

## Supplement

### 6. Extension of the salient feature preference model to $k$ -wise comparisons

We describe how to extend the salient feature preference model of Equation (2) from pairwise comparisons to  $k$ -wise comparisons when  $k > 2$ . We base our generalization on the Plackett-Luce model (Plackett, 1975; Luce, 1959), which is a generalization of the BTL model from pairwise comparisons to  $k$ -wise comparisons.

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

$$\mathbb{P}(t_1 >_B \dots >_B t_k) = \prod_{\ell=1}^k \frac{\exp\left(\langle U_{t_\ell}^{\tau(T_\ell)}, w^* \rangle\right)}{\sum_{j \in [k] \setminus [\ell-1]} \exp\left(\langle U_{t_j}^{\tau(T_\ell)}, w^* \rangle\right)}, \quad (5)$$

where “ $t_1 >_B \dots >_B t_k$ ” means item  $t_1$  is preferred to item  $t_2$  and so on and so forth.

We explain Equation (5): Given items  $T_\ell = (t_1, \dots, t_k)$ , first project each item’s features  $U_{t_i}$  onto the coordinate subspace spanned by the coordinates given by  $\tau(T_\ell)$ . Then the utility of item  $t_i$  in the presence of the other items in  $T$  is given by the inner product of its projected features with  $w^*$ :  $\langle (U_{t_i})^{\tau(T_\ell)}, w^* \rangle$ . The higher the utility an item has, the more likely the item will be ranked higher among the items in  $T_\ell$ . Now imagine a bag of balls where each ball corresponds to one of the items in  $T_\ell$ . We select balls from this bag without replacement where the probability of picking a ball is the ratio of its utility to the sum of the utilities of all the remaining balls. The order in which we select balls results in a ranking of the  $k$  items. This process is what Equation (5) represents.

In the pairwise comparison case ( $k = 2$ ) for two items  $T_\ell = (i, j)$ , Equation (5) reduces to Equation (2), which is the salient preference model. We can also extend the top- $t$  selection function naturally to accommodate  $k$ -wise comparisons.

### 7. Negative log-likelihood derivation

**Lemma 2.** *Under the set-up of Section 2, the negative log-likelihood of  $w \in \mathbb{R}^d$  is*

$$\mathcal{L}_m(w; U, S_m, \tau) = \sum_{\ell=1}^m \log\left(1 + \exp\left(\langle U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}, w \rangle\right)\right) - y_\ell \langle U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}, w \rangle. \quad (6)$$

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

$$\mathcal{L}_m(w; U, S_m, \tau) \quad (7)$$

$$= -\log P_w(S_m) \quad (8)$$

$$= -\log\left(\prod_{\ell=1}^m (\mathbb{P}(y_\ell = 1)^{y_\ell} \mathbb{P}(y_\ell = 0)^{1-y_\ell})\right) \text{ by independence and since } y_\ell \in \{0, 1\} \quad (9)$$

$$= -\sum_{i=1}^m y_\ell \log(\mathbb{P}(y_\ell = 1)) + (1 - y_\ell) \log(1 - \mathbb{P}(y_\ell = 1)) \quad (10)$$

$$= -\sum_{i=1}^m y_\ell \log\left(\frac{\exp\left(\langle U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}, w \rangle\right)}{1 + \exp\left(\langle U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}, w \rangle\right)}\right) \quad (11)$$

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

$$= \sum_{i=1}^m \log\left(1 + \exp\left(\langle U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}, w \rangle\right)\right) - y_\ell \langle U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}, w \rangle \quad (12)$$

□

## 8. Proof of Proposition 1

**Proposition 3** (Restatement of Proposition 1). *Given item features  $U \in \mathbb{R}^{d \times n}$ , the salient feature preference model with selection function  $\tau$  is identifiable if and only if  $\text{span}\{U_i^{\tau(i,j)} - U_j^{\tau(i,j)} : (i, j) \in P\} = \mathbb{R}^d$ .*

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

$$\mathbb{P}(i >_B j; w) = \mathbb{P}(i >_B j; w^*) \quad (13)$$

$$\Leftrightarrow \frac{\exp(\langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w \rangle)}{1 + \exp(\langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w \rangle)} = \frac{\exp(\langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w^* \rangle)}{1 + \exp(\langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w^* \rangle)} \quad (14)$$

$$\Leftrightarrow \exp(\langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w \rangle) = \exp(\langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w^* \rangle) \quad (15)$$

$$\Leftrightarrow \langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w \rangle = \langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w^* \rangle \quad (16)$$

$$\Leftrightarrow \langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w^* - w \rangle = 0. \quad (17)$$

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

$$\langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, w^* - (w^* - x) \rangle = \langle U_i^{\tau(i,j)} - U_j^{\tau(i,j)}, x \rangle \quad (18)$$

$$= 0. \quad (19)$$

Therefore, with  $w = w^* - x$ , Equation (17) is true and implies Equation (13) meaning

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

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

$\Leftarrow$  Now assume  $\text{span}\{U_i^{\tau(i,j)} - U_j^{\tau(i,j)} : (i, j) \in P\} = \mathbb{R}^d$ . We want to prove identifiability so suppose there exists  $w$  such that Equation (13) holds. We will show  $w = w^*$ . Let  $x \in \mathbb{R}^d$  where  $x = \sum_{(i,j) \in P} \alpha_{i,j} (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})$  for  $\alpha_{i,j} \in \mathbb{R}$ . Then by Equation (17),

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

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

## 9. Proof of Proposition 2

**Proposition 4** (Restatement of Proposition 2). *Under the set-up of Section 2,  $\lambda := \lambda_{\min}(\mathbb{E}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T) > 0$  if and only if the salient feature preference model with selection function  $\tau$  is identifiable.*

*Proof.* For both directions, we prove the contrapositive.

$\Rightarrow$  Assume  $\lambda_{\min}(\mathbb{E}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T) = 0$ . Recall the expectation is with respect to a uniformly at

random chosen pair of items. Let  $\mathbf{0} \in \mathbb{R}^d$  be the all 0 vector. Then there exists  $y \neq \mathbf{0} \in \mathbb{R}^d$  that has unit norm such that

$$(\mathbb{E}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T)y = \mathbf{0} \quad (20)$$

$$\implies y^T (\mathbb{E}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T)y = 0 \quad (21)$$

$$\implies \frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} y^T (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T y = 0 \text{ since } (i,j) \in P \text{ is chosen uniformly at random} \quad (22)$$

$$\implies \frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} \|(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T y\|_2^2 = 0 \quad (23)$$

$$\implies \|(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T y\|_2^2 = 0 \forall (i,j) \in P \quad (24)$$

$$\implies (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T y = \mathbf{0} \forall (i,j) \in P. \quad (25)$$

We now show  $y \notin \text{span}\{U_i^{\tau(i,j)} - U_j^{\tau(i,j)} : (i,j) \in P\}$ , which establishes the salient feature preference model is not identifiable by Proposition 1. By contradiction, suppose there exist  $\alpha_{i,j} \in \mathbb{R}$  such that

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

Then

$$1 = \langle y, y \rangle \quad (26)$$

$$= \left\langle \sum_{(i,j) \in P} \alpha_{i,j} (U_i^{\tau(i,j)} - U_j^{\tau(i,j)}), y \right\rangle \quad (27)$$

$$= \sum_{(i,j) \in P} \alpha_{i,j} \langle (U_i^{\tau(i,j)} - U_j^{\tau(i,j)}), y \rangle \quad (28)$$

$$= 0, \quad (29)$$

a contradiction.

$\Leftarrow$  Now suppose that the preference model is not identifiable. By Proposition 1,  $\text{span}\{U_i^{\tau(i,j)} - U_j^{\tau(i,j)} : (i,j) \in P\} \neq \mathbb{R}^d$ . In particular, there exists  $y \in \mathbb{R}^d$  such that  $y \neq \mathbf{0}$  and  $\langle y, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle = 0$  for all  $(i,j) \in P$ , i.e.  $y$  is in the orthogonal complement of  $\text{span}\{U_i^{\tau(i,j)} - U_j^{\tau(i,j)} : (i,j) \in P\}$ . Furthermore,

$$\frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T y = \mathbf{0} \quad (30)$$

$$\implies (\mathbb{E}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T)y = \mathbf{0}, \quad (31)$$

$$(32)$$

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

## 10. Proof of Theorem 1

Recall the set-up from the beginning of Section 2. There are  $n$  items where the features of the items are given by the columns of  $U \in \mathbb{R}^{d \times n}$  and let  $w^* \in \mathbb{R}^d$  be the judgment vector. Let  $\tau$  be the selection function. Let  $S_m = \{(i_\ell, j_\ell, y_\ell)\}_{\ell=1}^m$  be the  $m$  samples of independent pairwise comparisons where each pair of items  $(i_\ell, j_\ell)$  is chosen uniformly at random from all the

pairs of items  $P := \{(i, j) \in [n] \times [n] : i < j\}$ . Furthermore,  $y_\ell$  is 1 if the  $i_\ell$ -th item beats the  $j_\ell$ -th item and 0 otherwise where  $y_\ell \sim \text{Bernoulli} \left( \frac{\exp(\langle U_{i_\ell}^\tau(i_\ell, j_\ell) - U_{j_\ell}^\tau(i_\ell, j_\ell), w^* \rangle)}{1 + \exp(\langle U_{i_\ell}^\tau(i_\ell, j_\ell) - U_{j_\ell}^\tau(i_\ell, j_\ell), w^* \rangle)} \right)$ . We will not repeat these assumptions in the following lemmas.

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

**Theorem 3** (restatement of Theorem 1: sample complexity of estimating  $w^*$ ). *Let  $U$ ,  $w^*$ ,  $\tau$ , and  $S_m$  be defined as above. Let  $\hat{w}$  be the maximum likelihood estimator, i.e. the minimum of  $\mathcal{L}_m$  in Equation (3), restricted to the set  $\mathcal{W}(b^*)$ . The following expectations are taken with respect to a uniformly chosen random pair of items from  $P$ . For  $(i, j) \in P$ , let*

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

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

$$\beta := \max_{(i,j) \in P} \|U_i^\tau(i,j) - U_j^\tau(i,j)\|_\infty. \quad (33)$$

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

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

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

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

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

*Proof.* We use the proof technique of Theorem 4 in (Negahban et al., 2016). We use the notation  $\mathcal{L}_m(w)$  instead of  $\mathcal{L}_m(w; U, S_m, \tau)$  throughout the proof since it is clear from context.

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

$$\mathcal{L}_m(w^* + \Delta) - \mathcal{L}_m(w^*) - \langle \nabla \mathcal{L}_m(w^*), \Delta \rangle \quad (34)$$

$$\leq -\langle \nabla \mathcal{L}_m(w^*), \Delta \rangle \quad (35)$$

$$\leq \|\nabla \mathcal{L}_m(w^*)\|_2 \|\Delta\|_2, \quad (36)$$

by the Cauchy-Schwarz inequality.

Recall Taylor's theorem:

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

Now, we lower bound Equation (34). Let  $H_{\mathcal{L}_m}$  be the Hessian of  $\mathcal{L}_m$ . Then by Taylor's theorem, there exists  $\lambda \in [0, 1]$  such that

$$\frac{1}{m} (\mathcal{L}_m(w^* + \Delta) - \mathcal{L}_m(w^*) - \langle \nabla \mathcal{L}_m(w^*), \Delta \rangle) \quad (37)$$

$$= \frac{1}{2m} \Delta^T H_{\mathcal{L}_m}(w^* + \lambda\Delta) \Delta \quad (38)$$

$$= \frac{1}{2m} \sum_{\ell=1}^m h(\langle w^* + \lambda\Delta, U_{i_\ell}^\tau(i_\ell, j_\ell) - U_{j_\ell}^\tau(i_\ell, j_\ell) \rangle) \Delta^T (U_{i_\ell}^\tau(i_\ell, j_\ell) - U_{j_\ell}^\tau(i_\ell, j_\ell)) (U_{i_\ell}^\tau(i_\ell, j_\ell) - U_{j_\ell}^\tau(i_\ell, j_\ell))^T \Delta \quad (39)$$

where the Hessian  $H_{\mathcal{L}_m}$  is computed in Lemma 6 and  $h(x) := \frac{e^x}{(1+e^x)^2}$ .

Note

$$|\langle w^* + \lambda\Delta, U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)} \rangle| \quad (40)$$

$$= |(1-\lambda)\langle w^*, U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)} \rangle + \lambda\langle \hat{w}, U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)} \rangle| \quad (41)$$

$$\leq (1-\lambda)b^* + \lambda b^* \quad (42)$$

$$= b^* \quad (43)$$

where the second to last inequality is by definition of  $b^*$  and since  $\hat{w} \in \mathcal{W}(b^*)$ . Because  $h(x) = \frac{e^x}{(1+e^x)^2}$  is symmetric and decreases on  $[0, \infty)$  by Lemma 7, for any  $i, j \in [n]$ ,

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

Therefore,

$$\frac{1}{m} (\mathcal{L}_m(w^* + \Delta) - \mathcal{L}_m(w^*) - \langle \nabla \mathcal{L}_m(w^*), \Delta \rangle) \quad (44)$$

$$\geq \frac{\exp(b^*)}{2m(1 + \exp(b^*))^2} \sum_{\ell=1}^m \Delta^T (U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}) (U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)})^T \Delta. \quad (45)$$

By Lemma 4 and 5 and combining Equations (36) and (45), with probability at least  $1 - \delta$  if

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

$$\left( \frac{\exp(b^*)}{2(1 + \exp(b^*))^2} \right) \frac{\lambda}{2} \|\Delta\|_2^2 \leq \frac{1}{m} (\mathcal{L}_m(w^* + \Delta) - \mathcal{L}_m(w^*) - \langle \nabla \mathcal{L}_m(w^*), \Delta \rangle) \quad (46)$$

$$\leq \sqrt{\frac{3\beta^2 \log(4d/\delta)d + 4\sqrt{d}\beta \log(4d/\delta)}{6m}} \|\Delta\|_2 \quad (47)$$

$$\implies \|\Delta\|_2 \leq \frac{4(1 + \exp(b^*))^2}{\exp(b^*)\lambda} \sqrt{\frac{3\beta^2 \log(4d/\delta)d + 4\sqrt{d}\beta \log(4d/\delta)}{6m}}. \quad (48)$$

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

We now present the lemmas used in the prior proof.

**Lemma 4.** *Let  $\delta > 0$ . Under the model assumptions in this section, if*

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

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

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

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

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

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

so  $\frac{1}{m} \nabla \mathcal{L}_m(w^*) = \sum_{\ell=1}^m X_\ell$  by Lemma 6.

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

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

$$\mathbb{E}(X_\ell) = \frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} \mathbb{E}(X_\ell | \text{items } i, j \text{ are chosen}) \quad (49)$$

$$= \frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} \frac{1}{m} \left( U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \right) \left( \frac{\exp(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle)}{1 + \exp(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle)} - \mathbb{E}(y_{(i,j)}) \right) \quad (50)$$

$$= \frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} \frac{1}{m} \left( U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \right) \left( \frac{\exp(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle)}{1 + \exp(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle)} - \frac{\exp(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle)}{1 + \exp(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle)} \right) \quad (51)$$

$$= 0, \quad (52)$$

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

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

$$|X_\ell^{(k)}| \quad (53)$$

$$= \left| \frac{1}{m} \left( (U_{i_\ell}^{\tau(i_\ell, j_\ell)})^{(k)} - (U_{j_\ell}^{\tau(i_\ell, j_\ell)})^{(k)} \right) \left( \frac{\exp(\langle w^*, U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)} \rangle)}{1 + \exp(\langle w^*, U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)} \rangle)} - y_\ell \right) \right| \quad (54)$$

$$\leq \frac{1}{m} \left| \left( (U_{i_\ell}^{\tau(i_\ell, j_\ell)})^{(k)} - (U_{j_\ell}^{\tau(i_\ell, j_\ell)})^{(k)} \right) \right| \text{ since } \frac{\exp(\langle w^*, U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)} \rangle)}{1 + \exp(\langle w^*, U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)} \rangle)}, y_\ell \in [0, 1] \quad (55)$$

$$\leq \frac{1}{m} \max_{(i,j) \in P} \|U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\|_\infty \quad (56)$$

$$= \frac{\beta}{m}, \quad (57)$$

by definition of  $\beta$ .

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

$$\mathbb{E}((X_\ell^{(k)})^2) \tag{58}$$

$$= \frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} \mathbb{E}((X_\ell^{(k)})^2 | \text{items } i, j \text{ are chosen}) \tag{59}$$

$$= \frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} \frac{1}{m^2} \left( (U_i^{\tau(i,j)})^{(k)} - (U_j^{\tau(i,j)})^{(k)} \right)^2 \mathbb{E} \left( \left( p(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle) - y_{(i,j)} \right)^2 \right) \tag{60}$$

$$= \frac{1}{m^2 \binom{n}{2}} \sum_{(i,j) \in P} \left( (U_i^{\tau(i,j)})^{(k)} - (U_j^{\tau(i,j)})^{(k)} \right)^2 \tag{61}$$

$$\left( p(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle)^2 - 2\mathbb{E}(y_{(i,j)})p(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle) + \mathbb{E}((y_{(i,j)})^2) \right) \tag{62}$$

$$= \frac{1}{m^2 \binom{n}{2}} \sum_{(i,j) \in P} \left( (U_i^{\tau(i,j)})^{(k)} - (U_j^{\tau(i,j)})^{(k)} \right)^2 \left( - \left( p(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle) \right)^2 + \mathbb{E}((y_{(i,j)})^2) \right) \tag{63}$$

$$= \frac{1}{m^2 \binom{n}{2}} \sum_{(i,j) \in P} \left( (U_i^{\tau(i,j)})^{(k)} - (U_j^{\tau(i,j)})^{(k)} \right)^2 \left( -p(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle)^2 + \mathbb{E}(y_{(i,j)}) \right) \text{ since } y_{(i,j)} \in \{0, 1\} \tag{64}$$

$$= \frac{1}{m^2 \binom{n}{2}} \sum_{(i,j) \in P} \left( (U_i^{\tau(i,j)})^{(k)} - (U_j^{\tau(i,j)})^{(k)} \right)^2 \left( p(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle) - p(\langle w^*, U_i^{\tau(i,j)} - U_j^{\tau(i,j)} \rangle)^2 \right) \tag{65}$$

$$\leq \frac{\beta^2}{4m^2} \tag{66}$$

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

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

**Theorem 2** (Bernstein's inequality). *Let  $X_i$  be i.i.d. random variables such that  $\mathbb{E}(X_i) = 0$  and  $|X_i| \leq M$ . Then for any  $t > 0$ ,*

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

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

$$\mathbb{P} \left( \left| \frac{1}{m} \nabla \mathcal{L}_m(w^*)^{(k)} \right| > t \right) \leq 2 \exp \left( - \frac{\frac{1}{2}t^2}{\frac{\beta^2}{4m} + \frac{\beta t}{3m}} \right) \tag{67}$$

since  $\sum_{\ell=1}^m \mathbb{E}((X_\ell^{(k)})^2) \leq \frac{\beta^2}{4m}$  and  $|X_\ell^{(k)}| \leq \frac{\beta}{m}$ .

Since  $\|x\|_2 \leq \sqrt{d}\|x\|_\infty$  for any  $x \in \mathbb{R}^d$ ,

$$\mathbb{P} \left( \left\| \frac{1}{m} \nabla \mathcal{L}_m(w^*) \right\|_2 > t \right) \tag{68}$$

$$\leq \mathbb{P} \left( \frac{\sqrt{d}}{m} \|\nabla \mathcal{L}_m(w^*)\|_\infty > t \right) \tag{69}$$

$$= \mathbb{P} \left( \left\| \frac{1}{m} \nabla \mathcal{L}_m(w^*) \right\|_\infty > \frac{t}{\sqrt{d}} \right) \tag{70}$$

$$\leq 2d \exp \left( -\frac{\frac{1}{2} \frac{t^2}{d}}{\frac{\beta^2}{4m} + \frac{\beta \frac{t}{\sqrt{d}}}{3m}} \right) \text{ by union bound and inequality (67)} \tag{71}$$

$$= 2d \exp \left( -\frac{t^2}{\frac{d\beta^2}{2m} + \frac{2\beta t\sqrt{d}}{3m}} \right) \tag{72}$$

$$= 2d \exp \left( -\frac{6mt^2}{3d\beta^2 + 4\beta t\sqrt{d}} \right). \tag{73}$$

In other words, for  $t > 0$ , with probability at least  $1 - 2d \exp \left( -\frac{6mt^2}{3d\beta^2 + 4\beta t\sqrt{d}} \right)$ ,  $\left\| \frac{1}{m} \nabla \mathcal{L}_m(w^*) \right\|_2 \leq t$ .

Let

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

Set

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

If

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

then

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

which we establish below.



If

$$m \geq \frac{\alpha}{6} \quad (74)$$

$$\implies m \geq \frac{\alpha(4\beta \log(4d/\delta))^2 d}{6(4\beta \log(4d/\delta))^2 d} \quad (75)$$

$$\implies m \geq \frac{\alpha(4\beta \log(4d/\delta))^2 d}{6(\alpha - 3\beta^2 \log(4d/\delta)d)^2} \quad (76)$$

$$\implies m \geq \frac{\alpha d}{6 \left( \frac{\alpha - 3\beta^2 \log(4d/\delta)d}{4\beta \log(4d/\delta)} \right)^2} \quad (77)$$

$$\implies \left( \frac{\alpha - 3\beta^2 \log(4d/\delta)d}{4\beta \log(4d/\delta)} \right)^2 \geq \frac{\alpha d}{6m} \quad (78)$$

$$\implies \frac{\frac{\alpha}{\log(4d/\delta)} - 3\beta^2 d}{4\beta} \geq \sqrt{\frac{\alpha d}{6m}} \quad (79)$$

$$\implies \frac{\alpha}{\log(4d/\delta)} \geq 4\beta \sqrt{\frac{\alpha d}{6m}} + 3\beta^2 d \quad (80)$$

$$\implies \frac{\alpha}{4\beta \sqrt{\frac{\alpha d}{6m}} + 3\beta^2 d} \geq \log(4d/\delta) \quad (81)$$

$$\implies \frac{t^2 6m}{4\beta t \sqrt{d} + 3\beta^2 d} \geq \log(4d/\delta) \quad (82)$$

$$\implies 2d \exp\left(-\frac{6mt^2}{4\beta t \sqrt{d} + 3\beta^2 d}\right) \leq \frac{\delta}{2} \quad (83)$$

$$(84)$$

Therefore, if

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

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

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

□

**Lemma 5.** For  $(i, j) \in P$ , let  $Z_{(i,j)} = (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T$ . Let

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

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

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

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

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

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

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

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

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

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

where

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

*Proof.* Let

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

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

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

**Theorem 3** (Theorem 1.4 in (Tropp, 2012)). *Consider a finite sequence  $\{X_k\}$  of independent, random, self-adjoint matrices with dimension  $d$ . Assume that each random matrix satisfies  $\mathbb{E}X_k = 0$  and  $\lambda_{\max}(X_k) \leq R$  almost surely. Then for all  $t \geq 0$*

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_k X_k\right) \geq t\right) \leq d \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right), \quad (85)$$

where

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

Notice

$$\sigma_{\max}\left(\sum_{\ell=1}^m \mathbb{E}(X_\ell^2)\right) = m \sigma_{\max}(\mathbb{E}(X_1^2)) \text{ since each } X_\ell \text{ is distributed the same} \quad (86)$$

$$= \frac{m}{m^2} \eta \quad (87)$$

$$= \frac{1}{m} \eta. \quad (88)$$

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

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{\ell=1}^m -X_\ell\right) \geq t\right) \leq d \exp\left(\frac{-t^2/2}{\eta/m + \zeta t/(3m)}\right) \quad (89)$$

$$\leq d \exp\left(\frac{-3mt^2}{6\eta + 2\zeta t}\right). \quad (90)$$

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

$$\lambda_{\max}\left(\sum_{\ell=1}^m -X_{\ell}\right) \leq t \quad (91)$$

$$\implies \frac{\Delta^T}{\|\Delta\|_2} \left(\sum_{\ell=1}^m -X_{\ell}\right) \frac{\Delta}{\|\Delta\|_2} \leq t \quad (92)$$

$$\begin{aligned} \implies \Delta^T & \left(\mathbb{E}\left((U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T\right)\right) - \\ & \frac{1}{m} \sum_{\ell=1}^m (U_{i_{\ell}}^{\tau(i_{\ell},j_{\ell})} - U_{j_{\ell}}^{\tau(i_{\ell},j_{\ell})})(U_{i_{\ell}}^{\tau(i_{\ell},j_{\ell})} - U_{j_{\ell}}^{\tau(i_{\ell},j_{\ell})})^T \Delta \leq t \|\Delta\|_2^2 \end{aligned} \quad (93)$$

$$\begin{aligned} \implies \Delta^T & \left(\mathbb{E}\left((U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T\right)\right) \Delta - t \|\Delta\|_2^2 \\ & \leq \Delta^T \left(\frac{1}{m} \sum_{\ell=1}^m (U_{i_{\ell}}^{\tau(i_{\ell},j_{\ell})} - U_{j_{\ell}}^{\tau(i_{\ell},j_{\ell})})(U_{i_{\ell}}^{\tau(i_{\ell},j_{\ell})} - U_{j_{\ell}}^{\tau(i_{\ell},j_{\ell})})^T\right) \Delta \end{aligned} \quad (94)$$

$$\begin{aligned} \implies \|\Delta\|_2^2 & \frac{\Delta^T}{\|\Delta\|_2} \left(\mathbb{E}\left((U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T\right)\right) \frac{\Delta}{\|\Delta\|_2} - t \|\Delta\|_2^2 \\ & \leq \Delta^T \left(\frac{1}{m} \sum_{\ell=1}^m (U_{i_{\ell}}^{\tau(i_{\ell},j_{\ell})} - U_{j_{\ell}}^{\tau(i_{\ell},j_{\ell})})(U_{i_{\ell}}^{\tau(i_{\ell},j_{\ell})} - U_{j_{\ell}}^{\tau(i_{\ell},j_{\ell})})^T\right) \Delta \end{aligned} \quad (95)$$

$$\implies (\lambda - t) \|\Delta\|_2^2 \leq \Delta^T \left(\frac{1}{m} \sum_{\ell=1}^m (U_{i_{\ell}}^{\tau(i_{\ell},j_{\ell})} - U_{j_{\ell}}^{\tau(i_{\ell},j_{\ell})})(U_{i_{\ell}}^{\tau(i_{\ell},j_{\ell})} - U_{j_{\ell}}^{\tau(i_{\ell},j_{\ell})})^T\right) \Delta \quad (96)$$

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

Set  $t = \frac{\lambda}{2}$ . Since  $\lambda > 0$  by assumption, Equation (96) becomes

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

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

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

since

$$d \exp\left(\frac{-3m\frac{\lambda^2}{4}}{6\eta + 2\frac{\lambda}{2}\zeta t}\right) \leq \frac{\delta}{2} \quad (97)$$

$$\implies \frac{-3m\frac{\lambda^2}{4}}{6\eta + 2\frac{\lambda}{2}\zeta t} \leq -\log(2d/\delta) \quad (98)$$

$$\implies \frac{3m\frac{\lambda^2}{4}}{6\eta + 2\frac{\lambda}{2}\zeta t} \geq 2 \log(2d/\delta) \quad (99)$$

$$\implies m \geq \frac{8 \log(2d/\delta)(6\eta + \lambda\zeta)}{3\lambda^2}. \quad (100)$$

$$(101)$$

□

**Lemma 6** (Gradient and Hessian of Equation (3)). *Given samples  $S_m$ , features of the  $n$  items  $U \in \mathbb{R}^{d \times n}$ , and  $w \in \mathbb{R}^d$ ,*

$$\frac{1}{m} \nabla \mathcal{L}_m(w; U, S_m, \tau) \quad (102)$$

$$= \frac{1}{m} \sum_{\ell=1}^m \frac{\exp(\langle w, U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)} \rangle)}{1 + \exp(\langle w, U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)} \rangle)} \left( U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)} \right) - y_\ell \left( U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)} \right) \quad (103)$$

and

$$\frac{1}{m} H_{\mathcal{L}_m}(w; U, S_m, \tau) \quad (104)$$

$$= \frac{1}{m} \sum_{\ell=1}^m \frac{\exp(\langle w, U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)} \rangle)}{(1 + \exp(\langle w, U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)} \rangle))^2} (U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}) (U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)})^T \quad (105)$$

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

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

Note

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

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

We arrive at the desired result by the chain rule:

$$\frac{1}{m} \mathcal{L}_m(w; U, S_m, \tau) = \quad (106)$$

$$\frac{1}{m} \sum_{\ell=1}^m f'(g(w; U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)})) \nabla_w g(w; U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}) - y_\ell \nabla_w g(w; U_{i_\ell}^{\tau(i_\ell, j_\ell)} - U_{j_\ell}^{\tau(i_\ell, j_\ell)}). \quad (107)$$

**Hessian:** Note

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

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

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

which proves the claim.  $\square$

**Lemma 7.** Let  $h(x) = \frac{e^x}{(1+e^x)^2}$ . Then  $h(x)$  is symmetric and decreases on  $[0, \infty)$ .

*Proof.* Symmetry:

$$h(-x) = \frac{e^{-x}}{(1 + e^{-x})^2} \quad (108)$$

$$= \frac{e^{-x}}{e^{-2x}(e^x + 1)^2} \quad (109)$$

$$= \frac{e^x}{(e^x + 1)^2} \quad (110)$$

$$= h(x). \quad (111)$$

Decreasing on  $[0, \infty)$ :

Note

$$h'(x) = \frac{e^x(1 + e^x)^2 - e^{2x}2(1 + e^x)}{(1 + e^x)^4} \quad (112)$$

$$= \frac{e^x(1 + e^x) - e^{2x}2}{(1 + e^x)^3} \quad (113)$$

$$= \frac{e^x(1 - e^x)}{(1 + e^x)^3} \quad (114)$$

$$\leq 0 \quad (115)$$

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

## 11. Specific Selection Functions: Proofs of Corollaries 1.1 and 1.2

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

### 11.1. Proof of Corollary 1.1

The following lemma is a straight forward generalization from (Negahban et al., 2016), but we include the proof for completeness. We need this lemma to prove Corollary 1.1.

**Lemma 8.** *Let  $U \in \mathbb{R}^{d \times n}$ . Assume that the columns of  $U$  sum to 0:  $\sum_{i=1}^n U_i = 0$ . Then*

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

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

*Proof.* Let  $e_i \in \mathbb{R}^n$  denote the  $i$ -th standard basis vector,  $I_{n \times n}$  denote the  $n \times n$  identity matrix, and  $\mathbb{1} \in \mathbb{R}^n$  be the vector

of all ones. Since the expectation is over a uniformly chosen pair of items  $(i, j) \in P$ ,

$$\mathbb{E}((U_i - U_j)(U_i - U_j)^T) \quad (116)$$

$$= \mathbb{E}(U(e_i - e_j)(e_i - e_j)^T U^T) \quad (117)$$

$$= \frac{1}{\binom{n}{2}} U \left( \sum_{(i,j) \in P} e_i e_i^T - e_i e_j^T - e_j e_i^T + e_j e_j^T \right) U^T \quad (118)$$

$$= \frac{1}{\binom{n}{2}} U \left( (n-1) \sum_{i=1}^n e_i e_i^T - \sum_{(i,j) \in P} e_i e_j^T + e_j e_i^T \right) U^T \text{ since each item is in } n-1 \text{ comparisons} \quad (119)$$

$$= \frac{1}{\binom{n}{2}} U \left( (n-1) I_{n \times n} - \sum_{(i,j) \in P} e_i e_j^T + e_j e_i^T \right) U^T \quad (120)$$

$$= \frac{1}{\binom{n}{2}} U \left( (n-1) I_{n \times n} - (\mathbb{1}\mathbb{1}^T - I_{n \times n}) \right) U^T \text{ explained below} \quad (121)$$

$$= \frac{1}{\binom{n}{2}} U \left( n I_{n \times n} - \mathbb{1}\mathbb{1}^T \right) U^T \quad (122)$$

$$= \frac{1}{\binom{n}{2}} (n U U^T - U \mathbb{1}\mathbb{1}^T U^T) \quad (123)$$

$$= \frac{n}{\binom{n}{2}} U U^T \text{ since } U \mathbb{1} = \sum_{i=1}^n U_i = \mathbf{0} \text{ by assumption.} \quad (124)$$

Equation (121) is because  $e_i e_j^T$  is the matrix with a 1 in the  $i$ -th row and  $j$ -th column and 0 elsewhere and we are summing over all  $(i, j) \in [n] \times [n]$  where  $i < j$ . Thus, the sum equals  $\mathbb{1}\mathbb{1}^T - I_{n \times n}$ , which is the matrix with ones everywhere except for the diagonal.  $\square$

**Corollary 8.1** (Restatement of Corollary 1.1). *Assume the set-up stated in the beginning of Section 2. For the selection function  $\tau$ , suppose  $\tau(i, j) = [d]$  for any  $(i, j) \in P$ . In other words, all the features are used in each pairwise comparison. Assume  $n > d$ . Let  $\nu := \max\{\max_{(i,j) \in P} \|U_i - U_j\|_2^2, 1\}$ . Without loss of generality, assume the columns of  $U$  sum to zero:  $\sum_{i=1}^n U_i = \mathbf{0}$ . Then,*

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

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

and

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

Let

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

Let  $\delta > 0$ . Hence, if

$$m \geq \max \left\{ m_1, \frac{48 \log(2d/\delta) \binom{n}{2}^2}{3n^2 \lambda_{\min}(U U^T)^2} \left( \frac{\nu n \lambda_{\max}(U U^T)}{\binom{n}{2}} + \frac{n^2 \lambda_{\max}(U U^T)^2}{\binom{n}{2}^2} \right) + \frac{8 \log(2d/\delta) \binom{n}{2}}{3n \lambda_{\min}(U U^T)} \left( \nu + \frac{n \lambda_{\max}(U U^T)}{\binom{n}{2}} \right) \right\},$$

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

$$\|w^* - \hat{w}\|_2 \leq \frac{4(1 + \exp(b^*))^2 \binom{n}{2}}{\exp(b^*) n \lambda_{\min}(U U^T)} \sqrt{\frac{3\beta^2 \log(4d/\delta)d + 4\sqrt{d}\beta \log(4d/\delta)}{6m}}. \quad (125)$$

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

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

$$\mathbb{P}(\text{item } i \text{ beats item } j) = \frac{1}{1 + \exp(-\langle w^*, U_i - U_j \rangle)} \quad (126)$$

$$= \frac{1}{1 + \exp(-\langle w^*, (U_i - \bar{U}) - (U_j - \bar{U}) \rangle)}. \quad (127)$$

Let  $\tilde{U} = U(I - \frac{1}{n} \mathbb{1} \mathbb{1}^T)$  be the centered version of  $U$ , i.e. where we subtract  $\bar{U}$  from each column of  $U$ . Since  $n > d$  and by Proposition 9, if  $\lambda_{\min}(U) > 0$ , then  $\lambda_{\min}(\tilde{U}) > 0$  generically. Therefore, WLOG, we may assume  $\sum_{i=1}^n U_i = 0$ .

First, we simplify  $\lambda$ . By Lemma 8,

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

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

$$\lambda_{\max}(\mathbb{E}(U_i - U_j)(U_i - U_j)^T - (U_k - U_\ell)(U_k - U_\ell)^T) \quad (128)$$

$$= \lambda_{\max}\left(\frac{n}{\binom{n}{2}} UU^T - (U_k - U_\ell)(U_k - U_\ell)^T\right) \text{ by Lemma 8} \quad (129)$$

$$\leq \lambda_{\max}\left(\frac{n}{\binom{n}{2}} UU^T\right) + \lambda_{\max}((U_k - U_\ell)(U_k - U_\ell)^T) \quad (130)$$

$$= \lambda_{\max}\left(\frac{n}{\binom{n}{2}} UU^T\right) + \|(U_k - U_\ell)\|_2^2 \quad (131)$$

$$\leq \lambda_{\max}\left(\frac{n}{\binom{n}{2}} UU^T\right) + \nu, \quad (132)$$

$$(133)$$

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

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

$$\mathbb{E}(X - \mathbb{E}(X))^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2. \quad (134)$$

Furthermore, since  $\eta$  is the largest singular value of a symmetric matrix squared, the largest eigenvalue of that matrix is also equal to  $\eta$ . Therefore,  $\eta = \lambda_{\max}(\mathbb{E}((U_i - U_j)(U_i - U_j)^T(U_i - U_j)(U_i - U_j)^T) - \mathbb{E}((U_i - U_j)(U_i - U_j)^T)^2)$ . Most steps are explained below after the equations. Because the expectation is with respect to a uniformly at random pair of items

$(i, j) \in P$  and by Lemma 8,

$$\lambda_{\max} \left( \mathbb{E}((U_i - U_j)(U_i - U_j)^T(U_i - U_j)(U_i - U_j)^T) - \mathbb{E}((U_i - U_j)(U_i - U_j)^T)^2 \right) \quad (135)$$

$$= \lambda_{\max} \left( \frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} (U_i - U_j)(U_i - U_j)^T(U_i - U_j)(U_i - U_j)^T - \frac{n^2}{\binom{n}{2}^2} UU^T UU^T \right) \quad (136)$$

$$= \lambda_{\max} \left( \frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} ((U_i - U_j)^T(U_i - U_j)) (U_i - U_j)(U_i - U_j)^T - \frac{n^2}{\binom{n}{2}^2} UU^T UU^T \right) \quad (137)$$

$$= \lambda_{\max} \left( \frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} ((U_i - U_j)^T(U_i - U_j)) U(e_i - e_j)(e_i - e_j)^T U^T - \frac{n^2}{\binom{n}{2}^2} UU^T UU^T \right) \quad (138)$$

$$\leq \lambda_{\max} \left( \frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} ((U_i - U_j)^T(U_i - U_j)) U(e_i - e_j)(e_i - e_j)^T U^T \right) + \lambda_{\max} \left( \frac{n^2}{\binom{n}{2}^2} UU^T UU^T \right) \quad (139)$$

$$= \max_x \frac{x^T}{\|x\|} \left( \frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} ((U_i - U_j)^T(U_i - U_j)) U(e_i - e_j)(e_i - e_j)^T U^T \right) \frac{x}{\|x\|} + \lambda_{\max} \left( \frac{n^2}{\binom{n}{2}^2} UU^T UU^T \right) \quad (140)$$

$$= \max_x \left( \frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} ((U_i - U_j)^T(U_i - U_j)) \frac{x^T}{\|x\|} U(e_i - e_j)(e_i - e_j)^T U^T \frac{x}{\|x\|} \right) + \lambda_{\max} \left( \frac{n^2}{\binom{n}{2}^2} UU^T UU^T \right) \quad (141)$$

$$\leq \max_x \left( \frac{\nu}{\binom{n}{2}} \sum_{(i,j) \in P} \frac{x^T}{\|x\|} U(e_i - e_j)(e_i - e_j)^T U^T \frac{x}{\|x\|} \right) + \lambda_{\max} \left( \frac{n^2}{\binom{n}{2}^2} UU^T UU^T \right) \quad (142)$$

$$= \lambda_{\max} \left( \frac{\nu}{\binom{n}{2}} \sum_{(i,j) \in P} U(e_i - e_j)(e_i - e_j)^T U^T \right) + \lambda_{\max} \left( \frac{n^2}{\binom{n}{2}^2} UU^T UU^T \right) \quad (143)$$

$$= \frac{\nu n}{\binom{n}{2}} \lambda_{\max}(UU^T) + \frac{n^2}{\binom{n}{2}^2} \lambda_{\max}(UU^T)^2 \text{ by Lemma 8.} \quad (144)$$

$$(145)$$

Equation (137) is because  $(U_i - U_j)^T(U_i - U_j) \in \mathbb{R}$ . Equation (142) is because  $(U_i - U_j)^T(U_i - U_j) \geq 0$  and  $\frac{x^T}{\|x\|} U(e_i - e_j)(e_i - e_j)^T U^T \frac{x}{\|x\|} \geq 0$ .

Now that we have bounds on  $\eta$  and  $\zeta$  and a simplified form for  $\lambda$ , we apply Theorem 1, completing the proof.

Now we explain how to get from these results to those in the main paper with the order terms. The  $O(\cdot)$  upper bound on the estimation error is easy to see. The value of  $C_1$  is given at the end of the proof of Theorem 1. The only remaining term to explain from the main paper is the upper bound of  $\frac{8 \log(2d/\delta)(6\eta + \lambda\zeta)}{3\lambda^2}$ , which gives us a lower bound on the number of samples required.



In particular,

$$\frac{8 \log(2d/\delta)(6\eta + \lambda\zeta)}{3\lambda^2} \quad (146)$$

$$= \frac{48 \log(2d/\delta)\eta}{3\lambda^2} + \frac{8 \log(2d/\delta)\zeta}{3\lambda} \quad (147)$$

$$= \frac{48 \log(2d/\delta) \binom{n}{2}^2}{3n^2 \lambda_{\min}(UU^T)^2} \left( \frac{\nu n \lambda_{\max}(UU^T)}{\binom{n}{2}} + \frac{n^2 \lambda_{\max}(UU^T)^2}{\binom{n}{2}^2} \right) + \frac{8 \log(2d/\delta) \binom{n}{2}}{3n \lambda_{\min}(UU^T)} \left( \nu + \frac{n \lambda_{\max}(UU^T)}{\binom{n}{2}} \right) \quad (148)$$

$$= \frac{48 \log(2d/\delta)}{3\lambda_{\min}(UU^T)^2} \left( \frac{\binom{n}{2} \nu \lambda_{\max}(UU^T)}{n} + \lambda_{\max}(UU^T)^2 \right) + \frac{8 \log(2d/\delta)}{3\lambda_{\min}(UU^T)} \left( \frac{\binom{n}{2} \nu}{n} + \lambda_{\max}(UU^T) \right) \quad (149)$$

$$\leq \frac{48 \log(2d/\delta)}{3\lambda_{\min}(UU^T)^2} \left( \frac{\binom{n}{2} \nu \lambda_{\max}(UU^T)}{n} + n \lambda_{\max}(UU^T)^2 \right) + \frac{8 \log(2d/\delta)}{3\lambda_{\min}(UU^T)} \left( \frac{\binom{n}{2} \nu}{n} + n \lambda_{\max}(UU^T) \right) \quad (150)$$

$$\leq \frac{48 \log(2d/\delta)}{3\lambda_{\min}(UU^T)^2} (n\nu \lambda_{\max}(UU^T) + n \lambda_{\max}(UU^T)^2) + \frac{48 \log(2d/\delta)}{3\lambda_{\min}(UU^T)} (n\nu + n \lambda_{\max}(UU^T)) \quad (151)$$

$$\leq \frac{48 \log(2d/\delta) n \nu}{3} \left( \frac{\lambda_{\max}(UU^T)}{\lambda_{\min}(UU^T)^2} + \frac{\lambda_{\max}(UU^T)^2}{\lambda_{\min}(UU^T)^2} + \frac{1}{\lambda_{\min}(UU^T)} + \frac{\lambda_{\max}(UU^T)}{\lambda_{\min}(UU^T)} \right) \text{ since } \nu \geq 1 \quad (152)$$

$$\leq \frac{2 * 48 \log(2d/\delta) n \nu}{3} \left( \frac{\lambda_{\max}(UU^T)}{\lambda_{\min}(UU^T)^2} + \frac{\lambda_{\max}(UU^T)^2}{\lambda_{\min}(UU^T)^2} + \frac{1}{\lambda_{\min}(UU^T)} \right) \text{ since } \frac{\lambda_{\max}(UU^T)}{\lambda_{\min}(UU^T)} \geq 1 \quad (153)$$

$$= C_3 \log(2d/\delta) n \nu \left( \frac{\lambda_{\max}(UU^T)}{\lambda_{\min}(UU^T)^2} + \frac{\lambda_{\max}(UU^T)^2}{\lambda_{\min}(UU^T)^2} + \frac{1}{\lambda_{\min}(UU^T)} \right) \quad (154)$$

where  $C_3 = 2 * 48/3$ . We remark that the assumption that  $\nu \geq 1$  was made to simplify the upper bound and is not required.  $\square$

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

**Proposition 9.** *Given an arbitrary rank- $d$ ,  $d \times n$  matrix  $\tilde{U}$ , let  $U$  be its centered version, i.e.  $U = \tilde{U}(I - \frac{1}{n} \mathbb{1}\mathbb{1}^T)$ . Then  $\sigma_{\min}(U) = 0$  if and only if the all-ones vector is in the row space of  $\tilde{U}$ .*

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

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

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

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

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

## 11.2. Discussion of Corollary 1.1 as compared to related work

While our sample complexity theorem for MLE of the parameters of FBTL is novel to the best of our knowledge, there are some related results that merit a comparison. First, there is a result in (Saha and Rajkumar, 2018) that gives sample complexity results for a different estimator of FBTL parameters under a substantially different sampling model. In particular, they only allow pairs to be sampled from a graph, and then for each sampled pair they observe a fixed number of pairwise comparisons. In their results one can see that as the number of pairs sampled increases, their error upper bound increases and the probability of their resulting bound also decreases. In contrast, our analysis shows that our error bound decreases as  $m$  increases, and the probability of our resulting bound remains constant.

Second, we can also attempt a comparison to the bounds for BTL without features in (Negahban et al., 2012), despite the fact that with standard basis features, our bound does not apply because  $\lambda = 0$ . Assuming that  $\exp(b^*)/\lambda$  is a constant in our bound and that  $\nu\bar{\lambda}$  is a constant, we roughly have an error bound of  $O(1)$  given  $m = \Theta(n^2(\beta^2 + \beta)d \log(d/\delta))$  samples. The result in (Negahban et al., 2012) instead has that  $m = \Theta(d^2 \log d)$  gives an error bound of  $O(1)$  with probability  $1 - \frac{2}{d}$ , recalling that in their setting  $d = n$ . So if we can tighten bounds that require  $\beta$  in our proof, our results may compare favorably.

Recall the definition of  $\beta$  in Equation (4):  $\beta := \max_{(i,j) \in P} \|U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\|_{\infty}$ . In our proof, we use this to bound differences between feature vectors at Equation (66). In particular, we bound  $\frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} \left( (U_i^{\tau(i,j)})^{(k)} - (U_j^{\tau(i,j)})^{(k)} \right)^2 \leq \beta^2$ . If we instead directly made the assumption that

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

we could replace  $\beta$  with  $\tilde{\beta}$  directly in our bounds. Assume  $\tilde{\beta} \leq 1/n^2$ . Then our sample complexity would reduce to  $m = \Theta(d \log(d/\delta)) = \Theta(d \log(2d^2)) = \Theta(d \log(d))$  where recall  $\delta = \frac{2}{d}$ , beating the complexity in (Negahban et al., 2012). However, it is not clear in general what impact the assumption that  $\tilde{\beta} \leq 1/n^2$  would have on the minimum eigenvalue of  $UU^T$ . Indeed, the standard basis vectors are a special case where  $\tilde{\beta} \leq 1/n$ , and as we pointed out, for this special case  $\lambda = 0$ .

Third, although there are crucial differences between our model and the model in (Shah and Wainwright, 2017) that make a direct comparison impossible, we attempt to roughly compare results. The first difference is that they assume the feature vectors of the items are standard basis vectors, which means our bounds do not apply just as in the comparison with (Negahban et al., 2012). The second difference, perhaps the most crucial, is that we make different assumptions about how the intransitive pairwise comparisons are related to the ranking. In (Shah and Wainwright, 2017), the items are ranked based on the probability that one item beats any other item chosen uniformly at random. There are scenarios where the true ranking in our model is not the same as the true ranking in (Shah and Wainwright, 2017). The third difference is that we assume that pairs are drawn uniformly at random, whereas they assume each pair  $(i, j) \in P$  is drawn  $x_{i,j}$  times where  $x_{i,j} \sim \text{Binom}(r, p)$  for  $r, p > 0$ .

Their result (Theorem 2) roughly says with probability  $1/n^{13}$ , if the gap between a pair of consecutively ranked items' scores is at least  $\sqrt{\log n / (npr)}$ , then their algorithm learns the ranking exactly. We compare to our Corollary 1.3 with  $k = 1$  and  $\delta = \frac{1}{n^{13}}$  though again we emphasize an exact comparison is impossible because our model is not a special case of theirs or vice versa. Our corollary says with enough samples with high probability, we learn the ranking exactly. On average, their sampling method will see  $O(n^2rp)$  samples, so a reasonable way to compare results is to show the required number of samples in our method is comparable to  $O(n^2rp)$ . If we assume that  $\beta, \eta, \zeta, \lambda$ , and  $M$  are all constant,  $\alpha_k = \sqrt{\log n / (npr)}$  which is their assumed gap between scores, and  $d = n$ , the number of samples we require is  $\max\{n \log(n * n^{13}), \log(n), n \log(n * n^{13})npr / \log(n)\} = O(n^2pr)$ , matching their bounds.

Fourth, the set-up of (Heckel et al., 2019) is the same as (Shah and Wainwright, 2017) except it considers the adaptive setting. If the gaps of the utilities of consecutively ranked items are constant and denoted by  $\Delta$ , then under the same assumptions in the discussion about (Shah and Wainwright, 2017), our Corollary 1.3 is slightly better by a log factor than their Theorem 1a:  $O(\log(n/\delta)n/(\Delta^2))$  vs.  $O(\log(n/\delta)n \log(2 \log(2/\Delta))/(\Delta^2))$ . However, if many gaps between scores are large and only some gaps between scores are small, their adaptive method is better than our Corollary 1.3. This is not surprising since they can adaptively choose which pair to sample next based on the past pairwise comparisons, whereas we consider the passive setting.

### 11.3. Proof of Corollary 1.2

**Corollary 9.1** (Restatement of Corollary 1.2). *Assume the set-up stated in the beginning of Section 2. Assume that for any  $(i, j) \in P$ ,  $|\tau(i, j)| = 1$ . Partition  $P = \sqcup_{k=1}^d P_k$  into  $d$  sets where  $(i, j) \in P_k$  if  $\tau(i, j) = \{k\}$  for  $k \in [d]$ . Let  $\epsilon := \min_{(i,j) \in P} \|U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\|_{\infty}$ . Then*

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

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

and

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

Furthermore, let

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

and let

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

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

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

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

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

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

$$\mathbb{E}((U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T) = \frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T \quad (155)$$

$$= \frac{1}{\binom{n}{2}} \sum_{k=1}^d \sum_{(i,j) \in P_k} (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T \quad (156)$$

$$= \frac{1}{\binom{n}{2}} \sum_{k=1}^d \left( \sum_{(i,j) \in P_k} (U_i^{(k)} - U_j^{(k)})^2 \right) \text{diag}(e_k), \quad (157)$$

which is a diagonal matrix. Therefore,

$$\lambda = \frac{1}{\binom{n}{2}} \min_{k \in [d]} \left( \sum_{(i,j) \in P_k} (U_i^{(k)} - U_j^{(k)})^2 \right) \quad (158)$$

$$\geq \frac{\epsilon^2}{\binom{n}{2}} \min_{k \in [d]} |P_k|. \quad (159)$$

Second, we simplify and bound  $\zeta$ . Since  $|\tau(k, j)| = 1$  for all  $k, j \in P$ , let  $U_i^{\tau(k, j)}$  denote the coordinate of  $U_i$  corresponding to the only element in  $\tau(k, j)$ . Define  $e_{\tau(k, j)}$  similarly, which is one of the standard basis vectors. From the proof of bounding  $\lambda$  in Equations (155) to (157), we have  $\mathbb{E}((U_i^{\tau(i, j)} - U_j^{\tau(i, j)})(U_i^{\tau(i, j)} - U_j^{\tau(i, j)})^T) =$

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

$$\zeta = \max_{(\ell,p) \in P} \lambda_{\max}(\mathbb{E}((U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T) - (U_\ell^{\tau(\ell,p)} - U_p^{\tau(\ell,p)})(U_\ell^{\tau(\ell,p)} - U_p^{\tau(\ell,p)})^T) \quad (160)$$

$$= \max_{(\ell,p) \in P} \lambda_{\max} \left( \frac{1}{\binom{n}{2}} \sum_{k=1}^d \left( \sum_{(i,j) \in P_k} (U_i^{(k)} - U_j^{(k)})^2 \right) \text{diag}(e_k) - (U_\ell^{\tau(\ell,p)} - U_p^{\tau(\ell,p)})(U_\ell^{\tau(\ell,p)} - U_p^{\tau(\ell,p)})^T \right) \quad (161)$$

$$= \max_{(\ell,p) \in P} \lambda_{\max} \left( \frac{1}{\binom{n}{2}} \sum_{k=1}^d \left( \sum_{(i,j) \in P_k} (U_i^{(k)} - U_j^{(k)})^2 \right) \text{diag}(e_k) - (U_\ell^{(\tau(\ell,p))} - U_p^{(\tau(\ell,p))})^2 \text{diag}(e_{\tau(\ell,p)}) \right) \quad (162)$$

$$\leq \beta^2 \left( \max_{k \in [d]} \left( \frac{|P_k|}{\binom{n}{2}} + 1 \right) \right) \quad (163)$$

$$(164)$$

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

Third, we simplify  $\eta$ . First notice from the proof of bounding  $\lambda$  from Equations (155) to (157),

$$\left( \mathbb{E}(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T \right)^2 = \left( \frac{1}{\binom{n}{2}} \sum_{k=1}^d \left( \sum_{(i,j) \in P_k} (U_i^{(k)} - U_j^{(k)})^2 \right) \text{diag}(e_k) \right)^2 \quad (165)$$

$$= \frac{1}{\binom{n}{2}^2} \sum_{k=1}^d \left( \sum_{(i,j) \in P_k} (U_i^{(k)} - U_j^{(k)})^2 \right)^2 \text{diag}(e_k), \quad (166)$$

since the matrices above are diagonal.

Also,

$$\mathbb{E}(((U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T)^2) \quad (167)$$

$$= \mathbb{E}((U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T) \quad (168)$$

$$= \frac{1}{\binom{n}{2}} \sum_{k=1}^d \sum_{(i,j) \in P_k} (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T (U_i^{\tau(i,j)} - U_j^{\tau(i,j)})(U_i^{\tau(i,j)} - U_j^{\tau(i,j)})^T \quad (169)$$

$$= \frac{1}{\binom{n}{2}} \sum_{k=1}^d \left( \sum_{(i,j) \in P_k} (U_i^{(k)} - U_j^{(k)})^4 \right) \text{diag}(e_k), \quad (170)$$

For any random variable  $X$ , we have

$$\mathbb{E}(X - \mathbb{E}(X))^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2. \quad (171)$$

Therefore,

$$\eta = \sigma_{\max} \left( \frac{1}{\binom{n}{2}} \sum_{k=1}^d \left( \sum_{(i,j) \in P_k} (U_i^{(k)} - U_j^{(k)})^4 \right) \text{diag}(e_k) - \frac{1}{\binom{n}{2}^2} \sum_{k=1}^d \left( \sum_{(i,j) \in P_k} (U_i^{(k)} - U_j^{(k)})^2 \right)^2 \text{diag}(e_k) \right) \quad (172)$$

$$= \frac{1}{\binom{n}{2}} \sigma_{\max} \left( \sum_{k=1}^d \left( \sum_{(i,j) \in P_k} (U_i^{(k)} - U_j^{(k)})^4 - \frac{1}{\binom{n}{2}} \left( \sum_{(i,j) \in P_k} (U_i^{(k)} - U_j^{(k)})^2 \right)^2 \right) \text{diag}(e_k) \right) \quad (173)$$

$$\leq \frac{\beta^4}{\binom{n}{2}} \max_{k \in [d]} \left( |P_k| + \frac{|P_k|^2}{\binom{n}{2}} \right) \quad (174)$$

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

The remainder of the corollary follows by applying the bounds on  $\lambda, \zeta$  and  $\eta$  to Theorem 1.

Now we explain how to get from these results to those in the main paper with the order terms. The  $O(\cdot)$  upper bound on the estimation error is easy to see. The value of  $C_1$  is given at the end of the proof of Theorem 1. Finally, it is easy to see  $C_4 = 48/3$  in the main paper.  $\square$

#### 11.4. Tightening the bounds of Corollary 1.2

Still in the setting where the selection function chooses one coordinate per pair, assume  $|P_i| \approx |P_j|$  for all  $i, j \in [d]$ , where  $P_i$  is defined in Corollary 1.2. Then, as we have stated in the main text,  $\lambda, \eta, \zeta = O(1/d)$ , and so by Corollary 1.2,  $\Omega(d^3 \log(d/\delta))$  samples ensures the estimation error is  $O(1)$ . However, by tightening a bound used in the proof of Theorem 1, we can show  $\Omega(d^2 \log(d/\delta))$  samples ensures the estimation error is  $O(1)$ .

Recall the definition of  $\beta$  in Equation (4):  $\beta := \max_{(i,j) \in P} \|U_i^{\tau(i,j)} - U_j^{\tau(i,j)}\|_{\infty}$ . In our proof, we use this to bound differences between feature vectors at Equation (66). In particular, for  $k \in [d]$  we bound  $\frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} \left( (U_i^{\tau(i,j)})^{(k)} - (U_j^{\tau(i,j)})^{(k)} \right)^2 \leq \beta^2$ . For any  $k \in [d]$ , since  $|P_i| \approx |P_j|$  for all  $i, j \in [d]$ , each coordinate is chosen approximately  $\binom{n}{2}/d$  times. Therefore,  $\frac{1}{\binom{n}{2}} \sum_{(i,j) \in P} \left( (U_i^{\tau(i,j)})^{(k)} - (U_j^{\tau(i,j)})^{(k)} \right)^2 \leq \beta^2 d$  since only  $\binom{n}{2}/d$  of the  $\binom{n}{2}$  terms in the sum are non-zero. We can now replace  $\beta$  with  $\beta/\sqrt{d}$  in Corollary 1.2. Therefore,  $\Omega(d^2 \log(d/\delta))$  samples ensures the estimation error is  $O(1)$  since  $\lambda, \eta, \zeta = O(1/d)$ .

## 12. Proof of Corollary 1.3

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

**Corollary 9.2** (restatement of Corollary 1.3: sample complexity of learning the ranking). *Assume the set-up of Theorem 1. Pick  $k \in [\binom{n}{2}]$ . Let  $\alpha_k$  be the  $k$ -th smallest number in  $\{|\langle w^*, U_i - U_j \rangle| : (i, j) \in P\}$ . Let  $M := \max_{i \in [n]} \|U_i\|_2$ . Let  $\gamma^* : [n] \rightarrow [n]$  be the ranking obtained from  $w^*$  by sorting the items by their full-feature utilities  $\langle w^*, U_i \rangle$  where  $\gamma^*(i)$  is the position of item  $i$  in the ranking. Define  $\hat{\gamma}$  similarly but for the estimated ranking obtained from the MLE estimate  $\hat{w}$ . Let  $\delta > 0$ . Let*

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

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

and

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

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

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

$$\|w^* - \hat{w}\|_2 \leq \frac{4(1 + \exp(b^*))^2}{\exp(b^*)\lambda} \sqrt{\frac{3\beta^2 \log(4d/\delta)d + 4\sqrt{d}\beta \log(4d/\delta)}{6m}} \quad (175)$$

$$\leq \frac{\alpha_k}{2M} \quad (176)$$

by definition of  $m$ .

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

$$|\langle w^* - \hat{w}, U_i \rangle| \leq \|w^* - \hat{w}\|_2 \|U_i\|_2 \text{ by Cauchy-Schwarz} \quad (177)$$

$$\leq \frac{\alpha_k \|U_i\|_2}{2M} \quad (178)$$

$$\leq \frac{\alpha_k}{2}. \quad (179)$$

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

$$\langle w^*, U_i \rangle - \frac{\alpha_k}{2} \leq \langle \hat{w}, U_i \rangle \leq \langle w^*, U_i \rangle + \frac{\alpha_k}{2}. \quad (180)$$

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

By applying Equation (180) and using the fact  $\langle w^*, U_i \rangle - \langle w^*, U_j \rangle \leq 0$ , we have

$$\langle \hat{w}, U_i \rangle \leq \langle w^*, U_i \rangle + \frac{\alpha_k}{2} \text{ by Equation (180)} \quad (181)$$

$$= \langle w^*, U_i \rangle - \langle w^*, U_j \rangle + \langle w^*, U_j \rangle + \frac{\alpha_k}{2} \quad (182)$$

$$\leq -\alpha_k + \langle w^*, U_j \rangle + \frac{\alpha_k}{2} \text{ since } (i, j) \in P_{\alpha_k} \text{ and since } \langle w^*, U_i \rangle - \langle w^*, U_j \rangle \leq 0 \quad (183)$$

$$\leq \langle w^*, U_j \rangle - \frac{\alpha_k}{2} \quad (184)$$

$$\leq \langle \hat{w}, U_j \rangle \text{ by Equation (180)}. \quad (185)$$

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

Now we explain how to get from these results to those in the main paper with the order terms. The value of  $C_1$  and  $C_2$  are given at the end of the proof of Theorem 1. It is easy to see that  $C_5 = 64 * 4 * 2^4/6$ . □

## 13. Synthetic Experiments

Code is available at [https://github.com/Amandarg/salient\\_features](https://github.com/Amandarg/salient_features).

### 13.1. Plot of Parameters in Theorem 1

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

In the experiments, the ambient dimension  $d = 10$  and the number of items  $n = 100$ . We repeat the following 10 times: sample  $U$  and  $w^*$ , and use this  $U$  and  $w^*$  while varying  $t \in [d]$  to compute all of the parameters of interest and intransitivity

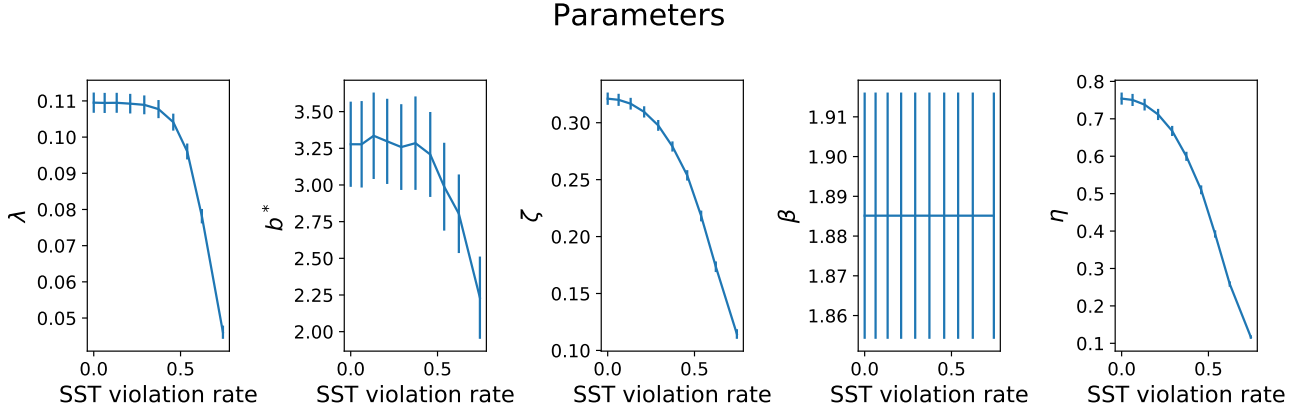


Figure 5. The parameters of Theorem 1 for the top- $t$  selection function as a function of the average strong stochastic transitivity violation rate over the 10 experiments. The average over 10 experiments where a new  $U$  and  $w^*$  are drawn each time is depicted. The bars represent the standard error over the 10 experiments.

rates. The  $x$ -axis of each plot is the average strong stochastic transitivity (SST) violation rate defined in Section 4.1 where the average is taken over the 10 experiments. From Figure 2, intransitives decrease as  $t$  increases, so the  $x$ -axis in Figures 5 and 6 could roughly, but not exactly, be replaced with  $t$ , where  $t$  is decreasing from 10 to 1. The  $y$ -axis on the plots depict the average value and the bars represent the standard error over the 10 experiments.

Figure 5 shows the parameters in Theorem 1. Larger  $\lambda$  means smaller sample complexity, whereas smaller  $b^*$ ,  $\zeta$ ,  $\beta$  and  $\eta$  means smaller sample complexity.

Recall in the Supplement re-statement of Theorem 1, the number of samples  $m$  required in the theorem is

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

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

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

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

### 13.2. Additional Synthetic Experiments and Details

First we define the Kendall tau correlation. It is used in both Sections 4.1 and 4.2, and is defined as follows. Let  $\gamma, \rho : [n] \rightarrow [n]$  be two rankings on  $n$  items where  $\gamma(i)$  and  $\rho(i)$  is the position of item  $i$  in the ranking. Let  $A = \sum_{(i,j) \in P} \mathbb{1}_{\{(\sigma(i) - \sigma(j))(\rho(i) - \rho(j)) > 0\}}$ , respectively  $D = \sum_{(i,j) \in P} \mathbb{1}_{\{(\sigma(i) - \sigma(j))(\rho(i) - \rho(j)) \leq 0\}}$ , be the number of pairs of items that  $\sigma$  and  $\rho$  agree, respectively disagree, on the relative ordering. Then the Kendall tau correlation of  $\rho$  and  $\gamma$  is

$$KT(\gamma, \rho) := \frac{A - D}{\binom{n}{2}}. \quad (186)$$

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

## Samples and Bound

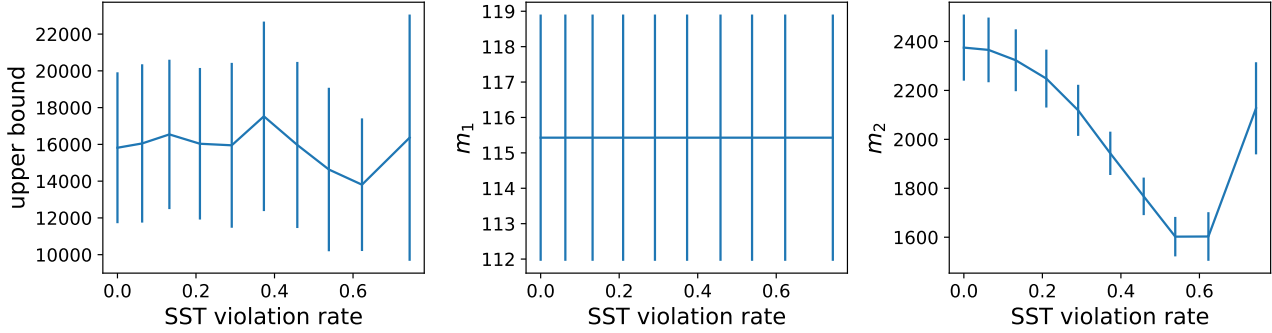


Figure 6. Number of samples  $m_1$  and  $m_2$  and upper bound on estimation error from Theorem 1 for the top- $t$  selection function as a function of the average strong stochastic transitivity violation rate over the 10 experiments. The average over 10 experiments where a new  $U$  and  $w^*$  are drawn is depicted. The bars represent the standard error over the 10 experiments.

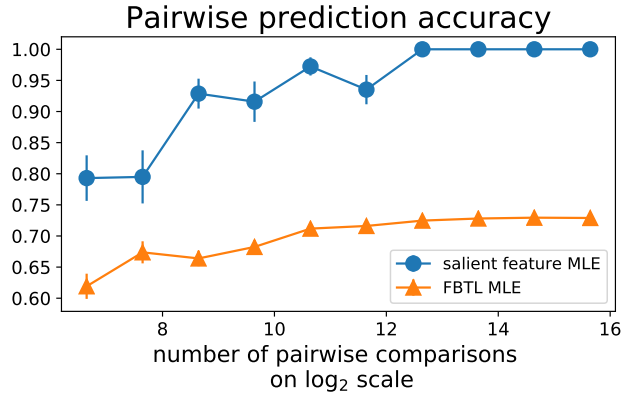


Figure 7. Pairwise prediction accuracy as a function of the number of samples, which are on the logarithmic scale, where the pairwise comparisons are sampled from the salient feature preference model with the top-1.

with the top-1 selection function, and repeat 10 times. Figure 7 shows the average pairwise prediction accuracy, which is defined as

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

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

Third, see Figures 8 and 9 for plots investigating model misspecification. In particular, we use the same experimental set-up as in Section 4.1 except that in Figure 9 the salient feature preference model with the top-3 selection function is used to generate the preference data. We fit the MLE for the salient feature preference model for the top- $t$  selection function for all  $t \in [d]$  for both plots. The FBTL model is equivalent to when  $t = 10$ .

In Figure 8, we see that the model is very sensitive to the choice of  $t$ . As we would expect,  $t = 2$  has the second smallest error when the number of samples exceed  $2^{10}$ .

In Figure 9, we see that the model is still sensitive to the choice of  $t$ , but not as sensitive as in Figure 8. In this case, we can not only overestimate  $t$ , i.e.  $t > 3$ , but underestimate  $t$ , i.e.  $t < 3$ . We see that  $t = 2$  and  $t = 4$ —the two values of  $t$  closest to the truth of  $t = 3$ —have roughly the same error. Interestingly,  $t = 1$  has the worst performance.



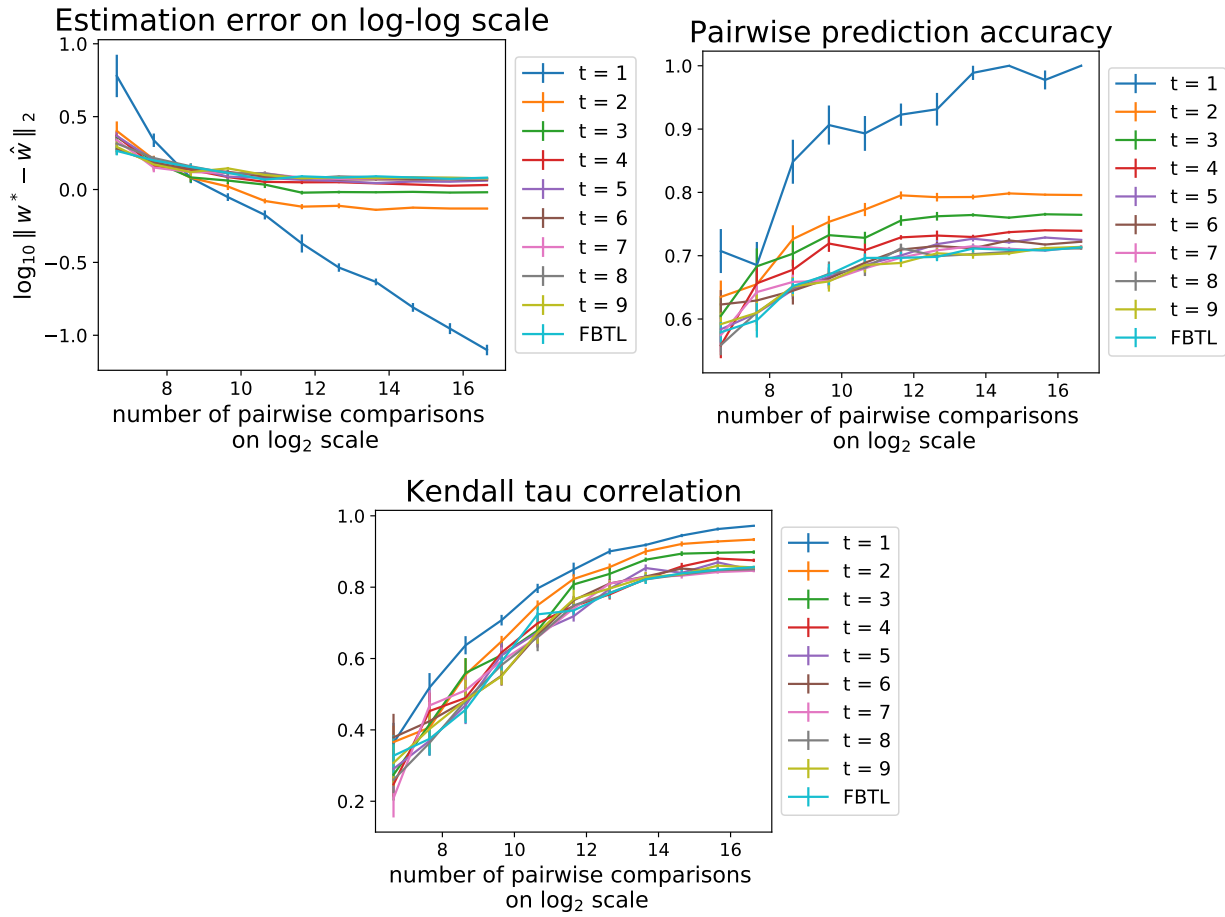


Figure 8. These plots investigate model misspecification. The true generative model for the pairwise preference data is the salient feature preference model with the top-1 selection function. The coordinates of  $U$  and  $w$  are sampled from a Gaussian as described in the main text. The MLEs for the salient feature preference model with the top- $t$  selection function for  $t \in [d]$  is shown.

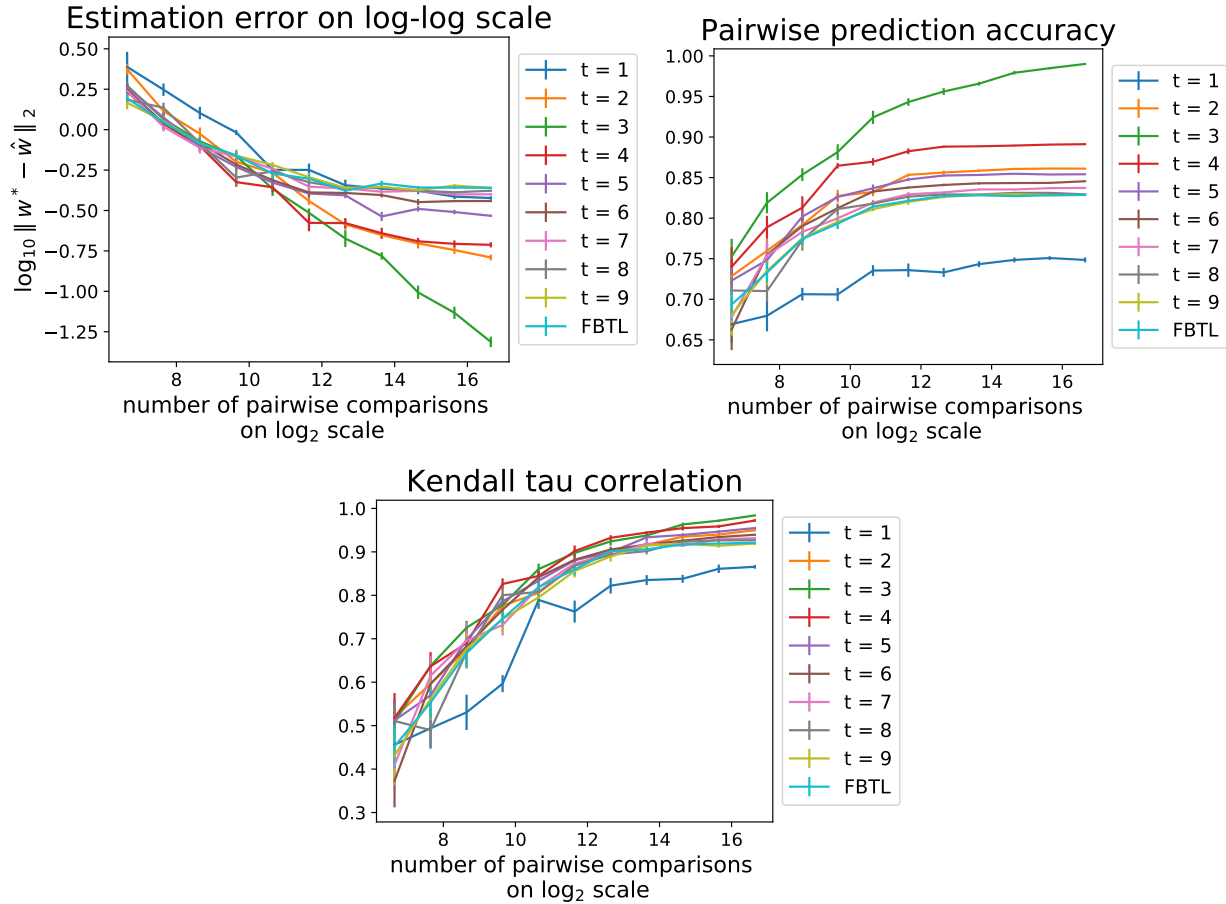


Figure 9. These plots investigate model misspecification. The true generative model for the pairwise preference data is the salient feature preference model with the top-3 selection function. The coordinates of  $U$  and  $w$  are sampled from a Gaussian as described in the main text. The MLEs for the salient feature preference model with the top- $t$  selection function for  $t \in [d]$  is shown.

## 14. Real Data Experiments

Code is available at [https://github.com/Amandarg/salient\\_features](https://github.com/Amandarg/salient_features).

### 14.1. Algorithm implementation

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

- **RankNet:** We use the RankNet implementation found at <https://github.com/airalcorn2/RankNet>, which uses Keras. However, we use the Adam optimizer with default parameters except with a learning rate of 0.0001. We also add an  $\ell_2$  penalty to the weights.
- **Salient feature preference model and FBTL:** We use `sklearn`'s logistic regression solver. In particular, we set `tol = 1e - 10` and `max_iter = 10000`. Furthermore, we do not fit an intercept. We use the default `liblinear` solver for real data experiments, and the `sag` solver for synthetic data experiments since we do not use regularization. All other parameters use the default values.
- **Ranking SVM:** We use `sklearn`'s `LinearSVC` solver with the same parameters as above. In particular, we do not fit an intercept.

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

### 14.2. District compactness experiments

We refer the reader to (Kaufman et al., 2017) for the full details about the district compactness data, but provide relevant details here. We obtained the data by contacting the authors.

### 14.3. Pairwise comparison description

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

### 14.4. $k$ -wise rankings for $k > 2$ description

There are 8 sets of  $k$ -wise ranking data. In many cases, the feature data for some districts are missing entirely, so in our own experiments, we throw out any district without feature data. Recall, we use the  $k$ -wise ranking data for validation and testing, so we also remove any districts present in the training set.

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

---

<sup>1</sup><https://arc-ts.umich.edu/greatlakes/>

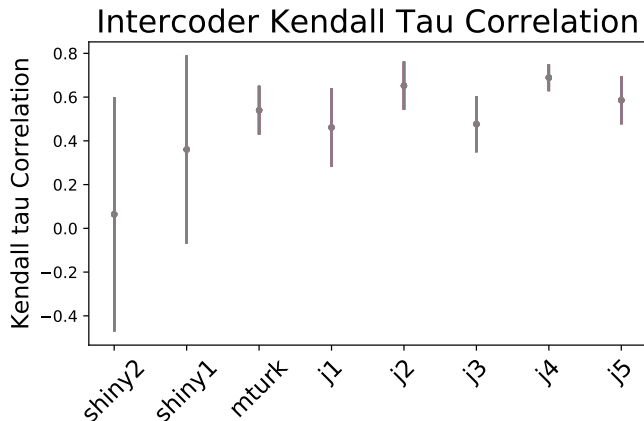


Figure 10. For each of the  $k$ -wise ranking data sets, the average agreement between people in terms of the Kendall tau correlation is shown.

See Figure 10 which depicts the average Kendall tau correlation between pairs of rankings in a  $k$ -wise ranking data set and the standard deviation. Recall the Kendall tau correlation,  $KT(\cdot, \cdot)$ , is defined in Equation (186). This plot shows roughly how much people agree with each other, where higher values mean more agreement. In particular, suppose there are  $N$   $k$ -wise rankings given by  $\sigma_1, \dots, \sigma_N$ . Then the average Kendall tau correlation for the  $N$  rankings is

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

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

The districts used in `shiny1` and `shiny2` are the same, and these districts also comprise one of the UG1 data sets as well. However, the districts in `mturk` are disjoint from the rest of the  $k$ -wise ranking sets. In addition, `mturk` has relatively low intercoder variability. For these two reasons, we decided to use `mturk` as our validation set. We decided to keep `shiny1` and `shiny2` separate since the original authors did and also since they are comprised of different groups of people resulting in different behavior, e.g., `shiny1` has a higher average intercoder Kendall tau correlation than `shiny2`.

#### 14.5. Data preprocessing

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

#### 14.6. Experiment details

The hyperparameters for the salient feature preference model with the top- $t$  selection function are  $t$  and the  $\ell_2$  regularization parameter  $\mu$ . The hyperparameter for FBTL is the  $\ell_2$  regularization parameter  $\mu$ . For Ranking SVM, the only hyperparameter is  $C$  which controls the penalty for violating the margin. We vary  $t \in [d]$  where  $d = 27$  since there are 27 features. We vary  $\mu$  and  $C$  in  $\{.00001, .0001, .001, .01, .1, 1, 10, 100, 1000, 10000, 100000, 1000000\}$ .

The hyperparameters for RankNet include the  $\ell_2$  regularization parameter  $\mu$  and number of nodes in the hidden layer. We use one hidden layer. We varied the number of nodes in the single hidden unit in  $\{5 * \dots\}$

$i : i \in [19]$ . We use a batch size of 250, and we use 800 epochs. Initially, we varied  $\mu$  also in  $\{.00001, .0001, .001, .01, .1, 1, 10, 100, 1000, 10000, 100000, 1000000\}$ , but as we will discuss in the next section we decided to vary  $\mu$  in  $\{.00001, .0001, .001, .01, .1, 1, 10\}$ .

#### 14.7. Best performing hyperparameters

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

For FBTL, the best performing hyperparameter is  $\mu = 100000$ . The average Kendall tau correlation of the estimated ranking to each individual ranking in `mturk` is 0.38 with a standard deviation of 0.05. The pairwise comparison accuracy on the training set is 56%, which is defined in Section 13.2 of the Supplement. Although the regularization strength is large, the norm of the estimated judgement vector is .015. The largest coordinate of the judgement vector in absolute value is .005 and the smallest is .0001.

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

Figure 11 shows how often each of the 27 features are selected by the top-2 selection function over unique pairwise comparisons in the training data. Notice that `var_xcoord` and `circle_area` are never selected. The learned weights for those features in the FBTL model when all the features are used are 2 of the top 3 features with the smallest weights, so these features play a relatively insignificant role when all the features are used any way.

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

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

#### 14.8. Zappos experiments

We refer the reader to (Yu and Grauman, 2014; 2017) for the full details about the UT\_Zappos50k data set but provide relevant details here. The data can be found at <http://vision.cs.utexas.edu/projects/finegrained/utzap50k/>.

#### 14.9. Pairwise comparison data description

The UT\_Zappos50K data set consists of pairwise comparisons on images of shoes and 960 extracted color and vision features for each shoe (Yu and Grauman, 2014; 2017). Given images of two different shoes and an attribute from {“open,” “pointy,” “sporty,” “comfort”}, respondents were asked to pick which shoe exhibits the attribute more. The data consists of both easier, coarse questions, i.e. based on comfort, pick between a slipper or high-heel, and also harder, fine grained questions i.e. based on comfort, pick between two slippers. Each pairwise comparison is asked to 5 different people, and the confidence of each person’s answer is also collected.

There are 2,863 unique pairwise comparisons involving 5,319 shoes for open, 2,700 unique pairwise comparisons involving 5,028 shoes for pointy, 2,766 unique pairwise comparisons involving 5,144 shoes for sporty, and 2,756 unique pairwise comparisons involving 5,129 shoes for comfort. For each attribute, 86% of unique pairwise comparisons involve an item that is in no other pairwise comparison regarding that attribute. Also, for each attribute, nearly 93% of items only appear in

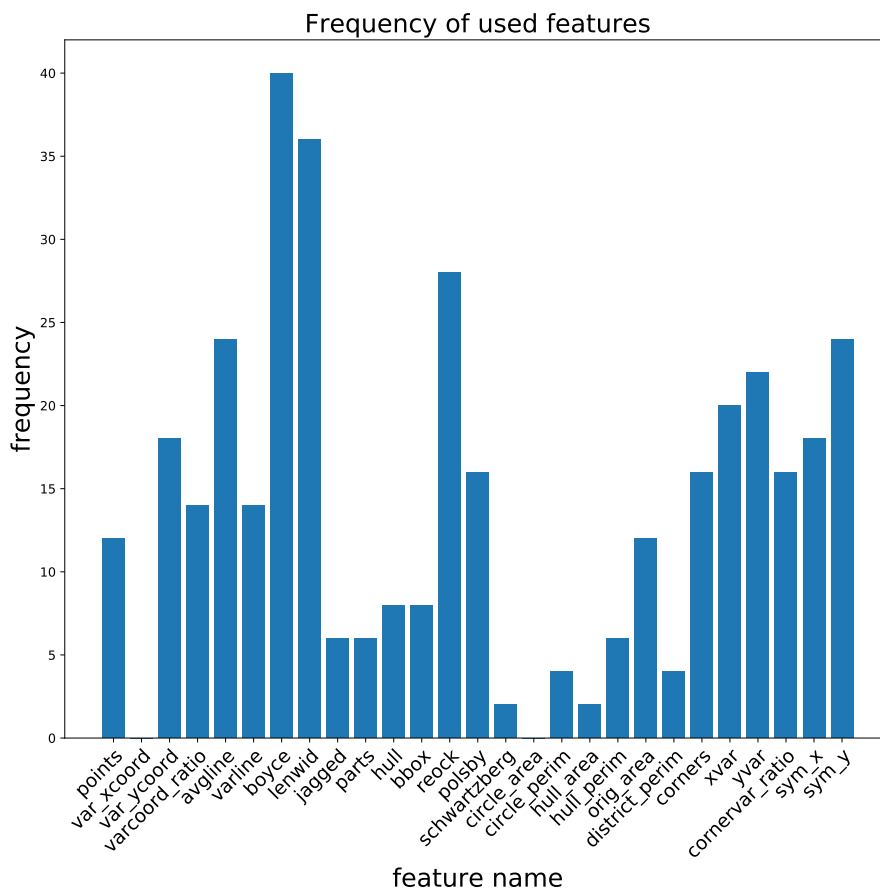


Figure 11. The frequency that the top-2 selection function chooses each feature over unique pairwise comparisons in the training data.

Table 3. Statistics about the best performing  $t$  for the salient feature preference model with the top- $t$  selection function on the validation set over 10 train/validation/test splits for UT\_Zappos50k.

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

Table 4. Statistics about the best performing  $\mu$  for the salient feature preference model on the validation set over 10 train/validation/test splits for UT\_Zappos50k.

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

one pairwise comparison. In light of this, an algorithm like (Chen and Joachims, 2016b) will likely not work well since (1) this model requires learning a set of parameters for each item and (2) the model does not work for unseen items, i.e., we must ensure that items in testing also appear in training to evaluate the model.

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

#### 14.10. Data pre-processing

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

#### 14.11. Experiment details

The hyperparameters for the salient feature preference model with the top- $t$  selection function are  $t$  and the  $\ell_2$  regularization parameter  $\mu$ . The hyperparameter for FBTL is the  $\ell_2$  regularization parameter  $\mu$ . For Ranking SVM, the only hyperparameter is  $C$  which controls the penalty for violating the margin. We vary  $t \in \{10 * i : i \in [99]\}$  since there are 990 features. We vary  $\mu$  and  $C$  in  $\{.000001, .00001, .0001, .001, .01, .1\}$ . For RankNet, the hyperparameters are  $\mu$  and the number of nodes in the hidden layer. We vary  $\mu$  in  $\{.05, .1, .15\}$  and the nodes in  $\{50, 250, 500\}$ . We choose these values of  $\mu$  to try since on validation sets, it appeared that any value less than .05 was over fitting (train accuracy was in the 90%*s* but validation accuracy was in the 70%*s*) and values above .15 were not learning a good model (train accuracy was in the 60%*s*). We only search over these hyperparameters due to time constraints. We use ten 70% train, 15% validation, and 15% test split.

#### 14.12. Best performing hyperparameters

Because the pairwise comparisons are either “coarse” or “fine-grained,” we pick the best hyperparameters based on the average of the pairwise comparison accuracy on the “coarse” questions and the “fine-grained” questions on the validation set. See Table 3 for statistics about the best performing  $t$  for the salient feature preference model with the top- $t$  selection function on the validation set over 10 train/validation/test splits. See Tables 4, 5, 7 for statistics about the best performing  $\mu$  for the salient feature preference model, FBTL model, and RankNet on the validation set over the 10 train/validation/test splits. See Table 6 for statistics about the best performing  $C$  for Ranking SVM on the validation set over the over the 10 train/validation/test splits. See Table 8 for the best performing number of nodes in the hidden layer on the validation set over the 10 splits. We also report the average pairwise accuracy, which has been defined in the main text, on the validation set for all algorithms in Table 9.

Table 5. Statistics about the best performing  $\mu$  for FBTL on the validation set over 10 train/validation/test splits for UT Zappos50k.

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

Table 6. Statistics about the best performing  $C$  for Ranking SVM on the validation set over 10 train/validation/test splits for UT Zappos50k.

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

Table 7. Statistics about the best performing  $\mu$  for RankNet on the validation set over 10 train/validation/test splits for UT Zappos50k.

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

Table 8. Statistics about the best performing number of nodes in the hidden layer for RankNet on the validation set over 10 train/validation/test splits for UT Zappos50k.

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

Table 9. Average pairwise prediction accuracy over 10 train/validation/test splits on the validation sets by attribute for UT Zappos50k.  $C$  stands for coarse and  $F$  stands for fine grained. The number in parenthesis is the standard deviation.

Model:	open- $C$	pointy- $C$	sparty- $C$	comfort- $C$	open- $F$	pointy- $F$	sparty- $F$	comfort- $F$
Salient features	0.75 (.01)	0.8 (.01)	0.79 (.02)	0.77 (.03)	0.64 (.03)	0.6 (.03)	0.62 (.03)	0.66 (.03)
FBTL	0.75 (.02)	0.8 (.01)	0.79 (.01)	0.77 (.02)	0.63 (.03)	0.59 (.03)	0.6 (.02)	0.62 (.03)
Ranking SVM	0.75 (.02)	0.8 (.02)	0.8 (.01)	0.77 (.02)	0.62 (.04)	0.59 (.03)	0.6 (.02)	0.62 (.04)
RankNet	0.75 (.02)	0.78 (.03)	0.78 (.01)	0.76 (.02)	0.67 (.03)	0.61 (.04)	0.61 (.02)	0.64 (.03)