Matrix estimation for individual fairness

Sarah H. Cen *
MIT EECS
shcen@mit.edu

Cindy Zhang*
MIT Mathematics
cindyyz@mit.edu

Devavrat Shah*
MIT EECS
devavrat@mit.edu

Abstract

In recent years, multiple notions of algorithmic fairness have arisen. Individual fairness (IF) is one such notion, which states that similar individuals should receive similar treatment. In this work, we establish a strong connection between matrix estimation and IF. We propose to use a popular matrix estimation technique known as singular value thresholding (SVT) to pre-process sparse, noisy data before applying a prediction algorithm. We show that using SVT provides an IF guarantee. We study performance under SVT and find that, under mild conditions, IF does not increase the prediction error, which indicates that there is little to no performance-fairness trade-off in settings for which matrix estimation is applicable. We illustrate these findings on a well known recommender system dataset. We confirm that the proposed method provides individually fair recommendations. We also show that the proposed method has desirable group fairness and interpretability properties.

1 Introduction

As data-driven decision-making becomes more ubiquitous, attention on the *fairness* of machine learning (ML) algorithms has risen. Because what is deemed to be fair is context-dependent (e.g., reflects a given value system), there is no universally accepted notion of fairness.

One notion of algorithmic fairness is *individual fairness (IF)*, which is distinct from notions of group fairness (e.g., parity in error rates). Stated informally, IF captures the idea that similar individuals should receive similar treatment. More precisely, for a given decision, individuals with similar decision-relevant attributes should experience each possible outcome at similar frequencies.

For example, suppose that two individuals in some world W are applying for the same jobs at the same time and have almost identical job qualifications. If these qualifications are the only factors that should be considered during the hiring process under the accepted social and legal norms of W, then IF requires that both individuals receive job offers at approximately the same rates. If IF does not hold, then the hiring decisions use the applicants' attributes in a way that W does not consider relevant to the hiring process and systematically benefits one individual at the expense of another.

In recent years, there has been ongoing debate on (a) how to enforce IF and (b) whether there is a performance-fairness trade-off. Both can be understood by observing that IF is usually enforced by adding a regularizer or constraint to an optimization problem, which requires *explicitly formalizing three quantities*: (1) what it means for two individuals to be similar, (2) what it means for two treatments to be similar, and (3) the acceptable level of dissimilarity between the outcomes that two similar individuals receive. There is no consensus on how to specify these three quantities. The debate reflects an important concern that the precise formulation of IF has serious consequences and that, in the limit, an adversarially designed formulation is no better than not enforcing IF at all. Moreover, because IF is typically formulated as a constraint (or, equivalently, a regularizer), there is

*All authors are from the Massachusetts Institute of Technology (MIT). Cen and Shah are with the Department of Electrical Engineering and Computer Science (EECS). Zhang is with the Department of Mathematics.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

concern that enforcing IF may hurt performance. Whether there is a performance-fairness trade-off remains an open question.

In light of this debate, we ask two questions. First, as IF is typically enforced as a constraint or regularizer, is there a class of problems for which IF can be enforced using an alternative approach? Second, is there necessarily a performance-fairness trade-off?

In this work, we establish a connection between IF and a popular matrix estimation (ME) technique known as *singular value thresholding (SVT)*. We propose to use SVT to pre-process sparse, noisy data before feeding it into an inference algorithm. We show that, as long as this algorithm is well behaved, the resulting prediction is individually fair, i.e., the proposed procedure naturally guarantees IF. We then study how IF affects prediction performance. We find that the proposed procedure mirrors a well-known ME method that has strong performance guarantees. This result implies that there is little to no performance-fairness trade-off in settings where ME can be applied, such as in collaborative filtering. Our contributions are summarized as follows.

Connection between individual fairness and singular value thresholding. ME is used in high-dimensional inference to handle sparse, noisy data. One of the most popular ME methods is singular value thresholding (SVT), which produces an estimate or prediction using only singular values above some threshold. In Section 4.2, we derive a set of thresholds under which pre-processing sparse, noisy data using SVT guarantees IF. In particular, we find that SVT naturally guarantees IF for sufficiently high thresholds. We then use the result to explore how IF affects predictions in sparse-data regimes.

The effect of individual fairness on performance. In the ME literature, universal singular value thresholding (USVT) is a principled application of SVT that has strong performance guarantees; specifically, it produces an estimator that is not only consistent but approximately minimax [14]. In Section 4.3, we show that there is a close connection between IF and USVT. One way to interpret this result is that, in general, IF does not harm performance because it places no further restrictions on ME than the performance-based measures already required by USVT.

Illustration. In Section 5, we demonstrate our findings on movie recommendation data. We visualize the output of USVT and show it provides individually fair recommendations. We also study how USVT affects recommendations across sensitive features, such as gender. Finally, we illustrate the interpretability of spectral matrix methods.

To our knowledge, this is the first work that establishes a theoretical link between IF and ME.

2 Background and related work

Matrix estimation (ME). ME studies the problem of estimating the entries of a matrix from noisy observations of a subset of the entries [13, 37, 28, 35, 17, 14, 15]. ME is a general method that can be applied to data that can be expressed in matrix form. Specifically, suppose there is a latent matrix, and one can only obtain noisy samples of a subset of its entries. The goal of ME is to estimate the values of every entry based on the noisy subsamples. ME is used, for example, by recommender systems to estimate a user's interest in different types of content [31, 41, 11]. In fact, the winning solution of the Netflix Prize was built on ME methods [30]. ME has also been used to study social networks [5, 1, 24]; to impute and forecast a time series [3, 4]; to aggregate information in crowdsourcing [40]; to improve robustness against adversarial attacks in deep learning [43]; and more.

Singular value thresholding (SVT). There is an extensive literature on ME and the closely related areas of matrix completion and matrix factorization. While there are various approaches [38], spectral methods are among the most popular [13, 34, 28, 29] One such method is SVT [12], which first factorizes the matrix of sparse, noisy observations, then reconstructs it using only the singular values that exceed a predetermined threshold. It is well-known that SVT is a shrinkage operator that provides a solution to a nuclear norm minimization problem. *Universal singular value thresholding* (USVT) builds on SVT by proposing an adaptive threshold that produces an estimator that is both consistent and approximately minimax [14]. We review SVT and USVT in Sections 4.1 and 4.3.

Individual fairness (IF). Stated informally, IF is the notion that similar individuals should receive similar treatment [18, 6]. As an example, suppose individuals A and B apply for job interviews at the same time with similar qualifications. Then, IF requires that A and B receive interview requests

at similar rates. IF is distinct from notions of group fairness (e.g., statistical parity in the outcomes across demographic groups), but there are conditions under which IF implies group fairness [18].[2]

Enforcing individual fairness. There has been significant debate on the formulation of IF, specifically how to define the "similarity" of individuals and the treatments that they receive [25, 22, 8]. Under IF, similarity is captured by the *choice* of distance metrics, and IF is enforced as a Lipschitz constraint based on the chosen metrics. This metric-based approach is both a strength and weakness. On the one hand, the metrics can be adapted to different contexts, allowing IF to be task-relevant and society-dependent [25]. On the other hand, devising context-aware metrics is ethically challenging because the chosen metrics implicitly encode a value system [21]. For example, if the difference in GPAs contributes more to the distance between two job applicants than differences in their recommendations, then the chosen metric places greater emphasis on GPAs than on recommendations. Several works propose to learn context-aware distance metrics from human feedback [25, 9, 32] or under the guidance of a fairness oracle [22, 8]. However, obtaining consensus on the distance metrics remains the primary barrier to implementing IF in practice.

Individual fairness and collaborative filtering. Concerns of fairness related to ME generally arise in the recommendation setting. ME is used as a *collaborative filtering* technique in that recommendations for a specific user leverage information about other users. Most works on the fairness of collaborative filtering study group fairness [27, 44, 9, 20, 36]. A small number of works examine notions of fairness related to individuals [39, 10, 42], but they are distinct from our notion of IF as formulated by Dwork et al. [18]. To our knowledge, we provide the first theoretical analysis connecting IF to ME and collaborative filtering, which can be found in Section 4

Accuracy and efficiency. Two common threads of interest in algorithmic fairness are the fairness-accuracy trade-off [19, 45, 33, 26] and the computational cost of fairness [36]. By establishing a connection between IF and USVT, we show in Section 4.3 that IF can be achieved without significant performance or computational costs in ME applications, including collaborative filtering.

Protected and sensitive attributes. We briefly comment on a consideration tangentially related to our work: the use of protected or sensitive attributes (e.g., gender, race, age). While some believe these attributes must be excluded from the feature set to ensure fairness, there is mounting evidence that doing so can *increase* disparities due to, for instance, proxy variables. Moreover, there are settings in which the use of a protected attribute is legally permitted[3] Determining when these attributes are *task-relevant* and to what extent they should influence algorithmic decision-making remain open questions. We briefly illustrate how our proposed method impacts predictions across sensitive attributes in Section 5.

3 Problem statement

3.1 Setup

Consider a setting with m individuals. Suppose there is an unknown ground truth matrix $A \in \mathbb{R}^{m \times n}$, where each row in A corresponds to an individual such that the i-th row $\mathbf{a}_i \in \mathbb{R}^n$ is an unknown n-dimensional feature vector that describes individual $i \in [m]$. Without loss of generality, suppose that $A_{ij} \in [-1, 1]$ for all $i \in [m]$ and $j \in [n]$[4].

Although A is unknown, suppose that it is possible to observe a noisy subsample of its entries. Formally, let $\Omega \subset [m] \times [n]$ denote the index set of observed entries and $\mathcal{Z} = [-1, 1] \cup \{\emptyset\}$. Let $Z \in \mathcal{Z}^{m \times n}$ denote the matrix of observations, where each entry of Z is a random variable, $\mathbb{E}Z_{ij} = A_{ij}$ and $Z_{ij} \in [-1, 1]$ if $(i, j) \in \Omega$; and $Z_{ij} = \emptyset$, otherwise. Intuitively, Z is the matrix of sparse, noisy observations while A contains perfect information.

[2] Under group fairness (GF), the goal is to achieve (approximate) parity in group-level statistics (e.g., to equalize true positive rates across race, gender, or age groups). GF has inspired a host of interesting research and insights [6, 16]. However, several works provide poignant examples that highlight the weaknesses of GF [18, 25]. For instance, an algorithm that treats all individuals equally poorly passes the GF requirement, and many of the GF definitions are conflicting, making them impossible to satisfy simultaneously [16].

[3] Under U.S. law, hiring decisions can legally use a protected attribute (e.g., age) if the business shows that it is an unavoidable factor in job performance and that there is no good alternative attribute to consider [7].

[4] For any A whose entries are finite such that $|A_{ij}| < \infty$ for all $i \in [m]$ and $j \in [n]$, one can always translate and rescale A to be between -1 and 1, then adjust the final result accordingly.

Consider the following inference task. Given a context $\mathbf{c} \in \mathcal{C}$, make a prediction $\mathbf{y} \in \mathcal{Y}$ for individual $i \in [m]$ using the observations Z. Let $\mathcal{F} = \{f : [m] \times \mathcal{Z}^{m \times n} \times \mathcal{C} \to \mathcal{Y}\}$ denote the class of algorithms that perform this inference task. Note that the output of f could be a deterministic value or a distribution over possible values. We provide examples of this setup in Section 3.3 below.

3.2 Individual fairness

Individual fairness (IF) is the notion that *similar individuals should receive similar treatments* [18]. IF is formulated as a (D, d)-Lipschitz constraint, as follows.

Definition 1. *Consider an observation matrix $Z \in \mathcal{Z}^{m \times n}$ and context $\mathbf{c} \in \mathcal{C}$. An algorithm $f \in \mathcal{F}$ is* individually fair on *Z if there exists a constant $L(Z, \mathbf{c}) \geq 0$, a metric d on $\mathcal{Z}^n$, and a metric D on $\mathcal{Y}$ such that:*

$$D(f(i, Z, \mathbf{c}), f(j, Z, \mathbf{c})) \leq L(Z, \mathbf{c}) d(\mathbf{z}_i, \mathbf{z}_j) \quad \forall i, j \in [m]. \tag{1}$$

f is approximately *individually fair on Z if there exists an additional constant $\delta(Z, \mathbf{c}) > 0$ such that:*

$$D(f(i, Z, \mathbf{c}), f(j, Z, \mathbf{c})) \leq L(Z, \mathbf{c}) d(\mathbf{z}_i, \mathbf{z}_j) + \delta(Z, \mathbf{c}) \quad \forall i, j \in [m]. \tag{2}$$

Definition 2. *Consider an observation matrix $Z \in \mathcal{Z}^{m \times n}$ and context $\mathbf{c} \in \mathcal{C}$. An algorithm $f \in \mathcal{F}$ is* individually fair on *A if there exists a constant $L(Z, \mathbf{c}) \geq 0$, a metric d on $[-1, 1]^n$, and a metric D on $\mathcal{Y}$ such that:*

$$D(f(i, Z, \mathbf{c}), f(j, Z, \mathbf{c})) \leq L(Z, \mathbf{c}) d(\mathbf{a}_i, \mathbf{a}_j) \quad \forall i, j \in [m]. \tag{3}$$

f is approximately *individually fair on A if there exists an additional constant $\delta(Z, \mathbf{c}) > 0$ such that:*

$$D(f(i, Z, \mathbf{c}), f(j, Z, \mathbf{c})) \leq L(Z, \mathbf{c}) d(\mathbf{a}_i, \mathbf{a}_j) + \delta(Z, \mathbf{c}) \quad \forall i, j \in [m]. \tag{4}$$

Intuitively, d captures what it means for two individuals to be similar in the input (or feature) space; D captures what it means for two individuals to receive similar treatment; and L and δ capture the acceptable level of dissimilarity in the treatment of two similar individuals.

Definitions 1 and 2 differ in terms of the input space over which d is applied. An algorithm f is individually fair on Z (resp., on A) if two individuals with similar observed features $\mathbf{z}_i$ and $\mathbf{z}_j$ (resp., similar true features $\mathbf{a}_i$ and $\mathbf{a}_j$) receive similar treatments.

Our objective. The objective of this work is to provide a procedure for achieving IF in settings with *sparse, noisy data*. We focus on a subclass of algorithms $\mathcal{F}(\mathcal{H}, \Pi) = \{f = h \circ \Pi : h \in \mathcal{H}\} \subset \mathcal{F}$, where $\mathcal{H} \subset \{h : [m] \times [-1, 1]^{m \times n} \times \mathcal{C} \to \mathcal{Y}\}$ and $\Pi : \mathcal{Z}^{m \times n} \to [-1, 1]^{m \times n}$. Intuitively, Π is an algorithm that takes in sparse, noisy data Z and produces an estimate $\Pi(Z)$ of the unknown matrix A. The algorithm h is then applied on top of Π such that $f(i, Z, \mathbf{c}) = h(i, \Pi(Z), \mathbf{c})$.

We ask two questions. First, given a smooth function h, how can one design Π in order to achieve IF? Second, how does the proposed Π affect the prediction performance of f?

3.3 Examples

The setup in Section 3.1 can be applied to many problems in which the training data and algorithmic inputs are *noisy, sparse, or both*. Consider the following examples and the implications of IF.

Example 1 (Recommendation)**.** *Consider a platform that provides personalized movie recommendations to its m users based on sparse, noisy observations of their preferences. Suppose that the movie preferences of each user $i \in [m]$ can be described by an unknown n-dimensional vector $\mathbf{a}_i \in \mathbb{R}^n$. Although $A = [\mathbf{a}_1, \ldots, \mathbf{a}_m]^\top$ is unknown, the platform receives occasional feedback from users in the form of ratings and can also observe the users' viewing behaviors. These sparse noisy observations are stored in Z, where $Z_{ij} = \emptyset$ implies that the platform does not have information about feature j for user i. The movies that a platform recommends may depend on the context $\mathbf{c}$. For example, the recommendations on a user's homepage are different from what is shown if the user types "foreign comedies" into the search bar. As an output, the recommendation algorithm could produce a multi-hot vector $\mathbf{y} \in \mathbb{R}^s$ over s available movies, where a high value for $\mathbf{y}_k$ indicates a high predicted likelihood that the user would enjoy movie $k \in [s]$. Note that $f \in \mathcal{F}$ can leverage other information (e.g., ratings by other users, as done in collaborative filtering). In this example, IF on Z requires that users with similar viewing and rating behaviors receive similar recommendations. IF on A implies that users with similar (unknown) movie preferences receive similar recommendations.*

Algorithm 1: Singular value thresholding (SVT)

Input: Observation matrix $Z \in \mathcal{Z}^{m \times n}$, threshold $\tau \geq 0$, increasing function $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$.

Output: Estimate $\hat{A} \in \mathbb{R}^{m \times n}$.

for $i \in [m]$, $j \in [n]$ **do**
  If $Z_{ij} = \emptyset$, $Z_{ij} \leftarrow 0$;
  $Z_{ij} \leftarrow \max(-1, \min(1, Z_{ij}))$;
end

Compute the singular value decomposition (SVD): $Z = \sum_{\ell=1}^{\min(m,n)} \sigma_\ell \mathbf{u}_\ell \mathbf{v}_\ell^T$ where $\sigma_\ell \geq 0$ is the ℓ-th singular value, $\mathbf{u}_\ell \in \mathbb{R}^{m \times 1}$ is the ℓ-th left singular vector, $\mathbf{v}_\ell \in \mathbb{R}^{n \times 1}$ is the ℓ-th right singular vector, and $\sigma_1, \sigma_2, \ldots$ are in decreasing order;

Threshold: Let $S(\tau) := \{\ell : \sigma_\ell > \tau\}$ be the set of components whose singular values exceed τ;

Produce estimate: $\hat{A} \leftarrow \min\left(1, \max\left(-1, \sum_{\ell \in S(\tau)} \alpha(\sigma_\ell) \mathbf{u}_\ell \mathbf{v}_\ell^T\right)\right)$;

Example 2 (Admissions)**.** *Consider an admissions setting in which there are m applicants. Suppose that, for the purposes of admissions, each applicant $i \in [m]$ is described by an unknown n-dimensional vector $\mathbf{a}_i \in \mathbb{R}^n$. Suppose each individual i submits an application $\mathbf{z}_i$, which contains sparse, noisy measurements of $\mathbf{a}_i$. For example, one's standardized test score in math is a noisy measurement of one's math abilities. Data sparsity can occur when one applicant includes information that another does not (e.g., one may list "debate club" on their resume while another does not, but this sparsity does not necessarily imply that the latter is worse at public speaking). In this setting, $\mathbf{c}$ could be the same across all applicants or reflect different admissions contexts. As an output, $f \in \mathcal{F}$ could produce an admissions score $\mathbf{y} \in [0, 1]$. In this example. IF on Z requires that applicants with similar admissions-relevant applications receive similar admissions scores. IF on A implies that applicants whose true (unknown) qualifications are similar receive similar admissions scores.*

4 Main results

Our main contributions are as follows. In Section 4.1, we propose to use an ME method known as SVT to pre-process sparse, noisy data before feeding the data into an inference algorithm. In Section 4.2, we show that, as long as this algorithm is well behaved, the resulting predictions are individually fair and provide a closed-form IF guarantee. We use these results to investigate how IF affects predictions in sparse-data regimes. We find that, when the data is too sparse, IF requires that all individuals are treated similarly. This finding is consistent with the IF notion that, when the algorithm does not have enough information to distinguish between individuals, it should treat them similarly or acquire more data. In Section 4.3, we show that the class of SVT methods that guarantee IF mirrors a well-known ME technique that has strong performance guarantees. This connection implies that IF does not impose a high performance cost in settings for which ME is appropriate.

4.1 Proposed procedure: De-noising with singular value thresholding

In this section, we provide a procedure for processing sparse, noisy data in order to perform the inference task described in Section 3.1. In the following two sections, we study the fairness and performance of the corresponding class of algorithms.

Matrix estimation (ME) studies the problem of estimating an unknown matrix from sparse, noisy observations of its entries. ME can be used for the inference task by providing an algorithm for Π. In this work, we consider a popular ME method known as **singular value thresholding (SVT)**. The SVT procedure is given in Algorithm 1 Intuitively, SVT detects and removes components of the observation matrix Z that are within the noise. The threshold τ determines the boundary between signal and noise, where a higher value for τ means that fewer components are kept.

Let $\Pi_\alpha^\tau : \mathcal{Z}^{m \times n} \to \mathbb{R}^{m \times n}$ denote the SVT operation such that $\hat{A} = \Pi_\alpha^\tau(Z)$ is the estimate of A that is obtained by running Algorithm 1 on data Z using the threshold τ and increasing function α. For notational convenience, let $\mathcal{F}(\mathcal{H}, \tau, \alpha) := \mathcal{F}(\mathcal{H}, \Pi_\alpha^\tau)$.

4.2 Equivalence between IF and SVT

In the previous section, we proposed to run the sparse, noisy data Z through SVT before applying an inference algorithm h on top of it. In this section, we show that using this procedure provides strong fairness guarantees by establishing an equivalence between SVT and IF.

Recall that σ_ℓ, $\mathbf{u}_\ell$, and $\mathbf{v}_\ell$, respectively, are the ℓ-th singular value, left singular vector, and right singular vector of Z. Recall further that $S(\tau) := \{\ell : \sigma_\ell > \tau\}$. Let $\kappa^{\max} = \max_{k\in[m]} \|\mathbf{z}_k\|_2^2$, $\kappa^{\min} = \min_{k\in[m]} \|\mathbf{z}_k\|_2^2$, and $\|M\|_{q,\infty} = \max_i \|\mathbf{m}_i\|_q$. Finally, let:

$$K_1 = (m-1)\sqrt{\kappa^{\max}} \sum_{\ell\in S_\tau} \frac{\alpha(\sigma_\ell)\,\|\mathbf{u}_\ell\|_\infty \|\mathbf{v}_\ell\|_\infty}{\sigma_\ell^2 - 2\kappa^{\max}} \qquad K_2 = \sqrt{\kappa^{\max}} \sum_{\ell\in S_\tau} \left(\frac{\sigma_\ell^2 - \kappa^{\min}}{\sigma_\ell^2 - \kappa^{\max}} - 1\right).$$

We now present our main results on the relationship between SVT and IF.

Theorem 1. *Suppose $D(h(i,\hat{A},\mathbf{c}), h(j,\hat{A},\mathbf{c})) \le L_1 \|\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j\|_1$ for all $i,j \in [m]$ and $h \in \mathcal{H}$. Suppose $m > 1$ and $\|\mathbf{z}_k\|_2^2 > 0$ for all $k \in [m]$. Then, for any $f \in \mathcal{F}(\mathcal{H},\tau,\alpha)$ where $\tau \ge \sqrt{2\kappa^{\max}}$,*

$$D(f(i,Z,\mathbf{c}), f(j,Z,\mathbf{c})) \le nL_1K_1 \|\mathbf{z}_i - \mathbf{z}_j\|_2 + nL_1K_1K_2,$$

for all $i,j \in [m]$. If $h(k,\hat{A},\mathbf{c}) = \hat{\mathbf{a}}_k^\top \mathbf{c}$, then

$$|f(i,Z,\mathbf{c}) - f(j,Z,\mathbf{c})| \le \|\mathbf{c}\|_1 L_1K_1 \|\mathbf{z}_i - \mathbf{z}_j\|_2 + \|\mathbf{c}\|_1 L_1K_1K_2.$$

for all $i,j \in [m]$. If $\|\mathbf{z}_i\|_2^2 = \|\mathbf{z}_j\|_2^2$, the same results hold with $K_2 = 0$.

Theorem 2. *Suppose $D(h(i,\hat{A},\mathbf{c}), h(j,\hat{A},\mathbf{c})) \le L_1 \|\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j\|_q$ for all $i,j \in [m]$ and $h \in \mathcal{H}$. Then, for all $i,j \in [m]$,*

$$D(f(i,Z,\mathbf{c}), f(j,Z,\mathbf{c})) \le L_1 \|\mathbf{a}_i - \mathbf{a}_j\|_q + 2L_1||\hat{A} - A||_{q,\infty}.$$

Theorems 1 and 2 illustrate the relationship between SVT and IF. Theorem 1 shows that, when τ is sufficiently large, SVT performs a smoothness operation and, in doing so, ensures that f is individually fair on Z. Theorem 2 states that f is also individually fair on A when the estimate $\hat{A}$ is sufficiently close to A. Note that Theorem 2 holds for any ME method. These results demonstrate that, in contrast to other black-box methods, applying ME provides a closed-form IF guarantee.[5]

Interpreting Theorem 1. In order to interpret this result, we study the average-case behavior of the constant K_1. Suppose that, for every row in A, $\lfloor np \rfloor$ of its entries are randomly observed, and the observations are binary such that $Z_{ij} \in \{-1,+1\}$ for $(i,j) \in \Omega$. Let $r = |S_\tau|$ and $\alpha(x) = x$. Suppose Z satisfies the standard incoherence condition with parameter μ_0, the singular values that survive SVT satisfy $\sigma_\ell \approx \gamma_\ell n$, and the context vector is always normalized such that $\|\mathbf{c}\|_1 = 1$. Then,

$$K_1 \lesssim m\sqrt{np} \sum_{\ell\in S_\tau} \frac{\gamma_\ell n\mu_0 r}{(\gamma_\ell^2 n^2 - 2np)\sqrt{mn}} \le \frac{\gamma_r \mu_0 r^2 \sqrt{mp}}{\gamma_r^2 n - 2p}.$$

In addition, when $m > 1$,

$$K_1 \gtrsim (m-1)\sqrt{np} \sum_{\ell\in S_\tau} \frac{\gamma_\ell n\mu_0 r}{\gamma_\ell^2 n^2 \sqrt{mn}} \ge \frac{\mu_0 r^2 \sqrt{mp}}{2\gamma_1 n}.$$

In other words, $K_1 \|\mathbf{z}_i - \mathbf{z}_j\|_2 = \Theta(r^2\sqrt{mp^2/n})$, which implies that there are two regimes of behavior as $m \to \infty$ (as more individuals are observed).[6] When $m = o(n/(p^2r^4))$, then SVT satisfies IF by making predictions that are similar across all individuals. On the other hand, when $m = \omega(n/(p^2r^4))$, then SVT satisfies IF while also making differentiated predictions across individuals.

To understand these two regimes, consider Examples 1 and 2. When $m = o(n/(p^2r^4))$, the observed data Z is very sparse. Intuitively, f does not have enough samples on which to train compared

[5]Note that the condition in both theorems that $D(h(i,\hat{A},\mathbf{c}), h(j,\hat{A},\mathbf{c})) \le L_1||\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j||_q$ for all $i,j \in [m]$ and $h \in \mathcal{H}$ is not strong. In fact, if it is not met, then there is no method Π such that f is IF.

[6]For more details on this analysis, please see the Appendix.

to the complexity of the problem. In response, SVT satisfies IF by producing predictions that are similar across all individuals. In the recommendation setting, this outcome would correspond to the platform not having enough information to learn each user's distinct preferences and, in response, recommending similar content to everyone. In the admissions setting, this outcome would correspond to not knowing enough about the applicants to fairly compare them side-by-side. Enforcing IF in this regime would not mean that the applicants cannot be ranked. Rather, suppose that f outputs a mixed strategy. Then, IF requires that the mixed strategy across all applicants is similar, i.e., that the admissions policy acknowledge the large amount of uncertainty in its ranking.[7] In the second regime—when $m = \omega(n/(p^2r^4))$—there are enough samples for f to train on compared to the complexity of the problem. Intuitively, compared to the first regime, f has enough information in order to detect patterns across individuals and reliably differentiate between them.

Interpreting Theorem 2. It is usually too strong of a requirement to demand that an algorithm f is individually fair on A, i.e., with respect to the unknown feature vectors $\mathbf{a}_i$ and $\mathbf{a}_j$. Even so, Theorem 2 shows that it is possible to achieve approximate IF on A, and the tightness of this guarantee depends on the accuracy of the ME method. Theorem 2 can be applied to any ME method for which there are upper bounds on $||\hat{A} - A||_{q,\infty}$ (cf. Theorem D.1 in Agarwal et al. [2]).

4.3 Performance under individual fairness

In Section 4.2, we show that pre-processing Z using ME can guarantee IF on Z as well as A. In this section, we show that IF and performance are closely tied under ME.

Performance under IF on Z. Recall from Theorem 1 that, as long as the threshold τ is sufficiently large, SVT achieves IF on Z. However, it is unclear if the threshold chosen for IF is a "good" one for prediction performance. We now show that the threshold chosen for IF on Z is within a constant factor of an adaptive threshold that is known to provide high accuracy.

We consider a well-known ME method known as **universal singular value thresholding (USVT)**. USVT refines SVT by proposing a universal formula for the threshold τ, thereby removing the need to tune τ by hand. Under mild assumptions on A and Ω, USVT has strong performance guarantees. In order to study performance, let the mean-squared error (MSE) of ME be:

$$\text{MSE}(\hat{A}) := \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}(\hat{A}_{ij} - A_{ij})^2.$$

Let $\mathcal{L}(f)$ denote the loss of algorithm $f \in \mathcal{F}$. Let $\|M\|_*$ and $\|M\|_2$ denote the nuclear and spectral norms, respectively, of matrix M. We begin with a well-known performance guarantee on USVT.

Proposition 3 (Modified from Theorem 1.1. in Chatterjee [14])**.** *Suppose that the entries of A are independent random variables. Suppose each entry of A is independently observed with probability $p \in [0,1]$. Let $\hat{p}$ be the proportion of observed values, $\alpha(x) = x/\hat{p}$, $\epsilon \in (0,1]$, and $w = (2+\eta)^2$ for $\eta \in (0,1)$. Let $\rho_1 = \max(m,n)$ and $\rho_2 = \min(m,n)$. Then, if $p \geq \rho_1^{\epsilon-1}$ and $\tau = \sqrt{w\rho_1\hat{p}}$,*

$$\textit{MSE}\left(\Pi_\alpha^\tau(Z)\right) \leq C(\eta)\min\left(\frac{\|A\|_*}{\rho_2\sqrt{\rho_1 p}}, \frac{\|A\|_*^2}{\rho_1\rho_2}, 1\right) + C(\epsilon,\eta)\exp(-c(\eta)\rho_1 p),$$

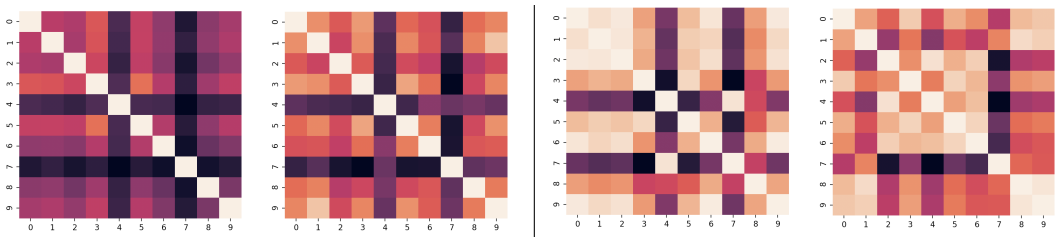
where $C(\eta), c(\eta) > 0$ depend only on η and $C(\epsilon,\eta)$ depends only on η and ϵ.[8]

Corollary 4. *Suppose $f = h \circ \Pi_\alpha^\tau$ and $\mathcal{L}(f) \leq L_2 \textit{MSE}(\Pi_\alpha^\tau(Z)) + \delta_2$, where $L_2 > 0$ and $\delta_2 \geq 0$. Then, under the same conditions as those in Proposition 3, $\lim_{m\to\infty}\mathcal{L}(f) = \delta_2$.*

[7] A proponent of IF would argue that this behavior is appropriate. If f does not have enough information to train on, then by tailoring its predictions to each individual's sparse noisy features, f may use spurious information to draw distinctions between individuals in an unfair manner. In the admissions example, a proponent of IF would argue that the admissions board can move out of the sparse regime by acquiring more information about applicants or standardizing the qualifications over which the applicants are evaluated.

[8] This upper bound can be improved when the additional condition that $\text{Var}(Z_{ij}) \leq \sigma^2$ for all i,j and $\sigma \leq 1$ holds. Then, if $\tau \geq \sqrt{wn\hat{q}}$, where $\hat{q} = \hat{p}\sigma^2 + \hat{p}(1-\hat{p})(1-\sigma^2)$, $q \geq n^{\epsilon-1}$, and $q = p\sigma^2 + p(1-p)(1-\sigma^2)$:

$$\text{MSE}(\hat{A}) \leq C(\eta)\min\left(\frac{\|A\|_*\sqrt{q}}{mp\sqrt{n}}, \frac{\|A\|_*^2}{mn}, 1\right) + C(\epsilon,\eta)\exp(-c(\eta)nq).$$



(a) $\|\mathbf{z}_i - \mathbf{z}_j\|_2$ (b) $\|\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j\|_2$ (c) $|(\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j)^\top \mathbf{c}_1|$ (d) $|(\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j)^\top \mathbf{c}_2|$

Figure 1: Visualizing individual fairness (IF) of universal singular value thresholding (USVT). Entry (i, j) in each heatmap is given by the expression in each caption. Darker colors indicate higher values.

Proposition 3 states that $\tau = \sqrt{w\rho_1 \hat{p}}$ should be the universal threshold for SVT because this choice of τ guarantees that the MSE of $\hat{A}$ decays at a rate of $o((mn)^{-1})$. As an immediate extension, Corollary 4 states that, if the loss of h when given perfect information A is small, then the loss of f under USVT is also small. In fact, if the loss of h is 0 when A is known, then $\delta_2 = 0$, and $\lim_{m\to\infty} \mathcal{L}(f) = 0$.

Chatterjee [14] also shows that the MSE of USVT is within a constant multiplicative factor and an exponentially small, additive term of the MSE of the minimax estimator, which implies that one cannot do much better than the USVT (cf. Theorem 1.2 in [14]).

The next result establishes a connection between the threshold of USVT and that required for IF in Theorem 1, showing that there are conditions under which there is no performance-fairness trade-off.

Lemma 5. *Suppose that the entries of A are independent Rademacher random variables and each entry is independently observed with probability $p \in [0, 1]$. Let $\rho_1 = \max(m, n) = \beta n$, where $\beta \geq 1$. Then, $\mathbb{P}(2\kappa^{\max} \geq w\rho_1 p) \leq m \exp\left(-2p^2\rho_1(w\beta - 1)^2/\beta\right)$.*

In other words, the minimum threshold required by IF is effectively within constant factor of that proposed by USVT. This finding has two implications. First, when the threshold under USVT is larger than that required by IF, USVT automatically guarantees IF on Z. Second, because USVT is designed to provide strong performance guarantees, as given in Proposition 3, requiring IF on Z imposes little to no performance cost other than that already incurred under USVT. [9]

Performance under IF on A. Recalling Theorem 2, ME is approximately individually fair on A and fully individually fair on A when $||\Pi(Z) - A||_{q,\infty} = 0$. Therefore, the relationship between IF on A and performance under ME is straightforward: the lower the estimation error $\|\Pi(Z) - A\|_{q,\infty}$, the more individually fair f is on A.

5 Experiments

We conducted three experiments using the MovieLens 1M dataset [23]. The first experiment studies whether the recommendations produced by our proposal are individually fair. The second makes a connection to group fairness by examining recommendations by gender. The third illustrates the interpretability of spectral ME methods.

The MovieLens 1M dataset [23] contains 6040 users and 3952 movies. Each entry Z_{ij} gives to user i's rating of movie j if $(i, j) \in \Omega$, and $Z_{ij} = \emptyset$ if user i does not rate movie j. USVT is applied on Z using $\eta = 0$, and $h(i, \hat{A}, \mathbf{c}) = \mathbf{c}^\top \hat{\mathbf{a}}_i$. The resulting prediction $f(i, Z, \mathbf{c})$ denotes user i's preference for the mixture of movies $\mathbf{c}$. Experimental details can be found in the Appendix.

Experiment 1: Individual fairness. Fig. 1 illustrates whether similar users are treated similarly (i.e., whether IF holds). An entry (i, j) is a light color if the norm given in each heatmap's caption is small, and a dark color if large. For example, if entry (i, j) in Fig. 1a is dark, then the rating behaviors of users i and j are very different on average.

[9] Note that although Proposition 3 and Corollary 4 place conditions on A and Ω, these assumptions do not take away from the results on the performance-fairness trade-off. Intuitively, these assumptions arise from the literature on designing ME methods to maximize accuracy. Therefore, to compare performance with and without IF, we must consider the same assumptions used to derive performance guarantees in existing ME works.

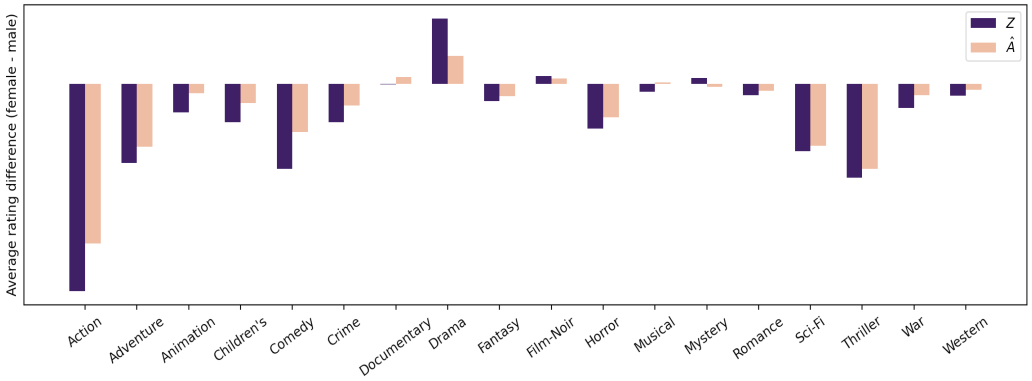
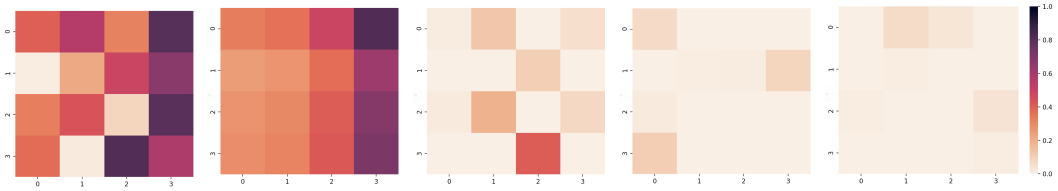


Figure 2: Average difference in ratings by females and males across movie genres, scaled by the frequency of the genres in the observation matrix Z (in blue) and in the matrix estimate $\hat{A}$ (in red).



(a) Original matrix (b) $\sigma_1 \mathbf{u}_1 \mathbf{v}_1^\top$ (c) $\sigma_2 \mathbf{u}_2 \mathbf{v}_2^\top$ (d) $\sigma_3 \mathbf{u}_3 \mathbf{v}_3^\top$ (e) $\sigma_4 \mathbf{u}_4 \mathbf{v}_4^\top$

Figure 3: Visualizing singular value decomposition (SVD). σ_ℓ, $\mathbf{u}_\ell$, and $\mathbf{v}_\ell$ denote the singular values, left singular vectors, and right singular vectors of the 4×4 matrix in (a), where $\sigma_1 \geq \ldots \geq \sigma_4$.

Intuitively, IF holds if the relative coloring in Fig. 1b matches that of Fig. 1a. Fig. 1c and 1d extends this analysis by illustrating the difference $|f(i, Z, \mathbf{c}_k) - f(j, Z, \mathbf{c}_k)|$ between the recommendations that users i and j receive under specific contexts $\mathbf{c}_1$ and $\mathbf{c}_2$. These plots confirm that USVT guarantees IF on Z. They also illustrate that how similarly two users are treated depends on the context $\mathbf{c}$. For example, two users may be similar in every way, except that one likes horror movies and the other does not. Then, the recommendations these users receive on Halloween may be different.

Experiment 2: Group fairness. This experiment studies the question: If female users rate a genre higher than male users on average, does the platform respect this trend? If not, then the method uses spurious information (e.g., assumptions about gender preferences) in its recommendations.

Fig. 2 illustrates that, for USVT, the answer is yes. Visually, the answer is yes when both bars face the same direction. A downward-facing bar implies that males rate the genre more highly than females on average, and vice versa for an upward-facing bar. Interestingly, the trend is flipped in a few genres but only in genres that contain a small number of movies. We also note that the bars for $\hat{A}$ are generally smaller in magnitude than those for Z, indicating that males and females are treated more similarly under ME than if Z is not pre-processed using ME.

Experiment 3: Interpretability of spectral ME. Fig. 3 visualizes the SVD operation—which lies at the core of most spectral ME methods—and gives an intuition for its meaning. Recall that a matrix $M \in \mathbb{R}^{m \times n}$ can be expressed as $M = \sum_{\ell \in [\min(m,n)]} \sigma_\ell \mathbf{u}_\ell \mathbf{v}_\ell^\top$, where $\sigma_1 \geq \ldots \geq \sigma_{\max(m,n)}$.

Fig. 3a visualizes a random 4×4 matrix M in heatmap form such that each entry corresponds to a box. Fig. 3b-3e visualize each component of the matrix's SVD. As expected, the components decrease in "importance". Recall that SVT effectively preserves and re-scales the larger components. Therefore, when M is a noisy matrix, USVT can be interpreted as preserving the main "signal" in a matrix and removing smaller components that are indistinguishable from noise. Fig. 3 also points to the interpretability of spectral methods. In the MovieLens dataset, the singular vectors $\mathbf{v}_1$ and $\mathbf{u}_1$ capture the most common movie preferences and how much each user aligns with these preferences, respectively, while the singular value σ_1 indicates how prevailing these preferences are. One can therefore probe these quantities when assessing how the algorithm generates predictions.

References

[1] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 670–688. IEEE, 2015.

[2] Anish Agarwal, Abdullah Alomar, and Devavrat Shah. On multivariate singular spectrum analysis. *arXiv preprint arXiv:2006.13448*, 2020.

[3] Anish Agarwal, Muhammad Jehangir Amjad, Devavrat Shah, and Dennis Shen. Model agnostic time series analysis via matrix estimation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(3):1–39, 2018.

[4] Muhammad Amjad, Devavrat Shah, and Dennis Shen. Robust synthetic control. *The Journal of Machine Learning Research*, 19(1):802–852, 2018.

[5] Animashree Anandkumar, Rong Ge, Daniel Hsu, and Sham Kakade. A tensor spectral approach to learning mixed membership community models. In *Conference on Learning Theory*, pages 867–881. PMLR, 2013.

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning: Limitations and opportunities, 2018.

[7] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[8] Yahav Bechavod, Christopher Jung, and Zhiwei Steven Wu. Metric-free individual fairness in online learning. *arXiv preprint arXiv:2002.05474*, 2020.

[9] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2212–2220, 2019.

[10] 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*, pages 405–414, 2018.

[11] Christian Borgs, Jennifer Chayes, Christina E Lee, and Devavrat Shah. Thy friend is my friend: Iterative collaborative filtering for sparse matrix estimation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4718–4729, 2017.

[12] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

[13] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[14] Sourav Chatterjee et al. Matrix estimation by universal singular value thresholding. *Annals of Statistics*, 43(1):177–214, 2015.

[15] Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[16] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[17] Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

[18] 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*, pages 214–226, 2012.

[19] Golnoosh Farnadi, Pigi Kouki, Spencer K Thompson, Sriram Srinivasan, and Lise Getoor. A fairness-aware hybrid recommender system. *arXiv preprint arXiv:1809.09030*, 2018.

[20] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921. IEEE, 2020.

[21] Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017.

[22] Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. *arXiv preprint arXiv:1802.06936*, 2018.

[23] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[24] Samuel B Hopkins and David Steurer. Efficient bayesian estimation from few samples: community detection and related problems. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 379–390. IEEE, 2017.

[25] Christina Ilvento. Metric learning for individual fairness. *arXiv preprint arXiv:1906.00250*, 2019.

[26] Rashidul Islam, Kamrun Naher Keya, Ziqian Zeng, Shimei Pan, and James Foulds. Neural fair collaborative filtering. *arXiv preprint arXiv:2009.08955*, 2020.

[27] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Enhancement of the neutrality in recommendation. In *Decisions@ RecSys*, pages 8–14. Citeseer, 2012.

[28] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

[29] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *The Journal of Machine Learning Research*, 11:2057–2078, 2010.

[30] Yehuda Koren. The bellkor solution to the netflix grand prize. *Netflix prize documentation*, 81(2009):1–10, 2009.

[31] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[32] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. Operationalizing individual fairness with pairwise fair representations. *arXiv preprint arXiv:1907.01439*, 2019.

[33] Weiwen Liu and Robin Burke. Personalizing fairness-aware re-ranking. *arXiv preprint arXiv:1809.02921*, 2018.

[34] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

[35] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13:1665–1697, 2012.

[36] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. Fairness in rankings and recommendations: An overview. *arXiv preprint arXiv:2104.05994*, 2021.

[37] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12), 2011.

[38] Jasson DM Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719, 2005.

[39] Dimitris Serbos, Shuyao Qi, Nikos Mamoulis, Evaggelia Pitoura, and Panayiotis Tsaparas. Fairness in package-to-group recommendations. In *Proceedings of the 26th International Conference on World Wide Web*, pages 371–379, 2017.

[40] Devavrat Shah and Christina Lee. Reducing crowdsourcing to graphon estimation, statistically. In *International Conference on Artificial Intelligence and Statistics*, pages 1741–1750. PMLR, 2018.

[41] Dogyoon Song, Christina E Lee, Yihua Li, and Devavrat Shah. Blind regression: Nonparametric regression for latent variable models via collaborative filtering. *Advances in Neural Information Processing Systems*, 29:2155–2163, 2016.

[42] Maria Stratigi, Jyrki Nummenmaa, Evaggelia Pitoura, and Kostas Stefanidis. Fair sequential group recommendations. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 1443–1452, 2020.

[43] Yuzhe Yang, Guo Zhang, Dina Katabi, and Zhi Xu. Me-net: Towards effective adversarial robustness with matrix estimation. *arXiv preprint arXiv:1905.11971*, 2019.

[44] Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. *arXiv preprint arXiv:1705.08804*, 2017.

[45] Ziwei Zhu, Xia Hu, and James Caverlee. Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1153–1162, 2018.

Appendix

A Preliminary results

In order to prove Theorem 1, we first begin with a few helpful results.

Lemma A.1. *Without loss of generality, suppose that* $\mathbf{z}_i^T\mathbf{z}_i \geq \mathbf{z}_j^T\mathbf{z}_j$. *If* $m > 1$, $\tau > \sqrt{\|\mathbf{z}_i\|_2^2 + \|\mathbf{z}_j\|_2^2}$, *and* $\|\mathbf{z}_k\|_2^2 > 0$ *for all* $k \in [m]$,

$$|u_{\ell i} - u_{\ell j}| \leq \frac{(m-1)|\mathbf{u}_\ell|_\infty \max_{k\neq i}\|\mathbf{z}_k\|_2}{\sigma_\ell^2 - \|\mathbf{z}_i\|_2^2 - \|\mathbf{z}_j\|_2^2}\left[\|\mathbf{z}_i - \mathbf{z}_j\|_2 + \left(\frac{\lambda_\ell - \|\mathbf{z}_j\|_2^2}{\lambda_\ell - \|\mathbf{z}_i\|_2^2} - 1\right)\|\mathbf{z}_i\|_2\right].$$

Proof. Since $\mathbf{u}_\ell$ is the ℓ-th left singular vector of Z, it is the ℓ-th eigenvector of ZZ^T. Let $B = ZZ^T$ and b_i be the i-th row of B. Therefore, $B\mathbf{u}_\ell = \lambda_\ell \mathbf{u}_\ell$, where λ_ℓ is the ℓ-th eigenvalue of B. Then,

$$\lambda_\ell u_{\ell i} = \mathbf{b}_i^\top \mathbf{u}_\ell = \sum_{k=1}^m B_{ik}u_{\ell k}.$$

Since $B_{ij} = \mathbf{z}_i^T\mathbf{z}_j$,

$$\lambda_\ell u_{\ell i} = \sum_{k=1}^m \mathbf{z}_i^T\mathbf{z}_k u_{\ell k} = \mathbf{z}_i^T\mathbf{z}_i u_{\ell i} + \sum_{k\neq i}^m \mathbf{z}_i^T\mathbf{z}_k u_{\ell k}$$

$$\implies (\lambda_\ell - \mathbf{z}_i^T\mathbf{z}_i)u_{\ell i} = \sum_{k\neq i}^m \mathbf{z}_i^T\mathbf{z}_k u_{\ell k}.$$

Without loss of generality, suppose that $\mathbf{z}_i^T\mathbf{z}_i \geq \mathbf{z}_j^T\mathbf{z}_j$. Then,

$$\begin{aligned}
u_{\ell i} - u_{\ell j} &= (\lambda_\ell - \mathbf{z}_i^T\mathbf{z}_i)^{-1}\sum_{k\neq i}^m \mathbf{z}_i^T\mathbf{z}_k u_{\ell k} - (\lambda_\ell - \mathbf{z}_j^T\mathbf{z}_j)^{-1}\sum_{k\neq j}^m \mathbf{z}_j^T\mathbf{z}_k u_{\ell k} \\
&= (\lambda_\ell - \mathbf{z}_i^T\mathbf{z}_i)^{-1}\left(\mathbf{z}_i^T\mathbf{z}_j u_{\ell j} + \sum_{k\neq i,j}^m \mathbf{z}_i^T\mathbf{z}_k u_{\ell k}\right) \\
&\quad - (\lambda_\ell - \mathbf{z}_j^T\mathbf{z}_j)^{-1}\left(\mathbf{z}_j^T\mathbf{z}_i u_{\ell i} + \sum_{k\neq i,j}^m \mathbf{z}_j^T\mathbf{z}_k u_{\ell k}\right) \\
&= (\lambda_\ell - \mathbf{z}_j^T\mathbf{z}_j)^{-1}\left(\mathbf{z}_i^T\mathbf{z}_j u_{\ell j} + \sum_{k\neq i,j}^m \mathbf{z}_i^T\mathbf{z}_k u_{\ell k}\right) \\
&\quad + \left((\lambda_\ell - \mathbf{z}_i^T\mathbf{z}_i)^{-1} - (\lambda_\ell - \mathbf{z}_j^T\mathbf{z}_j)^{-1}\right)\left(\mathbf{z}_i^T\mathbf{z}_j u_{\ell j} + \sum_{k\neq i,j}^m \mathbf{z}_i^T\mathbf{z}_k u_{\ell k}\right) \\
&\quad - (\lambda_\ell - \mathbf{z}_j^T\mathbf{z}_j)^{-1}\left(\mathbf{z}_j^T\mathbf{z}_i u_{\ell i} + \sum_{k\neq i,j}^m \mathbf{z}_j^T\mathbf{z}_k u_{\ell k}\right) \\
&= (\lambda_\ell - \mathbf{z}_j^T\mathbf{z}_j)^{-1}\left(\mathbf{z}_i^T\mathbf{z}_j(u_{\ell j} - u_{\ell i}) + \sum_{k\neq i,j}^m (\mathbf{z}_i - \mathbf{z}_j)^T\mathbf{z}_k u_{\ell k}\right) \\
&\quad + \left((\lambda_\ell - \mathbf{z}_i^T\mathbf{z}_i)^{-1} - (\lambda_\ell - \mathbf{z}_j^T\mathbf{z}_j)^{-1}\right)\left(\mathbf{z}_i^T\mathbf{z}_j u_{\ell j} + \sum_{k\neq i,j}^m \mathbf{z}_i^T\mathbf{z}_k u_{\ell k}\right),
\end{aligned}$$

which implies that

$$
\begin{aligned}
u_{\ell i} - u_{\ell j} &= (\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j)^{-1} \left(\mathbf{z}_i^T \mathbf{z}_j (u_{\ell j} - u_{\ell i}) + \sum_{k \neq i,j}^{m} (\mathbf{z}_i - \mathbf{z}_j)^T \mathbf{z}_k u_{\ell k} \right) \\
&\quad + \left((\lambda_\ell - \mathbf{z}_i^T \mathbf{z}_i)^{-1} - (\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j)^{-1} \right) \left(\mathbf{z}_i^T \mathbf{z}_j u_{\ell j} + \sum_{k \neq i,j}^{m} \mathbf{z}_i^T \mathbf{z}_k u_{\ell k} \right) \\
\implies (u_{\ell i} - u_{\ell j})(\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j + \mathbf{z}_i^T \mathbf{z}_j) &= \left(\sum_{k \neq i,j}^{m} (\mathbf{z}_i - \mathbf{z}_j)^T \mathbf{z}_k u_{\ell k} \right) \\
&\quad + \left(\frac{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j}{\lambda_\ell - \mathbf{z}_i^T \mathbf{z}_i} - 1 \right) \left(\mathbf{z}_i^T \mathbf{z}_j u_{\ell j} + \sum_{k \neq i,j}^{m} \mathbf{z}_i^T \mathbf{z}_k u_{\ell k} \right), \\
\implies (u_{\ell i} - u_{\ell j}) &= \frac{1}{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j + \mathbf{z}_i^T \mathbf{z}_j} \left(\sum_{k \neq i,j}^{m} (\mathbf{z}_i - \mathbf{z}_j)^T \mathbf{z}_k u_{\ell k} \right) \\
&\quad + \frac{1}{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j + \mathbf{z}_i^T \mathbf{z}_j} \left(\frac{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j}{\lambda_\ell - \mathbf{z}_i^T \mathbf{z}_i} - 1 \right) \left(\sum_{k \neq i}^{m} \mathbf{z}_i^T \mathbf{z}_k u_{\ell k} \right), \\
\implies (u_{\ell i} - u_{\ell j}) &\leq \frac{1}{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j - \mathbf{z}_i^T \mathbf{z}_i} \Bigg[\left(\sum_{k \neq i,j}^{m} (\mathbf{z}_i - \mathbf{z}_j)^T \mathbf{z}_k u_{\ell k} \right) \\
&\quad + \left(\frac{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j}{\lambda_\ell - \mathbf{z}_i^T \mathbf{z}_i} - 1 \right) \left(\sum_{k \neq i}^{m} \mathbf{z}_i^T \mathbf{z}_k u_{\ell k} \right) \Bigg],
\end{aligned}
$$

where the last line, in which $z_i^\top \mathbf{z}_j$ is switched with $-z_i^\top \mathbf{z}_i$ and the inequality is introduced, holds true if $\lambda_\ell - \mathbf{z}_j^\top \mathbf{z}_j - \mathbf{z}_i^\top \mathbf{z}_i > 0$. Applying the triangle inequality,

$$
\begin{aligned}
\left| \sum_{k \neq i,j}^{m} (\mathbf{z}_i - \mathbf{z}_j)^T \mathbf{z}_k u_{\ell k} \right| &\leq \left| \sum_{k \neq i,j}^{m} (\mathbf{z}_i - \mathbf{z}_j)^T \mathbf{z}_k \cdot \frac{\|\mathbf{z}_i - \mathbf{z}_j\|_2}{\|\mathbf{z}_i - \mathbf{z}_j\|_2} \right| |u_{\ell k}| && (5) \\
&= \left| \sum_{k \neq i,j}^{m} \frac{(\mathbf{z}_i - \mathbf{z}_j)^T}{\|\mathbf{z}_i - \mathbf{z}_j\|_2} \mathbf{z}_k \|\mathbf{z}_i - \mathbf{z}_j\|_2 \right| |u_{\ell k}| && (6) \\
&\leq \left| \sum_{k \neq i,j}^{m} \frac{\mathbf{z}_k^T}{\|\mathbf{z}_k\|_2} \mathbf{z}_k \|\mathbf{z}_i - \mathbf{z}_j\|_2 \right| |u_{\ell k}| && (7) \\
&\leq \sum_{k \neq i,j}^{m} \|\mathbf{z}_k\|_2 \|\mathbf{z}_i - \mathbf{z}_j\|_2 |u_{\ell k}| && (8) \\
&\leq \|\mathbf{z}_i - \mathbf{z}_j\|_2 \sum_{k \neq i,j}^{m} \|\mathbf{z}_k\|_2 |u_{\ell k}| && (9) \\
&\leq \|\mathbf{z}_i - \mathbf{z}_j\|_2 (m-2) \max_{k \neq i,j} \|\mathbf{z}_k\|_2 |\mathbf{u}_\ell|_\infty && (10)
\end{aligned}
$$

which implies that, when $\lambda_\ell - \mathbf{z}_j^\top \mathbf{z}_j - \mathbf{z}_i^\top \mathbf{z}_i > 0$,

$$
\begin{aligned}
|u_{\ell i} - u_{\ell j}| &\leq \frac{1}{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j - \mathbf{z}_i^T \mathbf{z}_i} \Bigg[\|\mathbf{z}_i - \mathbf{z}_j\|_2 (m-2) \max_{k \neq i,j} \|\mathbf{z}_k\|_2 |\mathbf{u}_\ell|_\infty \\
&\qquad + \left(\frac{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j}{\lambda_\ell - \mathbf{z}_i^T \mathbf{z}_i} - 1 \right) \sum_{k \neq i}^{m} |\mathbf{z}_i^T \mathbf{z}_k u_{\ell k}| \Bigg], \\
&\leq \frac{1}{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j - \mathbf{z}_i^T \mathbf{z}_i} \Bigg[\|\mathbf{z}_i - \mathbf{z}_j\|_2 (m-2) \max_{k \neq i,j} \|\mathbf{z}_k\|_2 |\mathbf{u}_\ell|_\infty \\
&\qquad + \left(\frac{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j}{\lambda_\ell - \mathbf{z}_i^T \mathbf{z}_i} - 1 \right) \sum_{k \neq i}^{m} |\mathbf{z}_i^T \mathbf{z}_k| |u_{\ell k}| \Bigg], \qquad (11) \\
&\leq \frac{1}{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j - \mathbf{z}_i^T \mathbf{z}_i} \Bigg[\|\mathbf{z}_i - \mathbf{z}_j\|_2 (m-2) \max_{k \neq i,j} \|\mathbf{z}_k\|_2 |\mathbf{u}_\ell|_\infty \\
&\qquad + \left(\frac{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j}{\lambda_\ell - \mathbf{z}_i^T \mathbf{z}_i} - 1 \right) \|\mathbf{z}_i\|_2 |\mathbf{u}_\ell|_\infty \sum_{k \neq i}^{m} \|\mathbf{z}_k\|_2 \Bigg], \qquad (12) \\
&\leq \frac{1}{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j - \mathbf{z}_i^T \mathbf{z}_i} \Bigg[\|\mathbf{z}_i - \mathbf{z}_j\|_2 (m-2) \max_{k \neq i,j} \|\mathbf{z}_k\|_2 |\mathbf{u}_\ell|_\infty \\
&\qquad + \left(\frac{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j}{\lambda_\ell - \mathbf{z}_i^T \mathbf{z}_i} - 1 \right) \|\mathbf{z}_i\|_2 |\mathbf{u}_\ell|_\infty (m-1) \max_{k \neq i} \|\mathbf{z}_k\|_2 \Bigg], \qquad (13) \\
&\leq \frac{(m-1)|\mathbf{u}_\ell|_\infty \max_{k \neq i} \|\mathbf{z}_k\|_2}{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j - \mathbf{z}_i^T \mathbf{z}_i} \left[\|\mathbf{z}_i - \mathbf{z}_j\|_2 + \left(\frac{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j}{\lambda_\ell - \mathbf{z}_i^T \mathbf{z}_i} - 1 \right) \|\mathbf{z}_i\|_2 \right], \qquad (14) \\
&= \frac{(m-1)|\mathbf{u}_\ell|_\infty \max_{k \neq i} \|\mathbf{z}_k\|_2}{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j - \mathbf{z}_i^T \mathbf{z}_i} \left[\|\mathbf{z}_i - \mathbf{z}_j\|_2 + \left(\frac{\lambda_\ell - \mathbf{z}_j^T \mathbf{z}_j}{\lambda_\ell - \mathbf{z}_i^T \mathbf{z}_i} - 1 \right) \|\mathbf{z}_i\|_2 \right], \qquad (15)
\end{aligned}
$$

which concludes the proof by noticing that $\lambda_\ell = \sigma_\ell^2$. □

Lemma A.2. *Suppose* $\hat{A} = \Pi_\alpha^\tau(Z)$. *Then,*

$$
|\hat{A}_{iq} - \hat{A}_{jq}| \leq \sum_{\ell \in S(\tau)} \alpha(\sigma_\ell) |u_{\ell i} - u_{\ell j}| |\mathbf{v}_\ell|_\infty. \qquad (16)
$$

Proof. Recall the SVT method from Algorithm 1. $\hat{A}$ can be written as:

$$
\hat{A} = \sum_{\ell \in S(\tau)} \alpha(\sigma_\ell) \mathbf{u}_\ell \mathbf{v}_\ell^T.
$$

The element (i, q) of $\hat{A}$ can be written as $\hat{a}_{iq} = \sum_{\ell \in S(\tau)} \alpha(\sigma_\ell) u_{\ell i} v_{\ell q}$, which implies that

$$
\begin{aligned}
|\hat{a}_{iq} - \hat{a}_{jq}| &= \left| \sum_{\ell \in S(\tau)} \alpha(\sigma_\ell)(u_{\ell i} - u_{\ell j}) v_{\ell q} \right|_1 \\
&\leq \sum_{\ell \in S(\tau)} |\alpha(\sigma_\ell)(u_{\ell i} - u_{\ell j}) v_{\ell q}| \\
&\leq \sum_{\ell \in S(\tau)} \alpha(\sigma_\ell) |u_{\ell i} - u_{\ell j}| |v_{\ell q}| \qquad (17) \\
&\leq \sum_{\ell \in S(\tau)} \alpha(\sigma_\ell) |u_{\ell i} - u_{\ell j}| |\mathbf{v}_\ell|_\infty. \qquad (18)
\end{aligned}
$$

□

B Proof of Theorem 1

Theorem 1. *Suppose* $D(h(i,\hat{A},\mathbf{c}),h(j,\hat{A},\mathbf{c})) \leq L_1 \left\|\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j\right\|_1$ *for all* $i,j \in [m]$ *and* $h \in \mathcal{H}$. *Suppose* $m > 1$ *and* $\|\mathbf{z}_k\|_2^2 > 0$ *for all* $k \in [m]$. *Then, for any* $f \in \mathcal{F}(\mathcal{H},\tau,\alpha)$ *where* $\tau \geq \sqrt{2\kappa^{\max}}$,

$$D(f(i,Z,\mathbf{c}),f(j,Z,\mathbf{c})) \leq nL_1K_1 \left\|\mathbf{z}_i - \mathbf{z}_i\right\|_2 + nL_1K_1K_2, \tag{19}$$

for all $i,j \in [m]$. *If* $h(k,\hat{A},\mathbf{c}) = \hat{\mathbf{a}}_k^\top \mathbf{c}$, *then*

$$|f(i,Z,\mathbf{c}) - f(j,Z,\mathbf{c})| \leq \|\mathbf{c}\|_1 L_1K_1 \left\|\mathbf{z}_i - \mathbf{z}_i\right\|_2 + \|\mathbf{c}\|_1 L_1K_1K_2. \tag{20}$$

for all $i,j \in [m]$. *If* $\|\mathbf{z}_i\|_2^2 = \|\mathbf{z}_j\|_2^2$, *the same results hold with* $K_2 = 0$.

Proof. In order to prove (19), note that $|\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j|_1 = \sum_{q=1}^n |\hat{a}_{iq} - \hat{a}_{jq,}|$. From Lemma A.2,

$$|\hat{A}_{iq} - \hat{A}_{jq}| \leq \sum_{\ell \in S(\tau)} \alpha(\sigma_\ell)|u_{\ell i} - u_{\ell j}||\mathbf{v}_\ell|_\infty. \tag{21}$$

Without loss of generality, suppose that $\mathbf{z}_i^T\mathbf{z}_i \geq \mathbf{z}_j^T\mathbf{z}_j$. By the conditions of the theorem, $\|\mathbf{z}_k\|_2^2 > 0$ for all $k \in [m]$, and $m > 1$. We can apply Lemma A.1, which gives:

$$\begin{aligned}
|\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j|_1 &= \sum_{q=1}^n |\hat{A}_{iq} - \hat{A}_{jq,}| \\
&\leq \sum_{q=1}^n \sum_{\ell \in S(\tau)} \alpha(\sigma_\ell)|u_{\ell i} - u_{\ell j}||\mathbf{v}_\ell|_\infty \\
&\leq n(m-1)\max_{k \neq i} \|\mathbf{z}_k\|_2 \sum_{\ell \in S(\tau)} \frac{\alpha(\sigma_\ell)|\mathbf{u}_\ell|_\infty|\mathbf{v}_\ell|_\infty}{\sigma_\ell^2 - \|\mathbf{z}_i\|_2^2 - \|\mathbf{z}_j\|_2^2} \\
&\qquad \left[\left\|\mathbf{z}_i - \mathbf{z}_j\right\|_2 + \left(\frac{\lambda_\ell - \|\mathbf{z}_j\|_2^2}{\lambda_\ell - \|\mathbf{z}_i\|_2^2} - 1\right)\|\mathbf{z}_i\|_2\right]
\end{aligned} \tag{22}$$

as long as $\sigma_\ell > \sqrt{\|\mathbf{z}_i\|_2^2 + \|\mathbf{z}_j\|_2^2}$. Since $D(h(i,\hat{A},\mathbf{c}),h(j,\hat{A},\mathbf{c})) \leq L'|\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j|_1$ for all $h \in \mathcal{H}$, $D(f(i,Z,\mathbf{c}),f(j,Z,\mathbf{c})) \leq L'|\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j|_1$ for all $f \in \mathcal{F}(\mathcal{H},\cdot,\cdot)$. Combining this fact with the observations that $\kappa^{\max} \geq \max_{k \neq i}\|\mathbf{z}_k\|_2^2$ and $\kappa^{\max} \geq \|\mathbf{z}_i\|_2^2, \|\mathbf{z}_j\|_2^2$ gives the upper bound (19) as long as $\tau \geq \sqrt{2\kappa^{\max}}$, where the requirement on τ is inherited from Lemma A.1.

Using the same logic as Lemma A.2 to prove (20),

$$\begin{aligned}
|\hat{\mathbf{a}}_i^T\mathbf{c} - \hat{\mathbf{a}}_j^T\mathbf{c}| &\leq \sum_{\ell \in S(\tau)} \alpha(\sigma_\ell)|u_{\ell i} - u_{\ell j}||\mathbf{v}_\ell^T\mathbf{c}| &(23)\\
&\leq |\mathbf{c}|_1 \sum_{\ell \in S(\tau)} \alpha(\sigma_\ell)|u_{\ell i} - u_{\ell j}||\mathbf{v}_\ell|_\infty &(24)
\end{aligned}$$

Without loss of generality, suppose that $\mathbf{z}_i^T\mathbf{z}_i \geq \mathbf{z}_j^T\mathbf{z}_j$. By the conditions of the theorem, $\|\mathbf{z}_k\|_2^2 > 0$ for all $k \in [m]$, and $m > 1$. We can apply Lemma A.1, which gives:

$$\begin{aligned}
&|\hat{\mathbf{a}}_i^T\mathbf{c} - \hat{\mathbf{a}}_j^T\mathbf{c}| \\
&\leq (m-1)|\mathbf{c}|_1 \max_{k \neq i}\|\mathbf{z}_k\|_2 \sum_{\ell \in S(\tau)} \frac{\alpha(\sigma_\ell)|\mathbf{u}_\ell|_\infty|\mathbf{v}_\ell|_\infty}{\sigma_\ell^2 - \|\mathbf{z}_i\|_2^2 - \|\mathbf{z}_j\|_2^2} \\
&\qquad \left[\left\|\mathbf{z}_i - \mathbf{z}_j\right\|_2 + \left(\frac{\lambda_\ell - \|\mathbf{z}_j\|_2^2}{\lambda_\ell - \|\mathbf{z}_i\|_2^2} - 1\right)\|\mathbf{z}_i\|_2\right] \\
&\leq (m-1)|\mathbf{c}|_1 \max_{k \neq i}\|\mathbf{z}_k\|_2 \sum_{\ell \in S(\tau)} \frac{\alpha(\sigma_\ell)|\mathbf{u}_\ell|_\infty|\mathbf{v}_\ell|_\infty}{\sigma_\ell^2 - \|\mathbf{z}_i\|_2^2 - \|\mathbf{z}_j\|_2^2}\left\|\mathbf{z}_i - \mathbf{z}_j\right\|_2 \\
&\qquad + (m-1)|\mathbf{c}|_1 \max_{k \neq i}\|\mathbf{z}_k\|_2 \sum_{\ell \in S(\tau)} \|\mathbf{z}_i\|_2 \frac{\alpha(\sigma_\ell)|\mathbf{u}_\ell|_\infty|\mathbf{v}_\ell|_\infty}{\sigma_\ell^2 - \|\mathbf{z}_i\|_2^2 - \|\mathbf{z}_j\|_2^2}\left(\frac{\lambda_\ell - \|\mathbf{z}_j\|_2^2}{\lambda_\ell - \|\mathbf{z}_i\|_2^2} - 1\right)
\end{aligned} \tag{25}$$

as long as $\sigma_\ell > \sqrt{\|\mathbf{z}_i\|_2^2 + \|\mathbf{z}_j\|_2^2}$. (25) gives (20) immediately by noticing that $\mathbf{z}_{k2}^2$ and $\kappa^{\max} \geq \|\mathbf{z}_i\|_2^2, \|\mathbf{z}_j\|_2^2$.

Finally, notice that when $\|\mathbf{z}_i\|_2^2 = \|\mathbf{z}_j\|_2^2$, the right-most terms in (19) and (25) are 0, which gives the final statement in Theorem 1 that exact IF for (20) holds when $\|\mathbf{z}_i\|_2^2 = \|\mathbf{z}_j\|_2^2$. □

In the main text, we provide an interpretation of Theorem 1. We study the average-case behavior of the IF Lipschitz constant $\|\mathbf{c}\|_1 L_1 K_1$ by studying the rate at which K_1 grows under common conditions. For example, the condition that for every row in A, $\lfloor np \rfloor$ of its entries are randomly observed, and the observations are binary such that $Z_{ij} \in \{-1, +1\}$ for $(i, j) \in \Omega$ can be mapped to the recommendation setting by interpreting $+1$ as a thumbs-up rating and a -1 as a thumbs-down rating. Moreover, condition that $\|\mathbf{c}\|_1 = 1$ simply requires that the context vector is normalized across contexts; this condition can be dropped without affecting our understanding of Theorem 1 because $\mathbf{c}$ appears on both sides of the inequality in Theorem 1. The incoherence condition is a standard assumption in matrix completion that characterizes a matrix's singular vectors.

By the analysis under Theorem 1, we find that $K_1 = \Theta(r^2\sqrt{mp}/n)$. Note that the probability that one entry is observed in $\mathbf{z}_i$ but not observed in $\mathbf{z}_j$ is $2(1-p)p$, and the probability that both are observed is p^2. Additionally, observe that whenever the same entry in both $\mathbf{z}_i$ and $\mathbf{z}_j$ is not observed, then they have the same value, and when only one is observed, they must have different values. Since we are interested in how large the upper bound grows, when the same entry in $\mathbf{z}_i$ and $\mathbf{z}_j$ is observed, we assume it takes a non-zero value. As such, $\|\mathbf{z}_i - \mathbf{z}_j\|_2$ grows at a rate of approximately $\Theta(\sqrt{np}$, which implies that $\|\mathbf{c}\|_1 K_1 \|\mathbf{z}_i - \mathbf{z}_j\|_2 = \Theta(r^2\sqrt{mp^2/n})$.

C Proof of Theorem 2

Theorem 2. *Suppose $D(h(i, \hat{A}, \mathbf{c}), h(j, \hat{A}, \mathbf{c})) \leq L_1 \|\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j\|_q$ for all $i, j \in [m]$ and $h \in \mathcal{H}$. Then, for all $i, j \in [m]$,*

$$D(f(i, Z, \mathbf{c}), f(j, Z, \mathbf{c})) \leq L_1 \|\mathbf{a}_i - \mathbf{a}_j\|_q + 2L_1||\hat{A} - A||_{q,\infty}.$$

Proof. Recall that $\|M\|_{q,\infty} = \max_i \|\mathbf{m}_i\|_q$. This result follows from the application of the triangle inequality.

$$\begin{aligned}
D(f(i, Z, \mathbf{c}), f(j, Z, \mathbf{c})) &= D(h(i, \hat{A}, \mathbf{c}), h(j, \hat{A}, \mathbf{c})) \\
&\leq L_1|\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j|_q \\
&\leq L_1(|\hat{\mathbf{a}}_i - \mathbf{a}_i|_q + |\hat{\mathbf{a}}_j - \mathbf{a}_j|_q + |\mathbf{a}_i - \mathbf{a}_j|_q) \\
&\leq L_1(2||\hat{A} - A||_{q,\infty} + |\mathbf{a}_i - \mathbf{a}_j|_q) \\
&\leq L_1|\mathbf{a}_i - \mathbf{a}_j|_q + 2L_1||\hat{A} - A||_{q,\infty},
\end{aligned}$$

which gives the result as stated. □

D Results from Section 4.3

Proposition 3 follows directly from Theorem 1.1. in Chatterjee [14] with small modifications (e.g., to ensure that the notation is consistent with the notation in this work).

Below, we prove Corollary 4 and Lemma 5.

Corollary 3. *Suppose $f = h \circ \Pi_\alpha^\tau$ and $\mathcal{L}(f) \leq L_2 MSE(\Pi_\alpha^\tau(Z)) + \delta_2$, where $L_2 > 0$ and $\delta_2 \geq 0$. Then, under the same conditions as those in Proposition 3, $\lim_{m\to\infty} \mathcal{L}(f) = \delta_2$.*

Proof. Recall Proposition 3. Since $\rho_1 \to \infty$ as $m \to \infty$, $\text{MSE}(\Pi_\alpha^\tau(Z)) \to 0$ as $m \to \infty$, and the limit in Corollary 4 follows directly. □

Lemma 4. *Suppose that the entries of A are independent Rademacher random variables and each entry is independently observed with probability $p \in [0, 1]$. Let $\rho_1 = \max(m, n) = \beta n$, where $\beta \geq 1$. Then, $\mathbb{P}(2\kappa^{\max} \geq w\rho_1 p) \leq m \exp\left(-2p^2\rho_1(w\beta - 1)^2/\beta\right)$.*

Proof. Note that, under the stated conditions, $\kappa^{\max}$ is equivalent to the maximum of m random independent binomial random variables $\text{Bin}(n, p)$, which implies that:

$$\mathbb{P}(\kappa^{\max} \geq w\rho_1 p) \leq m\mathbb{P}(B \geq w\rho_1 p),$$

where $B \sim \text{Bin}(n, p)$. Let $\rho_1 = \beta n$, where $\beta \geq 1$. Applying Hoeffding's inequality gives:

$$\begin{aligned}
\mathbb{P}(\kappa^{\max} \geq w\rho_1 p) &\leq \exp(-2n((1-p) - (n - w\rho_1 p)/n)^2) \\
&\leq \exp\left(-2n\left(\frac{n - np - n + w\rho_1 p}{n}\right)^2\right) \\
&\leq \exp\left(-\frac{2p^2(w\rho_1 - n)^2}{n}\right) \\
&\leq \exp\left(-2p^2 n(w\beta - 1)^2\right).
\end{aligned}$$

□

E Experimental Setup

Below are some additional details about the experimental setup. The accompanying code can be found in the Supplementary Material. All experiments were run on a personal laptop with a 1.4 GHz Quad-Core Intel Core i5 and 8 GB of memory. USVT, Experiment 1, or Experiment 3 run in under 10 seconds. Experiment 2 runs over about 6 minutes.

E.1 Applying USVT to MovieLens dataset

All experiments were conducted using the MovieLens 1M dataset [23]. Each entry in the MovieLens dataset corresponds to a user, a movie, and a rating the user provided for that movie. The dataset contains $m = 6040$ users and $n = 3952$ movies. The ratings are integers between 1 and 5, inclusive. Not all user-movie pairs have ratings.

For this work, the entries are placed in an $m \times n$ matrix Z. Recall that $A \in [-1, 1]^{m \times n}$ is an unknown ground truth matrix, where each entry (i, j) corresponds to the unknown interest user i has in movie j. Recall further that the observations Z are therefore a noisy, subsample of A. If there is a record of user i rating movie j a score of k, then $Z_{ij} = k$. Otherwise $Z_{ij} = \emptyset$. The index set of non-empty entries in Z form the set Ω referenced in Section 3. The entries Z_{ij} where $(i, j) \in \Omega$ are then normalized:

$$Z_{ij} \longrightarrow \frac{Z_{ij} - \bar{Z}}{Z_{\max} - Z_{\min}}$$

where $\bar{Z}$ is the mean of the non-empty entries, $Z_{\max}$ is the maximum value of non-empty entries and $Z_{\min}$ is the minimum value of non-empty entries in Z. Any entries for which $Z_{ij} = \emptyset$ (i.e., when $(i, j) \notin \Omega$), these entries are set to $Z_{ij} = \bar{Z}$. 80 percent of the entries in Z that belong to Ω are then selected at random and stored in the training set $Z_{\text{train}} \in \mathcal{Z}^{m \times n}$, and the remaining 20 percent are stored in the test set $Z_{\text{test}} \in \mathcal{Z}^{m \times n}$, where $\mathcal{Z} = [-1, 1] \cup \{\emptyset\}$.

The USVT algorithm (as presented in Section 4.3) is run on Z_{train} to produce the estimate $\hat{A}$. Recall that the USVT threshold is $\tau = \sqrt{w\rho_1 \hat{p}}$ from Proposition 3, where $\hat{p}$ is the proportion of values in Z_{train} that are observed ratings, $w = (2 + \eta)^2$ for $\eta \in (0, 1)$, and $\rho_1 = \max(m, n)$. For our experiments, we tested values of $\eta = \{0.01, 0.03, 0.05, 0.1, 0.15, 0.2\}$. In the results presented in Section 5, $\eta = 0.01$, which generally gave the lowest MSE when comparing $\hat{A}_{ij}$ to Z_{ij} for $(i, j) \notin \Omega$.

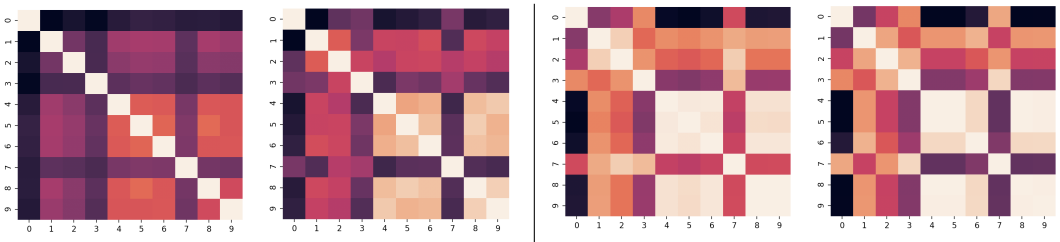
Note that our code differs from the precise USVT algorithm in two ways. First, instead of using $\hat{p}$, we use a common convention. We replaced $\hat{p}$ with an approximation of the variance of entries. In the literature, this convention is justified by observing that , when the absolute value of the entries in A and Z are bounded above by 1 and the entries of Z are i.i.d. Gaussian random variables centered at A observed with probability p, that the variance of each entry in Z is bounded above by p. Since the variance of entries in Z is unknown, we estimate the variance using the empirical variance scaled by a small constant factor (although the constant factor is not necessary and could be equivalently achieved by choosing a larger η). Second, we let $\alpha(x) = x$ instead of $\alpha(x) = x/\hat{p}$. Due to the fact that the data is very sparse, we found that normalizing by $1/\hat{p}$ resulted in very large values in $\hat{A}$.

E.2 Experiment 1: Individual fairness

The following are details about the experiment used to generate Fig. 1. To visualize IF properties of USVT, we produced heatmaps comparing the observed matrix Z, the estimated matrix $\hat{A}$, and the matrices $\hat{A}\mathbf{c}_1$ and $\hat{A}\mathbf{c}_2$ containing the recommendations that users receive under specific contexts $\mathbf{c}_1$ and $\mathbf{c}_2$. The context vectors are chosen to be $\mathbf{c}_1 = \mathbf{v}_1$ and $\mathbf{c}_2 = \mathbf{v}_2$, which are the first and second right-singular vectors, respectively.

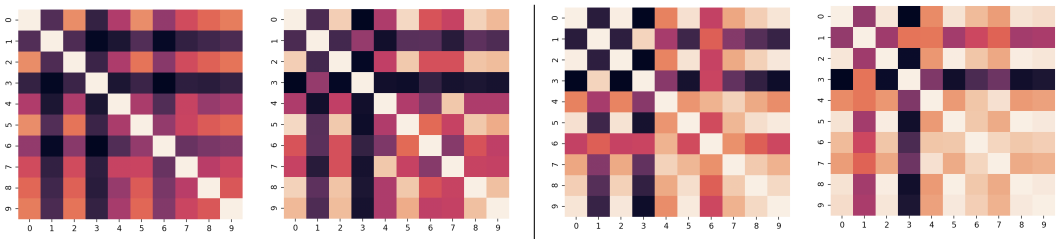
In Fig. 1, heatmaps for visualizing $\|\mathbf{z}_i - \mathbf{z}_j\|_2$, $\|\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j\|_2$, $|(\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j)^\top \mathbf{c}_1|$, $|(\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j)^\top \mathbf{c}_2|$ are displayed for 10 users selected uniformly at random. Each entry (i, j) in the heatmap is given by the corresponding expression—as a result, the heatmaps are symmetric. As explained in Section 5, similar relative coloring between heatmap (a) and the others indicates the level of IF on Z under the given contexts (or, in the case of (b), overall).

Further examples of Fig. 1 are provided below. They are generated using the process described above, and they differ in that the users displayed in each figure.



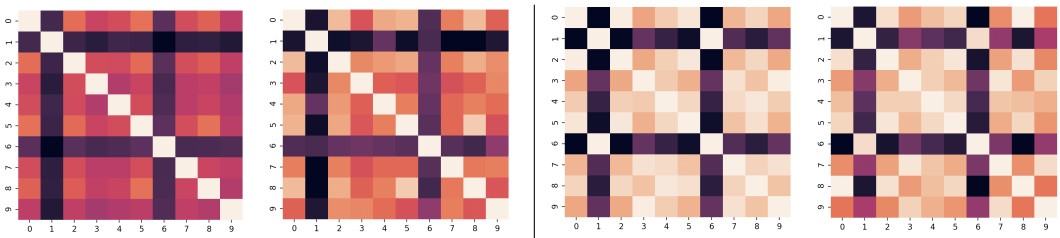
(a) $\|\mathbf{z}_i - \mathbf{z}_j\|_2$ (b) $\|\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j\|_2$ (c) $|(\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j)^\top \mathbf{c}_1|$ (d) $|(\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j)^\top \mathbf{c}_2|$

Figure 4: Visualizing IF of USVT.



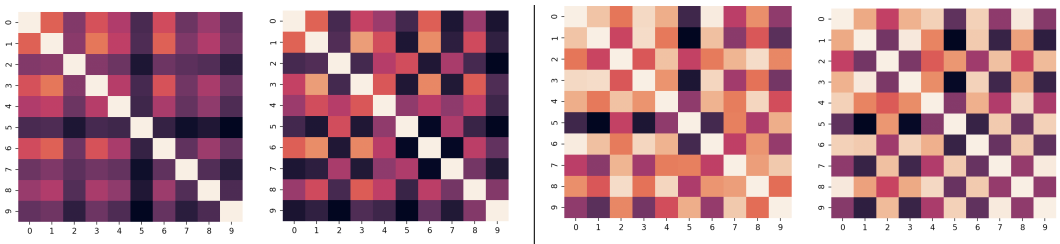
(a) $\|\mathbf{z}_i - \mathbf{z}_j\|_2$ (b) $\|\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j\|_2$ (c) $|(\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j)^\top \mathbf{c}_1|$ (d) $|(\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j)^\top \mathbf{c}_2|$

Figure 5: Visualizing IF of USVT.



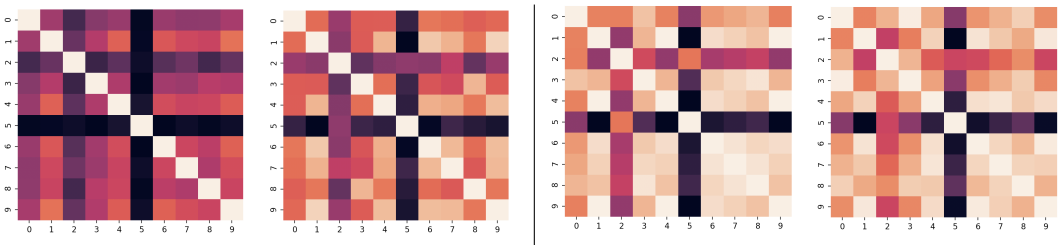
(a) $\|\mathbf{z}_i - \mathbf{z}_j\|_2$ (b) $\|\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j\|_2$ (c) $|(\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j)^\top \mathbf{c}_1|$ (d) $|(\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j)^\top \mathbf{c}_2|$

Figure 6: Visualizing IF of USVT.



(a) $\|\mathbf{z}_i - \mathbf{z}_j\|_2$ (b) $\|\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j\|_2$ (c) $|(\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j)^\top \mathbf{c}_1|$ (d) $|(\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j)^\top \mathbf{c}_2|$

Figure 7: Visualizing IF of USVT.



(a) $\|\mathbf{z}_i - \mathbf{z}_j\|_2$ (b) $\|\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j\|_2$ (c) $|(\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j)^\top \mathbf{c}_1|$ (d) $|(\hat{\mathbf{a}}_i - \hat{\mathbf{a}}_j)^\top \mathbf{c}_2|$

Figure 8: Visualizing IF of USVT.

E.3 Experiment 2: Sensitive attributes

The following are details about the experiment used to generate Fig. 2. To understand the properties of USVT with respect to demographic groups, we compared the average difference in ratings by females and males across movie genres in Z and $\hat{A}$.. This result comparison would illustrate how heavily the recommendation algorithm relies on gender rather than on rating behavior.

The average ratings by females for each movie genre and the average ratings by males for each movie genre are calculated separately over all entries Z_{ij}, and similarly over all entries $\hat{A}_{ij}$, where i ranges from 1 to m and j ranges from 1 to n. The average rating by males in each genre is then subtracted from the average rating by females in each genre, and each of the differences is then scaled by the number of movies (out of n total movies) that are of the given genre. The weighted differences are then displayed in a barplot, as seen in Fig. 2. Recall that USVT does not use information about an individual's gender, which implies that the result is acquired naturally, i.e., without a fairness intervention.

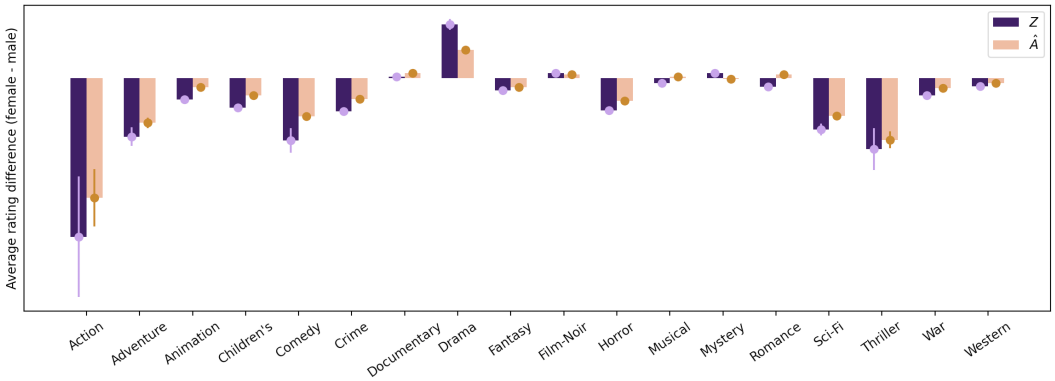


Figure 9: Average difference (with error bars) in ratings by females and males across movie genres, scaled by the frequency of the genres in the observation matrix Z (in purple) and in the matrix estimate $\hat{A}$ (in orange).

Fig. 9 visualizes the same experiment, except with error bars. To obtain error bars, the users are split into groups of 250 (24 groups). The same statistic as that shown in Fig. 2 is obtained for each group. The bar plots mark the mean of these statistics, and the error bars give $+/-$ one standard deviation.

E.4 Experiment 3: Spectral matrix methods

The procedure for generating Fig. 3 is straightforward. Randomly generate an $m \times n$ matrix M. The entries could, for example, be uniformly sampled between -1 and 1 or from a joint Gaussian distribution. Decompose M using an SVD operation, which can be used to express M as $M = \sum_{\ell \in [\min(m,n)]} \sigma_\ell \mathbf{u}_\ell \mathbf{v}_\ell^\top$, where $\sigma_1 \geq \ldots \geq \sigma_{\max(m,n)}$. Based on the SVD, one can then play around with the singular values and vectors to generate plots such as those given in Fig. 3.