Supplementary material for
Estimation from Pairwise Comparisons: Sharp Minimax Bounds with Topology Dependence

A Proof of Theorem 1

We split the proof into two parts, corresponding to the upper and lower bounds respectively. The proofs for different models involve some common techniques, and so we begin by introducing these auxiliary underlying results.

Recall the Laplacian L of the comparison graph. By virtue of being the Laplacian matrix of a graph with non-negative edges, L is symmetric and positive-semidefinite. By the singular value decomposition, we can write $L = U^T \Lambda U$ where $U \in \mathbb{R}^{d \times d}$ is an orthonormal matrix, and Λ is a diagonal matrix of nonnegative eigenvalues. We will let $L^\dagger$ denote the Moore-Penrose pseudoinverse of L. The Moore-Penrose pseudoinverse is given by $L^\dagger = U^T \Lambda^\dagger U$, where $\Lambda^\dagger$ is a diagonal matrix with entries

$$\Lambda^\dagger_{jj} = \begin{cases} (\Lambda_{jj}^{-1}) & \text{if } \Lambda_{jj} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The all ones vector lies in the nullspace of L, and we will assume without loss of genenerality that the last row of U is proportional to the all ones vector, and that $\Lambda_{dd} = \Lambda_{dd}^{-1} = 0$.

A.1 Auxiliary results for upper bounds

All of our upper bounds make use of a general result for bounding the error of an M-estimator, which we introduce here. Recall that Theorem 1 involves the minimax risk defined in the seminorm $\|v\|_L = \sqrt{v^T L v}$. It is also convenient to introduce the seminorm $\|u\|_{L^\dagger} = \sqrt{u^T L^\dagger u}$, where $L^\dagger$ is the Moore-Penrose pseudoinverse of L.

For future reference, we state and prove a lemma showing that these two seminorms satisfy a restricted form of the Cauchy-Schwarz inequality:

Lemma 3. *Any two vectors u and v such that $u \perp \text{nullspace}(L)$ or/and $v \perp \text{nullspace}(L)$ must satisfy*

$$|\langle u, v \rangle| \leq \|u\|_{L^\dagger} \, \|v\|_L. \tag{7}$$

Proof. Since $L = U^T \Lambda U$ and $L^\dagger = U^T \Lambda^\dagger U$, we have

$$\sqrt{v^T L v}\sqrt{u^T L^\dagger u} = \sqrt{v^T U \Lambda U^T v}\, \sqrt{u^T U \Lambda^\dagger U^T u} = \|\widetilde{v}\|_2 \|\widetilde{u}\|_2 \geq |\langle \widetilde{v}, \widetilde{u} \rangle|,$$

where we have defined $\widetilde{v} := \sqrt{\Lambda} U^T v$ and $\widetilde{u} := \sqrt{\Lambda^\dagger} U^T u$. Continuing on,

$$\langle \widetilde{v}, \widetilde{u} \rangle = v^T U \sqrt{\Lambda}\sqrt{\Lambda^\dagger} U^T u \;=\; v^T U U^T u,$$

where we have used the fact that u or/and v are orthogonal to the null space of L. Since U is orthonormal, we conclude that $\langle \widetilde{v}, \widetilde{u} \rangle = \langle v, u \rangle$, which completes the proof. □

We now are equipped to state and prove a general lemma on M-estimators. Given a loss function $\ell : \mathbb{R}^d \to \mathbb{R}$, consider the M-estimator

$$\widehat{w} \in \arg\min_{w \in \mathcal{W}} \ell(w), \qquad \text{where } \mathcal{W} \text{ is a subset of } \bar{\mathcal{W}} := \{w \in \mathbb{R}^d \mid \langle 1, w \rangle = 0\}. \tag{8}$$

We assume that ℓ is differentiable and strongly convex at w^* with respect to the seminorm $\|\cdot\|_L$, meaning that there is some constant $\gamma > 0$ such that

$$\ell(w^* + \Delta) - \ell(w^*) - \langle \nabla \ell(w^*), \Delta \rangle \geq \gamma \|\Delta\|_L^2 \tag{9}$$

for all perturbations $\Delta \in \mathbb{R}^d$ such that $(w^* + \Delta) \in \mathcal{W}$.

Lemma 4 (Upper bound for M-estimators)**.** *For any differentiable loss function satisfying the γ-strong convexity condition* (9) *and any vector $w^* \in \mathcal{W}$, we have*

$$\|\widehat{w} - w^*\|_L \leq \frac{1}{\gamma}\|\nabla \ell(w^*)\|_{L^\dagger}, \tag{10}$$

where $\|u\|_{L^\dagger} = \sqrt{u^T L^\dagger u}$ is the seminorm defined by the Moore-Penrose pseudoinverse of L.

Proof. Since $\widehat{w}$ and w^* are optimal and feasible, respectively, for the original optimization problem, we have $\ell(\widehat{w}) \leq \ell(w^*)$. Defining the error vector $\Delta = \widehat{w} - w^*$, adding and subtracting the quantity $\langle \nabla \ell(w^*), \Delta \rangle$ yields the bound

$$\ell(w^* + \Delta) - \ell(w^*) - \langle \nabla \ell(w^*), \Delta \rangle \leq -\langle \nabla \ell(w^*), \Delta \rangle.$$

By the γ-convexity condition, the left-hand side is lower bounded by $\gamma \|\Delta\|_L^2$. As for the right-hand side, note that Δ satisfies the constraint $\langle 1, \Delta \rangle = 0$, and thus is orthogonal to the nullspace of the Laplacian matrix L. Therefore, by Lemma 3, we have $|\langle \nabla \ell(w^*), \Delta \rangle| \leq \|\nabla \ell(w^*)\|_{L^\dagger} \, \|\Delta\|_L$. Combining the pieces yields the claimed inequality (10). □

A.2 Auxiliary results for lower bounds

Our lower bounds make use of a technical lemma, standard in minimax analysis. Suppose that our goal is to bound the minimax risk of estimating a parameter w over an indexed class of distributions $\mathcal{P} = \{\mathbb{P}_w \mid \theta \in \Omega\}$ in the square of a seminorm ρ. Consider a collection of vectors $\{w^1, \ldots, w^M\}$ contained within Ω such that, for all distinct pairs of indices $j, k \in [M]$,

$$\rho\big(w^j, w^k\big) \geq \delta \quad \text{and} \quad D(\mathbb{P}_{w^j} \parallel \mathbb{P}_{w^k}) \leq \beta. \tag{11}$$

We refer to any such subset as an (δ, β)-packing set.

Lemma 5 (Pairwise Fano minimax lower bound)**.** *Suppose that we can construct a (δ, β)-packing with cardinality M. Then the minimax error is lower bounded as*

$$\mathfrak{M}_n\big(\theta(\mathcal{P}); \rho^2\big) \geq \frac{\delta^2}{2}\big(1 - \frac{\beta + \log 2}{\log M}\big). \tag{12}$$

Note that the relevant seminorm for Theorem 1 is given by $\rho(w^1, w^2) = \|w^1 - w^2\|_L$. The following lemma will be employed to construct packings for the subsequent proofs.

Define the integer

$$M(\alpha) := \left\lfloor \exp\Big\{ \frac{d}{2}\big(\log 2 + 2\alpha \log 2\alpha + (1 - 2\alpha)\log(1 - 2\alpha)\big)\Big\}\right\rfloor. \tag{13}$$

Lemma 6. *For any pair $\delta > 0$ and $\alpha \in (0, \frac{1}{4})$, there exists a set of $M(\alpha)$ vectors of length d such that*

$$\alpha\delta^2 \leq \|w^j - w^k\|_L^2 \leq \delta^2 \qquad \textit{for all } j \neq k \in [M(\alpha)].,$$

and

$$\langle 1, w^j \rangle = 0 \qquad \textit{for all } j \in [M(\alpha)].$$

Proof: The Gilbert-Varshamov bound guarantees the existence of a binary code $\{z^1, \ldots, z^N\}$ in dimension $(d-1)$, minimum Hamming distance $\lceil \alpha d \rceil$, and the number of code words N at least

$$N \geq \frac{2^{d-1}}{\sum_{\ell=0}^{\lceil \alpha d \rceil - 1} \binom{d-1}{\ell}}.$$

Since $d \geq 2$ and $\alpha \in (0, \frac{1}{4})$, we have

$$\frac{\lceil \alpha d \rceil - 1}{d-1} \leq 2\alpha \leq \frac{1}{2}.$$

Applying standard bounds on the tail of the binomial distribution gives

$$\begin{aligned}\frac{1}{2^{d-1}} \sum_{\ell=0}^{\lceil \alpha d \rceil - 1} \binom{d-1}{\ell} &\leq \exp\Big(- (d-1) D_{\mathrm{KL}}(\frac{\lceil \alpha d \rceil - 1}{d-1} \| \frac{1}{2})\Big) \\ &\leq \exp\Big(- (d-1) D_{\mathrm{KL}}(2\alpha \| \frac{1}{2})\Big),\end{aligned}$$

and hence $N \geq M(\alpha)$.

Defining the d-length vectors $\widetilde{w}^j = \begin{bmatrix} z^j \\ 0 \end{bmatrix}$, this construction ensures that

$$\alpha d \ \leq \ \|\widetilde{w}^j - \widetilde{w}^k\|_2^2 \ \leq \ d \qquad \text{for all distinct } j, k \in [M(\alpha)].$$

Our desired packing $\{w^1, \ldots, w^{M(\alpha)}\}$ is then given by the vectors $w^j := \frac{\delta}{\sqrt{d}} U^T \sqrt{\Lambda^\dagger} \widetilde{w}^j$ for each $j \in [M(\alpha)]$. Given this definition, we have

$$\langle 1, \, w^j \rangle = \frac{\delta}{\sqrt{d}} 1^T U^T \sqrt{\Lambda^\dagger} \widetilde{w} \ = 0,$$

since the all-ones vector lies in the nullspace of the matrix $L^\dagger = U^T \Lambda^\dagger U$. On the other hand, for any pair of distinct vectors in this set, we have

$$\begin{aligned}(w^j - w^k)^T L (w^j - w^k) &= \frac{\delta^2}{d} (\widetilde{w}^j - \widetilde{w}^k)^T \sqrt{\Lambda^\dagger} U L U^T \sqrt{\Lambda^\dagger} (\widetilde{w}^j - \widetilde{w}^k) \\ &= \frac{\delta^2}{d} (\widetilde{w}^j - \widetilde{w}^k)^T \sqrt{\Lambda^\dagger} \Lambda \sqrt{\Lambda^\dagger} (\widetilde{w}^j - \widetilde{w}^k) \\ &= \frac{\delta^2}{d} \|\widetilde{w}^j - \widetilde{w}^k\|_2^2,\end{aligned}$$

where the last step makes use of the fact that the last coordinate of each vector $\widetilde{w}^j$ and $\widetilde{w}^k$ is zero. It follows that $\alpha \delta^2 \leq \|w^j - w^k\|_L^2 \leq \delta^2$, which completes the proof.

A.3 Proof of part (a): Paired linear model

We now turn to the proof of Theorem 1(a) on the minimax rate for the paired linear model (Paired Linear).

A.3.1 Upper bound

The maximum likelihood estimate in the paired linear model is a special case of the general M-estimator (8) with $\ell(w) := \frac{1}{2n} \sum_{i=1}^n \big(y_i - \langle x_i, \, w \rangle\big)^2$. For this quadratic objective function, it is easy to verify that the γ-convexity condition holds with $\gamma = 1$. In particular, note that the Hessian of ℓ is given by $L = X^T X / n$.

It remains to upper bound $\|\nabla \ell(w^*)\|_{L^\dagger}$. The paired-linear observation model (Paired Linear) can be written in a vectorized form as $y = X w^* + \varepsilon$, and hence $\nabla \ell(w^*) = X^T \varepsilon / n$. Consequently, we have

$$\|\nabla \ell(w^*)\|_{L^\dagger}^2 = \frac{1}{n^2} \varepsilon^T X L^\dagger X^T \varepsilon.$$

Observe that ε has independent zero-mean components, and each component $i \in [n]$ has its second moment bounded as $\mathbb{E}[\varepsilon_i^2] = \sigma^2$. Since $L = \frac{1}{n} X^T X$, we have

$$\mathbb{E}[\frac{1}{n} \varepsilon^T X L^\dagger X^T \varepsilon] = \sigma^2 \mathrm{tr}(X L^\dagger X^T) = \sigma^2 (d-1).$$

Applying Lemma 4 gives the desired result

$$\mathbb{E}[\|\Delta\|_L^2] \leq \sigma^2 \frac{d-1}{n}.$$

A.3.2 Lower bound

Based on the pairwise Fano lower bound stated earlier in Lemma 5, we need to construct a suitable (δ, β)-packing, where the seminorm $\rho(w^j, w^k) = \|w^j - w^k\|_L$ is defined by the Laplacian. Given the additive Gaussian noise observation model, we have

$$D(\mathbb{P}_{w^j} \parallel \mathbb{P}_{w^k}) = \frac{n}{2\sigma^2}\|w^j - w^k\|_L^2, \tag{14}$$

With the packing from Lemma 6, Lemma 5 guarantees that

$$\mathfrak{M}_n(\theta(\mathcal{P}); \|\cdot\|_L^2) \geq \frac{\alpha\delta^2}{2}\Big\{1 - \frac{\frac{n\delta^2}{2\sigma^2} + \log 2}{\log M(\alpha)}\Big\}.$$

Choosing $\delta^2 = 0.01\sigma^2\frac{d}{n}$ and setting $\alpha = 0.01$ proves the claim for $d > 9$.

For the case of $d \leq 9$, consider the packing set comprising the three d-length vectors $w^1 = [\frac{\delta}{\sqrt{2}} \quad -\frac{\delta}{\sqrt{2}} \quad 0 \cdots 0]^T$, $w^2 = -w^1$ and $w^3 = [0 \quad \cdots \quad 0]^T$, for some $\delta > 0$. From the calculations made for the general case above, we have $\min_{j,k}\|w^j - w^k\|_L^2 \geq \delta^2$ and $\max_{j,k} D_{\mathrm{KL}}(\mathbb{P}_{w^j}\|\mathbb{P}_{w^k}) \leq \frac{2n\delta^2}{\sigma^2}$. Choosing $\delta^2 = \frac{\sigma^2 \log 2}{4n}$ and applying Lemma 5 proves the claim.

A.4 Proof of part (b): Thurstone

We now turn to the proof of Theorem 1(b) of the minimax rate for the Thurstone model (Thurstone).

A.4.1 Upper bound

Let Φ and ϕ denote respectively the CDF and PDF of the standard Gaussian $N(0,1)$ distribution. For the Thurstone model, the rescaled negative log likelihood takes the form

$$\ell(w) = -\frac{1}{n}\sum_{i=1}^{n}\Big\{\mathbb{I}[y_i = 1]\log\Phi\big(\frac{\langle x_i,\, w\rangle}{\sigma}\big) + \mathbb{I}[y_i = -1]\log\Big(1 - \Phi\big(\frac{\langle x_i,\, w\rangle}{\sigma}\big)\Big)\Big\}.$$

and the MLE is obtained by constrained minimization over the set

$$\mathcal{W}_B := \big\{w \in \mathbb{R}^d \mid \langle 1,\, w\rangle = 0, \quad \text{and} \quad \|w\|_\infty \leq B\big\}. \tag{15}$$

Our first auxiliary result shows that the loss function ℓ is lower bounded by a quadratic form determined by the design matrix $X \in \mathbb{R}^{n\times d}$ whose i^{th} row is given by x_i^T.

Lemma 7. *For all pairs $v, w \in \mathcal{W}_B$, we have*

$$v^T\nabla^2\ell(w)v \geq \frac{c_1}{\sigma^2}\|Xv\|_2^2 \qquad \textit{where } c_1 = \tfrac{4}{\pi} - 1.$$

Proof. The Hessian can be written as

$$\nabla^2\ell(w) = \frac{1}{n\sigma^2}\sum_{i=1}^{n}\big[\mathbb{I}[y_i = 1]T_{i1} + \mathbb{I}[y_i = -1]T_{i2}\big]x_i x_i^T,$$

where

$$T_{i1} := \frac{\phi(w^Tx_i/\sigma)^2 - \Phi(w^Tx_i/\sigma)\phi'(w^Tx_i/\sigma)}{\Phi(w^Tx_i/\sigma)^2}, \quad \text{and}$$
$$T_{i2} := \frac{\phi(w^Tx_i/\sigma)^2 + (1 - \Phi(w^Tx_i/\sigma))\phi'(w^Tx_i/\sigma)}{(1 - \Phi(w^Tx_i/\sigma))^2}.$$

The scalars $\{T_{i1}, T_{i2}\}_{i=1}^n$ are always non-negative (since Φ is log-concave), and hence maximum likelihood is a convex optimization problem. In fact, we will show now that the scalars $\{T_{i1}, T_{i2}\}_{i=1}^n$ are all lower bounded by

$c_1 := \frac{4}{\pi} - 1$. Supposing this lower bound is true, the quantity of interest $v^T \nabla^2 \ell(w) v$ is bounded as

$$\begin{aligned} v^T \nabla^2 \ell(w) v &= v^T \frac{1}{n\sigma^2} \sum_{i=1}^{n} \big[\mathbb{I}[y_i = 1] T_{i1} + \mathbb{I}[y_i = -1] T_{i2}\big] x_i x_i^T v \\ &= \frac{1}{n\sigma^2} \sum_{i=1}^{n} \big[\mathbb{I}[y_i = 1] T_{i1} + \mathbb{I}[y_i = -1] T_{i2}\big] \langle v^T, x_i \rangle^2 \\ &\geq \frac{1}{n\sigma^2} \sum_{i=1}^{n} c_1 \langle v^T, x_i \rangle^2 \\ &= n \frac{c_1}{\sigma^2} \|v\|_L^2. \end{aligned}$$

We will now complete the proof of this lemma by proving the claimed lower bounds on $\{T_{i1}, T_{i2}\}_{i=1}^n$. Let us begin with T_{i2} for some $i \in [n]$. Since $\|w\|_\infty \leq B$ and since x_i is a difference vector, we have

$$\begin{aligned} T_{i2} &\geq \inf_{t \in [-2B/\sigma, 2B/\sigma]} \frac{\phi(t)^2 + (1 - \Phi(t))\phi'(t)}{(1 - \Phi(t))^2} \\ &= \inf_{t \in [-2B/\sigma, 2B/\sigma]} \big(\frac{\phi(t)}{1 - \Phi(t)}\big)^2 - t \frac{\phi(t)}{1 - \Phi(t)}. \end{aligned}$$

Applying standard bounds on the Gaussian distribution and making some algebraic manipulations gives

$$\begin{aligned} T_{i2} &\geq \big(\frac{t + \sqrt{t^2 + \frac{8}{\pi}}}{2}\big)^2 - t\big(\frac{t + \sqrt{t^2 + 4}}{2}\big) \\ &= \frac{2}{\pi} - \frac{t}{2}\big(\sqrt{t^2 + 4} - \sqrt{t^2 + \frac{8}{\pi}}\big) \\ &= \frac{2}{\pi} - \frac{t}{2} \frac{(t^2 + 4) - (t^2 + \frac{8}{\pi})}{\sqrt{t^2 + 4} + \sqrt{t^2 + \frac{8}{\pi}}} \\ &= \frac{2}{\pi} - \big(2 - \frac{4}{\pi}\big) \frac{t}{\sqrt{t^2 + 4} + \sqrt{t^2 + \frac{8}{\pi}}} \\ &\geq \frac{4}{\pi} - 1. \end{aligned}$$

For any $i \in [n]$, making use of the fact that $\Phi(-t) = 1 - \Phi(t)$ and $\phi(-t) = \phi(t)$, we have

$$\begin{aligned} T_{i1} &\geq \inf_{t \in [-2B/\sigma, 2B/\sigma]} \frac{\phi(t)^2 - \Phi(t)\phi'(t)}{\Phi(t)^2} \\ &= \inf_{t \in [-2B/\sigma, 2B/\sigma]} \frac{\phi(t)^2 + (1 - \Phi(t))\phi'(t)}{(1 - \Phi(t))^2} \\ &\geq \frac{4}{\pi} - 1. \end{aligned}$$

Here the final inequality results from the arguments made above for the case of T_{i2} . □

Defining the difference vector $\Delta := \widehat{w} - w^*$, Lemma 7 guarantees that

$$\ell(w^* + \Delta) - \ell(w^*) - \langle \nabla \ell(w^*), \Delta \rangle \geq \frac{c_1}{\sigma^2} \|\Delta\|_L^2.$$

Applying Lemma 4 gives

$$\|\Delta\|_L \leq \frac{\sigma^2}{c_1} \|\nabla \ell(w^*)\|_{L^\dagger}. \tag{16}$$

It remains to upper bound the quantity $\nabla\ell(w^*)^T L^\dagger \nabla\ell(w^*)$. Observe that the gradient takes the form

$$\nabla\ell(w^*) = \frac{-1}{n\sigma}\sum_{i=1}^{n}\big[\mathbb{I}[y_i=1]\frac{\phi(\langle w^*,\, x_i\rangle/\sigma)}{\Phi(\langle w^*,\, x_i\rangle/\sigma)} - \mathbb{I}[y_i=-1]\frac{\phi(\langle w^*,\, x_i\rangle/\sigma)}{1-\Phi(\langle w^*,\, x_i\rangle/\sigma)}\big]x_i.$$

Define a random vector $\theta\in\mathbb{R}^n$ with independent components as

$$\theta_i = \begin{cases} \frac{\phi(\langle w^*, x_i\rangle/\sigma)}{\Phi(\langle w^*, x_i\rangle/\sigma)} & \text{w.p.} \quad \Phi(\langle w^*,\, x_i\rangle/\sigma) \\ \frac{-\phi(\langle w^*, x_i\rangle/\sigma)}{1-\Phi(\langle w^*, x_i\rangle/\sigma)} & \text{w.p.} \quad 1-\Phi(\langle w^*,\, x_i\rangle/\sigma). \end{cases} \tag{17}$$

With this notation, we have $\nabla\ell(w^*) = \frac{-1}{n\sigma}X^T\theta$, and hence

$$\nabla\ell(w^*)^T L^\dagger\nabla\ell(w^*) = \frac{1}{n^2\sigma^2}\theta^T X L^\dagger X^T\theta.$$

Observe that the absolute value of every component of the random vector θ is upper bounded by

$$\sup_{w\in\mathcal{W}_B}\frac{\phi(w^Tx_i/\sigma)}{\Phi(w^Tx_i/\sigma)(1-\Phi(w^Tx_i/\sigma))} \;\le\; \frac{1}{\sqrt{2\pi}\Phi(2B/\sigma)(1-\Phi(2B/\sigma))} = \frac{1}{\sqrt{2\pi}\kappa}.$$

Furthermore, since each coordinate of θ is independent and of mean zero, for any positive-semidefinite matrix M it must be that

$$\mathbb{E}[\theta^T M\theta] \le \frac{1}{2\pi\kappa^2}\mathrm{tr}(M).$$

Recall that $L=\frac{1}{n}X^TX$ and $\mathrm{tr}(\frac{1}{n}XL^\dagger X^T) = d-1$. Consequently,

$$\mathbb{E}[\frac{1}{n^2\sigma^2}\theta^T XL^\dagger X^T\theta] \le \frac{1}{2\pi\kappa^2}\frac{d-1}{n\sigma^2}.$$

Substituting this inequality in (16) gives the desired result:

$$\mathbb{E}[\|\Delta\|_L^2] \le \frac{\sigma^2}{2\pi c_1^2\kappa^2}\frac{d-1}{n}.$$

A.4.2 Lower bound

As before, we let Φ and ϕ denote respectively the CDF and PDF of the standard Gaussian distribution. For any pair of weight vectors w^j and w^k, the KL divergence between the distributions $\mathbb{P}_{w^j}$ and $\mathbb{P}_{w^k}$ is given by

$$D_{\mathrm{KL}}(\mathbb{P}_{w^j}\|\mathbb{P}_{w^k}) = \sum_{i=1}^{n}\Phi(\langle w^j,\, x_i\rangle/\sigma)\log\frac{\Phi(\langle w^j,\, x_i\rangle/\sigma)}{\Phi(\langle w^k,\, x_i\rangle/\sigma)} + (1-\Phi(\langle w^j,\, x_i\rangle/\sigma))\log\frac{1-\Phi(\langle w^j,\, x_i\rangle/\sigma)}{1-\Phi(\langle w^k,\, x_i\rangle/\sigma)}.$$

Observe that for any $c>0$, it must be that $\log c\le c-1$. It follows that for any $a,b\in(0,1)$, $\log\frac{a}{b}\le\frac{a}{b}-1$ and hence $a\log\frac{a}{b}\le(a-b)\frac{a}{b}$. Applying this argument gives

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{P}_{w^j}\|\mathbb{P}_{w^k}) &\le \sum_{i=1}^{n}(\Phi(\langle w^j,\, x_i\rangle/\sigma)-\Phi(\langle w^k,\, x_i\rangle/\sigma))\frac{\Phi(\langle w^j,\, x_i\rangle/\sigma)}{\Phi(\langle w^k,\, x_i\rangle/\sigma)} \\ &\qquad - \Big\{\Phi(\langle w^j,\, x_i\rangle/\sigma)) - \Phi(\langle w^k,\, x_i\rangle/\sigma)\Big\}\frac{1-\Phi(\langle w^j,\, x_i\rangle/\sigma)}{1-\Phi(\langle w^k,\, x_i\rangle/\sigma)} \\ &\le \sum_{i=1}^{n}\frac{(\Phi(\langle w^j,\, x_i\rangle/\sigma)-\Phi(\langle w^k,\, x_i\rangle/\sigma))^2}{\Phi(\langle w^k,\, x_i\rangle/\sigma)(1-\Phi(\langle w^k,\, x_i\rangle/\sigma))}. \end{aligned}$$

Since $\|w\|_\infty \le B$, we have

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{P}_{w^j}\|\mathbb{P}_{w^k}) &\le \sum_{i=1}^n \frac{(\Phi(\langle w^j,\, x_i\rangle/\sigma) - \Phi(\langle w^k,\, x_i\rangle/\sigma))^2}{\Phi(2B/\sigma)(1-\Phi(2B/\sigma))} \\ &\le \sum_{i=1}^n \frac{\phi(0)^2}{\Phi(2B/\sigma)(1-\Phi(2B/\sigma))}(\langle w^j,\, x_i\rangle/\sigma - \langle w^k,\, x_i\rangle/\sigma)^2 \\ &= \frac{n}{2\pi\sigma^2\Phi(2B/\sigma)(1-\Phi(2B/\sigma))}(w^j - w^k)^T L (w^j - w^k). \end{aligned}$$

Lemma 6 guarantees the existence of a packing set $\{w^1, \ldots, w^{M(\alpha)}\}$ such that $\langle 1,\, w^j\rangle = 0$ for all $j \in [M(\alpha)]$, and moreover such that

$$\alpha\delta^2 \le \|w^j - w^k\|_L^2 \le \delta^2 \quad \text{for all distinct pairs } j,k \in [M(\alpha)].$$

In order to apply this packing, we need to verify that each vector w^j also satisfies the boundedness constraint $\|w^j\|_\infty \le B$. We claim that this boundedness condition holds when

$$\delta^2 = 0.01\frac{\sigma^2 d}{n} \times 2\pi\Phi(2B/\sigma)(1-\Phi(2B/\sigma)) \tag{18}$$

From the proof of Lemma 6, we have $w^j = \frac{\delta}{\sqrt{d}} U^T\sqrt{\Lambda^\dagger}\widetilde{w}^j$, where $\widetilde{w}^j$ has all its entries in $\{-1,0,1\}$. Consequently,

$$\|w^j\|_\infty \;\le\; \frac{\delta}{\sqrt{d}}\|\sqrt{\Lambda^\dagger}\widetilde{w}^j\|_2 \overset{(i)}{\le} \frac{\delta}{\sqrt{d}}\sqrt{\mathrm{tr}(\Lambda^\dagger)} \overset{(ii)}{=} \frac{\delta}{\sqrt{d}}\sqrt{\mathrm{tr}(L^\dagger)} \overset{(iii)}{\le} B$$

where inequality (i) follows from the fact that $\widetilde{w}^j$ has entries in $\{-1,0,1\}$; equality (ii) follows since $L^\dagger = U^T\Lambda^\dagger U$ by definition; and inequality (iii) follows from our choice (18) of δ and our assumption $n \ge \frac{c\sigma^2\kappa\mathrm{tr}(L^\dagger)}{B^2}$ on the sample size with $c = .01$. Finally, observe that

$$\max_{j,k} D_{\mathrm{KL}}(\mathbb{P}_{w^j}\|\mathbb{P}_{w^k}) \le \frac{n\delta^2}{2\pi\sigma^2\Phi(2B/\sigma)(1-\Phi(2B/\sigma))}, \quad \text{and} \quad \min_{j,k}\|w^j - w^k\|_L^2 \ge \alpha\delta^2.$$

We have thus constructed a packing suitable for application of Lemma 5, and doing so yields the lower bound

$$\|\widehat{w} - w^*\|_L^2 \ge \frac{\alpha}{2}\delta^2\Big\{1 - \frac{\frac{\delta^2 n}{2\pi\sigma^2\Phi(2B/\sigma)(1-\Phi(2B/\sigma))} + \log 2}{\log M(\alpha)}\Big\}.$$

Substituting our choice (18) of δ and setting $\alpha = 0.01$ proves the claim for $d > 9$.

For the case of $d \le 9$, consider the packing set comprising the three d-length vectors $w^1 = [\frac{\delta}{\sqrt{2}} \quad -\frac{\delta}{\sqrt{2}} \quad 0 \cdots 0]^T$, $w^2 = -w^1$ and $w^3 = [0 \quad \cdots \quad 0]^T$, for some $\delta > 0$. From the calculations made for the general case above, we have $\min_{j,k}\|w^j - w^k\|_L^2 \ge \delta^2$ and $\max_{j,k} D_{\mathrm{KL}}(\mathbb{P}_{w^j}\|\mathbb{P}_{w^k}) \le \frac{4n\delta^2}{2\pi\kappa\sigma^2}$. Choosing $\delta^2 = \frac{\kappa\sigma^2}{2n}$ and applying Lemma 5 proves the claim.

A.5 Proof of part (c): BTL model

We now turn to the proof of Theorem 1(c) on the minimax rate for the BTL model (BTL).

A.5.1 Upper bound

In this case, the maximum likelihood estimate is given by $\widehat{w} \in \arg\min_{w\in\mathcal{W}_B} \ell(w)$, where

$$\ell(w) = \frac{1}{n}\sum_{i=1}^n \log\big(1 + \exp\big(\frac{-y_i\langle w,\, x_i\rangle}{\sigma}\big)\big).$$

This loss function has gradient and Hessian, respectively, given by

$$\nabla \ell(w) = \frac{1}{n\sigma} \sum_{i=1}^{n} \frac{-y_i e^{\frac{-y_i \langle w,\, x_i \rangle}{\sigma}}}{1 + e^{\frac{-y_i \langle w,\, x_i \rangle}{\sigma}}} x_i, \quad \text{and} \quad \nabla^2 \ell(w) = \frac{1}{n^2\sigma^2} \sum_{i=1}^{n} \frac{e^{\frac{-y_i \langle w,\, x_i \rangle}{\sigma}}}{\big(1 + e^{\frac{-y_i \langle w,\, x_i \rangle}{\sigma}}\big)^2} x_i x_i^T.$$

By inspection, the Hessian is positive semi-definite, showing that ℓ is convex. Moreover, a simple calculation shows that any observation $y_i \in \{-1, 1\}$ and any differencing vector x_i, we have $\inf_{w \in \mathcal{W}_B} \frac{e^{\frac{-y_i \langle w,\, x_i \rangle}{\sigma}}}{\big(1+e^{\frac{-y_i \langle w,\, x_i \rangle}{\sigma}}\big)^2} \geq \frac{1}{\big(e^{\frac{B}{\sigma}} + e^{-\frac{B}{\sigma}}\big)^2}$. Thus, defining the difference vector $\Delta := \widehat{w} - w^*$, we find that

$$\ell(w^* + \Delta) - \ell(w^*) - \langle \nabla \ell(w^*),\, \Delta \rangle \geq \Delta^T \nabla^2 \ell(w^*) \Delta \;\geq\; \frac{1}{\sigma^2 \big(e^{\frac{B}{\sigma}} + e^{\frac{-B}{\sigma}}\big)^2}\, \Delta^T L \Delta.$$

Applying Lemma 4 gives

$$\|\Delta\|_L \leq \sigma^2 \big(e^{\frac{B}{\sigma}} + e^{\frac{-B}{\sigma}}\big)^2 \|\nabla \ell(w^*)\|_{L^\dagger}. \tag{19}$$

Now define a random vector $\theta \in \mathbb{R}^n$ with independent components

$$\theta_i = \begin{cases} \frac{-e^{\frac{\langle w^*,\, x_i \rangle}{\sigma}}}{1+e^{\frac{\langle w^*,\, x_i \rangle}{\sigma}}} & \text{with probability } \frac{1}{1+e^{\frac{\langle w^*,\, x_i \rangle}{\sigma}}} \\ \frac{e^{-\frac{\langle w^*,\, x_i \rangle}{\sigma}}}{1+e^{-\frac{\langle w^*,\, x_i \rangle}{\sigma}}} & \text{with probability } \frac{1}{1+e^{-\frac{\langle w^*,\, x_i \rangle}{\sigma}}} \end{cases}$$

With this notation, we have $\nabla \ell(w^*) = -\frac{1}{n\sigma} X^T \theta$, and hence

$$\nabla \ell(w^*)^T L^\dagger \nabla \ell(w^*) = \frac{1}{n^2\sigma^2} \theta^T X L^\dagger X^T \theta.$$

Observe that the absolute value of every component of the random vector θ is upper bounded by 1. Furthermore, since each coordinate of θ is independent and of mean zero, for any positive-semidefinite matrix M it must be that

$$\mathbb{E}[\theta^T M \theta] \leq \operatorname{tr}(M).$$

Recall that $L = \frac{1}{n} X^T X$ and $\operatorname{tr}(\frac{1}{n} X L^\dagger X^T) = d - 1$. Consequently,

$$\mathbb{E}[\frac{1}{n^2\sigma^2} \theta^T X L^\dagger X^T \theta] \leq \frac{d-1}{n\sigma^2}.$$

Substituting this inequality in (19) gives

$$\mathbb{E}[\|\Delta\|_L^2] \leq \sigma^2 \big(e^{\frac{B}{\sigma}} + e^{\frac{-B}{\sigma}}\big)^4 \frac{d-1}{n}.$$

Setting $e^{\frac{B}{\sigma}} + e^{\frac{-B}{\sigma}} \leq 2e^{\frac{B}{\sigma}}$ proves the claim.

A.5.2 Lower bound

Consider the function

$$\Psi(w, x) = \log\big(\exp\big(\frac{w_{a(x)}}{\sigma}\big) + \exp\big(\frac{w_{b(x)}}{\sigma}\big)\big) - \frac{w_{a(x)} + w_{b(x)}}{2\sigma},$$

where $a(x)$ and $b(x)$ denote the indices of the 1 and -1, respectively, in the differencing vector x.

Given a single observation pair (y, x) from the BTL model, the associated likelihood can be written as

$$\mathbb{P}[y; w, x] = \exp\big(\frac{y}{2\sigma} \langle w,\, x \rangle - \Psi(w, x)\big).$$

The Kullback-Leibler divergence between a pair $\mathbb{P}_{w^j}$ and $\mathbb{P}_{w^k}$ is given by

$$D(\mathbb{P}_{w^j} \parallel \mathbb{P}_{w^k}) = \frac{1}{2\sigma}\frac{1-e^{\frac{(w^j)^T x}{\sigma}}}{1+e^{\frac{(w^j)^T x}{\sigma}}}\langle w^j - w^k,\, x\rangle - \langle w^j - w^k,\, \nabla\Psi(w^j, x)\rangle + \frac{1}{2}(w^j - w^k)^T \nabla^2\Psi(\widetilde{w}, x)(w^j - w^k).$$

for some $\widetilde{w}$ on the line joining w^j and w^k. A straightforward computation yields

$$\nabla\Psi(w^j, x) = \frac{1}{2\sigma}\frac{1-e^{\frac{(w^j)^T x}{\sigma}}}{1+e^{\frac{(w^j)^T x}{\sigma}}}x, \quad \text{and} \quad \nabla^2\Psi(\widetilde{w}, x) = \frac{1}{2\sigma^2}\frac{1}{e^{\frac{(\widetilde{w})^T x}{\sigma}} + e^{\frac{-(\widetilde{w})^T x}{\sigma}} + 2}xx^T \;\leq\; \frac{1}{8\sigma^2}xx^T,$$

from which it follows that

$$D(\mathbb{P}_{w^j} \parallel \mathbb{P}_{w^k}) \leq \frac{1}{8\sigma^2}(w^j - w^k)^T xx^T (w^j - w^k).$$

Aggregating over all samples, and observing that the distribution of the observation is independent across samples, we get

$$D(\mathbb{P}_{w^j}(y) \parallel \mathbb{P}_{w^k}(y)) \leq \frac{n}{8\sigma^2}(w^j - w^k)^T L(w^j - w^k).$$

Lemma 6 guarantees the existence of a packing set $\{w^1, \ldots, w^{M(\alpha)}\}$ such that $\langle 1,\, w^j\rangle = 0$ for all $j \in [M(\alpha)]$, and moreover such that

$$\alpha\delta^2 \leq \|w^j - w^k\|_L^2 \leq \delta^2 \quad \text{for all distinct pairs } j, k \in [M(\alpha)].$$

In order to apply this packing, we need to verify that each vector w^j also satisfies the boundedness constraint $\|w^j\|_\infty \leq B$. We claim that this boundedness condition holds when

$$\delta^2 = 0.08\frac{\sigma^2 d}{n} \tag{20}$$

From the proof of Lemma 6, we have $w^j = \frac{\delta}{\sqrt{d}}U^T\sqrt{\Lambda^\dagger}\widetilde{w}^j$, where $\widetilde{w}^j$ has all its entries in $\{-1, 0, 1\}$. Consequently,

$$\|w\|_\infty \;\leq\; \frac{\delta}{\sqrt{d}}\|\sqrt{\Lambda^\dagger}\widetilde{w}^j\|_2 \overset{(i)}{\leq} \frac{\delta}{\sqrt{d}}\sqrt{\mathrm{tr}(\Lambda^\dagger)} \overset{(ii)}{=} \frac{\delta}{\sqrt{d}}\sqrt{\mathrm{tr}(L^\dagger)} \overset{(iii)}{\leq} B$$

where inequality (i) follows from the fact that $\widetilde{w}^j$ has entries in $\{-1, 0, 1\}$; equality (ii) follows since $L^\dagger = U^T\Lambda^\dagger U$ by definition; and inequality (iii) follows from our choice (20) of δ and our assumption $n \geq \frac{c\sigma^2\kappa\mathrm{tr}(L^\dagger)}{B^2}$ on the sample size with $c = 0.01$.

Finally, observe that

$$\max_{j,k} D_{\mathrm{KL}}(\mathbb{P}_{w^j}\|\mathbb{P}_{w^k}) \leq \frac{n\delta^2}{8\sigma^2}, \quad \text{and} \quad \min_{j,k}\|w^j - w^k\|_L^2 \geq \alpha\delta^2.$$

We have thus constructed a packing suitable for application of Lemma 5, and doing so yields the lower bound

$$\|\widehat{w} - w^*\|_L^2 \geq \frac{\alpha}{2}\delta^2\Big\{1 - \frac{\frac{n\delta^2}{8\sigma^2} + \log 2}{\log M(\alpha)}\Big\}.$$

Substituting our choice (20) of δ and setting $\alpha = 0.01$ proves the claim for $d > 9$.

For the case of $d \leq 9$, consider the packing set comprising the three d-length vectors $w^1 = [\frac{\delta}{\sqrt{2}} \quad -\frac{\delta}{\sqrt{2}} \quad 0 \cdots 0]^T$, $w^2 = -w^1$ and $w^3 = [0 \quad \cdots \quad 0]^T$, for some $\delta > 0$. From the calculations made for the general case above, we have $\min_{j,k}\|w^j - w^k\|_L^2 \geq \delta^2$ and $\max_{j,k} D_{\mathrm{KL}}(\mathbb{P}_{w^j}\|\mathbb{P}_{w^k}) \leq \frac{n\delta^2}{2\sigma^2}$. Choosing $\delta^2 = \frac{\sigma^2\log 2}{n}$ and applying Lemma 5 proves the claim.

B Proof of Theorem 2

In the cardinal case, when each coordinate is measured the same number of times, the CARDINAL model reduces to the well-studied normal location model, for which the MLE is known to be the minimax estimator and its risk is straightforward to characterize (see Lehmann and Casella [LC98] for instance).

In the ordinal case, the result follows from Theorem 1b, with $L = \frac{2}{d(d-1)}\big(dI - 11^T\big)$. Since $\langle 1,\, w^* \rangle = \langle 1,\, \widehat{w} \rangle = 0$, we have $\frac{\|w^* - \widehat{w}\|_L^2}{\lambda_{\max}(L)} \leq \|w^* - \widehat{w}\|_2^2 \leq \frac{\|w^* - \widehat{w}\|_L^2}{\lambda_2(L)}$. For our choice of L, we have $\lambda_2(L) = \lambda_{\max}(L) = \frac{2}{d-1}$. Substituting this relation in Theorem 1b gives the desired result.

C Materials and Methods for Experiments

We describe additional details of the experiments discussed in Section 6.

C.1 Simulations using Synthetic Data

Every data point in the plots using synthetic data is an average over 20 trial runs. In each run, the vector $w^* \in \mathbb{R}^d$ is constructed by first drawing an d-length vector from the distribution $N(0, I)$ and shifting it to satisfy $\langle w^*,\, 1 \rangle = 0$. In the simulations of Section 6.1.1 evaluating the effects of graph topology, each of the n samples are obtained in the following manner. Given the graph topology, an edge is selected uniformly at random, and the chosen edge determines the pair of items compared. The outcome of the comparison is generated as per the THURSTONE model. The value of σ is fixed to be 1. In the simulations of Section 6.3 evaluating the effects of model misspecification, each of the n samples are obtained as follows. The pair to be compared is chosen uniformly at random from the set of all $\binom{d}{2}$ pairs (i.e., samples from a complete topology). The outcome of this comparison is generated as per the THURSTONE model with a probability $(1 - \epsilon)$ and as per the BTL model with a probability ϵ. The value of σ is fixed to be 1 here as well. Given the n samples, inference is performed via the maximum likelihood estimator for the THURSTONE model, under the knowledge of the true σ.

C.2 Experiments on Amazon Mechanical Turk (MTurk)

Amazon Mechanical Turk (mturk.com), or MTurk in short, is an online "crowdsourcing" platform where individuals or businesses can put up a task, and any individual can log in and complete the tasks in exchange for a payment that is specified along with the task.

Each set of MTurk experiments described in Section 6 is a subset of the following set of seven experiments. The tasks were selected to have broad coverage of several important subjective judgment paradigms such as preference elicitation, knowledge elicitation, audio and visual perception and skill utilization.

(a) *Rating taglines for a product:* A product was described and taglines for this product were shown (Figure 6a). The worker had to rate each of these taglines in terms of its originality, clarity and relevance to this product.

(b) *Estimating areas of circles:* In each question, the worker was shown a circle in a bounding box (Figure 6b), and the worker was required to identify the fraction of the box's area that the circle occupied.

(c) *Finding spelling mistakes in text:* The worker had to identify the number of words that were misspelled in each paragraph shown (Figure 6c).

(d) *Estimating age of people from photographs:* The worker was shown photographs of people (Figure 6d) and was asked to estimate their ages.

(e) *Estimating distances between pairs of cities:* Pairs of cities were listed (Figure 6e) and for each pair, the worker had to estimate the distance between them.

(f) *Identifying sounds:* The worker was presented with audio clips, each of which was the sound of a single key on a piano (which corresponds to a single frequency). The worker had to estimate the frequency of the sound in each audio clip (Figure 6f).

(g) *Rating relevance of the results of a search query:* Results for the query 'Internet' for an image search were shown (Figure 1) and the worker had to rate the relevance of these results with respect to the given query.

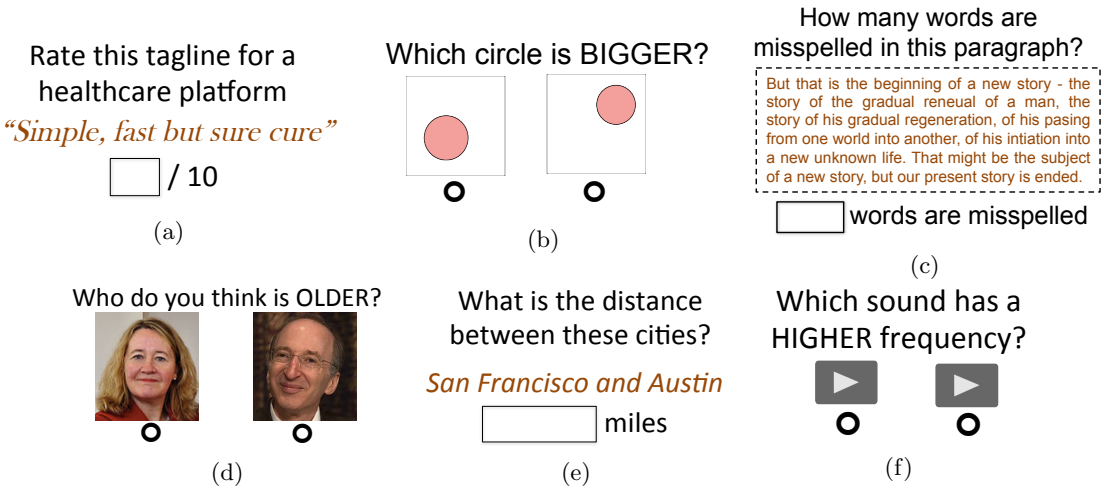

Figure 6. Screenshots of the tasks presented to the subjects. For each task, only one version (cardinal or ordinal) is shown here.

Here are some additional details about the experiments. Each experiment comprised of 100 tasks, all comprising the same set of questions but organized in either a cardinal or ordinal format at random. A worker were offered 20 cents per completed task. A worker was allowed to do no more than one task in an experiment. Workers were required to answer all the questions in a task. Only those workers who had 100 or more approved works prior to this and also had at least 95% approval rate were allowed. Workers from any country were allowed to participate, except for the task of estimating distances between cities (for which only USA-based workers were permitted since all questions involved American cities).

The analysis of Section 6.2.1 was performed in the following manner. Upon obtaining the data, we first reduced the cardinal data obtained from the experiments into ordinal form by comparing answers given by the subjects to consecutive questions. For five of the experiments ((b) through (f)), we had access to the "ground truth" solutions, using which we computed the fraction of answers that were incorrect in the ordinal and the cardinal-converted-to-ordinal data (any tie in the latter case was counted as half an error). For the two remaining experiments ((a) and (g)) for which there is no ground truth, we computed the 'error' as the fraction of (ordinal or cardinal-converted-to-ordinal) answers provided by the subjects that disagreed with each other.

The results of Figure 4a establishes the absence of a 'data processing inequality' between data converted from cardinal elicitations to ordinal and data obtained by directly eliciting ordinal information. This absence of data-processing inequality may be explained by the argument that the inherent evaluation process in the human subjects is not the same in the cardinal and ordinal cases: humans do *not* perform an ordinal evaluation by first performing cardinal evaluations and then comparing them (this is why it is frequently found to be easier to compare than score [Bar03, SBC05]).

The analysis presented in Section 6.2.2 and Section 6.1.2 was performed as follows. For the ordinal data, we evaluated the performance of the maximum likelihood estimators of the THURSTONE model, and for the cardinal data we evaluated the performance of the CARDINAL model. Note that the cardinal data was *not* converted to ordinal form in these two sections. The true and inferred vectors were first scaled to have their maximum elements equal to 1 and minimum elements equal to -1; this mimics the effect of knowing the scaling B from 'domain knowledge'. The (scaled) inferred vectors in either case were then compared with the (scaled) true vector in terms of the error $\frac{\|w^* - \widehat{w}\|_2^2}{d}$.