A Proofs

A.1 Proof of Theorem 3.1

Let $\mathbf{x}'$ be the output of the gradient descent iteration update for input $\mathbf{x}$ with step size η.

If $\mathbf{x} \in \mathcal{X}_\gamma$ and $\mathbf{x}' \in \mathcal{X}_\mu$, then

$$\begin{aligned} & f(\mathbf{x}') - f(\mathbf{x}^*) \\ = \; & f(\mathbf{x} - \eta \nabla f(\mathbf{x})) - f(\mathbf{x}^*) \\ \leq \; & f(\mathbf{x}) - \eta ||\nabla f(\mathbf{x})||^2 + \frac{\mu}{2}\eta^2 ||\nabla f(\mathbf{x})||^2 - f(\mathbf{x}^*) \\ = \; & f(\mathbf{x}) - f(\mathbf{x}^*) - \left(\eta - \frac{\mu}{2}\eta^2\right) ||\nabla f(\mathbf{x})||^2 \\ \leq \; & f(\mathbf{x}) - f(\mathbf{x}^*) - 2\gamma \left(\eta - \frac{\mu}{2}\eta^2\right) (f(\mathbf{x}) - f(\mathbf{x}^*)) \\ = \; & (1 - 2\gamma\eta + \gamma\mu\eta^2)(f(\mathbf{x}) - f(\mathbf{x}^*)) \end{aligned}$$

where the first inequality is by the assumption that f is μ-smooth on $\mathcal{X}_\mu$ and the second inequality is by the assumption that f satisfies the γ-PL inequality on $\mathcal{X}_\gamma$. Taking $\eta = 1/\mu$, which minimizes the above bound, establishes the claim of the theorem.

A.2 Proof of Theorem 3.2

Let $\mathbf{x}'$ be the output of the MM algorithm iteration update for input $\mathbf{x}$.

By the facts $f(\mathbf{x}') \leq g(\mathbf{x}'; \mathbf{x})$ and $g(\mathbf{x}'; \mathbf{x}) \leq g(\mathbf{z}; \mathbf{x})$ for all $\mathbf{z}$, for any $\eta \geq 0$,

$$\begin{aligned} & f(\mathbf{x}') - f(\mathbf{x}^*) \\ \leq \; & g(\mathbf{x}'; \mathbf{x}) - f(\mathbf{x}^*) \\ \leq \; & g(\mathbf{x} - \eta \nabla f(\mathbf{x}); \mathbf{x}) - f(\mathbf{x}^*) \\ = \; & f(\mathbf{x} - \eta \nabla f(\mathbf{x})) - f(\mathbf{x}^*) \\ & + g(\mathbf{x} - \eta \nabla f(\mathbf{x}); \mathbf{x}) - f(\mathbf{x} - \eta \nabla f(\mathbf{x})). \end{aligned}$$

Now, by the same arguments as in the proof of Theorem 3.1, if $\mathbf{x} \in \mathcal{X}_\gamma$ and $\mathbf{x} - \eta \nabla f(\mathbf{x}) \in \mathcal{X}_\mu$, we have

$$f(\mathbf{x} - \eta \nabla f(\mathbf{x})) - f(\mathbf{x}^*) \leq (1 - 2\gamma\eta + \gamma\mu\eta^2)(f(\mathbf{x}) - f(\mathbf{x}^*)).$$

Next, if $\mathbf{x} \in \mathcal{X}_\gamma$ and $\mathbf{x} - \eta \nabla f(\mathbf{x}) \in \mathcal{X}_\mu$,

$$\begin{aligned} & g(\mathbf{x} - \eta \nabla f(\mathbf{x}); \mathbf{x}) - f(\mathbf{x} - \eta \nabla f(\mathbf{x})) \\ \leq \; & \frac{\delta}{2}\eta^2 ||\nabla f(\mathbf{x})||^2 \\ \leq \; & \delta \eta^2 \gamma (f(\mathbf{x}) - f(\mathbf{x}^*)) \end{aligned}$$

where the first inequality is by the smoothness condition on the majorant surrogate function and the second inequality is by the assumption that f satisfies the PL inequality with parameter γ on $\mathcal{X}_\gamma$.

Putting the pieces together, we have

$$f(\mathbf{x}') - f(\mathbf{x}^*) \leq \left(1 - 2\gamma\eta + \gamma(\mu + \delta)\eta^2\right) (f(\mathbf{x}) - f(\mathbf{x}^*)).$$

Taking $\eta = 1/(\mu + \delta)$ (which minimizes the factor involving η in the last inequality) yields the asserted result.

A.3 Proof of Lemma 3.1

The Hessian of the negative log-likelihood function has the following elements:

$$\nabla^2(-\ell(\mathbf{w}))_{i,j} = \begin{cases} \sum_{v \neq i} m_{i,v} \frac{e^{w_i} e^{w_v}}{(e^{w_i} + e^{w_v})^2}, & \text{if } i = j \\ -m_{i,j} \frac{e^{w_i} e^{w_j}}{(e^{w_i} + e^{w_j})^2}, & \text{if } i \neq j. \end{cases} \tag{A.1}$$

We will show that for all $i \neq j$,

$$\frac{\partial^2}{\partial w_i \partial w_j}(-\ell(\mathbf{w})) \leq -c_\omega m_{i,j} \text{ for all } \mathbf{w} \in [-\omega, \omega]^n \tag{A.2}$$

and

$$-\frac{1}{4} m_{i,j} \leq \frac{\partial^2}{\partial w_i \partial w_j}(-\ell(\mathbf{w})) \text{ for all } \mathbf{w} \in \mathbb{R}^n. \tag{A.3}$$

From (A.2), we have $\nabla^2(-\ell(\mathbf{w})) \succeq c_\omega \mathbf{L}_\mathbf{M}$ for all $\mathbf{w} \in [-\omega, \omega]^n$. Hence, for all $\mathbf{w} \in [-\omega, \omega]^n$ and $\mathbf{x} \in \mathcal{X}$

$$\mathbf{x}^\top \nabla^2(-\ell(\mathbf{w}))\mathbf{x} \geq c_\omega \lambda_2(\mathbf{L}_\mathbf{M}) ||\mathbf{x}||^2$$

where $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}^\top \mathbf{1} = 0\}$. This shows that $-\ell$ is $c_\omega \lambda_2(\mathbf{L}_\mathbf{M})$-strongly convex on $\mathcal{X}$.

From (A.3), we have $\frac{1}{4}\mathbf{L}_\mathbf{M} \succeq \nabla^2(-\ell(\mathbf{w}))$ for all $\mathbf{w} \in \mathbb{R}^n$. Hence,

$$\mathbf{x}^\top \nabla^2(-\ell(\mathbf{w}))\mathbf{x} \leq \frac{1}{4}\lambda_n(\mathbf{L}_\mathbf{M}) ||\mathbf{x}||^2 \text{ for all } \mathbf{x} \in \mathbb{R}^n.$$

This shows that $-\ell$ is $\frac{1}{4}\lambda_n(\mathbf{L}_\mathbf{M})$-smooth on $\mathbb{R}^n$.

It remains to show that (A.2) and (A.3) hold. For (A.2), we need to show that $c_\omega \leq x_i x_j/(x_i + x_j)^2$ for all $\mathbf{x} \in [-\omega, \omega]^n$. Note that $x_i x_j/(x_i + x_j)^2 = z(1 - z)$ where $z := x_i/(x_i + x_j)$. Note that $z \in \Omega := [e^{-\omega}/(e^{-\omega} + e^{\omega}), 1 - e^{-\omega}/(e^{-\omega} + e^{\omega})]$ for all $\mathbf{x} \in [-\omega, \omega]^n$. The function $z(1 - z)$ achieves its minimum over the interval Ω at a boundary of Ω. Thus, it holds $\min_{z \in \Omega} z(1 - z) = c_\omega$. For (A.3), we can immediately note that for all $\mathbf{w} \in \mathbb{R}^n$,

$$\frac{w_i w_j}{(w_i + w_j)^2} = \frac{w_i}{w_i + w_j}\left(1 - \frac{w_i}{w_i + w_j}\right) \leq \frac{1}{4}.$$

A.4 Proof of Lemma 3.3

Let $\mathbf{y}$ be an arbitrary vector in $[-\omega, \omega]^n$. Let $r(\mathbf{x}; \mathbf{y}) = \underline{\ell}(\mathbf{x}; \mathbf{y}) - \ell(\mathbf{x})$ for $\mathbf{x} \in [-\omega, \omega]^n$. Then, we have

$$\begin{aligned} & r(\mathbf{y}; \mathbf{y}) = 0, \nabla_\mathbf{x} r(\mathbf{y}; \mathbf{y}) = 0, \text{ and} \\ & \nabla_x^2 r(\mathbf{x}; \mathbf{y}) = \nabla^2(-\ell(\mathbf{x})) + A \end{aligned} \tag{A.4}$$

where A is a $n \times n$ diagonal matrix with diagonal elements

$$A_{i,i} = -\sum_{j \in i} m_{i,j} \frac{e^{x_i}}{e^{y_i} + e^{y_j}} \geq -\frac{1}{2} e^{2\omega} ||\mathbf{M}||_\infty.$$

Since $\nabla^2(-\ell(\mathbf{x}))$ is a positive semi-definite matrix and A is a diagonal matrix, for all $\mathbf{x}, \mathbf{y} \in [-\omega, \omega]^n$, we have for all $\mathbf{w} \in [-\omega, \omega]^n$

$$\mathbf{x}^\top \nabla^2_{\mathbf{x}} r(\mathbf{w}; \mathbf{y}) \mathbf{x} \geq -||\mathbf{M}||_\infty \frac{e^{2\omega}}{2} ||\mathbf{x}||^2 = -\delta ||\mathbf{x}||^2, .$$

By limited Taylor expansion, for all $\mathbf{x} \in [-\omega, \omega]^n$,

$$\begin{aligned}
& r(\mathbf{x}; \mathbf{y}) \\
\geq & r(\mathbf{y}; \mathbf{y}) + (\mathbf{x} - \mathbf{y})^\top \nabla_{\mathbf{x}} r(\mathbf{y}; \mathbf{y}) \\
& + \frac{1}{2} \min_{0 \leq a \leq 1} (\mathbf{x} - \mathbf{y})^\top \nabla^2_{\mathbf{x}} r(a\mathbf{x} + (1-a)\mathbf{y}; \mathbf{y})(\mathbf{x} - \mathbf{y}) \\
= & \frac{1}{2} \min_{0 \leq a \leq 1} (\mathbf{x} - \mathbf{y})^\top \nabla^2_{\mathbf{x}} r(a\mathbf{x} + (1-a)\mathbf{y})(\mathbf{x} - \mathbf{y}; \mathbf{y}) \\
\geq & -\frac{\delta}{2} ||\mathbf{x} - \mathbf{y}||^2.
\end{aligned}$$

By the definition of $r(\mathbf{x}; \mathbf{y})$, we have $\bar{\ell}(\mathbf{x}; \mathbf{y}) - \ell(\mathbf{x}) \geq -\frac{\delta}{2} ||\mathbf{x} - \mathbf{y}||^2$.

A.5 Surrogate function (3.3) for the Bradley-Terry model is a first-order surrogate function

We show that the surrogate function $\underline{\ell}$ of the log-likelihood function ℓ of the Bradley-Terry model, given by (3.3), is a first-order surrogate function on $\mathcal{X}_\omega = [-\omega, \omega]^n$ with $\mu_0 = \frac{1}{2} e^{2\omega} d(\mathbf{M})$.

We need to show that the error function $h(\mathbf{x}; \mathbf{y}) = \ell(\mathbf{x}) - \underline{\ell}(\mathbf{x}; \mathbf{y})$ is a μ_0-smooth function on $\mathcal{X}_\omega$.

By a straightforward calculus, we note

$$\nabla^2 h(\mathbf{x}; \mathbf{y}) = \nabla^2 \ell(\mathbf{x}) + D(\mathbf{x}, \mathbf{y})$$

where $D(\mathbf{x}, \mathbf{y})$ is a diagonal matrix with diagonal elements

$$d_u = \sum_{j \neq u} m_{u,j} \frac{e^{x_u}}{e^{y_u} + e^{y_j}}.$$

We can take

$$\mu_0 = \max_{\mathbf{x}, \mathbf{y} \in \mathcal{X}_\omega} \max\{|\lambda_1(\nabla^2 h(\mathbf{x}; \mathbf{y}))|, |\lambda_n(\nabla^2 h(\mathbf{x}; \mathbf{y}))|\}.$$

For any $A = B + D$ where B is a $n \times n$ matrix and D is a $n \times n$ diagonal matrix with diagonal elements $d_1, d_2, \ldots, d_n$, we have

$$\lambda_1(B) + \min_u d_u \leq \lambda_i(A) \leq \lambda_n(B) + \max_u d_u.$$

It thus follows that

$$\begin{aligned}
\mu_0 \leq \max_{\mathbf{x}, \mathbf{y} \in \mathcal{X}_\omega} & \max\{|\lambda_1(\nabla^2 \ell(\mathbf{x}))| + \\
& \min_u d_u|, |\lambda_n(\nabla^2 \ell(\mathbf{x})) + \max_u d_u|\}.
\end{aligned}$$

Now note that for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}_\omega$,

$$-\frac{1}{2} d(\mathbf{M}) \leq \lambda_1(\nabla^2 \ell(\mathbf{x})) \leq \lambda_n(\nabla^2 \ell(\mathbf{x})) = 0$$

and

$$\frac{1}{2} e^{-2\omega} \min_u \sum_{j \in u} m_{u,j} \leq \min_u d_u \leq \max_u d_u \leq \frac{1}{2} e^{2\omega} d(\mathbf{M}).$$

We have

$$|\lambda_n(\nabla^2 \ell(\mathbf{x})) + \max_u d_u| = \max_u d_u \leq \frac{1}{2} e^{2\omega} d(\mathbf{M})$$

and

$$\begin{aligned}
& |\lambda_1(\nabla^2 \ell(\mathbf{x})) + \min_u d_u| \\
= & (\lambda_1(\nabla^2 \ell(\mathbf{x})) + \min_u d_u) \mathbf{I}_{\lambda_1(\nabla^2 \ell(\mathbf{x})) + \min_u d_u \geq 0} \\
& + (-\lambda_1(\nabla^2 \ell(\mathbf{x})) - \min_u d_u) \mathbf{I}_{\lambda_1(\nabla^2 \ell(\mathbf{x})) + \min_u d_u < 0} \\
\leq & \min_u d_u \mathbf{I}_{\lambda_1(\nabla^2 \ell(\mathbf{x})) + \min_u d_u \geq 0} \\
& - \lambda_1(\nabla^2 \ell(\mathbf{x})) \mathbf{I}_{\lambda_1(\nabla^2 \ell(\mathbf{x})) + \min_u d_u < 0} \\
\leq & \frac{1}{2} e^{2\omega} d(\mathbf{M}) \mathbf{I}_{\lambda_1(\nabla^2 \ell(\mathbf{x})) + \min_u d_u \geq 0} \\
& + \frac{1}{2} d(\mathbf{M}) \mathbf{I}_{\lambda_1(\nabla^2 \ell(\mathbf{x})) + \min_u d_u < 0} \\
\leq & \frac{1}{2} e^{2\omega} d(\mathbf{M}).
\end{aligned}$$

A.6 Proof of Lemma 3.4

We consider the log-a posteriori probability function $\rho(\mathbf{w}) = \ell(\mathbf{w}) + \ell_0(\mathbf{w}) + \text{const}$ where ℓ is the log-likelihood function given by (3.1) and ℓ_0 is the prior log-likelihood function given by (3.5). Note that $\nabla^2(-\ell_0(\mathbf{w}))$ is a diagonal matrix with diagonal elements equal to βe^{w_i}, for $i = 1, 2, \ldots, n$. It can be readily shown that for $\mathbf{w} \in \mathcal{W}_\omega$,

$$c_\omega \mathbf{L}_\mathbf{M} + e^{-\omega} \beta \mathbf{I}_n \preceq \nabla^2(-\rho(\mathbf{w})) \preceq \frac{1}{4} \mathbf{L}_\mathbf{M} + e^{\omega} \beta \mathbf{I}_n. \tag{A.5}$$

From (A.5), for all $\mathbf{w} \in \mathcal{W}_\omega$ and $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{x}^\top \nabla^2(-\rho(\mathbf{w})) \mathbf{x} \geq \lambda_1(e^{-\omega} \beta \mathbf{I}_n) ||\mathbf{x}||^2 = e^{-\omega} \beta ||\mathbf{x}||^2.$$

Hence, $-\rho$ is $e^{-\omega}\beta$-strongly convex on $\mathcal{W}_\omega$.

Similarly, from (A.5), for all $\mathbf{w} \in \mathcal{W}_\omega$, and $\mathbf{x} \in \mathbb{R}^n$,

$$\begin{aligned}
\mathbf{x}^\top \nabla^2(-\rho(\mathbf{w})) \mathbf{x} & \leq \lambda_n(\frac{1}{4} \mathbf{L}_\mathbf{M} + e^{\omega} \beta \mathbf{I}_n) ||\mathbf{x}||^2 \\
& \leq (\lambda_n(\frac{1}{4} \mathbf{L}_\mathbf{M}) + \lambda_n(e^{\omega} \beta \mathbf{I}_n)) ||\mathbf{x}||^2 \\
& = (\frac{1}{4} \lambda_n(\mathbf{L}_\mathbf{M}) + e^{\omega} \beta) ||\mathbf{x}||^2.
\end{aligned}$$

Hence, $-\rho$ is μ-smooth on $\mathcal{W}_\omega$ with $\mu = \frac{1}{4}\lambda_n(\mathbf{L_M}) + e^\omega\beta$.

B Comparison of Theorem 3.2 with Proposition 2.7 Mairal (2015)

Theorem B.1. *Suppose that f is a strongly convex function on $\mathcal{X}_\gamma$ and $\mathbf{x}^*$ is a minimizer of f and that it holds $\mathbf{x}^* \in \mathcal{X}_\gamma$. Assume that g is a first-order surrogate function of f on $\mathcal{X}_\mu$ with parameter $\mu_0 > 0$. Let $\mathbf{x}^{(t+1)}$ be the output of the MM algorithm for input $\mathbf{x}^{(t)}$. Then, if $\mathbf{x}^{(t)} \in \mathcal{X}_\gamma$ and $\mathbf{x}^{(t+1)} \in \mathcal{X}_\mu$, then we have*

$$f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^*) \leq c(f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*))$$

where

$$c = \begin{cases} \frac{\mu_0}{\gamma}, & \text{if } \gamma > 2\mu_0 \\ 1 - \frac{\gamma}{4\mu_0}, & \text{if } \gamma \leq 2\mu_0. \end{cases}$$

Proof. If g is a first-order surrogate function on $\mathcal{X}_\mu$ with parameter μ_0, then

$$f(\mathbf{x}') \leq f(\mathbf{z}) + \frac{\mu_0}{2}||\mathbf{z} - \mathbf{y}||^2$$

where $\mathbf{x}' = \arg\min_{\mathbf{z}'} g(\mathbf{z}'; \mathbf{y})$.

From this it follows that

$$\begin{aligned} & f(\mathbf{x}') \\ \leq & \min_{\mathbf{z}} \left\{ f(\mathbf{z}) + \frac{\mu_0}{2}||\mathbf{z} - \mathbf{x}^*||^2 \right\} \\ \leq & \min_{a\in[0,1]} \left\{ f(a\mathbf{x}^* + (1-a)\mathbf{x}) + \frac{\mu_0 a^2}{2}||\mathbf{x} - \mathbf{x}^*||^2 \right\} \\ \leq & \min_{a\in[0,1]} \left\{ af(\mathbf{x}^*) + (1-a)f(\mathbf{x}) + \frac{\mu_0 a^2}{2}||\mathbf{x} - \mathbf{x}^*||^2 \right\} \end{aligned}$$

where the last inequality is by convexity of f.

We have established that

$$\begin{aligned} & f(\mathbf{x}') - f(\mathbf{x}^*) \\ \leq & \min_{a\in[0,1]} \left\{ (1-a)(f(\mathbf{x}) - f(\mathbf{x}^*)) + \frac{\mu_0 a^2}{2}||\mathbf{x} - \mathbf{x}^*||^2 \right\}. \end{aligned}$$

By assumption that f is γ-strongly convex on $\mathcal{X}_\gamma$ and $\mathbf{x} \in \mathcal{X}_\gamma$, we have

$$f(\mathbf{x}) - f(\mathbf{x}^*) \geq \frac{\gamma}{2}||\mathbf{x} - \mathbf{x}^*||^2.$$

It follows that

$$\begin{aligned} & f(\mathbf{x}') - f(\mathbf{x}^*) \\ \leq & \min_{a\in[0,1]} \left\{ 1 - a + \frac{\mu_0 a^2}{\gamma} \right\} (f(\mathbf{x}) - f(\mathbf{x}^*)). \end{aligned}$$

It remains only to note that

$$\min_{a\in[0,1]} \left\{ 1 - a + \frac{\mu_0 a^2}{\gamma} \right\} = c.$$

□

The rate of convergence bound derived from Theorem 3.2 can be tighter than the rate of convergence bound derived from Theorem B.1.

To show this consider the Bradley-Terry model for which we have shown in Lemma 3.3 that the surrogate function $\underline{\ell}$ of the log-likelihood function ℓ satisfies condition of Theorem 3.2 on $[-\omega, \omega]^n$ with $\delta = \frac{1}{2}e^{2\omega}d(\mathbf{M})$. It also holds that surrogate function $\underline{\ell}$ is also a first-order surrogate function of ℓ on $[-\omega, \omega]^n$ with $\mu_0 = \frac{1}{2}e^{2\omega}d(\mathbf{M})$. Hence in this case, we have $\delta = \mu_0$.

The convergence rate bound of Theorem 3.2 is tighter than the convergence rate bound of Theorem B.1 if and only if $\mu + \delta < 4\mu_0$. Since $\delta = \mu_0$, this is equivalent to $\mu < 3\delta$. Since by Lemma 3.1 we can take $\mu = \frac{1}{2}d(\mathbf{M})$, the latter condition reads as

$$1 < 3e^\omega$$

which indeed holds true.

C Generalized Bradley-Terry models

C.1 Generalized Bradley-Terry models

Bradley-Terry model of paired comparisons According to the Bradley-Terry model, each paired comparison of items i and j has two possible outcomes: either i wins against j ($i \succ j$) or j wins against i ($j \succ i$). The distribution of the outcomes is given by

$$\Pr[i \succ j] = \frac{e^{w_i}}{e^{w_i} + e^{w_j}}$$

where $\mathbf{w} = (w_1, w_2, \ldots, w_n)^\top \in \mathbb{R}^n$ are model parameters.

Rao-Kupper model of paired comparisons with ties The Rao-Kupper model is such that each paired comparison of items i and j has three possible outcomes: either $i \succ j$ or $j \succ i$ or $i \equiv j$ (tie). The model is defined by the probability distribution of outcomes that is given by

$$\Pr[i \succ j] = \frac{e^{w_i}}{e^{w_i} + \theta e^{w_j}}$$

and

$$\Pr[i \equiv j] = \frac{(\theta^2 - 1)e^{w_i}e^{w_j}}{(e^{w_i} + \theta e^{w_j})(\theta e^{w_i} + e^{w_j})}$$

where $\mathbf{w} = (w_1, w_2, \ldots, w_n)^\top \in \mathbb{R}^n$ and $\theta \geq 1$ are model parameters.

The larger the value of parameter θ, the more mass is put on the tie outcome. For the value of parameter $\theta = 1$, the model corresponds to the Bradley-Terry model for paired comparisons.

Luce choice model The Luce choice model is a natural generalization of the Bradley-Terry model of paired comparisons to comparison sets of two or more items. For any given comparison set $S \subseteq N = \{1, 2, \ldots, n\}$ of two or more items, the outcome is a choice of one item $i \in S$ (an event we denote as $i \succeq S$) which occurs with probability

$$\Pr[i \succeq S] = \frac{e^{w_i}}{\sum_{j \in S} e^{w_j}}$$

where $\mathbf{w} = (w_1, w_2, \ldots, w_n)^\top \in \mathbb{R}_n$ are model parameters.

We will use the following definitions and notation. Let T be the set of ordered sequences of two or more items from N such that for each $y = (y_1, y_2, \ldots, y_k) \in T$, y_1 is an arbitrary item and $y_2, \ldots, y_k$ are sorted in lexicographical order. We can interpret each $y = (y_1, y_2, \ldots, y_k) \in T$ as a choice of item y_1 from the set of items $\{y_1, y_2, \ldots, y_k\}$. According to the Luce's choice model, the probability of outcome y is given by

$$\Pr[Y = (y_1, y_2, \ldots, y_k)] = \frac{e^{w_{y_1}}}{\sum_{j \in y} e^{w_j}}.$$

We denote with d_y the number of observed outcomes y in the input data. For each $y \in T$, let $|y|$ denote the number of items in y.

Plackett-Luce ranking model The Plackett-Luce ranking model is a model of full rankings: for each comparison set of items $S \subseteq N = \{1, 2, \ldots, n\}$, the set of possible outcomes contains all possible permutations of items in S. The distribution over possible outcomes is defined as follows. Let T be the set of all possible permutations of subsets of two or more items from N. Each $y = (y_1, y_2, \ldots, y_k) \in T$ corresponds to a permutation of the set of items $S = \{y_1, y_2, \ldots, y_k\}$. The probability of outcome y is given by

$$\begin{aligned}&\Pr[Y = (y_1, y_2, \ldots, y_k)] \\ =&\frac{e^{w_{y_1}}}{\sum_{j=1}^{k} e^{w_{y_j}}} \frac{e^{w_{y_2}}}{\sum_{j=2}^{k} e^{w_{y_j}}} \cdots \frac{e^{w_{y_{k-1}}}}{\sum_{j=k-1}^{k} e^{w_{y_j}}}\end{aligned}$$

where $\mathbf{w} = (w_1, w_2, \ldots, w_n)^\top \in \mathbb{R}^n$ are model parameters.

The model has an intuitive explanation as a sampling of items without replacement proportional to the item weights e^{w_i}. The Plackett-Luce ranking model corresponds to the Bradley-Terry model of paired comparisons when the comparison sets consist of two items. We denote with d_y the number of observed outcomes y in the input data.

In this section, we discuss how the results for Bradley-Terry model of paired comparisons can be extended to other instances of generalized Bradley-Terry models. In particular, we show this for the Rao-Kupper model of paired comparisons with tie outcomes, the Luce choice model and the Plackett-Luce ranking model.

C.2 Rao-Kupper model

The probability distribution of outcomes according to the Rao-Kupper model is defined in Section C.1. The log-likelihood function can be written as

$$\begin{aligned}\ell(\mathbf{w}) =& \sum_{i=1}^{n} \sum_{j \neq i} \bar{d}_{i,j} \left(w_i - \log(e^{w_i} + \theta e^{w_j})\right) + \\ & \frac{1}{2} \sum_{i=1}^{n} t_{i,j} \log(\theta^2 - 1)\end{aligned}$$

where $\bar{d}_{i,j}$ is the number of observed paired comparisons of items i and j such that either i wins against j or there is a tie outcome, and $t_{i,j}$ is the number of observed paired comparisons of items i and j with tie outcomes.

Lemma C.1. *The negative log-likelihood function for the Rao-Kupper model of paired comparisons with parameter $\theta > 1$ is γ-strongly convex on $\mathcal{W}_\omega = \{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}||_\infty \leq \omega$ and $\mathbf{w}^\top \mathbf{1} = 0\}$ and μ-smooth on $\mathbb{R}^n$ with*

$$\gamma = c_{\theta,\omega} \lambda_2(\mathbf{L_M}) \text{ and } \mu = \frac{1}{2} \lambda_n(\mathbf{L_M})$$

where $c_{\theta,\omega} = \theta / (\theta e^{-\omega} + e^{\omega})^2$.

Proof of Lemma C.1 is provided in Appendix C.5.

A surrogate minorant function for the log-likelihood function of the Rao-Kupper model is given as follows:

$$\begin{aligned}&\underline{\ell}(\mathbf{x}; \mathbf{y}) \\ =& \sum_{i=1}^{n} \sum_{j \neq i} \bar{d}_{i,j} \left(x_i - \frac{e^{x_i} + \theta e^{x_j}}{e^{y_i} + \theta e^{y_j}} - \log(e^{y_i} + \theta e^{y_j}) + 1 \right) \\ &+ \frac{1}{2} \sum_{i=1}^{n} t_{i,j} \log(\theta^2 - 1).\end{aligned}$$

The MM algorithm is defined by, for $i = 1, 2, \ldots, n$,

$$\begin{aligned}w_i^{(t+1)} =& \log \left(\sum_{j \neq i} \bar{d}_{i,j} \right) - \\ & \log \left(\sum_{j \neq i} \left(\frac{\bar{d}_{i,j}}{e^{w_i^{(t)}} + \theta e^{w_j^{(t)}}} + \frac{\theta \bar{d}_{j,i}}{e^{w_j^{(t)}} + \theta e^{w_i^{(t)}}} \right) \right).\end{aligned}$$

Lemma C.2. *For all* $\mathbf{x}, \mathbf{y} \in [-\omega, \omega]^n$, $\underline{\ell}(\mathbf{x}; \mathbf{y}) - \ell(\mathbf{x}) \geq -\frac{\delta}{2}||\mathbf{x} - \mathbf{y}||^2$ *where*

$$\delta = e^{2\omega} d(\mathbf{M}).$$

C.3 Luce choice model

The probability distribution of outcomes according to the Luce choice model is defined in Section C.1. The log-likelihood function can be written as:

$$\ell(\mathbf{w}) = \sum_{y \in T} d_y \left(w_{y_1} - \log \left(\sum_{j \in y} e^{w_j} \right) \right).$$

Lemma C.3. *The negative log-likelihood function for the Luce choice model with comparison sets of size* $k \geq 2$ *is* γ*-strongly convex and* μ*-smooth on* $\mathcal{W}_\omega = \{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}||_\infty \leq \omega \text{ and } \mathbf{w}^\top \mathbf{1} = 0\}$ *with*

$$\gamma = c_{\omega,k} \lambda_2(\mathbf{L_M}) \text{ and } \mu = d_{\omega,k} \lambda_n(\mathbf{L_M})$$

where

$$c_{\omega,k} = \begin{cases} 1/(e^{-\omega} + e^{\omega})^2, & \text{if } k = 2 \\ 1/((k-2)e^{2\omega} + 2)^2, & \text{if } k > 2 \end{cases} \quad \text{and}$$

$$d_{\omega,k} = \frac{1}{((k-2)e^{-2\omega} + 2)^2}.$$

Note that for every fixed $\omega > 0$, (a) $c_{\omega,k}/d_{\omega,k}$ is decreasing in k, (b) $1/e^{8\omega} \leq c_{\omega,k}/d_{\omega,k} \leq 1/e^{2\omega}$, and (c) $1/e^{8\omega}$ is the limit value of $c_{\omega,k}/d_{\omega,k}$ as k goes to infinity.

A minorant surrogate function for the log-likelihood function of the Luce choice model is given by

$$\underline{\ell}(\mathbf{x}; \mathbf{y}) = \sum_{y \in T} d_y \left(x_{y_1} - \frac{\sum_{j \in y} e^{x_j}}{\sum_{j \in y} e^{y_j}} - \log \left(\sum_{j \in y} e^{y_j} \right) + 1 \right).$$

The MM algorithm iteration can be written as: for $i = 1, 2, \ldots, n$,

$$w_i^{(t+1)} = \log \left(\sum_{y \in T} d_y \mathbb{I}_{i=y_1} \right) - \log \left(\sum_{y \in T} d_y \mathbb{I}_{i \in y} \frac{1}{\sum_{j \in y} e^{w_j^{(t)}}} \right)$$

where $\sum_{y \in T} d_y \mathbb{I}_{i=y_1}$ is the number of observed comparisons in which item i is the chosen item.

Lemma C.4. *For all* $\mathbf{x}, \mathbf{y} \in [-\omega, \omega]^n$, $\underline{\ell}(\mathbf{x}; \mathbf{y}) - \ell(\mathbf{x}) \geq -\frac{\delta}{2}||\mathbf{x} - \mathbf{y}||^2$ *where*

$$\delta = \frac{1}{k(k-1)} e^{2\omega} d(\mathbf{M}).$$

C.4 Plackett-Luce ranking model

The probability distribution of outcomes according to the Plackett-Luce ranking model is defined in Section C.1. The log-likelihood function can be written as follows:

$$\ell(\mathbf{w}) = \sum_{y \in T} d_y \sum_{r=1}^{|y|-1} \left(w_{y_r} - \log \left(\sum_{j=r}^{|y|} e^{w_{y_j}} \right) \right).$$

Lemma C.5. *The negative log-likelihood function for the Plackett-Luce ranking model with comparison sets of size* $k \geq 2$ *is* γ*-strongly convex and* μ*-smooth on* $\mathcal{W}_\omega = \{\mathbf{w} \in \mathbb{R}^n : ||\mathbf{w}||_\infty \leq \omega \text{ and } \mathbf{w}^\top \mathbf{1} = 0\}$ *with*

$$\gamma = \tilde{c}_{\omega,k} \lambda_2(\mathbf{L_M}) \text{ and } \mu = \tilde{d}_{\omega,k} \lambda_n(\mathbf{L_M})$$

where

$$\tilde{c}_{\omega,k} = \frac{1}{k^2} e^{-4\omega} \text{ and } \tilde{d}_{\omega,k} = \left(2 - \frac{1}{k}\right) e^{4\omega}.$$

Proof of Lemma C.5 is provided in Appendix C.6.

Note that for fixed values of ω and k, Lemma C.5 implies the convergence time $\log(d(\mathbf{M})/a(\mathbf{M}))$. Note, however, that for fixed $\omega > 0$, $\tilde{c}_{\omega,k}/\tilde{d}_{\omega,k}$ decreases to 0 with k and is of the order $1/k^2$. This is because in the derivation of parameters $\tilde{c}_{\omega,k}$ and $\tilde{d}_{\omega,k}$ we use (conservative) deterministic bounds. Following Hajek et al. (2014), one can derive bounds for γ and μ that hold with high probability, which are such that $\tilde{c}_{\omega,k}$ and $\tilde{d}_{\omega,k}$ scale with k in the same way.

The log-likelihood function of the Plackett-Luce ranking model admits the following minorization function:

$$\underline{\ell}(\mathbf{x}; \mathbf{y}) = \sum_{y \in T} d_y \sum_{r=1}^{|y|-1} \left(x_{y_r} - \frac{\sum_{j=r}^{|y|} e^{x_{y_j}}}{\sum_{j=r}^{|y|} e^{y_{y_j}}} - \log \left(\sum_{j=r}^{|y|} e^{y_{y_j}} \right) + 1 \right).$$

The MM algorithm is given by: for $i = 1, 2, \ldots, n$,

$$w_i^{(t+1)} = \log \left(\sum_{y \in T} d_y \mathbb{I}_{i \in S_{1,|y|-1}(y)} \right) - \log \left(\sum_{y \in T} d_y \sum_{r=1}^{|y|-1} \mathbb{I}_{i \in S_{r,|y|}(y)} \frac{1}{\sum_{j=r}^{|y|} e^{w_{y_j}^{(t)}}} \right)$$

where $S_{a,b}(y) = \{y_a, y_{a+1}, \ldots, y_b\}$.

Lemma C.6. *For all* $\mathbf{x}, \mathbf{y} \in [-\omega, \omega]^n$, $\underline{\ell}(\mathbf{x}; \mathbf{y}) - \ell(\mathbf{x}) \geq -\frac{\delta}{2}||\mathbf{x} - \mathbf{y}||^2$ *where*

$$\delta = \frac{1}{2} e^{2\omega} d(\mathbf{M}).$$

C.5 Proof of Lemma C.1

Let $t_{i,j}$ be the number of paired comparisons in the input data with tie outcome for items i and j. Note that $t_{i,j} = t_{j,i}$. The log-likelihood function can be written as follows:

$$\begin{aligned}
&\ell(\mathbf{w}) \\
= &\sum_{i=1}^{n}\sum_{j\neq i} d_{i,j}\left(w_i - \log(e^{w_i}+\theta e^{w_j})\right) \\
&+\frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i} t_{i,j}\left(w_i + w_j - \log(e^{w_i}+\theta e^{w_j})\right) \\
&-\frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i} t_{i,j}\left(\log(\theta e^{w_i}+e^{w_j}) - \log(\theta^2-1)\right).
\end{aligned}$$

Let $\bar{d}_{i,j}$ be the number of paired comparisons of items i and j such that $i \succeq j$, i.e. $\bar{d}_{i,j} = d_{i,j} + t_{i,j}$. By a straightforward calculus, we can write

$$\begin{aligned}
\ell(\mathbf{w}) = &\sum_{i=1}^{n}\sum_{j\neq i} \bar{d}_{i,j}\left(w_i - \log(e^{w_i}+\theta e^{w_j})\right) \\
&+\frac{1}{2}\sum_{i=1}^{n} t_{i,j}\log(\theta^2-1).
\end{aligned}$$

Now, we note when $i \neq j$,

$$\begin{aligned}
&\frac{\partial^2}{\partial w_i \partial w_j}(-\ell(\mathbf{w})) \\
=&-\bar{d}_{i,j}\frac{\theta e^{w_i}e^{w_j}}{(e^{w_i}+\theta e^{w_j})^2} - \bar{d}_{j,i}\frac{\theta e^{w_i}e^{w_j}}{(\theta e^{w_i}+e^{w_j})^2}
\end{aligned}$$

and

$$\frac{\partial^2}{\partial w_i^2}(-\ell(\mathbf{w})) = -\sum_{j\neq i}\frac{\partial^2}{\partial w_u \partial w_j}(-\ell(\mathbf{w})).$$

For any $i \neq j$, it indeed holds

$$\frac{\theta e^{w_i}e^{w_j}}{(e^{w_i}+\theta e^{w_j})^2} \leq \frac{1}{4}.$$

Hence, when $i \neq j$,

$$\frac{\partial^2}{\partial w_i \partial w_j}(-\ell(\mathbf{w})) \geq -\frac{1}{4}(\bar{d}_{i,j}+\bar{d}_{j,i}) \geq -\frac{1}{2}m_{i,j}.$$

It follows that $\frac{1}{2}\mathbf{L_M} \succeq \nabla^2(-\ell(\mathbf{w}))$ for all $\mathbf{w} \in \mathbb{R}^n$. Hence,

$$\mathbf{x}^\top\nabla^2(-\ell(\mathbf{w}))\mathbf{x} \leq \frac{1}{2}\lambda_n(\mathbf{L_M}) \text{ for all } \mathbf{x} \in \mathbb{R}^n.$$

This implies that $-\ell$ is a $\frac{1}{2}\lambda_n(\mathbf{L_M})$-smooth function on $\mathbb{R}^n$.

On the other hand, we can show that for all $\mathbf{w} \in [-\omega,\omega]^n$,

$$\frac{\theta e^{w_i}e^{w_j}}{(e^{w_i}+\theta e^{w_j})^2} \geq \frac{\theta}{(\theta e^{-\omega}+e^{\omega})^2} := c_{\theta,\omega}.$$

This can be noted as follows. Let $z = \theta e^{w_j}/(e^{w_i}+\theta e^{w_j})$. Note that

$$\frac{\theta e^{w_i}e^{w_j}}{(e^{w_i}+\theta e^{w_j})^2} = z(1-z)$$

and that $z \in \Omega := [1/(1+\theta e^{2\omega}), 1/(1+\theta e^{-2\omega})]$. The function $z(1-z)$ is convex and thus achieves its minimum value over the interval Ω at one of its boundary points. It can be readily checked that the minimum is achieved at $z^* = 1/(1+\theta e^{2\omega})$, which yields $z^*(1-z^*) = c_{\theta,\omega}$.

Hence, when $i \neq j$,

$$\frac{\partial^2}{\partial w_i \partial w_j}(-\ell(\mathbf{w})) \leq -c_{\theta,\omega}(\bar{d}_{i,j}+\bar{d}_{j,i}) \leq -c_{\theta,\omega}m_{i,j}.$$

It follows that $\nabla^2(-\ell(\mathbf{w})) \succeq c_{\theta,\omega}\mathbf{L_M}$. From this, we have that for all $\mathbf{w} \in [-\omega,\omega]^n$ and $\mathbf{x} \in \mathcal{X}$

$$\mathbf{x}^\top\nabla^2(-\ell(\mathbf{w}))\mathbf{x} \geq c_{\theta,\omega}\lambda_2(\mathbf{L_M})$$

where $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : ||\mathbf{x}||_\infty \leq \omega \text{ and } \mathbf{x}^\top\mathbf{1} = 0\}$. This implies that $-\ell$ is $c_{\theta,\omega}\lambda_2(\mathbf{L_M})$-strongly convex on $\mathcal{X}$.

C.6 Proof of Lemma C.5

It can be easily shown that for all $\mathbf{w} \in [-\omega,\omega]^n$, $S \subseteq N$ such that $|S| \geq 2$, and $u, v \in S$ such that $u \neq v$, we have

$$\frac{e^{-4\omega}}{|S|^2} \leq \frac{e^{w_u}e^{w_v}}{(\sum_{j\in S}e^{w_j})^2} \leq \frac{e^{4\omega}}{|S|^2}.$$

Combining with (A.1), we have

$$\begin{aligned}
&\frac{\partial^2}{\partial w_u \partial w_v}(-\ell(\mathbf{w})) \\
\leq &-\sum_{y\in T} d_y \frac{w_u w_v}{(\sum_{j=1}^{k} e^{w_{y_j}})^2} 1_{u,v\in\{y_1,y_2,\ldots,y_k\}} \\
\leq &-\frac{e^{-4\omega}}{k^2}\sum_{y\in T} d_\pi 1_{u,v\in\{y_1,y_2,\ldots,y_k\}} \\
= &-\frac{e^{-4\omega}}{k^2}m_{u,v}.
\end{aligned}$$

From this it follows that for all $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{x}^\top\mathbf{1} = 0$,

$$\mathbf{x}^\top\nabla^2(-\ell(\mathbf{w}))\mathbf{x} \geq \frac{e^{-4\omega}}{k^2}\lambda_2(\mathbf{L_M})||\mathbf{x}||^2. \qquad \text{(C.1)}$$

Similarly, we have

$$\begin{aligned}
& \frac{\partial^2}{\partial w_u \partial w_v}(-\ell(\mathbf{w})) \\
\geq & -\sum_{y\in T} d_y \sum_{l=1}^{k-1} \frac{w_u w_v}{(\sum_{j=l}^{k} e^{w_{y_j}})^2} 1_{u,v\in\{y_1,y_2,\ldots,y_k\}} \\
\geq & -e^{4\omega} \sum_{l=1}^{k-1} \frac{1}{(k-l+1)^2}\, m_{u,v} \\
= & -e^{4\omega} \sum_{l=2}^{k} \frac{1}{l^2}\, m_{u,v} \\
\geq & -e^{4\omega} \left(1 + \int_1^k \frac{dx}{x^2}\right) m_{u,v} \\
= & -e^{4\omega} \left(2 - \frac{1}{k}\right) m_{u,v}.
\end{aligned}$$

From this it follows that for all $\mathbf{x}$,

$$\mathbf{x}^\top \nabla^2(-\ell(\mathbf{w}))\mathbf{x} \leq e^{4\omega} \left(2 - \frac{1}{k}\right) \lambda_n(\mathbf{L_M})||\mathbf{x}||^2. \quad \text{(C.2)}$$

D Code and Dataset

The code and datasets for reproducing our experiments are available online:

`https://github.com/GDMMBT/GDMM`.