PNAS
www.pnas.org

Supplementary Information for

Algorithmic Monoculture and Social Welfare

Jon Kleinberg, Manish Raghavan

Manish Raghavan.
E-mail: manish@cs.cornell.edu

This PDF file includes:

Supporting Information Text

1. Random Utility Models satisfying Definition 1

Theorem 4. *Let f be the pdf of $\mathcal{E}$. The family of RUMs $\mathcal{F}_\theta$ given by ranking $x_i + \frac{\varepsilon_i}{\theta}$ with $\varepsilon_i \sim \mathcal{E}$ satisfies the conditions of Definition 1 if:*

- *f is differentiable*
- *f has positive support on $(-\infty, \infty)$*

Proof. We need to show that $\mathcal{F}_\theta$ satisfies the differentiability, asymptotic optimality, and monotonicity conditions in Definition 1.

Differentiability: The probability density of any realization of the n noise samples ε_i/θ is $\prod_{i=1}^n f(\varepsilon_i/\theta)$. Let $\varepsilon = [\varepsilon_1/\theta, \ldots, \varepsilon_n/\theta]$ be the vector of noise values and let $M(\pi) \subseteq \mathbb{R}^n$ be the region such that any $\varepsilon \in M(\pi)$ will produces the ranking π. The probability of any permutation π is

$$\Pr_\theta[\pi] = \int_{M(\pi)} \prod_{i=1}^n f\left(\frac{\varepsilon_i}{\theta}\right)\, d^n\mathbf{z}.$$

Because f is differentiable,

$$\frac{d}{d\theta} f\left(\frac{x}{\theta}\right) = f'\left(\frac{x}{\theta}\right) \cdot \left(-\frac{x}{\theta^2}\right)$$

Because $\Pr\theta(\pi)$ is an integral of the product of differentiable functions over a fixed region, it is differentiable.

Asymptotic optimality: We will show that for any pair of elements and any $\delta > 0$, there exists sufficiently large θ such that the probability that they incorrectly ranked is at most δ. We will conclude with a union bound over the $n-1$ pairs of adjacent candidates that there exists sufficiently large θ such that the probability of outputting the correct ranking must be at least $1-(n-1)\delta$.

Consider two candidates $x_i > x_{i+1}$. Let ν be the difference $x_i - x_{i+1}$. Then, they will be correctly ranked if

$$\frac{\varepsilon_i}{\theta} > -\frac{\nu}{2}$$
$$\frac{\varepsilon_{i+1}}{\theta} < \frac{\nu}{2}$$

Let $\overline{q}$ and $\underline{q}$ be the $1-\frac{\delta}{2}$ and $\frac{\delta}{2}$ quantiles of $\mathcal{E}$ respectively, and let $q = \max(|\overline{q}|, |\underline{q}|)$. For $\theta > \frac{2q}{\nu}$,

$$\begin{aligned}
\Pr\left[\frac{\varepsilon_i}{\theta} < -\frac{\nu}{2}\right] &= \Pr\left[\varepsilon_i < -\frac{\nu\theta}{2}\right] \\
&< \Pr\left[\varepsilon_i < -q\right] \\
&\leq \Pr\left[\varepsilon_i < \underline{q}\right] \\
&= \frac{\delta}{2} \\
\Pr\left[\frac{\varepsilon_{i+1}}{\theta} > \frac{\nu}{2}\right] &= \Pr\left[\varepsilon_{i+1} > \frac{\nu\theta}{2}\right] \\
&< \Pr\left[\varepsilon_{i+1} > q\right] \\
&\leq \Pr\left[\varepsilon_{i+1} > \overline{q}\right] \\
&= \frac{\delta}{2}
\end{aligned}$$

Thus, for sufficiently large θ, the probability that x_i and x_{i+1} are incorrectly ordered is at most δ.

Repeating this analysis for all $n-1$ pairs of adjacent elements, taking the maximum of all the θ's, and taking a union bound yields that the probability of incorrectly ordering any pair of elements is at most $(n-1)\delta$, meaning the probability of outputting the correct ranking is at least $1-(n-1)\delta$. Since δ is arbitrary, this probability can be made arbitrarily close to 1, satisfying the asymptotic optimality condition.

Monotonicity: The removal of any elements does not alter the distribution of the remaining elements, meaning that the distribution of $\pi^{(-S)}$ is equivalent to a RUM with $n - |S|$ elements. Thus, it suffices to show that for a RUM with positive support on $(-\infty, \infty)$, the probability of ranking the best candidate first strictly increases with θ.

Recall that by definition, the candidates are ranked according to $x_i + \frac{\varepsilon_i}{\theta}$. The probability that x_1 is ranked first is

$$\begin{aligned}
\Pr\left[x_1 + \frac{\varepsilon_1}{\theta} > \max_{2\leq i\leq n} x_i + \frac{\varepsilon_i}{\theta}\right] &= \Pr\left[\frac{\varepsilon_1}{\theta} > \max_{2\leq i\leq n} x_i - x_1 + \frac{\varepsilon_i}{\theta}\right] \\
&= \Pr\left[\varepsilon_1 > \max_{2\leq i\leq n} \theta(x_i - x_1) + \varepsilon_i\right] \\
&= \mathbb{E}_{\varepsilon_2,\ldots,\varepsilon_n} \Pr\left[\varepsilon_1 > \max_{2\leq i\leq n} \theta(x_i - x_1) + \varepsilon_i \mid \varepsilon_2, \ldots, \varepsilon_n\right] \qquad [1]
\end{aligned}$$

We want to show that Eq. (1) is increasing in θ. Intuitively, this is because as θ increases, the right hand side of the inequality inside the probability decreases. To prove this formally, it suffices to show that the subderivative of Eq. (1) with respect to θ only includes strictly positive numbers. First, we have

$$\frac{\partial}{\partial\theta}\mathbb{E}_{\varepsilon_2,\ldots,\varepsilon_n}\Pr\left[\varepsilon_1 > \max_{2\le i\le n}\theta(x_i-x_1)+\varepsilon_i \mid \varepsilon_2,\ldots,\varepsilon_n\right] \subset \mathbb{R}_{>0} \iff \frac{\partial}{\partial\theta}\Pr\left[\varepsilon_1 > \max_{2\le i\le n}\theta(x_i-x_1)+\varepsilon_i \mid \varepsilon_2,\ldots,\varepsilon_n\right] \subset \mathbb{R}_{>0}$$

Let F and f be the cumulative density function and probability density function of $\mathcal{E}$ respectively. Then,

$$\Pr\left[\varepsilon_1 > \max_{2\le i\le n}\theta(x_i-x_1)+\varepsilon_i \mid \varepsilon_2,\ldots,\varepsilon_n\right] = 1 - F\left(\max_{2\le i\le n}\theta(x_i-x_1)+\varepsilon_i\right)$$

Note that $F(\cdot)$ is strictly increasing (since f is assumed to have positive support on $(-\infty,\infty)$), so it suffices to show that

$$\frac{\partial}{\partial\theta}\max_{2\le i\le n}\theta(x_i-x_1)+\varepsilon_i \subset \mathbb{R}_{<0}$$

For any i,

$$\frac{d}{d\theta}\theta(x_i-x_1)+\varepsilon_i = x_i - x_1 < 0.$$

Thus, the subderivative of the max of such functions includes only strictly negative numbers, which completes the proof. □

2. 3-candidate RUM Counterexamples

A. Violating Definition 2. Here, we provide a noise mode $\mathcal{E}$, accuracy parameter θ, and candidate distribution $\mathcal{D}$ such that $U_{AH} < U_{AA}$.

Choose the noise distribution $\mathcal{E}$ and accuracy parameter θ such that

$$\frac{\varepsilon}{\theta} = \begin{cases} 1 & w.p.\frac{\delta}{2} \\ 0 & w.p. 1-\delta \\ -1 & w.p.\frac{\delta}{2} \end{cases}$$

Note that this distribution does not satisfy Definition 1 because it is neither differentiable nor supported on $(-\infty,\infty)$; however, we can provide a "smooth" approximation to this distribution by expressing it as the sum of arbitrarily tightly concentrated Gaussians with the same results.

We choose the candidate distribution $\mathcal{D}$ such that $x_1 - 1 > x_2 > x_3 > x_1 - 2$. For example,

$$x_1 = \frac{7}{4}$$
$$x_2 = \frac{1}{2}$$
$$x_3 = 0$$

Under this condition, assuming $x_3 = 0$ without loss of generality,

$$U_{AH}(\theta,\theta) - U_{AA}(\theta,\theta) = \frac{\delta^2}{32}\left(\delta^3 x_1 - 4\delta^2 x_1 + 4\delta x_1 + 2\delta^3 x_2 - 14\delta^2 x_2 + 20\delta x_2 - 8x_2\right)$$

Notice that the lowest-power δ term is $-\frac{\delta^2 x_2}{4}$. Therefore, for sufficiently small δ, this is negative. For example, plugging in the values given above with $\delta = .1$, $U_{AH}(\theta,\theta) - U_{AA}(\theta,\theta) \approx -0.00076$.

B. Violating Definition 3. Next, we'll give a 3-candidate RUM for which $U_{AH} < U_{HH}$ does not hold in general. Consider the following 3-candidate example.

$$x_1 = 3$$
$$x_2 = 2$$
$$x_3 = 0$$

Choose $\mathcal{E}$ and θ such that

$$\frac{\varepsilon}{\theta} = \begin{cases} 1 & \text{w.p. } \frac{1-\delta}{2} \\ -1 & \text{w.p. } \frac{1-\delta}{2} \\ 10 & \text{w.p. } \frac{\delta}{2} \\ -10 & \text{w.p. } \frac{\delta}{2} \end{cases}$$

Again, while this noise model doesn't satisfy Definition 1, we can approximate it arbitrarily closely with the sum of tightly concentrated Gaussians. Let the $\theta_A = 1.1\theta$ and $\theta_H = 0.9\theta$.

We will show that for these parameters, $U_{AH}(\theta_A, \theta_H) > U_{HH}(\theta_A, \theta_H)$, i.e., it is somehow better to choose after a better opponent than after a worse opponent. At a high level, the reasoning for this is as follows:

1. When choosing first, the only difference between the algorithm and the human evaluator is that the algorithm is more likely to choose x_2 than x_3. Both strategies have identical probabilities of selecting x_1.

2. When choosing second, the human evaluator's utility is higher when x_2 has already been chosen than when x_3 has already been chosen. This is because when x_2 is unavailable, the human evaluator is almost guaranteed to get x_1; when x_3 is unavailable, the human evaluator will choose x_2 with probability $\approx 1/4$.

Let τ and π be rankings generated by the algorithm and human evaluator respectively. First, we will show that

$$\Pr[\tau_1 = x_1] = \Pr[\pi_1 = x_1] \quad [2]$$
$$\Pr[\tau_1 = x_2] > \Pr[\pi_1 = x_2] \quad [3]$$

To do so, consider the realizations of $\varepsilon_1, \varepsilon_2, \varepsilon_3$ that result in different rankings under θ_A and θ_H. In fact, the only set of realizations that result in different rankings are when $\varepsilon_2/\theta = -1$ and $\varepsilon_3/\theta = 1$. Thus, the algorithm and human evaluator always rank x_1 in the same position, conditioned on a realization, which proves Eq. (2); the only difference is that the algorithm sometimes ranks x_2 above x_3 when the human evaluator does not. Moreover, whenever $\varepsilon_1/\theta = -10$, x_2 is more strictly more likely to be ranked first under the algorithm than the human evaluator, which proves Eq. (3).

Next, we must show that when choosing second, the human evaluator is better off when x_2 is unavailable than when x_3 is unavailable. This is clearly true because for the human evaluator,

$$\Pr\left[x_1 + \frac{\varepsilon_1\theta_H}{\theta} > x_3 + \frac{\varepsilon_3\theta_H}{\theta}\right] \approx 1 - O(\delta)$$
$$\Pr\left[x_1 + \frac{\varepsilon_1\theta_H}{\theta} > x_2 + \frac{\varepsilon_3\theta_H}{\theta}\right] \approx \frac{3}{4}$$

Thus, conditioned on x_2 being unavailable, the human evaluator gets utility ≈ 3, whereas when x_3 is unavailable, the human evaluator gets utility ≈ 2.75. Let u_{-i} be the expected utility for the human evaluator when x_i is unavailable. Putting this together, we get

$$\begin{aligned}
U_{AH}(\theta_A, \theta_H) - U_{HH}(\theta_A, \theta_H) &= \sum_{i=1}^{3} (\Pr[\tau_1 = x_i] - \Pr[\pi_1 = x_i]) u_{-i} \\
&= (\Pr[\tau_1 = x_1] - \Pr[\pi_1 = x_1])u_{-1} + (\Pr[\tau_1 = x_2] - \Pr[\pi_1 = x_2])u_{-2} + (\Pr[\tau_1 = x_3] - \Pr[\pi_1 = x_3])u_{-3} \\
&= (\Pr[\tau_1 = x_2] - \Pr[\pi_1 = x_2])u_{-2} + (\Pr[\tau_1 = x_3] - \Pr[\pi_1 = x_3])u_{-3} \quad [\Pr[\tau_1 = x_1 = \Pr[\pi_1 = x_1]] \\
&= (\Pr[\tau_1 = x_2] - \Pr[\pi_1 = x_2])(u_{-2} - u_{-3}) \qquad [\textstyle\sum_{i=1}^{3} \Pr[\tau_1 = x_i] = \sum_{i=1}^{3} \Pr[\pi_1 = x_i]] \\
&> 0
\end{aligned}$$

The last step follows from Eq. (3) and because $u_{-2} > u_{-3}$.

3. Proof of Theorem 2

A. Verifying Definition 2. By Eq. (2) from the main paper, we can equivalently show that for any θ, $U_{AH}(\theta, \theta) > U_{AA}(\theta, \theta)$. Let τ and π be the algorithmic and human-generated rankings respectively. Note that they're identically distributed because $\theta_A = \theta_H$. Define

$$Y \triangleq \begin{cases} \pi_1 & \pi_1 \neq \tau_1 \\ \pi_2 & \text{otherwise} \end{cases}$$

Note that $U_{AH}(\theta, \theta) = \mathbb{E}[Y]$ and $U_{AA}(\theta, \theta) = \mathbb{E}[\tau_2]$. We want to show that $U_{AH}(\theta, \theta) - U_{AA}(\theta, \theta) = \mathbb{E}[x_Y - x_{\tau_2}] > 0$. It is sufficient to show that for any k, $\mathbb{E}[Y - \tau_2 \mid \tau_1 = x_k] > 0$. Let $X_i = x_i + \varepsilon_i/\theta$. Note that for distinct i, j, k and $x_i > x_j$,

$$\begin{aligned}
\mathbb{E}[Y - \tau_2 \mid \tau_1 = x_k] > 0 &\Longleftarrow \frac{\Pr[Y = x_i \mid \tau_1 = x_k]}{\Pr[Y = x_j \mid \tau_1 = x_k]} > \frac{\Pr[\tau_2 = x_i \mid \tau_1 = x_k]}{\Pr[\tau_2 = x_j \mid \tau_1 = x_k]} \\
&\iff \Pr[Y = x_i \mid \tau_1 = x_k] > \Pr[\tau_2 = x_i \mid \tau_1 = x_k] \qquad [\text{numerator and denominator sum to 1}] \\
&\iff \Pr[X_i > X_j] > \Pr[X_i > X_j \mid X_k > X_i \cap X_k > X_j] \\
&\iff \Pr[X_i > X_j] > \mathbb{E}_{X_k}[\Pr[X_i > X_j \mid X_k = a, X_i < a, X_j < a]].
\end{aligned}$$

Thus, it suffices to show that for any a,

$$\Pr[X_i > X_j] > \Pr[X_i > X_j \mid X_i < a, X_j < a]. \quad [4]$$

Since $\Pr[X_i > X_j] = \lim_{a\to\infty} \Pr[X_i > X_j \mid X_i < a, X_j < a]$, it suffices to show that for all a,

$$\frac{d}{da}\Pr[X_i > X_j \mid X_i < a, X_j < a] \geq 0, \tag{5}$$

and that it is strictly positive for some a. In other words, the higher a is, the more likely i and j are to be correctly ordered. In Theorems 6 and 7, we show that Eq. (5) holds for both Laplacian and Gaussian noise respectively, which proves that RUMs based on both distributions satisfy Definition 2.

B. Verifying Definition 3. Next, we show that for both Laplacian and Gaussian distributions, $U_{AH}(\theta_A, \theta_H) < U_{HH}(\theta_A, \theta_H)$ for all $\theta_A > \theta_H$. In fact, for 3-candidate RUM families, we will show that this is always true for any *well-ordered* distribution, defined as follows.

Definition 4. *A noise model with density $f(\cdot)$ is **well-ordered** if for any $a > b$ and $c > d$,*

$$f(a-c)f(b-d) > f(a-d)f(b-c).$$

In other words, for a well-ordered noise model, given two numbers, two candidates are more likely to be correctly ordered than inverted conditioned on realizing those two numbers in some order. Lemma 1 shows that both Gaussian and Laplacian distributions are well-ordered.

Thus, it suffices to show that for any 3-candidate RUM with a well-ordered noise model, $U_{AH}(\theta_A, \theta_H) < U_{HH}(\theta_A, \theta_H)$ when $\theta_A > \theta_H$.

Theorem 5. *For 3 candidates with unique values $x_1 > x_2 > x_3$ and well-ordered i.i.d. noise with support $(-\infty, \infty)$, if $\theta_A > \theta_H$, then $U_{AH}(\theta_A, \theta_H) < U_{HH}(\theta_A, \theta_H)$.*

Proof. Define u_{-i} to be the expected utility of the maximum element of the human-generated ranking when i is not available. Because we're in the 3-candidate setting, we have

$$\begin{aligned} u_{-1} &= \lambda_1 x_2 + (1-\lambda_1)x_3 \\ u_{-2} &= \lambda_2 x_1 + (1-\lambda_2)x_3 \\ u_{-3} &= \lambda_3 x_1 + (1-\lambda_3)x_2 \end{aligned}$$

where $1/2 < \lambda_i < 1$. This is because the noise has support everywhere, so it is impossible to correctly rank any two candidates with probability 1, and any two candidates are more likely than not to be correctly ordered:

$$\Pr\left[\frac{\varepsilon_i}{\theta} - \frac{\varepsilon_j}{\theta} > -\delta\right] = \Pr[\varepsilon_i - \varepsilon_j \geq 0] + \Pr\left[0 > \frac{\varepsilon_i - \varepsilon_j}{\theta} > -\delta\right] > \frac{1}{2}$$

Note that $\lambda_2 > \lambda_1$ and $\lambda_2 > \lambda_3$, since

$$\lambda_2 = \Pr[\varepsilon_1 - \varepsilon_3 > -\theta(x_1 - x_3)] > \max\left\{\Pr[\varepsilon_1 - \varepsilon_3 > -\theta(x_2 - x_3)], \Pr[\varepsilon_1 - \varepsilon_3 > -\theta(x_1 - x_2)]\right\} = \max\{\lambda_1, \lambda_3\}.$$

Let $\tau \sim \mathcal{F}_{\theta_A}$ and $\pi \sim \mathcal{F}_{\theta_H}$. With this, we can write

$$\begin{aligned} U_{AH}(\theta_A, \theta_H) &= \sum_{i=1}^{3} \Pr[\tau_1 = i] u_{-i} \\ U_{HH}(\theta_A, \theta_H) &= \sum_{i=1}^{3} \Pr[\pi_1 = i] u_{-i} \end{aligned}$$

Define

$$\Delta p_i = \Pr[\tau_1 = i] - \Pr[\pi_1 = i]$$

Using Lemmas 2, and 3, we have

$$\begin{aligned} &\Delta p_1 > 0 && \text{[By monotonicity of RUM families, see Appendix 1]} \\ &\Delta p_1 \geq \Delta p_2 \\ &\Delta p_3 \leq 0 \end{aligned}$$

Also, $\Delta p_1 + \Delta p_2 + \Delta p_3 = 0$. We must show that

$$U_{AH}(\theta_A, \theta_H) - U_{HH}(\theta_A, \theta_H) = \sum_{i=1}^{3} \Delta p_i u_{-i} < 0.$$

We consider 2 cases.

Case 1: $\Delta p_2 \leq 0$.
Then, $\Delta p_1 = -(\Delta p_2 + \Delta p_3)$. This yields

$$\begin{aligned}\sum_{i=1}^{3} \Delta p_i u_{-i} &= \Delta p_1 u_{-1} + \Delta p_2 u_{-2} + \Delta p_3 u_{-3} \\ &\leq \Delta p_1 u_{-1} - \Delta p_1 \min(u_{-2}, u_{-3}) \\ &= \Delta p_1 \left(\lambda_1 x_2 + (1-\lambda_1)x_3 - \min\left\{\lambda_2 x_1 + (1-\lambda_2)x_3, \lambda_3 x_1 + (1-\lambda_3)x_2\right\}\right) \\ &\leq \Delta p_1 \left(\lambda_1 x_2 + (1-\lambda_1)x_3 - \min\left\{\lambda_2 x_1 + (1-\lambda_2)x_3, x_2\right\}\right)\end{aligned}$$

We can show that this is at most 0 regardless of which term attains the minimum. Because $\lambda_2 > \lambda_1$,

$$\begin{aligned}\lambda_1 x_2 + (1-\lambda_1)x_3 - \lambda_2 x_1 - (1-\lambda_2)x_3 &= \lambda_1 x_2 + x_3 - \lambda_1 x_3 - \lambda_2 x_1 - x_3 + \lambda_2 x_3 \\ &= \lambda_1 x_2 - \lambda_1 x_3 - \lambda_2 x_1 + \lambda_2 x_3 \\ &= \lambda_1(x_2 - x_3) + \lambda_2(x_3 - x_1) \\ &< \lambda_1(x_2 - x_3) + \lambda_1(x_3 - x_1) \\ &= \lambda_1(x_2 - x_1) \\ &< 0\end{aligned}$$

For the second term, we have

$$\lambda_1 x_2 + (1-\lambda_1)x_3 - x_2 = (1-\lambda_1)(x_3 - x_2) < 0.$$

Thus,

$$\sum_{i=1}^{3} \Delta p_i u_{-i} < 0.$$

Case 2: $\Delta p_2 > 0$. Note that $u_{-1} < x_2 < u_{-3}$. Then, using $\Delta p_3 = -(\Delta p_1 + \Delta p_2)$,

$$\begin{aligned}\sum_{i=1}^{3} \Delta p_i u_{-i} &= \Delta p_1 u_{-1} + \Delta p_2 u_{-2} + \Delta p_3 u_{-3} \\ &= \Delta p_1(u_{-1} - u_{-3}) + \Delta p_2(u_{-2} - u_{-3}) \\ &\leq \Delta p_2(u_{-1} - u_{-3}) + \Delta p_2(u_{-2} - u_{-3}) && [\Delta p_1 \geq \Delta p_2 \text{ and } u_{-1} < u_{-3}] \\ &= \Delta p_2(u_{-1} + u_{-2} - 2u_{-3}) \\ &\leq \Delta p_2(x_2 + x_1 - 2(\lambda_3 x_1 + (1-\lambda_3)x_2)) \\ &< \Delta p_2\left(x_2 + x_1 - 2\left(\frac{1}{2}x_1 + \frac{1}{2}x_2\right)\right) && [\lambda_3 > \tfrac{1}{2}] \\ &= 0\end{aligned}$$

Thus, $U_{AH}(\theta_A, \theta_H) < U_{HH}(\theta_A, \theta_H)$. □

C. Supplementary Lemmas for Random Utility Models.

Lemma 1. *Both Gaussian and Laplacian distributions are well-ordered.*

Proof. The Gaussian noise model is well-ordered:

$$\begin{aligned}f(a-c)f(b-d) &= \frac{1}{2\sigma^2\pi}\exp(-(a-c)^2 - (b-d)^2) \\ &= \frac{1}{2\sigma^2\pi}\exp(-(a-d)^2 - (b-c)^2 - 2(ac + bd - ad - bc)) \\ &= f(a-d)f(b-c)\exp(-2((a-b)(c-d))) \\ &< f(a-d)f(b-c)\end{aligned}$$

Laplacian noise is as well:

$$\begin{aligned}f(a-c)f(b-d) &= \frac{1}{4}\exp(-|a-c| - |b-d|) \\ f(a-d)f(b-c) &= \frac{1}{4}\exp(-|a-d| - |b-c|)\end{aligned}$$

It suffices to show that for $a > b$ and $c > d$, $|a-c| + |b-d| < |a-d| + |b-c|$. To show this, plot (a, b) and (c, d) in the (x, y) plane. Note that they're both below the $y = x$ line, and that the ℓ_1 distance between them is $|a-c| + |b-d|$. Moreover, the ℓ_1 distance between any two points must be realized by some Manhattan path, which is a combination of horizontal and vertical line segments. Consider the point (b, a), which is above the $y = x$ line. Any Manhattan path from (b, a) to (c, d) must cross the $y = x$ line at some point (w, w). Since (b, a) and (a, b) are equidistant from (w, w), for any Manhattan path from (b, a) to (c, d), there exists a Manhattan path from (a, b) to (c, d) passing through (w, w) of the same length, meaning the ℓ_1 distance from (a, b) to (c, d) is smaller than the ℓ_1 distance from (b, a) to (c, d). As a result, $|a-c| + |b-d| < |a-d| + |b-c|$. □

Next, we show a few basic facts. Let $f_A(r)$ be the density function of the joint realization $R = [X_1, \ldots, X_n] = [x_1 + \varepsilon_1/\theta_A, \ldots, x_n + \varepsilon_n/\theta_A]$ under the algorithmic ranking and $f_H(r)$ be the similarly defined density function under the human-generated ranking. Consider the "contraction" operation $r' = \text{cont}(r)$ such that $r'_i = x_i + (r_i - x_i) \cdot \frac{\theta_H}{\theta_A}$. Essentially, the contraction defines a coupling between $f_A(\cdot)$ and $f_H(\cdot)$, since for $r' = \text{cont}(r)$, $f_A(r')\, dr' = f_H(r)\, dr$. Let $\pi(r)$ be the ranking induced by r. Note that contraction cannot introduce any new inversions in $\pi(r)$—that is, if i is ranked above j in $\pi(r)$ for $i < j$, then i is ranked above j in $\pi(\text{cont}(r))$. Intuitively, this is because contraction pulls values closer to their means, and can therefore only correct existing inversions, not introduce new ones. This fact will allow us to prove some useful lemmas.

Lemma 2. *If F_θ is a RUM family satisfying Definition 1, then for $\tau \sim \mathcal{F}_{\theta_A}$ and $\pi \sim \mathcal{F}_{\theta_H}$,*

$$\Pr[\tau_1 = x_n] \leq \Pr[\pi_1 = x_n]$$

Proof. Consider any realization r. Because inversions can only be corrected, not generated, by contraction, if $\pi_1(r') = n$, then $\pi_1(r) = n$ where $r' = \text{cont}(r)$. Since r' and r have equal measure under f_A and f_H respectively, we have

$$\begin{aligned}
\Pr[\pi_1 = x_n] &= \int_{\mathbb{R}^n} f_H(r) \mathbb{1}_{\pi_1(r) = x_n}\, dr \\
&= \int_{\mathbb{R}^n} f_A(\text{cont}(r)) \mathbb{1}_{\pi_1(r) = x_n}\, d\,\text{cont}(r) \\
&\geq \int_{\mathbb{R}^n} f_A(\text{cont}(r)) \mathbb{1}_{\pi_1(\text{cont}(r)) = x_n}\, d\,\text{cont}(r) \\
&= \int_{\mathbb{R}^n} f_A(r) \mathbb{1}_{\pi_1(r) = x_n}\, dr \\
&= \Pr[\tau_1 = x_n]
\end{aligned}$$

□

Next, we prove the following result for well-ordered noise models.

Lemma 3. *For any $i > 1$, if the noise model $\mathcal{E}$ is well-ordered, for $\theta_A \geq \theta_H$, $\tau \sim \mathcal{F}_{\theta_A}$, and $\pi \sim \mathcal{F}_{\theta_H}$,*

$$\Pr[\tau_1 = x_1] - \Pr[\pi_1 = x_1] \geq \Pr[\tau_1 = x_i] - \Pr[\pi_1 = x_i]$$

Proof. For $j \neq i$, let $S_{j \to i} \subseteq \mathbb{R}^n$ be the set of realizations r such that $\pi_1(r) = x_j$ and $\pi_1(\text{cont}(r)) = x_i$. Note that $S_{j \to i} = \emptyset$ for $j < i$ because contraction cannot create inversions. Then, we have that

$$\Pr[\tau_1 = x_i] - \Pr[\pi_1 = x_i] = \sum_{j > i} \int_{\mathbb{R}^n} f_H(r) \mathbb{1}_{r \in S_{j \to i}}\, dr - \sum_{j < i} \int_{R^n} f_H(r) \mathbb{1}_{r \in S_{i \to j}}\, dr \leq \sum_{j > i} \int_{\mathbb{R}^n} f_H(r) \mathbb{1}_{r \in S_{j \to i}}\, dr$$

Define

$$\text{swap}_i(r) = r',$$

where

$$r'_j = \begin{cases} r_j & j \notin \{1, i\} \\ r_1 & j = i \\ r_i & j = 1 \end{cases}$$

Intuitively, the swap_i operation simply swaps the realizations in positions 1 and i. Note that this is a bijection. Also, if $r \in S_{j \to i}$, then $\text{swap}_i(r) \in S_{j \to 1}$, since

$$\begin{aligned}
&\text{cont}(\text{swap}_i(r))_1 \geq \text{cont}(r)_i \geq \max_j \text{cont}(r)_j \geq \max_{j \notin \{1, i\}} \text{cont}(\text{swap}_i(r))_j \\
&\text{cont}(\text{swap}_i(r))_1 \geq \text{cont}(r)_i \geq \text{cont}(r)_1 \geq \text{cont}(\text{swap}_i(r))_i
\end{aligned}$$

Furthermore for $r \in S_{j\to i}$, $f_H(r) \le f_H(\text{swap}_i(r))$ since

$$\frac{f_H(\text{swap}_i(r))}{f_H(r)} = \frac{f(r_i - x_1)f(r_1 - x_i)}{f(r_1 - x_1)f(r_i - x_i)} \ge 1$$

because the noise is well-ordered and $r \in S_{j\to i}$ implies $r_i > r_1$. Thus,

$$\begin{aligned}
\sum_{j>i} \int_{\mathbb{R}^n} f_H(r)\mathbb{1}_{r\in S_{j\to i}}\, dr &\le \sum_{j>i} \int_{\mathbb{R}^n} f_H(\text{swap}_i(r))\mathbb{1}_{r\in S_{j\to i}}\, dr \\
&\le \sum_{j>i} \int_{\mathbb{R}^n} f_H(\text{swap}_i(r))\mathbb{1}_{\text{swap}_i(r)\in S_{j\to 1}}\, dr \\
&\le \sum_{j>i} \int_{\mathbb{R}^n} f_H(r)\mathbb{1}_{r\in S_{j\to 1}}\, dr \\
&\le \sum_{j>1} \int_{\mathbb{R}^n} f_H(r)\mathbb{1}_{r\in S_{j\to 1}}\, dr \\
&= \Pr[\tau_1 = x_1] - \Pr[\pi_1 = x_1]
\end{aligned}$$

□

Finally, we show that Eq. (5) holds for both Laplacian and Gaussian noise.

Theorem 6. *For any $a \in \mathbb{R}$ and $X_i = x_i + \sigma\varepsilon_i$ where ε_i is Laplacian with unit variance,*

$$\frac{d}{da}\Pr[X_i > X_j \mid X_i < a, X_j < a] \ge 0.$$

Moreover, it is strictly positive for some a.

Proof. First, we must derive an expression for $\Pr[X_i > X_j \mid X_i < a, X_j < a]$. Recall that the Laplace distribution parameterized by μ and λ has pdf

$$f(x;\mu,\lambda) = \frac{\lambda}{2}\exp(-\lambda|x-\mu|)$$

and cdf

$$F(x;\mu,\lambda) = \begin{cases} \frac{1}{2}\exp\left(-\lambda(\mu - x)\right) & x < \mu \\ 1 - \frac{1}{2}\exp\left(-\lambda(x-\mu)\right) & x \ge \mu \end{cases}$$

Note that x_i and x_j be the respective means of X_i and X_j, with $x_i > x_j$. Because the Laplace distribution is piecewise defined, we must consider 3 cases and show that in all 3 cases, Eq. (5) holds. Note that

$$\Pr[X_i > X_j \mid X_i < a, X_j < a] = \frac{\int_{-\infty}^{a} f(x;x_i,\lambda)F(x;x_j,\lambda)\, dx}{F(a;x_i,\lambda)F(a;x_j,\lambda)} \qquad [6]$$

Case 1: $a \le x_j$.
Then, the numerator of Eq. (6) is

$$\begin{aligned}
\int_{-\infty}^{a} \frac{\lambda}{2}\exp\left(-\lambda(x_i - x)\right)\cdot\frac{1}{2}\exp(-\lambda(x_j - x))\, dx &= \frac{\lambda}{4}\int_{-\infty}^{a}\exp(-\lambda(x_i + x_j - 2x))\, dx \\
&= \frac{\lambda\exp(-\lambda(x_i+x_j))}{4}\int_{-\infty}^{a}\exp(2\lambda x)\, dx \\
&= \frac{\lambda\exp(-\lambda(x_i+x_j))}{4}\frac{1}{2\lambda}\exp(2\lambda a) \\
&= \frac{\exp(-\lambda(x_i+x_j-2a))}{8}
\end{aligned}$$

The denominator is

$$\frac{1}{2}\exp(-\lambda(x_i - a))\cdot\frac{1}{2}\exp(-\lambda(x_j - a)) = \frac{1}{4}\exp(-\lambda(x_i + x_j - 2a)).$$

Thus,

$$\Pr[X_i > X_j \mid X_i < a, X_j < a] = \frac{1}{2},$$

so its derivative is trivially nonnegative.

Case 2: $x_j < a \leq x_i$.
Then, the numerator of Eq. (6) is

$$\begin{aligned}
&\int_{-\infty}^{x_j} \frac{\lambda}{2} \exp\left(-\lambda(x_i - x)\right) \cdot \frac{1}{2} \exp(-\lambda(x_j - x))\, dx + \int_{x_j}^{a} \frac{\lambda}{2} \exp\left(-\lambda(x_i - x)\right) \left(1 - \frac{1}{2} \exp(-\lambda(x - x_j))\right)\, dx \\
&= \frac{\exp(-\lambda(x_i - x_j))}{8} + \frac{\lambda}{2} \int_{x_j}^{a} \exp(-\lambda(x_i - x))\, dx - \frac{\lambda}{4} \int_{x_j}^{a} \exp(-\lambda(x_i - x_j))\, dx \\
&= \frac{\exp(-\lambda(x_i - x_j))}{8} + \frac{\lambda}{2} \frac{1}{\lambda} (\exp(-\lambda(x_i - a)) - \exp(-\lambda(x_i - x_j))) - \frac{\lambda}{4}(a - x_j) \exp(-\lambda(x_i - x_j)) \\
&= \frac{1}{2} \exp(-\lambda(x_i - a)) - \left(\frac{3}{8} + \frac{\lambda}{4}(a - x_j)\right) \exp(-\lambda(x_i - x_j))
\end{aligned}$$

The denominator is

$$\left(1 - \frac{1}{2} \exp(-\lambda(a - x_j))\right) \cdot \frac{1}{2} \exp(\lambda(x_j - a)) = \frac{1}{2} \exp(-\lambda(x_i - a)) - \frac{1}{4} \exp(-\lambda(x_i - x_j))$$

We can factor out $\frac{1}{4} \exp(-\lambda(x_i - x_j))$ from both, so

$$\begin{aligned}
\Pr[X_i > X_j \mid X_i < a, X_j < a] &= \frac{2 \exp(\lambda(a - x_j)) - \left(\frac{3}{2} + \lambda(a - x_j)\right)}{2 \exp(\lambda(a - x_j)) - 1} \\
&= \frac{2 \exp(\lambda(a - x_j)) - 1 - \left(\frac{1}{2} + \lambda(a - x_j)\right)}{2 \exp(\lambda(a - x_j)) - 1} \\
&= 1 - \frac{\frac{1}{2} + \lambda(a - x_j)}{2 \exp(\lambda(a - x_j)) - 1}
\end{aligned}$$

Thus,

$$\begin{aligned}
&\frac{d}{da} \Pr[X_i > X_j \mid X_i < a, X_j < a] > 0 \\
\iff &\frac{d}{da} \frac{\frac{1}{2} + \lambda(a - x_j)}{2 \exp(\lambda(a - x_j)) - 1} < 0 \\
\iff &(2 \exp(\lambda(a - x_j)) - 1)\lambda < \left(\frac{1}{2} + \lambda(a - x_j)\right) 2\lambda \exp(\lambda(a - x_j)) \\
\iff &2 - \exp(-\lambda(a - x_j)) < 2\left(\frac{1}{2} + \lambda(a - x_j)\right) \\
\iff &1 - \exp(-\lambda(a - x_j)) < 2\lambda(a - x_j) \\
\iff &\exp(-\lambda(a - x_j)) > 1 - 2\lambda(a - x_j)
\end{aligned}$$

This is true because $\lambda(a - x_j) > 0$, and for $z > 0$,

$$\exp(-z) > 1 - z > 1 - 2z.$$

Case 3: $a > x_i$.
Then, the numerator of Eq. (6) is

$$\begin{aligned}
&\int_{-\infty}^{x_j} \frac{\lambda}{2} \exp\left(-\lambda(x_i - x)\right) \cdot \frac{1}{2} \exp(-\lambda(x_j - x))\, dx + \int_{x_j}^{x_i} \frac{\lambda}{2} \exp\left(-\lambda(x_i - x)\right) \left(1 - \frac{1}{2} \exp(-\lambda(x - x_j))\right)\, dx \\
&\quad + \int_{x_i}^{a} \frac{\lambda}{2} \exp(-\lambda(x - x_i)) \left(1 - \frac{1}{2} \exp(-\lambda(x - x_j))\right)\, dx \\
&\quad = \frac{1}{2} - \left(\frac{3}{8} + \frac{\lambda}{4}(x_i - x_j)\right) \exp(-\lambda(x_i - x_j)) + \frac{1}{2}(1 - \exp(-\lambda(a - x_i))) - \frac{\lambda}{4} \int_{x_i}^{a} \exp(-\lambda(2x - x_i - x_j))\, dx \\
&\quad = 1 - \left(\frac{3}{8} + \frac{\lambda}{4}(x_i - x_j)\right) \exp(-\lambda(x_i - x_j)) - \frac{1}{2} \exp(-\lambda(a - x_i)) \\
&\quad + \frac{1}{8} \exp(\lambda(x_i + x_j))(\exp(-2\lambda a) - \exp(-2\lambda x_i)) \\
&\quad = 1 - \left(\frac{1}{2} + \frac{\lambda}{4}(x_i - x_j)\right) \exp(-\lambda(x_i - x_j)) - \frac{1}{2} \exp(-\lambda(a - x_i)) + \frac{1}{8} \exp(-\lambda(2a - x_i - x_j))
\end{aligned}$$

The denominator is

$$\begin{aligned}&\left(1-\frac{1}{2}\exp(-\lambda(t-x_i))\right)\left(1-\frac{1}{2}\exp(-\lambda(t-x_j))\right)\\&=1-\frac{1}{2}\exp(-\lambda(a-x_i))-\frac{1}{2}\exp(-\lambda(a-x_j))+\frac{1}{4}\exp(-\lambda(2a-x_i-x_j))\end{aligned}$$

Thus,

$$\begin{aligned}&\Pr[X_i > X_j \mid X_i < a, X_j < a]\\&\quad=\frac{1-\left(\frac{1}{2}+\frac{\lambda}{4}(x_i-x_j)\right)\exp(-\lambda(x_i-x_j))-\frac{1}{2}\exp(-\lambda(a-x_i))+\frac{1}{8}\exp(-\lambda(2a-x_i-x_j))}{1-\frac{1}{2}\exp(-\lambda(a-x_i))-\frac{1}{2}\exp(-\lambda(a-x_j))+\frac{1}{4}\exp(-\lambda(2a-x_i-x_j))}\\&\quad\propto\frac{8-(4+2\lambda(x_i-x_j))\exp(-\lambda(x_i-x_j))-4\exp(-\lambda(a-x_i))+\exp(-\lambda(2a-x_i-x_j))}{4-2\exp(-\lambda(a-x_i))-2\exp(-\lambda(a-x_j))+\exp(-\lambda(2a-x_i-x_j))}\end{aligned}$$

We're interested in

$$\begin{aligned}&\frac{d}{da}\Pr[X_i > X_j \mid X_i < a, X_j < a] > 0\\\iff&(4-2\exp(-\lambda(a-x_i))-2\exp(-\lambda(a-x_j))+\exp(-\lambda(2a-x_i-x_j)\\&\quad\cdot(4\lambda\exp(-\lambda(a-x_i))-2\lambda\exp(-\lambda(2a-x_i-x_j)))\\&\quad>(8-4\exp(-\lambda(a-x_i))+\exp(-\lambda(2a-x_i-x_j))-(4+2\lambda(x_i-x_j))\exp(-\lambda(x_i-x_j)))\\&\quad\cdot(2\lambda\exp(-\lambda(a-x_i))+2\lambda\exp(-\lambda(a-x_j))-2\lambda\exp(-\lambda(2a-x_i-x_j)))\\\iff&16\exp(-\lambda(a-x_i))-8\exp(-\lambda(2a-x_i-x_j))-8\exp(-2\lambda(a-x_i))+4\exp(-\lambda(3a-2x_i-x_j))\\&\quad-8\exp(-\lambda(2a-x_i-x_j))+4\exp(-\lambda(3a-x_i-2x_j))+4\exp(-\lambda(3a-2x_i-x_j))\\&\quad-2\exp(-2\lambda(2a-x_i-x_j))\\&\quad>16\exp(-\lambda(a-x_i))+16\exp(-\lambda(a-x_j))-16\exp(-\lambda(2a-x_i-x_j))\\&\quad-8\exp(-2\lambda(a-x_i))-8\exp(-\lambda(2a-x_i-x_j))+8\exp(-\lambda(3a-2x_i-x_j))\\&\quad+2\exp(-\lambda(3a-2x_i-x_j))+2\exp(-\lambda(3a-x_i-2x_j))-2\exp(-2\lambda(2a-x_i-x_j))\\&\quad-2(4+2\lambda(x_i-x_j))\exp(-\lambda(a-x_j))-2(4+2\lambda(x_i-x_j))\exp(-\lambda(a+x_i-2x_j))\\&\quad+2(4+2\lambda(x_i-x_j))\exp(-2\lambda(a-x_j))\\\iff&\exp(-\lambda(3a-x_i-2x_j))\\&\quad>8\exp(-\lambda(a-x_j))-4\exp(-\lambda(2a-x_i-x_j))+\exp(-\lambda(3a-2x_i-x_j))\\&\quad-(4+2\lambda(x_i-x_j))\exp(-\lambda(a-x_j))-(4+2\lambda(x_i-x_j))\exp(-\lambda(a+x_i-2x_j))\\&\quad+(4+2\lambda(x_i-x_j))\exp(-2\lambda(a-x_j))\\\iff&\exp(-\lambda(2a-x_i-x_j))\\&\quad>8-4\exp(-\lambda(a-x_i))+\exp(-\lambda(2a-2x_i))\\&\quad-(4+2\lambda(x_i-x_j))-(4+2\lambda(x_i-x_j))\exp(-\lambda(x_i-x_j))+(4+2\lambda(x_i-x_j))\exp(-\lambda(a-x_j))\\\iff&\exp(-\lambda(2a-x_i-x_j))-8+4\exp(-\lambda(a-x_i))-\exp(-2\lambda(a-x_i))\\&\quad+(4+2\lambda(x_i-x_j))(1+\exp(-\lambda(x_i-x_j)))-(4+2\lambda(x_i-x_j))\exp(-\lambda(a-x_j))\\&\quad>0\end{aligned}\tag{7}$$

Note that for any $z \geq 0$, we have

$$\begin{aligned}(4+2z)(1+e^{-z})-8\geq 0&\iff(2+z)(1+e^{-z})\geq 4\\&\iff z+2e^{-z}+ze^{-z}\geq 2\end{aligned}$$

For $z = 0$, this holds with equality, and the left hand side is increasing since

$$\begin{aligned}\frac{d}{dx}z+2e^{-z}+ze^{-z}\geq 0&\iff 1-2e^{-z}+e^{-z}-ze^{-z}\geq 0\\&\iff 1\geq e^{-z}+ze^{-z}\\&\iff\frac{1}{1+z}\geq e^{-z}\\&\iff 1+z\leq e^{z}\end{aligned}$$

Therefore, choosing $z = \lambda(x_i - x_j)$ and plugging back to Eq. (7), we have

$$\begin{aligned}
&\exp(-\lambda(2a - x_i - x_j)) - 8 + 4\exp(-\lambda(a - x_i)) - \exp(-2\lambda(a - x_i)) \\
&+ (4 + 2\lambda(x_i - x_j))(1 + \exp(-\lambda(x_i - x_j))) - (4 + 2\lambda(x_i - x_j))\exp(-\lambda(a - x_j)) > 0 \\
\Longleftarrow\ &\exp(-\lambda(2a - x_i - x_j)) + 4\exp(-\lambda(a - x_i)) - \exp(-2\lambda(a - x_i)) - (4 + 2\lambda(x_i - x_j))\exp(-\lambda(a - x_j)) > 0 \\
\Longleftrightarrow\ &\exp(-\lambda(a - x_j)) + 4 - \exp(-\lambda(a - x_i)) - (4 + 2\lambda(x_i - x_j))\exp(-\lambda(x_i - x_j)) > 0 \\
\Longleftrightarrow\ &4(1 - \exp(-\lambda(x_i - x_j))) + \exp(-\lambda(a - x_i))(\exp(-\lambda(x_i - x_j)) - 1) - 2\lambda(x_i - x_j)\exp(-\lambda(x_i - x_j)) > 0 \\
\Longleftrightarrow\ &(4 - \exp(-\lambda(a - x_i)))(1 - \exp(-\lambda(x_i - x_j))) - 2\lambda(x_i - x_j)\exp(-\lambda(x_i - x_j)) > 0 \\
\Longleftarrow\ &3(1 - \exp(-\lambda(x_i - x_j))) - 2\lambda(x_i - x_j)\exp(-\lambda(x_i - x_j)) > 0 \qquad [\exp(-\lambda(a - x_i)) < 1]
\end{aligned}$$

Again letting $z = \lambda(x_i - x_j)$, this is true if and only if

$$\begin{aligned}
3(1 - e^{-z}) > 2ze^{-z} &\Longleftrightarrow 3(e^z - 1) > 2z \\
&\Longleftrightarrow 3e^z > 3 + 2z
\end{aligned}$$

which is true because $e^z > 1 + z$ for $z > 0$. This completes the proof for Case 3.

As a result, we have that

$$\frac{d}{da}\Pr[X_i > X_j \mid X_i < a, X_j < a] \geq 0$$

for all a, with strict inequality for some a, which proves the theorem. □

Theorem 7. *For any $a \in \mathbb{R}$ and $X_i = x_i + \sigma\varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}(0, 1)$,*

$$\frac{d}{da}\Pr[X_i > X_j \mid X_i < a, X_j < a] > 0.$$

Proof. Assume $\sigma = 1/\sqrt{2}$. This is without loss of generality because for any instance with arbitrary σ', there is an instance with $\sigma = 1/\sqrt{2}$ that yields the same distribution over rankings (simply by scaling all item values by σ/σ'). First, we have

$$\begin{aligned}
\Pr[X_i > X_j \mid X_i < a, X_j < a] &= \frac{\int_{-\infty}^{a} \Pr[X_i = x]\Pr[X_j < x]\,dx}{Pr[X_i < a]\Pr[X_j < a]} \\
&= \frac{\int_{-\infty}^{a} \exp(-(x - x_i)^2)/\sqrt{\pi} \cdot (1 + \operatorname{erf}(x - x_j))/2\,dx}{(1 + \operatorname{erf}(a - x_i))/2 \cdot (1 + \operatorname{erf}(a - x_j))/2} \\
&= \frac{2}{\sqrt{\pi}} \frac{\int_{-\infty}^{a} \exp(-(x - x_i)^2)(1 + \operatorname{erf}(x - x_j))\,dx}{(1 + \operatorname{erf}(a - x_i)) \cdot (1 + \operatorname{erf}(a - x_j))}
\end{aligned}$$

The derivative with respect to a is positive if and only if

$$\begin{aligned}
&(1 + \operatorname{erf}(a - x_i))(1 + \operatorname{erf}(a - x_j))\exp(-(a - x_i)^2)(1 + \operatorname{erf}(a - x_j)) \\
&> \int_{-\infty}^{a} \exp(-(x - x_i)^2)(1 + \operatorname{erf}(x - x_j))\,dx \\
&\cdot \frac{2}{\sqrt{\pi}}\left((1 + \operatorname{erf}(a - x_i))\exp(-(a - x_j)^2) + (1 + \operatorname{erf}(a - x_j))\exp(-(a - x_i)^2)\right) \qquad [8]
\end{aligned}$$

Let $t = a - x_i$ and $\delta = x_i - x_j$. Then, using the fact that

$$\begin{aligned}
\int_{-\infty}^{a} \exp(-(x - x_i)^2)(1 + \operatorname{erf}(x - x_j))\,dx &= \int_{-\infty}^{a - x_i} \exp(-x^2)\,dx + \int_{-\infty}^{a - x_i} \exp(-x^2)\operatorname{erf}(x + \delta)\,dx \\
&= \frac{\sqrt{\pi}}{2}(1 + \operatorname{erf}(a - x_i)) + \int_{-\infty}^{a - x_i} \exp(-x^2)\operatorname{erf}(x + \delta)\,dx,
\end{aligned}$$

Eq. (8) becomes

$$\begin{aligned}
&\frac{\sqrt{\pi}}{2} \cdot \frac{(1 + \operatorname{erf}(t))(1 + \operatorname{erf}(t + \delta))^2\exp(-t^2)}{(1 + \operatorname{erf}(t))\exp(-(t + \delta)^2) + (1 + \operatorname{erf}(t + \delta))\exp(-t^2)} > \frac{\sqrt{\pi}}{2}(1 + \operatorname{erf}(t)) + \int_{-\infty}^{t} \exp(-x^2)\operatorname{erf}(x + \delta)\,dx \\
\Longleftrightarrow\ &\frac{(1 + \operatorname{erf}(t))(1 + \operatorname{erf}(t + \delta))^2\exp(-t^2)}{(1 + \operatorname{erf}(t))\exp(-(t + \delta)^2) + (1 + \operatorname{erf}(t + \delta))\exp(-t^2)} - (1 + \operatorname{erf}(t)) - \frac{2}{\sqrt{\pi}}\int_{-\infty}^{t} \exp(-x^2)\operatorname{erf}(x + \delta)\,dx > 0 \qquad [9]
\end{aligned}$$

To show that this is true, we will use the fact that $f(t) > 0$ whenever the following conditions are met:

1. $f(t)$ is continuous and differentiable everywhere
2. $\lim_{t\to-\infty} f(t) = 0$
3. $\frac{d}{dt} f(t) > 0$

We'll show that these conditions hold for the LHS of Eq. (9).

$$\lim_{t\to-\infty} \frac{(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))^2\exp(-t^2)}{(1+\operatorname{erf}(t))\exp(-(t+\delta)^2)+(1+\operatorname{erf}(t+\delta))\exp(-t^2)} - (1+\operatorname{erf}(t)) - \frac{2}{\sqrt{\pi}}\int_{-\infty}^{t}\exp(-x^2)\operatorname{erf}(x+\delta)\,dx$$
$$= \lim_{t\to-\infty} \frac{(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))^2\exp(-t^2)}{(1+\operatorname{erf}(t))\exp(-(t+\delta)^2)+(1+\operatorname{erf}(t+\delta))\exp(-t^2)} \quad [10]$$

Observe that both the numerator and denominator of Eq. (10) are positive, so this limit must be at least 0. We can upper bound it by

$$\begin{aligned}\lim_{t\to-\infty} \frac{(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))^2\exp(-t^2)}{(1+\operatorname{erf}(t))\exp(-(t+\delta)^2)+(1+\operatorname{erf}(t+\delta))\exp(-t^2)} &\le \lim_{t\to-\infty} \frac{(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))^2\exp(-t^2)}{(1+\operatorname{erf}(t+\delta))\exp(-t^2)}\\ &= \lim_{t\to-\infty}(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))\\ &= 0\end{aligned}$$

Thus, the limit is 0. Now, we must show that the derivative is positive. The derivative is

$$\frac{d}{dt}\left[\frac{(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))^2\exp(-t^2)}{(1+\operatorname{erf}(t))\exp(-(t+\delta)^2)+(1+\operatorname{erf}(t+\delta))\exp(-t^2)} - (1+\operatorname{erf}(t)) - \frac{2}{\sqrt{\pi}}\int_{-\infty}^{t}\exp(-x^2)\operatorname{erf}(x+\delta)\,dx\right]$$
$$= \frac{d}{dt}\left[\frac{(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))^2\exp(-t^2)}{(1+\operatorname{erf}(t))\exp(-(t+\delta)^2)+(1+\operatorname{erf}(t+\delta))\exp(-t^2)}\right] - \frac{2}{\sqrt{\pi}}\exp(-t^2) - \frac{2}{\sqrt{\pi}}\exp(-t^2)\operatorname{erf}(t+\delta) \quad [11]$$

Taking this derivative and factoring out

$$\frac{2(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))\exp(4t^2)}{\sqrt{\pi}\left((\operatorname{erf}(t)+1)\,e^{t^2}+(\operatorname{erf}(\delta+t)+1)\,e^{(\delta+t)^2}\right)^2},$$

we get that Eq. (11) is positive if and only if

$$\begin{aligned}&\delta\sqrt{\pi}\exp((t+\delta)^2)(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta)) - \exp(2\delta t+t^2)(1+\operatorname{erf}(t+\delta)) + (1+\operatorname{erf}(t)) > 0\\ \iff&\delta\sqrt{\pi}\exp((t+\delta)^2)(1+\operatorname{erf}(t)) + \frac{1+\operatorname{erf}(t)}{1+\operatorname{erf}(t+\delta)} - \exp(2\delta t+t^2) > 0\\ \iff&\delta\sqrt{\pi}\exp(t^2)(1+\operatorname{erf}(t)) + \exp(-2\delta t-t^2)\frac{1+\operatorname{erf}(t)}{1+\operatorname{erf}(t+\delta)} - 1 > 0\\ \iff&(1+\operatorname{erf}(t))\left[\delta\sqrt{\pi}\exp(t^2) + \frac{\exp(-2\delta t-\delta^2)}{1+\operatorname{erf}(t+\delta)}\right] - 1 > 0\\ \iff&\frac{1+\operatorname{erf}(t)}{\exp(-t^2)}\left[\delta\sqrt{\pi} + \frac{\exp(-(t+\delta)^2)}{1+\operatorname{erf}(t+\delta)}\right] - 1 > 0\end{aligned} \quad [12]$$

Define

$$g(t) \triangleq \frac{1+\operatorname{erf}(t)}{\exp(-t^2)}.$$

Then, Eq. (12) is

$$\begin{aligned}&g(t)\left[\delta\sqrt{\pi} + \frac{1}{g(t+\delta)}\right] - 1 > 0\\ \iff&\frac{1}{g(t)} - \frac{1}{g(t+\delta)} < \delta\sqrt{\pi}\end{aligned}$$

By the Mean Value Theorem,

$$\frac{1}{g(t)} - \frac{1}{g(t+\delta)} = -\delta\left.\frac{d}{dt}\frac{1}{g(t)}\right|_{t=t^*}$$

for some $t \le t^* \le t+\delta$. Thus, it suffices to show that

$$\frac{d}{dt}\frac{1}{g(t)} > -\sqrt{\pi} \quad [13]$$

for all t. To do this, consider Mills Ratio (1)

$$R(t) \triangleq \exp(t^2/2) \int_t^\infty \exp(-x^2/2)\, dx.$$

Note that this is quite similar in functional form to $g(t)$, and with some manipulation, we can relate the two:

$$\begin{aligned}
R(t) &= \exp(t^2/2) \int_t^\infty \exp(-x^2/2)\, dx \\
R(\sqrt{2}t) &= \exp(t^2) \int_{\sqrt{2}t}^\infty \exp(-x^2/2)\, dx \\
&= \sqrt{2}\exp(t^2) \int_t^\infty \exp(-x^2)\, dx \\
&= \sqrt{2}\exp(t^2) \int_{-\infty}^{-t} \exp(-x^2)\, dx && [\exp(-x^2) \text{ is symmetric}] \\
R(-\sqrt{2}t) &= \sqrt{2}\exp(t^2) \int_{-\infty}^{t} \exp(-x^2)\, dx \\
&= \sqrt{2}\exp(t^2) \cdot \frac{\sqrt{\pi}}{2}(1 + \operatorname{erf}(t)) \\
&= \sqrt{\frac{\pi}{2}} \left(\frac{1 + \operatorname{erf}(t)}{\exp(-t^2)} \right) \\
R(-\sqrt{2}t) &= \sqrt{\frac{\pi}{2}} g(t)
\end{aligned}$$

Sampford (2, Eq. (3)) proved that $\frac{d}{dt}\frac{1}{R(t)} < 1$ for any t. Thus,

$$\frac{d}{dt}\frac{1}{g(t)} = \frac{d}{dt}\frac{1}{\sqrt{\frac{2}{\pi}}R(-\sqrt{2}t)} = \sqrt{\frac{\pi}{2}}\frac{d}{dt}\frac{1}{R(-\sqrt{2}t)} > \sqrt{\frac{\pi}{2}} \cdot 1 \cdot -\sqrt{2} = -\sqrt{\pi},$$

which proves Eq. (13) and completes the proof. □

4. Verifying that the Mallows Model Satisfies Definition 1

Theorem 8. *The family of distributions $\mathcal{F}_\theta$ produced by the Mallows Model with Kendall tau distance with $\theta = \phi - 1$ satisfies the conditions of Definition 1.*

Proof. We must show that $\mathcal{F}_\theta$ satisfies the differentiability, asymptotic optimality, and monotonicity conditions of Definition 1.

Differentiability: Let Π be the set of all permutations on n candidates. The probability of a realizing a particular permutation π under the Mallows model is

$$\Pr_\theta[\pi] = \frac{\phi^{-d(\pi,\pi^*)}}{\sum_{\pi' \in \Pi} \phi^{-d(\pi',\pi^*)}}$$

Both the numerator and denominator are differentiable with respect to $\theta = \phi - 1$, so $\Pr_\theta[\pi]$ is differentiable with respect to θ.

Asymptotic optimality: For the correct ranking π^*,

$$\Pr_\theta[\pi^*] = \frac{1}{Z},$$

where the normalizing constant Z is

$$Z = \sum_{\pi \in \Pi} \phi^{-d(\pi,\pi^*)}$$

In the limit,

$$\begin{aligned}
\lim_{\theta\to\infty} Z &= \lim_{\phi\to\infty} Z \\
&= \lim_{\phi\to\infty} \sum_{\pi\in\Pi} \phi^{-d(\pi,\pi^*)} \\
&= \lim_{\phi\to\infty} 1 + \sum_{\pi\neq\pi^*\in\Pi} \phi^{-d(\pi,\pi^*)} \\
&= 1 + \sum_{\pi\neq\pi^*\in\Pi} \lim_{\phi\to\infty} \phi^{-d(\pi,\pi^*)} \\
&= 1
\end{aligned}$$

because for any $\pi \neq \pi^*$, $d(\pi, \pi^*) \geq 1$. Therefore,

$$\lim_{\theta \to \infty} \Pr_{\theta}[\pi^*] = \lim_{\theta \to \infty} \frac{1}{Z} = 1$$

Monotonicity: We must show that for any $S \subset \mathbf{x}$, if $\pi_1^{(-S)}$ denotes the value of the top-ranked candidate according to π excluding candidates in S,

$$\mathbb{E}_{\mathcal{F}_{\theta'}}\left[\pi_1^{(-S)}\right] \geq \mathbb{E}_{\mathcal{F}_{\theta}}\left[\pi_1^{(-S)}\right].$$

For any $i \notin S$, let j be the smallest index such that $j > i$ and $j \notin S$. Consider any π such that $\pi_1^{(-S)} = x_j$. Then, swapping i and j yields a permutation $\hat{\pi}$ such that $\hat{\pi}_1^{(-S)} = x_i$. Moreover,

$$\Pr[\hat{\pi}] = \Pr[\pi] \cdot \phi^{\text{inv}(\pi) - \text{inv}(\hat{\pi})}.$$

Since $i < j$, $\text{inv}(\pi) - \text{inv}(\hat{\pi}) \geq 1$. Finally, note that swapping i and j is a bijection between $\{\pi : \pi_1^{(-S)} = x_j\}$ and $\{\pi : \pi_1^{(-S)} = x_i\}$. Thus,

$$\frac{\Pr[\pi_1^{(-S)} = x_i]}{\Pr[\pi_1^{(-S)} = x_j]} = \sum_{\pi : \pi_1^{(-S)} = x_j} \frac{\Pr[\pi]}{\Pr[\pi_1^{(-S)} = x_j]} \cdot \phi^{\text{inv}(\pi) - \text{inv}(\hat{\pi})}$$

Note that the terms $\frac{\Pr[\pi]}{\Pr[\pi_1^{(-S)} = x_j]}$ sum to 1, so this is sum is some polynomial in ϕ with nonnegative weights and integer powers of ϕ. As a result, it must have a positive derivative with respect to ϕ, i.e., for $i < j$,

$$\frac{d}{d\phi} \frac{\Pr[\pi_1^{(-S)} = x_i]}{\Pr[\pi_1^{(-S)} = x_j]} > 0$$

Let $\phi' > \phi$. Then,

$$\frac{\Pr_{\phi}[\pi_1^{(-S)} = x_i]}{\Pr_{\phi}[\pi_1^{(-S)} = x_j]} < \frac{\Pr_{\phi'}[\pi_1^{(-S)} = x_i]}{\Pr_{\phi'}[\pi_1^{(-S)} = x_j]}$$

Rearranging,

$$\frac{\Pr_{\phi}[\pi_1^{(-S)} = x_i]}{\Pr_{\phi'}[\pi_1^{(-S)} = x_i]} < \frac{\Pr_{\phi}[\pi_1^{(-S)} = x_j]}{\Pr_{\phi'}[\pi_1^{(-S)} = x_j]} \quad [14]$$

For $\theta' - \phi' - 1$ and $\theta = \phi - 1$,

$$\mathbb{E}_{\mathcal{F}_{\theta}}\left[\pi_1^{(-S)}\right] = \sum_{i \notin S} \Pr_{\phi}\left[\pi_1^{-(S)} = x_i\right] x_i$$

$$\mathbb{E}_{\mathcal{F}_{\theta'}}\left[\pi_1^{(-S)}\right] = \sum_{i \notin S} \Pr_{\phi'}\left[\pi_1^{-(S)} = x_i\right] x_i$$

By Lemma 4,

$$\mathbb{E}_{\mathcal{F}_{\theta'}}\left[\pi_1^{(-S)}\right] > \mathbb{E}_{\mathcal{F}_{\theta}}\left[\pi_1^{(-S)}\right],$$

which completes the proof. Note that we apply Lemma 4 indexing backwards from n to 1, ignoring elements in S, with $p_i = \Pr_{\phi}\left[\pi_1^{-(S)} = x_i\right]$ and $q_i = \Pr_{\phi'}\left[\pi_1^{-(S)} = x_i\right]$. Eq. (14) provides the condition that p_i/q_i is decreasing (as i decreases, since we are indexing backwards). □

5. Proof of Theorem 3

A. Verifying Definition 2. We must show that when $\pi, \tau \sim \mathcal{F}_{\theta}$,

$$\mathbb{E}\left[\pi_1 - \pi_2 \mid \pi_1 \neq \tau_1\right] > 0. \quad [15]$$

We begin by expanding:

$$\begin{aligned}
\mathbb{E}\left[\pi_1 - \pi_2 \mid \pi_1 \neq \tau_1\right] &= \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j) \Pr[\pi_1 = x_i \cap \pi_2 = x_j \mid \pi_1 \neq \tau_1] \\
&= \sum_{i=1}^{n-1} \sum_{j>i} (x_i - x_j) \left(\Pr[\pi_1 = x_i \cap \pi_2 = x_j \mid \pi_1 \neq \tau_1] - \Pr[\pi_1 = x_j \cap \pi_2 = x_i \mid \pi_1 \neq \tau_1]\right)
\end{aligned}$$

Since $x_i > x_j$ for $i < j$, it suffices to show that for all $i < j$,

$$\Pr[\pi_1 = x_i \cap \pi_2 = x_j \mid \pi_1 \neq \tau_1] \geq \Pr[\pi_1 = x_j \cap \pi_2 = x_i \mid \pi_1 \neq \tau_1], \quad [16]$$

and that this holds strictly for some $i < j$. We simplify Eq. (16) as follows:

$$\begin{aligned}
&\Pr[\pi_1 = x_i \cap \pi_2 = x_j \mid \pi_1 \neq \tau_1] > \Pr[\pi_1 = x_j \cap \pi_2 = x_i \mid \pi_1 \neq \tau_1] \\
&\iff \frac{\Pr[\pi_1 = x_i \cap \pi_2 = x_j \cap \pi_1 \neq \tau_1]}{\Pr[\pi_1 \neq \tau_1]} > \frac{\Pr[\pi_1 = x_j \cap \pi_2 = x_i \cap \pi_1 \neq \tau_1]}{\Pr[\pi_1 \neq \tau_1]} \\
&\iff \Pr[\pi_1 = x_i \cap \pi_2 = x_j \cap \pi_1 \neq \tau_1] > \Pr[\pi_1 = x_j \cap \pi_2 = x_i \cap \pi_1 \neq \tau_1] \\
&\iff \Pr[\pi_1 = x_i \cap \pi_2 = x_j \cap \tau_1 \neq x_i] > \Pr[\pi_1 = x_j \cap \pi_2 = x_i \cap \tau_1 \neq x_j] \\
&\iff \Pr[\pi_1 = x_i \cap \pi_2 = x_j] \Pr[\tau_1 \neq x_i] > \Pr[\pi_1 = x_j \cap \pi_2 = x_i] \Pr[\tau_1 \neq x_j] \quad [17]
\end{aligned}$$

We can simplify Eq. (17) using Lemmas 5 and 6. Let $|i - j|$ denote the difference in rank between x_i and x_j.

$$\begin{aligned}
&\Pr[\pi_1 = x_i \cap \pi_2 = x_j] \Pr[\tau_1 \neq x_i] - \Pr[\pi_1 = x_j \cap \pi_2 = x_i] \Pr[\tau_1 \neq x_j] \\
&= \Pr[\pi_1 = x_i \cap \pi_2 = x_j](1 - \Pr[\tau_1 = x_i]) - \phi^{-1} \Pr[\pi_1 = x_i \cap \pi_2 = x_j](1 - \Pr[\tau_1 = x_j]) \\
&= \Pr[\pi_1 = x_i \cap \pi_2 = x_j](1 - \Pr[\tau_1 = x_i]) - \phi^{-1} \Pr[\pi_1 = x_i \cap \pi_2 = x_j](1 - \phi^{-|i-j|} \Pr[\tau_1 = x_i]) \\
&= \Pr[\pi_1 = x_i \cap \pi_2 = x_j](1 - \Pr[\tau_1 = x_i] - \phi^{-1} - \phi^{-|i-j|-1} \Pr[\tau_1 = x_i]))
\end{aligned}$$

This is positive if and only if

$$\begin{aligned}
&1 - \Pr[\tau_1 = x_i] - \phi^{-1} - \phi^{-|i-j|-1} \Pr[\tau_1 = x_i] > 0 \\
&\iff \Pr[\tau_1 = x_i] \left(1 - \phi^{-|i-j|-1}\right) < 1 - \phi^{-1} \\
&\iff \Pr[\tau_1 = x_i] < \frac{1 - \phi^{-1}}{1 - \phi^{-|i-j|-1}} \\
&\iff \frac{1 - \phi^{-1}}{\phi^{i-1}(1 - \phi^{-n})} < \frac{1 - \phi^{-1}}{1 - \phi^{-|i-j|-1}} \\
&\iff \phi^{i-1}(1 - \phi^{-n}) > 1 - \phi^{-|i-j|-1}
\end{aligned}$$

This is weakly true for any $i < j$ because $\phi^{i-1} \geq 1$ and $|i - j| + 1 \leq n$, and it is strictly true for any i, j other than 1 and n. Thus, $\mathbb{E}[\pi_1 - \pi_2 \mid \pi_1 \neq \tau_1] > 0$.

B. Verifying Definition 3. Recall that Definition 3 is equivalent to $U_{AH}(\theta_A, \theta_H) < U_{HH}(\theta_A, \theta_H)$ for $\theta_A > \theta_H$. Let τ be the algorithmic ranking, and let π be a ranking from a human evaluator. Recall that $U_H(\theta_A, \theta_H) = \mathbb{E}[\pi_1]$. Throughout this proof, we will drop the (θ_A, θ_H) notation and simply write U_H, U_{AH}, and U_{HH}.

$$\begin{aligned}
U_{AH} &= \sum_{i=1}^{n} (\Pr[\pi_1 = x_i \cap \tau_1 \neq x_i] + \Pr[\pi_2 = x_i \cap \pi_1 = \tau_1]) x_i \\
&= \sum_{i=1}^{n} \Pr[\pi_1 = x_i \cap \tau_1 \neq x_i] x_i + \sum_{i=1}^{n} \Pr[\pi_2 = x_i \cap \pi_1 = \tau_1] x_i \\
&= \sum_{i=1}^{n} \left(Pr[\pi_1 = x_i] - \Pr[\pi_1 = x_i \cap \tau_1 = x_i]\right) x_i + \sum_{i=1}^{n} \sum_{j \neq i} \Pr[\pi_1 = x_j \cap \tau_1 = x_j \cap \pi_2 = x_i] x_i \\
&= U_H - \sum_{i=1}^{n} \Pr[\pi_1 = x_i \cap \tau_1 = x_i] x_i + \sum_{i=1}^{n} \Pr[\pi_1 = x_i \cap \tau_1 = x_i] \mathbb{E}[\pi_2 \mid \pi_1 = x_i \cap \tau_1 = x_i] \\
&= U_H + \sum_{i=1}^{n} \Pr[\pi_1 = x_i] \Pr[\tau_1 = x_i] \left(\mathbb{E}[\pi_2 \mid \pi_1 = x_i] - x_i\right)
\end{aligned}$$

Similarly, because two human evaluators are independent,

$$U_{HH} = U_H + \sum_{i=1}^{n} \Pr[\pi_1 = x_i]^2 \left(\mathbb{E}[\pi_2 \mid \pi_1 = x_i] - x_i\right).$$

Let $V_{-i} = \mathbb{E}[\pi_2 \mid \pi_1 = x_i]$. Note that conditioned on $\pi_1 = x_i$, the remaining elements of π_1 follow a Mallows model distribution over $n - 1$ candidates. Because the Mallows model is value-independent, increasing any item value increases the expected value of the top-ranked item (and in fact, the item ranked at any position). Thus, V_{-i} increases as i increases (since x_i, the value of

the unavailable candidate, decreases). Moreover, x_i is strictly decreasing in i, so $V_{-i} - x_i$ is strictly increasing in i. With this, we have

$$U_{AH} - U_H = \sum_{i=1}^{n} \Pr[\pi_1 = x_i] \Pr[\tau_1 = x_i] (V_{-i} - x_i)$$
$$U_{HH} - U_H = \sum_{i=1}^{n} \Pr[\pi_1 = x_i]^2 (V_{-i} - x_i)$$

Let $C_A = \Pr[\pi_1 = \tau_1] = \sum_{i=1}^{n} \Pr[\pi_1 = x_i] \Pr[\tau_1 = x_i]$, and similarly let $C_H = \sum_{i=1}^{n} \Pr[\pi_1 = x_i]^2$. $C_A > C_H$ by Lemma 4 with $y'_i = \Pr[\pi_1 = x_{n-i+1}]$, $p'_i = \Pr[\pi_1 = x_{n-i+1}]$ and $q'_i = \Pr[\tau_1 = x_{n-i+1}]$.

Let $p_i = \Pr[\pi'_1 = i] \Pr[\pi_1 = i]/C_A$, $q_i = \Pr[\pi'_1 = i]^2/C_H$, and $y_i = V_{-i} - x_i$. Then, we have

$$\frac{U_{AH} - U_H}{C_A} = \sum_{i=1}^{n} p_i y_i$$
$$\frac{U_{HH} - U_H}{C_H} = \sum_{i=1}^{n} q_i y_i$$

With $\phi_A = \theta_A + 1$ and $\phi_H = \theta_H + 1$,

$$\frac{p_i}{q_i} = \frac{C_H}{C_A} \cdot \frac{\frac{1-\phi_A^{-1}}{\phi_A^{i-1}(1-\phi_A^{-n})}}{\frac{1-\phi_H^{-1}}{\phi_H^{i-1}(1-\phi_H^{-n})}} \propto \frac{\phi_H^{i-1}}{\phi_A^{i-1}},$$

which is decreasing in i since $\phi_H < \phi_A$. By Lemma 4, $\sum_{i=1}^{n} p_i y_i < \sum_{i=1}^{n} q_i y_i$. Finally, note that $U_{HH} - U_H < 0$ by Lemma 7, so

$$\begin{aligned}
\sum_{i=1}^{n} p_i y_i &< \sum_{i=1}^{n} q_i y_i \\
\frac{U_{AH} - U_H}{C_A} &< \frac{U_{HH} - U_H}{C_H} \\
\frac{C_H(U_{AH} - U_H)}{C_A} &< U_{HH} - U_H \\
U_{AH} - U_H &< U_{HH} - U_H && [C_A > C_H, \text{ and } U_{HH} - U_H < 0] \\
U_{AH} &< U_{HH}
\end{aligned}$$

6. Supplementary Lemmas for the Mallows Model

Lemma 4. *Let $\{y_i\}_{i=1}^n$, $\{p_i\}_{i=1}^n$, and $\{q_i\}_{i=1}^n$ be sequences such that*

- *y_i is strictly increasing.*
- *$\sum_{i=1}^n p_i = \sum_{i=1}^n q_i = 1$.*
- *$\frac{p_i}{q_i}$ is decreasing.*

Then, $\sum_{i=1}^n p_i y_i < \sum_{i=1}^n q_i y_i$.

Proof. First, note that there exists j such that $p_i > q_i$ for $i < j$ and $p_i \leq q_i$ for $i \geq j$. To see this, let j be the smallest index such that $p_j \leq q_j$. Such a j must exist because p_i and q_i both sum to 1, so it cannot be the case that $p_i > q_i$ for all i. This implies $p_i/q_i \leq 1$, and since p_i/q_i is decreasing, $p_i \leq q_i$ for $i \geq j$.

Next, note that

$$\begin{aligned}
0 &= \sum_{i=1}^{n} (p_i - q_i) \\
&= \sum_{i=1}^{j-1} (p_i - q_i) + \sum_{i=j}^{n} (p_i - q_i),
\end{aligned}$$

meaning

$$\sum_{i=1}^{j-1} (p_i - q_i) = \sum_{i=j}^{n} (q_i - p_i).$$

Using this choice of j, we can write

$$\begin{aligned}
\sum_{i=1}^{n} p_i y_i - \sum_{i=1}^{n} q_i y_i &= \sum_{i=1}^{n} (p_i - q_i) y_i \\
&= \sum_{i=1}^{j-1} (p_i - q_i) y_i - \sum_{i=j}^{n} (q_i - p_i) y_i \\
&\leq \sum_{i=1}^{j-1} (p_i - q_i) y_{j-1} - \sum_{i=j}^{n} (q_i - p_i) y_j \\
&= \sum_{i=1}^{j-1} (p_i - q_i) y_{j-1} - \sum_{i=j}^{n} (q_i - p_i) y_j \\
&= \sum_{i=1}^{j-1} (p_i - q_i) y_{j-1} - \sum_{i=1}^{j-1} (p_i - q_i) y_j \\
&= \sum_{i=1}^{j-1} (p_i - q_i)(y_{j-1} - y_j) \\
&< 0
\end{aligned}$$

□

Lemma 5. *For $x_i > x_j$,*

$$\Pr[\pi_1 = x_i \cap \pi_2 = x_j] = \phi \Pr[\pi_1 = x_j \cap \pi_2 = x_i]. \quad [18]$$

Proof. Let π_{-ij} be a permutation of all of the candidates except x_i and x_j. Then, we have

$$\begin{aligned}
\Pr[\pi_1 = x_i \cap \pi_2 = x_j] &= \sum_{\pi_{-ij}} \Pr[\pi_1 = x_i \cap \pi_2 = x_j \mid \pi_{-ij}] \Pr[\pi_{-ij}] \\
&= \sum_{\pi_{-ij}} \phi \Pr[\pi_1 = x_j \cap \pi_2 = x_i \mid \pi_{-ij}] \Pr[\pi_{-ij}] \\
&= \phi \Pr[\pi_1 = x_j \cap \pi_2 = x_i]
\end{aligned}$$

Intuitively, given that x_i and x_j are in the top 2 positions, x_i followed by x_j is ϕ times more likely than x_j followed by x_i regardless of the remainder of the permutation, and therefore, x_i followed by x_j is ϕ times more likely overall. □

Lemma 6. *For $1 \leq i \leq n$,*

$$\Pr[\pi_1 = x_i] = \frac{1 - \phi^{-1}}{\phi^{i-1}(1 - \phi^{-n})}. \quad [19]$$

Proof. Let π_{-i} be a permutation over all items except i. Then,

$$\begin{aligned}
\Pr[\pi_1 = x_i] &= \sum_{\pi_{-i}} \Pr[\pi_1 = x_i \mid \pi_{-i}] \Pr[\pi_{-i}] \\
&= \sum_{\pi_{-i}} \phi^{-(i-1)} \Pr[\pi_{-i}] \\
&= \phi^{-(i-1)} \sum_{\pi_{-i}} \Pr[\pi_{-i}]
\end{aligned}$$

Note that $\Pr[\pi_{-i}]$ doesn't depend on *which* $n-1$ items are being ranked, so this term appears for any i. Moreover, $\sum_{i=1}^{n} \Pr[\pi_1 = x_i] = 1$. Therefore, we have

$$\Pr[\pi_1 = x_i] \propto \phi^{-(i-1)}.$$

Normalizing, we get

$$\begin{aligned}\Pr[\pi_1 = x_i] &= \frac{\phi^{-(i-1)}}{\sum_{j=1}^{n} \phi^{-(j-1)}} \\ &= \frac{\phi^{-(i-1)}}{\frac{1-\phi^{-n}}{1-\phi^{-1}}} \\ &= \frac{1-\phi^{-1}}{\phi^{i-1}(1-\phi^{-n})}\end{aligned}$$

Intuitively, any permutation over $n-1$ items is equally likely regardless of what those items are, and inserting any item at the front of this permutation yields a likelihood proportional to the number of additional inversions this causes, which is equal to the item's position on the list.* □

Lemma 7. *For the Mallows Model, $U_H(\theta_A, \theta_H) > U_{HH}(\theta_A, \theta_H)$.*

Proof. Intuitively, this is because selecting first is better than selecting second. To prove this, let π and τ be ranking generated by independent human evaluators under the Mallows Model, i.e., $\pi, \tau \sim \mathcal{F}_{\theta_H}$.

$$\begin{aligned}U_H(\theta_A, \theta_H) - U_{HH}(\theta_A, \theta_H) &= \mathbb{E}[\pi_1] - \mathbb{E}[\tau_1 \cdot \mathbb{1}_{\pi_1 \neq \tau_1} + \tau_2 \cdot \mathbb{1}_{\pi_1 = \tau_1}] \\ &= \mathbb{E}[(\pi_1 - \tau_2) \cdot \mathbb{1}_{\pi_1 = \tau_1}] \\ &= \mathbb{E}[(\pi_1 - \pi_2) \cdot \mathbb{1}_{\pi_1 = \tau_1}]\end{aligned}$$

For any $i < j$, conditioned on $\pi_1 = \tau_1$, they are more likely to be correctly ordered than not:

$$\begin{aligned}\mathbb{E}[(\pi_1 - \pi_2) \cdot \mathbb{1}_{\pi_1 = \tau_1}] &= \sum_{i<j} \left(\Pr[\pi_1 = x_i \cap \tau_1 = x_i \cap \pi_2 = x_j] - \Pr[\pi_1 = x_j \cap \tau_1 = x_j \cap \pi_2 = x_i]\right)(x_i - x_j) \\ &= \sum_{i<j} \left(\Pr[\pi_1 = x_i \cap \pi_2 = x_j]\Pr[\tau_1 = x_i] - \Pr[\pi_1 = x_j \cap \pi_2 = x_i]\Pr[\tau_1 = x_j]\right)(x_i - x_j) \\ &> \sum_{i<j} \left(\Pr[\pi_1 = x_i \cap \pi_2 = x_j]\Pr[\tau_1 = x_j] - \Pr[\pi_1 = x_j \cap \pi_2 = x_i]\Pr[\tau_1 = x_j]\right)(x_i - x_j) \\ &= \sum_{i<j} \left(\Pr[\pi_1 = x_i \cap \pi_2 = x_j] - \Pr[\pi_1 = x_j \cap \pi_2 = x_i]\right)(x_i - x_j) \\ &\geq \sum_{i<j} \left(\phi_H \Pr[\pi_1 = x_j \cap \pi_2 = x_i] - \Pr[\pi_1 = x_j \cap \pi_2 = x_i]\right)(x_i - x_j) \\ &> 0\end{aligned}$$

□

References

1. JP Mills, Table of the ratio: area to bounding ordinate, for any portion of normal curve. *Biometrika* **18**, 395–400 (1926).
2. MR Sampford, Some inequalities on Mill's ratio and related functions. *The Annals Math. Stat.* **24**, 130–132 (1953).

*Alternatively, we could prove this by showing that for any permutation with i in front, the permutation in which i and $i-1$ are swapped is ϕ times more likely, and thus, $i-1$ is ϕ times more likely to be in front than i.