb{1}^T)$. Then $\sigma_{\min}(U) = 0$ if and only if the all-ones vector is in the row space of $\widetilde{U}$.*

Proof. Suppose $\widetilde{U}$ contains the all-ones vector in its row space, and therefore let v be such that $\widetilde{U}^T v = \mathbb{1}$. Let $Q = (I - \frac{1}{n}\mathbb{1}\mathbb{1}^T)$. Then

$$U^T v = Q\widetilde{U}^T v = 0$$

since the all-ones vector is in the nullspace of Q, implying that $\sigma_{\min}(U) = 0$. For the other direction suppose $\sigma_{\min}(U) = 0$. Then there exists a vector $v \neq 0$ such that

$$0 = U^T v = Q\widetilde{U}^T v.$$

This implies either that $\widetilde{U}^T v = 0$ or $\widetilde{U}^T v$ is in the nullspace of Q. Since we assumed that $\widetilde{U}$ has full row rank, then it must be that $\widetilde{U}^T v = \mathbb{1}$, the only vector in the nullspace of Q. □

Discussion of Corollary 2.3.4 as compared to related work

While our sample complexity theorem for MLE of the parameters of FBTL is novel to the best of our knowledge, there are some related results that merit a comparison. First, there is a result in [SR18] that gives sample complexity results for a different estimator of FBTL parameters under a substantially different sampling model. In particular, they only allow pairs to be sampled from a graph, and then for each sampled pair they observe a fixed number of pairwise comparisons. In their results one can see that as the number of pairs sampled increases, their error upper bound increases and the probability of their resulting bound also decreases. In contrast, our analysis shows that our error bound decreases as m increases, and the probability of our resulting bound remains constant.

Second, we can also attempt a comparison to the bounds for BTL without features in [NRW$^+$12], despite the fact that with standard basis features, our bound does not apply because $\lambda = 0$. Assuming that $\exp(b^*)/\lambda$ is a constant in our bound and that $\nu\bar{\lambda}$ is a constant, we roughly have an error bound of $O(1)$ given $m = \Theta(n^2(\beta^2 + \beta)d\log(d/\delta))$ samples. The result in [NRW$^+$12] instead has that $m = \Theta(d^2 \log d)$ gives an error bound of $O(1)$ with probability $1 - \frac{2}{d}$, recalling that in their setting $d = n$. So if we can tighten bounds that require β in our proof, our results may compare favorably.

Recall the definition of β in Equation (2.4): $\beta := \max_{(i,j)\in P} \|U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\|_\infty$. In our proof, we use this to bound differences between feature vectors at Equation (2.67). In particular, we bound $\frac{1}{\binom{n}{2}} \sum_{(i,j)\in P} \left((U_i^{\tau(i,j)})^{(k)} - (U_j^{\tau(i,j)})^{(k)} \right)^2 \leq \beta^2$. If we instead directly made the assumption that

$$\widetilde{\beta}^2 := \frac{1}{\binom{n}{2}} \max_{k\in[d]} \sum_{(i,j)\in P} \left((U_i^{\tau(i,j)})^{(k)} - (U_j^{\tau(i,j)})^{(k)}\right)^2 ,$$

we could replace β with $\widetilde{\beta}$ directly in our bounds. Assume $\widetilde{\beta} \leq 1/n^2$. Then our sample complexity would reduce to $m = \Theta(d\log(d/\delta)) = \Theta(d\log(2d^2)) = \Theta(d\log(d))$ where recall $\delta = \frac{2}{d}$, beating the complexity in [NRW+12]. However, it is not clear in general what impact the assumption that $\widetilde{\beta} \leq 1/n^2$ would have on the minimum eigenvalue of UU^T. Indeed, the standard basis vectors are a special case where $\widetilde{\beta} \leq 1/n$, and as we pointed out, for this special case $\lambda = 0$.

Third, although there are crucial differences between our model and the model in [SW17] that make a direct comparison impossible, we attempt to roughly compare results. The first difference is that they assume the feature vectors of the items are standard basis vectors, which means our bounds do not apply just as in the comparison with [NRW+12]. The second difference, perhaps the most crucial, is that we make different assumptions about how the intransitive pairwise comparisons are related to the ranking. In [SW17], the items are ranked based on the probability that one items beats any other item chosen uniformly at random. There are scenarios where the true ranking in our model is not the same as the true ranking in [SW17]. The third difference is that we assume that pairs are drawn uniformly at random, whereas they assume each pair $(i, j) \in P$ is drawn $x_{i,j}$ times where $x_{i,j} \sim \text{Binom}(r, p)$ for $r, p > 0$.

Their result (Theorem 2) roughly says with probability $1/n^{13}$, if the gap between a pair of consecutively ranked items' scores is at least $\sqrt{\log n/(npr)}$, then their algorithm learns the ranking exactly. We compare to our Corollary 2.3.6 with $k = 1$ and $\delta = \frac{1}{n^{13}}$ though again we emphasize an exact comparison is impossible because our model is not a special case of theirs or vice versa. Our corollary says with enough samples with high probability, we learn the ranking exactly. On average, their sampling method will see $O(n^2rp)$ samples, so a reasonable way to compare results is to show the required number of samples in our method is comparable to $O(n^2rp)$. If we assume that $\beta, \eta, \zeta, \lambda$, and M are all constant, $\alpha_k = \sqrt{\log n/(npr)}$ which is their assumed gap between scores, and $d = n$, the number of samples we require is $\max\{n\log(n * n^{13}), \log(n), n\log(n * n^{13})npr/\log(n)\} = O(n^2pr)$, matching their bounds.

Fourth, the set-up of [HSR$^+$19] is the same as [SW17] except it considers the adaptive setting. If the gaps of the utilities of consecutively ranked items are constant and denoted by Δ, then under the same assumptions in the discussion about [SW17], our Corollary 2.3.6 is slightly better by a log factor than their Theorem 1a: $O(\log(n/\delta)n/(\Delta)^2))$ vs. $O(\log(n/\delta)n\log(2\log(2/\Delta))/(\Delta)^2))$. However, if many gaps between scores are large and only some gaps between scores are small, their adaptive method is better than our Corollary 2.3.6. This is not surprising since they can adaptively chose which pair to sample next based on the past pairwise comparisons, whereas we consider the passive setting.

Proof of Corollary 2.3.5

Corollary 2.6.15 (Restatement of Corollary 2.3.5)**.** *Assume the set-up stated in the beginning of Section 2.2. Assume that for any $(i,j) \in P$, $|\tau(i,j)| = 1$. Partition $P = \sqcup_{k=1}^{d} P_k$ into d sets where $(i,j) \in P_k$ if $\tau(i,j) = \{k\}$ for $k \in [d]$. Let $\epsilon := \min_{(i,j)\in P} \|U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\|_\infty$. Then*

$$\lambda \geq \frac{\epsilon^2}{\binom{n}{2}} \min_{k\in[d]} |P_k|,$$

$$\zeta \leq \beta^2 + \frac{\beta^2}{\binom{n}{2}} \max_{k\in[d]} |P_k|,$$

and

$$\eta \leq \frac{\beta^4}{\binom{n}{2}} \max_{k\in[d]} \left(|P_k| + \frac{|P_k|^2}{\binom{n}{2}} \right).$$

Furthermore, let

$$m_1 = \frac{3\beta^2 \log{(4d/\delta)} d + 4\sqrt{d}\beta \log{(4d/\delta)}}{6}$$

and let

$$m_3 := \frac{48\log(2d/\delta)\beta^4 \max_{k\in[d]} \left(\binom{n}{2}|P_k| + |P_k|^2\right)}{3\epsilon^4 \min_{k\in[d]} |P_k|^2} + \frac{8\log(2d/\delta)\beta^2 \left(\binom{n}{2} + \max_{k\in[d]} |P_k|\right)}{3\epsilon^2 \min_{k\in[d]} |P_k|}.$$

Let $\delta > 0$. If $m \geq \max\{m_1, m_3\}$, then with probability at least $1 - \delta$,

$$\|w^* - \hat{w}\|_2 \leq \frac{4(1+\exp(b^*))^2\binom{n}{2}}{\exp(b^*)\epsilon^2 \min_{k\in[d]} |P_k|}\sqrt{\frac{3\beta^2\log(4d/\delta)d + 4\sqrt{d}\beta\log(4d/\delta)}{6m}},$$

where the randomness is from the randomly chosen pairs and the outcomes of the pairwise comparisons.

Proof. Note that $|P_k| > 0$, so that $\lambda > 0$, for all $k \in [d]$ if the model is identifiable. Let $U_i^{(j)}$ be the j-th coordinate of the vector U_i, e_i be the i-th standard basis vector, and for a vector x, let $\operatorname{diag}(x)$ be the diagonal matrix whose (i,i)-th entry is the i-th entry of x.

First we simplify and bound λ. Since each pair of items are chosen uniformly at random,

$$\begin{aligned}
&\mathbb{E}((U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T) && (2.156)\\
&= \frac{1}{\binom{n}{2}}\sum_{(i,j)\in P}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T && (2.157)\\
&= \frac{1}{\binom{n}{2}}\sum_{k=1}^{d}\sum_{(i,j)\in P_k}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T && (2.158)\\
&= \frac{1}{\binom{n}{2}}\sum_{k=1}^{d}\left(\sum_{(i,j)\in P_k}(U_i^{(k)} - U_j^{(k)})^2\right)\operatorname{diag}(e_k), && (2.159)
\end{aligned}$$

which is a diagonal matrix. Therefore,

$$\begin{aligned}
\lambda &= \frac{1}{\binom{n}{2}}\min_{k\in[d]}\left(\sum_{(i,j)\in P_k}(U_i^{(k)} - U_j^{(k)})^2\right) && (2.160)\\
&\geq \frac{\epsilon^2}{\binom{n}{2}}\min_{k\in[d]}|P_k|. && (2.161)
\end{aligned}$$

Second, we simplify and bound ζ. Since $|\tau(k,j)| = 1$ for all $k, j \in P$, let $U_i^{(\tau(k,j))}$ denote the coordinate of U_i corresponding to the only element in $\tau(k,j)$. Define $e_{\tau(k,j)}$ similarly, which is one of the standard basis vectors. From the proof of bounding λ in Equation (2.157) to Equation (2.159), we have $\mathbb{E}((U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T) =$

$\frac{1}{\binom{n}{2}}\sum_{k=1}^{d}\left(\sum_{(i,j)\in P_k}(U_i^{(k)}-U_j^{(k)})^2\right)\text{diag}(e_k)$. Let $U(i,j) := U_i^{\tau(i,j)} - U_j^{\tau(i,j)}$. Thus,

$$\zeta = \max_{(\ell,p)\in P} \lambda_{\max}(\mathbb{E}(U(i,j)U(i,j)^T) - U(\ell,p)U(\ell,p))^T \tag{2.162}$$

$$= \max_{(\ell,p)\in P} \lambda_{\max}\left(\frac{1}{\binom{n}{2}}\sum_{k=1}^{d}\left(\sum_{(i,j)\in P_k} U(i,k)^2\right)\text{diag}(e_k) - U(\ell,p)U(\ell,p)^T\right) \tag{2.163}$$

$$= \max_{(\ell,p)\in P} \lambda_{\max}\left(\frac{1}{\binom{n}{2}}\sum_{k=1}^{d}\left(\sum_{(i,j)\in P_k} U(i,k)^2\right)\text{diag}(e_k) - U(\ell,p)^2\text{diag}(e_{\tau(\ell,p)})\right) \tag{2.164}$$

$$\leq \beta^2\left(\max_{k\in[d]}\left(\frac{|P_k|}{\binom{n}{2}}+1\right)\right) \tag{2.165}$$

$$\tag{2.166}$$

since the maximum eigenvalue of a diagonal matrix is bounded by the absolute value of its largest entry. We have also applied the triangle inequality and the definition of β since $|\tau(i,j)| = 1$ for all $(i,j) \in P$.

Third, we simplify η. First notice from the proof of bounding λ from Equation (2.157) to Equation (2.159),

$$\left(\mathbb{E}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T\right)^2 \tag{2.167}$$

$$= \left(\frac{1}{\binom{n}{2}}\sum_{k=1}^{d}\left(\sum_{(i,j)\in P_k}(U_i^{(k)}-U_j^{(k)})^2\right)\text{diag}(e_k)\right)^2 \tag{2.168}$$

$$= \frac{1}{\binom{n}{2}^2}\sum_{k=1}^{d}\left(\sum_{(i,j)\in P_k}(U_i^{(k)}-U_j^{(k)})^2\right)^2\text{diag}(e_k), \tag{2.169}$$

since the matrices above are diagonal.

Also,

$$\mathbb{E}(((U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T)^2) \tag{2.170}$$

$$= \mathbb{E}((U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T) \tag{2.171}$$

$$= \frac{1}{\binom{n}{2}} \sum_{k=1}^{d} \sum_{(i,j)\in P_k} (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T \tag{2.172}$$

$$= \frac{1}{\binom{n}{2}} \sum_{k=1}^{d} \left(\sum_{(i,j)\in P_k} (U_i^{(k)} - U_j^{(k)})^4 \right) \mathrm{diag}(e_k), \tag{2.173}$$

For any random variable X, we have

$$\mathbb{E}(X - \mathbb{E}(X))^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2. \tag{2.174}$$

Therefore,

$$\eta = \frac{1}{\binom{n}{2}} \sigma_{\max} \left(\sum_{k=1}^{d} \left(\sum_{(i,j)\in P_k} (U_i^{(k)} - U_j^{(k)})^4 - \frac{1}{\binom{n}{2}} \left(\sum_{(i,j)\in P_k} (U_i^{(k)} - U_j^{(k)})^2 \right)^2 \right) \mathrm{diag}(e_k) \right) \tag{2.175}$$

$$\leq \frac{\beta^4}{\binom{n}{2}} \max_{k\in[d]} \left(|P_k| + \frac{|P_k|^2}{\binom{n}{2}} \right) \tag{2.176}$$

since the largest singular value of a diagonal matrix is bounded by the largest entry of the diagonal in absolute value. We have also applied the triangle inequality and definition of β.

The remainder of the corollary follows by applying the bounds on λ, ζ and η to Theorem 2.3.2.

Now we explain how to get from these results to those in the main text of this chapter with the order terms. The $O(\cdot)$ upper bound on the estimation error is easy to see. The value of C_1 is given at the end of the proof of Theorem 2.3.2. Finally, it is easy to see $C_4 = 48/3$ in the main text of this chapter. □

Tightening the bounds of Corollary 2.3.5

Still in the setting where the selection function chooses one coordinate per pair, assume $|P_i| \approx |P_j|$ for all $i, j \in [d]$, where P_i is defined in Corollary 2.3.5. Then, as we have stated in the main text, $\lambda, \eta, \zeta = O(1/d)$, and so by Corollary 2.3.5, $\Omega(d^3 \log(d/\delta))$ samples ensures the estimation error is $O(1)$. However, by tightening a bound used in the proof of Theorem 2.3.2, we can show $\Omega(d^2 \log(d/\delta))$ samples ensures the estimation error is $O(1)$.

Recall the definition of β in Equation (2.4): $\beta := \max_{(i,j)\in P} \|U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\|_\infty$. In our proof, we use this to bound differences between feature vectors at Equation (2.67). In particular, for $k \in [d]$ we bound $\frac{1}{\binom{n}{2}} \sum_{(i,j)\in P} \left((U_i^{\tau(i,j)})^{(k)} - (U_j^{\tau(i,j)})^{(k)} \right)^2 \leq \beta^2$. For any $k \in [d]$, since $|P_i| \approx |P_j|$ for all $i, j \in [d]$, each coordinate is chosen approximately $\binom{n}{2}/d$ times. Therefore, $\frac{1}{\binom{n}{2}} \sum_{(i,j)\in P} \left((U_i^{\tau(i,j)})^{(k)} - (U_j^{\tau(i,j)})^{(k)} \right)^2 \leq \beta^2 \ d$ since only $\binom{n}{2}/d$ of the $\binom{n}{2}$ terms in the sum are non-zero. We can now replace β with $\beta/\sqrt{d}$ in Corollary 2.3.5. Therefore, $\Omega(d^2 \log(d/\delta))$ samples ensures the estimation error is $O(1)$ since $\lambda, \eta, \zeta = O(1/d)$.

2.6.7 Proof of Corollary 2.3.6

In this section, we present the full lower bounds on the number of samples and upper bound on the estimation error. The definitions of the constants that appear in the main text, i.e. C_5, appear at the end of the proof.

Corollary 2.6.16 (restatement of Corollary 2.3.6: sample complexity of learning the ranking)**.** *Assume the set-up of Theorem 2.3.2. Pick $k \in [\binom{n}{2}]$. Let α_k be the k-th smallest number in $\{|\langle w^*, U_i - U_j\rangle| : (i,j) \in P\}$. Let $M := \max_{i\in[n]} \|U_i\|_2$. Let $\gamma^* : [n] \to [n]$ be the ranking obtained from w^* by sorting the items by their full-feature utilities $\langle w^*, U_i\rangle$ where $\gamma^*(i)$ is the position of item i in the ranking. Define $\hat{\gamma}$ similarly but for the estimated ranking obtained from the MLE estimate $\hat{w}$. Let $\delta > 0$. Let*

$$m_1 = \frac{3\beta^2 \log(2d/\delta)d + 4\sqrt{d}\beta \log(2d^2/\delta)}{6},$$

$$m_2 = \frac{8\log(4d/\delta)(6\eta + \lambda\zeta)}{3\lambda^2},$$

and

$$m_3 = \frac{64M^2(1+\exp(b^*))^4(3\beta^2 \log(4d/\delta)d + 4\sqrt{d}\beta \log(4d/\delta))}{6\alpha_k^2 \exp(b^*)^2\lambda^2}.$$

If $m \geq \{m_1, m_2, m_3\}$, then with probability $1 - \frac{2}{d}$, $K(\gamma^*, \hat{\gamma}) \leq k - 1$, where $K(\gamma^*, \hat{\gamma}) = |\{(i,j) \in P : (\gamma^*(i) - \gamma^*(j))(\hat{\gamma}(i) - \hat{\gamma}(j)) < 0\}|$ is the Kendall tau distance between two *rankings.*

Proof. By Theorem 2.3.2, with probability $1-\delta$, we have

$$\|w^* - \hat{w}\|_2 \leq \frac{4(1+\exp(b^*))^2}{\exp(b^*)\lambda}\sqrt{\frac{3\beta^2 \log(4d/\delta)d + 4\sqrt{d}\beta\log(4d/\delta)}{6m}} \tag{2.177}$$

$$\leq \frac{\alpha_k}{2M} \tag{2.178}$$

by definition of m.

The estimated full feature utility for item i is no further than $\frac{\alpha_k}{2}$ to the true utility of item i:

$$|\langle w^* - \hat{w}, U_i\rangle| \leq \|w^* - \hat{w}\|_2 \|U_i\|_2 \text{ by Cauchy–Schwarz} \tag{2.179}$$

$$\leq \frac{\alpha_k \|U_i\|_2}{2M} \tag{2.180}$$

$$\leq \frac{\alpha_k}{2}. \tag{2.181}$$

Therefore for any $i \in [n]$,

$$\langle w^*, U_i\rangle - \frac{\alpha_k}{2} \leq \langle \hat{w}, U_i\rangle \leq \langle w^*, U_i\rangle + \frac{\alpha_k}{2}. \tag{2.182}$$

Let $P_{\alpha_k} := \{(i,j) \in P : |\langle w^*, U_i - U_j\rangle| \geq \alpha_k\}$ and let $(i,j) \in P_{\alpha_k}$. WLOG, suppose $\langle w^*, U_i\rangle - \langle w^*, U_j\rangle \leq 0$, i.e. $\gamma^*(i) - \gamma^*(j) \leq 0$, which means item j is ranked higher than item i in the true ranking given by γ. We want to show $\langle \hat{w}, U_i\rangle - \langle \hat{w}, U_j\rangle \leq 0$, i.e. $\hat{\gamma}(i) - \hat{\gamma}(j) \leq 0$, meaning that item j is ranked higher than item i in the estimated ranking given by $\hat{\gamma}$.

By applying Equation (2.182) and using the fact $\langle w^*, U_i\rangle - \langle w^*, U_j\rangle \leq 0$, we have

$$\langle \hat{w}, U_i\rangle \leq \langle w^*, U_i\rangle + \frac{\alpha_k}{2} \text{ by Equation (2.182)} \tag{2.183}$$

$$= \langle w^*, U_i\rangle - \langle w^*, U_j\rangle + \langle w^*, U_j\rangle + \frac{\alpha_k}{2} \tag{2.184}$$

$$\leq -\alpha_k + \langle w^*, U_j\rangle + \frac{\alpha_k}{2} \text{ since } (i,j) \in P_{\alpha_k} \text{ and since } \langle w^*, U_i\rangle - \langle w^*, U_j\rangle \leq 0 \tag{2.185}$$

$$\leq \langle w^*, U_j\rangle - \frac{\alpha_k}{2} \tag{2.186}$$

$$\leq \langle \hat{w}, U_j \rangle \text{ by Equation (2.182).} \tag{2.187}$$

Hence, $\langle \hat{w}, U_i \rangle - \langle \hat{w}, U_j \rangle \leq 0$ for every $i, j \in P_k$, meaning that for any $(i, j) \in P_k$, γ^* and $\hat{\gamma}$ agree on the relative ordering of item i and j. Furthermore, $|P_k| = \binom{n}{2} - (k-1)$. Therefore, $K(\gamma^*, \hat{\gamma}) \leq \binom{n}{2} - |P_k| = k - 1$.

Now we explain how to get from these results to those in the main text of this chapter with the order terms. The value of C_1 and C_2 are given at the end of the proof of Theorem 2.3.2. It is easy to see that $C_5 = 64 * 4 * 2^4/6$.

□

2.6.8 Synthetic Experiments

Code is available at `https://github.com/Amandarg/salient_features`.

Plot of Parameters in Theorem 2.3.2

In this section, the goal is to empirically illustrate how the top-t selection function and intransitivities effect the parameters b^*, ζ, η, β, and λ from Theorem 2.3.2 and hence the number of samples required and the exact upper bound on the estimation error. Just as in the synthetic experiment section, we sample each coordinate of U from $N(0, \frac{1}{\sqrt{d}})$ and each coordinate of w^* is sampled from $N(0, \frac{4}{\sqrt{d}})$.

In the experiments, the ambient dimension $d = 10$ and the number of items $n = 100$. We repeat the following 10 times: sample U and w^*, and use this U and w^* while varying $t \in [d]$ to compute all of the parameters of interest and intransitivity rates. The x-axis of each plot is the average strong stochastic transitivity (SST) violation rate defined in Section 2.4.1 where the average is taken over the 10 experiments. From Figure 2.2, intransitives decrease as t increases, so the x-axis in Figures 2.5 and 2.6 could roughly, but not exactly, be replaced with t, where t is decreasing from 10 to 1. The y-axis on the plots depict the average value and the bars represent the standard error over the 10 experiments.

Figure 2.5 shows the parameters in Theorem 2.3.2. Larger λ means smaller sample complexity, whereas smaller b^*, ζ, β and η means smaller sample complexity.

Recall in the Supplement re-statement of Theorem 2.3.2, the number of samples m

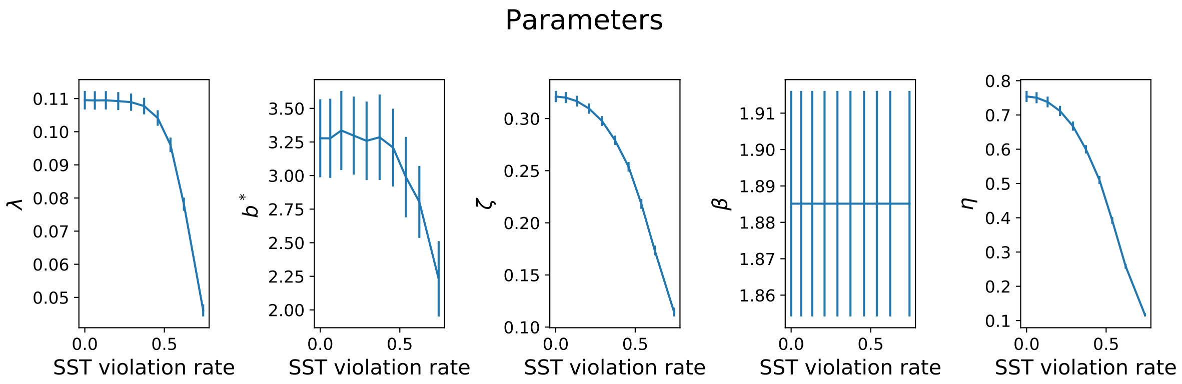


Figure 2.5: The parameters of Theorem 2.3.2 for the top-t selection function as a function of the average strong stochastic transitivity violation rate over the 10 experiments. The average over 10 experiments where a new U and w^* are drawn each time is depicted. The bars represent the standard error over the 10 experiments.

required in the theorem is

$$m \geq \max\left\{\frac{3\beta^2 \log(4d/\delta)d + 4\sqrt{d}\beta\log(4d/\delta)}{6}, \frac{8\log(2d/\delta)(6\eta+\lambda\zeta)}{3\lambda^2}\right\}.$$

Let $m_1 = \frac{3\beta^2 \log(4d/\delta)d+4\sqrt{d}\beta\log(4d/\delta)}{6}$ and $m_2 = \frac{8\log(2d/\delta)(6\eta+\lambda\zeta)}{3\lambda^2}$. Figure 2.6 shows m_1, m_2, and the bound from Theorem 2.3.2 with $\delta = \frac{1}{\delta} = \frac{1}{10}$ without the number of samples, i.e. the upper bound plot on the left does not include the number of samples in it. The plot shows

$$\frac{4(1+\exp(b^*))^2}{\exp(b^*)\lambda}\sqrt{\frac{3\beta^2\log(4d/\delta)d + 4\sqrt{d}\beta\log(4d/\delta)}{6}}$$

without the $\frac{1}{\sqrt{m}}$ term. Note that m_1 has constant average and standard error bars since with the dimension fixed, it is a function of β, which is constant in this case. Furthermore, this plot suggests that $m_1 << m_2$.

Additional Synthetic Experiments and Details

First we define the Kendall tau correlation. It is used in both Sections 2.4.1 and 2.4.2, and is defined as follows. Let $\gamma, \rho : [n] \to [n]$ be two rankings on n items where $\gamma(i)$ and $\rho(i)$ is the position of item i in the ranking. Let $A = \sum_{(i,j)\in P} \mathbb{1}_{\{(\sigma(i)-\sigma(j))(\rho(i)-\rho(j))>0\}}$,

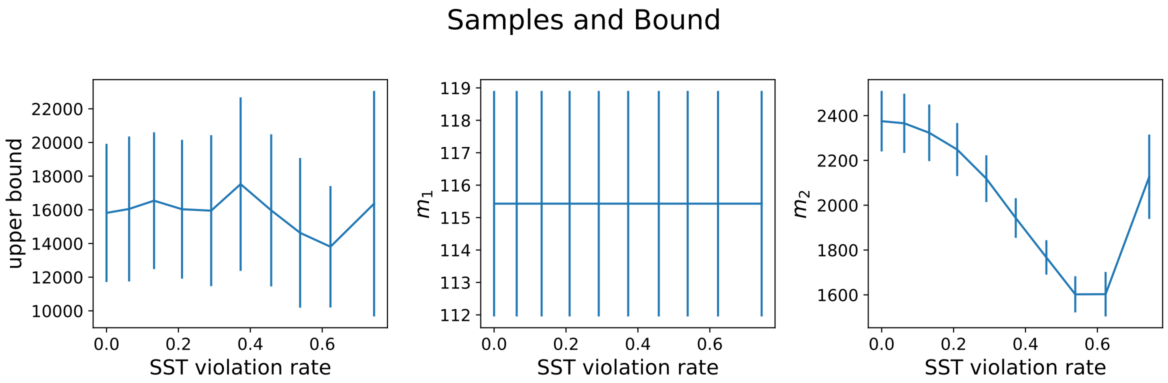


Figure 2.6: Number of samples m_1 and m_2 and upper bound on estimation error from Theorem 2.3.2 for the top-t selection function as a function of the average strong stochastic transitivity violation rate over the 10 experiments. The average over 10 experiments where a new U and w^* are drawn is depicted. The bars represent the standard error over the 10 experiments.

respectively $D = \sum_{(i,j)\in P} \mathbb{1}_{\{(\sigma(i)-\sigma(j))(\rho(i)-\rho(j))\leq 0\}}$, be the number of pairs of items that σ and ρ agree, respectively disagree, on the relative ordering. Then the Kendall tau correlation of ρ and γ is

$$KT(\gamma, \rho) := \frac{A - D}{\binom{n}{2}}. \tag{2.188}$$

Second, recall the set-up in Section 2.4: The ambient dimension $d = 10$, the number of items $n = 100$, and the top-1 selection function is used. The coordinates of U are drawn from $\mathcal{N}\left(0, \frac{1}{\sqrt{d}}\right)$,and the coordinates of w^* are drawn from $\mathcal{N}\left(0, \frac{4}{\sqrt{d}}\right)$. We sample m pairwise comparisons for $m \in \{2^i * (100) : i \in [11]\}$, fit the MLEs of the FBTL and salient preference model with the top-1 selection function, and repeat 10 times. Figure 2.7 shows the average pairwise prediction accuracy, which is defined as

$$\frac{|\{(i,j) \in P : (P_{ij} - .5)(\hat{P}_{ij} - 5) > 0\}|}{\binom{n}{2}}$$

where $\hat{P}_{ij}$ is the estimated pairwise probability that item i beats item j. The bars shows the standard error over the 10 experiments. The gap between the salient feature preference model MLE and the FBTL MLE is expected since the data is generated from the salient feature preference model.

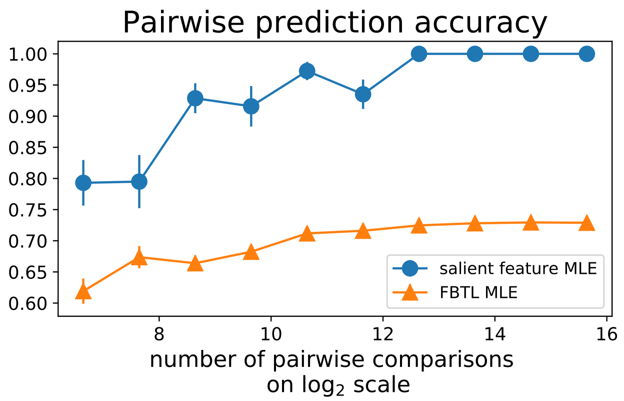


Figure 2.7: Pairwise prediction accuracy as a function of the number of samples, which are on the logarithmic scale, where the pairwise comparisons are sampled from the salient feature preference model with the top-1.

Third, see Figures 2.8 and 2.9 for plots investigating model misspecification. In particular, we use the same experimental set-up as in Section 2.4.1 except that in Figure 2.9 the salient feature preference model with the top-3 selection function is used to generate the preference data. We fit the MLE for the salient feature preference model for the top-t selection function for all $t \in [d]$ for both plots. The FBTL model is equivalent to when $t = 10$.

In Figure 2.8, we see that the model is very sensitive to the choice of t. As we would expect, $t = 2$ has the second smallest error when the number of samples exceed 2^{10}.

In Figure 2.9, we see that the model is still sensitive to the choice of t, but not as sensitive as in Figure 2.8. In this case, we can not only overestimate t, i.e. $t > 3$, but underestimate t, i.e. $t < 3$. We see that $t = 2$ and $t = 4$–the two values of t closest to the truth of $t = 3$–have roughly the same error. Interestingly, $t = 1$ has the worst performance.

2.6.9 Real Data Experiments

Code is available at `https://github.com/Amandarg/salient_features`.

Algorithm Implementation

In this section, we provide relevant details about how each algorithm is implemented.

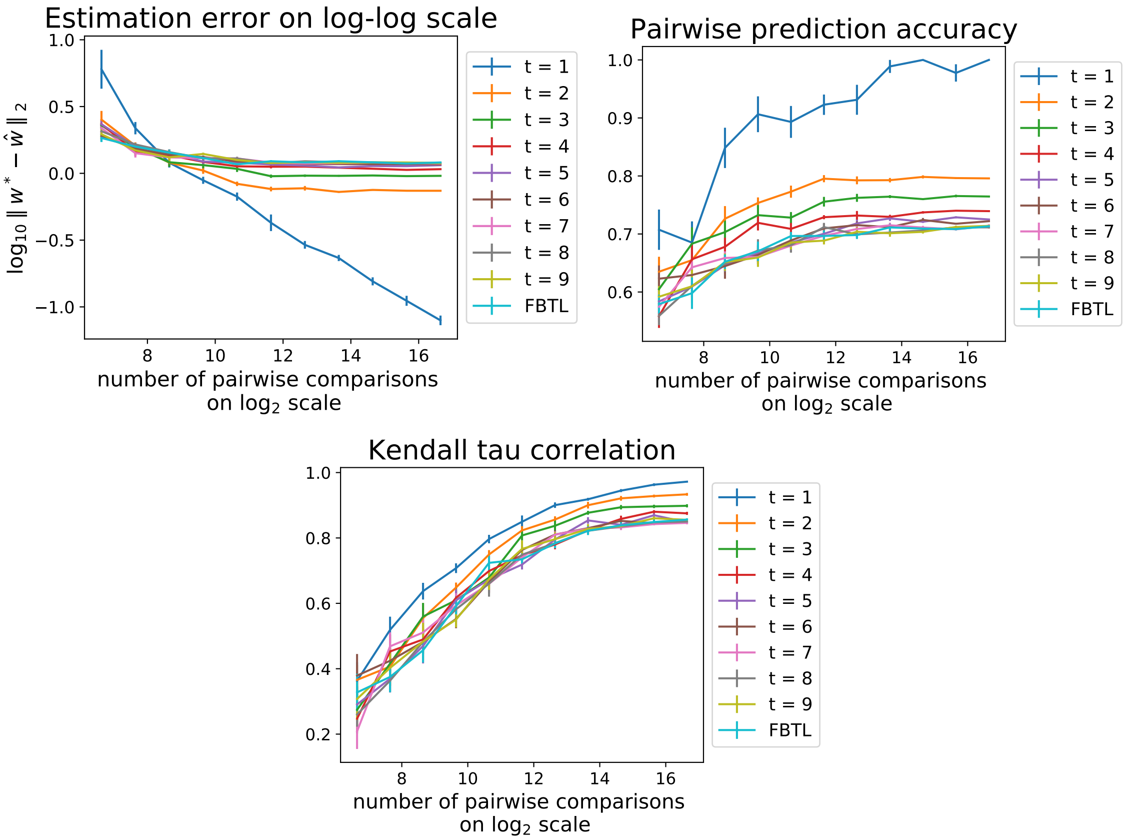


Figure 2.8: These plots investigate model misspecification. The true generative model for the pairwise preference data is the salient feature preference model with the top-1 selection function. The coordinates of U and w are sampled from a Gaussian as described in the main text. The MLEs for the salient feature preference model with the top-t selection function for $t \in [d]$ is shown.

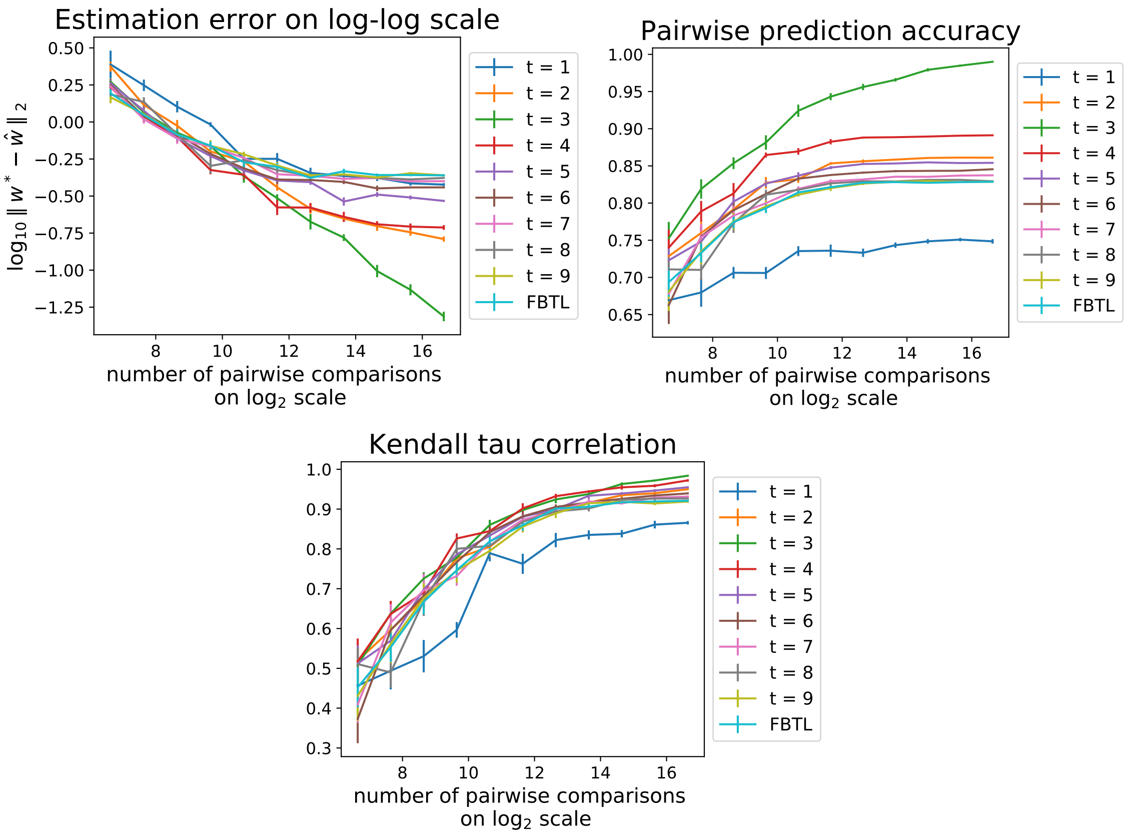


Figure 2.9: These plots investigate model misspecification. The true generative model for the pairwise preference data is the salient feature preference model with the top-3 selection function. The coordinates of U and w are sampled from a Gaussian as described in the main text. The MLEs for the salient feature preference model with the top-t selection function for $t \in [d]$ is shown.

- **RankNet:** We use the RankNet implementation found at `https://github.com/airalcorn2/RankNet`, which uses Keras. However, we use the Adam optimizer with default parameters except with a learning rate of 0.0001. We also add an ℓ_2 penalty to the weights.

- **Salient feature preference model and FBTL:** We use `sklearn`'s logistic regression solver. In particular, we set `tol` $= 1e - 10$ and `max_iter` $= 10000$. Furthermore, we do not fit an intercept. We use the default `liblinear` solver for real data experiments, and the `sag` solver for synthetic data experiments since we do not use regularization. All other parameters use the default values.

- **Ranking SVM:** We use `sklearn`'s `LinearSVC` solver with the same parameters as above. In particular, we do not fit an intercept.

The synthetic experiments were ran on a 2016 MacBook Pro with a 2.6 GhZ Quad-Core Intel Core i7 processor. The real data experiments were ran on the University of Michigan's Great Lakes Cluster [1].

District Compactness Experiments

We refer the reader to [KKK17] for the full details about the district compactness data, but provide relevant details here. We obtained the data by contacting the authors.

Pairwise comparison description There were three pairwise comparison studies. Due to data collection issues, only two of these pairwise comparison studies, called `shiny2pairs` and `shiny3pairs`, are available. In `shiny2pairs`, there are 3,576 pairwise for 298 people who each answered 12 pairwise comparisons. In `shiny3pairs`, there are 1,800 pairwise comparisons for 90 people who each answered 20 pairwise comparisons. There is no overlap in the districts used in `shiny2pairs` and `shiny3pairs`.

***k*-wise rankings for $k > 2$ description** There are 8 sets of k-wise ranking data. In many cases, the feature data for some districts are missing entirely, so in our own experiments, we throw out any district without feature data. Recall, we use the k-wise

[1] `https://arc-ts.umich.edu/greatlakes/`

ranking data for validation and testing, so we also remove any districts present in the training set.

- `Shiny1` contains rankings for 298 people on 20 districts, but the feature information for 10 districts are missing. The people are composed of undergraduate students, PhD students, law students, consultants, legislators involved in the redistricting process, and judges.
- `Shiny2` contains rankings on 20 districts for 103 people collected on Mturk. The feature information on 10 of the districts are missing however.
- `Mturk` contains another set of Mturk experiments collected on 100 districts and 13 people, which we use as our validation set. However, 34 of the districts also had pairwise comparison information collected about them, so we throw these out.
- `UG1-j1`, `UG1-j2`, `UG1-j3`, `UG1-j4`, and `UG1-j5` are 4 sets of 20-wise ranking data for 4 undergraduates at Harvard. The initial task was to rank 100 districts at once, but the resulting data set contains 5 sets of rankings on 20 districts. Out of the 100 districts used across the 5 sets of rankings, there are 38 districts with missing feature information.

See Figure 2.10 which depicts the average Kendall tau correlation between pairs of rankings in a k-wise ranking data set and the standard deviation. Recall the Kendall tau correlation, $KT(\cdot,\cdot)$, is defined in Equation (2.188). This plot shows roughly how much people agree with each other, where higher values mean more agreement. In particular, suppose there are N k-wise rankings given by $\sigma_1, \ldots, \sigma_N$. Then the average Kendall tau correlation for the N rankings is

$$\frac{1}{2\binom{N}{2}} \sum_{(i,j)\in[N]\times[N]} \mathrm{KT}(\sigma_i, \sigma_j)$$

and refer to this quantity as the average intercoder Kendall tau correlation. We see that people typically disagree on `shiny2` and `shiny1`, whereas people tend to agree more often on the rest of the k-wise data sets perhaps because there are fewer people.

The districts used in `shiny1` and `shiny2` are the same, and these districts also comprise one of the `UG1` data sets as well. However, the districts in `mturk` are disjoint from the rest

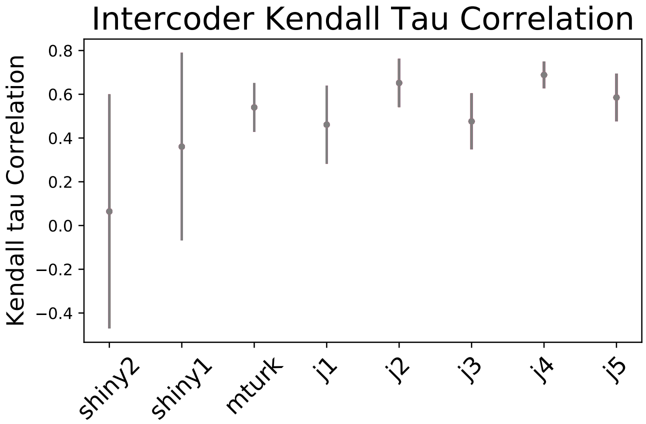


Figure 2.10: For each of the k-wise ranking data sets, the average agreement between people in terms of the Kendall tau correlation is shown.

of the k-wise ranking sets. In addition, `mturk` has relatively low intercoder variability. For these two reasons, we decided to use mturk as our validation set. We decided to keep `shiny1` and `shiny2` separate since the original authors did and also since they are comprised of different groups of people resulting in different behavior, e.g., `shiny1` has a higher average intercoder Kendall tau correlation than `shiny2`.

Data preprocessing We remove pairwise comparisons that were asked fewer than 5 times resulting in 5,150 pairwise comparisons over 94 unique pairs on 122 districts. There are 8 sets of k-wise comparison data that we use for validation and testing. We remove any districts in the k-wise ranking data that are present in the training data. We standardize the features of the districts by subtracting the mean and dividing by the standard deviation, where we use the mean and standard deviation from the training set. Standardizing the features is important for the salient feature preference model with the top-t selection function, so that each feature is roughly on the same scale. Otherwise, the top-t selection function might just choose the coordinates with the largest magnitude, and not the coordinates truly with the most variability.

Experiment details The hyperparameters for the salient feature preference model with the top-t selection function are t and the ℓ_2 regularization parameter μ. The hyperparameter for FBTL is the ℓ_2 regularization parameter μ. For Ranking SVM, the only hyperparameter is C which controls the penalty for violating the margin. We vary $t \in [d]$ where $d = 27$ since there are 27 features. We vary μ and C in

$$\{.00001, .0001, .001, .01, .1, 1, 10, 100, 1000, 10000, 100000, 1000000\}.$$

The hyperparameters for RankNet include the ℓ_2 regularization parameter μ and number of nodes in the hidden layer. We use one hidden layer. We varied the number of nodes in the single hidden unit in in $\{5 * i : i \in [19]\}$. We use a batch size of 250, and we use 800 epochs. Initially, we varied μ also in

$$\{.00001, .0001, .001, .01, .1, 1, 10, 100, 1000, 10000, 100000, 1000000\},$$

but as we will discuss in the next section we decided to vary μ in

$$\{.00001, .0001, .001, .01, .1, 1, 10\}.$$

Best performing hyperparameters Again, the validation set that was use is the `mturk` ranking data. Given $\hat{w}$, an estimate of w^*, we estimate the ranking by sorting each item's features with its inner product with $\hat{w}$. Then we pick the best hyperparameters by the largest average Kendall tau correlation of the estimated ranking with each individual ranking in `mturk`.

For FBTL, the best performing hyperparameter is $\mu = 100000$. The average Kendall tau correlation of the estimated ranking to each individual ranking in `mturk` is 0.38 with a standard deviation of 0.05. The pairwise comparison accuracy on the training set is 56%, which is defined in Section 2.6.8 of the Supplement. Although the regularization strength is large, the norm of the estimated judgement vector is .015. The largest coordinate of the judgement vector in absolute value is .005 and the smallest is .0001.

For the salient feature preference model with the top-t selection function the best performing hyperparameters are $t = 2$ and $\mu = .001$. The average Kendall tau correlation of the estimated ranking to each individual ranking in `mturk` is 0.54 with a standard deviation of 0.06. The pairwise comparison accuracy on the training set is 69%.

Figure 2.11 shows how often each of the 27 features are selected by the top-2 selection function over unique pairwise comparisons in the training data. Notice that `var xcoord` and `circle area` are never selected. The learned weights for those features in the FBTL model when all the features are used are 2 of the top 3 features with the smallest weights, so these features play a relatively insignificant role when all the features are used any way.

For RankNet, the best hyperparameters on the validation set are $\mu = .1$ and 75 nodes in the hidden layer. The average Kendall tau correlation of the estimated ranking to each individual ranking in `mturk` is 0.407 with a standard deviation of 0.05. The pairwise comparison accuracy on the training set is 59%. As we discussed in the previous section, we initially searched over larger values of μ. The best performing hyperparameters were $\mu = 10000$ and 40 nodes in the hidden layer. The pairwise comparison training accuracy was higher (69%) and the average Kendall tau correlation on the validation set was also higher (.48 with a standard deviation of .05). However, these hyperparameters were very unstable, i.e. training on the same data with the same hyperparameters sometimes gave a completely different model where the average Kendall tau correlation on the validation set or some of the test sets were sometimes negative.

For Ranking SVM, the best hyperparameter on the validation set is $C = 1000000$. The average Kendall tau correlation of the estimated ranking to each individual ranking in `mturk` is 0.38 with a standard deviation of 0.05. The pairwise comparison accuracy on the training set is 56%. Although C is large, the norm of the estimate of the judgement vector is .006, the largest entry in absolute value is .002, and the smallest is .0006, so it is finding a non-zero estimate for the judgement vector.

Zappos Experiments

We refer the reader to [YG14, YG17] for the full details about the `UT Zappos50k` data set but provide relevant details here. The data can be found at `http://vision.cs.utexas.edu/projects/finegrained/utzap50k/`.

Pairwise comparison data description The `UT Zappos50K` data set consists of pairwise comparisons on images of shoes and 960 extracted color and vision features for each shoe [YG14, YG17]. Given images of two different shoes and an attribute from {“open,” “pointy,” “sporty,” “comfort”}, respondents were asked to pick which shoe exhibits the attribute more. The data consists of both easier, coarse questions, i.e. based on comfort, pick between a slipper or high-heel, and also harder, fine grained questions

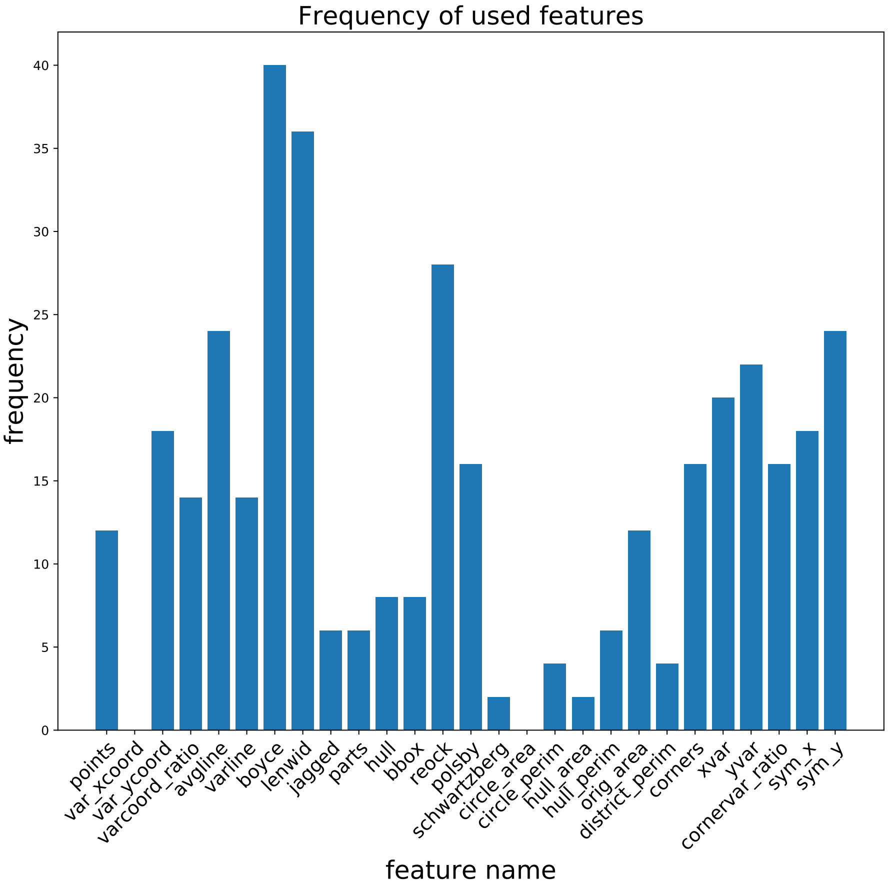


Figure 2.11: The frequency that the top-2 selection function chooses each feature over unique pairwise comparisons in the training data.

i.e. based on comfort, pick between two slippers. Each pairwise comparison is asked to 5 different people, and the confidence of each person's answer is also collected.

There are 2,863 unique pairwise comparisons involving 5,319 shoes for open, 2,700 unique pairwise comparisons involving 5,028 shoes for pointy, 2,766 unique pairwise comparisons involving 5,144 shoes for sporty, and 2,756 unique pairwise comparisons involving 5,129 shoes for comfort. For each attribute, 86% of unique pairwise comparisons involve an item that is in no other pairwise comparison regarding that attribute. Also, for each attribute, nearly 93% of items only appear in one pairwise comparison. In light of this, an algorithm like [CJ16b] will likely not work well since (1) this model requires learning a set of parameters for each item and (2) the model does not work for unseen items, i.e., we must ensure that items in testing also appear in training to evaluate the model.

Furthermore, for each of the attributes, there are no triplets of items (i, j, k) where pairwise comparison data has been collected on i vs. j, j vs. k, and k vs. i. Therefore, we cannot even test if there are intransitivities in this data.

Data pre-processing Respondents were given the option to declare a tie between two items. We do not train on any of these pairwise comparisons. To be clear, we use both the "coarse" and "fine-grained" comparisons during training. We standardize the features by subtracting the mean and dividing by the standard deviation, where we use the mean and standard deviation of the training set for each attribute since we train a model for each attribute.

Experiment details The hyperparameters for the salient feature preference model with the top-t selection function are t and the ℓ_2 regularization parameter μ. The hyperparameter for FBTL is the ℓ_2 regularization parameter μ. For Ranking SVM, the only hyperparameter is C which controls the penalty for violating the margin. We vary $t \in \{10 * i : i \in [99]\}$ since there are 990 features. We vary μ and C in $\{.000001, .00001, .0001, .001, .01, .1\}$. For RankNet, the hyperparameters are μ and the number of nodes in the hidden layer. We vary μ in $\{.05, .1, .15\}$ and the nodes in $\{50, 250, 500\}$. We choose these values of μ to try since on validation sets, it appeared that any value less than .05 was over fitting (train accuracy was in the 90%s but validation accuracy was in the 70%s) and values above .15 were not learning a good model (train

Table 2.3: Statistics about the best performing t for the salient feature preference model with the top-t selection function on the validation set over 10 train/validation/test splits for `UT Zappos50k`.

Attribute:	open	pointy	sporty	comfort
Min	440	310	110	40
Max	830	980	850	950
Average	663	614	550	563
Standard deviation	150	198	238	305

Table 2.4: Statistics about the best performing μ for the salient feature preference model on the validation set over 10 train/validation/test splits for `UT Zappos50k`.

Attribute:	open	pointy	sporty	comfort
Min	1000	100	1000	10
Max	10000	100000	10000	10000
Average	4600.0	12520.0	5500.0	5311.0
Standard deviation	4409.08	29389.65	4500.0	4700.46

accuracy was in the 60%s). We only search over these hyperparameters due to time constraints. We use ten 70% train, 15% validation, and 15% test split.

Best performing hyperparameters Because the pairwise comparisons are either "coarse" or "fine-grained," we pick the best hyperparameters based on the average of the pairwise comparison accuracy on the "coarse" questions and the "fine-grained" questions on the validation set. See Table 2.3 for statistics about the best performing t for the salient feature preference model with the top-t selection function on the validation set over 10 train/validation/test splits. See Tables 2.4, 2.5, 2.7 for statistics about the best performing μ for the salient feature preference model, FBTL model, and RankNet on the validation set over the 10 train/validation/test splits. See Table 2.6 for statistics about the best performing C for Ranking SVM on the validation set over the over the 10 train/validation/test splits. See Table 2.8 for the best performing number of nodes in the hidden layer on the validation set over the 10 splits. We also report the average pairwise accuracy, which has been defined in the main text, on the validation set for all algorithms in Table 2.9.

Table 2.5: Statistics about the best performing μ for FBTL on the validation set over 10 train/validation/test splits for `UT Zappos50k`.

Attribute:	open	pointy	sporty	comfort
Min	1000	100	1000	10
Max	100000	100000	100000	100000
Average	15400	12520	17200	24211
Standard deviation	28517	29389	27827	38131

Table 2.6: Statistics about the best performing C for Ranking SVM on the validation set over 10 train/validation/test splits for `UT Zappos50k`.

Attribute:	open	pointy	sporty	comfort
Min	10000	1000	10000	100
Max	100000	1000000	1000000	1000000
Average	70000	124300	163000	144010
Standard deviation	42426	294261	281888	288619

Table 2.7: Statistics about the best performing μ for RankNet on the validation set over 10 train/validation/test splits for `UT Zappos50k`.

Attribute:	open	pointy	sporty	comfort
Min	.05	.05	.05	.05
Max	.15	.1	.15	.15
Average	.075	.055	.085	.105
Standard deviation	.033	.015	.039	.041

Table 2.8: Statistics about the best performing number of nodes in the hidden layer for RankNet on the validation set over 10 train/validation/test splits for `UT Zappos50k`.

Attribute:	open	pointy	sporty	comfort
Min	50	50	50	250
Max	500	500	250	500
Average	335	205	190	350
Standard deviation	178.95	201.84	91.65	122.47

Table 2.9: Average pairwise prediction accuracy over 10 train/validation/test splits on the validation sets by attribute for `UT Zappos50k`. *C* stands for coarse and *F* stands for fine grained. *O* stands for open, *P* stand for pointy, *S* stands for sporty, and *Co* stands for comfort. The number in parenthesis is the standard deviation.

Model:	*O-C*	*P-C*	*S-C*	*Co-C*	*O-F*	*P-F*	*S-F*	*Co-F*
Salient features	.75(.01)	.8(.01)	.79(.02)	.77(.03)	.64(.03)	.6(.03)	.62(.03)	.66(.03)
FBTL	.75(.02)	.8(.01)	.79(.01)	.77(.02)	.63(.03)	.59(.03)	.6(.02)	.62(.03)
Ranking SVM	.75(.02)	.8(.02)	.8(.01)	.77(.02)	.62(.04)	.59(.03)	.6(.02)	.62(.04)
RankNet	.75(.02)	.78(.03)	.78(.01)	.76(.02)	.67(.03)	.61(.04)	.61(.02)	.64(.03)

Chapter 3

The Landscape of Non-Convex Quadratic Feasibility

The work in this chapter is joint with Lalit Jain and Laura Balzano, and parts of this work is published as *The Landscape of Non-Convex Quadratic Feasibility* at ICASSP 2018.

3.1 Introduction

In this chapter, we consider quadratic feasibility problems and present theory and experimental results utilizing first order methods for recovering a feasible point. We are motivated by a natural set of quadratic feasibility problems, namely ordinal embedding and collaborative ranking, that arise when using ordinal comparisons to find a Euclidean embedding for a set of items. These embeddings are useful for downstream machine learning applications such as rank aggregation, visualization, or recommender systems. We present both the ordinal embedding problem and collaborative ranking problem in full detail at the end of this section. Importantly, we will show both these embedding problems can be cast as the following homogeneous quadratic feasibility problem:

$$\begin{aligned} \text{find} \quad & x \in \mathbb{R}^n \\ \text{subject to} \quad & x^T P_i x > 0, \quad i = 1, \ldots, m \ , \end{aligned} \tag{3.1}$$

where $P_i \in \mathbb{R}^{n\times n}$ is a trace 0, symmetric matrix corresponding to the i-th constraint. Quadratic feasibility is a special case of quadratically constrained quadratic programming, which has been extensively studied. For instance, see the excellent survey [PB17]. In general, quadratic feasibility with indefinite P_i matrices is NP-hard.

We propose to solve Equation (3.1) by solving the following optimization problem that penalizes a candidate point when it does not satisfy a quadratic constraint:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \sum_{i\in[m]} \max\{0, 1 - x^T P_i x\}. \tag{3.2}$$

Similar to support vector machines, the hinge loss in Equation (3.2) captures a margin, which quantifies the amount a constraint is violated. Furthermore, the 1 in the objective of Equation (3.2) prevents first order methods from converging to the infeasible point $\hat{x} = \mathbf{0}$ and can be replaced with any positive constant. Since the constraint matrices are trace 0, they are indefinite, and thus Equation (3.2) is non-convex.

Assuming Equation (3.1) is feasible, there is a correspondence between feasible points and global minimizers of Equation (3.2). Indeed, any feasible point can be scaled to have an objective value of 0, the global minimum. Furthermore, any global minimizer corresponds to a feasible point. Thus our goal is to find a global minimum of the objective in Equation (3.2).

We propose to solve Equation (3.2) with a first order method, like stochastic gradient descent (SGD). First order methods are attractive in big data scenarios due to low memory and computation requirements. Although first order methods 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 [GJZ17] and Burer-Monteiro factorization for semidefinite programming [BVB16]. In these cases, a first order method can successfully avoid saddle points and so converges to global minima [LSJR16].

To the best of our knowledge, the local minimizers of the objective in Equation (3.2) have not been studied extensively making it unclear whether a first order method applied to Equation (3.2) finds a solution to Equation (3.1). We provide partial theoretical results towards understanding the optimization landscape of Equation (3.2) and compelling empirical results. We point out that [KS17] also recently proposed a similar method applying SGD to a smoothed version of Equation (3.2) that shows promising empirical results. However, they do not provide any theoretical results about the existence of non-global, local minima nor provide any assumptions regarding the success of recovering a feasible point of Equation (3.1) by applying a first order method to Equation (3.2). Furthermore, [TDN20] recently study a related problem dealing with finding a feasible point that satisfies a set of quadratic equalities with the ℓ_2-loss. They show that the optimization landscape is well-behaved, i.e. it has no spurious local minima and saddle points have strictly negative curvature, when the constraints are complex and drawn from a Gaussian distribution. The work in [BE06] identifies a sufficient condition for strong duality to hold when minimizing an indefinite quadratic function subject to two quadratic constraints. The work in [BTT96] shows that minimizing an indefinite quadratic over a sphere is equivalent to minimizing a convex function subject to linear constraints and minimizing an indefinite quadratic subject to finitely many convex quadratic constraints is equivalent to solving a minimax convex problem. In both cases, the solutions to the original non-convex problems are obtainable from their convex counterparts. These works are not applicable since we consider finding a feasible point to an arbitrary number of indefinite constraints.

Furthermore, the formulation of Equation (3.2) has been used in the specific case of ordinal embedding and collaborative ranking. For example, see [TVL14, JN11, AWC+07, PNZ+15b]. In both of these applications, extensive work has been done on bounding the sample complexity and determining the uniqueness of an embedding [JJN16, LN15b, PNZ+15b, OTX15b, AC+17], but little work has been done on theoretically understanding the proposed non-convex optimization problems and methods used to solve them.

Specifically, our work has three main contributions. First, 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 Equation (3.2); see Theorem 3.2.1. Second, in $\mathbb{R}^2$ under suitable assumptions, we show the objective of Equation (3.2) has no local minima; see Theorem 3.2.4. Third, we provide experiments showing the success of a first order method applied to Equation (3.2) for solving Equation (3.1).

3.1.1 Motivating Quadratic Feasibility Problems

Ordinal Embedding The first type of embedding model we consider is *ordinal embedding* (also known as non-metric multidimensional scaling [She62, Kru64]) and is based on Euclidean distance comparisons. Given ordinal constraints on distances of the form $\mathcal{T} = \{(i, j, k) : \text{item } i \text{ is closer to item } j \text{ than item } k\}$, the goal is to find n points, $\{x_1, x_2, \dots, x_n : x_i \in \mathbb{R}^d\}$, that satisfy Euclidean distance constraints. In particular, the constraint that corresponds to "item i is closer to item j than item k" is

$$||x_i - x_j||^2 < ||x_i - x_k||^2 \tag{3.3}$$

$$\iff \langle x_i, x_i\rangle - 2\langle x_i, x_j\rangle + \langle x_j, x_j\rangle < \langle x_i, x_i\rangle - 2\langle x_i, x_k\rangle + \langle x_k, x_k\rangle \tag{3.4}$$

$$\iff 0 < 2\langle x_i, x_j - x_k\rangle - \langle x_j, x_j\rangle + \langle x_k, x_k\rangle. \tag{3.5}$$

These are quadratic constraints, and finding a set of points that satisfies these constraints results in a quadratic feasibility problem. We now rewrite finding a set of points that satisfies the constraints in terms of finding a vector $X \in \mathbb{R}^{dn}$ that satisfies a set of constraints of the form $X^T P_{i,j,k} X > 0$ where $P_{i,j,k} \in \mathbb{R}^{nd\times nd}$. It turns out that $P_{i,j,k}$ is trace 0 and symmetric, which motivates our study of trace 0, symmetric matrices. Let

$$X := \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}. \tag{3.6}$$

Consider the constraint that "item i is closer to item k than item j." Let $P_{ijk} \in \mathbb{R}^{nd\times nd}$. Let $P_{ijk}^{(r,t)}$ denote the $(r, t)-$th entry of P_{ijk}. Define P_{ijk} as follows:

- $P_{ijk}^{(t,t)} = -1$ for $t \in \{(j-1)d + 1, \dots, jd\}$,
- $P_{ijk}^{(t,t)} = 1$ for $t \in \{(k-1)d + 1, \dots, kd\}$,
- $P_{ijk}^{(r,t)} = P_{ijk}^{(t,r)} = 1$ for $(t, r) \in \{((i-1)d + 1, (j-1)d + 1), ((i-1)d + 2, (j-1)d + 2), \dots, (id, jd)\}$,
- $P_{ijk}^{(r,t)} = P_{ijk}^{(t,r)} = -1$ for $(t, r) \in \{((i-1)d + 1, (k-1)d + 1), ((i-1)d + 2, (k-1)d + 2), \dots, (id, kd)\}$, and

- $P_{ijk}^{(r,t)} = 0$ for all other (r, t).

For example, if there are $n = 3$ points, $\{x, y, z\}$, in $d = 2$ dimensions and x is closer to z than y, the corresponding matrix is

$$P = \begin{pmatrix} 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \\ -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}$$

We check that this matrix does indeed capture the constraint. Let $X := (x_1, x_2, y_1, y_2, z_1, z_2)^T$, then

$$\begin{aligned} X^T P X &= X^T(-y_1 + z_1, -y_2 + z_2, -x_1 + y_1, -x_2 + y_2, x_1 - z_1, x_2 - z_2)^T \\ &= 2\langle x, z - y\rangle - \langle z, z\rangle + \langle y, y\rangle, \end{aligned}$$

which is exactly what we wanted to encode.

Therefore, we can formulate the ordinal embedding feasibility problem as follows:

$$\begin{aligned} \text{find} \quad & X \in \mathbb{R}^{nd} \\ \text{subject to} \quad & X^T P_{i,j,k} X > 0, \quad (i, j, k) \in \mathcal{T}\ . \end{aligned} \tag{3.7}$$

Collaborative Ranking The second type of embedding model is a low-rank approach to *collaborative ranking*. We assume there are n users and m items. Given preference constraints of the form

$$\mathcal{P} = \{(i, j, k) : \text{user } i \text{ prefers item } j \text{ to item } k\},$$

the goal of low-rank collaborative ranking is to find points $\{u_i \in \mathbb{R}^d\}_{i=1}^m$ corresponding to the items and points $\{w_i \in \mathbb{R}^d\}_{i=1}^n$ corresponding to the users that satisfy the following constraints:

$$\langle u_j, w_i\rangle > \langle u_k, w_i\rangle \text{ for } (i, j, k) \in \mathcal{P}.$$

For user j and item i, $\langle u_j, w_i \rangle$ represents the unknown score that user j assigns to item i. These scores define the preferences of user i where items with larger scores are preferred to items with smaller scores.

Finding $\{u_i \in \mathbb{R}^d\}_{i=1}^m$ and $\{w_i \in \mathbb{R}^d\}_{i=1}^n$ that satisfy the constraints in $\mathcal{P}$ is a quadratic feasibility problem once again. We will write this problem in terms of finding a vector $X \in \mathbb{R}^{(n+m)d}$ that satisfies $X^T L_{i,j,k} X > 0$ where $L_{i,j,k} \in \mathbb{R}^{(n+m)d \times (n+m)d}$. Again, it turns out $L_{i,j,k}$ is trace 0 and symmetric.

Let

$$X := \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}.$$

For $(i, j, k) \in \mathcal{P}$, let $L_{i,j,k} \in \mathbb{R}^{(n+m)d \times (n+m)d}$. Let $L_{i,j,k}^{(r,t)}$ denote the (r, t)th entry of $L_{i,j,k}$. Define $L_{i,j,k}$ as follows:

- $L_{i,j,k}^{(r,t)} = L_{i,j,k}^{(t,r)} = 1/2$ for $(r, t) \in \{((j-1)d+1, (m+i-1)d+1), ((j-1)d+2, (m+i-1)d+2) \ldots, ((j-1)d+d, (m+i-1)d+d)\}$
- $L_{i,j,k}^{(r,t)} = L_{i,j,k}^{(t,r)} = -1/2$ for $(r, t) \in \{((k-1)d+1, (m+i-1)d+1), ((k-1)d+2, (m+i-1)d+2) \ldots, ((k-1)d+d, (m+i-1)d+d)\}$
- $L_{i,j,k}^{(r,t)} = 0$ for all other r, t.

For instance, suppose that $n = 1$, $m = 2$, $d = 2$, i.e. there is only one person and two items and we seek an embedding in two dimensions. If this person says item 1 is better than item 2, then

$$L_{1,1,2} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & -1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1/2 \\ 1/2 & 0 & -1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & -1/2 & 0 & 0 \end{pmatrix}$$

We check that this matrix does indeed capture the constraint. Let

$$X := \begin{pmatrix} u_{11} \\ u_{12} \\ u_{21} \\ u_{22} \\ w_{11} \\ w_{12} \end{pmatrix}$$

Then

$$\begin{aligned} X^T L_{1,1,2} X &= X^T(1/2w_{1,1}, 1/2w_{1,1}, -1/2w_{1,1}, -1/2w_{1,2}, 1/2(u_{11} - u_{21}), u_{21} - u_{22})^T \\ &= 1/2(u_{11}w_{1,1} + u_{12}w_{1,1} - u_{21}w_{1,1} - u_{22}w_{1,2} + w_{11}(u_{11} - u_{21}) - w_{12}(u_{21} - u_{22})) \\ &= \langle u_1 - u_2, w \rangle, \end{aligned}$$

where $u_1 = (u_{11}, u_{12})^T$, $u_2 = (u_{21}, u_{22})^T$, and $w = (w_{11}, w_{12})^T$, which is exactly what we wanted to encode.

Therefore, the low-rank collaborative ranking problem can be re-written as

$$\begin{aligned} \text{find} \quad & X \in \mathbb{R}^{(n+m)d} \\ \text{subject to} \quad & X^T L_{i,j,k} X > 0, \quad (i,j,k) \in \mathcal{P}\ . \end{aligned} \tag{3.8}$$

3.2 Theory

3.2.1 Necessary Conditions for Minimizers

The following theorem gives necessary conditions for a point to be a non-global, local minimizer of optimization problem Equation (3.2). Notice that both the matrix constraints $P_{i,j,k}$ arising in ordinal embedding and $L_{i,j,k}$ arising in collaborative ranking discussed in the introduction are trace 0 and symmetric. Hence, we restrict our analysis to trace 0, symmetric matrices. We say a set of matrices $\{P_i \in \mathbb{R}^{n \times n}\}$ have a feasible point or share a feasible point if there is an $x \in \mathbb{R}^n$ so that $x^T P_i x > 0$ for all i.

Theorem 3.2.1. *Let $\{P_i \in \mathbb{R}^{n \times n}\}_{i=1}^m$ be a set of real, symmetric trace 0 matrices that share a feasible point. Assume x is not a global minimizer of Equation (3.2). If $x \in \mathbb{R}^n$ is a non-global, local minimizer of Equation (3.2), x must satisfy the following two equations:*

$$\sum_{i \in \{k : x^T P_k x < 1\}} x^T P_i x < 0, \tag{P1}$$

and

$$\sum_{i \in \{k : x^T P_k x < 1\}} x^T P_i x + \sum_{i \in \{k : x^T P_k x = 1\}} x^T P_i x \geq 0. \tag{P2}$$

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

Proof. First we set some notation. Let $L(x)$ be the objective of optimization problem Equation (3.2). 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, $L(x) = |I_x^{<1}| - x^T P_x^{<1} x = |I_x^{<1}| + |I_x^{=1}| - x^T P_x^{<1} x - x^T P_x^{=1} x$ 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. Since x' is not a global minimizer and a feasible point exists, $P_{x'}^{<1}$ exists and is non-zero. Hence, since $\text{trace}(P_{x'}^{<1}) = 0$ and $P_{x'}^{<1}$ is symmetric, $P_{x'}^{<1}$ is indefinite, i.e., it has positive and negative eigenvalues. Therefore, take u to be a unit eigenvector of $P_{x'}^{<1}$ with positive eigenvalue λ. Without loss of generality, we can assume $x'^T u \geq 0$. Otherwise, repeat the argument with $-u$.

Let $v_{\delta,\epsilon} = \epsilon x' + \delta u$ and $x := x' + v_{\delta,\epsilon}$. By choosing ϵ, δ sufficiently small, x is sufficiently close to x'. Again choosing ϵ, δ sufficiently small, if $j \in I_{x'}^{=1}$,

$$x^T P_j x = (1+\epsilon)^2 + 2(1+\epsilon)\delta u^T P_j x' + \delta^2 u^T P_j u > 1,$$

which implies $I_{x'}^{=1} \subseteq I_x^{>1}$, and similarly with small enough ϵ, δ, $I_{x'}^{>1} \subseteq I_x^{>1}$ and $I_x^{=1} = \varnothing$. Hence, $I_{x'}^{<1} = I_x^{<1}$. Since P_1 is not true and since $x'^T u \geq 0$,

$$\begin{align} L(x) &= |I_x^{<1}| - x^T P_x^{<1} x \tag{3.9}\\ &= |I_{x'}^{<1}| - x^T P_{x'}^{<1} x \tag{3.10}\\ &= |I_{x'}^{<1}| - (1+\epsilon)^2 x'^T P_{x'}^{<1} x' - \delta\lambda(2(1+\epsilon)x'^T u + \delta) \tag{3.11}\\ &< L(x'). \tag{3.12}\end{align}$$

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

$$\begin{align} L(x) &= |I_x^{<1}| - x^T P_x^{<1} x \tag{3.13}\\ &= |I_{x'}^{<1}| + |I_{x'}^{=1}| - (1-\epsilon)^2 x'^T \left(P_{x'}^{<1} + P_{x'}^{=1}\right) x' \tag{3.14}\\ &< |I_{x'}^{<1}| + |I_{x'}^{=1}| - \left(x'^T P_{x'}^{<1} x' + x'^T P_{x'}^{=1} x'\right) \tag{3.15}\\ &= L(x'), \tag{3.16}\end{align}$$

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

We note that for $x \in \mathbb{R}^n$, if $I_x^{=1} = \varnothing$, then $L(x)$ is a differentiable function in some sufficiently small neighborhood of x whose Hessian is indefinite since any trace 0 matrix is indefinite and the sum of trace 0 matrices is trace 0. From standard unconstrained optimization results all critical points of $L(x)$ in this neighborhood are saddle points. Therefore, a non-global, local minimizer x must have the property that $\{i : x^T P_i x = 1\} \neq \varnothing$, which gives an alternative proof to the second part of the theorem statement.

3.2.2 Two Dimensions

In this section, for trace 0 matrices in $\mathbb{R}^2$ sharing a feasible point, we show the objective of Equation (3.2) 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 by using the quadratic formula to find the feasible region of each constraint. There is a feasible point if the intersection of these regions is non-empty. However, this algorithm does not generalize to higher dimensions unlike solving Equation (3.2). We hope that our results in $\mathbb{R}^2$ generalize to higher dimensions.

Lemma 3.2.2. *Assume $A, B \in \mathbb{R}^{2\times 2}$ are linearly independent, trace 0 matrices. At any point x' on the curve $x^TBx = 1$, there is a tangent direction of $x^TBx = 1$ at x' which is a descent direction for $a - x^TAx$ 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 $a - x^TAx$ subject to $x^TBx = 1$, there exists $\lambda \in \mathbb{R}$ such that $Ax' = \lambda Bx'$ and $x'^TBx' = 1$. Hence, $x' \neq 0$. Since $A, B \in \mathbb{R}^{2\times 2}$, $\operatorname{tr}(A - \lambda B) = 0$, and they are independent, $A - \lambda B$ is invertible so no such x' or λ can exist. Therefore, we can always move along the curve $x^TBx = 1$ while decreasing $1 - x^TAx$. □

Lemma 3.2.3. *Assume $P_1, P_2, P_3 \in \mathbb{R}^{2\times 2}$ are trace 0, pairwise linearly independent matrices sharing a feasible point. Assume for some x', $x'^TP_1x' = x'^TP_2x' = 1$ and $x'^TP_3x' < 0$. Then at x', there is a tangent direction of $x^TP_1x = 1$ (respectively $x^TP_2x = 1$) which is an ascent direction of x^TP_2x (respectively x^TP_1x) and a descent direction for $\alpha - x^TP_3x$, where $\alpha \in \mathbb{R}$ is a constant.*

Proof. We have $P_3 = UDU^T$ where $D \in \mathbb{R}^{2\times 2}$ is diagonal and $U \in \mathbb{R}^{2\times 2}$ is orthogonal. Let $\hat{P}_i = U^TP_iU$ for $i \in \{1, 2, 3\}$, and $x'' = Ux'$. Then $x''^T\hat{P}_ix'' = x'^TP_ix'$, so x'' evaluates to the same value as x' on each matrix. Similarly, $x_F \in \mathbb{R}^2$ is a feasible point of P_i for $i \in \{1, 2, 3\}$ if and only if Ux_F is a feasible point of $\hat{P}_i$. It is easy to see that $\hat{P}_i$ are trace zero and pairwise linearly independent. Hence, the assumptions of the lemma still hold for $\hat{P}_i$ and in particular, $\hat{P}_3 = U^TP_3U = D$ is diagonal. Therefore, without loss of generality,

$$P_3 = \begin{pmatrix} e & 0 \\ 0 & -e \end{pmatrix},$$

$$P_1 = \begin{pmatrix} a & c \\ c & -a \end{pmatrix},$$

and

$$P_2 = \begin{pmatrix} b & d \\ d & -b \end{pmatrix}.$$

For $i \in [1, 2]$, let $f_i(x) = x^T P_i x$ and $f_3(x) = \alpha - x^T P_3 x$.

Let

$$T = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

Then a tangent vector to the curve $x^T P_i x = 1$ at x' is $UP_i x'$. A computation shows

$$\langle \nabla f_2(x'), TP_1 x' \rangle = -2(ad - bc)\|x'\|_2^2, \tag{3.17}$$

$$\langle \nabla f_1(x'), TP_2 x' \rangle = 2(ad - bc)\|x'\|_2^2. \tag{3.18}$$

Likewise

$$\langle \nabla f_3(x'), TP_1 x' \rangle = -2ce\|x'\|_2^2, \tag{3.19}$$

$$\langle \nabla f_3(x'), TP_2 x' \rangle = -2de\|x'\|_2^2. \tag{3.20}$$

Since P_1, P_2 are linearly independent with P_3, $c, d \neq 0$. By Lemma 3.2.5 (whose proof is delayed until the end of this subsection), $cd > 0$. Without loss of generality, assume $c, d > 0$ and $e > 0$. Since $\langle \nabla f_2(x'), TP_1 x' \rangle$ and $\langle \nabla f_1(x'), TP_2 x' \rangle$ have opposite signs, one is non-negative. WLOG say $\langle \nabla f_1(x'), TP_2 x' \rangle \geq 0$, so $TP_2 x'$ is an ascent direction of f_1 restricted to $x^T P_2 x = 1$.

Because $d, c, e > 0$, Equation (3.20) is negative, so $TP_2 x'$ 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 $TP_2 x'$, $x^T P_1 x$ increases by Equation (3.18) and $\alpha - x^T P_3 x$ decreases by Equation (3.20). If $c, d < 0$ or $e < 0$, then the same argument applies but with the tangent vector $-TP_i x'$. □

Theorem 3.2.4 (Arbitrary Number of Constraints)**.** *Let $\{P_i \in \mathbb{R}^{2\times 2}\}$ be real, symmetric, trace zero matrices that share a feasible point. Then every local minimizer of the objective* *of Equation (3.2) is a global minimizer.*

Proof. First we claim that no three of the curves $x^T P_i x = 1$ can intersect at a point. By the quadratic equation (for instance, see the proof of case 5 in Lemma 3.2.5), the solutions to $x^T P_i x = x^T P_j x$ are given by two lines of the form $x_2 = \alpha_1 x_1$ and $x_2 = \alpha_2 x_1$ such that $\alpha_1 = \frac{-1}{\alpha_2}$. Clearly, the solutions to $x^T P_i x = x^T P_j x = 1$ must be on these lines. Therefore, the solution set to $x^T P_k x = x^T P_i x = x^T P_j x = 1$ must be empty since any solution must satisfy both $x^T P_k x = x^T P_i x$ and $x^T P_i x = x^T P_j x$. There are two lines such that $x^T P_k x = x^T P_i x$ and two different lines such that $x^T P_i x = x^T P_j x$. The intersection of all these lines is the origin, which is not a point such that $x^T P_i x = 1$.

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 3.2.1. By contradiction, suppose $\hat{z} \in \mathbb{R}^2$ is a non-global, local minimizer of objective Equation (3.2). By Theorem 3.2.1, $\hat{z}^T P_{\hat{z}}^{<1} \hat{z} < 0$ 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 3.2.2 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 3.2.3. 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. □

We now return to the proof of Lemma 3.2.5 used in Lemma 3.2.3. First, we need two propositions.

Proposition 1. *Assume $P_1, P_2, P_3 \in \mathbb{R}^{2\times 2}$ are trace 0, pairwise linearly independent matrices such that*

$$P_1 = \begin{pmatrix} a & c \\ c & -a \end{pmatrix},$$

$$P_2 = \begin{pmatrix} b & d \\ d & -b \end{pmatrix},$$

and

$$P_3 = \begin{pmatrix} e & 0 \\ 0 & -e \end{pmatrix}.$$

where $b, c > 0$ and $a, d < 0$. If there is a feasible point (x_F, y_F) in the first quadrant (i.e., $x_F, y_F > 0$), then

$$\frac{c - \sqrt{c^2 + a^2}}{a} < \frac{d + \sqrt{d^2 + b^2}}{b}.$$

Proof. The main idea is to characterize the feasible regions of P_1 and P_2 separately, and then consider when these regions have a non-empty intersection in the first quadrant. Let $z := (x, y)^T$. By the quadratic formula, the solutions to $z^T P_1 z = 0$ are given by two lines:

$$y = \frac{(c \pm \sqrt{c^2 + a^2})x}{a}. \tag{3.21}$$

Since $c > 0$ and $a < 0$, the line $y = \frac{c - \sqrt{c^2+a^2}}{a}x$ has positive slope. Consider any point $z = (x, \frac{(c-\sqrt{c^2+a^2})x}{a} + \epsilon)^T$ such that $x, \epsilon > 0$, which characterizes any point in the first quadrant above the line $y = \frac{(c-\sqrt{c^2+a^2})x}{a}$. Then $z^T P_1 z > 0$:

$$\begin{aligned}
& z^T P_1 z && (3.22)\\
&= a\left(x^2 - \left(\frac{c - \sqrt{c^2 + a^2}}{a}x + \epsilon\right)^2\right) + 2cx\left(\frac{c - \sqrt{c^2 + a^2}}{a}x + \epsilon\right) && (3.23)\\
&= a\left(x^2 - \frac{(c - \sqrt{c^2 + a^2})^2}{a^2}x^2 - 2\frac{c - \sqrt{c^2 + a^2}}{a}x\epsilon - \epsilon^2\right) + 2cx\frac{c - \sqrt{c^2 + a^2}}{a}x + 2cx\epsilon && (3.24)\\
&= -\epsilon a\left(2\frac{c - \sqrt{c^2 + a^2}}{a}x + \epsilon\right) + 2cx\epsilon && (3.25)\\
&> 0 && (3.26)
\end{aligned}$$

where the second to last line uses the fact that for $z_1 = (x, \frac{c-\sqrt{c^2+a^2}}{a}x)^T$, $z_1^T P_1 z_1 = 0$ by construction and the last line is true since $x, c, \epsilon, \frac{c-\sqrt{c^2+a^2}}{a} > 0$ and $a < 0$.

Similarly, we can show that any $z = (x, \frac{(c-\sqrt{c^2+a^2})x}{a} - \epsilon)^T$ where $x > 0$ and $\epsilon > 0$ such that $\frac{(c-\sqrt{c^2+a^2})x}{a} - \epsilon > 0$, which characterizes any point in the first quadrant below the line $y = \frac{(c-\sqrt{2+a^2})x}{a}$, we have $z^T P_1 z < 0$.

A similar argument shows that the solutions to $z^T P_2 z = 0$ are the lines $y = \frac{d \pm \sqrt{d^2+b^2}}{b} x$, and the line $y = \frac{d+\sqrt{d^2+b^2}}{b} x$ has positive slope. Any point z in the first quadrant above this line has the property that $z^T P_2 z < 0$ and any point z in the first quadrant below this line has the property that $z^T P_2 z > 0$.

Therefore, P_1 and P_2 share a feasible point in the first quadrant if there is a point above the line $\frac{c-\sqrt{c^2+a^2}}{a} x$ which is also below the line $\frac{d+\sqrt{d^2+b^2}}{b} x$ meaning $\frac{c-\sqrt{c^2+a^2}}{a} < \frac{d+\sqrt{d^2+b^2}}{b}$. Furthermore, P_3 has a feasible point (x, y) in the first quadrant whenever $x > y$. Equivalently, any point in the first quadrant above the line $y = x$ is infeasible for P_3 and any point in the first quadrant below the line $y = x$ is feasible for P_3. Hence, it is easy to see that P_3 always shares a feasible point with P_2 since they both have feasible points arbitrarily close to the x-axis in the first quadrant. Lastly, P_1 shares a feasible point with P_3 if $\frac{c-\sqrt{c^2+a^2}}{a} < 1$ which is always true since $a < 0$ and $c > 0$:

$$\frac{c - \sqrt{c^2 + a^2}}{a} < 1 \tag{3.27}$$

$$\iff \quad c - a > \sqrt{a^2 + c^2} \tag{3.28}$$

$$\iff \quad c + |a| > \sqrt{a^2 + c^2}. \tag{3.29}$$

Therefore, P_1, P_2 and P_3 share a feasible point in the first quadrant if $\frac{c-\sqrt{c^2+a^2}}{a} < \frac{d+\sqrt{d^2+b^2}}{b}$. □

Proposition 2. *Let $a, d < 0$ and $b, c > 0$. Let*

$$f(x) = x + \sqrt{b^2c^2 + b^2a^2} - \sqrt{b^2c^2 + b^2a^2 + x^2 + 2bcx}.$$

Then $f(x) > 0$ if and only if $x > 0$.

Proof. The derivative of $f(x)$ is

$$f'(x) = 1 - \frac{2x + 2bc}{2\sqrt{b^2c^2 + b^2a^2 + x^2 + 2bcx}}.$$

We have

$$\begin{aligned}
f'(x) = 1 - \frac{2x+2bc}{2\sqrt{b^2c^2+b^2a^2+x^2+2bcx}} > 0 && (3.30)\\
\iff \sqrt{b^2c^2+b^2a^2+x^2+2bcx} > x+bc && (3.31)\\
\iff b^2c^2+b^2a^2+x^2+2bcx > (x+bc)^2 = x^2+2bcx+b^2c^2 && (3.32)\\
\iff b^2a^2 > 0. && (3.33)\\
&& (3.34)
\end{aligned}$$

Therefore, since $b^2a^2 > 0$, $f'(x) > 0$, meaning that f is always increasing. It is easy to see that $f(0) = 0$. Thus, since $f(0) = 0$ and f is increasing, $f(x) > 0$ if $x > 0$ and $f(x) < 0$ if $x < 0$. □

Lemma 3.2.5. *Assume $P_1, P_2, P_3 \in \mathbb{R}^{2\times 2}$ are trace 0, pairwise linearly independent matrices such that*

$$P_1 = \begin{pmatrix} a & c \\ c & -a \end{pmatrix},$$

$$P_2 = \begin{pmatrix} b & d \\ d & -b \end{pmatrix},$$

and

$$P_3 = \begin{pmatrix} e & 0 \\ 0 & -e \end{pmatrix}.$$

If P_1, P_2, and P_3 share a feasible point $z_F = (x_F, y_F)^T \in \mathbb{R}^2$ and there exists $z' = (x', y')^T \in \mathbb{R}^2$ such that $x'^T P_1 x' = x'^T P_2 x' = 1$ and $x'^T P_3 x' < 0$, then $cd > 0$.

Proof. We prove this lemma by showing z_F or z' cannot exist when $cd < 0$ by considering every case based on the signs of a, b, c, d, and $x_F y_F$. We eliminate having to consider the cases when $c = 0, d = 0, x_F y_F = 0, e \leq 0$ or $a = b = 0$. First, $c, d \neq 0$ since P_2 and P_3 are linearly independent with P_1. Second, since P_1 and P_2 are linearly independent, it cannot be the case that $a = b = 0$. Third, without loss of generality, we may assume $x_F y_F \neq 0$ since if there is a feasible point, there is always a feasible point (x'_F, y'_F) such that $x'_F \neq 0$ and $y'_F \neq 0$ since $z^T P_i z$ is a continuous function of z. Fourth, since there is a feasible point, $e \neq 0$. Without loss of generality, $e > 0$. To see this, we can multiply

a	b	c	d	$x_F y_F$	proof case number
$+$	$+$	$+$	$-$	$+$	1
$+$	$-$	$+$	$-$	$+$	2
$-$	$+$	$+$	$-$	$+$	5
$-$	$-$	$+$	$-$	$+$	2
$=0$	$+$	$+$	$-$	$+$	1
$=0$	$-$	$+$	$-$	$+$	2
$+$	$=0$	$+$	$-$	$+$	1
$-$	$=0$	$+$	$-$	$+$	2
$+$	$+$	$+$	$-$	$-$	3
$+$	$-$	$+$	$-$	$-$	6
$-$	$+$	$+$	$-$	$-$	4
$-$	$-$	$+$	$-$	$-$	4
$=0$	$+$	$+$	$-$	$-$	4
$=0$	$-$	$+$	$-$	$-$	4
$+$	$=0$	$+$	$-$	$-$	3
$-$	$=0$	$+$	$-$	$-$	7

Table 3.1: All the cases based on the signs of the variables we consider and the corresponding case number in the proof.

each matrix on the left and right by

$$T := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

and the assumptions of the lemma hold with $\tilde{z}' = Tx'$ and $\tilde{z}_F = Tz_F$. The entry in the first row and column of TP_3T is $-e > 0$.

See Table 3.1 for a summary of all the cases we will consider in the proof based on the signs of a, b, c, d, and $x_F y_F$ and the corresponding proof cases. We only consider the case when $c > 0$ and $d < 0$ in our proofs. However, an identical proof holds for when $d > 0$ and $c < 0$ by switching the roles of c and d.

By definition of z_F,

$$a(x_F^2 - y_F^2) + 2cx_F y_F > 0 \tag{3.35}$$

$$b(x_F^2 - y_F^2) + 2dx_F y_F > 0 \tag{3.36}$$

$$x_F^2 - y_F^2 > 0, \tag{3.37}$$

and by definition of z',

$$a(x'^2 - y'^2) + 2cx'y' = 1 \tag{3.38}$$
$$b(x'^2 - y'^2) + 2dx'y' = 1 \tag{3.39}$$
$$x'^2 - y'^2 < 0. \tag{3.40}$$

Case 1: Assume $a \geq 0$, $b \geq 0$, $c > 0$, $d < 0$, $x_F y_F > 0$. By Equation (3.39), $\frac{1-b(x'^2-y'^2)}{2d} = x'y'$. Since $b \geq 0$, $x'^2 - y'^2 < 0$ and $d < 0$, $x'y' < 0$. By Equation (3.38), $\frac{1-a(x'^2-y'^2)}{2x'y'} = c$. However, since $x'y' < 0$, $0 > \frac{1-a(x'^2-y'^2)}{2x'y'} = c$, a contradiction since $c > 0$.

Case 2: Assume $b \leq 0$, $c > 0$, $d < 0$, $x_F y_F > 0$ (this case is independent of the sign of a). By Equation (3.37) and since $b \leq 0, d < 0$ and $x_F y_F > 0$, $b(x_F^2 - y_F^2) + 2dx_F y_F < 0$, which contradicts Equation (3.36).

Case 3: Assume $a \geq 0$, $b \geq 0$, $c > 0$, $d < 0$, $x_F y_F < 0$. By Equation (3.38), $\frac{1-a(x'^2-y'^2)}{2c} = x'y'$. Since $a \geq 0, c > 0$ and $x'^2 - y'^2 < 0$, $x'y' > 0$. By Equation (3.39), $\frac{1-b(x'^2-y'^2)}{2d} = x'y'$. Since $b \geq 0$, $d < 0$, and $x'^2 - y'^2 < 0$, $x'y' < 0$, a contradiction.

Case 4: Assume $a \leq 0$, $c > 0$, $d < 0$, $x_F y_F < 0$ (this case is independent of the sign of b). By Equation (3.35), $\frac{-a(x_F^2-y_F^2)}{2c} < x_F y_F$. But $\frac{-a(x_F^2-y_F^2)}{2c} \geq 0$, so $0 < x_F y_F$, a contradiction since $x_F y_F < 0$.

Case 5: Assume $a < 0$, $b > 0$, $c > 0$, $d < 0$, $x_F y_F > 0$. We show Equation (3.40) cannot hold. If $x_F y_F > 0$, then either $x_F, y_F > 0$ or $x_F, y_F < 0$. In the latter case, multiplying x_F and y_F by -1 is a feasible point with positive coordinates. Therefore, by Proposition 1 and the assumptions of this lemma,

$$\frac{c - \sqrt{c^2 + a^2}}{a} < \frac{d + \sqrt{d^2 + b^2}}{b} \tag{3.41}$$
$$\Longrightarrow\ bc - b\sqrt{c^2 + a^2} > ad + a\sqrt{d^2 + b^2} \text{ since } a < 0, b > 0 \tag{3.42}$$
$$\Longrightarrow\ 0 > ad - bc + (a\sqrt{d^2 + b^2} + b\sqrt{c^2 + a^2}). \tag{3.43}$$

Let $ad - bc = \epsilon$. Then $d = \frac{\epsilon + bc}{a}$, so plugging this into the inequality above, we have

$$0 > ad - bc + a\sqrt{d^2 + b^2} + b\sqrt{c^2 + a^2} \tag{3.44}$$
$$= \epsilon + a\sqrt{\left(\frac{\epsilon + bc}{a}\right)^2 + b^2} + b\sqrt{c^2 + a^2} \tag{3.45}$$

$$= \epsilon - \sqrt{\epsilon^2 + 2\epsilon bc + b^2c^2 + b^2} + \sqrt{b^2c^2 + b^2a^2}. \tag{3.46}$$

Hence, $\epsilon - \sqrt{\epsilon^2 + 2\epsilon bc + b^2c^2 + b^2} + \sqrt{b^2c^2 + b^2a^2} < 0$, so $\epsilon = ad - bc < 0$ by Proposition 2. We will use $ad - bc < 0$ while computing the closed form for z' next.

Let $z'' = (x'', y'')^T$. Note that $a - b \neq 0$ since $a < 0$ and $b > 0$. Thus, by the quadratic formula, the solutions to $z''^T P_1 z'' = z''^T P_2 z''^T$ are given by the two lines $x'' = \alpha_+ y''$ and $x'' = \alpha_- y''$ where

$$\alpha_+ = \frac{((d-c) + \sqrt{(c-d)^2 + (a-b)^2})}{a-b} \tag{3.47}$$

and

$$\alpha_- = \frac{((d-c) - \sqrt{(c-d)^2 + (a-b)^2})}{a-b}. \tag{3.48}$$

Using this relationship between x'' and y'', we solve for when $z'^T P_1 z' = z'^T P_2 z' = 1$, i.e. Equation (3.38) and Equation (3.39). In particular, we need to solve for y' such that

$$a(\alpha_+ {y'}^2 - {y'}^2) + 2c\alpha_+ {y'}^2 = 1, \tag{3.49}$$

or

$$a(\alpha_- {y'}^2 - {y'}^2) + 2c\alpha_- {y'}^2 = 1, \tag{3.50}$$

If $a(\alpha_+^2 - 1) + 2c\alpha_+ > 0$, two solutions are

$$y' = \pm\sqrt{\frac{1}{a(\alpha_+^2 - 1) + 2c\alpha_+}}. \tag{3.51}$$

If $a(\alpha_-^2 - 1) + 2c\alpha_- > 0$, two solutions are

$$y' = \pm\sqrt{\frac{1}{a(\alpha_-^2 - 1) + 2c\alpha_-}}. \tag{3.52}$$

We now characterize when $a(\alpha_+^2 - 1) + 2c\alpha_+ > 0$ and $a(\alpha_-^2 - 1) + 2c\alpha_- > 0$. Let

$\beta = \sqrt{(a-b)^2 + (d-c)^2}$. Then,

$$
\begin{aligned}
& a(\alpha_-^2 - 1) + 2c\alpha_- > 0 && (3.53)\\
\iff\ & a\left(\frac{((d-c)-\beta)^2}{(a-b)^2} - 1\right) + 2c\left(\frac{(d-c)-\beta}{a-b}\right) > 0 && (3.54)\\
\iff\ & a\left(((d-c)-\beta)^2 - (a-b)^2\right) + 2c\left((a-b)(d-c) - (a-b)\beta\right) > 0 && (3.55)\\
\iff\ & a\left(2(d-c)^2 - 2(d-c)\beta\right) + 2c\left((a-b)(d-c) - (a-b)\beta\right) > 0 && (3.56)\\
\iff\ & 2a(d-c)^2 - 2\beta(c(a-b) + a(d-c)) + 2c(a-b)(d-c) > 0 && (3.57)\\
\iff\ & 2a(d-c)^2 - 2\beta(ad-bc) + 2c(a-b)(d-c) > 0 && (3.58)\\
\iff\ & (d-c)(ad-bc) - \beta(ad-bc) > 0 && (3.59)\\
\iff\ & (ad-bc)((d-c)-\beta) > 0 && (3.60)
\end{aligned}
$$

A similar calculation shows $a(\alpha_+^2 - 1) + 2c\alpha_+ > 0$ if and only if $(ad-bc)((d-c)+\beta) > 0$. From our earlier calculations, we know $ad - bc < 0$. Furthermore, it is easy to see that $(d-c)+\beta > 0$ and $(d-c)-\beta < 0$ since $d < 0$ and $c > 0$ and by definition of β. Therefore, $(ad-bc)((d-c)+\beta) < 0$ and $(ad-bc)((d-c)-\beta) > 0$, so the only solutions to $z'^T P_1 z' = z'^T P_2 z' = 1$ are $(\alpha_- y', y')$ where

$$y' = \pm\sqrt{\frac{1}{a(\alpha_-^2 - 1) + 2c\alpha_-}}. \qquad (3.61)$$

Now we show that at $z' = (\alpha_- y', y')$, $z'^T P_3 z' > 0$, contradicting the assumption of this lemma. Plugging x' and y' into Equation (3.40), we have

$$
\begin{aligned}
& \alpha_-^2 {y'}^2 - {y'}^2 < 0 && (3.62)\\
\iff\ & \alpha_-^2 < 1 && (3.63)\\
\iff\ & \frac{(d-c)^2 - 2(d-c)\sqrt{(c-d)^2 + (a-b)^2} + (c-d)^2 + (a-b)^2}{(a-b)^2} < 1 && (3.64)\\
\iff\ & (d-c)^2 - 2(d-c)\sqrt{(c-d)^2 + (a-b)^2} + (c-d)^2 + (a-b)^2 < (a-b)^2 && (3.65)\\
\iff\ & (d-c)^2 - 2(d-c)\sqrt{(c-d)^2 + (a-b)^2} + (c-d)^2 < 0 && (3.66)\\
\iff\ & (d-c) - \sqrt{(c-d)^2 + (a-b)^2} > 0 && (3.67)
\end{aligned}
$$

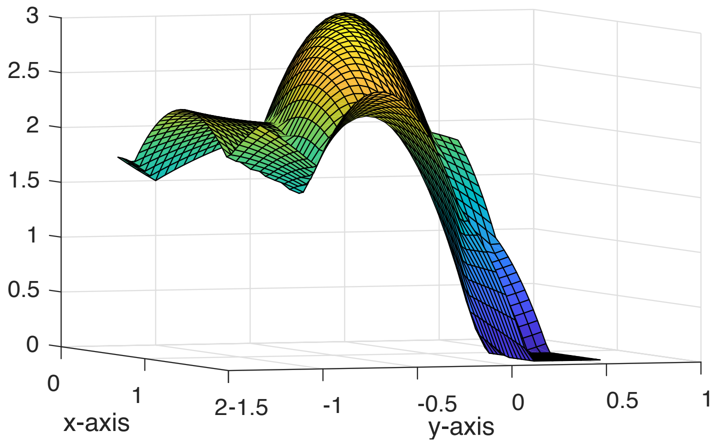


Figure 3.1: Existence of non-global, local minimum of objective of Equation (3.2) when trace 0 assumptions are not satisfied.

since $(d-c) < 0$. However, $(d-c) - \sqrt{(c-d)^2 + (a-b)^2} < 0$ since $d < 0$ and $c > 0$. Therefore, $z'^T P_3 z' > 0$, a contradiction.

Case 6: Assume $a > 0$, $b < 0$, $c > 0$, $d < 0$, $x_F y_F < 0$. A similar proof as case 5 holds by showing Equation (3.38), Equation (3.39), and Equation (3.40) cannot simultaneously hold.

Case 7: Assume $a < 0$, $b = 0$, $c > 0$, $d < 0$, and $x_F y_F < 0$. Then Equation (3.36) cannot hold since $b(x_F^2 - y_F^2) + 2dx_F y_F = 2dx_F y_F < 0$. $\square$

3.2.3 Importance of Assumptions

The trace 0 assumption of Theorem 3.2.4 is necessary. Otherwise, consider

$$Q_1 = \begin{pmatrix} 1 & 0 \\ 0 & -.5 \end{pmatrix},$$

$$Q_2 = \begin{pmatrix} .5 & 1 \\ 1 & 1 \end{pmatrix},$$

$$Q_3 = \begin{pmatrix} 0 & 1 \\ 1 & 5 \end{pmatrix},$$

which share a feasible point: $[1,1]^T$. Figure 3.1 shows that $x \approx [1.1, -.7]$ is a non-global, local minimizer of the objective of Equation (3.2) since the global minimum is 0. Therefore, proper initialization of first order methods and appropriate assumptions on the constraint matrices need to be more thoroughly studied to guarantee the success of solving Equation (3.2) with first order methods.

3.3 Experiments

In our experiments, we focus on validating SGD on Equation (3.2) 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 first order method like SGD. Therefore, we experiment with different step sizes and initializations at different scales. We remark that [KS17] contains an extensive set of experiments that validate using first order methods on a smoothed version of Equation (3.2) 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 Equation (3.2). 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 3.2 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.

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 3.3.

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 the success of SGD in both of the above experiments.

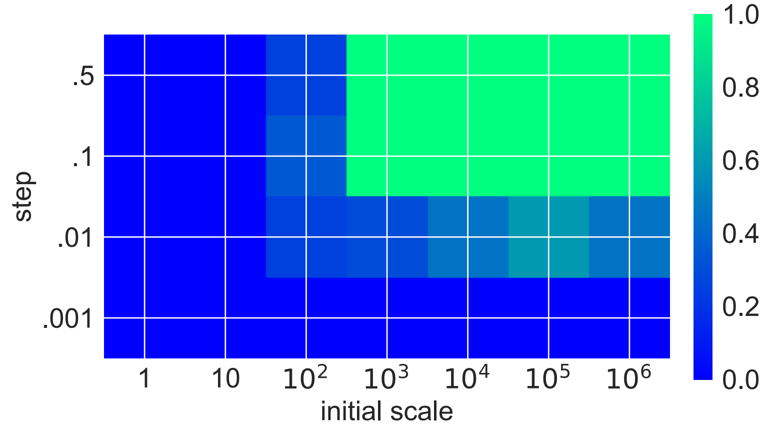


Figure 3.2: Success of recovering a feasible embedding.

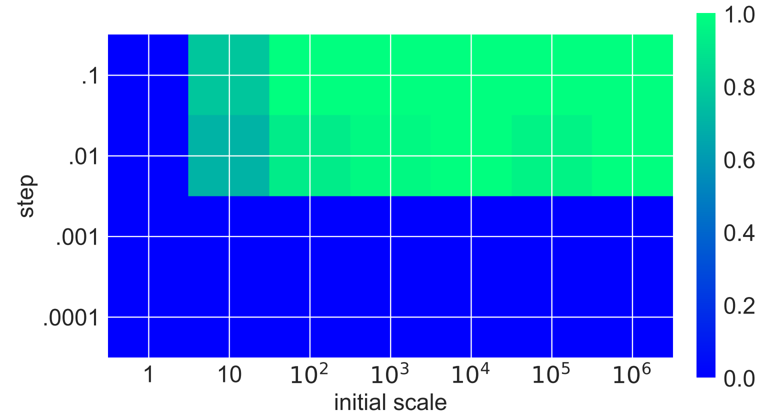


Figure 3.3: Success of general quadratic feasibility in $\mathbb{R}^{20}$.

3.4 Why [BY20] Is Not Applicable

Since the problem of quadratic feasibility is so well-studied, it is natural to ask whether existing results for non-convex problems apply to our setting. Some of the most state of the art results are given in [BY20]. In this section, we show the sufficient conditions identified in [BY20] for the Shor relaxation of the following problem related to quadratic feasibility are not satisfied:

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{minimize}} \quad & x^T x \\ \text{subject to} \quad & x^T A_i x \leq -1,\ i = 1, \ldots, m \\ & x^T x \leq M^2, \end{aligned} \tag{3.68}$$

where $M \in \mathbb{R}$, $A_i \in \mathbb{R}^{n\times n}$ are indefinite matrices, i.e., each have both positive and negative eigenvalues, whose non-zero eigenvalues have the same magnitude. We assume that there is a feasible point of Equation (3.68). This type of problem arises in the low-rank preference models that motivated the previous section.

The Shor relaxation of Equation (3.68), where $A \bullet B = \text{trace}(AB)$, is

$$\begin{aligned} \underset{X\in\mathbb{S}^{n\times n},\ x\in\mathbb{R}^{\ltimes}}{\text{minimize}} \quad & I \bullet X \\ \text{subject to} \quad & A_i \bullet X \leq -1,\ i = 1, \ldots, m \\ & I \bullet X \leq M^2, \\ & Y(x, X) \succeq 0, \end{aligned} \tag{3.69}$$

where $Y(x, X) = \begin{pmatrix} 1 & x^T \\ x & X \end{pmatrix}$.

Before that, there are three assumptions from the paper that we need to check our problem satisfies:

1. The feasible set of Equation (3.68) is non-empty. For our problem, we assume this is the case.

2. There exists $y \leq 0$ such that $\sum_{i=1}^m y_i A_i + y_{m+1} I \prec 0$. For our problem, set $y_i = 0$ for $i = 1, \ldots, m$ and $y_{m+1} = -1$, which satisfies this condition.

3. The interior of the feasible set of Equation (3.69) is nonempty. For our problem, since we know there is a feasible point x of Equation (3.68), there is a feasible point of Equation (3.69): $(xx^T, 0)$. At this feasible point, if any of the constraints are active, we just consider $(\alpha xx^T, 0)$ where $0 < \alpha < 1$. It is not difficult to see that a small neighborhood around this point is also a feasible point of Equation (3.69) since the first two constraints will be satisfied and the last constraint will also be satisfied since eigenvalues are continuous.

The sufficient condition identified in [BY20] for the Shor relaxation to be exact is stated for problems where A_i is diagonal, so we need to "diagonalize" the problem (as done in [BY20]). We consider the "lifted" problem with $n(m+1)$ variables (as opposed to n variables), where $A_i = Q_i D_i Q_i^T$ is the spectral decomposition of A_i. The diagonalized problem is as follows where in our problem D_i is a diagonal matrix whose non-zero terms all have the same magnitude:

$$\begin{aligned} \underset{x, y_i \in \mathbb{R}^n}{\text{minimize}} \quad & x^T x \\ \text{subject to} \quad & y_i^T D_i y_i \leq -1, \; i = 1, \ldots, m \\ & y_i = Q_i^T x, \; i = 1, \ldots, m \\ & x^T x + \sum_{i=1}^{m} y_i^T y_i \leq (m+1) M^2, \; i = 1, \ldots, m. \end{aligned} \tag{3.70}$$

Now we state the sufficient condition from [BY20], after which we will show that our problem in Equation (3.68) does not satisfy this sufficient condition.

Theorem 3.4.1 (sufficient condition with perturbation trick in [BY20])**.** *Consider the following linear system in $X, Y_i \in \mathbb{R}^{n \times n}$ for $i = 1, \ldots, m$ and $\epsilon > 0$:*

$$I \bullet X + \epsilon \sum_{i=1}^{m} I \bullet Y_i = -1 \tag{3.71}$$

$$D_i \bullet Y_i \leq 0, \; i = 1, \ldots, m \tag{3.72}$$

$$I \bullet X + \sum_{i=1}^{m} I \bullet Y_i \leq 0, \tag{3.73}$$

$$X, Y_i \text{ diagonal}, \; i = 1, \ldots, m \tag{3.74}$$

Furthermore, constrain all but one of the variables in X, Y_i to be non-negative. This gives $n(m+1)$ different linear systems. If there is a solution to each of these $n(m+1)$ systems, then the Shor relaxation of Equation (3.68) is exact. (This comes from Theorem 2 in [BY20] along with their perturbation trick. See equation (10) in the paper.)

Proposition 3. *When the non-zero values of D_i have the same magnitude for $i = 1, \ldots, n$, Equation (3.68) does not satisfy the sufficient condition in Theorem 3.4.1, i.e., there is not a solution to each of the $n(m+1)$ systems in Equation (3.71) - Equation (3.74).*

First, we need a lemma.

Lemma 3.4.2. *If $X', Y_1', \ldots, Y_m'$ is a solution to Equation (3.71) - Equation (3.74) with $[Y_1]_{jj}$ having any sign and all other variables constrained to be non-negative, then there exists another solution $X'', Y_1'', \ldots, Y_m''$ where $X'' = 0, Y_2'' = 0, \ldots, Y_m'' = 0$.*

Proof. Since $X', Y_1', \ldots, Y_m'$ is a solution such that $X', Y_2', \ldots, Y_m'$ have non-negative entries,

$$-1 - I \bullet X' - \epsilon \sum_{i=2}^{m} I \bullet Y_i' \leq -1 \neq 0.$$

Therefore, consider $Y_1'' = \frac{-1}{-1 - I \bullet X' - \epsilon \sum_{i=2}^{m} I \bullet Y_i'} Y_1'$. We claim $X'' = 0, Y_1'', Y_2'' = 0, \ldots, Y_m'' = 0$ is also a solution.

When $X'' = 0, Y_2'' = 0, \ldots, Y_m'' = 0$, clearly equations Equation (3.72) and Equation (3.74) for $i = 2, \ldots, m$ are satisfied. The rest of the equations Equation (3.71) - Equation (3.73) simplify to

$$\epsilon I \bullet Y_1 = -1 \tag{3.75}$$

$$D_1 \bullet Y_1 \leq 0 \tag{3.76}$$

$$I \bullet Y_1 \leq 0 \tag{3.77}$$

From Equation (3.71) since $X', Y_1', \ldots, Y_m'$ is a solution,

$$\epsilon I \bullet Y_1' = -1 - I \bullet X' - \epsilon \sum_{i=2}^{m} I \bullet Y_i',$$

so multiplying both sides by $\frac{-1}{-1-I\cdot X'-\epsilon\sum_{i=2}^m I\cdot Y_i'}$, we see

$$\epsilon I \bullet Y_1'' = -1.$$

Thus, Equation (3.75) is satisfied. Since Equation (3.75) is satisfied, clearly, Equation (3.77) is satisfied. Finally, $\frac{-1}{-1-I\cdot X'-\epsilon\sum_{i=2}^m I\cdot Y_i'} > 0$ since we have previously shown the denominator is negative. Thus, we see Equation (3.76) is also satisfied since $D_1 \bullet Y_1' \leq 0$. □

Proof of Proposition 3. Without loss of generality, suppose $[D_1]_{11}, [D_1]_{22}, \ldots, [D_1]_{kk} < 0$. We will show that there can be no solution to Equation (3.71) - Equation (3.74) when $[Y_1]_{11}$ is "free" but all other variables must be non-negative.

Seeking a solution, by Lemma 3.4.2, we may set $X, Y_2, \ldots, Y_m = 0$, and through a similar argument as Lemma 3.4.2, we may set $[Y_1]_{ii} = 0$ for $i = k+1, \ldots, n$.

At this point, we seek as solution to the following system in the variables $[Y_1]_{11}, \ldots, [Y_1]_{kk}$:

$$\begin{align} \epsilon\left([Y_1]_{11} + \cdots + [Y_1]_{kk}\right) &= -1 \tag{3.78}\\ [D_1]_{11}[Y_1]_{11} + \cdots + [D_1]_{11}[Y_1]_{kk} &\leq 0 \tag{3.79}\\ [Y_1]_{11} + \cdots + [Y_1]_{kk} &\leq 0 \tag{3.80}\\ [Y_1]_{22}, \ldots, [Y_1]_{kk} &\geq 0, \tag{3.81}\\ & \tag{3.82} \end{align}$$

where we have used the assumption that $[D_1]_{11} = \cdots = [D_1]_{kk}$.

However, because $[D_1]_{11} < 0$, we see that Equation (3.79) and Equation (3.80) cannot simultaneously hold since multiplying Equation (3.79) by $[D_1]_{11}$, we must satisfy

$$\begin{align} [Y_1]_{11} + \cdots + [Y_1]_{kk} &= \frac{-1}{\epsilon} < 0 \tag{3.83}\\ [Y_1]_{11} + \cdots + [Y_1]_{kk} &\geq 0 \tag{3.84}\\ & \tag{3.85} \end{align}$$

□

3.5 Conclusion

In this chapter, motivated by ordinal embedding and collaborative filtering, we studied the homogeneous non-convex quadratic feasibility problem. We posed this problem as an unconstrained non-convex optimization problem by penalizing a point for violating a quadratic constraint with the hinge loss, and we proposed to solve this problem with a first order method like stochastic gradient descent. Therefore, it is important to understand the optimization landscape, i.e. local minimizers. Assuming that the constraints are trace 0, symmetric matrices, we provided a necessary condition for a point to be a non-global, local minimizer and showed in the two dimensional case that all local minimizers are global minimizers.

Chapter 4

Training Individually Fair Machine Learning Models With Sensitive Subspace Robutness

The work in this chapter is joint with Mikhail Yurochkin and Yuekai Sun. Specifically, the theoretical results are Yuekai Sun's work and the experimental results were split with Mikhail Yurochkin. This work was published as *Training Individually Fair Machine Learning Models with Sensitive Subspace Robustness* at ICLR 2020.

4.1 Introduction

Machine learning (ML) models are gradually replacing humans in high-stakes decision making roles. For example, in Philadelphia, an ML model classifies probationers as high or low-risk [MS20]. In North Carolina, "analytics" is used to report suspicious activity and fraud by Medicaid patients and providers [MS20]. Although ML models appear to eliminate the biases of a human decision maker, they may perpetuate or even exacerbate biases in the training data [BS16]. Such biases are especially objectionable when it adversely affects underprivileged groups of users [BS16].

In response, the scientific community has proposed many mathematical definitions of algorithmic fairness and approaches to ensure ML models satisfy the definitions. Unfortunately, this abundance of definitions, many of which are incompatible [KMR17, Cho17], has hindered the adoption of this work by practitioners. There are two types of formal definitions of algorithmic fairness: group fairness and individual fairness. Most recent work on algorithmic fairness considers group fairness because it is more amenable to statistical analysis [RSZ17]. Despite their prevalence, group notions of algorithmic fairness suffer from certain shortcomings. One of the most troubling is there are many scenarios in which an algorithm satisfies group fairness, but its output is blatantly unfair from the point of view of individual users [DHP+12].

In this chapter, we consider individual fairness instead of group fairness. Intuitively, an individually fair ML model treats similar users similarly. Formally, an ML model is a map $h : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the input and output spaces. The leading notion of individual fairness is metric fairness [DHP+12]; it requires

$$d_y(h(x_1), h(x_2)) \leq L d_x(x_1, x_2) \text{ for all } x_1, x_2 \in \mathcal{X}, \tag{4.1}$$

where d_x and d_y are metrics on the input and output spaces and $L \geq 0$ is a Lipschitz constant. The fair metric d_x encodes our intuition of which samples should be treated similarly by the ML model. We emphasize that $d_x(x_1, x_2)$ being small does not imply x_1 and x_2 are similar in all respects. Even if $d_x(x_1, x_2)$ is small, x_1 and x_2 may differ in certain problematic ways, e.g. in their protected/sensitive attributes. This is why we refer to pairs of samples x_1 and x_2 such that $d_x(x_1, x_2)$ is small as *comparable* instead of similar.

Despite its benefits, individual fairness was dismissed as impractical because there is no widely accepted fair metric for many ML tasks. Fortunately, there is a line of recent work on learning the fair metric from data [Ilv20, WGL+19]. In this chapter, we consider two data-driven choices of the fair metric: one for problems in which the sensitive attribute is reliably observed, and another for problems in which the sensitive attribute is unobserved (see Appendix 4.6.2).

The rest of this chapter is organized as follows. In Section 4.2, we cast individual fairness as a form of robustness: robustness to certain sensitive perturbations to the inputs of an ML model. This allows us to leverage recent advances in adversarial ML to train individually fair ML models. More concretely, we develop an approach to audit ML models for violations of individual fairness that is similar to adversarial attacks [GSS15] and an approach to train ML models that passes such audits (akin to adversarial training [MMS+18]). We justify the approach theoretically (see Section 2.3) and empirically (see Section 4.4).

4.2 Fairness Through (Distributional) Robustness

To motivate our approach, imagine an auditor investigating an ML model for unfairness. The auditor collects a set of audit data and compares the output of the ML model on comparable samples in the audit data. For example, to investigate whether a resume screening system is fair, the auditor may collect a stack of resumes and change the names on the resumes of Caucasian applicants to names more common among the African-American population. If the system performs worse on the edited resumes, then the auditor may conclude the model treats African-American applicants unfairly. Such investigations are known as **correspondence studies**, and a prominent example is [BM04]'s celebrated investigation of racial discrimination in the labor market. In a correspondence study, the investigator looks for inputs that are comparable to the training examples (the edited resumes in the resume screening example) on which the ML model performs poorly. In the rest of this section, we formulate an optimization problem to find such inputs.

4.2.1 Fair Wasserstein Distances

Recall $\mathcal{X}$ and $\mathcal{Y}$ are the spaces of inputs and outputs. To keep things simple, we assume that the ML task at hand is a classification task, so $\mathcal{Y}$ is discrete. We also assume that we have a fair metric d_x of the form

$$d_x(x_1, x_2)^2 := \langle x_1 - x_2, \Sigma(x_1 - x_2)\rangle^{\frac{1}{2}},$$

where $\Sigma \in \mathbf{S}_+^{d\times d}$. For example, suppose we are given a set of K "sensitive" directions that we wish the metric to ignore; i.e., $d(x_1, x_2) \ll 1$ for any x_1 and x_2 such that $x_1 - x_2$ falls in the span of the sensitive directions. These directions may be provided by a domain expert or learned from data (see Section 4.4 and Appendix 4.6.2). In this case, we may choose Σ as the orthogonal complement projector of the span of the sensitive directions. We equip $\mathcal{X}$ with the fair metric and $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ with

$$d_z((x_1, y_1), (x_2, y_2)) := d_x(x_1, x_2) + \infty \cdot \mathbf{1}\{y_1 \neq y_2\}.$$

We consider d_z^2 as a transport cost function on $\mathcal{Z}$. This cost function encodes our intuition of which samples are comparable for the ML task at hand. We equip the space of probability distributions on $\mathcal{Z}$ with the fair Wasserstein distance

$$W(P, Q) = \inf_{\Pi \in \mathcal{C}(P,Q)} \int_{\mathcal{Z}\times\mathcal{Z}} c(z_1, z_2) d\Pi(z_1, z_2),$$

where $\mathcal{C}(P, Q)$ is the set of couplings between P and Q. The fair Wasserstein distance inherits our intuition of which samples are comparable through the cost function; i.e., the fair Wasserstein distance between two probability distributions is small if they are supported on comparable areas of the sample space.

4.2.2 Auditing Machine Learning Models for Algorithmic Bias

To investigate whether an ML model performs disparately on comparable samples, the auditor collects a set of audit data $\{(x_i, y_i)\}_{i=1}^n$ and solves the optimization problem

$$\max_{P:W(P,P_n)\leq\epsilon} \int_{\mathcal{Z}} \ell(z, h) dP(z), \tag{4.2}$$

where $\ell : \mathcal{Z} \times \mathcal{H} \to \mathbf{R}_+$ is a loss function, h is the ML model, P_n is the empirical distribution of the audit data, and $\epsilon > 0$ is a small tolerance parameter. We interpret ϵ as a moving budget that the auditor may expend to discover discrepancies in the performance of the ML model. This budget forces the auditor to avoid moving samples to incomparable areas of the sample space. We emphasize that Equation (4.2) detects *aggregate* violations of individual fairness. In other words, although the violations that the auditor's problem detects are individual in nature, the auditor's problem is only able to detect aggregate violations. We summarize the implicit notion of fairness in Equation (4.2) in a definition.

Definition 4.2.1 (distributionally robust fairness (DRF))**.** An ML model $h : \mathcal{X} \to \mathcal{Y}$ is (ϵ, δ)-distributionally robustly fair (DRF) WRT the fair metric d_x iff

$$\max_{P:W(P,P_n)\leq\epsilon} \int_{\mathcal{Z}} \ell(z, h) dP(z) \leq \delta. \tag{4.3}$$

Although Equation (4.2) is an infinite-dimensional optimization problem, it is possible to solve it exactly by appealing to duality. [BM19] showed that the dual of Equation (4.2) is

$$\begin{aligned}\sup\nolimits_{P:W(P,P_n)\le\epsilon}\mathbb{E}_P\big[\ell(Z,h)\big] = \inf\nolimits_{\lambda\ge 0}\{\lambda\epsilon+\mathbb{E}_{P_n}\big[\ell_\lambda^c(Z,h)\big]\},\\ \ell_\lambda^c((x_i,y_i),h) := \sup\nolimits_{x\in\mathcal{X}}\ell((x,y_i),h)-\lambda d_x^2(x,x_i).\end{aligned} \tag{4.4}$$

This is a univariate optimization problem, and it is amenable to stochastic optimization. We describe a stochastic approximation algorithm for Equation (4.4) in Algorithm 1. Inspecting the algorithm, we see that it is similar to the PGD algorithm for adversarial attack.

Algorithm 1 stochastic gradient method for Equation (4.4)

```
Require: starting point $\hat\lambda_1$, step sizes $\alpha_t > 0$
repeat
    draw mini-batch $(x_{t_1}, y_{t_1}), \ldots, (x_{t_B}, y_{t_B}) \sim P_n$
    $x_{t_b}^* \leftarrow \arg\max_{x\in\mathcal{X}} \ell((x, y_{t_b}), h) - \lambda d_x^2(x_{t_b}, x)$, $b \in [B]$
    $\hat\lambda_{t+1} \leftarrow \max\{0, \hat\lambda_t - \alpha_t(\epsilon - \frac{1}{B}\sum_{b=1}^B d_x^2(x_{t_b}, x_{t_b}^*))\}$
until converged
```

It is known that the optimal point of Equation (4.2) is the discrete measure

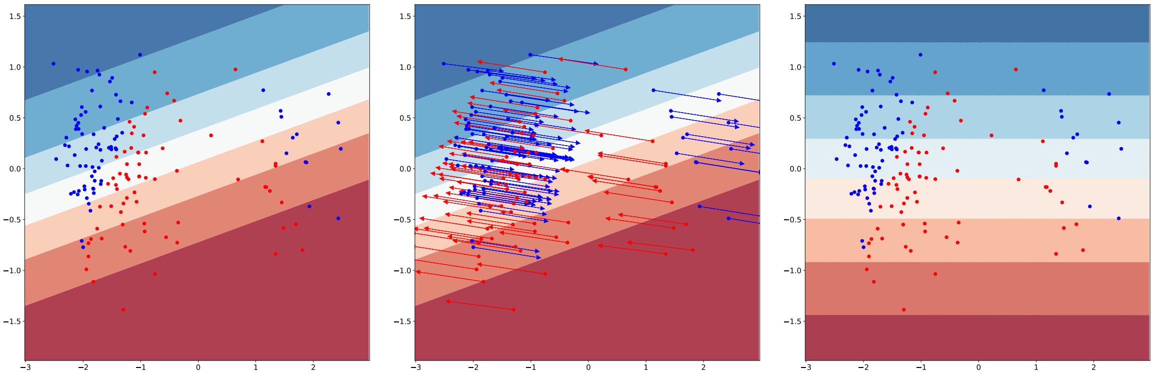
$$\frac{1}{n}\sum_{i=1}^n \delta_{(T_\lambda(x_i),y_i)},$$

where $T_\lambda : \mathcal{X}\to\mathcal{X}$ is the *unfair map*

$$T_\lambda(x_i) \leftarrow \arg\max\nolimits_{x\in\mathcal{X}}\ell((x,y_i),h)-\lambda d_x^2(x,x_i). \tag{4.5}$$

We call T_λ an unfair map because it reveals unfairness in the ML model by mapping samples in the audit data to comparable areas of the sample space that the system performs poorly on. We note that T_λ may map samples in the audit data to areas of the sample space that are not represented in the audit data, thereby revealing disparate treatment in the ML model not visible in the audit data alone. We emphasize that T_λ more than reveals disparate treatment in the ML model; it *localizes* the unfairness to certain areas of the sample space.

We present a simple example to illustrating fairness through robustness (a similar example appeared in [HSNL18]). Consider the binary classification dataset shown in



(a) unfair classifier (b) unfair map (c) classifier from SenSR

Figure 4.1: Figure (a) depicts a binary classification dataset in which the minority group shown on the right of the plot is underrepresented. This tilts the logistic regression decision boundary in favor of the majority group on the left. Figure (b) shows the unfair map of the logistic regression decision boundary. It maps samples in the minority group towards the majority group. Figure (c) shows an algorithmically fair classifier that treats the majority and minority groups identically.

Figure 4.1. There are two subgroups of observations in this dataset, and (sub)group membership is the protected attribute (e.g., the smaller group contains observations from a minority subgroup). In Figure 4.1a we see the decision heatmap of a vanilla logistic regression, which performs poorly on the blue minority subgroup. The two subgroups are separated in the horizontal direction, so the horizontal direction is the sensitive direction. Figure 4.1b shows that such classifier is unfair with respect to the corresponding fair metric, i.e. the *unfair map* Equation (4.5) leads to significant loss increase by transporting mass along the horizontal direction with very minor change of the vertical coordinate.

Comparison with metric fairness Before moving on to training individually fair ML models, we compare DRF with metric fairness Equation (4.1). Although we concentrate on the differences between the two definitions here, they are more similar than different: both formalize the intuition that the outputs of a fair ML model should perform similarly on comparable inputs. That said, there are two main differences between the two definitions. First, instead of requiring the output of the ML model to be similar on all inputs comparable to a training example, we require the output to be similar to the training label. Thus DRF not only enforces similarity of the output on comparable inputs, but also accuracy of the ML model on the training data. Second, DRF considers differences between datasets instead of samples by replacing the fair metric on inputs with the fair Wasserstein distance induced by the fair metric. The main benefits of this modifications are (i) it is possible to optimize Equation (4.2) efficiently, (ii) we can show this modified notion of individual fairness generalizes.

4.2.3 Fair Training with Sensitive Subspace Robustness

We cast the fair training problem as training supervised learning systems that are robust to sensitive perturbations. We propose solving the minimax problem

$$\inf_{h\in\mathcal{H}} \sup_{P:W(P,P_n)\leq\epsilon} \mathbb{E}_P\big[\ell(Z,h)\big] = \inf_{h\in\mathcal{H}} \inf_{\lambda\geq 0} \lambda\epsilon + \mathbb{E}_{P_n}\big[\ell_\lambda^c(Z,h)\big], \tag{4.6}$$

where ℓ_λ^c is defined in Equation (4.4). This is an instance of a distributionally robust optimization (DRO) problem, and it inherits some of the statistical properties of DRO. To see why Equation (4.6) encourages individual fairness, recall the loss function is a measure of the performance of the ML model. By assessing the performance of an ML model by its worse-case performance on hypothetical populations of users with perturbed sensitive attributes, minimizing Equation (4.6) ensures the system performs well on all such populations. In our toy example, minimizing Equation (4.6) implies learning a classifier that is insensitive to perturbations along the horizontal (i.e. sensitive) direction. In Figure 4.1c this is achieved by the algorithm we describe next.

To keep things simple, we assume the hypothesis class is parametrized by $\theta \in \Theta \subset \mathbf{R}^d$ and replace the minimization with respect to $\mathcal{H}$ by minimization with respect to θ. In light of the similarities between the DRO objective function and adversarial training, we borrow algorithms for adversarial training [MMS$^+$18] to solve Equation (4.6) (see Algorithm 2).

Algorithm 2 Sensitive Subspace Robustness (SenSR)

```
Require: starting point $\hat{\theta}_1$, step sizes $\alpha_t, \beta_t > 0$
repeat
    sample mini-batch $(x_1, y_1), \dots, (x_B, y_B) \sim P_n$
    $x^*_{t_b} \leftarrow \arg\max_{x \in \mathcal{X}} \ell((x, y_{t_b}), \theta) - \hat{\lambda}_t d_x^2(x_{t_b}, x)$, $b \in [B]$
    $\hat{\lambda}_{t+1} \leftarrow \max\{0, \hat{\lambda}_t - \alpha_t(\epsilon - \frac{1}{B}\sum_{b=1}^{B} d_x^2(x_{t_b}, x^*_{t_b}))\}$
    $\hat{\theta}_{t+1} \leftarrow \hat{\theta}_t - \frac{\beta_t}{B}\sum_{b=1}^{B} \partial_\theta \ell((x^*_{t_b}, y_{t_b}), \hat{\theta}_t)$
until converged
```

Related work Our approach to fair training is an instance of distributionally robust optimization (DRO). In DRO, the usual sample-average approximation of the expected cost function is replaced by $\widehat{L}_{\text{DRO}}(\theta) := \sup_{P \in \mathcal{U}} \mathbb{E}_P\big[\ell(Z, \theta)\big]$, where $\mathcal{U}$ is a (data dependent) uncertainty set of probability distributions. The uncertainty set may be defined by moment or support constraints [CSS07, DY10, GS10], f-divergences [BdDW$^+$12, LZ15, MMK$^+$16, ND16], and Wasserstein distances [SEK15, BKM19, EK15, LR18, SND18]. Most similar to our work is [HSNL18]: they show that DRO with a χ^2-neighborhood of the training data prevents representation disparity, i.e. minority groups tend to suffer higher losses because the training algorithm ignores them. One advantage of picking a Wasserstein uncertainty set is the set depends on the geometry of the sample space. This allows us to encode the correct notion of individual fairness for the ML task at hand in the Wasserstein distance.

Our approach to fair training is also similar to adversarial training [MMS$^+$18], which hardens ML models against adversarial attacks by minimizing adversarial losses of the form $\sup_{u \in \mathcal{U}} \ell(z + u, \theta)$, where $\mathcal{U}$ is a set of allowable perturbations [SZS$^+$14, GSS15, PMJ$^+$16, CW17, KGB17]. Typically, $\mathcal{U}$ is a scaled ℓ_p-norm ball: $\mathcal{U} = \{u : \|u\|_p \leq \epsilon\}$. Most similar to our work is [SND18]: they consider an uncertainty set that is a Wasserstein neighborhood of the training data.

There are a few papers that consider adversarial approaches to algorithmic fairness. [ZLM18] propose an adversarial learning method that enforces equalized odds in which the adversary learns to predict the protected attribute from the output of the classifier. [ES16] propose an adversarial method for learning classifiers that satisfy demographic parity. [MCPZ18] generalize their method to learn classifiers that satisfy other (group) notions of algorithmic fairness. [GPL$^+$19] propose to use adversarial logit pairing [KKG18] to achieve fairness in text classification using a pre-specified list of counterfactual tokens.

4.3 SenSR Trains Individually Fair Machine Learning Models

One of the main benefits of our approach is it provably trains individually fair ML models. Further, it is possible for the learner to certify that an ML model is individually fair *a posteriori*. As we shall see, both are consequences of uniform convergence results for the DR loss class. More concretely, we study how quickly the uniform convergence error

$$\delta_n := \sup_{\theta\in\Theta}\left\{\left|\sup_{P:W_*(P,P_*)\leq\epsilon}\mathbb{E}_P\big[\ell(Z,\theta)\big] - \sup_{P:W(P,P_n)\leq\epsilon}\mathbb{E}_P\big[\ell(Z,\theta)\big]\right|\right\}, \qquad (4.7)$$

where W_* is the Wasserstein distance on $\Delta(\mathcal{Z})$ with a transportation cost function c_* that is possibly different from c, vanishes, and P^* is the true distribution on $\mathcal{Z}$. We permit some discrepancy in the (transportation) cost function to study the effect of a data-driven choice of c. In the rest of this section, we regard c_* as the exact cost function and c as a cost function learned from human supervision. We start by stating our assumptions on the ML task:

(A1) the feature space $\mathcal{X}$ is bounded: $D := \max\{\mathrm{diam}(\mathcal{X}), \mathrm{diam}_*(\mathcal{X})\} < \infty$;

(A2) the functions in the loss class $\mathcal{L} = \{\ell(\cdot,\theta) : \theta\in\Theta\}$ are non-negative and bounded: $0 \leq \ell(z,\theta) \leq M$ for all $z\in\mathcal{Z}$ and $\theta\in\Theta$, and L-Lipschitz with respect to d_x:

$$\sup_{\theta\in\Theta}\{\sup_{(x_1,y),(x_2,y)\in\mathcal{Z}}|\ell((x_1,y),\theta) - \ell((x_2,y),\theta)|\} \leq Ld_x(x_1,x_2);$$

(A3) the discrepancy in the (transportation) cost function is uniformly bounded:

$$\sup_{(x_1,y),(x_2,y)\in\mathcal{Z}}|c((x_1,y),(x_2,y)) - c_*((x_1,y),(x_2,y))| \leq \delta_c D^2.$$

Assumptions A1 and A2 are standard (see [LR18, Assumption 1, 2, 3]) in the DRO literature. We emphasize that the constant L in Assumption A2 is **not** the constant

L in the definition of metric fairness; it may be much larger. Thus most models that satisfy the conditions of the loss class are not individually fair in a meaningful sense.

Assumption A3 deserves further comment. Under A1, A3 is mild. For example, if the exact fair metric is

$$d_x(x_1, x_2) = (x_1 - x_2)^T \Sigma_*(x_1 - x_2)^{\frac{1}{2}},$$

then the error in the transportation cost function is at most

$$\begin{aligned}
&|c((x_1, y), (x_2, y)) - c_*((x_1, y), (x_2, y))| \\
&\quad = |(x_1 - x_2)^T \Sigma(x_1 - x_2) - (x_1 - x_2)^T \Sigma_*(x_1 - x_2)| \\
&\quad \le D^2 \|\Sigma - \Sigma_*\|_2,
\end{aligned}$$

We see that the error in the transportation cost function vanishes in the large-sample limit as long as Σ is a consistent estimator of Σ_*.

We state the uniform convergence result in terms of the *entropy integral* of the loss class: $\mathfrak{C}(\mathcal{L}) = \int_0^\infty \sqrt{\log N_\infty(\mathcal{L}, r)}dr$, where $N_\infty(\mathcal{L}, r)$ as the r-covering number of the loss class in the uniform metric. The entropy integral is a measure of the complexity of the loss class.

Proposition 4.3.1 (uniform convergence)**.** *Under Assumptions A1–A3, Equation (4.7) satisfies*

$$\delta_n \le \frac{48\mathfrak{C}(\mathcal{L})}{\sqrt{n}} + \frac{48LD^2}{\sqrt{n\epsilon}} + \frac{L\delta_c D^2}{\sqrt{\epsilon}} + M(\frac{\log\frac{2}{t}}{2n})^{\frac{1}{2}} \tag{4.8}$$

with probability at least $1 - t$.

We note that Proposition 4.3.1 is similar to the generalization error bounds by [LR18]. The main novelty in Proposition 4.3.1 is allowing error in the transportation cost function. We see that the discrepancy in the transportation cost function may affect the rate at which the uniform convergence error vanishes: it affects the rate if δ_c is $\omega_P(\frac{1}{\sqrt{n}})$.

A consequence of uniform convergence is SenSR trains individually fair classifiers (if there are such classifiers in the hypothesis class). By individually fair ML model, we mean an ML model that has a small gap

$$\sup_{P:W_*(P,P_*)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] - \mathbb{E}_{P_*}\big[\ell(Z,\theta)\big], \tag{4.9}$$

The gap is the difference between the optimal value of the auditor's optimization problem

Equation (4.2) and the (non-robust) risk. A small gap implies the auditor cannot significantly increase the loss by moving samples from P_* to comparable samples.

Proposition 4.3.2. *Under the assumptions A1–A3, as long as there is $\bar{\theta} \in \Theta$ such that*

$$\sup_{P:W_*(P,P_*)\leq\epsilon} \mathbb{E}_P\big[\ell(Z,\bar{\theta})\big] \leq \delta^* \tag{4.10}$$

for some $\delta^* > 0$, $\widehat{\boldsymbol{\theta}} \in \arg\min_{\theta\in\Theta} \sup_{P:W(P,P_n)\leq\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big]$ satisfies

$$\sup_{P:W_*(P,P_*)\leq\epsilon} \mathbb{E}_P\big[\ell(Z,\widehat{\boldsymbol{\theta}})\big] - \mathbb{E}_{P_*}\big[\ell(Z,\widehat{\boldsymbol{\theta}})\big] \leq \delta^* + 2\delta_n,$$

where δ_n is the uniform convergence error Equation (4.7).

Proposition 4.3.2 guarantees Algorithm 2 trains an individually fair ML model. More precisely, if there are models in $\mathcal{H}$ that are (i) individually fair and (ii) achieve small test error, then Algorithm 2 trains such a model. It is possible to replace Equation (4.10) with other conditions, but a condition to its effect cannot be dispensed with entirely. If there are no individually fair models in $\mathcal{H}$, then it is not possible for Equation (4.6) to learn an individually fair model. If there are individually fair models in $\mathcal{H}$, but they all perform poorly, then the goal of learning an individually fair model is futile.

Another consequence of uniform convergence is Equation (4.9) is close to its empirical counterpart

$$\sup_{P:W(P,P_n)\leq\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] - \mathbb{E}_{P_n}\big[\ell(Z,\theta)\big]. \tag{4.11}$$

In other words, the gap *generalizes.* This implies Equation (4.11) is a *certificate of individual fairness*; i.e., it is possible for practitioners to check whether an ML model is individually fair by evaluating Equation (4.11).

Proposition 4.3.3. *Let*

$$R(\theta, P_n) = \sup_{P:W(P,P_n)\leq\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] - \mathbb{E}_{P_n}\big[\ell(Z,\theta)\big]$$

and

$$R(\theta, P_*) = \sup_{P:W_*(P,P_*)\leq\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] - \mathbb{E}_{P_*}\big[\ell(Z,\theta)\big]$$

Under the assumptions A1–A3, for any $\epsilon > 0$,

$$\sup_{\theta\in\Theta}\{R(\theta, P_n) - R(\theta, P_*)\} \leq 2\delta_n \textit{ with probability at least } 1 - t.$$

4.4 Computational Results

In this section, we present results from using SenSR to train individually fair ML models for two tasks: sentiment analysis and income prediction. We pick these two tasks to demonstrate the efficacy of SenSR on problems with structured (income prediction) and unstructured (sentiment analysis) inputs and in which the sensitive attribute (income prediction) is observed and unobserved (sentiment analysis). We refer to Appendix 4.6.3 and 4.6.4 for the implementation details.

4.4.1 Fair Sentiment Prediction with Word Embeddings

Table 4.1: Sentiment prediction experiments over 10 restarts

	Acc.,%	Race gap	Gend. gap	Cuis. gap
SenSR	94±1	0.30±.05	0.19±.03	**0.23**±.05
SenSR-E	93±1	**0.11**±.04	**0.04**±.03	1.11±.15
Baseline	**95**±1	7.01±.44	5.59±.37	4.10±.44
Project	94±1	1.00±.56	1.99±.58	1.70±.41
Sinha+	94±1	3.88±.26	1.42±.29	1.33±.18
Bolukb.+	94±1	6.85±.53	4.33±.46	3.44±.29

Problem formulation We study the problem of classifying the sentiment of words using positive (e.g. 'smart') and negative (e.g. 'anxiety') words compiled by [HL04]. We embed words using 300-dimensional GloVe [PSM14] and train a one layer neural network with 1000 hidden units. Such classifier achieves 95% test accuracy, however it entails major individual fairness violation. Consider an application of this sentiment classifier to summarizing customer reviews, tweets or news articles. Human names are typical in such texts and should not affect the sentiment score, hence we consider fair metric between any pair of names to be 0. Then sentiment score for all names should

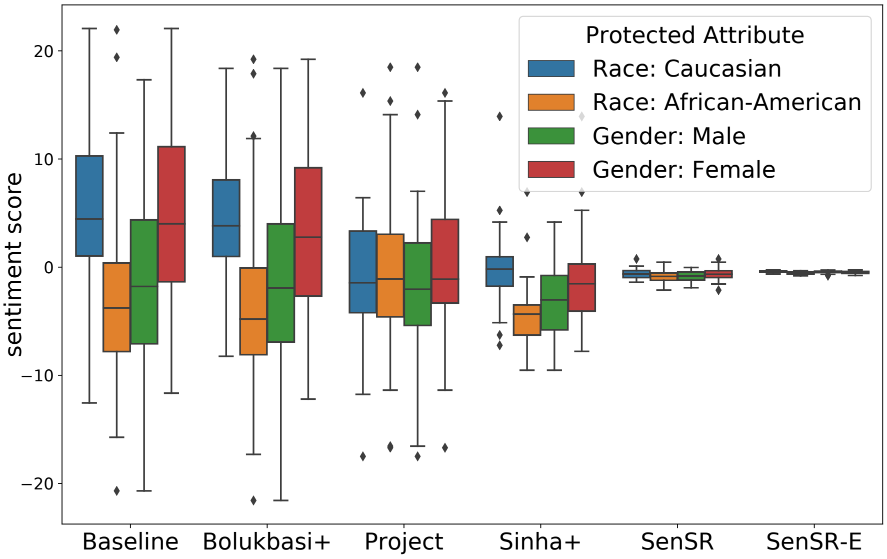


Figure 4.2: Box-plots of sentiment scores

be the same to satisfy the individual fairness. To make a connection to group fairness, following the study of [CBN17] that reveals the biases in word embeddings, we evaluate the fairness of our sentiment classifier using male and female names typical for Caucasian and African-American ethnic groups. We emphasize that to satisfy individual fairness, the sentiment of *any* name should be the same.

Comparison metrics To evaluate the gap between two groups of names, $\mathcal{N}_0$ for Caucasian (or female) and $\mathcal{N}_1$ for African-American (or male), we report $\frac{1}{|\mathcal{N}_0|}\sum_{n\in\mathcal{N}_0}(h(n)_1 - h(n)_0) - \frac{1}{|\mathcal{N}_1|}\sum_{n\in\mathcal{N}_1}(h(n)_1 - h(n)_0)$, where $h(n)_k$ is logits for class k of name n ($k = 1$ is the positive class). We use list of names provided in [CBN17], which consists of 49 Caucasian and 45 African-American names, among those 48 are female and 46 are male. The gap between African-American and Caucasian names is reported as Race gap, while the gap between male and female names is reported as Gend. gap in Table 4.1. As in [Spe17], we also compare sentiment difference of two sentences: "Let's go get Italian food" and "Let's go get Mexican food", i.e. cuisine gap (abbreviated Cuis. gap in Table 4.1), as a test of generalization beyond names. To embed these sentences we average their word embeddings.

Sensitive subspace We consider embeddings of 94 names that we use for evaluation as sensitive directions, which may be regarded as utilizing the expert knowledge, i.e. these names form a list of words that an expert believes should be treated equally. The fair metric is then defined using an orthogonal complement projector of the span of sensitive directions as we discussed in Section 4.2.1. When expert knowledge is not available, or we wish to achieve general fairness for names, we utilize a side dataset of popular baby names in New York City.[2] The dataset has 11k names, however only 32 overlap with the list of names used for evaluation. Embeddings of these names define a group of comparable samples that we use to learn sensitive directions with SVD (see Appendix 4.6.2 and Algorithm 3 for details). We take top 50 singular vectors to form the sensitive subspace. It is worth noting that, unlike many existing approaches in the fairness literature, we do not use any protected attribute information. Our algorithm only utilizes training words, their sentiments and a vanilla list of names.

Results From the box-plots in Figure 4.2, we see that both race and gender gaps are significant when using the baseline neural network classifier. It tends to predict Caucasian names as "positive", while the median for African-American names is negative; the median sentiment for female names is higher than that for male names. We considered three other approaches to this problem: the algorithm of [BCZ$^+$16] for pre-processing word embeddings; pre-processing via projecting out the sensitive subspace that we used for training SenSR (this is analogous to [PTB19]); training a distributionally robust classifier with Euclidean distance cost [SND18]. All approaches improved upon the baseline, however only SenSR can be considered individually fair. Our algorithm practically eliminates gender and racial gaps and achieves the notion of individual fairness as can be seen from almost equal predicted sentiment score for *all* names. We remark that using expert knowledge (i.e. evaluation names) allowed SenSR-E (E for expert) to further improve both group and individual fairness. However we warn practitioners that if the expert knowledge is too specific, generalization outside of the expert knowledge may not be very good. In Table 4.1 we report results averaged across 10 repetitions with 90%/10% train/test splits, where we also verify that accuracy trade-off with the baseline is minor. In the right column we present the generalization check, i.e. comparing a pair of sentences unrelated to names. Utilizing expert knowledge led to a fairness

[2] titled "Popular Baby Names" and available from `https://catalog.data.gov/dataset/`

Table 4.2: Summary of *Adult* classification experiments over 10 restarts

	B-Acc,%	S-Con.	GR-Con.	$\text{Gap}_G^{\text{RMS}}$	$\text{Gap}_R^{\text{RMS}}$	$\text{Gap}_G^{\max}$	$\text{Gap}_R^{\max}$
SenSR	78.9	**.934**	.984	**.068**	**.055**	**.087**	**.067**
Baseline	**82.9**	.848	.865	.179	.089	.216	.105
Project	82.7	.868	**1.00**	.145	.064	.192	.086
Adv. Debias.	81.5	.807	.841	.082	.070	.110	.078
CoCL	79.0	-	-	.163	.080	.201	.109

over-fitting effect, however we still see improvement over other methods. When utilizing the SVD of a larger dataset of names we observe better generalization. Our generalization check suggests that fairness over-fitting is possible, therefore datasets and procedure for verifying fairness generalization are needed.

4.4.2 Adult Income Prediction

Problem formulation Demonstrating the broad applicability of SenSR outside of natural language processing tasks, we apply SenSR to a classification task on the *Adult* [DG17a] data set to predict whether an individual makes at least \$50k based on features like gender and occupation for approximately 45,000 individuals. Models that predict income without fairness considerations can contribute to the problem of differences in pay between genders or races for the same work. Throughout this section, gender (male or female) and race (Caucasian or non-Caucasian) are binary.

Comparison metrics Arguably a classifier is individually unfair if the classifications for two data points that are the same on all features except demographic features are different. Therefore, to assess individual fairness, we report spouse consistency (S-Con.) and gender and race consistency (GR-Con.), which are measures of how often classifications change only because of differences in demographic features. For S-Con (resp. GR-con), we make 2 (resp. 4) copies of every data point where the only difference is that one is a husband and the other is a wife (resp. difference is in gender and race). S-Con (resp. GR-Con) is the fraction of corresponding pairs (resp. quadruples) that have the same classification. We also report various group fairness measures proposed by [DARW$^+$19] with respect to race or gender based on true positive rates, i.e. the ability of a classifier to correctly identify a given class. See Appendix 4.6.4 for the definitions. We report $\text{Gap}_R^{\text{RMS}}$, $\text{Gap}_G^{\text{RMS}}$, $\text{Gap}_R^{\max}$, and $\text{Gap}_G^{\max}$ where R refers to race, and G refers to gender. We use balanced accuracy (B-acc) instead of accuracy[3] to measure predictive ability since only 25% of individuals make at least \$50k.

Sensitive subspace Let $\{(x_i, x_{g_i})\}_{i=1}^m$ be the set of features $x_i \in \mathbb{R}^D$ of the data except the coordinate for gender is zeroed and where x_{g_i} indicates the gender of individual i. For $\gamma > 0$, let $w_g = \arg\min_{w\in\mathbb{R}^D} \frac{1}{m}\sum_{i=1}^m -x_{g_i}(w^T x_i) + \log(1 + e^{w^T x_i}) + \gamma\|w\|_2$, i.e. w_g is the learned hyperplane that classifies gender given by regularized logistic regression. Let $e_g \in \mathbb{R}^D$ (resp. e_r) be the vector that is 1 in the gender (resp. race) coordinate and 0 elsewhere. Then the sensitive subspace is the span of $[w_g, e_g, e_r]$. See Appendix 4.6.2 for details.

Results See Table 4.2 for the average[4] of each metric on the test sets over ten 80%/20% train/test splits for Baseline, Project (projecting features onto the orthogonal complement of the sensitive subspace before training), CoCL [DARW$^+$19], Adversarial Debiasing [ZLM18], and SenSR. With the exception of CoCL [DARW$^+$19], each classifier is a 100 unit single hidden layer neural network. The Baseline clearly exhibits individual and group fairness violations. While SenSR has the lowest B-acc, SenSR is the best by a large margin for S-Con. and has the best group fairness measures. We expect SenSR to do well on GR-consistency since the sensitive subspace includes the race and gender directions. However, SenSR's individually fair performance generalizes: the sensitive directions do not directly use the husband and wife directions, yet SenSR performs well on S-Con. Furthermore, SenSR outperforms Project on S-Con and group fairness measures illustrating that SenSR does much more than just ignoring the sensitive subspace. CoCL only barely improves group fairness compared to the baseline with a significant drop in B-acc and while Adversarial Debiasing also improves group fairness, it is worse than the baseline on individual fairness measures illustrating that group fairness does not imply individual fairness.

[3] Accuracy is reported in Table 4.4 in Appendix 4.6.4.

[4] The standard error is reported in the supplement. Each standard error is within 10^{-2}.

4.5 Summary

We consider the task of training ML systems that are fair in the sense that their performance is invariant under certain perturbations in a sensitive subspace. This notion of fairness is a variant of individual fairness [DHP+12]. One of the main barriers to the adoption of individual fairness is the lack of consensus on a fair metric for many ML tasks. To circumvent this issue, we consider two approaches to learning a fair metric from data: one for problems in which the sensitive attribute is observed, and another for problems in which the sensitive attribute is unobserved. Given a data-driven choice of fair metric, we provide an algorithm that provably trains individually fair ML models.

4.6 Supplement

4.6.1 Proofs

Proof of Proposition 4.3.1

By the duality result of [BM19], for any $\epsilon > 0$,

$$\begin{aligned}
&\sup_{P:W_*(P,P_*)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] - \sup_{P:W(P,P_n)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] \\
&\quad = \inf_{\lambda\ge 0}\big\{\lambda\epsilon + \mathbb{E}_{P_*}\big[\ell_\lambda^{c_*}(Z,\theta)\big]\big\} - (\lambda_n\epsilon + \mathbb{E}_{P_n}\big[\ell_{\lambda_n}^{c}(Z,\theta)\big]) \\
&\quad \le \mathbb{E}_{P_*}\big[\ell_{\lambda_n}^{c_*}(Z,\theta)\big] - \mathbb{E}_{P_n}\big[\ell_{\lambda_n}^{c}(Z,\theta)\big],
\end{aligned}$$

where $\lambda_n \in \arg\min_{\lambda\ge 0}\lambda\epsilon + \mathbb{E}_{P_n}\big[\ell_\lambda^c(Z,\theta)\big]$. By assumption A3,

$$\begin{aligned}
&|\ell_{\lambda_n}^{c_*}(z,\theta) - \ell_{\lambda_n}^{c}(z,\theta)| \\
&\quad = \left|\sup_{x_2\in\mathcal{X}} \ell((x_2,y),\theta) - \lambda_n c_*((x,y),(x_2,y)) - \sup_{x_2\in\mathcal{X}} \ell((x_2,y),\theta) - \lambda_n c((x,y),(x_2,y))\right| \\
&\quad \le \sup_{x_2\in\mathcal{X}} \lambda_n |c_*((x,y),(x_2,y)) - c((x,y),(x_2,y))| \\
&\quad \le \lambda_n\delta_c \cdot D^2.
\end{aligned}$$

This implies

$$\begin{aligned}
&\sup_{P:W_*(P,P_*)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] - \sup_{P:W(P,P_n)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] \\
&\quad\le \mathbb{E}_{P_*}\big[\ell^{c_*}_{\lambda_n}(Z,\theta)\big] - \mathbb{E}_{P_n}\big[\ell^{c_*}_{\lambda_n}(Z,\theta)\big] + \lambda_n\delta_c D^2.
\end{aligned}$$

This bound is crude; it is possible to obtain sharper bounds under additional assumptions on the loss and transportation cost functions. We avoid this here to keep the result as general as possible.

Similarly,

$$\begin{aligned}
&\sup_{P:W(P,P_n)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] - \sup_{P:W_*(P,P_*)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] \\
&\quad\le \mathbb{E}_{P_n}\big[\ell^{c}_{\lambda_*}(Z,\theta)\big] - \mathbb{E}_{P_*}\big[\ell^{c_*}_{\lambda_*}(Z,\theta)\big] \\
&\quad\le \mathbb{E}_{P_n}\big[\ell^{c_*}_{\lambda_*}(Z,\theta)\big] - \mathbb{E}_{P_*}\big[\ell^{c_*}_{\lambda_*}(Z,\theta)\big] + \lambda_*\delta_c D^2,
\end{aligned}$$

where $\lambda_* \in \arg\min_{\lambda\ge 0}\{\lambda\epsilon + \mathbb{E}_{P_*}\big[\ell^{c_*}_{\lambda}(Z,\theta)\big]\}$.

Lemma 4.6.1 ([LR18]). *Let $\tilde{\lambda} \in \arg\min_{\lambda\ge 0}\lambda\epsilon + \mathbb{E}_P\big[\ell^{c}_{\lambda}(Z,\theta)\big]$. As long as the function in the loss class are L-Lipschitz with respect to d_x (see Assumption A2), $\tilde{\lambda} \le \frac{L}{\sqrt{\epsilon}}$.*

Proof. By the optimality of $\tilde{\lambda}$,

$$\begin{aligned}
\tilde{\lambda}\epsilon &\le \tilde{\lambda}\epsilon + \mathbb{E}_P\big[\sup_{x_2\in\mathcal{X}} \ell((x_2,Y),\theta) - \tilde{\lambda}d_x(X,x_2)^2 - \ell((X,Y),\theta)\big] \\
&= \tilde{\lambda}\epsilon + \mathbb{E}_P\big[\ell^{c}_{\tilde{\lambda}}(Z,\theta) - \ell(Z,\theta)\big] \\
&\le \lambda\epsilon + \mathbb{E}_P\big[\ell^{c}_{\lambda}(Z,\theta) - \ell(Z,\theta)\big] \\
&= \lambda\epsilon + \mathbb{E}_P\big[\sup_{x_2\in\mathcal{X}} \ell((x_2,Y),\theta) - \ell((X,Y),\theta) - \lambda d_x(X,x_2)^2\big]
\end{aligned}$$

for any $\lambda \ge 0$. By Assumption A2, the right side is at most

$$\begin{aligned}
\tilde{\lambda}\epsilon &\le \lambda\epsilon + \mathbb{E}_P\big[\sup_{x_2\in\mathcal{X}} Ld_x(X,x_2) - \lambda d_x(X,x_2)^2\big] \\
&\le \lambda\epsilon + \sup_{t\ge 0} Lt - \lambda t^2
\end{aligned}$$

We minimize the right side with respect to t (set $t = \frac{L}{2\lambda}$) and λ (set $\lambda = \frac{L}{2\sqrt{\epsilon}}$) to obtain $\tilde{\lambda}\epsilon \le L\sqrt{\epsilon}$. □

By Lemma 4.6.1, we have

$$\begin{aligned}
&\sup_{P:W_*(P,P_*)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] - \sup_{P:W(P,P_n)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] \\
&\le \mathbb{E}_{P_*}\big[\ell^{c_*}_{\lambda_n}(Z,\theta)\big] - \mathbb{E}_{P_n}\big[\ell^{c_*}_{\lambda_n}(Z,\theta)\big] + \frac{L\delta_c D^2}{\sqrt{\epsilon}} \text{ and} \\
&\sup_{P:W(P,P_n)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] - \sup_{P:W_*(P,P_*)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] \\
&\le \mathbb{E}_{P_n}\big[\ell^{c_*}_{\lambda_*}(Z,\theta)\big] - \mathbb{E}_{P_*}\big[\ell^{c_*}_{\lambda_*}(Z,\theta)\big] + \frac{L\delta_c D^2}{\sqrt{\epsilon}}.
\end{aligned}$$

We combine the preceding bounds to obtain

$$\begin{aligned}
&\left|\sup_{P:W(P,P_n)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] - \sup_{P:W_*(P,P_*)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big]\right| \\
&\quad\le \sup_{f\in\mathcal{L}^{c_*}} \big|\textstyle\int_{\mathcal{Z}} f(z)d(P_n-P_*)(z)\big| + \frac{L\delta_c D^2}{\sqrt{\epsilon}},
\end{aligned}$$

where $\mathcal{L}^{c_*} = \{\ell^{c_*}_\lambda(\cdot,\theta) : \lambda\in[0,\frac{L}{\sqrt{\epsilon}}], \theta\in\Theta\}$ is the DR loss class. In the rest of the proof, we bound $\sup_{f\in\mathcal{L}^{c_*}} \big|\int_{\mathcal{Z}} f(z)d(P_*-P_n)(z)\big|$ with standard techniques from statistical learning theory. Assumption A2 implies the functions in $\mathcal{F}$ are bounded:

$$0 \le \ell((x_1,y_1),\theta) - \cancel{\lambda d_x(x_1,x_1)} \le \ell^c_\lambda(z_1,\theta) \le \sup_{x_2\in\mathcal{X}} \ell((x_2,y_1),\theta) \le M.$$

This implies has bounded differences, so δ_n concentrates sharply around its expectation. By the bounded-differences inequality and a symmetrization argument,

$$\sup_{f\in\mathcal{L}^{c_*}} \big|\textstyle\int_{\mathcal{Z}} f(z)d(P_n-P_*)(z)\big| \le 2\mathfrak{R}_n(\mathcal{L}^{c_*}) + M(\frac{\log\frac{2}{t}}{2n})^{\frac{1}{2}}$$

WP at least $1-t$, where $\mathfrak{R}_n(\mathcal{F})$ is the Rademacher complexity of $\mathcal{F}$:

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)\Big].$$

Lemma 4.6.2. *The Rademacher complexity of the DR loss class is at most*

$$\mathfrak{R}_n(\mathcal{L}^c) \leq \frac{24\mathfrak{C}(\mathcal{L})}{\sqrt{n}} + \frac{24LD^2}{\sqrt{n\epsilon}}.$$

Proof. To study the Rademacher complexity of $\mathcal{L}^c$, we first show that the $\mathcal{L}^c$-indexed Rademacher process $X_f \triangleq \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)$ is sub-Gaussian with respect to to a pseudometric. Let $f_1 = \ell^c_{\lambda_1}(\cdot, \theta_1)$ and $f_2 = \ell^c_{\lambda_2}(\cdot, \theta_2)$. Define

$$d_{\mathcal{L}^c}(f_1, f_2) \triangleq \|\ell(\cdot, \theta_1) - \ell(\cdot, \theta_2)\|_\infty + D^2|\lambda_1 - \lambda_2|.$$

We check that X_f is sub-Gaussian with respect to $d_{\mathcal{L}^c}$:

$$\begin{aligned}
&\mathbb{E}\big[\exp(t(X_{f_1} - X_{f_2}))\big] \\
&= \mathbb{E}\big[\exp\big(\frac{t}{n}\sum_{i=1}^n \sigma_i(\ell^c_{\lambda_1}(Z_i, \theta_1) - \ell^c_{\lambda_2}(Z_i, \theta_2))\big)\big] \\
&= \mathbb{E}\big[\exp\big(\frac{t}{n}\sigma(\ell^c_{\lambda_1}(Z, \theta_1) - \ell^c_{\lambda_2}(Z, \theta_2))\big)\big]^n \\
&= \mathbb{E}\big[\exp\big(\frac{t}{n}\sigma(\sup_{x_1\in\mathcal{X}}\inf_{x_2\in\mathcal{X}} \ell((x_1, Y), \theta_1) - \lambda_1 d_x(x_1, X)^2 - \ell((x_2, Y), \theta_2) + \lambda_2 d_x(X, x_2)^2))\big)\big]^n \\
&= \mathbb{E}\big[\exp\big(\frac{t}{n}\sigma(\sup_{x_1\in\mathcal{X}} \ell((x_1, Y), \theta_1) - \ell((x_1, Y), \theta_2) + (\lambda_2 - \lambda_1)d_x(x_1, X)^2))\big)\big]^n \\
&\quad \leq \exp\big(\tfrac{1}{2}t^2 d_{\mathcal{L}^c}(f_1, f_2)\big).
\end{aligned}$$

Let $N(\mathcal{L}^c, d_{\mathcal{L}^c}, \epsilon)$ be the ϵ-covering number of $(\mathcal{L}^c, d_{\mathcal{L}^c})$. We observe

$$N(\mathcal{L}^c, d_{\mathcal{L}^c}, \epsilon) \leq N(\mathcal{L}, \|\cdot\|_\infty, \tfrac{\epsilon}{2}) \cdot N([0, \tfrac{L}{\sqrt{\epsilon}}], |\cdot|, \tfrac{\epsilon}{2D^2}) \tag{4.12}$$

By Dudley's entropy integral,

$$
\begin{aligned}
\mathfrak{R}_n(\mathcal{L}^c) &\leq \frac{12}{\sqrt{n}} \int_0^\infty \log N(\mathcal{L}^c, d_{\mathcal{L}^c}, \epsilon)^{\frac{1}{2}} d\epsilon \\
&\leq \frac{12}{\sqrt{n}} \int_0^\infty \big(\log N(\mathcal{L}, \|\cdot\|_\infty, \tfrac{\epsilon}{2}) + N\big([0, \tfrac{L}{\sqrt{\epsilon}}], |\cdot|, \tfrac{\epsilon}{2D^2}\big)\big)^{\frac{1}{2}} d\epsilon \\
&\leq \frac{12}{\sqrt{n}} \bigg(\int_0^\infty \log N(\mathcal{L}, \|\cdot\|_\infty, \tfrac{\epsilon}{2})^{\frac{1}{2}} d\epsilon + \int_0^\infty N\big([0, \tfrac{L}{\sqrt{\epsilon}}], |\cdot|, \tfrac{\epsilon}{2D^2}\big)^{\frac{1}{2}} d\epsilon \bigg) \\
&\leq \frac{24\mathfrak{C}(\mathcal{L})}{\sqrt{n}} + \frac{24LD^2}{\sqrt{n\epsilon}} \int_0^{\frac{1}{2}} \log(\tfrac{1}{\epsilon}) d\epsilon
\end{aligned}
$$

where we recalled Equation (4.12) in the second step. We evalaute the integral on the right side to arrive at the stated bound: $\int_0^{\frac{1}{2}} \log(\frac{1}{\epsilon}) d\epsilon < 1$. □

By Lemma 4.6.2,

$$
\sup_{f \in \mathcal{L}^{c*}} \left| \int_{\mathcal{Z}} f(z) d(P_n - P_*)(z) \right| \leq \frac{48\mathfrak{C}(\mathcal{L})}{\sqrt{n}} + \frac{48LD^2}{\sqrt{n\epsilon}} + M(\frac{\log \frac{2}{t}}{2n})^{\frac{1}{2}},
$$

which implies

$$
\begin{aligned}
&\left| \sup_{P:W(P,P_n)\leq\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] - \sup_{P:W_*(P,P_*)\leq\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] \right| \\
&\quad \leq \frac{48\mathfrak{C}(\mathcal{L})}{\sqrt{n}} + \frac{48LD^2}{\sqrt{n\epsilon}} + \frac{L\delta_c D^2}{\sqrt{\epsilon}} + M(\frac{\log \frac{2}{t}}{2n})^{\frac{1}{2}}.
\end{aligned}
$$

WP at least $1 - t$.

Proofs of Propositions 4.3.2 and 4.3.3

Proof of Proposition 4.3.2. It is enough to show

$$
\sup\nolimits_{P:W_*(P,P_*)\leq\epsilon} \mathbb{E}_P\big[\ell(Z,\widehat{\boldsymbol{\theta}})\big] \leq \delta^* + 2\delta_n
$$

because the loss function is non-negative. We have

$$\begin{aligned}\sup_{P:W_*(P,P_*)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\widehat{\boldsymbol{\theta}})\big] &\le \sup_{P:W(P,P_n)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\widehat{\boldsymbol{\theta}})\big] + \delta_n \\ &\le \sup_{P:W(P,P_n)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\bar{\theta})\big] + \delta_n \\ &\le \sup_{P:W_*(P,P_*)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\bar{\theta})\big] + 2\delta_n \\ &\le \delta^* + 2\delta_n.\end{aligned}$$

□

Proof of Proposition 4.3.3.

$$\begin{aligned}&\sup_{P:W_*(P,P_n)\le\epsilon} \big(\mathbb{E}_P\big[\ell(Z,\theta)\big] - \mathbb{E}_{P_n}\big[\ell(Z,\theta)\big]\big) - \sup_{P:W(P,P_*)\le\epsilon} \big(\mathbb{E}_P\big[\ell(Z,\theta)\big] - \mathbb{E}_{P_*}\big[\ell(Z,\theta)\big]\big) \\ &= \sup_{P:W_*(P,P_*)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] - \sup_{P:W(P,P_n)\le\epsilon} \mathbb{E}_P\big[\ell(Z,\theta)\big] + \mathbb{E}_{P_*}\big[\ell(Z,\theta)\big] - \mathbb{E}_{P_n}\big[\ell(Z,\theta)\big] \\ &\le \delta_n + \mathbb{E}_{P_*}\big[\ell(Z,\theta)\big] - \mathbb{E}_{P_n}\big[\ell(Z,\theta)\big]\end{aligned}$$

The loss function is bounded, so it is possible to bound $\mathbb{E}_{P_*}\big[\ell(Z,\theta)\big] - \mathbb{E}_{P_n}\big[\ell(Z,\theta)\big]$ by standard uniform convergence results on bounded loss classes. □

4.6.2 Data-Driven Fair Metrics

Learning the fair metric from observations of the sensitive attribute

Here we assume the sensitive attribute is discrete and is observed for a small subset of the training data. Formally, we assume this subset of the training data has the form $\{(X_i, K_i, Y_i)\}$, where K_i is the sensitive attribute of the i-th subject. To learn the sensitive subspace, we fit a softmax regression model to the data

$$\Pr(K_i = l \mid X_i) = \frac{\exp(a_l^T X_i + b_l)}{\sum_{l=1}^k \exp(a_l^T X_i + b_l)},\ l = 1, \dots, k,$$

and take the span of $A = \Big[a_1 \dots a_k\Big]$ as the sensitive subspace to define the fair metric as

$$d_x(x_1, x_2)^2 = (x_1 - x_2)^T (I - P_{\mathrm{ran}(A)})(x_1 - x_2). \tag{4.13}$$

This approach readily generalizes to sensitive attributes that are not discrete-valued: replace the softmax model by an appropriate generalized linear model.

In many applications, the sensitive attribute is part of a user's demographic information, so it may not be available due to privacy restrictions. This does not preclude the proposed approach because the sensitive attribute is only needed to learn the fair metric and is neither needed to train the classifier nor at test time.

Learning the fair metric from comparable samples

In this section, we consider the task of learning a fair metric from supervision in a form of comparable samples. This type of supervision has been considered in the literature on debiasing learned representations. For example, method of [BCZ$^+$16] for removing gender bias in word embeddings relies on sets of words whose embeddings mainly vary in a gender subspace (e.g. (king, queen)).

To keep things simple, we focus on learning a generalized Mahalanobis distance

$$d_x(x_1, x_2) = (\varphi(x_1) - \varphi(x_2))^T \widehat{\mathbf{\Sigma}} (\varphi(x_1) - \varphi(x_2))^{\frac{1}{2}}, \tag{4.14}$$

where $\varphi(x) : \mathcal{X} \to \mathbf{R}^d$ is a *known* feature map and $\widehat{\mathbf{\Sigma}} \in \mathbf{S}_+^{d \times d}$ is a covariance matrix. Our approach is based on a factor model

$$\varphi_i = A_* u_i + B_* v_i + \epsilon_i,$$

where $\varphi_i \in \mathbf{R}^d$ is the learned representation of x_i, $u_i \in \mathbf{R}^K$ (resp. $v_i \in \mathbf{R}^L$) is the sensitive/irrelevant (resp. relevant) attributes of x_i to the task at hand, and ϵ_i is an error term. For example, in [BCZ$^+$16], the learned representations are the embeddings of words in the vocabulary, and the sensitive attribute is the gender bias of the words. The sensitive and relevant attributes are generally unobserved.

Recall our goal is to obtain $\widehat{\mathbf{\Sigma}}$ so that Equation (4.14) is small whenever $v_1 \approx v_2$. One possible choice of $\widehat{\mathbf{\Sigma}}$ is the projection matrix onto the orthogonal complement of $\text{ran}(A)$, which we denote by $P_{\text{ran}(A)}$. Indeed,

$$d_x(x_1, x_2)^2 = (\varphi_1 - \varphi_2)^T (I - P_{\text{ran}(A)})(\varphi_1 - \varphi_2) \tag{4.15}$$

$$\approx (v_1 - v_2)^T B_*^T (I - P_{\mathrm{ran}(A)}) B_* (v_1 - v_2), \tag{4.16}$$

which is small whenever $v_1 \approx v_2$. Although $\mathrm{ran}(A)$ is unknown, it is possible to estimate it from the learned representations and groups of comparable samples by factor analysis.

The factor model attributes variation in the learned representations to variation in the sensitive and relevant attributes. We consider two samples comparable if their relevant attributes are similar. In other words, if $\mathcal{I} \subset [n]$ is (the indices of) a group of comparable samples, then

$$H\Phi_{\mathcal{I}} = HU_{\mathcal{I}}A_*^T + \cancelto{\approx 0}{HV_{\mathcal{I}}B_*^T} + HE_{\mathcal{I}} \approx HU_{\mathcal{I}}A_*^T + HE_{\mathcal{I}}, \tag{4.17}$$

where $H = I_{|\mathcal{I}|} - \frac{1}{|\mathcal{I}|}1_{|\mathcal{I}|}1_{|\mathcal{I}|}^T$ is the centering or de-meaning matrix and the rows of $\Phi_{\mathcal{I}}$ (resp. $U_{\mathcal{I}}$, $V_{\mathcal{I}}$) are φ_i (resp. u_i, v_i). If this group of samples have identical relevant attributes, i.e., $V_{\mathcal{I}} = 1_{|\mathcal{I}|}v^T$ for some v, then $HV_{\mathcal{I}}$ vanishes exactly. As long as u_i and ϵ_i are uncorrelated (e.g, $\mathbb{E}\big[u_i\epsilon_i^T\big] = 0$), Equation (4.17) implies

$$\mathbb{E}\big[\Phi_{\mathcal{I}}^T H\Phi_{\mathcal{I}}\big] \approx A\mathbb{E}\big[U_{\mathcal{I}}^T HU_{\mathcal{I}}\big]A^T + \mathbb{E}\big[E_{\mathcal{I}}^T HE_{\mathcal{I}}\big],$$

This suggests estimating $\mathrm{ran}(A)$ from the learned representations and groups of comparable samples by factor analysis. We summarize our approach in Algorithm 3.

Algorithm 3 estimating $\widehat{\boldsymbol{\Sigma}}$ for the fair metric

```
Input: $\{\varphi_i\}_{i=1}^n$, comparable groups $\mathcal{I}_1, \dots, \mathcal{I}_G$
$\widehat{A}^T \in \arg\min_{W_g, A}\{\frac{1}{2}\sum_{g=1}^G \|H_g\Phi_{\mathcal{I}_g} - W_gA^T\|_F^2\}$        ▷ factor analysis
$Q \leftarrow \mathtt{qr}(\widehat{A})$        ▷ get orthonormal basis of $\mathrm{ran}(\widehat{A})$
$\widehat{\boldsymbol{\Sigma}} \leftarrow I_d - QQ^T$
```

4.6.3 SenSR Implementation Details

This section is to accompany the implementation of the SenSR algorithm and is best understood by reading it along with the code implemented using TensorFlow.[5] We discuss choices of learning rates and few specifics of the code. Words in *italics* correspond

[5] https://github.com/IBM/sensitive-subspace-robustness

to variables in the code and following notation in parentheses defines corresponding name in Table 4.3, where we summarize all hyperparameter choices.

Handling class imbalance Datasets we study have imbalanced classes. To handle it, on every *epoch*(E) (i.e. number of epochs) we subsample a *batch_size*(B) training samples enforcing equal number of observations per class. This procedure can be understood as data augmentation.

Perturbations specifics Our implementation of SenSR algorithm has two inner optimization problems — subspace perturbation and full perturbation (when $\epsilon > 0$). Subspace perturbation can be viewed as an initialization procedure for the attack. We implement both using Adam optimizer [KB15] inside the computation graph for better efficiency, i.e. defining corresponding perturbation parameters as Variables and re-setting them to zeros after every epoch. This is in contrast with a more common strategy in the adversarial robustness implementations, where perturbations (i.e. attacks) are implemented using tf.gradients with respect to the input data defined as a Placeholder.

Learning rates As mentioned above, in addition to regular Adam optimizer for learning the parameters we invoke two more for the inner optimization problems of SenSR. We use same learning rate of 0.001 for the parameters optimizer, however different learning rates across datasets for *subspace_step*(s) and *full_step*(f). Two other related parameters are number of steps of the inner optimizations: *subspace_epoch*(se) and *full_epoch*(fe). We observed that setting subspace perturbation learning rate too small may prevent our algorithm from reducing unfairness, however setting it big does not seem to hurt. On the other hand, learning rate for full perturbation should not be set too big as it may prevent algorithm from solving the original task. Note that full perturbation learning rate should be smaller than perturbation budget *eps*(ϵ) — we always use $\epsilon/10$. In general, malfunctioning behaviors are immediately noticeable during training and can be easily corrected, therefore we did not need to use any hyperparameter optimization tools.

Table 4.3: SenSR hyperparameter choices in the experiments

	E	B	s	se	ϵ	f	fe
Sentiment	4K	1K	0.1	10	0.1	0.01	10
Adult	12K	1K	10	50	10^{-3}	10^{-4}	40

4.6.4 Additional Adult Experiment Details

Preprocessing

The continuous features in *Adult* are the following: `age`, `fnlwgt`, `capital-gain`, `capital-loss`, `hours-per-week`, and `education-num`. The categorical features are the following: `workclass`, `education`, `marital-stataus`, `occupation`, `relationship`, `race`, `sex`, and `native-country`. See [DG17a] for a description of each feature. We remove `fnlwgt` and `education` but keep `education-num`, which is a integer representation of education. We do not use `native-country`, but use `race` and `sex` as predictive features. We treat `race` as binary: individuals are either White or non-White. For every categorical feature, we use one hot encoding. For every continuous feature, we standardize, i.e., subtract the mean and divide by the standard deviation. We remove anyone with missing data leaving 45,222 individuals.

This data is imbalanced: 25% make at least $50k per year. Furthermore, there is demographic imbalance with respect to race and gender as well as class imbalance on the outcome when conditioning on race or gender: 86% of individuals are white of which 26% make at least $50k a year; 67% of individuals are male of which 31% make at least $50k a year; 11% of females make at least $50k a year; and 15% of non-whites make at least $50k a year.

Full experimental results

See Tables 4.4 and 4.5 for the full experiment results. The tables report the average and the standard error for each metric on the test set for 10 train and test splits.

Sensitive subspace

To learn the hyperplane that classifies females and males, we use our implementation of regularized logistic regression with a batch size of 5k, 5k epochs, and .1 ℓ_2 regularization.

Table 4.4: Summary of *Adult* classification experiments over 10 restarts

	Accuracy	B-TPR	$\text{Gap}_G^{\text{RMS}}$	$\text{Gap}_R^{\text{RMS}}$	$\text{Gap}_G^{\max}$	$\text{Gap}_R^{\max}$
SenSR	.787±.003	.789±.003	**.068**±.004	**.055**±.003	**.087**±.005	**.067**±.004
Baseline	**.813**±.001	**.829**±.001	.179±.004	.089±.003	.216±.003	.105±.003
Project	**.813**±.001	.827±.001	.145±.004	.064±.003	.192±.004	.086±.004
Adv. Debias.	.812±.001	.815±.002	.082±.005	.070±.006	.110±.006	.078±.005
CoCL	-	.790	.163	.080	.201	.109

Table 4.5: Summary of individual fairness metrics in *Adult* classification experiments over 10 restarts

	Spouse Consistency	Gender and Race Consistency
SenSR	**.934**±.012	.984±.000
Baseline	.848±.008	.865±.004
Project	.868±.005	**1**±0
Adv. Debias.	.807±.002	.841±.012

Hyperparameters and training

For each model, we use the same 10 train/test splits where use 80% of the data for training. Because of the class imbalance, each minibatch is sampled so that there are an equal number of training points from both the "income at least $50k class" and the "income below $50k class."

Baseline, Project, and SenSR See Table 4.3 for the hyperparameters we used when training Baseline, Project, and SenSR (Baseline and Project use a subset). Hyperparameters are defined in Appendix 4.6.3.

Advesarial debiasing We used [ZLM18]'s adversarial debiasing implementation in IBM's AIF360 package [BDH+18] where the source code was modified so that each minibatch is balanced with respect to the binary labels just as we did with our experiments and dropout was not used. Hyperparameters are the following: adversary loss weight = .001, num epochs = 500, batch size = 1000, and privileged groups are defined by binary gender and binary race.

Group fair metrics

Let $\mathcal{C}$ be a set of classes, A be a binary protected attribute and $Y, \hat{Y} \in \mathcal{C}$ be the true class label and the predicted class label. Then for $a \in \{0, 1\}$ and $c \in \mathcal{C}$ define $\text{TPR}_{a,c} = \mathbb{P}(\hat{Y} = c | A = a, Y = c)$; $\text{Gap}_{A,c} = \text{TPR}_{0,c} - \text{TPR}_{1,c}$; $\text{Gap}_A^{\text{RMS}} = \sqrt{\frac{1}{|C|} \sum_{c \in C} \text{Gap}_{A,c}^2}$; $\text{Gap}_A^{\max} = \arg\max_{c \in C} |\text{Gap}_{A,c}|$; Balanced Acc $= \frac{1}{|C|} \sum_{c \in C} \mathbb{P}(\hat{Y} = c | Y = c)$.

For Adult, we report $\text{Gap}_R^{\text{RMS}}$, $\text{Gap}_G^{\text{RMS}}$, $\text{Gap}_R^{\max}$, and $\text{Gap}_G^{\max}$ where $\mathcal{C}$ is composed of the two classes that correspond to whether someone made at least \$50k, R refers to race, and G refers to gender.

Chapter 5

Individually Fair Rankings

The work in this chapter is joint with Hamid Eftekhari, Mikhail Yurochkin and Yuekai Sun. Specifically, the theoretical work was mainly done by Hamid Eftekhari and the experimental work was mainly done by me. This work is currently under review.

5.1 Introduction

Information retrieval (IR) systems are everywhere in today's digital world, and ranking models are an integral part of many IR systems. In light of their ubiquity, issues of bias and unfairness in ranking models have come to the fore of the public's attention. In many applications, the items to be ranked are individuals, so bias in the output of ranking models may directly affect people's lives. For example, gender bias in job search engines directly affect the career success of female applicants [Das18].

There is a rapidly growing literature on detecting and mitigating algorithmic bias in machine learning (ML). The ML community has developed many formal definitions of algorithmic fairness along with algorithms to enforce these definitions [DHP^{+}12, HPPS16, BHJ^{+}18, KLRS17, RSZ17, YBS20]. Unfortunately, these issues have received less attention in the IR community. In particular, compared to the myriad of mathematical definitions of algorithmic fairness in the ML community, there are only a few definitions of algorithmic fairness for ranking. A recent review of fair ranking [Cas19] identifies two characteristics of fair rankings:

1. sufficient exposure of items from disadvantaged groups in rankings: Rankings should display a diversity of items. In particular, rankings should take care to display

items from disadvantaged groups to avoid allocative harms to items from such groups.

2. consistent treatment of similar items in rankings: Items with similar relevant attributes should be ranked similarly.

There is a line of work on fair ranking by [SJ18, SJ19] that focuses on the first characteristic. In this chapter, we complement this line of work by focusing on the second characteristic. In particular, we (i) specialize the notion of individual fairness in ML to rankings and (ii) devise an efficient algorithm for enforcing this notion in practice. We focus on the second characteristic since, in some sense, consistent treatment of similar items implies sufficient exposure: if there are items from disadvantaged groups that are similar to relevant items from advantaged groups, then a ranking model that treats similar items consistently will provide sufficient exposure to the items from disadvantaged groups.

5.1.1 Related work

Our work addresses the fairness of a learning-to-rank (LTR) system with respect to the items being ranked. The majority of work in this area requires a fair ranking to fairly allocate exposure (measured by the rank of an item in a ranking) to items. One line of work [YS17, ZBC$^+$17, CSV18, GAK19, CMV20, YGS19] requires a fair ranking to place a minimum number of minority group items in the top k ranks. Another line of work models the exposure items receive based on rank position and allocates exposure based on these exposure models and item relevance [SJ18, ZC20, BGW18, SJ19, SZR$^+$19]. There is some work that consider other fairness notions. The work of [KVR19] proposes error-based fairness criteria, and the framework of [AJSD19] can handle arbitrary fairness constraints given by an oracle. In contrast, we propose a fundamentally new definition: an individually fair ranking is invariant to sensitive perturbations of the features of the items. For example, consider ranking a set of job candidates, and consider the counterfactual set of candidates obtained from the original set by flipping each candidate's gender. We require that a fair LTR model produces the same ranking for both the original and counterfactual set.

The work in [ZBC$^+$17,CSV18,SJ18,BGW18,GAK19,CMV20,YGS19,WZW18,AJSD19] propose post-processing algorithms to obtain a fair ranking, i.e., algorithms that fairly re-rank items based on estimated relevance scores or rankings from potentially biased LTR models. However, post-processing techniques are insufficient since they can be mislead by biased estimated relevance scores [ZC20, SJ19] with the exception of the work in [CMV20] which assumes a specific bias model and provably counteracts this bias. In contrast, like [ZC20, SJ19], we propose an in-processing algorithm.

We consider individual fairness as opposed to group fairness [YS17, ZBC+17, CSV18, SJ18, ZC20, GAK19, SZR+19, KVR19, CMV20, YGS19, WZW18, AJSD19]. The merits of individual fairness over group fairness have been well established, e.g., group fair models can be blatantly unfair to individuals [DHP+12]. In fact, we show empirically that individual fairness is sufficient for group fairness but not vice versa. The work in [BGW18, SJ19] also considers individually fair LTR models. However, our notion of individual fairness is fundamentally different since we utilize a fair metric on queries like in the seminal work that introduced individual fairness [DHP+12] instead of measuring the similarity of items through relevance alone. To see the benefit of our approach, consider the job applicant example. If the training data does not contain highly ranked minority candidates, then at test time our LTR model will be able to correctly rank a minority candidate who should be highly ranked, which is not necessarily true for the work in [BGW18, SJ19].

5.2 Problem formulation

A query $q \in \mathcal{Q}$ to a ranker consists of a candidate set of n items that needs to be ranked $d^q \triangleq \{d_1^q, \dots, d_n^q\}$ and a set of relevance scores $\mathrm{rel}^q \triangleq \{\mathrm{rel}^q(d) \in \mathbb{R}\}_{d \in d^q}$. Each item is represented by a feature vector $\varphi(d) \in \mathcal{X}$ that describes the match between item d and query q where $\mathcal{X}$ is the feature space of the item representations. We consider stochastic ranking policies $\pi(\cdot \mid q)$ that are distributions over rankings r (i.e. permutations) of the candidate set. Our notation for rankings is $r(d)$: the rank of item d in ranking r (and $r^{-1}(j)$ is the j-ranked item). A policy generally consists of two components: a scoring model and a sampling method. The scoring model is a smooth ML model h_θ parameterized by θ (e.g. a neural network) that outputs a vector of scores: $h_\theta(\varphi(d^q)) \triangleq (h_\theta(\varphi(d_1^q)), \dots, h_\theta(\varphi(d_n^q)))$. The sampling method defines a distribution on rankings of the candidate set from the scores. For example, the Plackett-Luce [Pla75] model defines the probability of the ranking $r = \langle d_1, \ldots, d_n \rangle$ as

$$\pi_\theta(r \mid q) = \prod_{j=1}^{n} \frac{\exp(h_\theta(\varphi(d_j)))}{\exp(h_\theta(\varphi(d_j))) + \cdots + \exp(h_\theta(\varphi(d_n)))}. \tag{5.1}$$

To sample a ranking from the Placket-Luce model, items from a query are chosen without replacement where the probability of selecting items is given by the softmax of the scores of remaining items. The order in which the items are sampled defines the order of the ranking from best to worst. The goal of the LTR problem is finding a policy that has maximum expected utility:

$$\pi^* \triangleq \arg\max\nolimits_\pi \mathbb{E}_{q\sim Q}\big[U(\pi \mid q)\big] \text{ where } U(\pi \mid q) \triangleq \mathbb{E}_{r\sim\pi(\cdot\mid q)}\big[\Delta(r, \mathrm{rel}^q)\big], \tag{5.2}$$

where Q is the distribution of queries, $U(\pi \mid q)$ is the utility of a policy π for query q, and Δ is a ranking metric (e.g. normalized discounted cumulative gain). In practice, we solve the empirical version of Equation (5.2):

$$\widehat{\pi} \triangleq \arg\max\nolimits_\pi \frac{1}{N}\sum_{i=1}^{N}\big[U(\pi \mid q_i)\big], \tag{5.3}$$

where $\{q_i\}_{i=1}^N$ is a training set. If the policy is parameterized by θ, it is not hard to evaluate the gradient of the utility with respect to θ with the log-derivative trick:

$$\begin{aligned}\partial_\theta U(\pi_\theta \mid q) &= \partial_\theta \mathbb{E}_{r\sim\pi_\theta(\cdot\mid q)}\big[\Delta(r, \mathrm{rel}^q)\big] = \int \Delta(r, \mathrm{rel}^q)\partial_\theta \pi_\theta(r \mid q)dr \\ &= \int \Delta(r, \mathrm{rel}^q)\partial_\theta\{\log \pi_\theta(r \mid q)\}\pi_\theta(r \mid q)dr = \mathbb{E}_{r\sim\pi_\theta(\cdot\mid q)}\big[\Delta(r, \mathrm{rel}^q)\partial_\theta \log \pi_\theta(r \mid q)\big].\end{aligned}$$

In practice, we (approximately) evaluate $\partial_\theta U(\pi_\theta \mid q)$ by sampling from $\pi_\theta(\cdot \mid q)$. This set-up is mostly adopted from [YDJ19].

5.2.1 Fair Ranking via Invariance Regularization

We cast the fair ranking problem as training ranking policies that are invariant under certain sensitive perturbations to the queries. Let d_Q be a fair metric on queries that encode which queries should be treated similarly by the LTR model. For example, a

LTR model should similarly rank a set of job candidates and the counterfactual set of job candidates obtained from the original set via flipping the gender of each candidate. Hence, these two queries should be close according to $d_{\mathcal{Q}}$. We propose Sensitive Set Transport Invariant Ranking (SenSTIR) to enforce individual fairness in ranking via the following optimization problem:

$$\pi^* \triangleq \arg\max_{\pi} \mathbb{E}_{q \sim Q}\big[U(\pi \mid q)\big] - \rho R(\pi), \tag{SenSTIR}$$

such that $\rho > 0$ is a regularization parameter and

$$R(\pi) \triangleq \left\{ \begin{array}{ll} \sup_{\Pi \in \Delta(\mathcal{Q} \times \mathcal{Q})} & \mathbb{E}_{(q,q') \sim \Pi}\big[d_{\mathcal{R}}(\pi(\cdot \mid q), \pi(\cdot \mid q'))\big] \\ \text{subject to} & \mathbb{E}_{(q,q') \sim \Pi}\big[d_{\mathcal{Q}}(q, q')\big] \leq \epsilon \\ & \Pi(\cdot, \mathcal{Q}) = Q \end{array} \right\} \tag{5.4}$$

is an invariance regularizer where $d_{\mathcal{R}}$ is a metric on ranking policies, $\Delta(\mathcal{Q} \times \mathcal{Q})$ is the set of probability distributions on $\mathcal{Q} \times \mathcal{Q}$ where $\mathcal{Q}$ is the set of queries, and $\epsilon > 0$. At a high-level, individual fairness requires ML models to have similar outputs for similar inputs. This property is exactly what the regularizer encourages: the LTR model is encouraged to assign similar ranking policies (with respect to $d_{\mathcal{R}}$) to similar queries (with respect to $d_{\mathcal{Q}}$). The problem of enforcing invariance for individual fairness has been considered in supervised learning [YBS20, YS20]. However, these methods are not readily applicable to the LTR setting because of two main challenges: (i) defining a fair distance $d_{\mathcal{Q}}$ on queries, i.e., *sets* of items, and (ii) ensuring the resulting optimization problem is differentiable.

Optimal transport distance $d_{\mathcal{Q}}$ between queries We appeal to the machinery of optimal transport to define an appropriate metric $d_{\mathcal{Q}}$ on queries, i.e., *sets* of items. First, we need a fair metric on items $d_{\mathcal{X}}$ that encodes our intuition of which items should be treated similarly. Such a metric also appears in the traditional individual fairness definition [DHP+12] for classification and regression problems. Learning an individually fair metric is an important problem of its own that is actively studied in the recent literature [Ilv20, WGL+19, YBS20, MYBS20b]. In the experiment section, the fair metric on items $d_{\mathcal{X}}$ is learned from data using existing methods. The key idea is to view queries, i.e., *sets* of items, as distributions on $\mathcal{X}$ so that a metric between distributions can be used. In particular, to define $d_{\mathcal{Q}}$ from $d_{\mathcal{X}}$, we utilize an optimal transport distance between queries with $d_{\mathcal{X}}$ as the transport cost:

$$d_{\mathcal{Q}}(q,q') \triangleq \begin{cases} \inf_{\Pi \in \Delta(\mathcal{X}\times\mathcal{X})} & \int_{\mathcal{X}\times\mathcal{X}} d_{\mathcal{X}}(x,x')d\Pi(x,x') \\ \text{subject to} & \Pi(\cdot,\mathcal{X}) = \frac{1}{n}\sum_{j=1}^{n}\delta_{\varphi(d_j^q)} \\ & \Pi(\mathcal{X},\cdot) = \frac{1}{n}\sum_{j=1}^{n}\delta_{\varphi(d_j^{q'})} \end{cases}, \tag{5.5}$$

where $\Delta(\mathcal{X}\times\mathcal{X})$ is the set of probability distributions on $\mathcal{X}\times\mathcal{X}$ where $\mathcal{X}$ is the feature space of item representations and δ is the Dirac delta function.

5.3 Algorithm

In order to apply stochastic optimization to Equation (SenSTIR), we appeal to duality. In particular, using Theorem 2.3 of [YS20], if $d_{\mathcal{R}}(\pi(\cdot \mid q), \pi(\cdot \mid q')) - \lambda d_{\mathcal{Q}}(q,q')$ is continuous in (q,q') for all λ, then the invariance regularizer R can be written as

$$R(\pi) = \inf_{\lambda\geq 0}\{\lambda\epsilon + \mathbb{E}_{q\sim Q}[r_\lambda(\pi,q)]\}, \text{where} \tag{5.6}$$

$$r_\lambda(\pi,q) \triangleq \sup_{q'\in\mathcal{Q}}\{d_{\mathcal{R}}(\pi(\cdot \mid q), \pi(\cdot \mid q')) - \lambda d_{\mathcal{Q}}(q,q')\}. \tag{5.7}$$

In order to compute $r_\lambda(\pi,q)$, we can use gradient ascent on $u(q' \mid \pi,q,\lambda) \triangleq d_{\mathcal{R}}(\pi(\cdot \mid q), \pi(\cdot \mid q')) - \lambda d_{\mathcal{Q}}(q,q')$. We start by computing the gradient of $d_{\mathcal{Q}}(q,q')$ with respect to $x' \triangleq \varphi(d^{q'})$. Let $x \triangleq \varphi(d^q)$. Let $\Pi^\star(q,q')$ be the optimal transport plan for the problem defining $d_{\mathcal{Q}}(q,q')$, that is

$$d_{\mathcal{Q}}(q,q') = \int_{\mathcal{X}\times\mathcal{X}} d_{\mathcal{X}}(x,x')d\Pi^\star(x,x'),\ \Pi^\star(\cdot,\mathcal{X}) = \frac{1}{n}\sum_{j=1}^{n}\delta_{\varphi(d_j^q)},\ \Pi^\star(\mathcal{X},\cdot) = \frac{1}{n}\sum_{j=1}^{n}\delta_{\varphi(d_j^{q'})}.$$

The probability distribution $\Pi^\star(q,q')$ can be viewed as a coupling matrix where $\Pi^\star_{i,j} \triangleq \Pi^\star(\varphi(d_i^q), \varphi(d_j^{q'}))$. Using this notation we have

$$\partial_{x'_j} d_{\mathcal{Q}}(q,q') = \sum_{i=1}^{n} \Pi^\star_{i,j}\partial_2 d_{\mathcal{X}}(\varphi(d_i^q), \varphi(d_j^{q'})), \tag{5.8}$$

where $\partial_2 d_{\mathcal{X}}$ denotes the derivative of $d_{\mathcal{X}}$ with respect to its second input. If $d_{\mathcal{R}}(\pi_\theta(\cdot \mid q), \pi_\theta(\cdot \mid q')) = \|h_\theta(\varphi(d^q)) - h_\theta(\varphi(d^{q'}))\|_2^2/2$, then by Equation (5.8), a single iteration of gradient ascent on $d_{\mathcal{Q}}$ with step size γ for x' is

$$x_j'^{(l+1)} = x_j'^{(l)} + \gamma \left(\partial_{x'_j} h_\theta(x'^{(l)})^T (h_\theta(x'^{(l)}) - h_\theta(x)) - \lambda \sum_{i=1}^{n} \Pi^\star_{i,j} \partial_2 d_{\mathcal{X}}(x_i, x_j'^{(l)}) \right). \quad (5.9)$$

In our experiments, we use this choice of $d_{\mathcal{R}}$, which has been widely used, e.g., robustness in image classification [KKG18, YWHD19] and fairness [YS20]. However, our theory and set-up do not preclude other metrics. We can now present Algorithm 4, an alternating, stochastic algorithm, to solve Equation (SenSTIR).

Algorithm 4 SenSTIR: Sensitive Set Transport Invariant Ranking

```
**Require:** Initial Parameters: $\theta_0, \lambda_0, \epsilon, \rho$; Step Sizes: $\gamma, \alpha_t, \eta_t > 0$, Training queries: $\hat{Q}$
**repeat**
    Sample mini-batch $(q_{t_i}, \mathrm{rel}^{q_{t_i}})_{i=1}^B$ from $\hat{Q}$
    $q'_{t_i} \leftarrow \arg\max_{q'} \{\frac{1}{2}\|h_{\theta_t}(\varphi(d^{q_{t_i}})) - h_{\theta_t}(\varphi(d^{q'}))\|_2^2 - \lambda_t d_{\mathcal{Q}}(q_{t_i}, q')\}$, $i \in [B]$    ▷ Using (5.9)
    $\lambda_{t+1} \leftarrow \max\{0, \lambda_t + \alpha_t \rho(\epsilon - \frac{1}{B}\sum_{i=1}^B d_{\mathcal{Q}}(q_{t_i}, q'_{t_i}))\}$
    $\theta_{t+1} \leftarrow \theta_t + \eta_t(\frac{1}{B}\sum_{i=1}^B \partial_\theta\{U(\pi_{\theta_t} \mid q_{t_i})\} - \rho(\partial_\theta h_{\theta_t}(q'_{t_i}) - \partial_\theta h_{\theta_t}(q_{t_i}))^T (h_{\theta_t}(q'_{t_i}) - h_{\theta_t}(q_{t_i}))$
**until** converged
```

5.4 Theoretical Results

In this section, we study the generalization performance of the invariance regularizer $R(h_\theta) := R(\pi_\theta)$, which is an instance of a hierarchical optimal transport problem that does not have known uniform convergence results in the literature. Furthermore, the regularizer is not a separable function of the training examples so classical proof techniques are not applicable. To state the result, suppose that $\hat{d}_{\mathcal{X}}$ is an approximation of the fair metric $d_{\mathcal{X}}$ between items that is learned from data. The corresponding learned metric on queries is defined by

$$\hat{d}_{\mathcal{Q}}(q,q') \triangleq \begin{cases} \inf_{\Pi\in\Delta(\mathcal{X}\times\mathcal{X})} & \int_{\mathcal{X}\times\mathcal{X}} \hat{d}_{\mathcal{X}}(x,x')d\Pi(x,x') \\ \text{subject to} & \Pi(\cdot,\mathcal{X}) = \frac{1}{n}\sum_{j=1}^{n}\delta_{\varphi(d_j^q)} \\ & \Pi(\mathcal{X},\cdot) = \frac{1}{n}\sum_{j=1}^{n}\delta_{\varphi(d_j^{q'})} \end{cases}, \tag{5.10}$$

and the empirical regularizer is defined by

$$\hat{R}(h_\theta) \triangleq \begin{cases} \sup_{\Pi\in\Delta(\mathcal{Q}\times\mathcal{Q})} & \mathbb{E}_\Pi\big[d_{\mathcal{Y}}(h_\theta(\varphi(d^q)), h_\theta(\varphi(d^{q'}))\big] \\ \text{subject to} & \mathbb{E}_\Pi\big[\hat{d}_{\mathcal{Q}}(q,q')\big] \le \epsilon \\ & \Pi(\cdot,\mathcal{Q}) = \hat{Q} \end{cases}, \tag{5.11}$$

where $\hat{Q}$ is the distribution of training queries and $d_{\mathcal{Y}}$ is a metric on $\mathcal{Y} \triangleq \{h_\theta(\varphi(d^q)) \mid q \in \mathcal{Q}\}$.

Define a class of loss functions $\mathcal{D}$ by $\mathcal{D} \triangleq \{d_{h_\theta} : \mathcal{Q}\times\mathcal{Q} \to \mathbf{R}_+ \mid h_\theta \in \mathcal{H}\}$, where $d_h(q,q') \triangleq d_{\mathcal{Y}}(h(\varphi(d^q)), h(\varphi(d^{q'})))$ and $\mathcal{H}$ is the hypothesis class of scoring functions.

Let $N(\mathcal{D}, d, \epsilon)$ be the ϵ-covering of the class $\mathcal{D}$ with respect to a metric d. The entropy integral of $\mathcal{D}$ (w.r.t. the uniform metric) measures the complexity of the class and is defined by

$$J(\mathcal{D}) \triangleq \int_0^\infty \sqrt{\log N(\mathcal{D}, \|\cdot\|_\infty, \epsilon)}d\epsilon. \tag{5.12}$$

Assumption A1. Bounded diameters: $\sup_{x,x'\in\mathcal{X}} d_{\mathcal{X}}(x,x') \le D_{\mathcal{X}}$, $\sup_{y,y'\in\mathcal{Y}} d_{\mathcal{Y}}(y,y') \le D_{\mathcal{Y}}$.

Assumption A2. Estimation error of $d_{\mathcal{X}}$ is bounded: $\sup_{x,x'\in\mathcal{X}} |\hat{d}_{\mathcal{X}}(x,x') - d_{\mathcal{X}}(x,x')| \le \eta_d$.

Theorem 5.4.1. *If assumptions A1 and A2 hold and $J(\mathcal{D})$ is finite, then with probability at least $1-t$*

$$\sup_{h_\theta\in\mathcal{H}} |\hat{R}(h_\theta) - R(h_\theta)| \le \frac{48(J(\mathcal{D}) + \epsilon^{-1}D_{\mathcal{X}}D_{\mathcal{Y}})}{\sqrt{n}} + D_{\mathcal{Y}}\left(\frac{\log\frac{2}{t}}{2n}\right)^{\frac{1}{2}} + \frac{D_{\mathcal{Y}}\eta_d}{\epsilon},$$

where n is the number of training queries. A proof of the theorem is given in the supplement. The key technical challenge is leveraging the transport geometry on the query space to obtain a uniform bound on the convergence rate. This theorem implies that for a trained ranking model $\hat{h}_\theta$, the error term $|\hat{R}(\hat{h}_\theta) - R(\hat{h}_\theta)|$ is small for large n. Therefore, one can certify that the value of the regularizer $R(\hat{h}_\theta)$ is small on yet unseen (test) data by ensuring that the value of $\hat{R}(\hat{h}_\theta)$ is small on training data.

5.5 Computational results

In this section, we demonstrate the efficacy of SenSTIR for learning individually fair LTR models. One key conclusion is that enforcing individual fairness is sufficient to achieve group fairness but not vice versa. See Section B of the supplement for full details about the experiments.

Fair metric Following [YBS20], the individually fair metric $d_{\mathcal{X}}$ on $\mathcal{X}$ is defined in terms of a *sensitive subspace* A that is learned from data. In particular, $d_{\mathcal{X}}$ is the Euclidean distance of the data projected onto the orthogonal complement of A. This metric encodes variation due to sensitive information about individuals in the subspace and ignores it when computing the fair distance. For example, A can be formed by fitting linear classifiers to predict sensitive information, like gender or age, of individuals and taking the span of the vectors orthogonal to the corresponding decision boundaries. In each experiment, we explain how A is learned.

Baselines For all methods, we learn linear score functions h_θ and maximize normalized discounted cumulative gain (NDCG), i.e., Δ in Equation 5.2 is NDCG. We compare SenSTIR to (1) vanilla training without fairness ("Baseline"), i.e., $\rho = 0$, (2) preprocessing by first projecting the data onto the orthogonal complement of the sensitive subspace and then using vanilla training ("Project"), (3) "Fair-PG-Rank" [SJ19], a recent approach for training fair LTR models, and (4) randomly sampling the linear weights from a standard normal ("Random") to give context to NDCG.

5.5.1 Synthetic

We use synthetic data considered in prior fair ranking work [SJ19]. Each query contains 10 majority or minority items in $\mathbb{R}^2$ such that 8 items per query are majority group items in expectation. For each item, z_1 and z_2 are drawn uniformly from $[0, 3]$. The relevance of an item is $z_1 + z_2$ clipped between 0 and 5. A majority item's feature vector is $(z_1, z_2)^T$, whereas a minority item's feature vector is corrupted and given by $(z_1, 0)^T$.

Fair Metric The sensitive subspace is spanned by the hyperplane learned by logistic regression to predict whether an item is in the majority group. Recall, the fair metric is the Euclidean distance of the projection of the data onto the orthogonal complement of this subspace. Since this hyperplane is nearly equal to $(0, 1)^T$, the biased feature z_2 is ignored in the fair metric.

Results Figure 5.1 illustrates SenSTIR for $\rho \in \{0, .0003, .001\}$ with $\epsilon = .001$. Each point is colored by its relevance, and the contours show predicted scores where redder (respectively bluer) regions indicates higher (respectively lower) predicted scores. Minority items are on the horizontal z_1-axis because of their corrupted features. When $\rho = 0$, i.e., fairness is not enforced, this score function badly violates individual fairness since there are pairs of items close in the fair metric but with wildly different predicted scores because the biased feature z_2 is used. For example, the bottom blue star is a minority item with nearly the same relevance as the top black star majority item; however, the majority item's predicted score is much higher. When ρ is increased, the contours learned by SenSTIR eventually become vertical, thereby ignoring the biased feature z_2 and achieving individual fairness. When $\rho = .001$, the scores of the blue and black star are nearly equal because they are very close in the fair metric and the fair regularization strength is large enough.

Figure 5.2 illustrates another individual fairness property of SenSTIR that Fair-PG-Rank does not satisfy: ranking stability with respect to sensitive perturbations of the features. For each test query q, let $q' \neq q$ be the closest test query in terms of the fair distance $d_{\mathcal{Q}}$. We can view q' as a counterfactual query in the test set. For each query q, we sample 10 rankings corresponding to q and 10 counterfactual rankings corresponding to q' based on the learned ranking policy. The (i, j)-th entry of a heatmap in Figure 5.2 is the proportion of times the i-th ranked item for query q is ranked j-th in the counterfactual ranking. To satisfy individual fairness, the original and counterfactual rankings should be similar, meaning the heatmaps should be close to diagonal. Even though the baseline is relatively stable for highly and lowly ranked items, these items still change positions under the counterfactual rankings more than 50% of the time. Although Fair-PG-Rank satisfies group fairness, it is worse than the baseline in terms of counterfactual stability, i.e., individual fairness. In contrast, as ρ increases, SenSTIR

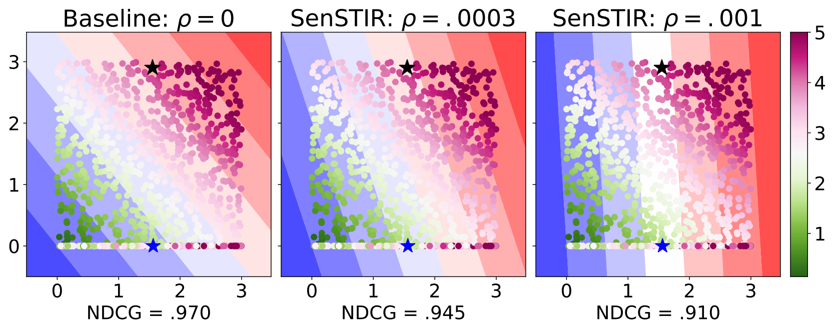


Figure 5.1: The points represent items shaded by their relevances, and the contours represent the predicted scores. The minority items lie on the horizontal z_1-axis because their z_2 value is corrupted to 0. The blue star and black star correspond to minority and majority items that are close in the fair metric with nearly the same relevance. However, they have wildly different predicted scores under the baseline. Using SenSTIR, as ρ increases, they eventually have the same predicted scores.

becomes stable.

5.5.2 German Credit

Following [SJ19], we adapt the German Credit classification data set [DG17b], which is susceptible to gender and age biases, to a LTR task. This data set contains 1000 individuals with binary labels indicating creditworthiness. Features include demographics like gender and age as well as information about savings accounts, housing, and employment.

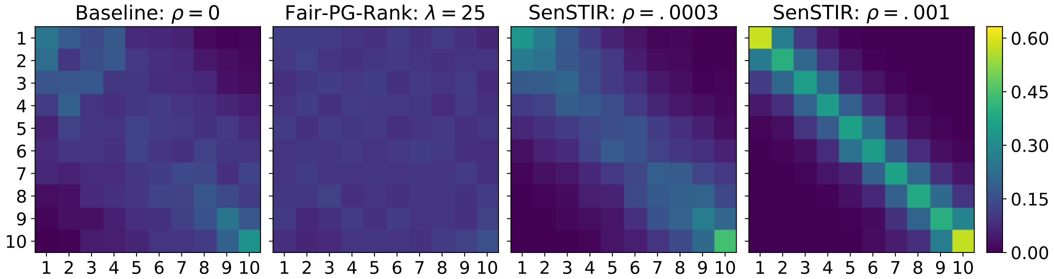


Figure 5.2: The (i, j)-th entries of these heatmaps represent the proportion of times that the i-th ranked item is moved to position j under the corresponding counterfactual ranking. With large enough ρ, SenSTIR ranks the original queries and counterfactual queries similarly as desired.

To simulate LTR data, individuals are sampled with replacement to build queries of size 10. Each individual has a binary relevance, and on average 4 individuals are relevant in each query. To apply Fair-PG-Rank, age is the binary protected attribute where the two groups are those younger than 25 and those 25 and older, a split proposed by [KC09]. For the fair metric, the sensitive subspace is spanned by the ridge regression coefficients for predicting age based on all other features and the standard basis vector corresponding to age.

Comparison metrics See Section B of the supplement for the precise definitions of these metrics. To assess accuracy, following [SJ19], we report the average stochastic test NDCG by sampling 25 rankings for each query from the learned ranking policy. To assess individual fairness, we use ranking stability with respect to demographic perturbations, which is the natural analogue of an evaluation metric for individual fairness in classification [YS20, YBS20]. In particular, for each query, we create a counterfactual query by flipping the (binary) gender of each individual in the query, and deterministically rank by sorting the items by their scores. We report the average Kendall's tau correlation (higher implies better individual fairness) between a test query's ranking and its counterfactual ranking. To assess group fairness and fairly compare to Fair-PG-Rank based on their fairness definition, we report the average stochastic disparity of group exposure also with 25 sampled rankings per query. This metric measures the asymmetric differences of the ratio of exposure a group receives to its relevance per query and favors the group with less relevance for a given query. Let G_1 (respectively G_0) be the set of older (respectively younger) people for a query q. For $i \in \{0,1\}$, let $M_{G_i} = (1/|G_i|)\sum_{d\in G_i} \mathrm{rel}^q(d)$. If $M_{G_0} > M_{G_1}$, let $G_A = G_0$, $G_D = G_1$ and $G_A = G_1$, $G_D = G_0$ otherwise. The stochastic disparity of group exposure for a set of rankings $\{r_i\}_{i=1}^N$ corresponding to a query is

$$\max\left\{0, \frac{\frac{1}{N|G_A|}\sum_{d\in G_A}\sum_{i=1}^N \frac{1}{\log_2(r_i(d)+1)}}{M_{G_A}} - \frac{\frac{1}{N|G_D|}\sum_{d\in G_D}\sum_{i=1}^N \frac{1}{\log_2(r_i(d)+1)}}{M_{G_D}}\right\}. \tag{5.13}$$

Results Figure 5.3 illustrates the fairness versus accuracy trade-off on the test set. The error bars represent the standard error over 10 random train/test splits. Both SenSTIR and Fair-PG-Rank enforce fairness through regularization, so we vary the regularization strength (ρ for SenSTIR with ϵ constant). Based on the NDCG of "Random", the

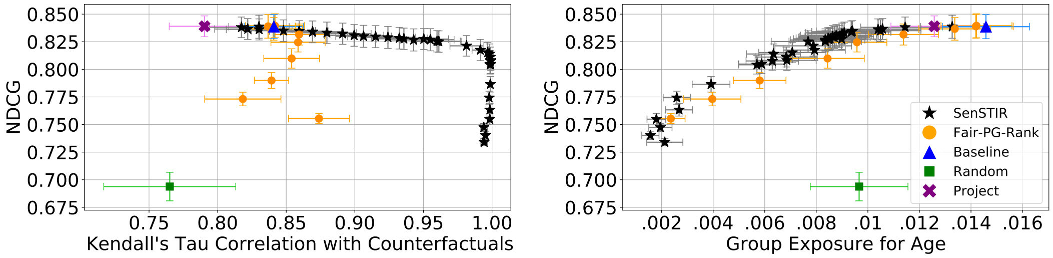


Figure 5.3: Individual (left) and group fairness (right) versus accuracy for the German credit data set

regularization strength ranges are reasonable for both methods. The left plot in Figure 5.3 shows the average Kendall's tau correlation (higher is better) between test queries and their gender counterfactuals versus the average stochastic NDCG. The maximum Kendall's tau correlation is 1, which SenSTIR achieves with relatively high NDCG. We emphasize that the sensitive subspace that SenSTIR utilizes to define the fair query metric directly relates to age, not gender. However, age is correlated with gender, so this metric shows the individually fair properties of SenSTIR generalize beyond age. Furthermore, SenSTIR gracefully trades off NDCG for individual fairness unlike Fair-PG-Rank. "Project" is worse in terms of individual fairness than vanilla training without enforcing fairness. Without direct age information, perhaps "Project" must more heavily rely on gender to learn accurate rankings, which illustrates that SenSTIR's generalization properties from age to gender are non-trivial. Disparity of group exposure (where smaller numbers are better) versus NDCG is depicted on the right plot of Figure 5.3. This group fairness metric is exactly what Fair-PG-Rank regularizes with. On average, for the same value of NDCG, SenSTIR typically outperforms Fair-PG-Rank showing that individual fairness can be sufficient for group fairness but not vice versa. While "Project" improves mildly upon the baseline, it shows being "age" blind does not result in group fair rankings.

5.5.3 Microsoft Learning To Rank

The demographic biases are real in the German Credit data, but the LTR task is simulated. There are no standard LTR data sets with demographic biases, so we consider Microsoft's Learning to Rank (MSLR) data set [QL13] with an artificial algorithmic fairness concern

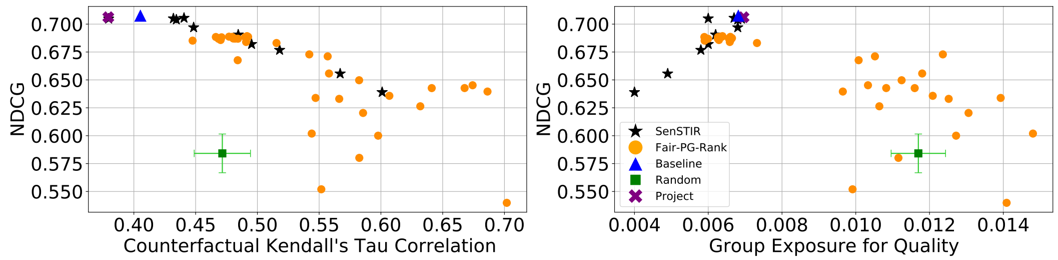


Figure 5.4: Individual (left) and group fairness (right) versus NDCG for the MSLR data set

dealing with webpage quality following [YDJ19]. The data set consists of query-web page pairs from a search engine with nearly 140 features with integral relevance scores. To apply Fair-PG-Rank, following [YDJ19], the protected binary attribute is whether a web page is high or low quality defined by the 40th percentile of quality scores (feature 133). For the fair metric, the sensitive subspace is spanned by the ridge regression coefficients for predicting the quality score (feature 133) based on all features and the standard basis vector corresponding to the quality score.

Comparison metrics Again we use average stochastic NDCG to measure accuracy, and the dispartiy of group exposure where the groups are high and low quality web pages. To assess individual fairness, we use the same set-up as in the German Credit experiments except the counterfactual for each test query q is the closest query $q' \neq q$ with respect to the fair metric over the train and test set.

Results Figure 5.4 shows the fairness and accuracy trade-off on the test set. Fair-PG-Rank becomes unstable with large fair regularization as it can drop below a random ranking in NDCG. The left plot shows the Kendall's tau correlation between test queries and their counterfactuals. SenSTIR gracefully trades-off NDCG with Kendall's tau correlation unlike Fair-PG-Rank. The right plot shows that SenSTIR also smoothly trades-off group fairness for NDCG. In contrast, as the regularization strength increases, both NDCG and group exposure worsen for Fair-PG-Rank, which was also observed by [YDJ19].

5.6 Conclusion

We proposed SenSTIR, an algorithm to learn provably individually fair LTR models with an optimal transport-based regularizer. This regularizer encourages the LTR model to produce similar ranking policies, i.e., distributions over rankings, for similar queries where similarity is defined by a fair metric. Our notion of a fair ranking is complementary to prior definitions that require allocating exposure to items fairly with respect to merit. In fact, we empirically showed that enforcing individual fairness can lead to allocating exposure fairly for groups but allocating exposure fairly for groups does not necessarily lead to individually fair LTR models.

5.7 Supplement

5.7.1 Proofs of Theoretical Results

Theorem 5.7.1 (Theorem 5.4.1)**.** *If assumptions A1 and A2 hold and $J(\mathcal{D})$ is finite, then with probability at least $1-t$*

$$\sup_{h\in\mathcal{H}} |\hat{R}(h) - R(h)| \leq \frac{48(J(\mathcal{D}) + \varepsilon^{-1} D_{\mathcal{X}} D_{\mathcal{Y}})}{\sqrt{n}} + D_{\mathcal{Y}} \left(\frac{\log\frac{2}{t}}{2n}\right)^{\frac{1}{2}} + \frac{D_{\mathcal{Y}}\eta_d}{\varepsilon}.$$

Proof. For queries q, q' let

$$\Delta(q,q') = \{\Pi \in \Delta(\mathcal{X}\times\mathcal{X}) : \Pi(\mathcal{X},\cdot) = \frac{1}{n}\sum_{j=1}^{n}\delta_{\varphi(d_j^q)},\ \Pi(\cdot,\mathcal{X}) = \frac{1}{n}\sum_{j=1}^{n}\delta_{\varphi(d_j^{q'})}\}.$$

Let $\Pi^* \in \arg\min_{\Pi\in\Delta(q,q')}\mathbb{E}_\Pi[d_{\mathcal{X}}(X,X')]$ and observe that by assumption A2 and the definition of d_Q and $\hat{d}_Q$ we have

$$\begin{aligned}
\hat{d}_{\mathcal{Q}}(q,q') - d_{\mathcal{Q}}(q,q') &= \inf_{\Pi\in\Delta(q,q')} \mathbb{E}_\Pi[\hat{d}_{\mathcal{X}}(X,X')] - \inf_{\Pi\in\Delta(q,q')} \mathbb{E}_\Pi[d_{\mathcal{X}}(X,X')] \\
&= \inf_{\Pi\in\Delta(q,q')} \mathbb{E}_\Pi[\hat{d}_{\mathcal{X}}(X,X')] - \mathbb{E}_{\Pi^*}[d_{\mathcal{X}}(X,X')] \\
&\leq \mathbb{E}_{\Pi^*}[\hat{d}_{\mathcal{X}}(X,X')] - \mathbb{E}_{\Pi^*}[d_{\mathcal{X}}(X,X')] \\
&= \mathbb{E}_{\Pi^*}[\hat{d}_{\mathcal{X}}(X,X') - d_{\mathcal{X}}(X,X')]
\end{aligned}$$

$$\leq \eta_d.$$

Similarly,

$$d_Q(q,q') - \hat{d}_Q(q,q') \leq \mathbb{E}_{\hat{\Pi}^*}[d_{\mathcal{X}}(X,X') - \hat{d}_{\mathcal{X}}(X,X')] \leq \eta_d.$$

It follows that

$$|\hat{d}_Q(q,q') - d_Q(q,q')| \leq \eta_d. \tag{5.14}$$

Next, we will bound the difference $|\hat{R}(h) - R(h)|$. To lighten the notation, we write h, h' for $h = h(\phi(d^q)), h' = h(\phi(d^{q'}))$. From the dual representation of $R(h)$ and $\hat{R}(h)$ we have

$$\begin{aligned}
\hat{R}(h) - R(h) &= \inf_{\lambda \geq 0}\{\lambda\epsilon + \mathbb{E}_{q\sim\hat{Q}}[\hat{r}_\lambda(h,q)]\} - \inf_{\lambda \geq 0}\{\lambda\epsilon + \mathbb{E}_{q\sim Q}[r_\lambda(h,q)]\} &&(5.15)\\
&= \inf_{\lambda \geq 0}\{\lambda\epsilon + \mathbb{E}_{q\sim\hat{Q}}[\hat{r}_\lambda(h,q)]\} - \lambda^*\epsilon - \mathbb{E}_{q\sim\hat{Q}}[\hat{r}_{\lambda^*}(h,q)] &&(5.16)\\
&\leq \mathbb{E}_{q\sim\hat{Q}}[\hat{r}_{\lambda^*}(h,q)] - \mathbb{E}_{q\sim Q}[r_{\lambda^*}(h,q)] &&(5.17)\\
&= \mathbb{E}_{q\sim\hat{Q}}[r_{\lambda^*}(h,q)] - \mathbb{E}_{q\sim Q}[r_{\lambda^*}(h,q)] + \mathbb{E}_{q\sim\hat{Q}}[\hat{r}_{\lambda^*}(h,q) - r_{\lambda^*}(h,q)]. &&(5.18)
\end{aligned}$$

To bound the last term, note that

$$\begin{aligned}
|\hat{r}_{\lambda^*}(h,q) - r_{\lambda^*}(h,q)| &= \sup_{q'}\{\hat{d}_{\mathcal{Y}}(h,h') - \lambda^* d_{\mathcal{Q}}(q,q')\} - \sup_{q'}\{\hat{d}_{\mathcal{Y}}(h,h') - \lambda^* d_{\mathcal{Q}}(q,q')\} &&(5.19)\\
&\leq \lambda^* \sup_{q'}\{|d_{\mathcal{Q}}(q,q') - \hat{d}_{\mathcal{Q}}(q,q')| &&(5.20)\\
&\leq \lambda^*\eta_d. &&(5.21)
\end{aligned}$$

Combining (5.21) and (5.18) yields

$$\hat{R}(h) - R(h) \leq \mathbb{E}_{q\sim\hat{Q}}[r_{\lambda^*}(h,q)] - \mathbb{E}_{q\sim Q}[r_{\lambda^*}(h,q)] + \lambda^*\eta_d. \tag{5.22}$$

Using a similar argument,

$$R(h) - \hat{R}(h) \leq \mathbb{E}_{q\sim Q}[r_{\hat{\lambda}^*}(h,q)] - \mathbb{E}_{q\sim \hat{Q}}[r_{\hat{\lambda}^*}(h,q)] + \hat{\lambda}^* \eta_d. \tag{5.23}$$

To find an upper bound on λ^*, observe that $r_\lambda(h,q) \geq 0$ for all $h \in \mathcal{H}, \lambda \geq 0$, as

$$\begin{aligned} r_\lambda(h,q) &= \sup_{q' \in \mathcal{X}} \{ d_{\mathcal{Y}}(h,h') - \lambda d_{\mathcal{Q}}(q,q') \} \\ &\geq d_{\mathcal{Y}}(h,h) - \lambda d_{\mathcal{Q}}(q,q) = 0. \end{aligned}$$

Thus

$$\lambda^* \varepsilon \leq \lambda^\star \varepsilon + \mathbb{E}_{q\sim \mathcal{Q}}[r_\lambda(h,q)] = R(h) \leq D_{\mathcal{Y}}.$$

Rearranging the above yields $\lambda^* \leq \frac{D_{\mathcal{Y}}}{\varepsilon}$ and the same upper bound is also valid for $\hat{\lambda}^\star$ by the same argument.

Combining inequalities (5.22,5.23) and the bound on $\lambda^*, \hat{\lambda}^*$, we can write

$$|\hat{R}(h) - R(h)| \leq \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{q\sim \hat{Q}} f(q) - \mathbb{E}_{q\sim Q} f(q) \right| + \frac{D_{\mathcal{Y}} \eta_d}{\varepsilon},$$

where $\mathcal{F} = \{ r_\lambda(h,\cdot) : \lambda \in [0,L],\, h \in \mathcal{H} \}$. A standard concentration argument proves

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E}_{q\sim \hat{Q}} f(q) - \mathbb{E}_{q\sim Q} f(q) \right| \leq \frac{48(J(\mathcal{D}) + \varepsilon^{-1} D_{\mathcal{X}} D_{\mathcal{Y}}}{\sqrt{n}} + D_{\mathcal{Y}}(\frac{\log \frac{2}{t}}{2n})^{\frac{1}{2}}$$

with probability at least $1-t$. This completes the proof of the theorem. □

The main technical novelty in this proof is the bound on λ_* in terms of the diameter of the output space. This restricts the set of possible c-transformed loss function class, thereby allowing us to appeal to standard techniques from empirical process theory to obtain uniform convergence results. Prior work in this area (e.g. [LR18]) relies on smoothness properties of the loss instead of the geometric properties of the output space, but this precludes non-smooth output metrics.

5.7.2 Experiments

All experiments were ran a cluster of CPUS. We do not require a GPU.

Data sets and pre-processing

Synthetic Synthetic data is generated as described in the main text such that there are 100 queries in the training set and 100 queries in the test set.

German Credit The German Credit data set [DG17b] consists of 1000 individuals with binary labels indicating if they are credit worthy or not. We use the version of the German Credit data set that [SJ19] used found at `https://www.kaggle.com/uciml/german-credit`. In particular, this version of the Geramn Credit data set only uses the following features: `age` (integer), `sex` (binary, does not include any marital status information unlike the original data set), `job` (categorical), `housing` (categorical), `savings account` (categorical), `checking account` (integer), `credit amount` (integer), `duration` (integer), and `purpose` (categorical). See [DG17b] for an explanation of each feature.

Categorical features are the only features with missing data, so we treat missing data as its own category. The following features are standardized by subtracting the mean and dividing by the standard deviation (before this data is turned into LTR data): `age`, `duration`, and `credit amount`. The remaining binary and categorical features are one hot encoded.

We use an 80/20 train/test split of the original 1000 data points, and then sample from the training/testing set with replacement to build the LTR data as discussed in the main text. For our experiments, we use 10 random train/test splits.

Microsoft Learning to Rank The Microsoft Learning to Rank data set [QL13] consists of query-web page pairs each of which has 136 features and integral relevance scores in $[0, 4]$. We use Fold 1's train/validation/test split. Following [YDJ19], we use the data in Fold 1 and adopt the given train/validation/test split. The data and feature descriptions can be found at `https://www.microsoft.com/en-us/research/project/mslr/`. We remove the `QualityScore` feature (feature 132) since we use the `QualityScore2` (feature 133) feature to learn the fair metric, and it appears based on the description of these features, they are very similar. We standardize the remaining features (except for the features corresponding to `Boolean model`, i.e. features 96-100, which are binary) by subtracting the mean and dividing by the standard deviation. Following [YDJ19], we remove any queries with less than 20 web pages. Furthermore, we only consider queries that have at least one web page with a relevance of 4. For each query, we sample 20 web pages without replacement until at least one of the 20 sampled web pages has a relevance of 4. After pre-processing, there are 33,060 train queries, 11,600 validation queries, and 11,200 test queries.

Comparison Metrics

Let r be a ranking (i.e. permutation) of a set of n items that are enumerated such that $r(i) \in [n]$ is the position of the i-th item in the ranking and $r^{-1}(i) \in [n]$ is the item that is ranked i-th. Let $\mathrm{rel}_q(i)$ be the relevance of item i given a query q.

Normalized Discounted Cumulative Gain (NDCG) Let S_n be the set of all rankings on n items. The discounted cumulative gain (DCG) of a ranking r is

$$\mathrm{DCG}(r) = \sum_{i=1}^{n} \frac{2^{\mathrm{rel}_q(r^{-1}(i))} - 1}{\log_2(i+1)}.$$

The NDCG of a ranking r is

$$\frac{\mathrm{DCG}(r)}{\max_{r' \in S_n} \mathrm{DCG}(r')}.$$

Because we learn a distribution over rankings and the number of rankings is too large, we cannot compute the expected value of the NDCG for a given query. Thus, for each query in the test set, we sample N rankings (where $N = 10$ for synthetic data, $N = 25$ for German credit data, and $N = 32$ for Microsoft Learning to Rank data) from the Placket-Luce distribution, compute the NDCG for each of these rankings, and then take an average. We refer to this quantity as the *stochastic NDCG*.

Kendall's tau correlation Let r and r' be two rankings on n items. Then

$$\mathrm{KT}(r, r') := \frac{1}{\binom{n}{2}} \sum_{\{i<j : i,j \in [n]\}} \mathrm{sign}(r(i) - r(j))\mathrm{sign}(r'(i) - r'(j))$$

is the Kendall's tau correlation between two rankings.

(Disparity of) Group exposure This definition was first proposed by [SJ19]. Assume each item belongs to one of two groups. Let G_1 (respectively G_0) be the set of items for a query q that belongs to group 1 (respectively group 0). For $i \in \{0,1\}$, let $M_{G_i} = \frac{1}{|G_i|}\sum_{d\in G_i} \mathrm{rel}_q(d)$, which is referred to as the merit of group i for query q. For a ranking r and for $i \in \{0,1\}$, let $v_r(G_i) = \frac{1}{|G_i|}\sum_{d\in G_i} \frac{1}{\log_2(r(d)+1)}$. Because we learn a distribution over rankings and the number of rankings is too large, we cannot compute the expected value of $v_r(G_i)$ over this distribution. Instead, we sample N rankings (where again $N = 10$ for synthetic data, $N = 25$ for German credit data, and $N = 32$ for Microsoft Learning to Rank data) from the Placket-Luce model. Let R_q be the set of these N sampled rankings for query q. Then the stochastic disparity of group exposure for query q is

$$\begin{cases} \max\left\{0, \frac{\frac{1}{N}\sum_{r\in R_q} v_r(G_0)}{M_{G_0}} - \frac{\frac{1}{N}\sum_{r\in R_q} v_r(G_1)}{M_{G_1}}\right\} & \text{if } M_{G_0} \geq M_{G_1} > 0 \\ \max\left\{0, \frac{\frac{1}{N}\sum_{r\in R_q} v_r(G_1)}{M_{G_1}} - \frac{\frac{1}{N}\sum_{r\in R_q} v_r(G_0)}{M_{G_0}}\right\} & \text{if } 0 < M_{G_0} < M_{G_1} \\ 0 & \text{if } M_{G_0} = 0 \text{ or } M_{G_1} = 0. \end{cases}$$

In the language of [SJ19], we use the identity function for merit, and set the position bias at position j to be $\frac{1}{\log_2(1+j)}$ just as they did.

SenSTIR implementation details

We implement SenSTIR in TensorFlow and use the Python `POT` package to compute the fair distance between queries and to compute Equation (5.9), which requires solving optimal transport problems. Throughout this section, variable names from our code are italicized, and the abbreviation we use to refer to these variables/hyperparameters are followed in parenthesis.

Fair regularizer optimization Recall that in all of the experiments, the fair metric $d_{\mathcal{X}}$ on items is the Euclidean distance of the data projected onto the orthogonal complement of a subspace. In order to optimize for the fair regularizer in Equation (SenSTIR), first we optimize over this subspace, and we refer to this step as the *subspace attack*. Note, the distance between the original queries and the resulting adversarial queries in the subspace is 0. Second, we use the resulting adversarial queries in the subspace as an initialization to the *full attack*, i.e. we find adversarial queries that have a non-zero fair distance to the original queries. We implement both using the Adam optimizer [KB15].

Learning rates As mentioned above, we use the Adam optimizer to optimize the fair regularizer. For the subspace attack, we set the learning rate to *adv_step*(as) and train for *adv_epoch*(ae) epochs, and for the full attack, we set the learning rate to *l2_attack*(fs) and train for *adv_epoch_full*(fe) epochs. We also use the Adam optimizer with a learning rate of .001 to learn the parameters of the score function h_θ.

Fair start Our code allows training the baseline (i.e. when $\rho = 0$) for a percentage–given by *fair_start*(frs)–of the total number of epochs before the optimization includes the fair regularizer.

Using baseline for variance reduction Following [SJ19], in the gradient estimate of the empirical version of $\mathbb{E}_{q\sim Q}\big[U(\pi \mid q)\big]$ in Equation (SenSTIR), we subtract off a baseline term $b(q)$ for each query q, where $b(q)$ is the average utility $U(\pi \mid q)$ over the Monte Carlo samples for the query q. This counteracts the high variance in the gradient estimate [Wil92].

Other hyperparameters In Tables 5.1 and 5.2, E stands for the total number of epochs used to update the score function h_θ, B stands for the batch size, $l2$ stands for the ℓ_2 regularization strength of the weights, and MC stands for the number of Monte Carlo samples used to estimate the gradient of the empirical version of $\mathbb{E}_{q\sim Q}\big[U(\pi \mid q)\big]$ in Equation (SenSTIR) for each query.

Hyperparameters

For the synthetic data, we use one train/test split. For the German experiments, we use 10 random train/test splits all of which use the same hyperparameters. For the Microsoft experiments, we pick hyperparameters on the validation set (where the range of hyperparameters considered are reported below) based on the trade-off of stochastic

NDCG and individual (respectively group) fairness for SenSTIR (respectively Fair-PG-Rank), and report the comparison metrics on the test set.

Fair metric For the synthetic data experiments, we use `sklearn`'s logistic regression solver to classify majority and minority individuals with 1/100 ℓ_2 regularization strength. For German and Microsoft, we use `sklearn`'s `RidgeCV` solver with the default hyperparameters to predict age and quality web page score, respectively. For the German experiments, when predicting age, each individual is represented in the training data exactly once, regardless of the number of queries that an individual appears in.

SenSTIR For every experiment, all weights are initialized by picking numbers in $[-.0001, .0001]$ uniformly at random, λ in Algorithm 4 is always initialized with 2, and the learning rate for Adam for the score function h_θ is always .001. For synthetic data, the fair regularization strength ρ varied in $\{.0003, .001\}$. For German, ρ is varied in $\{.001, .01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.06, 0.07, 0.08, 0.09, .1, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.28, 0.37, 0.46, 0.55, 0.64, 0.73, 0.82, 0.91, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 50, 100\}$. For Microsoft, ρ is varied in $\{.00001, .0001, .001, .01, .04, .07, .1, .33, .66, 1.\}$. We report results for all choices of ρ.

See Table 5.1 for the remaining values of hyperparameters where the column names have been defined in the previous section except for ϵ, which refers to ϵ in the definition of the fair regularizer. For Microsoft, the best performing hyperparameters on the validation set are reported where the ℓ_2 regularization parameter for the weights are varied in $\{.001, .0001, 0\}$, as is varied in $\{.01, .001\}$, ae and fe are varied in $\{20, 40\}$, and ϵ is varied in $\{1, .1, .01\}$.

Table 5.1: SenSTIR hyperparameter choices

	E	B	as	ae	ϵ	fs	fe	frs	$l2$	MC
Synthetic	2K	1	0.001	20	0.001	0.001	20	0	0	10
German	20K	10	.01	20	1	0.001	20	.1	0	25
Microsoft	68K	10	.01	40	.01	0.001	40	.1	0.001	32

Baseline and Project For the baseline (i.e. $\rho = 0$ with no fair regularization) and project baseline, we use the same number of epochs, batch sizes, Monte Carlo samples, and ℓ_2 regularization as in Table 5.1 for SenSTIR. Furthermore, we use the same weight initialization and learning rate for Adam as in the SenSTIR experiments.

Fair-PG-Rank We use the implementation found at `https://github.com/ashudeep/Fair-PGRank` for the synthetic and German experiments, whereas we use our own implementation for the Microsoft experiments because we could not get their code to run on this data. They use Adam for optimization, and the learning rate is .1 for the synthetic data and .001 for German and Microsoft. Let λ refer to the Fair-PG-Rank fair regularization strength. For synthetic, $\lambda = 25$. For German, λ is varied in $\{.1, 1, 1.5, 2, 2.5, 3, 3.5, 4\}$. For Microsoft, λ is varied in $\{.001, .01, .1, .5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 50, 100, 500, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950, 1000\}$. We report results for all choices of λ. See Table 5.2 which summarizes the remaining hyperparameter choices.

Table 5.2: Fair-PG-Rank hyperparameter choices

	E	B	$l2$	MC
Synthetic	5	1	0	10
German	100	1	0	25
Microsoft	68K	10	.01	32

Chapter 6

Conclusion and Future Work

In this thesis, we addressed three key challenges of preference learning: intransitivity, non-convexity, and algorithmic bias. We proposed new models and algorithms with empirical validation and theoretical guarantees using tools from statistical learning theory and optimization. We now conclude with a brief summary of each chapter and propose future directions.

6.1 Intransitivity

Although many preference models assume preferences are transitive, real preference data often exhibit systematic intransitivity. Chapter 2 proposed the *salient feature preference model*, which reconciles intransitive pairwise comparisons with a global ranking. Inspired by social science theories on intransitivity, we assumed pairwise contextual effects prevent the global ranking from being perfectly reflected in the pairwise comparison data. These pairwise contextual effects arise since the two items in each pairwise comparison are compared in isolation of the rest of the items, and the salient feature preference model accounts for these contextual effects. We proposed to learn the parameters of our model from pairwise comparison data via maximum likelihood estimation and analyzed its sample complexity. Furthermore, we demonstrated strong performance of our model on real preference data that contain intransitivity.

There are two main avenues to future work. First, our general framework about contextual effects can be applied to other machine learning problems that use human judgement data. The salient feature preference model for pairwise comparisons assumes humans make preference judgements about a small number of items based only on a small subset of salient features that "stand out," not on all the features. This framework can be applied to the ordinal embedding problem considered in Chapter 3 where respondents are asked "Is item A more similar to item B or item C?". In addition, as discussed in Chapters 4 and 5, we often need to learn the individually fair metric from data in order to train fair models. There have been algorithms recently proposed either to learn the metric from pairwise comparison data [MYBS20a] or to learn an individually fair classifier from pairwise comparison data [JKN$^+$19]. The work in [JKN$^+$19] considers training a fair recidivism model, so the type of data collected is of the form "Should these two people be given the same recidivism score?". It is reasonable to believe that contextual effects–due to which two people are being compared and their demographics–affect human judgements.

Second, we assumed the selection function τ is known, so one future direction is to learn τ while simultaneously learning the other model parameters. Recall that given a pair of items i and j, $\tau(i,j)$ determines which features the probabilistic outcome of a comparison between i and j depends on. However, it can be unreasonable to assume that τ is known, and poor choices of τ can lead to poor performance of our model. For instance, in the experimental section of Chapter 2, we proposed the top-t selection function, which returns the t coordinates with the largest magnitude difference. An adversary could include a meaningless feature that has high variance so that the top-t selection function always picks this meaningless feature.

Assuming τ selects a k-sparse subset of coordinates for every pair of items, we now describe one set-up to facilitate learning τ. As in Chapter 2, suppose there are n items, $U_i \in \mathbb{R}^d$ is the known feature vector of item $i \in [n]$, and $w^* \in \mathbb{R}^d$ is the unknown judgement vector that we wish to estimate. Let $\tau : [n] \times [n] \to \{s \subseteq [d] : |s| = k\}$ be the unknown selection function that we also wish to estimate. Assume that $[n] \times [n]$ is partitioned into N enumerated sets such that $(i,j), (o,p) \in [n] \times [n]$ are in the same set of the partition if and only if $\tau(i,j) = \tau(o,p)$. We comment on why we need this partition and how to obtain such a partition at the end of this subsection. Let $W^* \in \mathbb{R}^{d \times N}$ where the ℓ-th column of W^*, denoted W^*_ℓ, corresponds to the ℓ-th set in the enumeration of the partition. Let (i_ℓ, j_ℓ) be in the ℓ-th set of the enumeration. Define $W^*_\ell := w^*_{\tau(i_\ell, j_\ell)}$ where the q-th coordinate of $w^*_{\tau(i_\ell, j_\ell)}$ is the q-th coordinate of w^* if $q \in \tau(i_\ell, j_\ell)$. See Figure 6.1 for an illustration of the relationship between w^* and W^*.

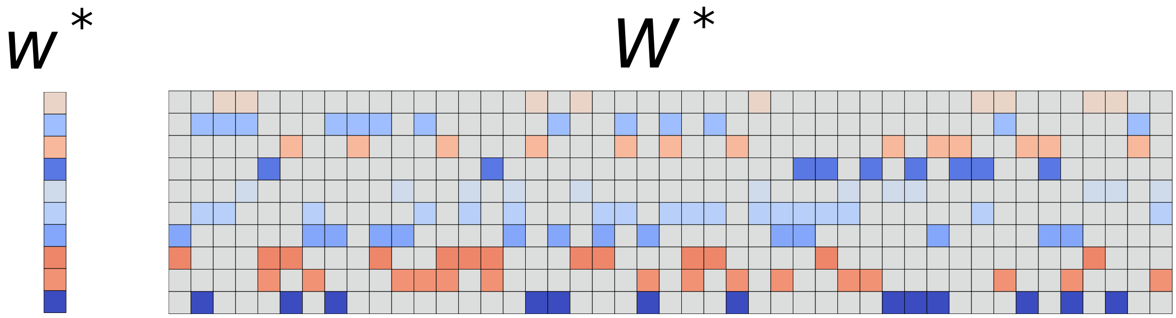


Figure 6.1: On the left, $w^* \in \mathbb{R}^{20}$ is depicted. On the right $W^* \in \mathbb{R}^{20\times 50}$ is depicted where the grey entries indicate 0 values. Each column of W^* is at most 3-sparse. Along each row of W^*, the non-zero entries are all equal and match the corresponding coordinate in w^*.

Suppose we observe m independent pairwise comparisons $S_m = \{(i_r, j_r, \ell_r, y_r) : i_r, j_r \in [n], \ell_r \in [N], y_r \in \{0,1\}\}_{r=1}^m$ where (i_r, j_r) is in the ℓ_r-th set in the enumeration, and y_r indicates a human judgement between items i_r and j_r. We assume that $y_r \sim \text{Bern}(P_{(i_r,j_r)})$ where

$$\begin{aligned} P_{(i_r,j_r)} &= \mathbb{P}(y_r = 1) && (6.1)\\ &= \mathbb{P}(\text{item } i_r \text{ is better than item } j_r) && (6.2)\\ &= \frac{1}{1+\exp\left(\langle W^*_{\ell_r}, U_{i_r} - U_{j_r}\rangle\right)}. && (6.3) \end{aligned}$$

Determining the sample complexity of estimating W^* and ultimately w^* is of great interest. Towards this goal, for $W \in \mathbb{R}^{d\times N}$, let

$$f(W) = \frac{1}{m}\sum_{r=1}^{m} \log\left(1+\exp\left(\left\langle W_{\tau(i_r,j_r)}, U_{i_r} - U_{j_r}\right\rangle\right)\right) - y_r \left\langle W_{\tau(i_r,j_r)}, U_{i_r} - U_{j_r}\right\rangle \qquad (6.4)$$

be the log-loss. We can estimate W^* by solving

$$\hat{W} = \text{argmin}_{W\in\mathbb{R}^{d\times N}} f(W) + \Omega(W) \qquad (6.5)$$

where $\Omega(W)$ is a regularizer on W. If $\Omega(W) = \sum_{i=1}^N \|W_i\|_1$, then Equation (6.5) is ℓ_1-regularized logistic regression. In this case, the objective of Equation (6.5) is separable into N separate problems over the columns of W. If each column of W^*

is k-sparse, we expect the sample complexity of estimating W^* with Equation (6.5) to be $O(Nk\log(d))$ [PV12]. However, we have additional information: the non-zero entries of each row of W are equal. Can we use this information to come up with a regularizer–perhaps one that encourages sparsity and clustering–to improve the sample complexity of estimating W^*? How do we estimate w^* from the estimate of W^*? Even determining a meaningful lower bound on the sample complexity of estimating W^* and w^* outside of the regularization framework is of interest.

We now return to why we require a partition of the pairs of items and one idea to obtain such a partition. Ideally, we would like a model that not only allows us to estimate a ranking of the items but also allows us to predict the probability that item i beats item j for an unseen pair of items i and j. If there is no additional structure on τ, we cannot predict the probability that item i beats item j for an unseen pair of items i and j since we do not know what $\tau(i, j)$, i.e., what features are relevant in the pairwise comparison. The partition of the items is one way to add structure to τ.

In order to obtain the partition from data, one potential solution is to use a clustering algorithm on $\{|U_i - U_j| : i, j \in [n] \times [n]\}$ since the probability that item i beats item j depends on $U_i - U_j$ as seen in Equation (6.1). For two pairs of items (i, j) and (o, p), it is reasonable to believe that if $|U_i - U_j|$ is close to $|U_o - U_p|$, then the features that "stand out" the most are the same in both pairwise comparisons. For an unseen pair of items (i, j), we can predict the probability that item i beats item j in a two-step process. First we determine what cluster $|U_i - U_j|$ belongs to. Second, given any pair of items in the training data (o, p) such that $|U_o - U_p|$ is in the same cluster as $|U_i - U_j|$, we estimate $\tau(i, j)$ with the sparsity pattern of the estimate of $W^*_{\tau(o,p)}$.

6.2 Non-convexity

Estimating the parameters of some preference models, like ordinal embedding and matrix factorization models for collaborative filtering, requires optimizing a non-convex function. In Chapter 3, we considered a particular class of non-convex homogeneous quadratic feasibility problems that encompasses the aforementioned preference models. Each feasibility problem entails finding a point that satisfies a set of quadratic inequalities. We proposed to find a feasible point by minimizing a non-convex function that penalizes a point each time it violates a quadratic inequality with the hinge loss. We empirically demonstrated that with proper initialization stochastic subgradient descent reliably finds feasible points despite the non-convex nature of the problem, which suggests that every local minimizer is a global minimizer. Motivated by this empirical finding, we theoretically studied the optimization landscape, i.e., local and global minimizers, of the non-convex optimization problem and paid special attention to the two-dimensional case.

There are two main directions of future work. First, generalizing our theoretical results from two dimensions to higher dimensions is of great interest. Our experiments on synthetic data suggest that the optimization landscape in higher dimensions is well-conditioned since stochastic subgradient descent finds feasible points. Even if it may not be the case that every minimizer is a global minimizer in higher dimensions, we saw in particular that stochastic subgradient descent succeeds in finding feasible points *when the initialization points have large norm.* Therefore, investigating the interplay of stochastic subgradient descent and the geometry of the optimization landscape outside an origin-centered ball with sufficiently large radius is of interest. Towards this goal, Lemma 3.2.2 can be generalized to arbitrary dimensions so long as A and B both have full rank. Suppose the associated matrices $\{P_i \in \mathbb{R}^{n \times n}\}$ to the non-convex feasibility problem and their sums have full rank. This property holds for random matrices drawn from a continuous distribution like in our experiments. By the generalization of Lemma 3.2.2 and Theorem 3.2.1, any non-global, local minimizer $x \in \mathbb{R}^n$ must be a intersection point of two or more curves of the form $x^T P_i x = 1$. Hence, there are only finitely many non-global, local minimizers, so the norms of non-global, local minimizers are bounded. Thus, if stochastic subgradient descent can avoid a sufficiently large ball around the origin, it will avoid all non-global, local minimizers and should succeed in finding feasible points.

Second, generalizing the problem setting to incorporate noise is important. Our problem setting assumed that there is always a feasible point to a set of quadratic equations, which we exploited throughout our proofs. In the preference modeling setting that motivated our work, this assumption means that people's preferences are observed perfectly, which is unrealistic. In order to pose the problem with noise, we could assume that there is a link function that relates the geometry of the space to noisy human judgements like in [JJN16]. In addition to the hinge loss, the logistic loss may be interesting to consider in the noisy setting.

6.3 Algorithmic Bias

It is well-established that machine learning algorithms can perpetuate or exacerbate historical and societal biases. In Chapter 4 we considered the classification setting and proposed *SenSR*, an algorithm that learns individually fair machine learning models by enforcing model invariance with respect to feature perturbations defined by a fair metric. For instance, suppose the model selects applicants for a job interview. Given two applicants that only differ in gender, a model trained with SenSR will either select both applicants for an interview or neither applicant. Furthermore, we also proposed an algorithm to learn a fair metric from data, thereby operationalizing individual fairness. This algorithm estimates a *sensitive subspace* of the feature space, and the fair metric is defined to be the Euclidean distance in the orthogonal complement of this subspace. The sensitive subspace corresponds to a region of the feature space that decisions should not be based on, like gender and features correlated with gender.

By using ideas from Chapter 4, in Chapter 5 we proposed *SenSTIR*, an algorithm that learns individually fair learning-to-rank (LTR) systems. We required individually fair LTR systems to be stable with respect to sensitive perturbations of the features. For example, an individually fair LTR system should identically rank a set of people and a counterfactual set of people that is obtained from the original set by flipping each person's gender. In both chapters, we studied the statistical properties of our algorithms and empirically demonstrated the ability of our algorithms to mitigate biases on real-world data sets.

One main direction of future work is to identify sufficient conditions under which a LTR system that satisfies our definition of individual fairness necessarily must satisfy the group exposure notions of fairness [SJ19, ZC20, BGW18]. In fact, we saw empirical evidence that a LTR system that satisfies our notion of individual fairness necessarily allocated exposure fairly to groups in both the German credit and Microsoft LTR data sets. However, we saw the converse is not necessarily true: group fair LTR systems are not necessarily individually fair. See Figures 5.3 and 5.4. Therefore, a better theoretical understanding of the interplay of individual fairness and group fairness for LTR systems could suggest that it is typically preferable to enforce individual fairness over group fairness since an individually fair LTR system may inherit all the properties of a group fair system but not vice versa.

Another direction of future work is to apply SenSTIR or SenSR with different fair metrics. Both the fairness performance and accuracy performance of SenSR and SenSTIR heavily depend on the fair metric. Although utilizing the fair metrics learned in Chapters 4 and 5 with SenSTIR or SenSR results in an arguably fair model, using these fair metrics might unnecessarily lower the accuracy of the model in comparison to using a fair metric that takes into account the causal nature of bias. For example, consider the German credit data experiments in Chapter 5. We are given demographic features like age and other features like credit history, and the goal is to rank a query of credit applicants from most to least creditworthy. The fair metric used in Chapter 5 tries to ignore variation in the data due to someone's age by ignoring the subspace spanned by the learned Ridge regression weights to predict someone's age. Arguably, credit history is a decent indication of whether someone is worthy of credit, but older people tend to have longer credit histories. Therefore, the fair metric will try to ignore credit length because it is correlated with age, but we may be losing too much information about credit history–which helps us learn accurate rankings–with this metric. In contrast, it might be worthwhile to use a fair metric based on the methods in [BNSV18] to account for the causal nature of bias in the fair metric, e.g. this metric may be able to retain more information about credit length while still maintaining fairness. Furthermore, as we have discussed previously, there are relatively new data sets of human judgements in the recidivism prediction domain that have been or can be used to learn a fair metric from data using standard metric learning techniques [WGL+19, JKN+19].

Bibliography

[AC$^+$17] Ery Arias-Castro et al. Some theory for ordinal embedding. *Bernoulli*, 23(3):1663–1693, 2017.

[Agr12] Alan Agresti. *Categorical data analysis.* John Wiley & Sons, 2012.

[AJSD19] Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das. Designing fair ranking schemes. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, page 1259–1276, New York, NY, USA, 2019. Association for Computing Machinery.

[ALMK16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias, May 23 2016. Name - ProPublica; Copyright - Copyright ProPublica May 23, 2016; Last updated - 2017-11-23.

[ASBP13] Ehsan Abbasnejad, Scott Sanner, Edwin V Bonilla, and Pascal Poupart. Learning community-based preferences via dirichlet process mixtures of gaussian processes. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[AWC$^+$07] Sameer Agarwal, Josh Wills, Lawrence Cayton, Gert Lanckriet, David Kriegman, and Serge Belongie. Generalized non-metric multidimensional scaling. In *Artificial Intelligence and Statistics*, pages 11–18, 2007.

[BCZ$^+$16] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA, 2016. Curran Associates Inc.

[BdDW$^+$12] Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust Solutions of Optimization Problems Affected by Uncertain Probabilities. *Management Science*, 59(2):341–357, November 2012.

[BDH$^+$18] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018.

[BE06] Amir Beck and Yonina C Eldar. Strong duality in nonconvex quadratic optimization with two quadratic constraints. *SIAM Journal on optimization*, 17(3):844–860, 2006.

[BGS13] Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Salience and consumer choice. *Journal of Political Economy*, 121(5):803–843, 2013.

[BGW18] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of Attention: Amortizing Individual Fairness in Rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 405–414, Ann Arbor, MI, USA, June 2018. Association for Computing Machinery.

[BHJ$^+$18] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 2018.

[BKM19] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[BKT16] Austin R Benson, Ravi Kumar, and Andrew Tomkins. On the relevance of irrelevant alternatives. In *Proceedings of the 25th International Confer-*

ence on World Wide Web, pages 963–973. International World Wide Web Conferences Steering Committee, 2016.

[BM04] Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4):991–1013, September 2004.

[BM19] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[BNSV18] Amanda Bower, Laura Niss, Yuekai Sun, and Alexander Vargo. Debiasing representations by removing unwanted variation due to protected attributes. *FAT-ML workshop at ICML*, July 2018.

[BP09] Thomas C Brown and George L Peterson. An enquiry into the method of paired comparison: reliability, scaling, and thurstone's law of comparative judgment. *Gen Tech. Rep. RMRS-GTR-216WWW. Fort Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Research Station. 98 p.*, 216, 2009.

[BS16] Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. *SSRN Electronic Journal*, 2016.

[BSR$^+$05] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

[BT52] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[BTT96] Aharon Ben-Tal and Marc Teboulle. Hidden convexity in some nonconvex quadratically constrained quadratic programming. *Math. Program.*, 72:51–63, 01 1996.

[BVB16] 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*, pages 2757–2765, 2016.

[BY20] Samuel Burer and Yinyu Ye. Exact semidefinite formulations for a class of (random and non-random) nonconvex quadratic programs. *Math. Program.*, 181(1):1–17, 2020.

[Cas19] Carlos Castillo. Fairness and Transparency in Ranking. *ACM SIGIR Forum*, 52(2):64–71, January 2019.

[Cat12] Manuela Cattelan. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, pages 412–433, 2012.

[CBN17] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, April 2017.

[Cho17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[CJ16a] Shuo Chen and Thorsten Joachims. Modeling intransitivity in matchup and comparison data. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pages 227–236, New York, NY, USA, 2016. ACM.

[CJ16b] Shuo Chen and Thorsten Joachims. Predicting matchups and preferences in context. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 775–784. ACM, 2016.

[CLRS09] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2009.

[CMV20] L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. Interventions for ranking in the presence of implicit bias. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 369–380, New York, NY, USA, 2020. Association for Computing Machinery.

[CSS07] Xin Chen, Melvyn Sim, and Peng Sun. A Robust Optimization Perspective on Stochastic Programming. *Operations Research*, 55:1058–1071, 2007.

[CSV18] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. Ranking with fairness constraints. *International Colloquium on Automata, Languages, and Programming*, 2018.

[CW17] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (sp)*, pages 39–57. IEEE, 2017.

[DARW$^+$19] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019.

[Das18] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, October 2018.

[DG17a] Dheeru Dua and Casey Graff. UCI machine learning repository. `https://archive.ics.uci.edu/ml/datasets/adult`, 2017.

[DG17b] Dheeru Dua and Casey Graff. UCI machine learning repository. `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`, 2017.

[DHP$^+$12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery.

[DY10] Erick Delage and Yinyu Ye. Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems. *Operations Research*, 58:595–612, 2010.

[EK15] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations. May 2015.

[ES16] Harrison Edwards and Amos Storkey. Censoring Representations with an Adversary. *ICLR*, November 2016.

[GAK19] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2221–2231, 2019.

[GJZ17] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1233–1242. JMLR.org, 2017.

[GNH19] Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects*. National Institute of Standards and Technology, 2019.

[Gob20] Jordan Goblet. Atp men's tour. `https://www.kaggle.com/jordangoblet/atp-tour-20002016`, 2020.

[GPL$^+$19] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226. ACM, 2019.

[GRGP01] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

[GS09] John Guiver and Edward Snelson. Bayesian inference for plackett-luce ranking models. In *proceedings of the 26th annual international conference on machine learning*, pages 377–384, 2009.

[GS10] Joel Goh and Melvyn Sim. Distributionally Robust Optimization and Its Tractable Approximations. *Operations Research*, 58(4-part-1):902–917, August 2010.

[GSS15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *ICLR*, December 2015.

[Hil20] Kashmir Hill. Wrongfully accused by an algorithm. *New York Times (Online)*, Jun 24 2020.

[HL04] Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 10, Seattle, WA, August 2004.

[HPPS16] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3315–3323. Curran Associates, Inc., 2016.

[HSNL18] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[HSR$^+$19] Reinhard Heckel, Nihar B Shah, Kannan Ramchandran, Martin J Wainwright, et al. Active ranking from pairwise comparisons and when parametric assumptions do not help. *The Annals of Statistics*, 47(6):3099–3126, 2019.

[Ilv20] Christina Ilvento. Metric Learning for Individual Fairness. *Foundations of Responsible Computing*, June 2020.

[JJN16] Lalit Jain, Kevin G Jamieson, and Rob Nowak. Finite sample prediction and recovery bounds for ordinal embedding. In *Advances in Neural Information Processing Systems*, pages 2711–2719, 2016.

[JKN$^+$19] Christopher Jung, Michael J. Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. Eliciting and enforcing subjective individual fairness. *CoRR*, abs/1905.10660, 2019.

[JN11] 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.

[Joa02] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

[KA09] Toshihiro Kamishima and Shotaro Akaho. Efficient clustering for orders. In *Mining complex data*, pages 261–279. Springer, 2009.

[KB15] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[KC09] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6. IEEE, 2009.

[Kel20] Ionas Kelepouris. NBA games stats from 2014 to 2018. `https://www.kaggle.com/ionaskel/nba-games-stats-from-2014-to-2018`, 2020.

[KGB17] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial Machine Learning at Scale. *ICLR*, November 2017.

[KKG18] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial Logit Pairing. *arXiv:1803.06373 [cs, stat]*, March 2018.

[KKK17] Aaron Kaufman, Gary King, and Mayya Komisarchik. How to measure legislative district compactness if you only know it when you see it. *American Journal of Political Science*, 2017.

[KLRS17] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. Curran Associates, Inc., 2017.

[KMR17] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science*, September 2017.

[KMU17] Jon Kleinberg, Sendhil Mullainathan, and Johan Ugander. Comparison-based choices. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 127–144. ACM, 2017.

[KNRW18] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. volume 80 of *ICML*, pages 2564–2572, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[Kru64] Joseph B Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129, 1964.

[KS17] Aritra Konar and Nicholas D. Sidiropoulos. First-order methods for fast feasibility pursuit of non-convex QCQPs. *IEEE Transactions on Signal Processing*, 65(22):5927–5941, 11 2017.

[KVR19] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. Fare: Diagnostics for fair ranking using pairwise error metrics. In *The World Wide Web Conference*, WWW '19, page 2936–2942, New York, NY, USA, 2019. Association for Computing Machinery.

[LCF$^+$18] Yao Li, Minhao Cheng, Kevin Fujii, Fushing Hsieh, and Cho-Jui Hsieh. Learning from group comparisons: Exploiting higher order interactions. In *Advances in Neural Information Processing Systems 31*, pages 4981–4990. Curran Associates, Inc., 2018.

[LLY17] Phil Lopes, Antonios Liapis, and Georgios N Yannakakis. Modelling affect for horror soundscapes. *IEEE Transactions on Affective Computing*, 10(2):209–222, 2017.

[LN15a] 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.

[LN15b] 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*, pages 1473–1479. IEEE, 2015.

[LR18] Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2687–2696. Curran Associates, Inc., 2018.

[LSJR16] Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. volume 49 of *Proceedings of Machine Learning Research*, pages 1246–1257, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

[Luc59] R Duncan Luce. *Individual choice behavior: A theoretical analysis.* Courier Corporation, 1959.

[LZ15] H. Lam and Enlu Zhou. Quantifying uncertainty in sample average approximation. In *2015 Winter Simulation Conference (WSC)*, pages 3846–3857, December 2015.

[MCPZ18] David Madras, Elliot Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. In *ICML*, 2018.

[MM08] Joshua E Menke and Tony R Martinez. A bradley–terry artificial neural network model for individual ratings in group competitions. *Neural computing and Applications*, 17(2):175–186, 2008.

[MMK$^+$16] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional Smoothing with Virtual Adversarial Training. *ICLR*, July 2016.

[MMS$^+$18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *ICLR*, June 2018.

[MS20] Cade Metz and Adam Satariano. An Algorithm That Grants Freedom, or Takes It Away. *The New York Times*, February 2020.

[MU19] Rahul Makhijani and Johan Ugander. Parametric models for intransitivity in pairwise rankings. In *The World Wide Web Conference*, pages 3056–3062, 2019.

[MYBS20a] Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Two simple ways to learn individual fairness metric from data. In *Proceedings of Machine Learning and Systems 2020*, pages 10708–10718. 2020.

[MYBS20b] Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Two simple ways to learn individual fairness metrics from data. In *International Conference on Machine Learning*, 2020.

[ND16] Hongseok Namkoong and John C. Duchi. Stochastic Gradient Methods for Distributionally Robust Optimization with F-divergences. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 2216–2224, Barcelona, Spain, 2016. Curran Associates Inc.

[NOS16] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2016.

[NR17] UN Niranjan and Arun Rajkumar. Inductive pairwise ranking: going beyond the n log (n) barrier. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[NRW$^+$12] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, Bin Yu, et al. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[OTX15a] Sewoong Oh, Kiran K Thekumparampil, and Jiaming Xu. Collaboratively learning preferences from ordinal data. In *Advances in Neural Information Processing Systems*, pages 1909–1917, 2015.

[OTX15b] Sewoong Oh, Kiran K Thekumparampil, and Jiaming Xu. Collaboratively learning preferences from ordinal data. In *Advances in Neural Information Processing Systems*, pages 1909–1917, 2015.

[PB17] Jaehyun Park and Stephen Boyd. General heuristics for nonconvex quadratically constrained quadratic programming. *arXiv preprint*, 2017.

[PGH19] Karlson Pfannschmidt, Pritha Gupta, and Eyke Hüllermeier. Learning choice functions. *preprint*, abs/1901.10860, 2019.

[Pla75] Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

[PMJ$^+$16] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS P)*, pages 372–387, 2016.

[PNZ$^+$15a] Dohyung Park, Joe Neeman, Jin Zhang, Sujay Sanghavi, and Inderjit Dhillon. Preference completion: Large-scale collaborative ranking from pairwise comparisons. In *International Conference on Machine Learning*, pages 1907–1916, 2015.

[PNZ$^+$15b] Dohyung Park, Joe Neeman, Jin Zhang, Sujay Sanghavi, and Inderjit Dhillon. Preference completion: Large-scale collaborative ranking from pairwise comparisons. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1907–1916, Lille, France, 07–09 Jul 2015. PMLR.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[PTB19] Flavien Prost, Nithum Thain, and Tolga Bolukbasi. Debiasing embeddings for reduced gender bias in text classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 69–75, Florence, Italy, August 2019. Association for Computational Linguistics.

[PV12] Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2012.

[QL13] Tao Qin and Tie-Yan Liu. Introducing LETOR 4.0 datasets. `http://arxiv.org/abs/1306.2597`, 2013.

[RA14] Arun Rajkumar and Shivani Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 118–126, Bejing, China, 22–24 Jun 2014. PMLR.

[RBM06] Jörg Rieskamp, Jerome R Busemeyer, and Barbara A Mellers. Extending the bounds of rationality: Evidence and theories of preferential choice. *Journal of Economic Literature*, 44(3):631–661, 2006.

[RGLA15] Arun Rajkumar, Suprovat Ghoshal, Lek-Heng Lim, and Shivani Agarwal. Ranking from stochastic pairwise preferences: Recovering condorcet winners and tournament solution sets at the top. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 665–673. JMLR.org, 2015.

[ROS20] Nir Rosenfeld, Kojin Oshiba, and Yaron Singer. Predicting choice with set-dependent aggregation. In *Proceedings of the 37th International Conference on Machine learning*, 2020.

[RSZ17] Ya'acov Ritov, Yuekai Sun, and Ruofei Zhao. On conditional parity as a notion of non-discrimination in machine learning. *arXiv:1706.08519 [cs, stat]*, June 2017.

[RU16] Stephen Ragain and Johan Ugander. Pairwise choice markov chains. In *Advances in Neural Information Processing Systems*, pages 3198–3206, 2016.

[SEK15] Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally Robust Logistic Regression. *NIPS*, September 2015.

[She62] Roger N Shepard. The analysis of proximities: multidimensional scaling with an unknown distance function. i. *Psychometrika*, 27(2):125–140, 1962.

[She64] Roger N Shepard. Attention and the metric structure of the stimulus space. *Journal of mathematical psychology*, 1(1):54–87, 1964.

[Sie20] Scott Sievert. New yorker caption contest data. `https://github.com/nextml/caption-contest-data/blob/master/contests/responses/508-round2-dueling-responses.csv.zip`, 2020.

[SJ18] Ashudeep Singh and Thorsten Joachims. Fairness of Exposure in Rankings. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*, pages 2219–2228, 2018.

[SJ19] Ashudeep Singh and Thorsten Joachims. Policy learning for fairness in ranking. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5426–5436. Curran Associates, Inc., 2019.

[SND18] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[Spe17] Robyn Speer. How to make a racist ai without really trying. `https://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/`, 2017.

[SPU19] Arjun Seshadri, Alexander Peysakhovich, and Johan Ugander. Discovering context effects from raw choice data. *International Conference on Machine Learning*, 2019.

[SR18] Aadirupa Saha and Arun Rajkumar. Ranking with features: Algorithm and a graph theoretic analysis. `https://arxiv.org/pdf/1808.03857.pdf`, 2018.

[SW17] Nihar B Shah and Martin J Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *The Journal of Machine Learning Research*, 18(1):7246–7283, 2017.

[SZR$^+$19] Piotr Sapiezynski, Wesley Zeng, Ronald E. Robertson, Alan Mislove, and Christo Wilson. Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists. In *Proceedings of the Workshop on Fairness, Accountability, Transparency, Ethics, and Society on the Web (FATES'19)*, San Francisco, CA, USA, May 2019.

[SZS$^+$14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[TDN20] Parth Thaker, Gautam Dasarathy, and Angelia Nedic. On the sample complexity and optimization landscape for quadratic feasibility problems. In *IEEE International Symposium on Information Theory*, pages 1438–1443, 06 2020.

[TL14] Yoshikazu Terada and Ulrike Luxburg. Local ordinal embedding. In *International Conference on Machine Learning*, pages 847–855, 2014.

[Tor65] Warren S Torgerson. Multidimensional scaling of similarity. *Psychometrika*, 30(4):379–393, 1965.

[Tro12] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[TS93] Amos Tversky and Itamar Simonson. Context-dependent preferences. *Management science*, 39(10):1179–1189, 1993.

[Tve72] Amos Tversky. Elimination by aspects: A theory of choice. *Psychological review*, 79(4):281, 1972.

[Tve77] Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

[TVL14] Yoshikazu Terada and Ulrike Von Luxburg. Local ordinal embedding. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–847–II–855. JMLR.org, 2014.

[Vig19] Neil Vigdor. Apple card investigated after gender discrimination complaints, Nov 10 2019.

[Wat20] Waterproofpaper.com. Printable map of us state capitals, 2020. [Online; accessed July 22, 2020].

[WGL$^+$19] Hanchen Wang, Nina Grgic-Hlaca, Preethi Lahoti, Krishna P. Gummadi, and Adrian Weller. An Empirical Study on Learning Fairness Metrics for COMPAS Data with Human Supervision. *NeurIPS HCML Workshop*, October 2019.

[Wil92] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[WZW18] Yongkai Wu, Lu Zhang, and Xintao Wu. On discrimination discovery and removal in ranked data using causal graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, page 2536–2544, New York, NY, USA, 2018. Association for Computing Machinery.

[YBS20] Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ML models with sensitive subspace robustness. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia, 2020.

[YBW15] Dehui Yang and Michael B. Wakin. Modeling and recovering non-transitive pairwise comparison matrices. *2015 International Conference on Sampling Theory and Applications, SampTA 2015*, pages 39–43, 07 2015.

[YDJ19] Himank Yadav, Zhengxiao Du, and Thorsten Joachims. Fair Learning-to-Rank from Implicit Feedback. *arXiv:1911.08054 [cs, stat]*, November 2019.

[YG14] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *Computer Vision and Pattern Recognition (CVPR)*, Jun 2014.

[YG17] A. Yu and K. Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *International Conference on Computer Vision (ICCV)*, Oct 2017.

[YGS19] Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. Balanced ranking with diversity constraints. IJCAI International Joint Conference on Artificial Intelligence, pages 6035–6042, 2019.

[YS17] Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, SSDBM '17, New York, NY, USA, 2017. Association for Computing Machinery.

[YS20] Mikhail Yurochkin and Yuekai Sun. SenSeI: Sensitive set invariance for enforcing individual fairness. `https://arxiv.org/abs/2006.14168`, 2020.

[YWHD19] Fanny Yang, Zuowen Wang, and Christina Heinze-Deml. Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness. In *Advances in Neural Information Processing Systems 32*, pages 14785–14796. Curran Associates, Inc., 2019.

[ZBC$^+$17] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578, 2017.

[ZC20] Meike Zehlike and Carlos Castillo. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference 2020*, WWW '20, page 2849–2855, New York, NY, USA, 2020. Association for Computing Machinery.

[ZLM18] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA, 2018. Association for Computing Machinery.