A On the Convexity of the PL Negative Log-Likelihood

The Hessian of Eq. (3), given by Eq. (48), is not in general positive semidefinite (PSD) (Hunter, 2004). A simple counterexample is as follows: consider $n = 2$ samples and a single observation, i.e., $M = 1$. The Hessian in this case is negative-definite for all $\pi_1, \pi_2 > 0$. Thus, Problem (4) with objective (3) is in general non-convex in $(\boldsymbol{\beta}, b) \in \mathbb{R}^{p+1}$. On the other hand, (3) under parametrization $\pi_i = e^{\theta_i}$, $i \in \mathcal{N}$ is convex as a consequence of the convexity of the log of the sum of exponentials, which is well known (see (Boyd and Vandenberghe, 2004)). The convexity of Problem (5) w.r.t. $\boldsymbol{\beta}$ follows by this observation and also the fact that the composition of convex and affine is convex.

B Proof of Theorem 3.1 (Maystre and Grossglauser, 2015)

We start by showing that $\frac{\partial \mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\pi})}{\partial \pi_i} = 0$, $i \in \mathcal{N}$ is the optimality condition to minimize Eq. (3). Consider the reparametrization $\pi_i = e^{\theta_i}$, $i \in \mathcal{N}$. Eq. (3) under this reparametrization is given by:

$$\mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\theta}) = \sum_{\ell=1}^{M} \left(\log \sum_{j \in A_\ell} e^{\theta_j} - \theta_\ell \right), \tag{21}$$

which is convex w.r.t. $\boldsymbol{\theta} = [\theta_i]_{i \in \mathcal{N}}$, i.e., even though Eq. (3) is not convex w.r.t. $\boldsymbol{\pi}$, it is convex under the reparametrization $\pi_i = e^{\theta_i}$, $i \in \mathcal{N}$. This implies that $\frac{\partial \mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\theta})}{\partial \theta_i} = 0$, $i \in \mathcal{N}$ is the optimality condition to minimize Eq. (21) w.r.t. $\boldsymbol{\theta}$. By the chain rule, this condition can be written in terms of $\pi_i = e^{\theta_i}$, $i \in \mathcal{N}$ as:

$$\frac{\partial \mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\theta})}{\partial \theta_i} = \frac{\partial \mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\pi})}{\partial \pi_i} e^{\theta_i} = 0 \quad \forall i \in \mathcal{N}. \tag{22}$$

Note that $e^{\theta_i} > 0$, $i \in \mathcal{N}$. Then, $\frac{\partial \mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\theta})}{\partial \theta_i} = 0$ is equivalent to $\frac{\partial \mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\pi})}{\partial \pi_i} = 0$, $i \in \mathcal{N}$, i.e., $\pi_i = e^{\theta_i}$, $i \in \mathcal{N}$ satisfies Eq.(22) if and only if θ_i, $i \in \mathcal{N}$ is the minimizer of Eq. (21). Hence, the stationarity condition $\frac{\partial \mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\pi})}{\partial \pi_i} = 0$, $i \in \mathcal{N}$ is also the optimality condition for problem (6).

The optimality condition is given explicitly by:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \pi_i} = \sum_{\ell \in W_i} \left(\frac{1}{\sum_{t \in A_\ell} \pi_t} - \frac{1}{\pi_i} \right) \\ + \sum_{\ell \in L_i} \frac{1}{\sum_{t \in A_\ell} \pi_t} = 0 \quad \forall i \in \mathcal{N}, \end{aligned} \tag{23}$$

where $W_i = \{\ell \,|\, i \in A_\ell, c_\ell = i\}$ is the set of observations where sample $i \in \mathcal{N}$ is chosen and $L_i = \{\ell \,|\, i \in A_\ell, c_\ell \neq i\}$ is the set of observations where sample $i \in \mathcal{N}$ is not chosen. Multiplying both sides of Eq. (23) with π_i, $i \in \mathcal{N}$, we have:

$$\sum_{\ell \in L_i} \left(\frac{\pi_i}{\sum_{t \in A_\ell} \pi_t} \right) - \sum_{\ell \in W_i} \left(\frac{\sum_{j \neq i \in A_\ell} \pi_j}{\sum_{t \in A_\ell} \pi_t} \right) = 0, \tag{24}$$

for all $i \in \mathcal{N}$. Note that $\sum_{\ell \in W_i} \sum_{j \neq i \in A_\ell} \cdot = \sum_{j \neq i} \sum_{\ell \in W_i \cap L_j} \cdot$ and $\sum_{\ell \in L_i} \cdot = \sum_{j \neq i} \sum_{\ell \in W_j \cap L_i} \cdot$. Accordingly, we rewrite Eq. (24) as:

$$\begin{aligned} &\sum_{j \neq i} \sum_{\ell \in W_j \cap L_i} \left(\frac{\pi_i}{\sum_{t \in A_\ell} \pi_t} \right) \\ &= \sum_{j \neq i} \sum_{\ell \in W_i \cap L_j} \left(\frac{\pi_j}{\sum_{t \in A_\ell} \pi_t} \right) \quad \forall i \in \mathcal{N}. \end{aligned} \tag{25}$$

Then, an optimal solution $\boldsymbol{\pi} \in \mathbb{R}_+^n$ to Eq. (6) satisfies:

$$\sum_{j \neq i} \pi_j \lambda_{ji}(\boldsymbol{\pi}) = \sum_{j \neq i} \pi_i \lambda_{ij}(\boldsymbol{\pi}) \quad \forall i \in \mathcal{N}, \tag{26}$$

where $\lambda_{ji}(\boldsymbol{\pi})$, $i, j \in \mathcal{N}, i \neq j$ are given by Eq. (8).

C Alternating Directions Method of Multipliers

We employ Alternating Directions Method of Multipliers (ADMM) to solve the problem in Eq.(11) (Boyd et al., 2011). ADMM is a primal-dual algorithm designed for problems with decoupled objectives, i.e., objectives that can be written as a sum of functions where each function depends on only one of the optimized variables. In our case, we solve Eq.(11) for $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\pi}$, and the objective $\mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\pi})$ is a function of $\boldsymbol{\pi}$ only.

ADMM solves a constrained optimization problem by minimizing the *augmented Lagrangian*, rather than the standard Lagrangian. The difference of augmented Lagrangian from the standard Lagrangian is the additional quadratic penalty on the equality constraint. This additional penalty is shown to greatly improve convergence properties of the algorithm (Boyd et al., 2011). The augmented Lagrangian of Eq. (11) is:

$$\begin{aligned} L_\rho(\tilde{\boldsymbol{\beta}}, \boldsymbol{\pi}, \boldsymbol{y}) = \mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\pi}) \\ + \boldsymbol{y}^T (\tilde{\boldsymbol{X}} \tilde{\boldsymbol{\beta}} - \boldsymbol{\pi}) + \frac{\rho}{2} \| \tilde{\boldsymbol{X}} \tilde{\boldsymbol{\beta}} - \boldsymbol{\pi} \|_2^2, \end{aligned} \tag{27}$$

where $\rho > 0$ is the penalty parameter, $\boldsymbol{y} \in \mathbb{R}^n$ is the dual variable, $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}, b) \in \mathbb{R}^{p+1}$ and $\tilde{\boldsymbol{X}} = [\boldsymbol{X} | \mathbf{1}] \in \mathbb{R}^{n \times (p+1)}$, so that $\boldsymbol{\pi} = \tilde{\boldsymbol{X}} \tilde{\boldsymbol{\beta}}$.

ADMM alternates between optimizing the primal variables $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\pi}$, and the dual variable $\boldsymbol{y}$. Applying

ADMM on problem (11) yields the following iterative algorithm:

$$\tilde{\boldsymbol{\beta}}^{k+1} = \underset{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^{p+1}}{\arg\min}\, \boldsymbol{y}^{k^T}(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} - \boldsymbol{\pi}^k) + \tfrac{\rho}{2}\,\|\,\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} - \boldsymbol{\pi}^k\,\|_2^2$$
$$= (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T(\boldsymbol{\pi}^k - \frac{1}{\rho}\boldsymbol{y}^k), \tag{28a}$$
$$\boldsymbol{\pi}^{k+1} = \underset{\boldsymbol{\pi}\in\mathbb{R}^n_+}{\arg\min}\big(\mathcal{L}(\mathcal{D}\,|\,\boldsymbol{\pi}) + \boldsymbol{y}^{k^T}(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}^{k+1} - \boldsymbol{\pi}) + \frac{\rho}{2}\,\|\,\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}^{k+1} - \boldsymbol{\pi}\,\|_2^2\,\big), \tag{28b}$$
$$\boldsymbol{y}^{k+1} = \boldsymbol{y}^k + \rho(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}^{k+1} - \boldsymbol{\pi}^{k+1}). \tag{28c}$$

For convenience in calculations, the augmented Lagrangian in (27) can be written in a different form, by introducing a scaled dual variable $\boldsymbol{u} = \frac{1}{\rho}\boldsymbol{y}$ and combining the linear and quadratic terms. By doing so, Eq. (27) is equivalent to the final form of the augmented Lagrangian:

$$L_\rho(\tilde{\boldsymbol{\beta}}, \boldsymbol{\pi}, \boldsymbol{u}) = \mathcal{L}(\mathcal{D}\,|\,\boldsymbol{\pi}) + \frac{\rho}{2}\,\|\,\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} - \boldsymbol{\pi} + \boldsymbol{u}\,\|_2^2 - \frac{\rho}{2}\,\|\,\boldsymbol{u}\,\|_2^2\,. \tag{29}$$

Having formed the final augmented Lagrangian in Eq. (29), applying ADMM on problem (11) yields the iterative steps:

$$\tilde{\boldsymbol{\beta}}^{k+1} = \underset{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^{p+1}}{\arg\min}\,\|\,\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} - \boldsymbol{\pi}^k + \boldsymbol{u}^k\,\|_2^2$$
$$= (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T(\boldsymbol{\pi}^k - \boldsymbol{u}^k), \tag{30a}$$
$$\boldsymbol{\pi}^{k+1} = \underset{\boldsymbol{\pi}\in\mathbb{R}^n_+}{\arg\min}\big(\mathcal{L}(\mathcal{D}\,|\,\boldsymbol{\pi}) + \tfrac{\rho}{2}\,\|\,\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}^{k+1} - \boldsymbol{\pi} + \boldsymbol{u}^k\,\|_2^2\,\big), \tag{30b}$$
$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}^{k+1} - \boldsymbol{\pi}^{k+1}. \tag{30c}$$

For convex problems, there are well-established convergence properties for ADMM. If the objective is closed, proper, and convex, and the standard Lagrangian of the problem has a saddle point, then the ADMM iterations are guaranteed to converge to a point where (a) the equality constraint is satisfied, and (b) objective and dual variable attain optimal values. Moreover, in many applications, ADMM has been shown to converge to a modest accuracy in a few tens of iterations (Boyd et al., 2011). For nonconvex problems, there are few convergence analyses for ADMM, which focus on a restricted class of problems (Guo et al., 2017). In general, ADMM is not guaranteed to converge for non-convex problems, and even if it does, it may not converge to the optimal point of the problem. Nevertheless, ADMM is extensively used to also solve nonconvex problems similar to the one we study (Chartrand and Wohlberg, 2013; Guo et al., 2017; Hong, 2018; Wang et al., 2019).

D Proofs

D.1 Proof of Lemma 4.1

At the k-th iteration of ADMM, gradient of the augmented Lagrangian in (29) w.r.t. $\boldsymbol{\pi}$ is:

$$\nabla_{\boldsymbol{\pi}} L_\rho(\tilde{\boldsymbol{\beta}}^{k+1}, \boldsymbol{\pi}, \boldsymbol{u}^k) = \nabla_{\boldsymbol{\pi}}\mathcal{L} + \rho(\boldsymbol{\pi} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}^{k+1} - \boldsymbol{u}^k). \tag{31}$$

To simplify the rest of the calculations, we introduce $\boldsymbol{\sigma} = \rho(\boldsymbol{\pi} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}^{k+1} - \boldsymbol{u}^k) \in \mathbb{R}^n$. Then, the stationarity condition $\frac{\partial L_\rho(\tilde{\boldsymbol{\beta}}^{k+1}, \boldsymbol{\pi}, \boldsymbol{u}^k)}{\partial \pi_i} = 0$, $i \in \mathcal{N}$, is equivalent to:

$$\frac{\partial L_\rho(\tilde{\boldsymbol{\beta}}^{k+1}, \boldsymbol{\pi}, \boldsymbol{u}^k)}{\partial \pi_i} = \frac{\partial \mathcal{L}}{\partial \pi_i} + \sigma_i = 0 \quad \forall i \in \mathcal{N}. \tag{32}$$

Setting $\frac{\partial \mathcal{L}}{\partial \pi_i}$ from Eq. (23) to Eq. (32), we have:

$$\frac{\partial L_\rho(\tilde{\boldsymbol{\beta}}^{k+1}, \boldsymbol{\pi}, \boldsymbol{u}^k)}{\partial \pi_i} = \sum_{\ell\in W_i}\left(-\frac{1}{\pi_i} + \frac{1}{\sum_{t\in A_\ell}\pi_t}\right) + \sum_{\ell\in L_i}\frac{1}{\sum_{t\in A_\ell}\pi_t} + \sigma_i = 0, \tag{33}$$

for all $i \in \mathcal{N}$. Multiplying both sides of Eq. (33) with $-\pi_i$, $i \in \mathcal{N}$, we have:

$$\sum_{\ell\in W_i}\left(\frac{\sum_{j\neq i\in A_\ell}\pi_j}{\sum_{t\in A_\ell}\pi_t}\right) - \sum_{\ell\in L_i}\left(\frac{\pi_i}{\sum_{t\in A_\ell}\pi_t}\right) - \pi_i\sigma_i = 0 \quad \forall i \in \mathcal{N}. \tag{34}$$

Note that $\sum_{\ell\in W_i}\sum_{j\neq i\in A_\ell}\cdot = \sum_{j\neq i}\sum_{\ell\in W_i\cap L_j}\cdot$ and $\sum_{\ell\in L_i}\cdot = \sum_{j\neq i}\sum_{\ell\in W_j\cap L_i}\cdot$. Accordingly, we rewrite Eq. (34) as:

$$\sum_{j\neq i}\sum_{\ell\in W_i\cap L_j}\left(\frac{\pi_j}{\sum_{t\in A_\ell}\pi_t}\right) - \sum_{j\neq i}\sum_{\ell\in W_j\cap L_i}\left(\frac{\pi_i}{\sum_{t\in A_\ell}\pi_t}\right) - \pi_i\sigma_i = 0 \;\forall i \in \mathcal{N}. \tag{35}$$

Then, the stationarity condition $\frac{\partial L_\rho(\tilde{\boldsymbol{\beta}}^{k+1}, \boldsymbol{\pi}, \boldsymbol{u}^k)}{\partial \pi_i} = 0$, $i \in \mathcal{N}$ is equivalent to:

$$\sum_{j\neq i}\pi_j\lambda_{ji}(\boldsymbol{\pi}) - \sum_{j\neq i}\pi_i\lambda_{ij}(\boldsymbol{\pi}) = \pi_i\sigma_i \quad \forall i \in \mathcal{N}, \tag{36}$$

where $\lambda_{ji}(\boldsymbol{\pi})$, $i, j \in \mathcal{N}, i \neq j$ are given by Eq. (8).

D.2 Proof of Theorem 4.2

Summing Eq. (15) for $i \in \mathcal{N}$, we get:

$$\sum_i\sum_j(\pi_j\lambda_{ji}(\boldsymbol{\pi}) - \pi_i\lambda_{ij}(\boldsymbol{\pi}))\,\mathbb{1}_{j\neq i} = \sum_i\pi_i\sigma_i = 0. \tag{37}$$

Since the Plackett-Luce scores are non-negative, i.e. $\pi_i \geq 0$, $i \in \mathcal{N}$, Eq. (37) implies that $\boldsymbol{\sigma} \equiv [\sigma_i]_{i\in\mathcal{N}}$ contains both positive and negative elements. Let $(\mathcal{N}_+, \mathcal{N}_-)$ be a partition of $\mathcal{N}$ such that $\sigma_i \geq 0$ for all $i \in \mathcal{N}_+$ and $\sigma_i < 0$ for all $i \in \mathcal{N}_-$. Then, for $i \in \mathcal{N}_+$ in Eq. (15), we have:

$$\sum_{j\neq i} \pi_j \lambda_{ji}(\boldsymbol{\pi}) = \pi_i \left(\sum_{j\neq i} \lambda_{ij}(\boldsymbol{\pi}) + \sigma_i \right), \ \forall i \in \mathcal{N}_+, \quad (38)$$

where $\lambda_{ij}(\boldsymbol{\pi}) + \sigma_i \geq 0$, $i \in \mathcal{N}_+$ and $j \in \mathcal{N}$. Eq. (38) shows that from each state $i \in \mathcal{N}_+$ into the states in $\mathcal{N}_-$, there exists a total of σ_i "additional outgoing rate", compared to Eq. (7). At the same time, for $i \in \mathcal{N}_-$ in Eq. (15), we have:

$$\begin{aligned} &\sum_{j\in\mathcal{N}_+} \pi_j \lambda_{ji}(\boldsymbol{\pi}) + \sum_{j\in\mathcal{N}_-|j\neq i} \pi_j \lambda_{ji}(\boldsymbol{\pi}) \\ &\quad = \pi_i \sum_{j\neq i} \lambda_{ij}(\boldsymbol{\pi}) + \pi_i \sigma_i, \quad \forall i \in \mathcal{N}_-. \end{aligned} \quad (39)$$

Since $\pi_i \sigma_i < 0$, for $i \in \mathcal{N}_-$, we distribute these terms into the first sum on the left hand side. Then, Eq. (39) is equivalent to:

$$\begin{aligned} &\sum_{j\in\mathcal{N}_+} \pi_j \left(\lambda_{ji}(\boldsymbol{\pi}) - \frac{\pi_i \sigma_i c_j}{\pi_j} \right) + \sum_{j\in\mathcal{N}_-|j\neq i} \pi_j \lambda_{ji}(\boldsymbol{\pi}) \\ &= \pi_i \sum_{j\neq i} \lambda_{ij}(\boldsymbol{\pi}), \quad \forall i \in \mathcal{N}_-, \end{aligned} \quad (40)$$

where $\sum_{j\in\mathcal{N}_+} c_j = 1$.

To determine the $\{c_j\}_{j\in\mathcal{N}_+}$, recall from Eq. (38) that from each state $j \in \mathcal{N}_+$ into the states $i \in \mathcal{N}_-$, there exists a total of σ_j additional outgoing rate. Then, Eq. (40) implies that $\sum_{i\in\mathcal{N}_-} -\frac{\pi_i \sigma_i c_j}{\pi_j} = \sigma_j$, i.e., $c_j = \frac{-\pi_j \sigma_j}{\sum_{i\in\mathcal{N}_-} \pi_i \sigma_i}$, $j \in \mathcal{N}_+$. Using $-\sum_{i\in\mathcal{N}_-} \pi_i \sigma_i = \sum_{i\in\mathcal{N}_+} \pi_i \sigma_i$ from Eq. (37), we confirm that $\sum_{j\in\mathcal{N}_+} c_j = \frac{-\sum_{j\in\mathcal{N}_+} \pi_j \sigma_j}{\sum_{i\in\mathcal{N}_-} \pi_i \sigma_i} = 1$, and rewrite $\{c_j\}_{j\in\mathcal{N}_+}$ as:

$$c_j = \frac{-2\pi_j \sigma_j}{\sum_{i\in\mathcal{N}_-} \pi_i \sigma_i - \sum_{i\in\mathcal{N}_+} \pi_i \sigma_i}, \quad \forall j \in \mathcal{N}_+. \quad (41)$$

Finally, setting $\{c_j\}_{j\in\mathcal{N}_+}$ into Eq. (40), we have:

$$\begin{aligned} &\sum_{j\in\mathcal{N}_+} \pi_j \left(\lambda_{ji}(\boldsymbol{\pi}) + \frac{2\pi_i \sigma_i \sigma_j}{\sum_{t\in\mathcal{N}_-} \pi_t \sigma_t - \sum_{t\in\mathcal{N}_+} \pi_t \sigma_t} \right) \\ &+ \sum_{j\in\mathcal{N}_-|j\neq i} \pi_j \lambda_{ji}(\boldsymbol{\pi}) = \pi_i \sum_{j\neq i} \lambda_{ij}(\boldsymbol{\pi}), \quad \forall i \in \mathcal{N}_-, \end{aligned} \quad (42)$$

where $\lambda_{ji}(\boldsymbol{\pi}) + \frac{2\pi_i \sigma_i \sigma_j}{\sum_{t\in\mathcal{N}_-} \pi_t \sigma_t - \sum_{t\in\mathcal{N}_+} \pi_t \sigma_t} \geq 0$, $j \in \mathcal{N}_+$ and $i \in \mathcal{N}_-$.

Eq. (15), partitioned as Eq. (38) and Eq. (42), is the balance equations of a continuous-time MC with transition rates given by:

$$\mu_{ji}(\boldsymbol{\pi}) = \begin{cases} \lambda_{ji}(\boldsymbol{\pi}) + \frac{2\pi_i \sigma_i \sigma_j}{\sum_{t\in\mathcal{N}_-} \pi_t \sigma_t - \sum_{t\in\mathcal{N}_+} \pi_t \sigma_t} \\ \qquad \text{if } j \in \mathcal{N}_+ \text{ and } i \in \mathcal{N}_- \\ \lambda_{ji}(\boldsymbol{\pi}) \quad \text{otherwise.} \end{cases} \quad (43)$$

Hence, $\boldsymbol{\pi}$ is the stationary distribution of this MC (Gallager, 2013).

D.3 Proof of Theorem 4.3

We use the following definition.

Definition D.1 (Diagonal dominance)**.** *A matrix $\boldsymbol{H}$ is diagonally dominant if $|\boldsymbol{H}_{ii}| \geq \sum_{j\neq i} |\boldsymbol{H}_{ij}|$, $i \in \mathcal{N}$, i.e., for every row, magnitude of the diagonal element is larger than the sum of magnitudes of all off-diagonal elements (Horn and Johnson, 2012).*

Eq. (33) is equivalent to:

$$\begin{aligned} \frac{\partial L_\rho(\tilde{\boldsymbol{\beta}}^{k+1}, \boldsymbol{\pi}, \boldsymbol{u}^k)}{\partial \pi_i} =& \sum_{\ell\in W_i} -\frac{1}{\pi_i} + \sum_{\ell|i\in A_\ell} \frac{1}{\sum_{t\in A_\ell} \pi_t} \\ &+ \rho(\boldsymbol{\pi} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}^{k+1} - \boldsymbol{u}^k)_i \ , \end{aligned} \quad (44)$$

for all $i \in \mathcal{N}$. At the k-th iteration of (13), let $\nabla^2 L_\rho(\tilde{\boldsymbol{\beta}}^{k+1}, \boldsymbol{\pi}, \boldsymbol{u}^k)$ be the Hessian of the augmented Lagrangian w.r.t. $\boldsymbol{\pi}$. Differentiating Eq. (44) w.r.t. π_j, $\nabla^2 L_\rho(\tilde{\boldsymbol{\beta}}^{k+1}, \boldsymbol{\pi}, \boldsymbol{u}^k)$ has the following form:

$$\begin{aligned} &\nabla^2_{ij} L_\rho(\tilde{\boldsymbol{\beta}}^{k+1}, \boldsymbol{\pi}, \boldsymbol{u}^k) \\ &= \begin{cases} \sum_{\ell\in W_i} \frac{1}{\pi_i^2} - \sum_{\ell|i\in A_\ell} \frac{1}{(\sum_{t\in A_\ell} \pi_t)^2} + \rho, & i = j \\ -\sum_{\ell|i,j\in A_\ell} \frac{1}{(\sum_{t\in A_\ell} \pi_t)^2}, & i \neq j. \end{cases} \end{aligned} \quad (45)$$

Consider $\rho \geq \frac{2}{\epsilon^2} \max_i \sum_{\ell|i\in A_\ell} \frac{1}{|A_\ell|^2}$. By Assumption 4.1, we have:

$$\begin{aligned} &\rho \geq \frac{2}{\epsilon^2} \sum_{\ell|i\in A_\ell} \frac{1}{|A_\ell|^2} \ \forall i \in \mathcal{N}, \\ &\Leftrightarrow \ \rho \geq \sum_{\ell|i\in A_\ell} \frac{2}{\left(\sum_{t\in A_\ell} \pi_t\right)^2} \ \forall i \in \mathcal{N}, \\ &\Leftrightarrow \ \rho + \sum_{\ell\in W_i} \frac{1}{\pi_i^2} > \sum_{\ell|i\in A_\ell} \frac{2}{\left(\sum_{t\in A_\ell} \pi_t\right)^2} \ \forall i \in \mathcal{N}, \quad (46a) \\ &\Leftrightarrow \ \rho + \sum_{\ell\in W_i} \frac{1}{\pi_i^2} > \sum_{\ell|i\in A_\ell} \frac{1}{\left(\sum_{t\in A_\ell} \pi_t\right)^2} \\ &\qquad + \sum_{j\neq i} \sum_{\ell|i,j\in A_\ell} \frac{1}{\left(\sum_{t\in A_\ell} \pi_t\right)^2} \ \forall i \in \mathcal{N}. \quad (46b) \end{aligned}$$

Eq. (46a) implies that all diagonal elements of $\nabla^2 L_\rho(\tilde{\boldsymbol{\beta}}^{k+1}, \boldsymbol{\pi}, \boldsymbol{u}^k)$ are positive. Also, by Eq. (46b),

$\nabla^2 L_\rho(\tilde{\boldsymbol{\beta}}^{k+1}, \boldsymbol{\pi}, \boldsymbol{u}^k)$ is diagonally dominant (c.f. Definition D.1). Thus, $\nabla^2 L_\rho(\tilde{\boldsymbol{\beta}}^{k+1}, \boldsymbol{\pi}, \boldsymbol{u}^k)$ is positive definite (Horn and Johnson, 2012), i.e., $L_\rho(\tilde{\boldsymbol{\beta}}^{k+1}, \boldsymbol{\pi}, \boldsymbol{u}^k)$ is convex w.r.t. $\boldsymbol{\pi}$. As a result, under Assumption 4.1, for $\rho \geq \frac{2}{\epsilon^2} \max_i \sum_{\ell | i \in A_\ell} \frac{1}{|A_\ell|^2}$, a stationary $\boldsymbol{\pi} > \mathbf{0}$ satisfying condition (14) is also a minimizer of step (13b).

D.4 Proof of Theorem 4.4

We make use of the following lemmas.

Lemma D.1 (Zeng et al. (2018)). *Logarithm and polynomials are Kurdyka–Łojasiewicz (KL) functions. Moreover, sums, products, compositions, and quotients (with denominator bounded away from 0) of KL functions are also KL.*

Lemma D.2 (Guo et al. (2017)). *Consider the optimization problem:*

$$\begin{aligned} &\underset{\tilde{\boldsymbol{\beta}},\, \boldsymbol{\pi}}{\text{minimize}} && g(\boldsymbol{\pi}) \\ &\text{subject to} && \tilde{\boldsymbol{X}} \tilde{\boldsymbol{\beta}} = \boldsymbol{\pi}, \end{aligned} \tag{47}$$

and solve Eq. (47) *via Alternating Direction Method of Multipliers (ADMM) (Boyd et al., 2011). Let* $\{(\boldsymbol{\pi}^k, \boldsymbol{u}^k, \tilde{\boldsymbol{\beta}}^k)\}_{k \in \mathbb{N}}$ *be the sequence generated by the ADMM algorithm, and* ρ *be the penalty parameter of ADMM. Suppose that there exists* $\kappa > 0$ *such that* $\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} \succeq \kappa \boldsymbol{I}$*, and the sequence* $\{(\boldsymbol{\pi}^k, \boldsymbol{u}^k, \tilde{\boldsymbol{\beta}}^k)\}_{k \in \mathbb{N}}$ *is bounded.*

If there exist solutions for the minimization steps of ADMM w.r.t. both $\boldsymbol{\pi}$ *and* $\tilde{\boldsymbol{\beta}}$*,* $g(\boldsymbol{\pi})$ *is a continuous differentiable function with an L-Lipschitz continuous gradient at* $\boldsymbol{\pi}^k$*,* $k \in \mathbb{N}$ *where* $L > 0$*, and the augmented Lagrangian of Eq.* (47) *is a KL function, then, for* $\rho > 2L$*,* $\{(\boldsymbol{\pi}^k, \boldsymbol{u}^k, \tilde{\boldsymbol{\beta}}^k)\}_{k \in \mathbb{N}}$ *converges to a point that satisfies the Karush-Kuhn-Tucker (KKT) conditions of Eq.* (47)*.*

To begin with, there exist solutions for the minimization steps in (13): $\tilde{\boldsymbol{\beta}}$ update has the closed form solution given by Eq. (13a) and $\boldsymbol{\pi}$ update admits a minimizer for large enough ρ by Lemma 4.3.

By Assumption 4.1, $\nabla_{\boldsymbol{\pi}} \mathcal{L}$ given by Eq. (23) exists, i.e. $\mathcal{L}$ is continuous differentiable at $\boldsymbol{\pi}^k$, $k \in \mathbb{N}$ generated by (13b). Let $\nabla^2(\mathcal{L})$ be the Hessian of $\mathcal{L}$. Differentiating Eq. (23) w.r.t. π_j, $\nabla^2(\mathcal{L})$ has the following form:

$$\begin{aligned} &\nabla^2_{ij}(\mathcal{L}) \\ &= \begin{cases} \sum_{\ell \in W_i} \frac{1}{\pi_i^2} - \sum_{\ell | i \in A_\ell} \frac{1}{(\sum_{t \in A_\ell} \pi_t)^2}, & i = j \\ - \sum_{\ell | i,j \in A_\ell} \frac{1}{(\sum_{t \in A_\ell} \pi_t)^2}, & i \neq j. \end{cases} \end{aligned} \tag{48}$$

Consider $L = \frac{\max_i |W_i|}{\epsilon^2}$, where W_i is the set of observations where sample $i \in \mathcal{N}$ is chosen. By Assumption 4.1, we have:

$$\begin{aligned} L = \frac{\max_i |W_i|}{\epsilon^2} &\geq \sum_{\ell \in W_i} \frac{1}{\pi_i^2} \;\; \forall i \in \mathcal{N}, \\ \Leftrightarrow \; L - \sum_{\ell \in W_i} \frac{1}{\pi_i^2} + \sum_{\ell | i \in A_\ell} \frac{1}{\left(\sum_{t \in A_\ell} \pi_t\right)^2} &\geq \sum_{\ell | i \in A_\ell} \frac{1}{\left(\sum_{t \in A_\ell} \pi_t\right)^2} \;\; \forall i \in \mathcal{N}, \\ \Leftrightarrow \; L - \sum_{\ell \in W_i} \frac{1}{\pi_i^2} + \sum_{\ell | i \in A_\ell} \frac{1}{\left(\sum_{t \in A_\ell} \pi_t\right)^2} &\geq \sum_{j \neq i} \sum_{\ell | i,j \in A_\ell} \frac{1}{\left(\sum_{t \in A_\ell} \pi_t\right)^2} \;\; \forall i \in \mathcal{N}. \end{aligned} \tag{49}$$

Now, consider the matrix $L\boldsymbol{I}_{n \times n} - \nabla^2(\mathcal{L})$. By Eq. (49), $L\boldsymbol{I}_{n \times n} - \nabla^2(\mathcal{L})$ is diagonally dominant (c.f. Definition D.1) and all of its diagonal elements are positive, i.e., $\nabla^2(\mathcal{L})$ is upper bounded by $L\boldsymbol{I}_{n \times n}$. Thus, the objective function of Eq. (11), i.e. $\mathcal{L}$, has an L-Lipschitz continuous gradient at $\boldsymbol{\pi}^k$, $k \in \mathbb{N}$, where $L = \frac{\max_i |W_i|}{\epsilon^2} > 0$.

Moreover, the augmented Lagrangian given by Eq. (29) is a sum of three functions: logarithm of the ratio of two polynomials where the denominator is bounded away from 0 for all $\boldsymbol{\pi}^k$, $k \in \mathbb{N}$ by Assumption 4.1, and two other polynomial functions. By Lemma D.1, these three functions and their sum is KL on the set $\{\boldsymbol{\pi}^k \,|\, \pi_i^k > \epsilon \,, i \in \mathcal{N}, k \in \mathbb{N}\}$. As a result, the augmented Lagrangian of Eq. (11) is a KL function. Putting it all together, by Lemma D.2, for $\rho > \frac{2 \max_i |W_i|}{\epsilon^2}$, the sequence $\{(\boldsymbol{\pi}^k, \boldsymbol{u}^k, \tilde{\boldsymbol{\beta}}^{k+1})\}_{k \in \mathbb{N}}$ generated by (13) converges to a point that satisfies the KKT conditions (Nocedal and Wright, 2006) of Problem (11).

E Extension to the Logistic Case

We describe here how to apply our approach to regress model parameters in the logistic case. Recall that Problem (5) is, in this case, convex, and can thus be solved by Newton's method. Nevertheless, we would like to accelerate its computation via a spectral method akin to ILSR. Following the steps we took in the affine case, we re-write (5) as:

$$\text{Minimize} \quad \mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\pi}) \tag{50a}$$

$$\text{subject to:} \quad \log \boldsymbol{\pi} = \boldsymbol{X}\boldsymbol{\beta}, \quad \boldsymbol{\pi} \geq \mathbf{0}, \tag{50b}$$

where $\log \boldsymbol{\pi} = [\log \pi_i]_{i \in \mathcal{N}}$ is the $\mathbb{R}^n$ vector generated by applying log to $\boldsymbol{\pi}$ element-wise. The augmented Lagrangian corresponding to Eq. (50) is:

$$\begin{aligned} L_\rho(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{u}) = \mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\pi}) \\ + \frac{\rho}{2} \left\| \boldsymbol{X}\boldsymbol{\beta} - \log \boldsymbol{\pi} + \boldsymbol{u} \right\|_2^2 - \frac{\rho}{2} \left\| \boldsymbol{u} \right\|_2^2, \end{aligned} \tag{51}$$

Algorithm 2 PLADMM-log

procedure ADMM($\boldsymbol{X}$, $\mathcal{D} = \{(c_\ell, A_\ell) \,|\, \ell \in \mathcal{M}\}$, ρ)
  Initialize $\boldsymbol{\beta}$ via Eq. (55); $\boldsymbol{\pi} \leftarrow [e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}]_{i \in \mathcal{N}}$; $\boldsymbol{u} \leftarrow \mathbf{0}$
  repeat
    $\boldsymbol{\pi} \leftarrow \text{ILSRX}(\rho, \boldsymbol{\pi}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{u})$
    $\boldsymbol{u} \leftarrow \boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta} - \log \boldsymbol{\pi}$
    $\boldsymbol{\beta} \leftarrow (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T (\log \boldsymbol{\pi} - \boldsymbol{u})$
  until convergence
return $\boldsymbol{\beta}$, $\boldsymbol{\pi}$
end procedure
procedure ILSRX($\rho, \boldsymbol{\pi}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{u}$)
  repeat
    $\sigma_i \leftarrow \rho \frac{(\log \pi_i - \boldsymbol{x}_i^T \boldsymbol{\beta} - u_i)}{\pi_i}, \quad i \in \mathcal{N}$
    Calculate $\boldsymbol{M}(\boldsymbol{\pi}) = [\mu_{ji}(\boldsymbol{\pi})]_{i,j \in \mathcal{N}}$ via Eq. (16)
    $\boldsymbol{\pi} \leftarrow \text{ssd}\,(\boldsymbol{M}(\boldsymbol{\pi}))$
  until convergence
return $\boldsymbol{\pi}$
end procedure

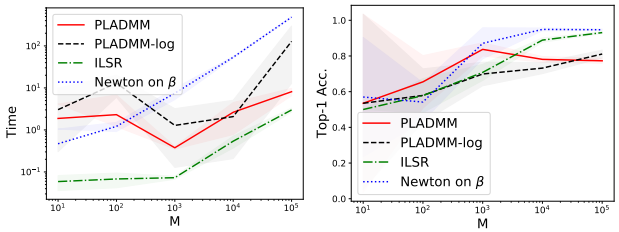

Figure 2: Convergence time (Conv. Time) and top-1 test accuracy (Top-1 Acc.) of PLADMM, PLADMM-log, ILSR, and Newton on $\boldsymbol{\beta}$ evaluated on synthetic datasets vs. the number of observations $M \in \{10, 100, 1000, 10000, 100000\}$. Observations are partitioned w.r.t. observation CV (c.f. Sec. 5), where number of samples is $n = 1000$, number of features is $p = 100$, and query size is $|A_\ell| = 2$.

and applying ADMM on problem (50) yields:

$$\boldsymbol{\beta}^{k+1} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min}\, L_\rho(\boldsymbol{\beta}, \boldsymbol{\pi}^k, \boldsymbol{u}^k) = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T (\log \boldsymbol{\pi}^k - \boldsymbol{u}^k), \tag{52a}$$

$$\boldsymbol{\pi}^{k+1} = \underset{\boldsymbol{\pi} \in \mathbb{R}^n_+}{\arg\min}\, \mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\pi}) + \frac{\rho}{2} \| \boldsymbol{X}\boldsymbol{\beta}^{k+1} - \log \boldsymbol{\pi} + \boldsymbol{u}^k \|_2^2, \tag{52b}$$

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \boldsymbol{X}\boldsymbol{\beta}^{k+1} - \log \boldsymbol{\pi}^{k+1}. \tag{52c}$$

Mutatis mutandis, following the same manipulations in Lemma 4.1, a stationary point of the objective in each step (52b) can be cast as the stationary distribution of the continuous-time MC with transition rates $\mu_{ji}(\boldsymbol{\pi})$, $i, j \in \mathcal{N}$, given by Eq. (16), the only difference being that vector $\boldsymbol{\sigma} = [\sigma_i]_{i \in \mathcal{N}}$ is now given by:

$$\sigma_i = \rho \frac{(\log \pi_i - \boldsymbol{x}_i^T \boldsymbol{\beta} - u_i)}{\pi_i}, \quad i \in \mathcal{N}. \tag{53}$$

Having adjusted the transition matrix $\boldsymbol{M}(\boldsymbol{\pi})$ thusly, $\boldsymbol{\pi}$ can again be obtained by repeated iterations of (18).

The resulting algorithm, which we refer to as Plackett-Luce ADMM-log (PLADMM-log), is summarized in Algorithm 2; the algorithm is almost identical to Algorithm 1, using $\log \boldsymbol{\pi}$ instead of $\boldsymbol{\pi}$, defining $\boldsymbol{\sigma}$ via (53), and having a different initialization. We discuss the latter below.

Initialization. Similar to the initialization of PLADMM (c.f. Eq. (20)), we initialize $\boldsymbol{\beta}$ so that the initial scores obey the Plackett-Luce model, mirroring the approach by Saha and Rajkumar (2018). Defining P_{ij}, $i, j \in \mathcal{N}$ the same way, and using the logistic parametrization in Sec.3, we have that:

$$\frac{P_{ij}}{P_{ji}} = \frac{\pi_i}{\pi_j} = e^{\boldsymbol{\beta}^T (\boldsymbol{x}_i - \boldsymbol{x}_j)}. \tag{54}$$

Accordingly, we initialize $\boldsymbol{\beta}$ as:

$$\boldsymbol{\beta}^0 = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \sum_{(i,j) \in \mathcal{D}} \left(\boldsymbol{\beta}^T (\boldsymbol{x}_i - \boldsymbol{x}_j) - \log \left(\frac{\hat{P}_{ij}}{\hat{P}_{ji}} \right) \right)^2, \tag{55}$$

where $\hat{P}_{ij}$, $i, j \in \mathcal{N}$, are again empirical estimates obtained from dataset $\mathcal{D}$. Given $\boldsymbol{\beta}^0$, we generate the initial Plackett-Luce scores via the logistic parametrization $\boldsymbol{\pi}^0 = [e^{\boldsymbol{x}_i^T \boldsymbol{\beta}^0}]_{i \in \mathcal{N}}$. Finally, we initialize the dual variable as $\boldsymbol{u}^0 = \mathbf{0}$.

F Experiments

F.1 Datasets

Synthetic Datasets. We generate the feature vectors $\boldsymbol{x}_i \in \mathbb{R}^p$, $i \in \mathcal{N}$ from $\mathcal{N}(\mathbf{0}, \sigma_x^2 \boldsymbol{I}_{p \times p})$ and a common parameter vector $\boldsymbol{\beta} \in \mathbb{R}^p$ from $\mathcal{N}(\mathbf{0}, \sigma_\beta^2 \boldsymbol{I}_{p \times p})$. Then, we generate the Plackett-Luce scores via the logistic parametrization $\boldsymbol{\pi} = [e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}]_{i \in \mathcal{N}}$. We normalize the resulting scores, so that $\mathbf{1}^\top \boldsymbol{\pi} = 1$. We set $\sigma_x^2 = \sigma_\beta^2 = 0.8$ in all experiments. Given $\boldsymbol{\pi}$, we generate each observation in $\mathcal{D}$ as follows: we first select $|A_\ell| = 2$ samples out of n samples uniformly at random. Then, we generate the choice c_l, $l \in \mathcal{M}$ from the Plackett-Luce model given by Eq. (1).

Filter Aesthetic Comparison (FAC). The Filter Aesthetic Comparison (FAC) dataset (Sun et al., 2017) contains 1280 unfiltered images pertaining to 8 different categories. Twenty-two different image filters are applied to each image. Labelers are provided with two filtered images and are asked to identify which image has better quality. We select $n = 1000$ images within one category, as only the filtered image pairs that are within the same category are compared. The resulting dataset contains $M = 728$ pairwise comparisons. Moreover, for each image, we extract features via a

state-of-the-art convolutional neural network architecture, namely GoogLeNet (Szegedy et al., 2015), with weights pre-trained on the ImageNet dataset (Deng et al., 2009). We select $p = 50$ of these features by Principal Component Analysis (Jolliffe, 1986).

Retinopathy of Prematurity (ROP). The Retinopathy of Prematurity (ROP) dataset contains $n = 100$ retina images with $p = 143$ features (Ataer-Cansızoğlu, 2015). Experts are provided with two images and are asked to choose the image with higher severity of the ROP disease. Five experts independently label 5941 image pairs; the resulting dataset contains $M = 29705$ pairwise comparisons. Note that some pairs are labelled more than once by different experts.

SUSHI. The SUSHI Preference dataset (Kamishima et al., 2009) contains $n = 100$ sushi ingredients with $p = 18$ features. Each of the 5000 customers independently ranks 10 ingredients according to her preferences. We select the rankings provided by 10 customers, where an ingredient is ranked higher if it precedes the other ingredients in a customer's ranked list. We generate two datasets: triplet Sushi containing $M = 1200$ rankings of $|A_\ell| = 3$ ingredients, and pairwise Sushi containing $M = 450$ pairwise comparisons.

F.2 Algorithms

We implement four algorithms that regress Plackett-Luce scores from features, which we call as *feature methods*.

PLADMM. PLADMM solves the problem in Eq. (11) and is summarized in Algorithm 1. We compute the stationary distribution at each iteration of ILSRX (c.f. Eq. (18)) using the power method (Lei et al., 2016). As the stopping criterion, we use $\|\boldsymbol{\pi}^k - \boldsymbol{\pi}^{k-1}\|_2 < r_{\texttt{tol}} \| \boldsymbol{\pi}^k\|_2$ and $\|\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}^k - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}^{k-1}\|_2 < r_{\texttt{tol}} \|\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}^k\|_2$. We set the relative tolerance $r_{\texttt{tol}} = 10^{-4}$ for all experiments. We use the same relative tolerance for the stopping criterion of the power method. We set $\rho = 1$ in our experiments, which is a standard choice in the ADMM literature (Boyd et al., 2011). In our experiments, we consistently observe that Eq.(37) is satisfied. That is why, we use $c_j = \frac{-\pi_j \sigma_j}{\sum_{i \in \mathcal{N}_-} \pi_i \sigma_i}$, $j \in \mathcal{N}_+$ instead of Eq.(41) to calculate the transition rates (16).

PLADMM-log. PLADMM-log solves the problem in Eq. (50) and is summarized in Algorithm 2. As the stopping criterion, we use $\|\boldsymbol{\pi}^k - \boldsymbol{\pi}^{k-1}\|_2 < r_{\texttt{tol}} \| \boldsymbol{\pi}^k\|_2$ and $\|e^{\boldsymbol{X}\boldsymbol{\beta}^k} - e^{\boldsymbol{X}\boldsymbol{\beta}^{k-1}}\|_2 < r_{\texttt{tol}} \|e^{\boldsymbol{X}\boldsymbol{\beta}^k}\|_2$, where exponentiation is applied elementwise.

SLSQP. SLSQP solves the problem in Eq. (4) via the sequential least-squares quadratic programming (SLSQP) algorithm (Nocedal and Wright, 2006). We initialize SLSQP the same as PLADMM (c.f. Algorithm 1). As stopping criterion, we use $\| \boldsymbol{\pi}^k - \boldsymbol{\pi}^{k-1} \|_2 < r_{\texttt{tol}} \| \boldsymbol{\pi}^k \|_2$, where $\boldsymbol{\pi}^k = \boldsymbol{X}\boldsymbol{\beta}^k + b^k\mathbf{1}$, $k \in \mathbb{N}$. Each iteration of SLSQP is $O\Big(\sum_{\ell \in \mathcal{D}} \big(|A_\ell|(p+1)\big) + (p+1)^2\Big)$ for constructing the gradient of Eq. (3) w.r.t. $\tilde{\boldsymbol{\beta}}$ and updating $\tilde{\boldsymbol{\beta}}$, respectively.

Newton on $\boldsymbol{\beta}$. Newton on $\boldsymbol{\beta}$ solves the convex problem in Eq. (5) via Newton's method (Nocedal and Wright, 2006). We initialize Newton on $\boldsymbol{\beta}$ the same as PLADMM-log (c.f. Algorithm 2). As stopping criterion, we use $\| \boldsymbol{\pi}^k - \boldsymbol{\pi}^{k-1} \|_2 < \ r_{\texttt{tol}} \ \| \boldsymbol{\pi}^k \|_2$, where $\boldsymbol{\pi}^k = [e^{\boldsymbol{x}_i^T \boldsymbol{\beta}^k}]_{i \in \mathcal{N}}$, $k \in \mathbb{N}$. Each iteration of Newton on $\boldsymbol{\beta}$ is $O\Big(\sum_{\ell \in \mathcal{D}} \big(|A_\ell|\, p^2\big) + p^2\Big)$ for constructing the Hessian of Eq. (3) w.r.t. $\boldsymbol{\beta}$ and updating $\boldsymbol{\beta}$, respectively.

We implement three algorithms that learn the Plackett-Luce scores from the choice observations alone, which we call as *featureless methods*.

ILSR. Iterative Luce Spectral Ranking (ILSR) algorithm solves the problem in Eq. (6) and is described by the iterations in Eq.(10). We initialize ILSR with $\boldsymbol{\pi}^0 = \frac{1}{n}\mathbf{1}$. We compute the stationary distribution at each iteration of ILSR using the power method. As the stopping criterion, we use $\|\boldsymbol{\pi}^k - \boldsymbol{\pi}^{k-1}\|_2 < \ r_{\texttt{tol}} \ \|\boldsymbol{\pi}^k\|_2$. Each iteration of ILSR is $O\Big(\sum_{\ell \in \mathcal{D}} \big(|A_\ell|\big) + n^2\Big)$ for constructing the transition matrix $\boldsymbol{\Lambda}(\boldsymbol{\pi})$ (c.f. Eq.(9)) and finding the stationary distribution $\boldsymbol{\pi}$, respectively.

MM. The Minorization-Maximization (MM) algorithm (Hunter, 2004) solves the problem in Eq. (6). We initialize MM with $\boldsymbol{\pi}^0 = \frac{1}{n}\mathbf{1}$. As the stopping criterion, we use $\|\boldsymbol{\pi}^k - \boldsymbol{\pi}^{k-1}\|_2 < \ r_{\texttt{tol}} \ \|\boldsymbol{\pi}^k\|_2$. Each iteration of MM is $O\Big(\sum_{\ell \in \mathcal{D}} \big(|A_\ell|\big)\Big)$.

Newton on $\boldsymbol{\theta}$. Newton on $\boldsymbol{\theta}$ algorithm solves the problem in Eq. (6) by reparametrizing the scores as $\pi_i = e^{\theta_i}$, $i \in \mathcal{N}$. It solves the resulting convex problem by Newton's method (Nocedal and Wright, 2006). We initialize Newton on $\boldsymbol{\theta}$ with $\boldsymbol{\theta}^0 = [\theta_i^0]_{i \in \mathcal{N}} = \mathbf{0}$. As stopping criterion, we use $\|\boldsymbol{\pi}^k - \boldsymbol{\pi}^{k-1}\|_2 < r_{\texttt{tol}} \|\boldsymbol{\pi}^k\|_2$, where $\pi_i^k = e^{\theta_i^k}$, $i \in \mathcal{N}$, $k \in \mathbb{N}$. Each iteration of Newton on $\boldsymbol{\theta}$ is $O\Big(\sum_{\ell \in \mathcal{D}} \big(|A_\ell|^2\big) + n^2\Big)$ for constructing the Hessian of Eq. (3) w.r.t. $\boldsymbol{\theta}$ and updating $\boldsymbol{\theta}$, respectively.

F.3 Top-1 Accuracy and Kendall-Tau Correlation

We measure the prediction performance by Top-1 accuracy (Top-1 Acc.) and Kendall-Tau correlation (KT) on the test set. Let the test set be $\mathcal{D}_{\texttt{choice}} = \{(c_\ell, A_\ell) \,|\, \ell \in \{1, ..., M_{\texttt{test}}\}\}$ for the choice setting and $\mathcal{D}_{\texttt{rank}} = \{(\alpha^\ell, A_\ell) \,|\, \ell \in \{1, ..., M_{\texttt{test}}\}\}$ for the rank-

ing setting, where $\alpha^\ell = \alpha^\ell_1 \succ \alpha^\ell_2 \succ \cdots \succ \alpha^\ell_{|A_\ell|}$ is an ordered sequence of the samples in A_ℓ. For both settings, given A_ℓ, we predict the ℓ-th choice as $\hat{c}_\ell = \arg\max_{i \in A_\ell} \pi_i$. We calculate the Top-1 accuracy (Top-1 Acc.) for the choice setting as:

$$\text{Top-1 Acc.} = \frac{\sum_{\ell=1}^{M_{\texttt{test}}} \mathbb{1}(\hat{c}_\ell = c_\ell)}{M_{\texttt{test}}} \in [0, 1], \tag{56}$$

and for the ranking setting as:

$$\text{Top-1 Acc.} = \frac{\sum_{\ell=1}^{M_{\texttt{test}}} \mathbb{1}(\hat{c}_\ell = \alpha^\ell_1)}{M_{\texttt{test}}} \in [0, 1]. \tag{57}$$

For the ranking setting, given A_ℓ, we also predict the ranking as $\hat{\alpha}^\ell = \arg\text{sort}[\pi_i]_{i \in A_\ell}$, i.e. sequence of the samples in A_ℓ ordered w.r.t. their scores. We calculate Kendall-tau correlation (KT) (Kendall, 1938) as a measure of the correlation between each true ranking α^ℓ and predicted ranking $\hat{\alpha}^\ell$, $\ell \in \{1, ..., M_{\texttt{test}}\}$. For observation ℓ, let $T_\ell = \sum_{t=1}^{|A_\ell|} \sum_{s=1}^{|A_\ell|} \mathbb{1}(\hat{\alpha}^\ell_t \succ \hat{\alpha}^\ell_s \wedge \alpha^\ell_t \succ \alpha^\ell_s)$ be the number correctly predicted ranking positions, and $F_\ell = \sum_{t=1}^{|A_\ell|} \sum_{s=1}^{|A_\ell|} \mathbb{1}(\hat{\alpha}^\ell_t \succ \hat{\alpha}^\ell_s \wedge \alpha^\ell_s \succ \alpha^\ell_t)$ be the number incorrectly predicted ranking positions. Then, KT is computed by:

$$\text{KT} = \frac{\sum_{\ell=1}^{M_{\texttt{test}}} (T_\ell - F_\ell) / \binom{|A_\ell|}{2}}{M_{\texttt{test}}} \in [-1, 1], \tag{58}$$

where $\binom{|A_\ell|}{2}$ is the number of sample pairs in a query of size $|A_\ell|$.

F.4 Impact of Number of Observations

Fig. 2 shows the convergence time (Time) and top-1 test accuracy (Top-1 Acc.) of PLADMM, PLADMM-log, ILSR, and Newton on $\boldsymbol{\beta}$ when trained on synthetic datasets with number of observations $M \in \{10, 100, 1000, 10000, 100000\}$. Observations are partitioned w.r.t. observation CV (c.f. Sec. 5), where number of samples is $n = 1000$, number of parameters is $p = 100$, and size of each query is $|A_\ell| = 2$. As $n > p$, PLADMM benefits from being able to regress n scores from a smaller number of p parameters and leads to significantly better Top-1 Acc compared to ILSR in Fig. 2. Especially when M is not enough to learn $n = 1000$ scores, but to learn $p = 100$ parameters, PLADMM gains the most performance advantage over ILSR, up to 13% Top-1 Acc. Moreover, PLADMM and PLADMM-log are consistently faster than Newton on $\boldsymbol{\beta}$, for all number of observations $M > 100$. Particularly, for $M = 100000$, PLADMM and PLADMM-log converge $4 - 60$ times faster than Newton on $\boldsymbol{\beta}$.

Dataset	Method	Training Metrics		Performance Metrics on the Test Set	
		Time (s) ↓	Iter. ↓	Top-1 Acc. ↑	KT ↑
FAC	**PLADMM**	0.352 ± 0.044	4 ± 0	0.68 ± 0.048	0.35 ± 0.089
	PLADMM-log	0.17 ± 0.033	4 ± 0	0.691 ± 0.054	0.378 ± 0.11
	ILSR (no $\boldsymbol{X}$)	0.066 ± 0.012	2 ± 0	0.591 ± 0.067	-0.13 ± 0.164
	MM (no $\boldsymbol{X}$)	10.7 ± 0.501	500 ± 0	0.544 ± 0.046	0.046 ± 0.087
	Newton on $\boldsymbol{\theta}$ (no $\boldsymbol{X}$)	9.152 ± 1.284	17 ± 3	0.5 ± 0.0	0.0 ± 0.0
	Newton on $\boldsymbol{\beta}$	1.531 ± 0.169	6 ± 1	0.701 ± 0.04	0.398 ± 0.08
	SLSQP	22.73 ± 19.151	160 ± 135	0.689 ± 0.063	0.375 ± 0.125
ROP	**PLADMM**	1.953 ± 0.217	4 ± 0	0.896 ± 0.005	0.791 ± 0.009
	PLADMM-log	0.359 ± 0.027	1 ± 0	0.904 ± 0.005	0.807 ± 0.01
	ILSR (no $\boldsymbol{X}$)	0.716 ± 0.058	2 ± 0	0.891 ± 0.005	0.781 ± 0.009
	MM (no $\boldsymbol{X}$)	356.497 ± 29.11	500 ± 0	0.905 ± 0.004	0.81 ± 0.008
	Newton on $\boldsymbol{\theta}$ (no $\boldsymbol{X}$)	85.42 ± 6.849	9 ± 0	0.906 ± 0.004	0.811 ± 0.008
	Newton on $\boldsymbol{\beta}$	55.718 ± 6.293	2 ± 0	0.904 ± 0.005	0.808 ± 0.009
	SLSQP	9.595 ± 7.136	2 ± 1	0.683 ± 0.049	0.366 ± 0.098
Pairwise Sushi	**PLADMM**	0.061 ± 0.002	4 ± 0	0.669 ± 0.034	0.338 ± 0.068
	PLADMM-log	0.764 ± 1.192	58 ± 30	0.634 ± 0.075	0.267 ± 0.15
	ILSR (no $\boldsymbol{X}$)	0.027 ± 0.003	2 ± 0	0.763 ± 0.039	0.521 ± 0.084
	MM (no $\boldsymbol{X}$)	5.191 ± 0.345	490 ± 31	0.773 ± 0.048	0.543 ± 0.094
	Newton on $\boldsymbol{\theta}$ (no $\boldsymbol{X}$)	2.342 ± 0.689	18 ± 5	0.735 ± 0.095	0.465 ± 0.185
	Newton on $\boldsymbol{\beta}$	0.176 ± 0.17	2 ± 2	0.685 ± 0.044	0.369 ± 0.087
	SLSQP	16.198 ± 8.728	245 ± 134	0.64 ± 0.06	0.28 ± 0.119
Triplet Sushi	**PLADMM**	0.127 ± 0.007	4 ± 0	0.569 ± 0.035	0.218 ± 0.045
	PLADMM-log	0.804 ± 0.349	36 ± 18	0.487 ± 0.034	0.19 ± 0.072
	ILSR (no $\boldsymbol{X}$)	0.054 ± 0.003	2 ± 0	0.678 ± 0.036	0.454 ± 0.06
	MM (no $\boldsymbol{X}$)	15.349 ± 0.617	500 ± 0	0.715 ± 0.035	0.522 ± 0.059
	Newton on $\boldsymbol{\theta}$ (no $\boldsymbol{X}$)	5.122 ± 0.34	14 ± 1	0.73 ± 0.036	0.496 ± 0.089
	Newton on $\boldsymbol{\beta}$	1.12 ± 0.659	3 ± 2	0.605 ± 0.058	0.285 ± 0.062
	SLSQP	21.738 ± 39.761	107 ± 197	0.521 ± 0.043	0.191 ± 0.059

Table 4: Evaluations on real datasets partitioned w.r.t. observation CV (c.f. Sec. 5). We report the convergence time in seconds (Time), number of iterations until convergence (Iter), top-1 accuracy on the test set (Top-1 Acc.), and Kendall-Tau correlation on the test set (KT). ILSR, MM, and Newton on $\boldsymbol{\theta}$ learn the Plackett-Luce scores $\boldsymbol{\pi}$ from the choice observations alone and do not use the features $\boldsymbol{X}$. Newton on $\boldsymbol{\beta}$ and sequential least squares quadratic programming (SLSQP) regress $\boldsymbol{\pi}$ from $\boldsymbol{X}$. (c.f. Sec. F.2).