Supplementary Material for "Preference Completion: Large-scale Collaborative Ranking from Pairwise Comparisons"

A. Proof of Theorem 3.1

We write $L(X)$ for the function being optimized; i.e.,

$$L(X) = \sum_{(i,j,k)\in\Omega} \mathcal{L}(Y_{i,j,k}(X_{i,j} - X_{i,k})).$$

Note that for any fixed X, $\mathbb{P}_{X^*}L(X) = mR(X)$ (where $\mathbb{P}_{X^*}$ denotes the expectation taken with respect to future samples from $\mathbb{P}_{X^*}$, as distinct from $\mathbb{E}$ which denotes the expectation over the samples used to generate $\hat{X}$). Let K be the set of $d_1 \times d_2$ matrices with nuclear norm at most 1. The proof of Theorem 3.1 proceeds in three main steps.

1. By some algebraic of manipulations L, we reduce the problem to showing a uniform law of large numbers for the family of functions $\{L(X) : X \in \sqrt{\lambda d_1 d_2} K\}$.
2. Using symmetrization and duality properties of K, we reduce the problem to bounding the norm of a matrix M whose entries are sums of random signs.
3. We bound the norm of M using various concentration inequalities and a theorem of Seginer (Seginer, 2000).

Since $\hat{X}$, by definition, minimizes $L(\hat{X})$, for any $\tilde{X} \in \sqrt{\lambda d_1 d_2} K$ we can bound

$$\begin{aligned}\mathbb{P}_{X^*}[L(\hat{X}) - L(\tilde{X})] &\le \mathbb{P}_{X^*}[L(\hat{X})] - L(\hat{X}) - \Big(\mathbb{P}_{X^*}[L(\tilde{X})] - L(\tilde{X})\Big)\\ &\le 2 \sup_{X\in\sqrt{\lambda d_1 d_2}k} |\mathbb{P}_{X^*}L(X) - L(X)|.\end{aligned}$$

In other words, it suffices to show a uniform law of large numbers for $\{L(X) : X \in \sqrt{\lambda d_1 d_2} K\}$.

Let $\epsilon_{i,j,k}$ be i.i.d. ± 1-valued variables and let $\xi_{i,j,k}$ be the indicator that $(i,j,k) \in \Omega$. By Giné-Zinn's symmetrization (as in (Davenport et al., 2013)),

$$\begin{aligned}&\mathbb{E} \sup_{X\in\sqrt{\lambda d_1 d_2}K} |\mathbb{P}_{X^*}L(X) - L(X)|\\ &\le 2\mathbb{E} \sup_{X\in\sqrt{\lambda d_1 d_2}K} \left| \sum_{i,j,k\in\Omega} \epsilon_{i,j,k}\mathcal{L}(Y_{i,j,k}(X_{i,j} - X_{i,k}))\right|.\end{aligned}$$

Since $\mathcal{L}$ is 1-Lipschitz, we obtain

$$\begin{aligned}\mathbb{E} \sup_{X\in\sqrt{\lambda d_1 d_2}K} |\mathbb{P}_{X^*}[L(X)] - L(X)| &\le 2\mathbb{E} \sup_{X\in\sqrt{\lambda d_1 d_2}K} \left| \sum_{i,j,k\in\Omega} \epsilon_{i,j,k} Y_{i,j,k}(X_{i,j} - X_{i,k})\right|\\ &= 2\mathbb{E} \sup_{X\in\sqrt{\lambda d_1 d_2}K} \left| \sum_{i,j,k} \xi_{i,j,k}\epsilon_{i,j,k}(X_{i,j} - X_{i,k})\right|,\end{aligned}$$

where in the last line, we recognized that $\epsilon_{i,j,k} Y_{i,j,k}$ has the same distribution as $\epsilon_{i,j,k}$. Now, let M denote the matrix where $M_{ij} = \sum_k (\xi_{i,j,k} \epsilon_{i,j,k} - \xi_{i,k,j} \epsilon_{i,k,j})$. Then

$$\sum_{i,j,k} \xi_{i,j,k} \epsilon_{i,j,k} (X_{i,j} - X_{i,k}) = \mathrm{tr}(M^T X)$$

and so

$$\sup_{X \in \sqrt{\lambda d_1 d_2} K} \sum_{i,j,k} \xi_{i,j,k} \epsilon_{i,j,k} (X_{i,j} - X_{i,k}) = \sup_{X \in \sqrt{\lambda d_1 d_2} K} \mathrm{tr}(M^T X) = \sqrt{\lambda d_1 d_2} \|M\|.$$

Putting everything together, we have (for any $\tilde{X} \in \sqrt{\lambda d_1 d_2} K$)

$$\mathbb{E} \left[\mathbb{P}_{X^*}[L(\hat{X})] - \mathbb{P}_{X^*}[L(\tilde{X})] \right] \leq 4 \sqrt{\lambda d_1 d_2} \mathbb{E} \|M\|.$$

Together with the following lemma (which we prove in Appendix B), this completes the proof of Theorem 3.1

Lemma A.1. *With* $p = \frac{m}{d_1 d_2}$,

$$\mathbb{E}\|M\| \leq C \kappa \sqrt{p(d_1 + d_2)} \log(d_1 d_2).$$

B. Proof of Lemma A.1

We will decompose M into two parts, $M = M^{(1)} - M^{(2)}$, with

$$M^{(1)}_{ij} = \sum_{k \neq j} \xi_{i,j,k} \epsilon_{i,j,k}$$
$$M^{(2)}_{ij} = \sum_{k \neq j} \xi_{i,k,j} \epsilon_{i,k,j}.$$

Then $\|M\| \leq \|M^{(1)}\| + \|M^{(2)}\|$. Since $M^{(1)}$ and $M^{(2)}$ have the same distribution,

$$\mathbb{E}\|M\| \leq 2 \mathbb{E}\|M^{(1)}\|,$$

and so we are reduced to studying $M^{(1)}$, which has i.i.d. entries. Now, we apply Seginer's theorem (Seginer, 2000):

$$\mathbb{E}\|M^{(1)}\| \leq C \left(\mathbb{E} \max_i \|M^{(1)}_{i*}\|_2 + \mathbb{E} \max_j \|M^{(1)}_{*j}\|_2 \right), \tag{1}$$

where $M^{(1)}_{i*}$ denotes the ith row of $M^{(1)}$ and $M^{(1)}_{*j}$ denotes the jth column, and $\| \cdot \|_2$ denotes the Euclidean norm.

We will separate the task of bounding $\mathbb{E} \max_i \|M^{(1)}_{i*}\|_2$ into two parts: if $\|x\|_0$ denotes the number of non-zero coordinates in x and $\|x\|_\infty$ denotes $\max_j |x_j|$ then $\|x\|_2 \leq \sqrt{\|x\|_0} \|x\|_\infty$; with the Cauchy-Schwarz inequality, this implies that

$$\left(\mathbb{E} \left[\max_i \|M^{(1)}_{i*}\|_2 \right] \right)^2 \leq \mathbb{E} \left[\max_i \|M^{(1)}_{i*}\|_0 \right] \mathbb{E} \left[\max_i \|M^{(1)}_{i*}\|_\infty^2 \right] \tag{2}$$

First, we will show that every row of $M^{(1)}$ is sparse. Let $Z_{ij} = \sum_{k \neq j} \xi_{i,j,k}$ and let Y_{ij} be the indicator that $Z_{ij} > 0$. Recalling that $\mathbb{E}\xi_{i,j,k} = p_{i,j,k}$, we have (by Assumption 3.1) $\mathbb{E} Z_{ij} \leq \kappa p$. Since Z_{ij} takes non-negative integer values, we have $\Pr(Y_{ij} = 1) = \Pr(Z_{ij} > 0) \leq \kappa p$. By Bernstein's inequality, for any fixed i

$$\Pr(\|M^{(1)}_{i*}\|_0 \geq \kappa d_2 p + t) \leq \Pr(\sum_{j=1}^{d_2} Y_{ij} \geq \kappa d_2 p + t) \leq \exp\left(-\frac{t^2/2}{\kappa p d_2 + t/3} \right).$$

Integrating by parts, we have

$$\mathbb{E} \left[\|M^{(1)}_{i*}\|_0 \right] \leq \kappa d_2 p + \int_{\kappa d_2 p}^{\infty} \Pr(\|M^{(1)}_{i*}\|_0 \geq t) \, dt \leq \kappa d_2 p + \frac{3}{8}.$$

Next, we will consider the size of the elements in $M^{(1)}$. First of all, $M_{ij}^{(1)} \leq Z_{ij}$ (this fairly crude bound will lose us a factor of $\sqrt{\log(d_1 d_2)}$). Now, Bernstein's inequality applied to Z_{ij} gives

$$\Pr(M_{ij}^{(1)} \geq \kappa p + t) \leq \Pr(Z_{ij} \geq \kappa p + t) \leq \exp\left(-\frac{t^2/2}{\kappa p + t/3}\right).$$

Taking a union bound over i and j, if $t \geq C\kappa \log(d_1 d_2)$ then

$$\Pr(\max_{ij} M_{ij}^{(1)} \geq t) \leq d_1 d_2 \exp(-ct) \leq \exp(-c't).$$

Integrating by parts,

$$\mathbb{E}\left[\max_{ij} M_{ij}^{(1)}\right] \leq \kappa \log^2(d_1 d_2) + \int_{\kappa \log^2(d_1 d_2)}^{\infty} \Pr(\max_{ij} M_{ij}^{(1)} \geq \sqrt{t})\, dt \leq \kappa \log^2(d_1 d_2) + C.$$

Going back to (2), we have shown that

$$\mathbb{E} \max_i \|M_{i*}^{(1)}\| \leq C\kappa \sqrt{pd_2} \log(d_1 d_2).$$

The same argument applies to $M_{*j}^{(1)}$ (but with $\sqrt{pd_1}$ instead of $\sqrt{pd_2}$), and so we conclude from (1) that

$$\mathbb{E}\|M^{(1)}\| \leq C\kappa \sqrt{p(d_1 + d_2)} \log(d_1 d_2).$$

C. Proof of Theorem 3.2

C.1. A sketch of the proof

The proof of Theorem 3.2 uses Fano's inequality.

1. We construct matrices $X^1, \dots, X^\ell$. These matrices all have small nuclear norm, and for every pair i, j the KL-divergence between the induced observation distributions is $\Theta(\log \ell)$. We construct these matrices randomly, using concentration inequalities and a union bound to show that we can take ℓ of the order $\sqrt{\lambda m(d_1 + d_2)}$.
2. We apply Fano's inequality to show that if we generate data according to a randomly chosen X^i, then any algorithm has a reasonable chance to choose a different X^j (using the fact that the KL-divergence is $O(\log \ell)$). Since the KL-divergence is $\Omega(\log \ell)$, this implies that the algorithm incurs a substantial penalty whenever it makes a wrong choice.

In any application of Fano's inequality, the key is to construct a large number of admissible models that are close to one another in KL-divergence. Specifically, if we can construct distributions $\mathbb{P}_1, \dots, \mathbb{P}_\ell$ with $D(\mathbb{P}_i \| \mathbb{P}_j) + 1 \leq \frac{1}{2} \log \ell$ for all i, j, then given a single sample from some $\mathbb{P}_i$, no algorithm can accurately identify which $\mathbb{P}_i$ it came from. In order to apply this denote by $\mathbb{P}_{X,m}$ the distribution of the data when the true parameters are X. We will construct $X^1 \dots, X^\ell \in \sqrt{\lambda d_1 d_2} K$ such that for all $i \neq j$,

$$D(\mathbb{P}_{X^i,m} \| \mathbb{P}_{X^j,m}) + 1 \leq \frac{1}{2} \log \ell, \tag{3}$$

$$R_j(X^i) \geq R_j(X^j) + c\frac{\log \ell}{m} \tag{4}$$

for some constant $c > 0$, where R_j denotes the expected risk when the true parameters are given by X^j. Given a single observation from some $\mathbb{P}_{X^j,m}$, (3) will imply (by Fano's inequality) that no algorithm can correctly identify which X^j was the true parameter. On the other hand, (4) will imply that if the algorithm makes a mistake – say it chooses X^i for $i \neq j$ – then its risk will be $c\frac{\log \ell}{m}$ larger than the best in the class. In particular, if we can prove (3) and (4) with $\log \ell \sim \sqrt{\lambda m(d_1 + d_2)}$ then it will imply Theorem 3.2.

We construct a set of matrices satisfying (3) and (4) using a probabilistic method. Supposing that $d_2 \geq d_1$, we choose a parameter $\gamma > 0$ and set B to be an integer that is approximately $\lambda \gamma^{-2}$. We define X^1 by filling its top $B \times d_2$ block with independent, uniform $\pm\gamma$ entries, and then copying that top block B/d_1 times to fill the matrix. Then let $X^2, \dots, X^\ell$ be independent copies of X^1. First of all, each $X^i \in \sqrt{\lambda d_1 d_2} K$ because $\|X^i\|_* \le \sqrt{\mathrm{rank}(X^i)}\|X^i\|_F \le \sqrt{\lambda d_1 d_2}$.

Now, let us consider $D(\mathbb{P}_{X^1,m} \| \mathbb{P}_{X^2,m})$. For a single i, j, k triple, there is probability $1/4$ of having $X^1_{i,j} - X^1_{i,k}$ different from $X^2_{i,j} - X^2_{i,k}$, in which case they differ by 4γ. If γ is bounded above, each different entry contributes $\Theta(\alpha^2\gamma^2)$ to the KL-divergence between $\mathbb{P}_{X^1,m}$ and $\mathbb{P}_{X^2,m}$. Since about m entries are observed in $\mathbb{P}_{X^1,m}$, we see that

$$D(\mathbb{P}_{X^1,m} \| \mathbb{P}_{X^2,m}) \asymp m\gamma^2. \tag{5}$$

On the other hand, $R_1(X^1)$ and $R_1(X^2)$ differ by $\Theta(\gamma^2)$, because for a constant fraction of triples i, j, k, the chance that $Y_{i,j,k}$ is 1 differs by $O(\gamma)$ in X^1 and X^2, and on the event that $Y_{i,j,k}$ differs in these two models the loss differs by another $O(\gamma)$ factor.

Applying standard concentration inequalities, we show that one can apply the union bound to $\ell = \exp(cBd_2)$ of these matrices. In view of (3) and (5), we need to take $Bd_2 = \frac{\lambda^2}{\gamma^2 d_1} \asymp m\gamma^2$. Eliminating γ, we end up with $\log \ell \asymp \sqrt{\lambda m / d_1}$ (which is within a constant factor of $\sqrt{\lambda m(d_1 + d_2)}$ under our assumption that $d_2 \ge d_1$).

C.2. Some concentration lemmas

We begin by quoting some standard concentration results (see, e.g. (Vershynin, 2012)).

Definition C.1. *A random variable X is σ^2-subgaussian if $\mathbb{E}e^{\theta X} \le e^{\theta^2\sigma^2/2}$ for all $\theta > 0$. A random variable X is L-subexponential if $\mathbb{E}e^{\theta X} \le (1 - \theta^2 L^2)$ for $\theta < 1/L$.*

One can easily show that the product of two subgaussian variables is subexponential:

Lemma C.2. *If X is σ^2-subgaussian and Y is τ^2-subgaussian then XY is $C\sigma\tau$-subexponential for a universal constant C.*

Moreover, one has a Bernstein-type inequality for sums of independent subexponential variables.

Lemma C.3. *If $X_1, \dots, X_k$ are i.i.d. L-subexponential then*

$$\Pr(\sum_i X_i \ge t) \le \exp\left(-\frac{ct^2}{L^2 k + Lt}\right).$$

C.3. Construction of a packing set

Let $0 < \gamma < 1$ be some parameter to be determined such that $B := \lambda\gamma^{-2}$ is an integer.

Proposition C.4. *Suppose that $\mathcal{L}'(0) < 0$. For every sufficiently small γ (depending on $\mathcal{L}$), there exists a set $\mathcal{X} \subset \sqrt{\lambda d_1 d_2} K$ of $\exp(cBd_2)$ $d_1 \times d_2$ matrices such that for any two $X^1, X^2 \in \mathcal{X}$,*

$$\frac{1}{d_1 d_2^2} \sum_{i=1}^{d_1} \sum_{j,k=1}^{d_2} \mathbb{E}_{X^1}[\mathcal{L}(Y(X^2_{ij} - X^2_{ik})) - \mathcal{L}(Y(X^1_{ij} - X^1_{ik}))] \ge c\gamma^2$$

and for any m,

$$\frac{1}{m} D(\mathbb{P}_{X^1,m} \| \mathbb{P}_{X^2,m}) \le C\gamma^2,$$

where $0 < c < C$ are universal constants.

Following Davenport et al., we construct this set $\mathcal{X}$ randomly: let X be a random $B \times d_2$ matrix, where each element is chosen independently to be either γ or $-\gamma$.

Lemma C.5. *Let X^1 and X^2 be independent copies of X. Then with probability at least $1 - \exp(-cBd_2)$,*

$$\sum_{i=1}^{B} \sum_{j,k=1}^{d_2} (X^1_{ij} - X^1_{ik} - X^2_{ij} + X^2_{ik})^2 \ge 2\gamma^2 B d_2^2,$$

where $c > 0$ is a universal constant.

Before proving Lemma C.5, let us see how it implies Proposition C.4. First of all, for X a random $B \times d_2$ matrix as above, let $\tilde{X}$ be the $d_1 \times d_2$ matrix obtained by stacking $\lceil d_1/B \rceil$ copies of X, and filling out any remaining entries by zeros. Then, for random X and Y, with high probability

$$\begin{aligned}\sum_{i=1}^{d_1} \sum_{j,k=1}^{d_2} (\tilde{X}^1_{ij} - \tilde{X}^1_{ik} - \tilde{X}^2_{ij} + \tilde{X}^2_{ik})^2 &= \lceil d_1/B \rceil \sum_{i=1}^{B} \sum_{j,k=1}^{d_2} (X^1_{ij} - X^1_{ik} - X^2_{ij} + X^2_{ik})^2 \\ &\asymp \gamma^2 d_1 d_2^2, \end{aligned} \tag{6}$$

where the lower bound for the last line came from Lemma C.5, and the upper bound just came from the observation that each term in the sum is bounded by $16\gamma^2$. Let $\mathcal{X}$ be the set obtained by choosing $\exp(cBd_2/4)$ random copies of $\tilde{X}$ in this way. The high-probability estimate in Lemma C.5 implies that with high probability, *every* pair $\tilde{X}^1, \tilde{X}^2$ in $\mathcal{X}$ satisfies (6). Now,

$$\begin{aligned} D(\mathbb{P}_{X^1,m} \| \mathbb{P}_{X^2,m}) &= \mathbb{E}_\Omega \left[\sum_{(i,j,k)\in\Omega} D(f(X^1_{ij} - X^1_{ik}) \| f(X^2_{ij} - X^2_{ik})) \right] \\ &\asymp \frac{m}{d_1 d_2^2} \sum_{i,j,k} (X^1_{ij} - X^1_{ik} - X^2_{ij} + X^2_{ik})^2, \end{aligned}$$

where $f(x) = e^x/(1+e^x)$ is the logistic function, and the last line follows from a Taylor expansion of $D(f(x)\|f(y))$ around $x = y$, because all the X^1_{ij} and X^2_{ij} are bounded by $\gamma < 1$. Together with (6), this proves the first inequality in Proposition C.4; the second inequality follows because each term of the form $D(f(X_{ij} - X_{ik}) \| f(Y_{ij} - Y_{ik}))$ is bounded by a constant times γ^2. This proves the second inequality of Proposition C.4.

By Taylor expansion again, if γ is sufficiently small (depending on $\mathcal{L}$) then

$$\mathcal{L}(Y_{i,j,k}(X^2_{i,j} - X^2_{i,k})) - \mathcal{L}(Y_{i,j,k}(X^1_{i,j} - X^1_{i,k})) \asymp Y_{i,j,k}(X^1_{i,j} - X^1_{i,k} - X^2_{i,j} + X^2_{i,k}).$$

Now, if i,j,k is a triple for which $2\gamma = X^1_{i,j} - X^1_{i,k} > X^2_{i,j} - X^2_{i,k}$ (and under the event of Lemma C.5, there are at least cBd_2^2 such triples) then $\mathbb{E}_{X^1}[Y_{i,j,k}] \asymp \gamma$ and so

$$\mathbb{E}_{X^1}[\mathcal{L}(Y_{i,j,k}(X^2_{i,j} - X^2_{i,k})) - \mathcal{L}(Y_{i,j,k}(X^1_{i,j} - X^1_{i,k}))] \asymp \gamma^2.$$

The same holds when i,j,k is a triple for which $-2\gamma = X^1_{i,j} - X^1_{i,k} < X^2_{i,j} - X^2_{i,k}$. Finally, if i,j,k is a triple such that $X^1_{i,j} - X^1_{i,k} = X^2_{i,j} - X^2_{i,k}$ then the expectation is zero. Summing over all triples, we see that on the event that Lemma C.5 holds,

$$\frac{1}{Bd_2^2} \sum_{i,j,k} \mathbb{E}_{X^1}[\mathcal{L}(Y_{i,j,k}(X^2_{i,j} - X^2_{i,k})) - \mathcal{L}(Y_{i,j,k}(X^1_{i,j} - X^1_{i,k}))] \geq c\gamma^2.$$

After summing over all $\lceil d_1/B \rceil$ blocks, this proves the first inequality of Proposition C.4.

Proof of Lemma C.5. We expand the square:

$$\begin{aligned} \sum_{ijk} (X_{ij} - X_{ik} - Y_{ij} + Y_{ik})^2 &= 2 \sum_{ijk} X_{ij}^2 + Y_{ij}^2 + 2X_{ij}Y_{ik} - X_{ij}X_{ik} - Y_{ij}Y_{ik} - 2X_{ij}Y_{ij} \\ &= 4\gamma^2 Bd_2^2 + 2 \sum_{ijk} 2X_{ij}Y_{ik} - X_{ij}X_{ik} - Y_{ij}Y_{ik} - 2X_{ij}Y_{ij}. \end{aligned} \tag{7}$$

We may study each of the cross-terms separately: for the $X_{ij}Y_{ik}$ term, note that $\sum_j X_{ij}$ and $\sum_k Y_{ik}$ are both $\gamma^2 d_2$-subgaussian (by Hoeffding's inequality). Hence, $\sum_{jk} X_{ij}Y_{ik}$ is $C\gamma^2 d_2$-subexponential (by Lemma C.2) and so by Lemma C.3,

$$\Pr\left(\left| \sum_{ijk} X_{ij}Y_{ik} \right| \geq \frac{1}{8}\gamma^2 Bd_2^2 \right) \leq 2\exp(-cBd_2).$$

The similar argument applies to the $X_{ij}X_{ik}$ term: $\sum_j X_{ij}$ is $\gamma^2 d_2$-subgaussian and so $\sum_{ijk} X_{ij}X_{ik} = \sum_i (\sum_j X_{ij})^2$ is $C\gamma^2 d_2$-subexponential; hence

$$\Pr\left(\left|\sum_{ijk} X_{ij}X_{ik}\right| \geq \frac{1}{8}\gamma^2 B d_2^2\right) \leq 2\exp(-cBd_2).$$

Of course, the $Y_{ij}Y_{ik}$ term is identical. Finally, note that $\sum_{ijk} X_{ij}Y_{ij} = d_2 \sum_{ij} X_{ij}Y_{ij}$. Since the terms in this sum are i.i.d., we may apply Hoeffding's inequality to obtain

$$\Pr\left(\left|\sum_{ijk} X_{ij}Y_{ij}\right| \geq \frac{1}{8}\gamma^2 B d_2^2\right) = \Pr\left(\left|\sum_{ij} X_{ij}Y_{ij}\right| \geq \frac{1}{8}\gamma^2 B d_2\right) \leq 2\exp(-cB^2 d_2^2).$$

Putting everything together, we see that with high probability, the total of all the cross-terms in (7) is at most half of the first term. □

C.4. Completing the proof

Let C denote the constant from Proposition C.4. Assume that $d_1 \leq d_2$ and that m is large enough so

$$\sqrt{\frac{d_2}{m}} \leq 8C\sqrt{\lambda} \leq \sqrt{\frac{m}{d_2}}. \tag{8}$$

Note that under the assumptions $\lambda \geq 1$ and $m \geq d_1 + d_2$ from Theorem 3.2, the lower bound of (8) is satisfied. Moreover, if the upper bound of (8) is not satisfied then we may decrease λ until it is; the conclusion of Theorem 3.2 will not be affected because as long as (8) fails, the minimum in Theorem 3.2 will be 1.

By the lower bound in (8), there is an integer B such that

$$B \leq \sqrt{\frac{\lambda m}{d_2}} \leq 2B;$$

fix this B and define γ by

$$\gamma^2 = \lambda / B \asymp \sqrt{\frac{\lambda d_2}{m}}.$$

By the upper bound in (8), $\gamma \leq 1$.

Now, Fano's inequality states that if we first select a random $X \in \mathcal{X}$ and then draw a sample from $\mathbb{P}_{X,m}$, then any algorithm trying to identify X can succeed with probability at most

$$\frac{\min\{D(\mathbb{P}_{X,m} \| \mathbb{P}(Y,m)) : X, Y \in \mathcal{X}\} + 1}{\log|\mathcal{X}|} \leq \frac{2Cm\gamma^2}{Bd_2} \leq \frac{1}{2}.$$

Finally, note that by the first inequality in Proposition C.4, the error incurred by choosing the wrong $X \in \mathcal{X}$ is at least $c\gamma^2 \asymp \sqrt{\frac{\lambda d_2}{m}}$.

Now, we have so far only discussed the case $d_2 \geq d_1$. The case $d_1 \leq d_2$ is not exactly equivalent because our model is not symmetric in its treatment of users and items. However, the proof of Theorem 3.2 does not change very much. We take horizontally stacked blocks of size $d_1 \times B$ instead of $B \times d_2$. The main difference is in the calculation leading to (6): there are extra cross-terms appearing due to the fact that items in different blocks need to be compared with one another. However, all of these additional terms may be controlled with Lemmas C.2 and C.3 in much the same way as the existing terms are controlled.

D. Comparison to Stochastic Gradient Descent

Another practical algorithm to optimize (3) is Stochastic Gradient Descent (SGD). We have experimented SGD on the same datasets in Table 1. We ran the algorithm with the same regularization parameters and different step sizes. The statistical results for SGD were observed to be no better than AltSVM, and hence we did not present them in the main paper.

Datasets	N	NDCG@10
ML1m	20	0.6852
	50	0.7666
	100	0.7728
ML10m	20	0.6977
	50	0.7452
	100	0.7659

Table 1. NDCG@10 of SGD on different datasets, for different numbers of observed ratings per user.

Precision@	SGD with $C = 5000$
1	0.1556
2	0.1498
5	0.1236
10	0.1031
100	0.0441

Table 2. Precision@K for SGD of (3) on the binarized MovieLens1m dataset.

Let us first describe the SGD procedure. At each step, ones chooses a triple $(i, j, k) \in \Omega$ uniformly at random and run a SGD step, which can be written as

$$u_i^+ \leftarrow u_i - \eta \cdot \left\{ g \cdot (v_j - v_k) + \frac{\lambda}{|\Omega_i|} u_i \right\}$$

$$v_j^+ \leftarrow v_j - \eta \cdot \left\{ g \cdot u_i + \frac{\lambda}{|\Omega^j|} v_j \right\}$$

$$v_j^+ \leftarrow v_j - \eta \cdot \left\{ -g \cdot u_i + \frac{\lambda}{|\Omega^k|} v_k \right\}$$

where $\Omega^{(j)}$ denotes the number of comparisons in Ω which involve item j. η is a step size and $g \in \partial \mathcal{L}(u_i^\top (v_j - v_k))$.

The following tables show the statistical result of SGD. The step size is chosen by $\eta = \frac{\alpha}{1+\beta t}$ as suggested in (Yun et al., 2014). α and β were the powers of 10^{-1}, and the best result is reported. The results are comparable to AltSVM, but it did not achieve better results. We note that this is the best result from several different step sizes, while AltSVM does not have any other parameter to choose except for the regularization parameter.

References

Davenport, Mark A, Plan, Yaniv, Berg, Ewout van den, and Wootters, Mary. 1-bit matrix completion. *arXiv preprint arXiv:1209.3672*, 2013.

Seginer, Yoav. The expected norm of random matrices. *Combinatorics Probability and Computing*, 9(2):149–166, 2000.

Vershynin, Roman. *Compressed sensing: theory and applications*, chapter Introduction to the non-asymptotic analysis of random matrices. Cambridge University Press, 2012.

Yun, Hyokun, Yu, Hsiang-Fu, Hsieh, Cho-Jui, Viswanathan, S. V. N., and Dhillon, Inderjit S. NOMAD: Non-locking, stochastic multi-machine algorithm for asynchronous and descentralized matrix completion. In *VLDB*, 2014.