Supplementary Materials

Before starting with the proofs, we introduce some useful notations to be used later. Recall $\mathcal{I} = \{S \subseteq [n]_+ : S = \mathcal{S}_+, \mathcal{S} \in \boldsymbol{\mathcal{S}} \backslash \varnothing\}$ is the set of all non-empty offered assortments augmented by the no-purchase option, and we denote $\mathcal{S}_+$ by S for the convenience of notation. Define $\mathcal{I}_k = \{S \subseteq \mathcal{I} : |S| = k\}$ to be the subsets of $\mathcal{I}$ with cardinality equal to k, $2 \le k \le n+1$. Correspondingly, we let $\boldsymbol{\mathcal{E}}_k = \{S \in \mathcal{I}_k : \mathcal{E}_S = 1\}$ for $k = 2, 3, \ldots, n+1$. Denote by $\mathbb{I}\{\cdot\}$ an indicator function of statements, which is equal to 1 if the statement inside $\{\cdot\}$ holds true and 0 otherwise. For a sequence of random variables $\{X_n\}_{n=1}^{\infty}$, we write $X_n = o_P(1)$ if $X_n \xrightarrow{P} 0$, and we write $X_n = O_P(1)$ if for all $\epsilon > 0$, there exists an $M > 0$ such that $\sup_n \mathbb{P}(|X_n| > M) < \epsilon$. For a symmetric matrix $\mathbf{M}$, we define $\lambda_{\min,\perp}(\mathbf{M})$ to be its smallest eigenvalue restricted to the eigenvectors orthogonal to $\mathbf{1}$:

$$\lambda_{\min,\perp}(\mathbf{M}) = \min\{\lambda : \lambda = \mathbf{v}^\top \mathbf{M}\mathbf{v}, \text{ for all } \mathbf{v} \text{ such that } \|\mathbf{v}\|_2 = 1 \text{ and } \mathbf{v}^\top \mathbf{1} = 0\}.$$

EC.1. Additional Empirical Evaluation Experiments

We conduct additional numerical experiments to assess the robustness of the proposed method with respect to the connectivity structure of the observed assortment graph. Specifically, we first report the values of $\kappa_{\mathcal{E}}$ under a range of data-generating schemes, from uniformly generated assortments to revenue-ordered assortments and intermediate cases. We then examine how the connectivity parameter $\kappa_{\mathcal{E}}$ in Assumption 1 affects the empirical coverage of the proposed confidence interval. Finally, we report the corresponding weighted-connectivity quantities, including $\underline{\lambda}$ and $\kappa_{\mathcal{E}}$, for several publicly available choice data sets.

EC.1.1. Empirical Values of Connectivity Parameter Across Historical-Assortment Designs

To provide a more intuitive understanding of how $\kappa_{\mathcal{E}}$ varies across different data-generating regimes, we conduct an additional simulation that varies only the historical-assortment generation process and reports the corresponding values of $\kappa_{\mathcal{E}}$. We set $n = 30$, as in the main simulation study, and generate $L = \lceil 10n \log n \rceil = 1,021$ observed historical assortments in each Monte Carlo replication. Each offer set includes the no-purchase option. Let $\rho_0 \in \{0, 0.25, 0.5, 0.75, 1\}$ denote the probability that an offer set is revenue ordered. Conditional on ρ_0, each historical offer set is generated

independently as follows: with probability $1-\rho_0$, we draw a nonempty subset of $[n]$ uniformly at random; with probability ρ_0, we draw a revenue-ordered assortment $[k]$ with k uniformly distributed on $\{1,\dots,n\}$. For each generated historical-assortment structure, we construct the weighted comparison-graph Laplacian in Assumption 1 and compute $\kappa_{\mathcal{E}}$. We use 200 Monte Carlo replications for each value of ρ_0.

Table EC.1 Empirical Values of $\kappa_{\mathcal{E}}$ under Random, Mixture, and Revenue-Ordered Historical-Assortment Designs

Historical-assortment design	ρ_0	Mean of $\kappa_{\mathcal{E}}$	SD of $\kappa_{\mathcal{E}}$
Uniform random	0.00	2.22	0.04
25% revenue-ordered	0.25	3.08	0.11
50% revenue-ordered	0.50	5.03	0.32
75% revenue-ordered	0.75	10.68	0.95
Revenue-ordered	1.00	98.90	17.72

Table EC.1 shows that $\kappa_{\mathcal{E}}$ increases substantially as the historical-assortment design moves from uniformly random assortments toward revenue-ordered assortments. This pattern is consistent with the role of $\kappa_{\mathcal{E}}$ in the theory: uniformly random assortments induce a relatively well-balanced comparison graph, whereas revenue-ordered offer sets concentrate comparisons along a nested structure and lead to a more ill-conditioned historical-assortment design. In the next section, we further examine the robustness of the proposed tests with respect to changes in $\kappa_{\mathcal{E}}$.

EC.1.2. Sensitivity to Observed-Assortment Connectivity

To examine the empirical role of the connectivity parameter $\kappa_{\mathcal{E}}$, we conduct an additional experiment under revenue-ordered historical assortments. We set $n = 30$ and $K^* = 7$, and use the same total sample-size scalings $L/(n\log n) \in \{10, 20, 40, 80\}$ as in Section 4.3.

All historical assortments are revenue-ordered. To isolate the effect of the global connectivity structure from the amount of local information near the true optimal assortment, we allocate 40% of the observations evenly across the assortments $[5],\dots,[9]$, which bracket $[K^*] = [7]$. The remaining 60% of the observations are allocated across the tail assortments $[10],\dots,[30]$, with the sample size assigned to prefix assortment $[k]$ proportional to γ^{30-k}, where

$$\gamma \in \{0, 0.1, 0.25, 0.5, 0.75, 0.9, 0.98, 1\}.$$

This design varies the conditioning of the weighted comparison graph while keeping the amount of local boundary information approximately fixed. To make the threshold nontrivial to recover, we

calibrate the two boundary revenue contrasts so that $\Delta_7 = -0.02$ and $\Delta_8 = 0.02$; thus, the adjacent revenue contrasts around the true threshold are close to zero.

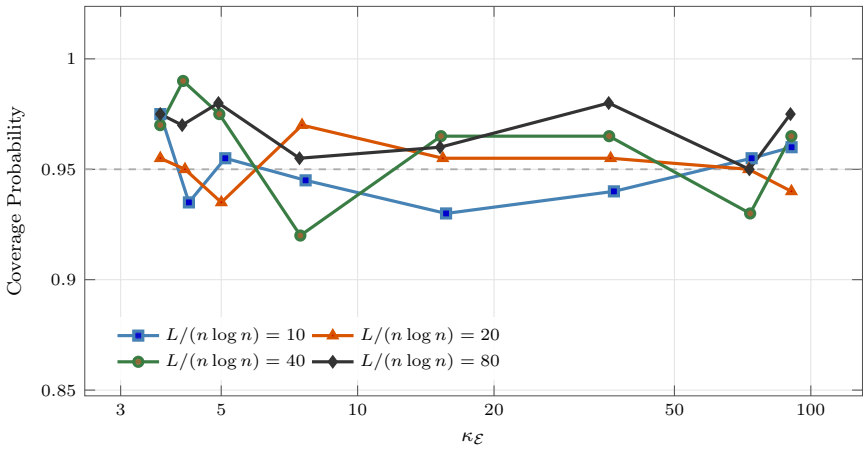
For each pair $(\gamma, L/(n\log n))$, we run 200 Monte Carlo replications using the regularization rule from Section 4.3 and 200 bootstrap draws. We evaluate the empirical coverage of the proposed confidence interval by

$$\mathbb{I}\{K^* \in [\widehat{K}_L, \widehat{K}_U]\}.$$

Across the designs considered here, the resulting values of $\kappa_{\mathcal{E}}$ range from 3.67 to 90.68. Figure EC.1 plots the empirical coverage probability against $\kappa_{\mathcal{E}}$. The coverage remains close to the nominal 95% level over the examined range of $\kappa_{\mathcal{E}}$.

While the theoretical results in Section 3 show that the error bound and sample-size requirements become more stringent as $\kappa_{\mathcal{E}}$ grows, such dependence serves only as a worst-case, nonasymptotic diagnostic. It does not imply that empirical coverage must decrease monotonically with $\kappa_{\mathcal{E}}$ in every finite-sample design. The present experiment shows that, across a broad observed range of $\kappa_{\mathcal{E}}$, the proposed confidence interval remains well calibrated. This stability is consistent with the fact that the confidence-interval construction incorporates increased uncertainty through the estimated standard errors and bootstrap critical values.

Figure EC.1 Empirical coverage of the proposed confidence interval as a function of $\kappa_{\mathcal{E}}$. The coverage remains close to the nominal level 0.95 across the examined range of $\kappa_{\mathcal{E}}$.



EC.1.3. Empirical Values of Observed-Assortment Connectivity on Public Product-Choice Data Sets

We further evaluate $\kappa_{\mathcal{E}}$ on public product-choice data sets commonly used in demand-estimation and choice-modeling applications, where observed choices are naturally modeled by multinomial-logit-type demand models and observed prices serve as revenue proxies. For each data set, we construct the weighted comparison graph from the observed or reconstructed choice sets, using the multiplicity of each historical assortment as its weight and adding the no-purchase option to each offer set, in accordance with (6).

The Tuna data set is a brand-choice data set for canned tuna. It is available through the `Ecdat` package (Croissant 2023) and originates from the retail-pricing application in Kim et al. (1995). The data contain prices and promotional variables for five tuna alternatives; under the standard brand-choice construction, all five alternatives are available in each observed choice occasion. The Ta Feng grocery data, accessed through the Choice-Learn construction (Auriau et al. 2024), provide transaction-level grocery purchases and prices. The public assortment construction used in Choice-Learn retains 25 products and marks all retained products as available in each unit-level choice. The Hotel 1 data of Bodea et al. (2009), distributed through the CRAN `RMM` package (Kim et al. 2022), are the closest match to our revenue-management setting: each observation is a hotel booking decision with an observed exposed room-type choice set and room-specific nightly rates. Finally, the tablet conjoint data from the ColumbiaX business analytics course materials (ColumbiaX 2019) consist of choice-based conjoint tasks in which respondents choose among tablet profiles described by product attributes and price, and we treat each distinct product-price profile as a separate product.

Table EC.2 reports the resulting values of $\underline{\lambda}$ and $\kappa_{\mathcal{E}}$. The Tuna and Ta Feng constructions use complete assortments; hence, their comparison graphs are fully connected and yield $\kappa_{\mathcal{E}} = 1$. The Hotel 1 data contain 12 distinct observed historical assortments and yield $\underline{\lambda} = 0.0124$ and $\kappa_{\mathcal{E}} = 12.41$. The tablet conjoint data contain genuinely varying product sets and yield a much smaller $\underline{\lambda} = 0.000304$ and a much larger $\kappa_{\mathcal{E}} = 620.51$, reflecting substantially weaker weighted connectivity in the induced comparison graph.

Overall, the public-data evidence supports the role of $\kappa_{\mathcal{E}}$ as a diagnostic for the informativeness of historical assortment variation. When the public construction exposes every product in every choice occasion, as in Tuna and Ta Feng, the comparison graph is perfectly balanced and $\kappa_{\mathcal{E}} = 1$. In contrast, data sets with genuinely varying offer sets, such as Hotel 1 and the tablet conjoint data, generate nontrivial and sometimes large values of $\kappa_{\mathcal{E}}$, together with smaller values of $\underline{\lambda}$, reflecting

weaker weighted connectivity. Together with the sensitivity experiment in Section EC.1.2, these results provide an empirical interpretation of $\kappa_{\mathcal{E}}$ and illustrate how the proposed method behaves across different connectivity regimes.

Table EC.2 Empirical Values of Observed-Assortment Connectivity on Public Product-Choice Data Sets

Data set	Assortment scheme	n	L	$\underline{\lambda}$	$\kappa_{\mathcal{E}}$
Tuna	Full assortment	5	13,705	0.1667	1.00
Ta Feng grocery	Full assortment	25	14,215	0.0385	1.00
Hotel 1	Partially observed	10	1,100	0.0124	12.41
Tablet conjoint	Partially observed	350	2,055	0.000304	620.51

EC.2. Proof of Theorem 2

The proof extends the results presented in Theorem 6 of Chen et al. (2019) to the context of the multinomial logit model. The major technical difference lies in the characterization of the new eigenstructure of the Hessian matrix, which is more complex than that of the BTL model. We employ a different approach to control the degenerating rate of the eigenvalues of the Hessian matrix and utilize the new eigenstructure to derive the convergence rate. Note that if we take $\boldsymbol{\theta}' = \mathbf{P}_{\mathbf{1}}^{\perp}\boldsymbol{\theta}$, then the convex problem (9) can be equivalently written as

$$\min_{\boldsymbol{\theta}' \in \mathbb{R}^{n+1}, \mathbf{1}^{\top}\boldsymbol{\theta}'=0} \mathcal{L}_{\tau}(\boldsymbol{\theta}'; \boldsymbol{x}) := \mathcal{L}(\boldsymbol{\theta}'; \boldsymbol{x}) + \frac{\tau}{2}\|\boldsymbol{\theta}'\|_2^2. \tag{EC.1}$$

Later we will show that the minimizer of $\mathcal{L}(\boldsymbol{\theta}'; \boldsymbol{x}) + \frac{\tau}{2}\|\boldsymbol{\theta}'\|_2^2$ belongs to the subspace perpendicular to $\mathbf{1}$, and thus we can simplify (EC.1) as

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{n+1}} \mathcal{L}_{\tau}(\boldsymbol{\theta}; \boldsymbol{x}) := \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x}) + \frac{\tau}{2}\|\boldsymbol{\theta}\|_2^2. \tag{EC.2}$$

It can be observed that the regularized MLE to (EC.2) will be $\widehat{\boldsymbol{\theta}}' := \mathbf{P}_{\mathbf{1}}^{\perp}\widehat{\boldsymbol{\theta}}$, where $\widehat{\boldsymbol{\theta}}$ is the regularized MLE of problem (9). In the following proof, we will consider solving (EC.2) instead. We will show the L_2 convergence rate of the resulting MLE of (EC.2), i.e., $\|\widehat{\boldsymbol{\theta}}' - \mathbf{P}_{\mathbf{1}}^{\perp}\boldsymbol{\theta}^*\|_2 = \|\mathbf{P}_{\mathbf{1}}^{\perp}\widehat{\boldsymbol{\theta}} - \mathbf{P}_{\mathbf{1}}^{\perp}\boldsymbol{\theta}^*\|_2$. Since $\widehat{\boldsymbol{\theta}}'$ can be easily transformed back to $\widehat{\boldsymbol{\theta}}$ by taking $\widehat{\theta}_i = \widehat{\theta}'_i - \widehat{\theta}'_0$ for $i \in [n]_+$, the L_2 convergence rates $\|\widehat{\boldsymbol{\theta}}' - \mathbf{P}_{\mathbf{1}}^{\perp}\boldsymbol{\theta}^*\|_2$ and $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$ will be different only by a factor of constant. Besides, since $\boldsymbol{\theta}^*$ and $\mathbf{P}_{\mathbf{1}}^{\perp}\boldsymbol{\theta}^*$ essentially specify the same MNL model, without loss of generality, we will replace $\boldsymbol{\theta}^*$ by $\mathbf{P}_{\mathbf{1}}^{\perp}\boldsymbol{\theta}^*$ throughout the proofs, and we will abuse the notation and let $\boldsymbol{\theta}^*$ denote $\boldsymbol{\theta}^{*\prime} = \mathbf{P}_{\mathbf{1}}^{\perp}\boldsymbol{\theta}^*$ and let $\widehat{\boldsymbol{\theta}}$ denote $\widehat{\boldsymbol{\theta}}' = \mathbf{P}_{\mathbf{1}}^{\perp}\widehat{\boldsymbol{\theta}}$, i.e., the resulting regularized MLE solving problem (EC.2). Before delving into the details, we need the following lemmas to help with the proof.

The following lemmas depict the smoothness and convexity of the log-likelihood function,

LEMMA EC.1. *Suppose that Assumption 1 holds. For all $\boldsymbol{\theta} \in \mathbb{R}^{n+1}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_\infty \le C$ for some $C \ge 0$, we have*

$$\lambda_{\min,\perp}\left(\nabla^2 \mathcal{L}_\tau(\boldsymbol{\theta};\boldsymbol{x})\right) \ge \tau + \frac{1}{(\kappa_{\boldsymbol{\theta}} e^{2C})^2}\underline{\lambda},\ \lambda_{\max}\left(\nabla^2 \mathcal{L}_\tau(\boldsymbol{\theta};\boldsymbol{x})\right) \le \tau + (\kappa_{\boldsymbol{\theta}} e^{2C})^2 \bar{\lambda}. \tag{EC.3}$$

Please see Section EC.9.3 for the proof of Lemma EC.1.

LEMMA EC.2. *Under Assumption 1 and the conditions that $n\bar{\lambda}L \gtrsim \log n$ and $\tau \asymp \sqrt{\bar{\lambda}\log n/L}$, the following event*

$$\mathcal{A}_g := \left\{ \|\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*;\boldsymbol{x})\|_2 \lesssim \sqrt{\frac{n\bar{\lambda}\log n}{L}} \right\} \tag{EC.4}$$

occurs with probability at least $1 - O(n^{-10})$.

Please refer to Section EC.9.4 for the proof of Lemma EC.2.

LEMMA EC.3. *Under Assumption 1, we have that for any $\boldsymbol{\theta} \in \mathbb{R}^{n+1}$ the following event holds*

$$\mathcal{A}_h := \left\{ \|\nabla^2 \mathcal{L}_\tau(\boldsymbol{\theta};\boldsymbol{x})\|_2 \le \tau + (n+1)^2\bar{\lambda}/4 \right\}. \tag{EC.5}$$

Please see Section EC.9.5 for the proof of Lemma EC.3. Now we begin with the proof. As in the proof of Chen et al. (2019), we let $\boldsymbol{\theta}^T$ be the output of the following gradient descent

- Initialize $\boldsymbol{\theta}^0 = \boldsymbol{\theta}^*$,
- for $t = 0, 1, \ldots, T-1$,

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta \nabla \mathcal{L}_\tau(\boldsymbol{\theta}^t;\boldsymbol{x}),$$

where η is taken to be $1/(\tau + (n+1)^2\bar{\lambda}/4)$.

Step I

By Lemma EC.3, we know that $\mathcal{L}_\tau(\boldsymbol{\theta};\boldsymbol{x})$ is $(\tau + (n+1)^2\bar{\lambda}/4)$-smooth and τ-strongly convex for any $\boldsymbol{\theta} \in \mathbb{R}^{n+1}$, then by Theorem 3.10 in Bubeck (2015) we have

$$\|\boldsymbol{\theta}^t - \widehat{\boldsymbol{\theta}}\|_2 \le \rho^t \|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}\|_2,$$

where $\rho = 1 - \frac{\tau}{\tau + (n+1)^2\bar{\lambda}/4}$. Note that since $\mathbf{1}^\top \boldsymbol{\theta}^* = 0$ and $\mathbf{1}^\top \nabla \mathcal{L}(\boldsymbol{\theta};\boldsymbol{x}) = \mathbf{0}$ for any $\boldsymbol{\theta} \in \mathbb{R}^{n+1}$, it follows by simple induction that $\mathbf{1}^\top \boldsymbol{\theta}^t = 0$ for all $t \ge 1$, and hence $\widehat{\boldsymbol{\theta}}^\top \mathbf{1} = 0$ due to the convergence of $\boldsymbol{\theta}^t$ to $\widehat{\boldsymbol{\theta}}$. The following lemma provides an initial upper bound for $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}\|_2$.

LEMMA EC.4. *On the event $\mathcal{A}_g$ as defined in (EC.4), there exists a constant $c > 0$ such that*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \le c\sqrt{n}.$$

The proof is deferred to Section EC.9.6. Recall that $\tau \asymp \sqrt{\frac{\bar{\lambda}\log n}{L}}$, then with Lemma EC.4, under the event $\mathcal{A}_g$, for large enough T we have

$$\begin{aligned}\|\boldsymbol{\theta}^T - \widehat{\boldsymbol{\theta}}\|_2 \le \rho^T \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 &\lesssim \left(1 - \frac{\tau}{\tau + (n+1)^2\bar{\lambda}/4}\right)^T \sqrt{n} \\ &\le \exp\left(-\frac{T\tau}{\tau + (n+1)^2\bar{\lambda}/4}\right)\sqrt{n} \le \exp\left(-c_3 T\sqrt{\frac{\log n}{n^4\bar{\lambda}L}}\right)\sqrt{n} \\ &\lesssim \sqrt{\frac{n\kappa_{\mathcal{E}}\log n}{\underline{\lambda}L}}.\end{aligned}$$

Step II

Now we show the convergence rate of $\|\boldsymbol{\theta}^T - \boldsymbol{\theta}^*\|_2$ by induction. More specifically, we propose that with probability at least $1 - O(n^{-10})$, the following holds true for some constant $C_1 > 0$ for any $t \in [T]_+$,

$$\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|_2 \le C_1\sqrt{\frac{n\kappa_{\mathcal{E}}\log n}{\underline{\lambda}L}}. \tag{EC.6}$$

Note that the base case $t = 0$ holds true trivially because of our setting of $\boldsymbol{\theta}^0 = \boldsymbol{\theta}^*$. Now suppose that (EC.6) holds true for the t-th iteration, we will show that under the event $\mathcal{A}_g$ (occurring with probability at least $1 - O(n^{-10})$), the following holds

$$\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^*\|_2 \le C_1\sqrt{\frac{n\kappa_{\mathcal{E}}\log n}{\underline{\lambda}L}}.$$

First note that if (EC.6) holds for the t-th iteration, we have

$$\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|_\infty \le \|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|_2 \le C_1\sqrt{\frac{n\kappa_{\mathcal{E}}\log n}{\underline{\lambda}L}}.$$

The rest of the proof follows the proof of Lemma 14 in Chen et al. (2019) with modifications. From the gradient descent algorithm we know that

$$\begin{aligned}\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^* &= \boldsymbol{\theta}^t - \eta\nabla\mathcal{L}_\tau\left(\boldsymbol{\theta}^t;\boldsymbol{x}\right) - \boldsymbol{\theta}^* \\ &= \boldsymbol{\theta}^t - \eta\nabla\mathcal{L}_\tau\left(\boldsymbol{\theta}^t;\boldsymbol{x}\right) - \left[\boldsymbol{\theta}^* - \eta\nabla\mathcal{L}_\tau\left(\boldsymbol{\theta}^*;\boldsymbol{x}\right)\right] - \eta\nabla\mathcal{L}_\tau\left(\boldsymbol{\theta}^*;\boldsymbol{x}\right) \\ &= \left\{\boldsymbol{I}_n - \eta\int_0^1 \nabla^2\mathcal{L}_\tau(\boldsymbol{\theta}(v);\boldsymbol{x})\mathrm{d}v\right\}\left(\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\right) - \eta\nabla\mathcal{L}_\tau\left(\boldsymbol{\theta}^*;\boldsymbol{x}\right),\end{aligned}$$

where $\boldsymbol{\theta}(\nu) := \boldsymbol{\theta}^* + \nu(\boldsymbol{\theta}^t - \boldsymbol{\theta}^*)$. Then we can see that if $C_1\sqrt{n\kappa_{\mathcal{E}}\log n/(\underline{\lambda}L)} \le \epsilon$ for a sufficiently small constant ϵ, we have that $\|\boldsymbol{\theta}(\nu) - \boldsymbol{\theta}^*\|_\infty \le \nu\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|_\infty \le \epsilon$. Then by Lemma EC.1 we have that for all $\nu \in [0,1]$

$$\tau + \frac{1}{2\kappa_{\boldsymbol{\theta}}^2}\underline{\lambda} \le \lambda_{\min,\perp}\left(\nabla^2\mathcal{L}_\tau(\boldsymbol{\theta}(\nu);\boldsymbol{x})\right) \le \lambda_{\max}\left(\nabla^2\mathcal{L}_\tau(\boldsymbol{\theta}(\nu);\boldsymbol{x})\right) \le \tau + 2\kappa_{\boldsymbol{\theta}}^2\bar{\lambda}.$$

Now we denote $\mathbf{B} = \int_0^1 \nabla^2\mathcal{L}_\tau(\boldsymbol{\theta}(\nu);\boldsymbol{x})\mathrm{d}\nu$, then we have

$$\begin{aligned}
\mathbf{B}\mathbf{1}_{n+1} &= \int_0^1 \nabla^2\mathcal{L}_\tau(\boldsymbol{\theta}(\nu);\boldsymbol{x})\mathbf{1}_{n+1}\mathrm{d}\nu = \mathbf{0};\\
\lambda_{\min,\perp}(\mathbf{B}) &= \min_{\mathbf{v}\in\mathbf{S}^n,\mathbf{v}^\top\mathbf{1}=0}\int_0^1 \mathbf{v}^\top\nabla^2\mathcal{L}_\tau(\boldsymbol{\theta}(\nu);\boldsymbol{x})\mathbf{v}\mathrm{d}\nu \ge \int_0^1 \lambda_{\min,\perp}\left(\nabla^2\mathcal{L}_\tau(\boldsymbol{\theta}(\nu);\boldsymbol{x})\right)\mathrm{d}\nu\\
&\ge \tau + \frac{1}{2\kappa_{\boldsymbol{\theta}}^2}\underline{\lambda};\\
\lambda_{\max}(\mathbf{B}) &= \max_{\mathbf{v}\in\mathbf{S}^n}\int_0^1 \mathbf{v}^\top\nabla^2\mathcal{L}_\tau(\boldsymbol{\theta}(\nu);\boldsymbol{x})\mathbf{v}\mathrm{d}\nu \le \tau + 2\kappa_{\boldsymbol{\theta}}^2\bar{\lambda}.
\end{aligned}$$

Besides, by the definition of η we have that $1 - \eta\left(\tau + \frac{1}{2\kappa_{\boldsymbol{\theta}}^2}\underline{\lambda}\right) \ge 1 - \eta\left(\tau + 2\kappa_{\boldsymbol{\theta}}^2\bar{\lambda}\right) \ge 0$. Combining the above results we have

$$\begin{aligned}
\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^*\|_2 &\le \left(1 - \frac{1}{2\kappa_{\boldsymbol{\theta}}^2}\eta\underline{\lambda}\right)\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|_2 + \eta\|\nabla\mathcal{L}_\tau\left(\boldsymbol{\theta}^*;\boldsymbol{x}\right)\|_2\\
&\le \left(1 - \frac{1}{2\kappa_{\boldsymbol{\theta}}^2}\eta\underline{\lambda}\right)C_1\sqrt{\frac{n\kappa_{\mathcal{E}}\log n}{\underline{\lambda}L}} + C\eta\sqrt{\frac{n\bar{\lambda}\log n}{L}}\\
&\le C_1\sqrt{\frac{n\kappa_{\mathcal{E}}\log n}{\underline{\lambda}L}},
\end{aligned}$$

so long as C_1 is large enough.

Step III

Combining the results in Step I and Step II, as T goes to infinity we have that with probability at least $1 - O(n^{-10})$,

$$\|\boldsymbol{\theta}^T - \boldsymbol{\theta}^*\|_2 \lesssim \sqrt{\frac{n\kappa_{\mathcal{E}}\log n}{\underline{\lambda}L}},$$

and

$$\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2 \le \|\boldsymbol{\theta}^T-\widehat{\boldsymbol{\theta}}\|_2+\|\boldsymbol{\theta}^T-\boldsymbol{\theta}^*\|_2 \lesssim \sqrt{\frac{n\kappa_{\mathcal{E}}\log n}{\underline{\lambda}L}}.$$

EC.3. Proof of Theorem 3

Before we begin with the proof, we first propose the following lemmas that provide bounds for several terms to be used in the proof.

LEMMA EC.5. *Under the same conditions as Theorem 2, for any $\boldsymbol{\theta}\in\mathbb{R}^{n+1}$ such that $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_\infty < C$ for some constant $C>0$, with probability at least $1-O(n^{-10})$ we have the following bounds*

$$\|\nabla\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})\|_\infty \lesssim \sqrt{\frac{\bar{\lambda}\log n}{L}}+\frac{\log n}{L}, \tag{EC.7}$$

$$\|\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})-\nabla\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})-\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\|_\infty \lesssim \frac{\kappa_{\mathcal{E}}^2 n\log n}{L}, \tag{EC.8}$$

$$\|\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})-\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})\|_\infty \lesssim \sqrt{\frac{\kappa_{\mathcal{E}}^2 n\bar{\lambda}\log n}{L}}, \tag{EC.9}$$

$$\left\|\begin{pmatrix}\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) & \mathbf{1}\\ \mathbf{1}^\top & 0\end{pmatrix}^{-1}-\begin{pmatrix}\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x}) & \mathbf{1}\\ \mathbf{1}^\top & 0\end{pmatrix}^{-1}\right\|_2 \lesssim \frac{n}{\underline{\lambda}}\sqrt{\frac{\kappa_{\mathcal{E}}^3\log n}{\underline{\lambda}L}}. \tag{EC.10}$$

The proof of Lemma EC.5 is deferred to Section EC.9.7. The following corollary of Lemma EC.5 helps characterize the Moore-Penrose inverse of the Hessian matrix.

COROLLARY EC.1. *For any $\boldsymbol{\theta}\in\mathbb{R}^{n+1}$, we have that*

$$\begin{pmatrix}\nabla^2\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x}) & \mathbf{1}\\ \mathbf{1}^\top & 0\end{pmatrix}^{-1}=\begin{pmatrix}\nabla^2\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x})^\dagger & \frac{1}{n+1}\mathbf{1}\\ \frac{1}{n+1}\mathbf{1}^\top & 0\end{pmatrix}.$$

Besides, for $\boldsymbol{\theta}\in\mathbb{R}^{n+1}$ such that $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_\infty < C$ for some constant $C>0$, we have that

$$\frac{1}{\bar{\lambda}} \lesssim [\nabla^2\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x})^\dagger]_{jj} \lesssim \frac{1}{\underline{\lambda}} \quad \textit{for } j\in[n+1], \quad \textit{and} \quad \left|[\nabla^2\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x})^\dagger]_{jk}\right| \lesssim \frac{1}{\underline{\lambda}} \quad \textit{for } j\neq k.$$

See Section EC.9.8 for the proof of Corollary EC.1. The following lemma provides a decomposition of the error of the debiased MLE $\widehat{\boldsymbol{\theta}}^d$ into the leading term and the remainder term.

LEMMA EC.6. *Under the same conditions as Theorem 2 , we have the following decomposition for the debiased MLE*

$$\widehat{\boldsymbol{\theta}}^d-\boldsymbol{\theta}^* = -\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})^\dagger\nabla\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})+\mathbf{R}_0, \tag{EC.11}$$

where the residual term

$$\|\mathbf{R}_0\|_2 \lesssim \frac{\kappa_{\mathcal{E}}^2 n\sqrt{n}\log n}{\underline{\lambda}L}\left(n\sqrt{\frac{\kappa_{\mathcal{E}}^3\log n}{\underline{\lambda}L}}+1\right),$$

with probability at least $1-O(n^{-10})$.

See Section EC.9.9 for the proof of Lemma EC.6. Now we can begin with the proof. We abbreviate $\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x})$ to $\mathcal{L}(\boldsymbol{\theta})$ for the convenience of notation in the following proof. For $k\in[n]$, we define the mapping $\mathbf{v}_k(\cdot):\mathbb{R}^{n+1}\to\mathbb{R}^{n+1}$

$$\mathbf{v}_k(\boldsymbol{\theta}) = \Big(\underbrace{(0-r_k)e^{\theta_0},(r_1-r_k)e^{\theta_1},(r_2-r_k)e^{\theta_2},\ldots,(r_{k-1}-r_k)e^{\theta_{k-1}}}_{\text{first } k \text{ entries}},0,\ldots,0\Big)^\top,$$

and for the convenience of notation we let $\mathbf{v}_k=\mathbf{v}_k(\boldsymbol{\theta}^*)$. Define the function

$$g_k(\boldsymbol{\theta}) = \sum_{i=1}^{k} r_i\exp(\theta_i) - (\sum_{i=0}^{k}\exp(\theta_i))r_k,$$

and we have $\nabla g(\boldsymbol{\theta})=\mathbf{v}_k(\boldsymbol{\theta})$ and $\nabla^2 g(\boldsymbol{\theta})=\operatorname{diag}\big(\mathbf{v}_k(\boldsymbol{\theta})\big)$, where

$$\mathbf{v}_k(\boldsymbol{\theta}) = \Big(\underbrace{(0-r_k)e^{\theta_0},(r_1-r_k)e^{\theta_1},(r_2-r_k)e^{\theta_2},\ldots,(r_{k-1}-r_k)e^{\theta_{k-1}}}_{\text{first } k \text{ entries}},0,\ldots,0\Big)^\top.$$

Then we have

$$\begin{aligned}\widehat{\Delta}_k-\Delta_k &= g_k(\widehat{\boldsymbol{\theta}}^d)-g_k(\boldsymbol{\theta}^*) = \nabla g(\boldsymbol{\theta}^*)^\top(\widehat{\boldsymbol{\theta}}^d-\boldsymbol{\theta}^*)+\frac{1}{2}(\widehat{\boldsymbol{\theta}}^d-\boldsymbol{\theta}^*)^\top\nabla^2 g(\widetilde{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}^d-\boldsymbol{\theta}^*)\\ &= \mathbf{v}_k(\boldsymbol{\theta}^*)^\top(-\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\nabla\mathcal{L}(\boldsymbol{\theta}^*)+\mathbf{R}_0)+r_1,\end{aligned}$$

where $\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2\le\|\widehat{\boldsymbol{\theta}}^d-\boldsymbol{\theta}^*\|_2$, $\|\mathbf{R}_0\|_2\lesssim\frac{\kappa_{\mathcal{E}}^2 n\sqrt{n}\log n}{\underline{\lambda}L}\left(n\sqrt{\frac{\kappa_{\mathcal{E}}^3\log n}{\underline{\lambda}L}}+1\right)$ with probability at least $1-O(n^{-10})$ from Lemma EC.6, and

$$|r_1| = \frac{1}{2}(\widehat{\boldsymbol{\theta}}^d-\boldsymbol{\theta}^*)^\top\nabla^2 g(\widetilde{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}^d-\boldsymbol{\theta}^*)\lesssim\|\mathbf{v}_k\|_\infty\|\widehat{\boldsymbol{\theta}}^d-\boldsymbol{\theta}^*\|_2^2.$$

Note that from the proof of Lemma EC.6 we can see that $\mathbf{R}_0^\top\mathbf{1}=0$, and hence $|\mathbf{v}_k^\top\mathbf{R}_0|\le\|\mathbf{R}_0\|_2\|\mathbf{P}_{\mathbf{1}}^\perp\mathbf{v}_k\|_2$, where $\mathbf{P}_{\mathbf{1}}^\perp$ is the projection matrix to the perpendicular space of $\mathbf{1}$. Also from

Lemma EC.5 and the proof of Lemma EC.6, we know that under the condition that $\underline{\lambda}L \gtrsim \kappa_{\mathcal{E}}^3 n^2 \log n$, with high probability we have

$$\|\widehat{\boldsymbol{\theta}}^d - \boldsymbol{\theta}^*\|_2 \lesssim \frac{1}{\underline{\lambda}} \times \sqrt{\frac{n\bar{\lambda}\log n}{L}} + \frac{\kappa_{\mathcal{E}}^2 n\sqrt{n}\log n}{\underline{\lambda}L} \lesssim \sqrt{\frac{\kappa_{\mathcal{E}} n \log n}{\underline{\lambda}L}},$$

and thus we have $|r_1| \lesssim \kappa_{\mathcal{E}} n \log n \|\mathbf{v}_k\|_\infty/(\underline{\lambda}L)$. Now we consider the term $\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \nabla \mathcal{L}(\boldsymbol{\theta}^*)$. We have

$$\begin{aligned} -\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \nabla \mathcal{L}(\boldsymbol{\theta}^*) &= \frac{1}{L}\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \sum_{\mathcal{S}\in\mathcal{E}} \sum_{\ell=1}^{L_{\mathcal{S}}} \left\{ \sum_{i\in\mathcal{S}_+} \left(x_{\mathcal{S}}^{(i,\ell)} - \frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}} \right) \boldsymbol{e}_i \right\} \\ &= \frac{1}{L} \sum_{\mathcal{S}\in\mathcal{E}} \sum_{\ell=1}^{L_{\mathcal{S}}} y_{\mathcal{S}}^{(\ell)}, \end{aligned}$$

where $y_{\mathcal{S}}^{(\ell)} := \mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \left\{ \sum_{i\in\mathcal{S}_+} \left(x_{\mathcal{S}}^{(i,\ell)} - \frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}} \right) \boldsymbol{e}_i \right\}$ are independent for ℓ and S given $\mathcal{E}$. Then we have

$$\mathrm{Var}\left(\sum_{\mathcal{S}\in\mathcal{E}} \sum_{\ell=1}^{L_{\mathcal{S}}} y_{\mathcal{S}}^{(\ell)} \right) = L \cdot \mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*) \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k = L \cdot \mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k \gtrsim \frac{L \cdot \|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2^2}{\bar{\lambda}}.$$

Now consider $\left| y_{\mathcal{S}}^{(\ell)} \right|$, and we have

$$\left| y_{\mathcal{S}}^{(\ell)} \right| \lesssim \|\nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k\|_2 \left\| \sum_{i\in\mathcal{S}_+} \left(x_{\mathcal{S}}^{(i,\ell)} - \frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}} \right) \boldsymbol{e}_i \right\|_2 \lesssim \|\nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger\|_2 \|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2 \lesssim \frac{1}{\underline{\lambda}} \|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2,$$

where the last inequality is due to the fact that $\nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{1} = \mathbf{0}$. Then in turn we also have

$$\begin{aligned} \sum_{\mathcal{S}\in\mathcal{E}} \sum_{\ell=1}^{L_{\mathcal{S}}} \mathbb{E}\left| y_{\mathcal{S}}^{(\ell)} \right|^3 &\lesssim \left(\max_{\mathcal{S}\in\mathcal{E}} \max_{\ell\in[L_{\mathcal{S}}]} \left| y_{\mathcal{S}}^{(\ell)} \right| \right) \sum_{\mathcal{S}\in\mathcal{E}} \sum_{\ell=1}^{L_{\mathcal{S}}} \mathbb{E}\left| y_{\mathcal{S}}^{(\ell)} \right|^2 \\ &\lesssim \frac{L}{\underline{\lambda}} \mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k \|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2 \lesssim \frac{L}{\underline{\lambda}^2} \cdot \|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2^3. \end{aligned}$$

Thus under the condition that $\kappa_{\mathcal{E}}^3/(\underline{\lambda}L) = o(1)$, we have

$$\max_k \left(\sum_{\mathcal{S}\in\mathcal{E}} \sum_{\ell=1}^{L_{\mathcal{S}}} \mathrm{Var}\left(y_{\mathcal{S}}^{(\ell)}\right) \right)^{-3/2} \left(\sum_{\mathcal{S}\in\mathcal{E}} \sum_{\ell=1}^{L_{\mathcal{S}}} \mathbb{E}\left|y_{\mathcal{S}}^{(\ell)}\right|^3 \right) = o(1).$$

Then by Lyapunov's Central Limit Theorem, we have

$$\sqrt{L}(\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k)^{-1/2} \left(\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \nabla \mathcal{L}(\boldsymbol{\theta}^*) \right) \xrightarrow{d} N(0,1), \text{ for all } k \in [n].$$

Besides, under the condition that

$$\max_k \left\{ \kappa_{\mathcal{E}}^3 n \log n \sqrt{\frac{\kappa_{\mathcal{E}} \cdot n}{\underline{\lambda} L}} + \frac{\kappa_{\mathcal{E}} \log n \|\mathbf{v}_k\|_\infty}{\|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2} \sqrt{\frac{\kappa_{\mathcal{E}} \cdot n^2}{\underline{\lambda} L}} \right\} = o(1), \tag{EC.12}$$

we have $\sqrt{L}(\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k)^{-1/2}(\mathbf{v}_k^\top \mathbf{R}_0 + r_1) = o_P(1)$, and by Slutsky's Theorem the claim follows.

Now to simplify the conditions, we turn to study the norm $\|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2$ for $k \in [n]$. First note that $\mathbf{P}_{\mathbf{1}} \mathbf{v}_k = (n+1)^{-1}\Delta_k \cdot \mathbf{1}$. Due to the non-decreasingness of Δ_k, we consider the scenarios of $k \le K^*$ and $k > K^*$ separately.

- $k \le K^*$. From Theorem 1 we know that $\Delta_k \le 0$ when $k \le K^*$. Besides, since $(r_i - r_k)u_i^* \ge 0$ for $i < k$, we know that $|r_k u_0^*| \ge |\Delta_k|$. Thus we have that

$$\begin{aligned}
\|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2^2 &= \left(-r_k u_0^* - \frac{\Delta_k}{n+1}\right)^2 + \sum_{i=1}^{k-1} \left\{ (r_i - r_k)u_i^* - \frac{\Delta_k}{n+1} \right\}^2 + \frac{n-k+1}{(n+1)^2}\Delta_k^2 \\
&= (r_k u_0^*)^2 + \sum_{i=1}^{k-1} \left((r_i - r_k)u_i^* \right)^2 - \frac{\Delta_k^2}{n+1} \ge (r_k u_0^*)^2 - \frac{\Delta_k^2}{n+1} \\
&\ge (r_k u_0^*)^2 - \frac{(r_k u_0^*)^2}{n+1} = \frac{n}{n+1}(r_k u_0^*)^2 \ge \frac{(r_k u_0^*)^2}{2}.
\end{aligned}$$

- $k > K^*$. When $k > K^*$, we know that $\Delta_K \ge 0$ from Theorem 1. Then we have

$$\begin{aligned}
\|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2^2 &= \left(-r_k u_0^* - \frac{\Delta_k}{n+1}\right)^2 + \sum_{i=1}^{k-1} \left\{ (r_i - r_k)u_i^* - \frac{\Delta_k}{n+1} \right\}^2 + \frac{n-k+1}{(n+1)^2}\Delta_k^2 \\
&\ge \left(-r_k u_0^* - \frac{\Delta_k}{n+1}\right)^2 \ge (r_n u_0^*)^2.
\end{aligned}$$

Combining the above results we have that $\|\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{v}_k\|_2 \gtrsim r_n u_0^*$, and in turn we have that

$$\max_k \frac{\|\mathbf{v}_k\|_\infty}{\|\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{v}_k\|_2} \lesssim \frac{r_1 \max_i u_i^*}{r_n u_0^*} \le \kappa_{\boldsymbol{\theta}} \frac{r_1}{r_n} = \kappa_r \kappa_{\boldsymbol{\theta}} \lesssim \kappa_r.$$

Therefore, under the condition that $\kappa_r \lesssim \sqrt{n}$, the condition (EC.12) simplifies to

$$\max_k \left\{ \kappa_{\mathcal{E}}^3 n \log n \sqrt{\frac{\kappa_{\mathcal{E}} \cdot n}{\underline{\lambda} L}} + \kappa_r \kappa_{\mathcal{E}} \log n \sqrt{\frac{\kappa_{\mathcal{E}} \cdot n^2}{\underline{\lambda} L}} \right\} \lesssim \kappa_{\mathcal{E}}^3 n \log n \sqrt{\frac{\kappa_{\mathcal{E}} \cdot n}{\underline{\lambda} L}} = o(1).$$

EC.4. Proof of Corollary 1

By Slutsky's Theorem, it suffices for us to show that

$$\max_k \left\{ |\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}}; \boldsymbol{x})^\dagger \widehat{\mathbf{v}}_k - \mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*; \boldsymbol{x})^\dagger \mathbf{v}_k | / \mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*; \boldsymbol{x})^\dagger \mathbf{v}_k \right\} = o_P(1).$$

From (EC.10), we know that $\|\nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}}; \boldsymbol{x})^\dagger - \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*; \boldsymbol{x})^\dagger\|_2 \lesssim \frac{n}{\underline{\lambda}} \sqrt{\frac{\kappa_{\mathcal{E}}^3 \log n}{\underline{\lambda} L}}$ with high probability. Also from Theorem 2, we know that $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \lesssim \sqrt{\frac{n \kappa_{\mathcal{E}} \log n}{\underline{\lambda} L}}$, and in turn

$$\|\widehat{\mathbf{v}}_k - \mathbf{v}_k\|_2 = O_P\left(\|\nabla^2 g(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2 \right) = O_P\left(\|\operatorname{diag}(\mathbf{v}_k)\|_2 \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \right) = O_P\left(\|\mathbf{v}_k\|_\infty \sqrt{\frac{n \kappa_{\mathcal{E}} \log n}{\underline{\lambda} L}} \right).$$

Combining the previous results, we have

$$\begin{aligned}
&|\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}}; \boldsymbol{x})^\dagger \widehat{\mathbf{v}}_k - \mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*; \boldsymbol{x})^\dagger \mathbf{v}_k| \le |\widehat{\mathbf{v}}_k^\top (\nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}}; \boldsymbol{x})^\dagger - \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*; \boldsymbol{x})^\dagger) \widehat{\mathbf{v}}_k| \\
&\quad + |(\widehat{\mathbf{v}}_k - \mathbf{v}_k) \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*; \boldsymbol{x})^\dagger \mathbf{v}_k| + |(\widehat{\mathbf{v}}_k - \mathbf{v}_k) \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*; \boldsymbol{x})^\dagger \widehat{\mathbf{v}}_k| \\
&\quad = O_P\left(\|\nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}}; \boldsymbol{x})^\dagger - \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*; \boldsymbol{x})^\dagger\|_2 \|\mathbf{P}_{\mathbf{1}}^{\perp} \mathbf{v}_k\|_2^2 + \|\mathbf{P}_{\mathbf{1}}^{\perp} \mathbf{v}_k\|_2 \|\nabla^2 \mathcal{L}(\boldsymbol{\theta}^*; \boldsymbol{x})^\dagger\|_2 \|\widehat{\mathbf{v}}_k - \mathbf{v}_k\|_2 \right) \\
&\quad \lesssim \frac{n \|\mathbf{P}_{\mathbf{1}}^{\perp} \mathbf{v}_k\|_2^2}{\underline{\lambda}} \sqrt{\frac{\kappa_{\mathcal{E}}^3 \log n}{\underline{\lambda} L}} + \frac{\|\mathbf{P}_{\mathbf{1}}^{\perp} \mathbf{v}_k\|_2 \|\mathbf{v}_k\|_\infty}{\underline{\lambda}} \sqrt{\frac{n \kappa_{\mathcal{E}} \log n}{\underline{\lambda} L}}.
\end{aligned}$$

Recall from the proof of Theorem 3 that $\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})^\dagger \mathbf{v}_k \gtrsim \|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2^2/\bar{\lambda}$, then under the condition that $\kappa_r \lesssim \sqrt{n}$ and $n^2\kappa_{\mathcal{E}}^5 \log n/(\underline{\lambda} L) = o(1)$, we have that with high probability

$$
\begin{aligned}
&\max_k \left\{ \frac{|\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})^\dagger \widehat{\mathbf{v}}_k - \mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})^\dagger \mathbf{v}_k|}{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})^\dagger \mathbf{v}_k} \right\} \\
&\lesssim \max_k \left\{ \frac{n\|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2^2}{\underline{\lambda}} \sqrt{\frac{\kappa_{\mathcal{E}}^3 \log n}{\underline{\lambda} L}} + \frac{\|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2 \|\mathbf{v}_k\|_\infty}{\underline{\lambda}} \sqrt{\frac{n\kappa_{\mathcal{E}} \log n}{\underline{\lambda} L}} \right\} \frac{\bar{\lambda}}{\|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2^2} \\
&= \max_k \kappa_{\mathcal{E}} \cdot \left\{ n\sqrt{\frac{\kappa_{\mathcal{E}}^3 \log n}{\underline{\lambda} L}} + \frac{\|\mathbf{v}_k\|_\infty}{\|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2} \sqrt{\frac{n\kappa_{\mathcal{E}} \log n}{\underline{\lambda} L}} \right\} \lesssim n\kappa_{\mathcal{E}}^2 \sqrt{\frac{\kappa_{\mathcal{E}} \log n}{\underline{\lambda} L}} + \kappa_r \kappa_{\mathcal{E}} \sqrt{\frac{n\kappa_{\mathcal{E}} \log n}{\underline{\lambda} L}} \\
&\lesssim n\kappa_{\mathcal{E}}^2 \sqrt{\frac{\kappa_{\mathcal{E}} \log n}{\underline{\lambda} L}} = o(1),
\end{aligned}
\tag{EC.13}
$$

and by Slutsky's theorem, the claim follows.

EC.5. Proof of Lemma 2

In the following proof, we denote $\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x})$ by $\mathcal{L}(\boldsymbol{\theta})$ for the convenience of notation. When $\boldsymbol{\mathcal{E}}$ is collected by a random scheme such that Assumption 1 holds with high probability (e.g., the setting of Lemma EC.7), establishing (21) reduces to showing that, with high probability over $\boldsymbol{\mathcal{E}}$, the following event holds.

$$
\mathcal{A}_{\boldsymbol{\mathcal{E}}} := \left\{ \sup_{\boldsymbol{\theta}^* \in \boldsymbol{\Theta}} \sup_{\alpha \in (0,1)} |\mathbb{P}_{\boldsymbol{\theta}^*} (T > c_W(\alpha, \boldsymbol{\mathcal{E}}) \mid \boldsymbol{\mathcal{E}}) - \alpha| = o(1) \right\}. \tag{EC.14}
$$

In the remainder of the proof, we work in this more general setting where Assumption 1 holds with high probability with respect to the randomness of $\boldsymbol{\mathcal{E}}$. For the convenience of notation, for two positive sequences x_n and y_n, we use $x_n \lesssim y_n$ or $x_n = O(y_n)$ to imply that there exists a constant $C > 0$ independent of $\boldsymbol{\theta}^*$ such that $x_n \le C y_n$ for large enough n.

For a given $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$, we define the following two auxiliary statistics,

$$
T_0 := \max_{k \in [n]} \left\{ \frac{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger}{\sqrt{L \mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k}} \sum_{\mathcal{S} \in \boldsymbol{\mathcal{E}}} \sum_{\ell=1}^{L_{\mathcal{S}}} \left\{ \sum_{i \in \mathcal{S}_+} \left(x_{\mathcal{S}}^{(i,\ell)} - \frac{e^{\theta_i^*}}{\sum_{j \in \mathcal{S}_+} e^{\theta_j^*}} \right) \boldsymbol{e}_i \right\} \right\}, \tag{EC.15}
$$

and

$$
W_0 := \max_{k \in [n]} \sum_{\ell=1}^{L} \left\{ \frac{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger}{\sqrt{L \mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k}} \sum_{\mathcal{S} \in \boldsymbol{\mathcal{E}}} \sum_{\ell=1}^{L_{\mathcal{S}}} \left\{ \sum_{i \in \mathcal{S}_+} \left(x_{\mathcal{S}}^{(i,\ell)} - \frac{e^{\theta_i^*}}{\sum_{j \in \mathcal{S}_+} e^{\theta_j^*}} \right) \boldsymbol{e}_i \right\} z_{\mathcal{S},\ell} \right\}. \tag{EC.16}
$$

We define $y_k^{(\ell,\mathcal{S})} = \sqrt{\frac{1}{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k}} \mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \left\{ \sum_{i\in\mathcal{S}_+} \left(x_{\mathcal{S}}^{(i,\ell)} - \frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}} \right) \boldsymbol{e}_i \right\}$, for $\mathcal{S} \in \boldsymbol{\mathcal{E}}$, $\ell = 1,\ldots,L_{\mathcal{S}}$ and $k = 1,\ldots,n$. Then $W_0 = \max_{k\in[n]} \frac{1}{\sqrt{L}} \sum_{\mathcal{S}\in\boldsymbol{\mathcal{E}}} \sum_{\ell=1}^{L_{\mathcal{S}}} y_k^{(\ell,\mathcal{S})} \cdot z_{\mathcal{S},\ell}$. We will show that the following three conditions hold with high probability with respect to the randomness of $\boldsymbol{\mathcal{E}}$ uniformly for all $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$.

(1) There exist ζ_1 and ζ_2 independent of $\boldsymbol{\theta}^*$ with $\zeta_1\sqrt{\log n} + \zeta_2 = o(1)$ such that $\mathbb{P}(|T - T_0| > \zeta_1) < \zeta_2$, $\mathbb{P}(\mathbb{P}(|W - W_0| > \zeta_1 \,|\, \boldsymbol{x}) > \zeta_2) < \zeta_2$,

(2) $c \leqslant \min_{k\in[n]} \frac{1}{L} \sum_{S\in\boldsymbol{\mathcal{E}}} \sum_{\ell=1}^{L_{\mathcal{S}}} \mathbb{E}[(y_k^{(\ell,\mathcal{S})})^2]$ for some constant $c > 0$,

(3) $\max_{k\in[n]} |y_k^{(\ell,\mathcal{S})}| \lesssim B$ with $B^2(\log(nL))^7/L = o(1)$ where B is not necessarily a constant.

We first verify condition (2). For any $k \in [n]$, since $y_k^{(\ell,\mathcal{S})}$'s are independent conditional on $\mathcal{E}$, we have

$$\begin{aligned}\frac{1}{L}\sum_{\mathcal{S}\in\boldsymbol{\mathcal{E}}}\sum_{\ell=1}^{L_{\mathcal{S}}} \mathbb{E}\big[(y_k^{(\ell,\mathcal{S})})^2\big] &= \frac{1}{L\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k} \mathrm{Var}\left[\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \sum_{\mathcal{S}\in\boldsymbol{\mathcal{E}}} \sum_{\ell=1}^{L_{\mathcal{S}}} \sum_{i\in\mathcal{S}_+} \left(x_{\mathcal{S}}^{(i,\ell)} - \frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}} \right) \boldsymbol{e}_i \right] \\ &= \frac{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k}{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k} = 1,\end{aligned}$$

and thus condition (2) holds trivially. As for condition (3), with high probability with respect to the randomness of $\boldsymbol{\mathcal{E}}$, we have that

$$\begin{aligned}\left| y_k^{(\ell,\mathcal{S})} \right| &= \sqrt{\frac{1}{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k}} \left| \mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \left\{ \sum_{i\in\mathcal{S}_+} \left(x_{\mathcal{S}}^{(i,\ell)} - \frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}} \right) \boldsymbol{e}_i \right\} \right| \\ &= \sqrt{\frac{1}{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k}} \left| (\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k)^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \left\{ \sum_{i\in\mathcal{S}_+} x_{\mathcal{S}}^{(i,\ell)} \boldsymbol{e}_i - \sum_{i\in\mathcal{S}_+} \frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}} \boldsymbol{e}_i \right\} \right| \\ &\lesssim \frac{1}{\|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2/\sqrt{\bar{\lambda}}} \|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2 \left\| \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \right\|_2 \left\| \sum_{i\in\mathcal{S}_+} x_{\mathcal{S}}^{(i,\ell)} \boldsymbol{e}_i - \sum_{i\in\mathcal{S}_+} \frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}} \boldsymbol{e}_i \right\|_2 \\ &\lesssim \sqrt{\kappa_{\boldsymbol{\mathcal{E}}}/\underline{\lambda}}.\end{aligned}$$

Thus we can take $B = \sqrt{\kappa_{\boldsymbol{\mathcal{E}}}/\underline{\lambda}}$. Note that

$$n\bar{\lambda} = n\|\mathbf{L}_{\boldsymbol{\mathcal{E}}}\|_2 \geq n \cdot (\mathbf{L}_{\boldsymbol{\mathcal{E}}})_{1,1} \gtrsim n \cdot \sum_{\mathcal{S}\in\boldsymbol{\mathcal{E}}} \frac{L_{\mathcal{S}}}{L \cdot |\mathcal{S}_+|} \gtrsim \sum_{\mathcal{S}\in\boldsymbol{\mathcal{E}}} \frac{L_{\mathcal{S}}}{L} = 1,$$

where the second inequality is due to the fact that $0 \in \mathcal{S}_+$ for any $\mathcal{E}$ by definition. Then condition (3) holds uniformly for $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ with high probability regarding $\mathcal{E}$ under the condition that $\kappa_{\mathcal{E}}^2 \log^7 n/(\underline{\lambda} L) = o(1)$.

Now we move on to verify condition (1). First recall from previous proof that the following hold with probability at least $1 - O(n^{-10})$,

$$\|\widehat{\boldsymbol{\theta}}^d - \boldsymbol{\theta}^*\|_2 \lesssim \sqrt{\frac{\kappa_{\mathcal{E}} n \log n}{\underline{\lambda} L}}, \quad \|\widehat{\mathbf{v}}_k - \mathbf{v}_k\|_2 \lesssim \|\mathbf{v}_k\|_\infty \sqrt{\frac{n\kappa_{\mathcal{E}} \log n}{\underline{\lambda} L}},$$
$$\|\nabla \mathcal{L}(\widehat{\boldsymbol{\theta}}) - \nabla \mathcal{L}(\boldsymbol{\theta}^*)\|_\infty \lesssim \sqrt{\frac{\kappa_{\mathcal{E}}^2 n \bar{\lambda} \log n}{L}}, \quad \|\nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger - \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger\|_2 \lesssim \frac{n}{\underline{\lambda}} \sqrt{\frac{\kappa_{\mathcal{E}}^3 \log n}{\underline{\lambda} L}},$$
$$\left|\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_k - \mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k\right| \lesssim \frac{n\|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2^2}{\underline{\lambda}} \sqrt{\frac{\kappa_{\mathcal{E}}^3 \log n}{\underline{\lambda} L}} + \frac{\|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2 \|\mathbf{v}_k\|_\infty}{\underline{\lambda}} \sqrt{\frac{n\kappa_{\mathcal{E}} \log n}{\underline{\lambda} L}}.$$

And based on the above upper bounds we have

$$\left|\frac{1}{\sqrt{\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_k}} - \frac{1}{\sqrt{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k}}\right| \lesssim \frac{1}{\sqrt{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k}} \left|\frac{\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_k - \mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k}{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k}\right|$$
$$\lesssim \frac{1}{\sqrt{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k}} \left(n\sqrt{\frac{\kappa_{\mathcal{E}}^3 \log n}{\underline{\lambda} L}} + \frac{\|\mathbf{v}_k\|_\infty}{\|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2} \sqrt{\frac{n\kappa_{\mathcal{E}} \log n}{\underline{\lambda} L}}\right),$$
$$\|\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger - \mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger\|_2 \lesssim \|(\widehat{\mathbf{v}}_k - \mathbf{v}_k)^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger\|_2 + \|\mathbf{v}_k^\top (\nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger - \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger)\|_2$$
$$\lesssim \frac{\|\mathbf{v}_k\|_\infty}{\underline{\lambda}} \sqrt{\frac{n\kappa_{\mathcal{E}} \log n}{\underline{\lambda} L}} + \frac{n \cdot \|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2}{\underline{\lambda}} \sqrt{\frac{\kappa_{\mathcal{E}}^3 \log n}{\underline{\lambda} L}},$$

and in turn, we have

$$\begin{aligned}\left\|\frac{\widehat{\mathbf{v}}_k^\top\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger}{\sqrt{\widehat{\mathbf{v}}_k^\top\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger\widehat{\mathbf{v}}_k}}-\frac{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger}{\sqrt{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\mathbf{v}_k}}\right\|_2 &\lesssim \frac{\|\widehat{\mathbf{v}}_k^\top\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger-\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\|_2}{\sqrt{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\mathbf{v}_k}}\\ &+\left|\frac{\|\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\|_2}{\sqrt{\widehat{\mathbf{v}}_k^\top\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger\widehat{\mathbf{v}}_k}}-\frac{\|\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\|_2}{\sqrt{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\mathbf{v}_k}}\right|\\ &\lesssim \frac{1}{\sqrt{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\mathbf{v}_k}}\left(\frac{\|\mathbf{v}_k\|_\infty}{\underline{\lambda}}\sqrt{\frac{n\kappa_{\mathcal{E}}\log n}{\underline{\lambda}L}}+\frac{n\cdot\|\mathbf{P}_{\mathbf{1}}^\perp\mathbf{v}_k\|_2}{\underline{\lambda}}\sqrt{\frac{\kappa_{\mathcal{E}}^3\log n}{\underline{\lambda}L}}\right).\end{aligned}$$

Now we consider

$$\begin{aligned}W-W_0\le\max_{k\in[n]}\sum_{S\in\boldsymbol{\mathcal{E}}}\sum_{\ell=1}^{L_S}\sum_{i\in\mathcal{S}_+}\Bigg\{&\frac{\widehat{\mathbf{v}}_k^\top\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger}{\sqrt{L\widehat{\mathbf{v}}_k^\top\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger\widehat{\mathbf{v}}_k}}\left(x_S^{(i,\ell)}-\frac{e^{\widehat{\theta}_i}}{\sum_{j\in\mathcal{S}_+}e^{\widehat{\theta}_j}}\right)\\ &-\frac{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger}{\sqrt{L\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\mathbf{v}_k}}\left(x_S^{(i,\ell)}-\frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+}e^{\theta_j^*}}\right)\Bigg\}\boldsymbol{e}_i z_{S,\ell},\end{aligned}$$

where the RHS is the maximum of n centered Gaussian random variables conditioning on the data $\boldsymbol{x}$ and the comparison graph $\mathcal{E}$. We have the following upper bound on the conditional variance

$$\begin{aligned}&\max_{k\in[n]}\frac{1}{L}\sum_{S\in\boldsymbol{\mathcal{E}}}\sum_{\ell=1}^{L_S}\left\{\sum_{i\in\mathcal{S}_+}\left\{\frac{\widehat{\mathbf{v}}_k^\top\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger}{\sqrt{\widehat{\mathbf{v}}_k^\top\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger\widehat{\mathbf{v}}_k}}\left(x_S^{(i,\ell)}-\frac{e^{\widehat{\theta}_i}}{\sum_{j\in\mathcal{S}_+}e^{\widehat{\theta}_j}}\right)-\frac{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger}{\sqrt{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\mathbf{v}_k}}\left(x_S^{(i,\ell)}-\frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+}e^{\theta_j^*}}\right)\right\}\boldsymbol{e}_i\right\}^2\\ &\lesssim\underbrace{\max_{k\in[n]}\frac{1}{L}\sum_{S\in\boldsymbol{\mathcal{E}}}\sum_{\ell=1}^{L_S}\left\{\sum_{i\in\mathcal{S}_+}\left\{\frac{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger}{\sqrt{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\mathbf{v}_k}}\left(\frac{e^{\widehat{\theta}_i}}{\sum_{j\in\mathcal{S}_+}e^{\widehat{\theta}_j}}-\frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+}e^{\theta_j^*}}\right)\right\}\boldsymbol{e}_i\right\}^2}_{\mathrm{I}}\\ &+\underbrace{\max_{k\in[n]}\frac{1}{L}\sum_{S\in\boldsymbol{\mathcal{E}}}\sum_{\ell=1}^{L_S}\left\{\sum_{i\in\mathcal{S}_+}\left\{\left(\frac{\widehat{\mathbf{v}}_k^\top\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger}{\sqrt{\widehat{\mathbf{v}}_k^\top\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger\widehat{\mathbf{v}}_k}}-\frac{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger}{\sqrt{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\mathbf{v}_k}}\right)\left(x_S^{(i,\ell)}-\frac{e^{\widehat{\theta}_i}}{\sum_{j\in\mathcal{S}_+}e^{\widehat{\theta}_j}}\right)\right\}\boldsymbol{e}_i\right\}^2}_{\mathrm{II}}.\end{aligned}$$

For term I, recall from previous proof that by the mean theorem, with probability at least $1-O(n^{-10})$, for all $S\in\mathcal{E}$ and $i\in S$ we have that,

$$\left|\frac{e^{\widehat{\theta}_i}}{\sum_{j\in\mathcal{S}_+}e^{\widehat{\theta}_j}}-\frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+}e^{\theta_j^*}}\right|\lesssim\frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+}e^{\theta_j^*}}\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_\infty\lesssim\frac{1}{|\mathcal{S}_+|}\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_\infty,$$

and thus with probability at least $1-O(n^{-10})$, we have

$$\begin{aligned}
\mathrm{I}&=\max_{k\in[n]}\frac{1}{L}\sum_{S\in\mathcal{E}}\sum_{\ell=1}^{L_S}\left\{\sum_{i\in\mathcal{S}_+}\left\{\frac{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger}{\sqrt{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\mathbf{v}_k}}\left(\frac{e^{\widehat{\theta}_i}}{\sum_{j\in\mathcal{S}_+}e^{\widehat{\theta}_j}}-\frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+}e^{\theta_j^*}}\right)\right\}\boldsymbol{e}_i\right\}^2\\
&\lesssim\max_{k\in[n]}\frac{1}{L}\sum_{S\in\mathcal{E}}\sum_{\ell=1}^{L_S}\frac{1}{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\mathbf{v}_k}\left\{(\mathbf{P}_{\mathbf{1}}^\perp\mathbf{v}_k)^\top\sum_{i\in\mathcal{S}_+}\left\{\left(\frac{e^{\widehat{\theta}_i}}{\sum_{j\in\mathcal{S}_+}e^{\widehat{\theta}_j}}-\frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+}e^{\theta_j^*}}\right)\left[\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\right]_i\right\}\right\}^2\\
&\lesssim\max_{k\in[n]}\frac{1}{L}\sum_{S\in\mathcal{E}}\sum_{\ell=1}^{L_S}\frac{1}{\|\mathbf{P}_{\mathbf{1}}^\perp\mathbf{v}_k\|_2^2/\bar{\lambda}}\|\mathbf{P}_{\mathbf{1}}^\perp\mathbf{v}_k\|_2^2\|\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\|_2^2\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_\infty^2\\
&\lesssim\frac{\kappa_{\mathcal{E}}\cdot\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_\infty^2}{\underline{\lambda}}\cdot n\bar{\lambda}\lesssim\frac{\kappa_{\mathcal{E}}^3n^2\log n}{\underline{\lambda}L},
\end{aligned}$$

where the last but one inequality follows from

$$\sum_{S\in\mathcal{E}}\frac{L_S}{L}\lesssim n\cdot\sum_{S\in\mathcal{E}}\frac{L_S}{L\cdot|\mathcal{S}_+|}\lesssim n\cdot(\mathbf{L}_{\mathcal{E}})_{1,1}\le n\cdot\|\mathbf{L}_{\mathcal{E}}\|_2=n\bar{\lambda},\tag{EC.17}$$

and for term II we have

$$\begin{aligned}
\mathrm{II}&=\max_{k\in[n]}\frac{1}{L}\sum_{S\in\mathcal{E}}\sum_{\ell=1}^{L_S}\left\{\sum_{i\in\mathcal{S}_+}\left\{\left(\frac{\widehat{\mathbf{v}}_k^\top\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger}{\sqrt{\widehat{\mathbf{v}}_k^\top\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger\widehat{\mathbf{v}}_k}}-\frac{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger}{\sqrt{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\mathbf{v}_k}}\right)\left(x_S^{(i,\ell)}-\frac{e^{\widehat{\theta}_i}}{\sum_{j\in\mathcal{S}_+}e^{\widehat{\theta}_j}}\right)\right\}\boldsymbol{e}_i\right\}^2\\
&\le\max_{k\in[n]}\frac{1}{L}\sum_{S\in\mathcal{E}}\sum_{\ell=1}^{L_S}\left\|\frac{\widehat{\mathbf{v}}_k^\top\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger}{\sqrt{\widehat{\mathbf{v}}_k^\top\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger\widehat{\mathbf{v}}_k}}-\frac{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger}{\sqrt{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\mathbf{v}_k}}\right\|_2^2\left\|\sum_{i\in\mathcal{S}_+}\left(x_S^{(i,\ell)}-\frac{e^{\widehat{\theta}_i}}{\sum_{j\in\mathcal{S}_+}e^{\widehat{\theta}_j}}\right)\boldsymbol{e}_i\right\|_2^2\\
&\lesssim\max_{k\in[n]}\frac{n\bar{\lambda}}{\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\mathbf{v}_k}\left(\frac{\|\mathbf{v}_k\|_\infty}{\underline{\lambda}}\sqrt{\frac{n\kappa_{\mathcal{E}}\log n}{\underline{\lambda}L}}+\frac{n\cdot\|\mathbf{P}_{\mathbf{1}}^\perp\mathbf{v}_k\|_2}{\underline{\lambda}}\sqrt{\frac{\kappa_{\mathcal{E}}^3\log n}{\underline{\lambda}L}}\right)^2\\
&\lesssim n\max_{k\in[n]}\left(\frac{\|\mathbf{v}_k\|_\infty}{\|\mathbf{P}_{\mathbf{1}}^\perp\mathbf{v}_k\|_2}\sqrt{\frac{n\kappa_{\mathcal{E}}^3\log n}{\underline{\lambda}L}}+n\kappa_{\mathcal{E}}\sqrt{\frac{\kappa_{\mathcal{E}}^3\log n}{\underline{\lambda}L}}\right)^2\lesssim\frac{n^3\kappa_{\mathcal{E}}^5\log n}{\underline{\lambda}L},
\end{aligned}$$

where the last but one inequality follows from (EC.17). Denote $\epsilon_0 = \frac{n^3\kappa_{\mathcal{E}}^5 \log n}{\underline{\lambda} L}$, and define the event $E_{\mathcal{E}}$ to be

$$E_{\mathcal{E}} = \Bigg\{ \max_{k\in[n]} \frac{1}{L} \sum_{S\in\boldsymbol{\mathcal{S}}} \sum_{\ell=1}^{L_S} \Bigg\{ \sum_{i\in\mathcal{S}_+} \Bigg\{ \frac{\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger}{\sqrt{\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_k}} \Big(x_S^{(i,\ell)} - \frac{e^{\widehat{\theta}_i}}{\sum_{j\in\mathcal{S}_+} e^{\widehat{\theta}_j}} \Big) - \frac{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger}{\sqrt{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k}} \Big(x_S^{(i,\ell)} - \frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}} \Big) \Bigg\} \boldsymbol{e}_i \Bigg\}^2 \le C_0\epsilon_0 \Bigg\},$$

where the constant C_0 is chosen properly such that $\sup_{\boldsymbol{\theta}^*\in\boldsymbol{\Theta}} \mathbb{P}_{\boldsymbol{\theta}^*}((E_{\mathcal{E}})^{\mathrm{c}}) \le n^{-10}$. By maximal inequality we have that under the event $E_{\mathcal{E}}$, there exists a large enough constant $C > 0$ such that

$$\mathbb{P}_{\boldsymbol{\theta}^*}\Big(W - W_0 \ge C\sqrt{\epsilon_0 \log n}\,|\,\boldsymbol{x},\mathcal{E}\Big) \le \mathbb{P}_{\boldsymbol{\theta}^*}\Bigg(\max_{k\in[n]} \sum_{S\in\boldsymbol{\mathcal{S}}} \sum_{\ell=1}^{L_S} \sum_{i\in\mathcal{S}_+} \Bigg\{ \frac{\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger}{\sqrt{L\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_k}} \Big(x_S^{(i,\ell)} - \frac{e^{\widehat{\theta}_i}}{\sum_{j\in\mathcal{S}_+} e^{\widehat{\theta}_j}} \Big) - \frac{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger}{\sqrt{L\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k}} \Big(x_S^{(i,\ell)} - \frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}} \Big) \Bigg\} \boldsymbol{e}_i z_{S,\ell} \ge C\sqrt{\epsilon_0 \log n}\,|\,\boldsymbol{x},\mathcal{E}\Bigg) \le n^{-10}/2.$$

Very similarly, under the event $\mathbb{E}_{\mathcal{E}}$, we can also show that

$$\mathbb{P}_{\boldsymbol{\theta}^*}\Big(W_0 - W \ge C\sqrt{\epsilon_0 \log n}\,|\,\boldsymbol{x},\mathcal{E}\Big) \le n^{-10}/2,$$

and in turn

$$\mathbb{P}_{\boldsymbol{\theta}^*}\Big(|W - W_0| \ge C\sqrt{\epsilon_0 \log n}\,|\,\boldsymbol{x},\mathcal{E}\Big) \le n^{-10}.$$

Combining the above results we have

$$\sup_{\boldsymbol{\theta}^*\in\boldsymbol{\Theta}} \mathbb{P}_{\boldsymbol{\theta}^*}\Big(\mathbb{P}_{\boldsymbol{\theta}^*}\Big(|W - W_0| \ge C\sqrt{\epsilon_0 \log n}\,|\,\boldsymbol{x},\mathcal{E}\Big) > n^{-10}\Big) \le \sup_{\boldsymbol{\theta}^*\in\boldsymbol{\Theta}} \mathbb{P}_{\boldsymbol{\theta}^*}((E_{\mathcal{E}})^{\mathrm{c}}) \le n^{-10}.$$

Now we consider $|T - T_0|$. Under the condition that $\kappa_r \lesssim \sqrt{n}$, with probability at least $1 - O(n^{-10})$, we have that

$$
\begin{aligned}
|T - T_0| &\lesssim \max_{k\in[n]} \left\{ (\widehat{\Delta}_k - \Delta_k) \left| \frac{\sqrt{L}}{\sqrt{\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_k}} - \frac{\sqrt{L}}{\sqrt{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k}} \right| + \frac{\|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2 \|\mathbf{R}_0\|_2 + \|\mathbf{v}_k\|_\infty \|\widehat{\boldsymbol{\theta}}^d - \boldsymbol{\theta}^*\|_2^2}{\sqrt{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k / L}} \right\} \\
&\lesssim \max_{k\in[n]} \left\{ \|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2 \|\widehat{\boldsymbol{\theta}}^d - \boldsymbol{\theta}^*\|_2 \left| \frac{\sqrt{L}}{\sqrt{\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_k}} - \frac{\sqrt{L}}{\sqrt{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k}} \right| + \frac{\|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2 \|\mathbf{R}_0\|_2 + \|\mathbf{v}_k\|_\infty \|\widehat{\boldsymbol{\theta}}^d - \boldsymbol{\theta}^*\|_2^2}{\sqrt{\mathbf{v}_k^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_k / L}} \right\} \\
&\lesssim \max_{k\in[n]} \left\{ \sqrt{\frac{\kappa_{\mathcal{E}}^2 n \log n}{L}} \left(n \sqrt{\frac{\kappa_{\mathcal{E}}^3 \log n}{\underline{\lambda}}} + \frac{\|\mathbf{v}_k\|_\infty}{\|\mathbf{P}_{\mathbf{1}}^\perp \mathbf{v}_k\|_2} \sqrt{\frac{n \kappa_{\mathcal{E}} \log n}{\underline{\lambda}}} \right) + \kappa_{\mathcal{E}}^3 n \log n \sqrt{\frac{\kappa_{\mathcal{E}} \cdot n}{\underline{\lambda} L}} \right\} \\
&\leq \kappa_{\mathcal{E}}^3 n \log n \sqrt{\frac{\kappa_{\mathcal{E}} \cdot n}{\underline{\lambda} L}}.
\end{aligned}
$$

Thus if we take $\zeta_1 = C_1 \kappa_{\mathcal{E}}^3 n \log n \sqrt{\frac{\kappa_{\mathcal{E}} \cdot n}{\underline{\lambda} L}}$ and $\zeta_2 = C_2 n^{-10}$ with large enough constants C_1 and C_2 independent of $\boldsymbol{\theta}^*$, under the condition that $\kappa_{\mathcal{E}}^3 n \log^{3/2} n \sqrt{\frac{\kappa_{\mathcal{E}} \cdot n}{\underline{\lambda} L}} = o(1)$, condition (1) also holds, and by Corollary 3.1 in Chernozhukov et al. (2013) we have that with high probability with respect to $\mathcal{E}$ the event (EC.14) holds.

Then by Jensen's inequality and the dominated convergence theorem, we have

$$
\begin{aligned}
&\lim_{n,L\to\infty} \sup_{\boldsymbol{\theta}^*\in\boldsymbol{\Theta}} \sup_{\alpha\in(0,1)} |\mathbb{P}_{\boldsymbol{\theta}^*}(T > c_W(\alpha, \mathcal{E})) - \alpha| \\
&= \lim_{n,L\to\infty} \sup_{\boldsymbol{\theta}^*\in\boldsymbol{\Theta}} \sup_{\alpha\in(0,1)} \left| \int \mathbb{P}_{\boldsymbol{\theta}^*}(T > c_W(\alpha, \mathcal{E}) \mid \mathcal{E}) \, \mathrm{d}\mu(\mathcal{E}) - \alpha \right| \\
&= \lim_{n,L\to\infty} \sup_{\boldsymbol{\theta}^*\in\boldsymbol{\Theta}} \sup_{\alpha\in(0,1)} \left| \int (\mathbb{P}_{\boldsymbol{\theta}^*}(T > c_W(\alpha, \mathcal{E}) \mid \mathcal{E}) - \alpha) \, \mathrm{d}\mu(\mathcal{E}) \right| \\
&\leq \lim_{n,L\to\infty} \int \sup_{\boldsymbol{\theta}^*\in\boldsymbol{\Theta}} \sup_{\alpha\in(0,1)} |\mathbb{P}_{\boldsymbol{\theta}^*}(T > c_W(\alpha, \mathcal{E}) \mid \mathcal{E}) - \alpha| \, \mathrm{d}\mu(\mathcal{E}) \\
&= \lim_{n,L\to\infty} \int_{\mathcal{A}_{\mathcal{E}}} \sup_{\boldsymbol{\theta}^*\in\boldsymbol{\Theta}} \sup_{\alpha\in(0,1)} |\mathbb{P}_{\boldsymbol{\theta}^*}(T > c_W(\alpha, \mathcal{E}) \mid \mathcal{E}) - \alpha| \, \mathrm{d}\mu(\mathcal{E}) \\
&\quad + \lim_{n,L\to\infty} \int_{\mathcal{A}_{\mathcal{E}}^c} \sup_{\boldsymbol{\theta}^*\in\boldsymbol{\Theta}} \sup_{\alpha\in(0,1)} |\mathbb{P}_{\boldsymbol{\theta}^*}(T > c_W(\alpha, \mathcal{E}) \mid \mathcal{E}) - \alpha| \, \mathrm{d}\mu(\mathcal{E}) \\
&\leq \int_{\mathcal{A}_{\mathcal{E}}} \lim_{n,L\to\infty} \sup_{\boldsymbol{\theta}^*\in\boldsymbol{\Theta}} \sup_{\alpha\in(0,1)} |\mathbb{P}_{\boldsymbol{\theta}^*}(T > c_W(\alpha, \mathcal{E}) \mid \mathcal{E}) - \alpha| \, \mathrm{d}\mu(\mathcal{E}) \\
&\quad + \lim_{n,L\to\infty} \int_{\mathcal{A}_{\mathcal{E}}^c} \mathrm{d}\mu(\mathcal{E}) = 0.
\end{aligned}
$$

Thus $\sup_{\boldsymbol{\theta}^*\in\boldsymbol{\Theta}} \sup_{\alpha\in(0,1)} |\mathbb{P}_{\boldsymbol{\theta}^*}(T > c_W(\alpha, \mathcal{E})) - \alpha| = o(1)$ as $n, L \to \infty$.

EC.6. Proof of Theorem 4

We will first show that $[\widehat{K}_L, \widehat{K}_U]$ is a valid confidence interval for K^* based on Lemma 2. Note that by definition, if $\widehat{K}_U < n$, we have that $\widehat{\Delta}_{\widehat{K}_U+1} > c_W(\alpha/2, \boldsymbol{\mathcal{E}})\sqrt{\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_k / L}$. Then we have

$$
\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}^*}(\widehat{K}_L \le K^* \le \widehat{K}_U) \ge 1 - \mathbb{P}_{\boldsymbol{\theta}^*}(\widehat{K}_L > K^*) - \mathbb{P}_{\boldsymbol{\theta}^*}(\widehat{K}_U < K^*) = 1 - \mathbb{P}_{\boldsymbol{\theta}^*}(\widehat{K}_L > K^*) - \mathbb{P}_{\boldsymbol{\theta}^*}(\widehat{K}_U + 1 \le K^*) \\
&= 1 - \mathbb{P}_{\boldsymbol{\theta}^*}(\Delta_{\widehat{K}_L} \ge 0) - \mathbb{P}_{\boldsymbol{\theta}^*}\left(\left\{\Delta_{\widehat{K}_U+1} < 0\right\} \cap \{\widehat{K}_U < n\}\right) \\
&= 1 - \mathbb{P}_{\boldsymbol{\theta}^*}\left(\{\Delta_{\widehat{K}_L} \ge 0\} \cap \left\{\widehat{\Delta}_{\widehat{K}_L} < -c_W(\alpha/2, \boldsymbol{\mathcal{E}})\sqrt{\widehat{\mathbf{v}}_{\widehat{K}_L}^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_{\widehat{K}_L} / L}\right\}\right) \\
&\quad - \mathbb{P}_{\boldsymbol{\theta}^*}\left(\{\Delta_{\widehat{K}_U+1} < 0\} \cap \left\{\widehat{K}_U < n,\, \widehat{\Delta}_{\widehat{K}_U+1} > c_W(\alpha/2, \boldsymbol{\mathcal{E}})\sqrt{\widehat{\mathbf{v}}_{\widehat{K}_U+1}^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_{\widehat{K}_U+1} / L}\right\}\right) \\
&\ge 1 - \mathbb{P}_{\boldsymbol{\theta}^*}\left(\min_{k > K^*} \frac{\widehat{\Delta}_k - \Delta_k}{\sqrt{\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_k / L}} \le -c_W(\alpha/2, \boldsymbol{\mathcal{E}})\right) - \mathbb{P}_{\boldsymbol{\theta}^*}\left(\max_{k \le K^*} \frac{\widehat{\Delta}_k - \Delta_k}{\sqrt{\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_k / L}} \ge c_W(\alpha/2, \boldsymbol{\mathcal{E}})\right) \\
&\ge 1 - \mathbb{P}_{\boldsymbol{\theta}^*}\left(\min_{k \in [n]} \frac{\widehat{\Delta}_k - \Delta_k}{\sqrt{\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_k / L}} \le -c_W(\alpha/2, \boldsymbol{\mathcal{E}})\right) - \mathbb{P}_{\boldsymbol{\theta}^*}\left(\max_{k \in [n]} \frac{\widehat{\Delta}_k - \Delta_k}{\sqrt{\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_k / L}} \ge c_W(\alpha/2, \boldsymbol{\mathcal{E}})\right).
\end{aligned}
$$

By Lemma 2, we know that

$$
\lim_{n,L\to\infty} \sup_{\boldsymbol{\theta}^* \in \boldsymbol{\Theta}} \mathbb{P}_{\boldsymbol{\theta}^*}\left(\max_{k \in [n]} \frac{\widehat{\Delta}_k - \Delta_k}{\sqrt{\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_k / L}} \ge c_W(\alpha/2, \boldsymbol{\mathcal{E}})\right) = \frac{\alpha}{2}.
$$

Also, following similar proof and the symmetry of Gaussian distribution we have

$$
\begin{aligned}
&\lim_{n,L\to\infty} \sup_{\boldsymbol{\theta}^* \in \boldsymbol{\Theta}} \mathbb{P}_{\boldsymbol{\theta}^*}\left(\min_{k \in [n]} \frac{\widehat{\Delta}_k - \Delta_k}{\sqrt{\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_k / L}} \le -c_W(\alpha/2, \boldsymbol{\mathcal{E}})\right) \\
&= \lim_{n,L\to\infty} \sup_{\boldsymbol{\theta}^* \in \boldsymbol{\Theta}} \mathbb{P}_{\boldsymbol{\theta}^*}\left(\max_{k \in [n]} -\frac{\widehat{\Delta}_k - \Delta_k}{\sqrt{\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_k / L}} \ge c_W(\alpha/2, \boldsymbol{\mathcal{E}})\right) = \frac{\alpha}{2}.
\end{aligned}
$$

Thus by taking limits and the supremum on both sides of the inequality, we have that (19) follows.

Now we move on to show (20). By the equivalence of (1) and (5), we have that

$$\sup_{\mathcal{S}^*\in\mathcal{S}_0}\mathbb{P}_{\boldsymbol{\theta}^*}(\text{Reject } H_0)=\sup_{\boldsymbol{\theta}^*:K^*\in\mathcal{K}_0}\mathbb{P}_{\boldsymbol{\theta}^*}(\text{Reject } H_0)=\sup_{\boldsymbol{\theta}^*:K^*\in\mathcal{K}_0}\mathbb{P}_{\boldsymbol{\theta}^*}([\widehat{K}_L,\widehat{K}_U]\cap\mathcal{K}_0=\varnothing)$$
$$=\sup_{\boldsymbol{\theta}^*:K^*\in\mathcal{K}_0}\mathbb{P}_{\boldsymbol{\theta}^*}\Big(\{K^*\in\mathcal{K}_0\}\cap\big\{[\widehat{K}_L,\widehat{K}_U]\cap\mathcal{K}_0=\varnothing\big\}\Big)\le\sup_{\boldsymbol{\theta}^*:K^*\in\mathcal{K}_0}\mathbb{P}_{\boldsymbol{\theta}^*}\Big(\{K^*\in\mathcal{K}_0\}\cap\{K^*\notin[\widehat{K}_L,\widehat{K}_U]\}\Big)$$
$$\le 1-\inf_{\boldsymbol{\theta}^*:K^*\in\mathcal{K}_0}\mathbb{P}_{\boldsymbol{\theta}^*}(K^*\in[\widehat{K}_L,\widehat{K}_U])\le 1-\inf_{\boldsymbol{\theta}^*\in\boldsymbol{\Theta}}\mathbb{P}_{\boldsymbol{\theta}^*}(K^*\in[\widehat{K}_L,\widehat{K}_U]).$$

Then from (19), by taking limits on both sides we have that (20) holds.

EC.7. Proof of Theorem 5

Similar as in the proof of Lemma 2, we work in this more general setting where $\boldsymbol{\mathcal{E}}$ is allowed to be random and Assumption 1 holds with high probability with respect to the randomness of $\boldsymbol{\mathcal{E}}$.

We will show the result in (1) first. We begin by establishing an upper bound for T and $c_W(\alpha/2,\boldsymbol{\mathcal{E}})$ with high probability. Conditional on $\boldsymbol{\mathcal{E}}$, define the matrix $\boldsymbol{\Sigma}_\Delta\in\mathbb{R}^{n\times n}$ as follows:

$$[\boldsymbol{\Sigma}_\Delta]_{jk}=\frac{\mathbf{v}_j^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\mathbf{v}_k}{\sqrt{\left(\mathbf{v}_j^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\mathbf{v}_j\right)\cdot\left(\mathbf{v}_k^\top\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^\dagger\mathbf{v}_k\right)}},\quad j,k\in[n],$$

and let $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_n)^\top\in\mathbb{R}^n$ be a multivariate Gaussian random vector conditional on $\boldsymbol{\mathcal{E}}$, such that $\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_\Delta)|\boldsymbol{\mathcal{E}}$.

By the verification of conditions (1)-(3) in the proof of Lemma 2 and applying Theorem 2.2 in Chernozhukov et al. (2013), with probability at least $1-o(1)$ with respect to the randomness of $\boldsymbol{\mathcal{E}}$, we have:

$$\max\left\{\sup_{t\in\mathbb{R}}\left|\mathbb{P}(T\le t|\boldsymbol{\mathcal{E}})-\mathbb{P}(\max_{j\in[n]}\xi_j\le t|\boldsymbol{\mathcal{E}})\right|,\ \sup_{t\in\mathbb{R}}\left|\mathbb{P}(W\le t|\boldsymbol{\mathcal{E}})-\mathbb{P}(\max_{j\in[n]}\xi_j\le t|\boldsymbol{\mathcal{E}})\right|\right\}\le\rho_n,\quad\text{(EC.18)}$$

where $\rho_n=o(1)$ is a rate independent of $\boldsymbol{\mathcal{E}}$. Applying the maximal inequality for $\boldsymbol{\xi}$ and taking $t=2\sqrt{\log n}$, we have that $\mathbb{P}(\max_{j\in[n]}\xi_j\le 2\sqrt{\log n}|\boldsymbol{\mathcal{E}})\le n^{-1}$. Then applying the law of total probability and (EC.18), with probability at least $1-o(1)$, we have that $\max(T,W)\le 2\sqrt{\log n}$, which also gives $c_W(\alpha/2,\boldsymbol{\mathcal{E}})\le 2\sqrt{\log n}$ with high probability.

Next, we use the upper bounds for T and $c_W(\alpha/2,\boldsymbol{\mathcal{E}})$ to construct an upper bound for the Type II error. For any $\boldsymbol{\theta}^*\in\boldsymbol{\Theta}$ such that $\mathcal{S}^*\notin\boldsymbol{\mathcal{S}}_0$ and $\delta(\boldsymbol{\theta}^*,\mathcal{K}_0)\ge C\delta_n$, we have that

$$\mathbb{P}_{\boldsymbol{\theta}^*}(\text{Reject H}_0)=\mathbb{P}_{\boldsymbol{\theta}^*}(\mathcal{K}_0\cap[\widehat{K}_L,\widehat{K}_U]=\varnothing)\ge\mathbb{P}_{\boldsymbol{\theta}^*}\left(\left\{\mathcal{K}_0^{\mathrm{L}}<\widehat{K}_L\right\}\cap\left\{\mathcal{K}_0^{\mathrm{R}}>\widehat{K}_U\right\}\right).\quad\text{(EC.19)}$$

We first consider the event $\{\mathcal{K}_0^{\mathrm{L}} < \widehat{K}_L\}$. Since $\mathbb{P}_{\boldsymbol{\theta}^*}(\mathcal{K}_0^{\mathrm{L}} < \widehat{K}_L) = 1$ when $\mathcal{K}_0^{\mathrm{L}} = -1$, we focus on the case where $\mathcal{K}_0^{\mathrm{L}} \geq 1$. By the definition of $\widehat{K}_L$, if the following holds:

$$\widehat{\Delta}_{\mathcal{K}_0^{\mathrm{L}}+1} < -c_W\big(\frac{\alpha}{2}, \boldsymbol{\mathcal{E}}\big)\sqrt{\frac{\widehat{\mathbf{v}}_{\mathcal{K}_0^{\mathrm{L}}+1}^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_{\mathcal{K}_0^{\mathrm{L}}+1}}{L}},$$

then we have

$$\widehat{K}_L = \max\left\{k : \widehat{\Delta}_k < -c_W\big(\frac{\alpha}{2}, \boldsymbol{\mathcal{E}}\big)\sqrt{\frac{\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_k}{L}}\right\} \geq \mathcal{K}_0^{\mathrm{L}} + 1 > \mathcal{K}_0^{\mathrm{L}}.$$

This indicates that

$$\begin{aligned}
\mathbb{P}(\mathcal{K}_0^{\mathrm{L}} \geq \widehat{K}_L) &\leq \mathbb{P}\left(\widehat{\Delta}_{\mathcal{K}_0^{\mathrm{L}}+1} \geq -c_W\big(\frac{\alpha}{2}, \boldsymbol{\mathcal{E}}\big)\sqrt{\frac{\widehat{\mathbf{v}}_{\mathcal{K}_0^{\mathrm{L}}+1}^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_{\mathcal{K}_0^{\mathrm{L}}+1}}{L}}\right) \\
&= \mathbb{P}\left(\Delta_{\mathcal{K}_0^{\mathrm{L}}+1} + \widehat{\Delta}_{\mathcal{K}_0^{\mathrm{L}}+1} - \Delta_{\mathcal{K}_0^{\mathrm{L}}+1} \geq -c_W\big(\frac{\alpha}{2}, \boldsymbol{\mathcal{E}}\big)\sqrt{\frac{\widehat{\mathbf{v}}_{\mathcal{K}_0^{\mathrm{L}}+1}^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_{\mathcal{K}_0^{\mathrm{L}}+1}}{L}}\right) \\
&\leq \mathbb{P}\left(\Delta_{\mathcal{K}_0^{\mathrm{L}}+1} \geq -\Big(c_W\big(\frac{\alpha}{2}, \boldsymbol{\mathcal{E}}\big) + T\Big)\sqrt{\frac{\widehat{\mathbf{v}}_{\mathcal{K}_0^{\mathrm{L}}+1}^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_{\mathcal{K}_0^{\mathrm{L}}+1}}{L}}\right) \\
&\leq \mathbb{P}\left(\Delta_{\mathcal{K}_0^{\mathrm{L}}+1} \geq -4\sqrt{\log n}\cdot\sqrt{\frac{\widehat{\mathbf{v}}_{\mathcal{K}_0^{\mathrm{L}}+1}^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_{\mathcal{K}_0^{\mathrm{L}}+1}}{L}}\right) + o(1) \\
&\leq \mathbb{P}\left(\Delta_{\mathcal{K}_0^{\mathrm{L}}+1} \geq -5\sqrt{\log n}\cdot\sqrt{\frac{\mathbf{v}_{\mathcal{K}_0^{\mathrm{L}}+1}^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_{\mathcal{K}_0^{\mathrm{L}}+1}}{L}}\right) + o(1),
\end{aligned}$$

where the last inequality follows from (EC.13) under the condition that $\kappa_r \lesssim \sqrt{n}$ and $n^2\kappa_{\boldsymbol{\mathcal{E}}}^5 \log n/(\underline{\lambda} L) = o(1)$. Furthermore, when $\sup_{\boldsymbol{\theta}^* \in \boldsymbol{\Theta}} \kappa_{\boldsymbol{\theta}^*} = O(1)$, when Assumption 1 holds, we have that $\sup_{\boldsymbol{\theta}^* \in \boldsymbol{\Theta}} \|\nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger\|_2 \lesssim \underline{\lambda}^{-1}$ and $\|\mathbf{v}_k\|_2^2 \asymp \sum_{j=0}^{k-1}(r_j - r_k)^2$ for all $k \in [n]$. Hence there exists a properly large constant $C' > 0$ such that as long as $C > C' > 0$, we have

$$\sup_{\boldsymbol{\theta}^* : \mathcal{S}^* \notin \boldsymbol{\mathcal{S}}_0, \delta(\boldsymbol{\theta}^*, \mathcal{K}_0) \geq C\delta_n} \mathbb{P}(\mathcal{K}_0^{\mathrm{L}} \geq \widehat{K}_L) \leq \mathbb{P}\left(\Big[\sum_{j=0}^{\mathcal{K}_0^{\mathrm{L}}}(r_j - r_{\mathcal{K}_0^{\mathrm{L}}+1})^2\Big]^{-1/2}\Delta_{\mathcal{K}_0^{\mathrm{L}}+1} \geq -C'\sqrt{\frac{\log n}{\underline{\lambda} L}}\right) + o(1) = o(1).$$

Next, we consider the event $\{\mathcal{K}_0^{\mathrm{R}} > \widehat{K}_U\}$. When $\mathcal{K}_0^{\mathrm{R}} = n+1$, we have $\mathbb{P}_{\boldsymbol{\theta}^*}(\mathcal{K}_0^{\mathrm{R}} > \widehat{K}_U) = 1$, and we will focus on when $\mathcal{K}_0^{\mathrm{R}} \le n$. By the definition of $\widehat{K}_U$, if the following holds:

$$\widehat{\Delta}_{\mathcal{K}_0^{\mathrm{R}}} > c_W\big(\frac{\alpha}{2}, \mathcal{E}\big)\sqrt{\frac{\widehat{\mathbf{v}}_{\mathcal{K}_0^{\mathrm{R}}}^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_{\mathcal{K}_0^{\mathrm{R}}}}{L}},$$

then we will have

$$\widehat{K}_U = \min\left\{k : \widehat{\Delta}_k > c_W\big(\frac{\alpha}{2}, \mathcal{E}\big)\sqrt{\frac{\widehat{\mathbf{v}}_k^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_k}{L}}\right\} - 1 \le \mathcal{K}_0^{\mathrm{R}} - 1 < \mathcal{K}_0^{\mathrm{R}},$$

which in turn indicates that

$$\mathbb{P}(\mathcal{K}_0^{\mathrm{R}} \le \widehat{K}_U) \le \mathbb{P}\left(\widehat{\Delta}_{\mathcal{K}_0^{\mathrm{R}}} \le c_W\big(\frac{\alpha}{2}, \mathcal{E}\big)\sqrt{\frac{\widehat{\mathbf{v}}_{\mathcal{K}_0^{\mathrm{R}}}^\top \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})^\dagger \widehat{\mathbf{v}}_{\mathcal{K}_0^{\mathrm{R}}}}{L}}\right).$$

Then following similar arguments regarding the event $\{\mathcal{K}_0^{\mathrm{L}} < \widehat{K}_L\}$, we have that $\sup_{\boldsymbol{\theta}^*:\mathcal{S}^* \notin \boldsymbol{\mathcal{S}}_0, \delta(\boldsymbol{\theta}^*, \mathcal{K}_0) \ge C\delta_n} \mathbb{P}(\mathcal{K}_0^{\mathrm{R}} \le \widehat{K}_U) = o(1)$. Combining with (EC.19), we have that

$$\inf_{\boldsymbol{\theta}^*:\mathcal{S}^* \notin \boldsymbol{\mathcal{S}}_0, \delta(\boldsymbol{\theta}^*, \mathcal{K}_0) \ge C\delta_n} \mathbb{P}_{\boldsymbol{\theta}^*}(\text{Reject } \mathrm{H}_0) \to 1.$$

Now we move on to show the results in (2). By Theorem 1, the optimal assortment estimate $\widehat{\mathcal{S}}$ is of the form $\widehat{\mathcal{S}} = \{1, 2, \ldots, \widehat{K}\}$, where $\widehat{K} = \max\{k \in [n] : \widehat{\Delta}_k < 0\}$. Also, it can be checked that $\widehat{\Delta}_k$ is also non-decreasing with respect to $k \in [n]$. Hence we have the following equivalence

$$\mathbb{P}(\widehat{\mathcal{S}} = \mathcal{S}^*) = \mathbb{P}(\widehat{K} = K^*) = \mathbb{P}\big(\{\widehat{\Delta}_{K^*} < 0\} \cap \{\widehat{\Delta}_{K^*+1} \ge 0\}\big).$$

By Theorem 3, we have that

$$\begin{aligned}
\mathbb{P}(\widehat{\Delta}_{K^*} < 0) &= \mathbb{P}\left\{\sqrt{\frac{L}{\mathbf{v}_{K^*}^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_{K^*}}}\left(\widehat{\Delta}_{K^*} - \Delta_{K^*}\right) < -\Delta_{K^*}\sqrt{\frac{L}{\mathbf{v}_{K^*}^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_{K^*}}}\right\} \\
&= \Phi\left(-\Delta_{K^*}\sqrt{\frac{L}{\mathbf{v}_{K^*}^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_{K^*}}}\right) + o(1), \\
\mathbb{P}(\widehat{\Delta}_{K^*+1} \ge 0) &= \mathbb{P}\left\{\sqrt{\frac{L}{\mathbf{v}_{K^*+1}^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_{K^*+1}}}\left(\widehat{\Delta}_{K^*+1} - \Delta_{K^*+1}\right) \ge -\Delta_{K^*+1}\sqrt{\frac{L}{\mathbf{v}_{K^*+1}^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_{K^*+1}}}\right\} \\
&= \Phi\left(\Delta_{K^*+1}\sqrt{\frac{L}{\mathbf{v}_{K^*+1}^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)^\dagger \mathbf{v}_{K^*+1}}}\right) + o(1),
\end{aligned}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Under Assumption 1, for all $k \in [n]$, we have that

$$\bar{\lambda}^{-1}\|\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{v}_k\|_2^2 \lesssim \mathbf{v}_k^{\top}\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^{\dagger}\mathbf{v}_k \lesssim \underline{\lambda}^{-1}\sum_{j=0}^{k-1}(r_j - r_k)^2, \tag{EC.20}$$

where $\mathbf{P}_{\mathbf{1}}^{\perp}$ is the projection matrix to the perpendicular space of $\mathbf{1}$. Furthermore,

$$\begin{aligned}\|\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{v}_k\|_2^2 = \|\mathbf{v}_k\|_2^2 - \|\mathbf{P}_{\mathbf{1}}\mathbf{v}_k\|_2^2 &= \sum_{j=1}^{k} v_{kj}^2 - \frac{1}{n+1}\Big(\sum_{j=1}^{k} v_{kj}\Big)^2 \\ &\geq \sum_{j=1}^{k} v_{kj}^2 - \frac{\underline{k}+1}{n+1}\Big(\sum_{j=1}^{k} v_{kj}^2\Big) \gtrsim \Big(1 - \frac{\underline{k}+1}{n+1}\Big)\sum_{j=0}^{k-1}(r_j - r_k)^2,\end{aligned} \tag{EC.21}$$

where v_{kj} denotes the j-th entry of $\mathbf{v}_k$, $\underline{k} := |\{j \in [n] : r_j > r_k\}|$ is the number of items with revenues larger than r_k, and the last but one inequality follows from Cauchy–Schwarz inequality. Combining (EC.20) and (EC.21) gives

$$\bar{\lambda}^{-1}\Big(1 - \frac{\underline{k}+1}{n+1}\Big)\sum_{j=0}^{k-1}(r_j - r_k)^2 \lesssim \mathbf{v}_k^{\top}\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)^{\dagger}\mathbf{v}_k \lesssim \underline{\lambda}^{-1}\sum_{j=0}^{k-1}(r_j - r_k)^2, \quad \text{for all } k \in [n].$$

Therefore, when $\min_{k\in\{K^*,K^*+1\}}[\sum_{j=0}^{k-1}(r_j - r_k)^2]^{-1/2}|\Delta_k| \geq C\delta_n$ for some large enough $C > 0$, by Chernoff bound, we have that

$$\mathbb{P}(\widehat{\mathcal{S}} = \mathcal{S}^*) \geq 1 - \mathbb{P}\big(\widehat{\Delta}_{K^*} \geq 0\big) - \mathbb{P}\big(\widehat{\Delta}_{K^*+1} < 0\big) \geq 1 - O(n^{-1}) - o(1) = 1 - o(1).$$

On the other hand, if $\min_{k\in\{K^*,K^*+1\}}[\sum_{j=0}^{k-1}(r_j - r_k)^2]^{-1/2}|\Delta_k| \leq c(\kappa_{\mathcal{E}}\log n/[1 - (\underline{K}+2)/(n+1)])^{-1/2}\delta_n$ for some small enough constant $c > 0$, where $\underline{K} := |\{j \in [n] : r_j > r_{K^*}\}|$, we have that

$$\mathbb{P}(\widehat{\mathcal{S}} \neq \mathcal{S}^*) \geq \max\{\mathbb{P}\big(\widehat{\Delta}_{K^*} \geq 0\big), \mathbb{P}\big(\widehat{\Delta}_{K^*+1} < 0\big)\} \geq 1/3.$$

EC.8. Analogous Result to the Bradley–Terry–Luce (BTL) Model

Lemma EC.7 establishes conditions under which Assumption 1 holds when the sets $\mathcal{S} \in \mathcal{E}$ are sampled independently. We emphasize that this random-sampling assumption is not required, and we include it solely to facilitate comparison of optimal scaling rates with prior work.

LEMMA EC.7. *Let $\mathcal{S} = \{\mathcal{S} \subseteq [n] : \mathcal{S} \neq \emptyset\}$ denote the family of all nonempty offer sets. Suppose each $\mathcal{S} \in \mathcal{S}$ is included in $\mathcal{E}$ independently with probability $p_{\mathcal{S}}$ and yields sample size $L_{\mathcal{S}}$. Let $\underline{p} := \min_{\mathcal{S}\in\mathcal{S}} p_{\mathcal{S}}$ and assume there exists a constant $C \geq 1$ such that*

$$\max_{i\in[n]_+} \frac{\sum_{\mathcal{S}\in\mathcal{S}} \frac{p_{\mathcal{S}}}{\underline{p}} |\mathcal{S}_+|^{-1} \mathbb{I}\{i \in \mathcal{S}_+\}}{\sum_{\mathcal{S}\in\mathcal{S}} |\mathcal{S}_+|^{-1} \mathbb{I}\{i \in \mathcal{S}_+\}} \leq C, \qquad 2^n \underline{p} \geq C\, n \log n. \tag{EC.22}$$

Define

$$M_n = \sum_{k=2}^{n+1} \frac{1}{k^2} \binom{n-2}{k-3} \asymp \frac{2^n}{n^2}, \qquad N_n = \sum_{k=2}^{n+1} \frac{1}{k^2} \binom{n-1}{k-2} \asymp \frac{2^{n+1}}{n^2}.$$

Let $\underline{L} := \min_{\mathcal{S}\in\mathcal{S}} L_{\mathcal{S}}$ and $\bar{L} := \max_{\mathcal{S}\in\mathcal{S}} L_{\mathcal{S}}$. Then there exist absolute constants $C_0 \geq c_0 > 0$ and $C_1 > 0$ such that, with probability at least $1 - O(n^{-10})$,

$$c_0 \frac{\underline{L}}{L} (nM_n + N_n)\, \underline{p} \leq \underline{\lambda} \leq \bar{\lambda} \leq C_0 \frac{\bar{L}}{L} nN_n\, \underline{p}, \quad \kappa_{\mathcal{E}} \leq C_1 \frac{\bar{L}}{\underline{L}}. \tag{EC.23}$$

REMARK EC.1. The proof of Lemma EC.7 is deferred to Section EC.9.2. The lower bound on $\underline{p}$ in (EC.22) guarantees that the comparison graph is connected with high probability. To understand the heterogeneity condition in (EC.22), consider the simple sufficient case where $\max_{\mathcal{S}\in\mathcal{S}} p_{\mathcal{S}} \leq C\underline{p}$ for some $C \geq 1$, then the first inequality in (EC.22) holds trivially. Moreover, consider a more heterogeneous sufficient case where for each $i \in [n]_+$, there exists a subset $\widetilde{\mathcal{S}}_i \subset \mathcal{S}_i$ with cardinality $|\widetilde{\mathcal{S}}_i| \leq n$, and suppose that $\max_{\mathcal{S}\in\widetilde{\mathcal{S}}_i} p_{\mathcal{S}}/\underline{p} \leq 2^n/n^2$ and $\max_{\mathcal{S}\in\mathcal{S}_i\setminus\widetilde{\mathcal{S}}_i} p_{\mathcal{S}}/\underline{p} \leq C'$ for some constant $C' \geq 1$. Then, because $\sum_{\mathcal{S}\in\mathcal{S}_i} (|\mathcal{S}|+1)^{-1} \asymp 2^n/n$ for all $i \in [n]_+$, we have that

$$\begin{aligned}\max_{i\in[n]_+} \sum_{\mathcal{S}\in\mathcal{S}_i} \frac{(|\mathcal{S}|+1)^{-1}}{\sum_{\mathcal{S}'\in\mathcal{S}_i}(|\mathcal{S}'|+1)^{-1}} \cdot \frac{p_{\mathcal{S}}}{\underline{p}} &\leq \max_{i\in[n]_+} \frac{\sum_{\mathcal{S}\in\widetilde{\mathcal{S}}_i}(|\mathcal{S}|+1)^{-1} \cdot \frac{p_{\mathcal{S}}}{\underline{p}} + \sum_{\mathcal{S}\in\mathcal{S}_i\setminus\widetilde{\mathcal{S}}_i}(|\mathcal{S}|+1)^{-1} \cdot \frac{p_{\mathcal{S}}}{\underline{p}}}{\sum_{\mathcal{S}\in\mathcal{S}_i}(|\mathcal{S}|+1)^{-1}} \\ &\lesssim \frac{2^n}{n^2} \cdot \frac{n}{2^n/n} + C' \lesssim C' + 1,\end{aligned}$$

and (EC.22) holds. Hence, the heterogeneity condition in (EC.22) allows for $p_{\mathcal{S}}$ to differ by up to an exponential order of n for a small proportion of sets $\mathcal{S} \in \mathcal{S}_i$, which can be useful in practical scenarios where certain assortments are offered much more frequently than others.

EC.9. Proof of Technical Lemmas

EC.9.1. Proof of Lemma 1

The matrix $\mathbf{L}_{\mathcal{E}}$ can be rewritten as follows,

$$\mathbf{L}_{\mathcal{E}} = \sum_{k=2}^{n+1} \frac{m_{k-1}}{k^2} \operatorname{diag}\left(k\mathbf{I}_k - \mathbf{1}_k\mathbf{1}_k^\top, \mathbf{0}_{n+1-k}\right),$$

where $\mathbf{0}_m$ denotes the $m \times m$ zero matrix. It is immediate that $\mathbf{L}_{\mathcal{E}}\mathbf{1} = \mathbf{0}$ and $\lambda_{n+1}(\mathbf{L}_{\mathcal{E}}) = 0$.

We show $\underline{\lambda} \asymp m_n/n$ first. Recall $\lambda_{\min,\perp}$ is the smallest eigenvalue restricted to the eigenvectors orthogonal to $\mathbf{1}$. Then $\underline{\lambda} = \lambda_n(\mathbf{L}_{\mathcal{E}}) = \lambda_{\min,\perp}(\mathbf{L}_{\mathcal{E}})$. For any $\mathbf{v} \in \mathbb{R}^{n+1}$ such that $\|\mathbf{v}\|_2 = 1$ and $\mathbf{v}^\top\mathbf{1} = 0$, the following holds

$$\mathbf{v}^\top\mathbf{L}_{\mathcal{E}}\mathbf{v} \geq \frac{m_n}{(n+1)^2} \cdot \mathbf{v}^\top\Big((n+1)\cdot\mathbf{I}_{n+1} - \mathbf{1}_{n+1}\mathbf{1}_{n+1}^\top\Big)\mathbf{v} = \frac{m_n}{(n+1)^2}\cdot(n+1) \gtrsim m_n/n,$$

which implies that $\underline{\lambda} \gtrsim m_n/n$. On the other hand, taking $\mathbf{v}_n = \boldsymbol{e}_{n+1} - \frac{1}{n+1}\mathbf{1}_{n+1}$, we have that $\mathbf{v}_n^\top\mathbf{1} = 0$ and that $\mathbf{v}_n^\top\mathbf{L}_{\mathcal{E}}\mathbf{v}_n = \boldsymbol{e}_{n+1}^\top\mathbf{L}_{\mathcal{E}}\boldsymbol{e}_{n+1} = m_n/(n+1)$. Hence $\underline{\lambda} = \lambda_{\min,\perp}(\mathbf{L}_{\mathcal{E}}) \lesssim \mathbf{v}_n^\top\mathbf{L}_{\mathcal{E}}\mathbf{v}_n = m_n/(n+1)$, which in turn gives that $\underline{\lambda} \asymp m_n/n$

Next, we show that $\bar{\lambda} \asymp \sum_{k=1}^n m_k/k$. We have that the following holds

$$\sum_{k=1}^n m_k/k \lesssim \boldsymbol{e}_1^\top\mathbf{L}_{\mathcal{E}}\boldsymbol{e}_1 \leq \bar{\lambda} = \lambda_1(\mathbf{L}_{\mathcal{E}}) \leq \sum_{k=2}^{n+1}\frac{m_{k-1}}{k^2}\|k\mathbf{I}_k - \mathbf{1}_k\mathbf{1}_k^\top\|_2 \lesssim \sum_{k=1}^n m_k/k,$$

which gives $\bar{\lambda} \asymp \sum_{k=1}^n m_k/k$.

When $\sum_{k=1}^{n-1} m_k/k \lesssim n_n/n$, $\bar{\lambda} \asymp \underline{\lambda}$ follows trivially, and the proof is complete.

EC.9.2. Proof of Lemma EC.7

Let $\mathbf{A}_{\mathcal{S}} = \mathcal{E}_{\mathcal{S}} \cdot \frac{1}{|\mathcal{S}_+|^2}\sum_{\substack{i,j\in\mathcal{S}_+\\ i<j}}(\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top$, where $\mathcal{E}_{\mathcal{S}} = \mathbb{I}\{\mathcal{S} \in \boldsymbol{\mathcal{E}}\}$, then we have that

$$\frac{\underline{L}}{L}\sum_{\mathcal{S}\in\boldsymbol{\mathcal{S}}}\mathbf{A}_{\mathcal{S}} \preceq \mathbf{L}_{\mathcal{E}} \preceq \frac{\bar{L}}{L}\sum_{\mathcal{S}\in\boldsymbol{\mathcal{S}}}\mathbf{A}_{\mathcal{S}}. \tag{EC.24}$$

Denote $\widetilde{\mathbf{L}}_{\mathcal{E}} = \sum_{\mathcal{S}\in\mathcal{S}} \mathbf{A}_{\mathcal{S}}$. Recall $M_n = \sum_{k=2}^{n+1} \frac{1}{k^2}\binom{n-2}{k-3}$ and $N_n = \sum_{k=2}^{n+1} \frac{1}{k^2}\binom{n-1}{k-2}$, and we have that

$$\begin{aligned}
\mathbb{E}(\widetilde{\mathbf{L}}_{\mathcal{E}}) &= \underline{p}\sum_{\mathcal{S}\in\mathcal{S}} \frac{p_{\mathcal{S}}}{\underline{p}} \cdot \frac{1}{|\mathcal{S}_+|^2} \sum_{\substack{i,j\in\mathcal{S}_+\\ i<j}} (\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \succeq \underline{p}\begin{pmatrix} nN_n & -N_n\mathbf{1}_n^\top \\ -N_n\mathbf{1}_n & (nM_n+N_n)\mathbf{I}_n - M_n\mathbf{1}_n\mathbf{1}_n^\top \end{pmatrix} \\
&= \underline{p}\left((n+1)M_n\mathbf{I}_{n+1} - M_n\mathbf{1}_{n+1}\mathbf{1}_{n+1}^\top + (N_n - M_n)\begin{pmatrix} n & -\mathbf{1}_n^\top \\ -\mathbf{1}_n & \mathbf{I}_n \end{pmatrix}\right) \\
&= \underline{p}\left((n+1)M_n\mathbf{I}_{n+1} - M_n\mathbf{1}_{n+1}\mathbf{1}_{n+1}^\top + (N_n - M_n)\left(\mathbf{I}_{n+1} + \mathrm{diag}(n, \mathbf{1}_n\mathbf{1}_n^\top) - \mathbf{1}_{n+1}\mathbf{1}_{n+1}^\top\right)\right) \\
&= \underline{p}\left((nM_n + N_n)\mathbf{I}_{n+1} + (N_n - M_n)\,\mathrm{diag}(n, \mathbf{1}_n\mathbf{1}_n^\top) - N_n\mathbf{1}_{n+1}\mathbf{1}_{n+1}^\top\right).
\end{aligned}$$

Consider the matrix $(N_n - M_n)\,\mathrm{diag}(n, \mathbf{1}_n\mathbf{1}_n^\top) - N_n\mathbf{1}_{n+1}\mathbf{1}_{n+1}^\top$. Define $\mathbf{w}_1 = (1, \mathbf{0}^\top)^\top$, $\mathbf{w}_2 = (0, \mathbf{1}_n^\top/\sqrt{n})^\top$ and $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2)$. Then it can be seen that $\mathbf{1}_{n+1} = \mathbf{w}_1 + \sqrt{n}\mathbf{w}_2$ and we have that

$$\begin{aligned}
&(N_n - M_n)\,\mathrm{diag}(n, \mathbf{1}_n\mathbf{1}_n^\top) - N_n\mathbf{1}_{n+1}\mathbf{1}_{n+1}^\top = n(N_n - M_n)\mathbf{W}\mathbf{W}^\top - \mathbf{W}\begin{pmatrix} N_n & \sqrt{n}N_n \\ \sqrt{n}N_n & N_n \end{pmatrix}\mathbf{W}^\top \\
&= \mathbf{W}\begin{pmatrix} (n-1)N_n - nM_n & -\sqrt{n}N_n \\ -\sqrt{n}N_n & -nM_n \end{pmatrix}\mathbf{W}^\top = \mathbf{W}\mathbf{G}\,\mathrm{diag}\left(n(N_n - M_n), -nM_n - N_n\right)\mathbf{G}^\top\mathbf{W}^\top,
\end{aligned}$$

where $\mathbf{G} = \frac{1}{\sqrt{n+1}}\begin{pmatrix} \sqrt{n} & 1 \\ -1 & \sqrt{n} \end{pmatrix}$ and the last equality follows from the following eigendecomposition:

$$\begin{pmatrix} (n-1)N_n - nM_n & -\sqrt{n}N_n \\ -\sqrt{n}N_n & -nM_n \end{pmatrix} = \mathbf{G}\,\mathrm{diag}\left(n(N_n - M_n), -nM_n - N_n\right)\mathbf{G}^\top.$$

Then combining the above results we have that

$$\mathbb{E}(\widetilde{\mathbf{L}}_{\mathcal{E}}) \succeq \underline{p}\left((nM_n + N_n)\mathbf{I}_{n+1} + \mathbf{W}\mathbf{G}\,\mathrm{diag}\left(n(N_n - M_n), -nM_n - N_n\right)\mathbf{G}^\top\mathbf{W}^\top\right), \tag{EC.25}$$

which indicates that $\lambda_n\big(\mathbb{E}(\widetilde{\mathbf{L}}_{\mathcal{E}})\big) \geq (nM_n + N_n)\underline{p}$. Also since $\mathbb{E}(\widetilde{\mathbf{L}}_{\mathcal{E}})\mathbf{1}_{n+1} = \mathbf{0}$, we have that $\mathbf{1}_{n+1}$ corresponds to the $(n+1)$-th eigenvalue of $\mathbb{E}(\widetilde{\mathbf{L}}_{\mathcal{E}})$. Moreover, we have the following upper bound holds,

$$\begin{aligned}
\mathbb{E}(\widetilde{\mathbf{L}}_{\mathcal{E}}) &= \underline{p}\sum_{\mathcal{S}\in\mathcal{S}} \frac{p_{\mathcal{S}}}{\underline{p}} \cdot \frac{1}{|\mathcal{S}_+|^2} \sum_{\substack{i,j\in\mathcal{S}_+\\ i<j}} (\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \\
&\preceq \underline{p}\sum_{\mathcal{S}\in\mathcal{S}} \frac{p_{\mathcal{S}}}{\underline{p}} \cdot |\mathcal{S}_+|^{-1}\mathbf{I}_{\mathcal{S}_+} = \underline{p}\cdot\mathrm{diag}(s_0, s_1, \ldots, s_n),
\end{aligned} \tag{EC.26}$$

where $\mathbf{I}_{\mathcal{S}_+}$ is a diagonal matrix with $(\mathbf{I}_{\mathcal{S}_+})_{ii} = 1$ for $i \in \mathcal{S}_+$ and $(\mathbf{I}_{\mathcal{S}_+})_{ii} = 0$ for $i \notin \mathcal{S}_+$, and $s_i = \sum_{\mathcal{S}\in\mathcal{S}} \frac{p_{\mathcal{S}}}{\underline{p}} \cdot |\mathcal{S}_+|^{-1}\mathbb{I}\{i \in \mathcal{S}_+\}$ for $i \in [n]_+$; the last inequality follows from the fact that for any $\mathcal{S} \in \mathcal{S}$ we have

$$\sum_{\substack{i,j\in\mathcal{S}_+\\ i<j}} (\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top = |\mathcal{S}_+| \cdot \mathbf{I}_{\mathcal{S}_+} - \mathbf{1}_{\mathcal{S}_+}\mathbf{1}_{\mathcal{S}_+}^\top \preceq |\mathcal{S}_+| \cdot \mathbf{I}_{\mathcal{S}_+},$$

where $\mathbf{1}_{\mathcal{S}_+} \in \mathbb{R}^{n+1}$ with $(\mathbf{1}_{\mathcal{S}_+})_i = 1$ for $i \in \mathcal{S}_+$ and $(\mathbf{1}_{\mathcal{S}_+})_i = 0$ for $i \notin \mathcal{S}_+$. When the condition (EC.22) holds, we have that $s_i \le C\sum_{\mathcal{S}\in\mathcal{S}} |\mathcal{S}_+|^{-1}\mathbb{I}\{i \in \mathcal{S}_+\} \lesssim (n+1)N_n \asymp 2^n/n$ for any $i \in [n]_+$, and in turn we have $\lambda_1\big(\mathbb{E}(\widetilde{\mathbf{L}}_{\mathcal{E}})\big) \lesssim (n+1)N_n\underline{p}$. Hence

$$(nM_n + N_n)\underline{p} \le \lambda_n\big(\mathbb{E}(\widetilde{\mathbf{L}}_{\mathcal{E}})\big) \le \lambda_1\big(\mathbb{E}(\widetilde{\mathbf{L}}_{\mathcal{E}})\big) \lesssim (n+1)N_n\underline{p}.$$

Let $\mathbf{R}_{\mathcal{E}} \in \mathbb{R}^{(n+1)\times n}$ denote the stacking of the top n normalized eigenvectors of $\mathbb{E}(\widetilde{\mathbf{L}}_{\mathcal{E}})$, which is unique up to rotation, and we have $\mathbf{R}_{\mathcal{E}}^\top\mathbf{R}_{\mathcal{E}} = \mathbf{I}_n$ and $\mathbf{R}_{\mathcal{E}}^\top\mathbf{1}_{n+1} = \mathbf{0}$. Then from (EC.25) and (EC.26), it can be seen that

$$\operatorname{diag}\big((n+1)N_n\underline{p}, (nM_n + N_n)\underline{p}\mathbf{I}_{n-1}\big) \preceq \mathbb{E}(\mathbf{R}_{\mathcal{E}}^\top\widetilde{\mathbf{L}}_{\mathcal{E}}\mathbf{R}_{\mathcal{E}}) = \mathbf{R}_{\mathcal{E}}^\top\mathbb{E}(\widetilde{\mathbf{L}}_{\mathcal{E}})\mathbf{R}_{\mathcal{E}} \preceq C'nN_n\underline{p} \cdot \mathbf{I}_n,$$

where $C' > 0$ is a large enough constant. We have that

$$\lambda_{\min}(\mathbf{A}_{\mathcal{S}}) \ge 0, \quad \lambda_{\max}(\mathbf{A}_{\mathcal{S}}) \le |\mathcal{S}_+|/|\mathcal{S}_+|^2 = 1/|\mathcal{S}_+| \le 1,$$

and

$$\lambda_{\min}(\mathbf{R}_{\mathcal{E}}^\top\mathbf{A}_{\mathcal{S}}\mathbf{R}_{\mathcal{E}}) \ge 0, \quad \lambda_{\max}(\mathbf{R}_{\mathcal{E}}^\top\mathbf{A}_{\mathcal{S}}\mathbf{R}_{\mathcal{E}}) \le \lambda_{\max}(\mathbf{A}_{\mathcal{S}}) \le 1.$$

Then by Theorem 5.1.1 in Tropp (2015), under the condition that $2^n\underline{p} \ge Cn\log n$ for some large enough constant $C > 0$, there exists a constant $c > 0$ such that for large enough n we have

$$\begin{aligned}\mathbb{P}\left(\lambda_{\min,\perp}(\widetilde{\mathbf{L}}_{\mathcal{E}}) \le (nM_n + N_n)\underline{p}/2\right) &\le \mathbb{P}\left(\lambda_{\min}(\mathbf{R}_{\mathcal{E}}^\top\widetilde{\mathbf{L}}_{\mathcal{E}}\mathbf{R}_{\mathcal{E}}) \le \lambda_{\min}\left(\mathbb{E}(\mathbf{R}_{\mathcal{E}}^\top\widetilde{\mathbf{L}}_{\mathcal{E}}\mathbf{R}_{\mathcal{E}})\right)/2\right)\\ &\le n\exp(-cnM_n\underline{p}) \le n^{-10},\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}\left(\lambda_{\max}(\widetilde{\mathbf{L}}_{\mathcal{E}}) \ge 3C'nN_n\underline{p}/2\right) &\le \mathbb{P}\left(\lambda_{\max}(\widetilde{\mathbf{L}}_{\mathcal{E}}) \ge 3\lambda_{\max}\left(\mathbb{E}(\widetilde{\mathbf{L}}_{\mathcal{E}})\right)/2\right)\\ &\le (n+1)\exp(-cnM_n\underline{p}) \le n^{-10}.\end{aligned}$$

Then we know that (EC.23) holds with probability at least $1 - O(n^{-10})$ as the result of (EC.24).

EC.9.3. Proof of Lemma EC.1

First, we can rewrite

$$\nabla^2 \mathcal{L}(\boldsymbol{\theta};\boldsymbol{x}) = \sum_{\mathcal{S}\in\mathcal{E}} \frac{L_{\mathcal{S}}}{L} \left\{ \sum_{\substack{i,l\in\mathcal{S}_+ \\ i<l}} \frac{e^{\theta_i} e^{\theta_l}}{\left(\sum_{j\in\mathcal{S}_+} e^{\theta_j}\right)^2} (\boldsymbol{e}_i - \boldsymbol{e}_l)(\boldsymbol{e}_i - \boldsymbol{e}_l)^\top \right\}.$$

Then we have

$$\frac{e^{\theta_i}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j}} = \frac{1}{\sum_{j\in\mathcal{S}_+} e^{\theta_j-\theta_i}} = \frac{1}{\sum_{j\in\mathcal{S}_+} e^{(\theta_j-\theta_j^*)+(\theta_j^*-\theta_i^*)+(\theta_i^*-\theta_i)}} \geq \frac{1}{\sum_{j\in\mathcal{S}_+} e^{\log\kappa_{\boldsymbol{\theta}}+2\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|_\infty}} = \frac{1}{|\mathcal{S}_+|\kappa_{\boldsymbol{\theta}} e^{2C}}.$$

Similarly, we also have

$$\frac{e^{\theta_i}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j}} = \frac{1}{\sum_{j\in\mathcal{S}_+} e^{(\theta_j-\theta_j^*)+(\theta_j^*-\theta_i^*)+(\theta_i^*-\theta_i)}} \leq \frac{1}{\sum_{j\in\mathcal{S}_+} e^{-\log\kappa_{\boldsymbol{\theta}}-2\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|_\infty}} = \frac{\kappa_{\boldsymbol{\theta}} e^{2C}}{|\mathcal{S}_+|}.$$

Since $(\boldsymbol{e}_i - \boldsymbol{e}_l)(\boldsymbol{e}_i - \boldsymbol{e}_l)^\top \succeq \mathbf{0}$ for any $i < l$, we have that

$$\left(\frac{e^{\theta_i} e^{\theta_l}}{\left(\sum_{j\in\mathcal{S}_+} e^{\theta_j}\right)^2} - \frac{1}{|\mathcal{S}_+|^2(\kappa_{\boldsymbol{\theta}} e^{2C})^2} \right) (\boldsymbol{e}_i - \boldsymbol{e}_l)(\boldsymbol{e}_i - \boldsymbol{e}_l)^\top \succeq \mathbf{0},$$

$$\text{and} \quad \left(\frac{e^{\theta_i} e^{\theta_l}}{\left(\sum_{j\in\mathcal{S}_+} e^{\theta_j}\right)^2} - \frac{(\kappa_{\boldsymbol{\theta}} e^{2C})^2}{|\mathcal{S}_+|^2} \right) (\boldsymbol{e}_i - \boldsymbol{e}_l)(\boldsymbol{e}_i - \boldsymbol{e}_l)^\top \preceq \mathbf{0}.$$

Hence,

$$\begin{aligned}
\nabla^2 \mathcal{L}(\boldsymbol{\theta};\boldsymbol{x}) &= \sum_{\mathcal{S}\in\mathcal{E}} \frac{L_{\mathcal{S}}}{L} \left\{ \sum_{\substack{i,l\in\mathcal{S}_+ \\ i<l}} \frac{e^{\theta_i} e^{\theta_l}}{\left(\sum_{j\in\mathcal{S}_+} e^{\theta_j}\right)^2} (\boldsymbol{e}_i - \boldsymbol{e}_l)(\boldsymbol{e}_i - \boldsymbol{e}_l)^\top \right\} \\
&\succeq \sum_{\mathcal{S}\in\mathcal{E}} \frac{L_{\mathcal{S}}}{L} \left\{ \sum_{\substack{i,l\in\mathcal{S}_+ \\ i<l}} \frac{1}{|\mathcal{S}_+|^2(\kappa_{\boldsymbol{\theta}} e^{2C})^2} (\boldsymbol{e}_i - \boldsymbol{e}_l)(\boldsymbol{e}_i - \boldsymbol{e}_l)^\top \right\} \\
&= \frac{1}{(\kappa_{\boldsymbol{\theta}} e^{2C})^2} \sum_{\mathcal{S}\in\mathcal{E}} \frac{L_{\mathcal{S}}}{L} \left\{ \sum_{\substack{i,l\in\mathcal{S}_+ \\ i<l}} \frac{1}{|\mathcal{S}_+|^2} (\boldsymbol{e}_i - \boldsymbol{e}_l)(\boldsymbol{e}_i - \boldsymbol{e}_l)^\top \right\} = \frac{1}{(\kappa_{\boldsymbol{\theta}} e^{2C})^2} \mathbf{L}_{\mathcal{E}},
\end{aligned}$$

and similarly $\nabla^2 \mathcal{L}(\boldsymbol{\theta};\boldsymbol{x}) \preceq (\kappa_{\boldsymbol{\theta}} e^{2C})^2 \mathbf{L}_{\mathcal{E}}$. Then by the fact that $\nabla^2 \mathcal{L}_\tau(\boldsymbol{\theta};\boldsymbol{x}) = \nabla^2 \mathcal{L}(\boldsymbol{\theta};\boldsymbol{x}) + \tau \mathbf{I}_{n+1}$, the inequalities in (EC.3) follow.

EC.9.4. Proof of Lemma EC.2

We know that

$$\begin{aligned}\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*;\boldsymbol{x}) &= \tau\boldsymbol{\theta}^* - \sum_{S\in\mathcal{E}}\left\{\sum_{i\in S_+}\left(x_S^{(i)} - \frac{L_S}{L}\frac{e^{\theta_i^*}}{\sum_{j\in S_+}e^{\theta_j^*}}\right)\boldsymbol{e}_i\right\}\\ &= \tau\boldsymbol{\theta}^* - \frac{1}{L}\sum_{S\in\mathcal{E}}\sum_{\ell=1}^{L_S}\underbrace{\left\{\sum_{i\in S_+}\left(x_S^{(i,\ell)} - \frac{e^{\theta_i^*}}{\sum_{j\in S_+}e^{\theta_j^*}}\right)\boldsymbol{e}_i\right\}}_{:=z_S^{(\ell)}},\end{aligned}$$

and we have $\mathbb{E}\left(\boldsymbol{z}_S^{(\ell)}\right)=\mathbf{0}$ and

$$\left\|\boldsymbol{z}_S^{(\ell)}\right\|_2^2 = \sum_{i\in S_+}\left(x_S^{(i,\ell)} - \frac{e^{\theta_i^*}}{\sum_{j\in S_+}e^{\theta_j^*}}\right)^2 \le \sum_{i\in S_+}(x_S^{(i,\ell)})^2 + \sum_{i\in S_+}\frac{(e^{\theta_i^*})^2}{\left(\sum_{j\in S_+}e^{\theta_j^*}\right)^2} \le 1 + \sum_{i\in S_+}\frac{e^{\theta_i^*}}{\sum_{j\in S_+}e^{\theta_j^*}} = 2,$$

$$\mathbb{E}\left(\boldsymbol{z}_S^{(\ell)}\boldsymbol{z}_S^{(\ell)\top}\right) = \sum_{\substack{i,l\in S_+\\ i<l}}\frac{e^{\theta_i^*}e^{\theta_l^*}}{\left(\sum_{j\in S_+}e^{\theta_j^*}\right)^2}(\boldsymbol{e}_i-\boldsymbol{e}_l)(\boldsymbol{e}_i-\boldsymbol{e}_l)^\top \preceq \kappa_{\boldsymbol{\theta}}^2\sum_{\substack{i,l\in S_+\\ i<l}}\frac{1}{|S_+|^2}(\boldsymbol{e}_i-\boldsymbol{e}_l)(\boldsymbol{e}_i-\boldsymbol{e}_l)^\top;$$

$$\mathbb{E}\left(\boldsymbol{z}_S^{(\ell)\top}\boldsymbol{z}_S^{(\ell)}\right) = \sum_{i\in S_+}\frac{e^{\theta_i^*}}{\sum_{j\in S_+}e^{\theta_j^*}} - \sum_{i\in S_+}\left(\frac{e^{\theta_i^*}}{\sum_{j\in S_+}e^{\theta_j^*}}\right)^2 \le 1,$$

and thus under Assumption 1 we have

$$\left\|\frac{1}{L^2}\sum_{S\in\mathcal{E}}\sum_{\ell=1}^{L_S}\mathbb{E}\left(\boldsymbol{z}_S^{(\ell)}\boldsymbol{z}_S^{(\ell)\top}\right)\right\|_2 \lesssim \frac{\kappa_{\boldsymbol{\theta}}^2}{L}\left\|\sum_{S\in\mathcal{E}}\frac{L_S}{L}\left\{\frac{1}{|S_+|^2}\sum_{\substack{i,l\in S_+\\ i<l}}(\boldsymbol{e}_i-\boldsymbol{e}_j)(\boldsymbol{e}_i-\boldsymbol{e}_j)^\top\right\}\right\|_2 \lesssim \frac{1}{L}\|\mathbf{L}_{\mathcal{E}}\|_2 = \frac{\bar{\lambda}}{L};$$

$$\left|\frac{1}{L^2}\sum_{S\in\mathcal{E}}\sum_{\ell=1}^{L_S}\mathbb{E}\left(\boldsymbol{z}_S^{(\ell)\top}\boldsymbol{z}_S^{(\ell)}\right)\right| \le \left|\frac{1}{L}\sum_{S\in\mathcal{E}}\frac{L_S}{L}\right| \le \left|\frac{n+1}{L}\sum_{S\in\mathcal{E}}\frac{L_S}{L\cdot|S_+|}\right| \lesssim \frac{n}{L}(\mathbf{L}_{\mathcal{E}})_{1,1} \le \frac{n}{L}\|\mathbf{L}_{\mathcal{E}}\|_2 = \frac{n\bar{\lambda}}{L}.$$

Thus we can take $V=\frac{n\bar{\lambda}}{L}$ and $B=1/L$, then by the matrix Bernstein inequality (Tropp 2012), with probability at least $1-O(n^{-10})$ we have that

$$\|\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*;\boldsymbol{x}) - \tau\boldsymbol{\theta}^*\|_2 \lesssim \sqrt{V\log n} + B\log n = \sqrt{\frac{n\bar{\lambda}\log n}{L}} + \frac{\log n}{L} \lesssim \sqrt{\frac{n\bar{\lambda}\log n}{L}},$$

where the last inequality follows by the condition that $n\bar{\lambda}L \gtrsim \log n$.

Thus when $\tau \asymp \sqrt{\frac{\bar{\lambda}\log n}{L}}$, since we know that $\|\boldsymbol{\theta}^*\|_2 \le \sqrt{n+1}\log \kappa_{\boldsymbol{\theta}}$, with probability at least $1-O(n^{-10})$ we have

$$\|\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*;\boldsymbol{x})\|_2 \le \|\tau\boldsymbol{\theta}^*\|_2 + \|\nabla \mathcal{L}_\tau(\boldsymbol{\theta}^*;\boldsymbol{x}) - \tau\boldsymbol{\theta}^*\|_2 \lesssim \sqrt{\frac{n\bar{\lambda}\log n}{L}}.$$

EC.9.5. Proof of Lemma EC.3

Note that for any $\boldsymbol{\theta} \in \mathbb{R}^{n+1}$ and any $\mathcal{S} \in \boldsymbol{\mathcal{S}}$, we have that

$$e^{\theta_i}e^{\theta_l}\Big/\Big(\sum_{j\in\mathcal{S}_+} e^{\theta_j}\Big)^2 \le e^{\theta_i}\Big/\Big(\sum_{j\in\mathcal{S}_+} e^{\theta_j}\Big)\Big(1-e^{\theta_i}\Big/\Big(\sum_{j\in\mathcal{S}_+} e^{\theta_j}\Big)\Big) \le 1/4, \quad \forall i,l \in \mathcal{S}_+,$$

and thus

$$\begin{aligned}
\nabla^2 \mathcal{L}(\boldsymbol{\theta};\boldsymbol{x}) &= \sum_{S\in\boldsymbol{\mathcal{E}}} \frac{L_{\mathcal{S}}}{L}\left\{\sum_{\substack{i,l\in\mathcal{S}_+\\ i<l}} \frac{e^{\theta_i}e^{\theta_l}}{\left(\sum_{j\in\mathcal{S}_+} e^{\theta_j}\right)^2}(\boldsymbol{e}_i-\boldsymbol{e}_l)(\boldsymbol{e}_i-\boldsymbol{e}_l)^\top\right\}\\
&\preceq \frac{1}{4}\sum_{S\in\boldsymbol{\mathcal{E}}} \frac{L_{\mathcal{S}}}{L}\left\{\sum_{\substack{i,l\in\mathcal{S}_+\\ i<l}} (\boldsymbol{e}_i-\boldsymbol{e}_l)(\boldsymbol{e}_i-\boldsymbol{e}_l)^\top\right\}\\
&\preceq \frac{(n+1)^2}{4}\sum_{S\in\boldsymbol{\mathcal{E}}} \frac{L_{\mathcal{S}}}{L\cdot|\mathcal{S}_+|^2}\left\{\sum_{\substack{i,l\in\mathcal{S}_+\\ i<l}} (\boldsymbol{e}_i-\boldsymbol{e}_l)(\boldsymbol{e}_i-\boldsymbol{e}_l)^\top\right\} = \frac{(n+1)^2}{4}\cdot\mathbf{L}_{\boldsymbol{\mathcal{E}}}.
\end{aligned}$$

Thus under Assumption 1, we have that

$$\|\nabla^2 \mathcal{L}_\tau(\boldsymbol{\theta};\boldsymbol{x})\|_2 \le \tau + \frac{(n+1)^2}{4}\|\mathbf{L}_{\boldsymbol{\mathcal{E}}}\|_2 = \tau + \frac{(n+1)^2\bar{\lambda}}{4}.$$

EC.9.6. Proof of Lemma EC.4

By the optimality of $\widehat{\boldsymbol{\theta}}$ and Taylor's expansion, we have

$$\mathcal{L}_\tau(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) = \mathcal{L}_\tau(\boldsymbol{\theta}^*;\boldsymbol{x}) + (\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)^\top\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*;\boldsymbol{x}) + \frac{1}{2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)^\top\nabla^2\mathcal{L}_\tau(\widetilde{\boldsymbol{\theta}};\boldsymbol{x})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*) \le \mathcal{L}_\tau(\boldsymbol{\theta}^*;\boldsymbol{x}),$$

where $\widetilde{\boldsymbol{\theta}}$ lies between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$. Thus

$$\begin{aligned}\frac{1}{2}\lambda_{\min}\left(\nabla^2\mathcal{L}_\tau(\widetilde{\boldsymbol{\theta}};\boldsymbol{x})\right)\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2^2 &\le \frac{1}{2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)^\top\nabla^2\mathcal{L}_\tau(\widetilde{\boldsymbol{\theta}};\boldsymbol{x})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\\ &\le -(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)^\top\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*;\boldsymbol{x}) \le \|\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*;\boldsymbol{x})\|_2\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2,\end{aligned}$$

and thus by Lemma EC.2

$$\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2 \le \frac{2\|\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*;\boldsymbol{x})\|_2}{\lambda_{\min}\left(\nabla^2\mathcal{L}_\tau(\widetilde{\boldsymbol{\theta}};\boldsymbol{x})\right)} \lesssim \frac{2\|\nabla\mathcal{L}_\tau(\boldsymbol{\theta}^*;\boldsymbol{x})\|_2}{\tau} \lesssim \sqrt{n}.$$

EC.9.7. Proof of Lemma EC.5

In this section, we provide the proof for the upper bounds given in Lemma EC.5.

EC.9.7.1. Proof of EC.7 Note that for any $i \in [n]_+$, we have that $\left|x_{\mathcal{S}}^{(i,\ell)} - \frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+}e^{\theta_j^*}}\right| \le 1$, $\mathrm{Var}(x_{\mathcal{S}}^{(i,\ell)}) \le e^{\theta_i^*}/\left(\sum_{j\in\mathcal{S}_+}e^{\theta_j^*}\right) \lesssim 1/|\mathcal{S}_+|$, and

$$\frac{1}{L^2}\sum_{\mathcal{S}\in\boldsymbol{\mathcal{E}}:i\in\mathcal{S}_+}\sum_{\ell=1}^{L_{\mathcal{S}}}\mathrm{Var}\left(x_{\mathcal{S}}^{(i,\ell)} - \frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+}e^{\theta_j^*}}\right) \lesssim \frac{1}{L}\sum_{\mathcal{S}\in\boldsymbol{\mathcal{E}}:i\in\mathcal{S}_+}\frac{L_{\mathcal{S}}}{L\cdot|\mathcal{S}_+|} \lesssim \frac{1}{L}(\mathbf{L}_{\boldsymbol{\mathcal{E}}})_{i+1,i+1} \le \frac{\bar{\lambda}}{L}.$$

By Bernstein's inequality, conditional on $\boldsymbol{\mathcal{E}}$, with probability at least $1-O(n^{-11})$, we have

$$\left|\frac{1}{L}\sum_{\mathcal{S}\in\boldsymbol{\mathcal{E}}:i\in\mathcal{S}_+}\sum_{\ell=1}^{L_{\mathcal{S}}}\left(x_{\mathcal{S}}^{(i,\ell)} - \frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+}e^{\theta_j^*}}\right)\right| \lesssim \sqrt{\frac{\bar{\lambda}\log n}{L}} + \frac{\log n}{L}.$$

Hence taking the union bound over $i \in [n]_+$, with probability at least $1-O(n^{-10})$, we have

$$\|\nabla\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})\|_\infty = \max_{i\in[n]_+}\left|\frac{1}{L}\sum_{\mathcal{S}\in\boldsymbol{\mathcal{E}}:i\in\mathcal{S}_+}\sum_{\ell=1}^{L_{\mathcal{S}}}\left(x_{\mathcal{S}}^{(i,\ell)} - \frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+}e^{\theta_j^*}}\right)\right| \lesssim \sqrt{\frac{\bar{\lambda}\log n}{L}} + \frac{\log n}{L}.$$

EC.9.7.2. Proof of EC.8 We first consider $\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) - \nabla\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})$. For $i \in [n]_+$, by the mean value theorem we have

$$\begin{aligned}\left[\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) - \nabla\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})\right]_{i+1} &= \sum_{\mathcal{S}\in\mathcal{E}:i\in\mathcal{S}_+} \frac{L_{\mathcal{S}}}{L}\left(\frac{e^{\widehat{\theta}_i}}{\sum_{j\in\mathcal{S}_+} e^{\widehat{\theta}_j}} - \frac{e^{\theta^*_i}}{\sum_{j\in\mathcal{S}_+} e^{\theta^*_j}}\right)\\ &= \sum_{\mathcal{S}\in\mathcal{E}:i\in\mathcal{S}_+} \frac{L_{\mathcal{S}}}{L}\left\{\frac{e^{\widetilde{\theta}_i}\left(\sum_{j\in\mathcal{S}_+\setminus\{i\}} e^{\widetilde{\theta}_j}\right)}{\left(\sum_{j\in\mathcal{S}_+} e^{\widetilde{\theta}_j}\right)^2}\boldsymbol{e}_i - \sum_{l\in\mathcal{S}_+\setminus\{i\}} \frac{e^{\widetilde{\theta}_i}e^{\widetilde{\theta}_l}}{\left(\sum_{j\in\mathcal{S}_+} e^{\widetilde{\theta}_j}\right)^2}\boldsymbol{e}_l\right\}^\top (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*),\end{aligned}$$

where $\widetilde{\boldsymbol{\theta}}$ lies on the line between $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$. Then we have

$$\begin{aligned}&\left|\left[\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) - \nabla\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x}) - \nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right]_{i+1}\right|\\ &\le \left\|\sum_{\mathcal{S}\in\mathcal{E}:i\in\mathcal{S}_+} \frac{L_{\mathcal{S}}}{L}\left\{\frac{e^{\widetilde{\theta}_i}\left(\sum_{l\in\mathcal{S}_+\setminus\{i\}} e^{\widetilde{\theta}_l}\right)}{\left(\sum_{j\in\mathcal{S}_+} e^{\widetilde{\theta}_j}\right)^2}\boldsymbol{e}_i - \sum_{l\in\mathcal{S}_+\setminus\{i\}} \frac{e^{\widetilde{\theta}_i}e^{\widetilde{\theta}_l}}{\left(\sum_{j\in\mathcal{S}_+} e^{\widetilde{\theta}_j}\right)^2}\boldsymbol{e}_l\right\} - [\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})]_{i+1}\right\|_1 \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty.\end{aligned}$$

Now consider any $\mathcal{S} \in \mathcal{E}$ such that $i \in \mathcal{S}_+$. Under the condition that $\sqrt{n\kappa_{\mathcal{E}}\log n/(\underline{\lambda}L)} \le c$ for some small enough constant $c > 0$, by Theorem 2 there exists a small enough constant $\epsilon > 0$ such that $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \le \epsilon$ with probability at least $1 - O(n^{-10})$. Then with probability at least $1 - O(n^{-10})$, for all $\widetilde{\boldsymbol{\theta}}$ lies between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ and all $l \in \mathcal{S}_+$, we have

$$\frac{e^{\widetilde{\theta}_l}}{\sum_{j\in\mathcal{S}_+} e^{\widetilde{\theta}_j}} = \frac{e^{\theta^*_l + (\widetilde{\theta}_l - \theta^*_l)}}{\sum_{j\in\mathcal{S}_+} e^{\theta^*_j + (\widetilde{\theta}_j - \theta^*_j)}} \le e^{2\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty}\left(\frac{e^{\theta^*_l}}{\sum_{j\in\mathcal{S}_+} e^{\theta^*_j}}\right) \le e^{2\epsilon}\left(\frac{e^{\theta^*_l}}{\sum_{j\in\mathcal{S}_+} e^{\theta^*_j}}\right) \lesssim \frac{e^{\theta^*_l}}{\sum_{j\in\mathcal{S}_+} e^{\theta^*_j}}.$$

Then for any $l \in \mathcal{S}_+$, by the mean value theorem,

$$\begin{aligned}\left|\frac{e^{\widetilde{\theta}_l}}{\sum_{j\in\mathcal{S}_+} e^{\widetilde{\theta}_j}} - \frac{e^{\theta^*_l}}{\sum_{j\in\mathcal{S}_+} e^{\theta^*_j}}\right| &\lesssim \frac{e^{\theta^*_l}}{\sum_{j\in\mathcal{S}_+} e^{\theta^*_j}}\left(1 - \frac{e^{\theta^*_l}}{\sum_{j\in\mathcal{S}_+} e^{\theta^*_j}}\right)|\widetilde{\theta}_l - \theta^*_l| + \frac{e^{\theta^*_l}}{\sum_{j\in\mathcal{S}_+} e^{\theta^*_j}} \sum_{j\in\mathcal{S}_+\setminus\{l\}} \frac{e^{\theta^*_j}}{\sum_{j\in\mathcal{S}_+} e^{\theta^*_j}}|\widetilde{\theta}_j - \theta^*_j|\\ &\lesssim \frac{e^{\theta^*_l}}{\sum_{j\in\mathcal{S}_+} e^{\theta^*_j}}\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \le \frac{e^{\theta^*_l}}{\sum_{j\in\mathcal{S}_+} e^{\theta^*_j}}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty.\end{aligned}$$

In turn for any $l \in \mathcal{S}_+ \backslash \{i\}$, by we have

$$\begin{aligned}
&\left| \frac{e^{\widetilde{\theta}_i} e^{\widetilde{\theta}_l}}{\left(\sum_{j\in\mathcal{S}_+} e^{\widetilde{\theta}_j}\right)^2} - \frac{e^{\theta_i^*} e^{\theta_l^*}}{\left(\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}\right)^2} \right| \le \left| \frac{e^{\widetilde{\theta}_i} e^{\widetilde{\theta}_l}}{\left(\sum_{j\in\mathcal{S}_+} e^{\widetilde{\theta}_j}\right)^2} - \left(\frac{e^{\widetilde{\theta}_i}}{\sum_{j\in\mathcal{S}_+} e^{\widetilde{\theta}_j}}\right)\left(\frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}}\right) \right| \\
&\quad + \left| \left(\frac{e^{\widetilde{\theta}_i}}{\sum_{j\in\mathcal{S}_+} e^{\widetilde{\theta}_j}}\right)\left(\frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}}\right) - \frac{e^{\theta_i^*} e^{\theta_l^*}}{\left(\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}\right)^2} \right| \\
&\lesssim \left(\frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}}\right)\left(\frac{e^{\theta_l^*}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}}\right) \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty.
\end{aligned}$$

Thus with probability at least $1 - O(n^{-10})$, for all $i \in [n]_+$ we have that

$$\begin{aligned}
&\left\| \sum_{\mathcal{S}\in\mathcal{E}, i\in\mathcal{S}_+} \frac{L_{\mathcal{S}}}{L} \left\{ \frac{e^{\widetilde{\theta}_i}\left(\sum_{l\in\mathcal{S}_+\backslash\{i\}} e^{\widetilde{\theta}_l}\right)}{\left(\sum_{j\in\mathcal{S}_+} e^{\widetilde{\theta}_j}\right)^2} \boldsymbol{e}_i - \sum_{l\in\mathcal{S}_+\backslash\{i\}} \frac{e^{\widetilde{\theta}_i} e^{\widetilde{\theta}_l}}{\left(\sum_{j\in\mathcal{S}_+} e^{\widetilde{\theta}_j}\right)^2} \boldsymbol{e}_l \right\} - [\nabla^2 \mathcal{L}(\boldsymbol{\theta}^*; \boldsymbol{x})]_{i+1} \right\|_1 \\
&\lesssim \sum_{\mathcal{S}\in\mathcal{E}, i\in\mathcal{S}_+} \frac{L_{\mathcal{S}}}{L} \sum_{l\in\mathcal{S}_+\backslash\{i\}} \left| \frac{e^{\widetilde{\theta}_i} e^{\widetilde{\theta}_l}}{\left(\sum_{j\in\mathcal{S}_+} e^{\widetilde{\theta}_j}\right)^2} - \frac{e^{\theta_i^*} e^{\theta_l^*}}{\left(\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}\right)^2} \right| \\
&\lesssim \sum_{\mathcal{S}\in\mathcal{E}, i\in\mathcal{S}_+} \frac{L_{\mathcal{S}}}{L} \sum_{l\in\mathcal{S}_+\backslash\{i\}} \left(\frac{e^{\theta_i^*}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}}\right)\left(\frac{e^{\theta_l^*}}{\sum_{j\in\mathcal{S}_+} e^{\theta_j^*}}\right) \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \\
&\lesssim \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \sum_{\mathcal{S}\in\mathcal{E}, i\in\mathcal{S}_+} \frac{L_{\mathcal{S}}}{L \cdot |\mathcal{S}_+|} \lesssim \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty (\mathbf{L}_{\mathcal{E}})_{i+1,i+1} \lesssim \bar{\lambda} \cdot \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty.
\end{aligned}$$

Putting the above analysis together, with probability at least $1 - O(n^{-10})$, we have that for all $i \in [n]_+$,

$$\left| \left[\nabla \mathcal{L}(\widehat{\boldsymbol{\theta}}; \boldsymbol{x}) - \nabla \mathcal{L}(\boldsymbol{\theta}^*; \boldsymbol{x}) - \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*; \boldsymbol{x})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \right]_{i+1} \right| \lesssim \bar{\lambda} \cdot \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty^2 \lesssim \frac{\kappa_{\mathcal{E}}^2 n \log n}{L},$$

and hence (EC.8) holds.

REMARK EC.2. Recall that

$$\begin{aligned}&\left[\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})-\nabla\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})\right]_{i+1}\\&=\sum_{\mathcal{S}\in\mathcal{E},i\in\mathcal{S}_+}\frac{L_{\mathcal{S}}}{L}\left\{\frac{e^{\widetilde{\theta}_i}\left(\sum_{j\in\mathcal{S}_+\setminus\{i\}}e^{\widetilde{\theta}_j}\right)}{\left(\sum_{j\in\mathcal{S}_+}e^{\widetilde{\theta}_j}\right)^2}\boldsymbol{e}_i-\sum_{l\in\mathcal{S}_+\setminus\{i\}}\frac{e^{\widetilde{\theta}_i}e^{\widetilde{\theta}_l}}{\left(\sum_{j\in\mathcal{S}_+}e^{\widetilde{\theta}_j}\right)^2}\boldsymbol{e}_l\right\}^\top(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*),\end{aligned}$$

where $\widetilde{\boldsymbol{\theta}}$ lies on the line between $\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*$. Thus with probability at least $1-O(n^{-10})$, under the condition that $\sqrt{n\kappa_{\mathcal{E}}\log n/(\underline{\lambda}L)}\le c$ for some small enough constant $c>0$, for all $i\in[n]_+$ we have that

$$\begin{aligned}&\left|\left[\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})-\nabla\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})\right]_{i+1}\right|\\&\le\left\|\sum_{\mathcal{S}\in\mathcal{E},i\in\mathcal{S}_+}\frac{L_{\mathcal{S}}}{L}\left\{\frac{e^{\widetilde{\theta}_i}\left(\sum_{j\in\mathcal{S}_+\setminus\{i\}}e^{\widetilde{\theta}_j}\right)}{\left(\sum_{j\in\mathcal{S}_+}e^{\widetilde{\theta}_j}\right)^2}\boldsymbol{e}_i-\sum_{l\in\mathcal{S}_+\setminus\{i\}}\frac{e^{\widetilde{\theta}_i}e^{\widetilde{\theta}_l}}{\left(\sum_{j\in\mathcal{S}_+}e^{\widetilde{\theta}_j}\right)^2}\boldsymbol{e}_l\right\}\right\|_1\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_\infty\\&\lesssim\left\{\sum_{\mathcal{S}\in\mathcal{E},i\in\mathcal{S}_+}\frac{L_{\mathcal{S}}}{L}\cdot\frac{e^{\widetilde{\theta}_i}\left(\sum_{j\in\mathcal{S}_+\setminus\{i\}}e^{\widetilde{\theta}_j}\right)}{\left(\sum_{j\in\mathcal{S}_+}e^{\widetilde{\theta}_j}\right)^2}\right\}\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_\infty\lesssim\left(\sum_{\mathcal{S}\in\mathcal{E},i\in\mathcal{S}_+}\frac{L_{\mathcal{S}}}{L\cdot|\mathcal{S}_+|}\right)\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_\infty\\&\lesssim\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_\infty(\mathbf{L}_{\mathcal{E}})_{i+1,i+1}\lesssim\bar{\lambda}\cdot\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_\infty.\end{aligned}$$

Then with probability at least $1-O(n^{-10})$ we have

$$\left\|\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})-\nabla\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})\right\|_\infty\lesssim\bar{\lambda}\cdot\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_\infty\lesssim\sqrt{\frac{\kappa_{\mathcal{E}}^2n\bar{\lambda}\log n}{L}}.\tag{EC.27}$$

EC.9.7.3. Proof of EC.9 For any $i\in[n]_+$, very similar to the proof in Section EC.9.7.2, with probability at least $1-O(n^{-10})$, for all $i\in[n]_+$ we have

$$\begin{aligned}&\|[\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})]_{i+1}-[\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})]_{i+1}\|_1\\&=\left\|\sum_{\mathcal{S}\in\mathcal{E},i\in\mathcal{S}_+}\left\{\frac{e^{\widehat{\theta}_i}\left(\sum_{l\in\mathcal{S}_+\setminus\{i\}}e^{\widehat{\theta}_l}\right)}{\left(\sum_{j\in\mathcal{S}_+}e^{\widehat{\theta}_j}\right)^2}\boldsymbol{e}_i-\sum_{l\in\mathcal{S}_+\setminus\{i\}}\frac{e^{\widehat{\theta}_i}e^{\widehat{\theta}_l}}{\left(\sum_{j\in\mathcal{S}_+}e^{\widehat{\theta}_j}\right)^2}\boldsymbol{e}_l\right\}-[\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})]_{i+1}\right\|_1\\&\lesssim\bar{\lambda}\cdot\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_\infty,\end{aligned}$$

and hence by Theorem 2 with probability at least $1 - O(n^{-10})$ we have

$$\|\nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}}; \boldsymbol{x}) - \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*; \boldsymbol{x})\|_\infty = \max_{i \in [n+1]} \|[\nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}}; \boldsymbol{x})]_{i+1} - [\nabla^2 \mathcal{L}(\boldsymbol{\theta}^*; \boldsymbol{x})]_{i+1}\|_1 \lesssim \sqrt{\frac{\kappa_{\mathcal{E}}^2 n \bar{\lambda} \log n}{L}}.$$

EC.9.7.4. Proof of EC.10 For the simplicity of notation, we let $\widetilde{\lambda}_1 \geq \ldots \geq \widetilde{\lambda}_n \geq \widetilde{\lambda}_{n+1} = 0$ be the eigenvalues of $\nabla^2 \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x})$ in descending order, and let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ and $\frac{1}{\sqrt{n+1}}\mathbf{1}$ be the normalized eigenvectors corresponding to $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_{n+1}$. From the proof of Lemma EC.1, we know that for any $\boldsymbol{\theta} \in \mathbb{R}^{n+1}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_\infty \leq C$ for some constant $C > 0$, we have $\frac{1}{(\kappa_{\boldsymbol{\theta}} e^{2C})^2} \mathbf{L}_{\mathcal{E}} \preceq \nabla^2 \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x}) \preceq (\kappa_{\boldsymbol{\theta}} e^{2C})^2 \mathbf{L}_{\mathcal{E}}$, which in turn gives that $\widetilde{\lambda}_1 \asymp \bar{\lambda}$ and $\widetilde{\lambda}_n \asymp \underline{\lambda}$. Also recall that $\nabla^2 \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x})\mathbf{1} = \mathbf{0}$, and hence for any constant $C > 0$, we have the eigen-decomposition

$$\begin{aligned}
\begin{pmatrix} \nabla^2 \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x}) & C\mathbf{1} \\ C\mathbf{1}^\top & 0 \end{pmatrix} &= \begin{pmatrix} \nabla^2 \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x}) & \mathbf{0} \\ \mathbf{0} & 0 \end{pmatrix} + C \begin{pmatrix} \mathbf{0} & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{pmatrix} = \sum_{i=1}^{n} \widetilde{\lambda}_i \begin{pmatrix} \boldsymbol{v}_i \\ 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_i \\ 0 \end{pmatrix}^\top \\
&+ C\sqrt{n+1} \begin{pmatrix} \frac{1}{\sqrt{2(n+1)}}\mathbf{1} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2(n+1)}}\mathbf{1} \\ \frac{1}{\sqrt{2}} \end{pmatrix}^\top - C\sqrt{n+1} \begin{pmatrix} \frac{1}{\sqrt{2(n+1)}}\mathbf{1} \\ -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2(n+1)}}\mathbf{1} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}^\top.
\end{aligned}$$

It is not hard to verify by the relationship between $\boldsymbol{v}_i$'s and $\mathbf{1}$ that the above representation is a valid eigen-decomposition of $\begin{pmatrix} \nabla^2 \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x}) & C\mathbf{1} \\ C\mathbf{1}^\top & 0 \end{pmatrix}$. Thus we can see that the $n+2$ eigenvalues of $\begin{pmatrix} \nabla^2 \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x}) & C\mathbf{1} \\ C\mathbf{1}^\top & 0 \end{pmatrix}$ are $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n$ and $\pm C\sqrt{n+1}$, and $\begin{pmatrix} \nabla^2 \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x}) & C\mathbf{1} \\ C\mathbf{1}^\top & 0 \end{pmatrix}^{-1}$ would take the following form

$$\begin{aligned}
\begin{pmatrix} \nabla^2 \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x}) & C\mathbf{1} \\ C\mathbf{1}^\top & 0 \end{pmatrix}^{-1} &= \sum_{i=1}^{n} \widetilde{\lambda}_i^{-1} \begin{pmatrix} \boldsymbol{v}_i \\ 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_i \\ 0 \end{pmatrix}^\top + C^{-1}(n+1)^{-1/2} \begin{pmatrix} \frac{1}{\sqrt{2(n+1)}}\mathbf{1} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2(n+1)}}\mathbf{1} \\ \frac{1}{\sqrt{2}} \end{pmatrix}^\top \\
&- C^{-1}(n+1)^{-1/2} \begin{pmatrix} \frac{1}{\sqrt{2(n+1)}}\mathbf{1} \\ -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2(n+1)}}\mathbf{1} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}^\top = \begin{pmatrix} \nabla^2 \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x})^\dagger & \frac{1}{C(n+1)}\mathbf{1} \\ \frac{1}{C(n+1)}\mathbf{1}^\top & 0 \end{pmatrix},
\end{aligned}$$

where $\nabla^2\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x})^\dagger = \sum_{i=1}^n \widetilde{\lambda}_i^{-1}\boldsymbol{v}_i\boldsymbol{v}_i^\top$. Then for any $\boldsymbol{\theta}\in\mathbb{R}^{n+1}$ such that $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_\infty \le C$ for some constant $C>0$ and a constant $C_n>0$ (dependent on n) such that $\underline{\lambda}\le C_n\sqrt{n}$, we have

$$\left\|\begin{pmatrix}\nabla^2\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x}) & C_n\mathbf{1}\\ C_n\mathbf{1}^\top & 0\end{pmatrix}^{-1}\right\|_2 = \max\left(\frac{1}{\lambda_n(\nabla^2\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x}))}, \frac{1}{C_n\sqrt{n+1}}\right) \lesssim \frac{1}{\underline{\lambda}}.$$

Thus by EC.9, with probability at least $1-O(n^{-10})$ we have

$$\begin{aligned}
&\left\|\begin{pmatrix}\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) & \mathbf{1}\\ \mathbf{1}^\top & 0\end{pmatrix}^{-1} - \begin{pmatrix}\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x}) & \mathbf{1}\\ \mathbf{1}^\top & 0\end{pmatrix}^{-1}\right\|_2 = \left\|\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})^\dagger - \nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})^\dagger\right\|_2\\
&= \left\|\begin{pmatrix}\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) & C_n\mathbf{1}\\ C_n\mathbf{1}^\top & 0\end{pmatrix}^{-1} - \begin{pmatrix}\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x}) & C_n\mathbf{1}\\ C_n\mathbf{1}^\top & 0\end{pmatrix}^{-1}\right\|_2\\
&\le \left\|\begin{pmatrix}\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) C_n\mathbf{1}\\ C_n\mathbf{1}^\top & 0\end{pmatrix}^{-1}\right\|_2 \left\|\begin{pmatrix}\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x}) C_n\mathbf{1}\\ C_n\mathbf{1}^\top & 0\end{pmatrix}^{-1}\right\|_2 \left\|\begin{pmatrix}\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) C_n\mathbf{1}\\ C_n\mathbf{1}^\top & 0\end{pmatrix} - \begin{pmatrix}\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x}) C_n\mathbf{1}\\ C_n\mathbf{1}^\top & 0\end{pmatrix}\right\|_2\\
&\lesssim \frac{\sqrt{n}}{\underline{\lambda}^2}\|\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) - \nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})\|_\infty \lesssim \frac{n}{\underline{\lambda}}\sqrt{\frac{\kappa_{\mathcal{E}}^3\log n}{\underline{\lambda}L}}.
\end{aligned}$$

EC.9.8. Proof of Corollary EC.1

From the proof of Lemma EC.5 it can be seen that

$$\begin{pmatrix}\nabla^2\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x}) & \mathbf{1}\\ \mathbf{1}^\top & 0\end{pmatrix}^{-1} = \begin{pmatrix}\nabla^2\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x})^\dagger & \frac{1}{n+1}\mathbf{1}\\ \frac{1}{n+1}\mathbf{1}^\top & 0\end{pmatrix}$$

holds true for any $\boldsymbol{\theta}\in\mathbb{R}^{n+1}$, where $\nabla^2\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x})^\dagger$ is the Moore-Penrose inverse of $\nabla^2\mathcal{L}(\boldsymbol{\theta};\boldsymbol{x})$. Now define $\boldsymbol{\Gamma} = (\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n,\frac{1}{\sqrt{n+1}}\mathbf{1})$, and denote by $\boldsymbol{\gamma}_j = (\gamma_{j1},\ldots,\gamma_{jn},\frac{1}{\sqrt{n+1}})$ the j-th row of $\boldsymbol{\Gamma}$, $j\in[n+1]$. Then since $\boldsymbol{\Gamma}$ is orthonormal, it can be seen that $\|\boldsymbol{\gamma}_j\|_2^2 = \sum_{k=1}^n\gamma_{jk}^2 + \frac{1}{n+1} = 1$ and $\sum_{k=1}^n\gamma_{jk}^2 \asymp 1$. Thus for any $\boldsymbol{\theta}\in\mathbb{R}^{n+1}$ such that $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_\infty < C$ for some constant $C>0$, for $j,k\in[n+1]$, $j\ne k$,

we have

$$\frac{1}{\bar{\lambda}} \lesssim [\nabla^2 \mathcal{L}(\boldsymbol{\theta};\boldsymbol{x})^\dagger]_{jj} = \boldsymbol{e}_j^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta};\boldsymbol{x})^\dagger \boldsymbol{e}_j = \boldsymbol{\gamma}_j^\top \begin{pmatrix} \widetilde{\lambda}_1^{-1} & \dots & 0 \\ \vdots & \ddots & \\ & & \widetilde{\lambda}_n^{-1} \\ 0 & & 0 \end{pmatrix} \boldsymbol{\gamma}_j = \sum_{i=1}^n \gamma_{ji}^2 \widetilde{\lambda}_i^{-1} \lesssim \frac{1}{\underline{\lambda}},$$

$$\left|[\nabla^2 \mathcal{L}(\boldsymbol{\theta};\boldsymbol{x})^\dagger]_{jk}\right| = \left|\boldsymbol{e}_j^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta};\boldsymbol{x})^\dagger \boldsymbol{e}_k\right| \le \|\nabla^2 \mathcal{L}(\boldsymbol{\theta};\boldsymbol{x})^\dagger\|_2 \lesssim \frac{1}{\underline{\lambda}}.$$

EC.9.9. Proof of Lemma EC.6

By Corollary EC.1 and the definition of $\widehat{\boldsymbol{\theta}}^d$, we have that

$$\begin{aligned}\widehat{\boldsymbol{\theta}}^d &= \widehat{\boldsymbol{\theta}}' - \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})^\dagger \nabla \mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) = \widehat{\boldsymbol{\theta}}' - (\mathbf{I}_{n+1}, \mathbf{0}) \begin{pmatrix} \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{I}_{n+1} \\ \mathbf{0} \end{pmatrix} \nabla \mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) \\ &= (\mathbf{I}_{n+1}, \mathbf{0}) \begin{pmatrix} \widehat{\boldsymbol{\theta}}' \\ 0 \end{pmatrix} - (\mathbf{I}_{n+1}, \mathbf{0}) \begin{pmatrix} \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{pmatrix}^{-1} \begin{pmatrix} \nabla \mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) \\ 0 \end{pmatrix},\end{aligned}$$

then left-multiply both sides by $(\mathbf{I}_{n+1}, \mathbf{0})^\top$ we have

$$\begin{aligned}\begin{pmatrix} \widehat{\boldsymbol{\theta}}^d \\ 0 \end{pmatrix} &= \operatorname{diag}(\mathbf{I}_{n+1}, 0) \begin{pmatrix} \widehat{\boldsymbol{\theta}}' \\ 0 \end{pmatrix} - \operatorname{diag}(\mathbf{I}_{n+1}, 0) \begin{pmatrix} \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{pmatrix}^{-1} \begin{pmatrix} \nabla \mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \widehat{\boldsymbol{\theta}}' \\ 0 \end{pmatrix} - \begin{pmatrix} \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{pmatrix}^{-1} \begin{pmatrix} \nabla \mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) \\ 0 \end{pmatrix},\end{aligned}$$

where the second equality is due to the fact that $\mathbf{1}_{n+1}^\top \nabla \mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) = 0$. Hence we can see that $\widehat{\boldsymbol{\theta}}^d$ can be treated as the sub-vector for the first $n+1$ coordinates of an augmented Newton-debiased estimator. Recall that we state in the proof of Theorem 2, we will abuse the notation and let $\widehat{\boldsymbol{\theta}}$ denote $\widehat{\boldsymbol{\theta}}' = \mathbf{P}_1^\perp \widehat{\boldsymbol{\theta}}$ and let $\boldsymbol{\theta}^*$ denote $\boldsymbol{\theta}^{*\prime} = \mathbf{P}_1^\perp \boldsymbol{\theta}^*$. Then we have the following decomposition

$$\begin{pmatrix}\widehat{\boldsymbol{\theta}}^d-\boldsymbol{\theta}^*\\ 0\end{pmatrix}=\underbrace{\left(\begin{pmatrix}\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) & \mathbf{1}\\ \mathbf{1}^\top & 0\end{pmatrix}^{-1}-\begin{pmatrix}\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x}) & \mathbf{1}\\ \mathbf{1}^\top & 0\end{pmatrix}^{-1}\right)\begin{pmatrix}-\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})+\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\\ 0\end{pmatrix}}_{I_1}$$
$$+\underbrace{\begin{pmatrix}\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x}) & \mathbf{1}\\ \mathbf{1}^\top & 0\end{pmatrix}^{-1}\begin{pmatrix}\nabla\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})-\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})+\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\\ 0\end{pmatrix}}_{I_2}$$
$$+\begin{pmatrix}\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x}) & \mathbf{1}\\ \mathbf{1}^\top & 0\end{pmatrix}^{-1}\begin{pmatrix}-\nabla\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})\\ 0\end{pmatrix}.$$

By Lemma EC.5, the following bounds for the terms I_1 and I_2 hold with probability at least $1-O(n^{-10})$,

$$\begin{aligned}\|I_1\|_2 &\le \left\|\begin{pmatrix}\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) & \mathbf{1}\\ \mathbf{1}^\top & 0\end{pmatrix}^{-1}-\begin{pmatrix}\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x}) & \mathbf{1}\\ \mathbf{1}^\top & 0\end{pmatrix}^{-1}\right\|_2\left\|-\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})+\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\right\|_2\\ &\lesssim \sqrt{n}\left\|\begin{pmatrix}\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x}) & \mathbf{1}\\ \mathbf{1}^\top & 0\end{pmatrix}^{-1}-\begin{pmatrix}\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x}) & \mathbf{1}\\ \mathbf{1}^\top & 0\end{pmatrix}^{-1}\right\|_2\left\|-\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})+\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\right\|_\infty\\ &\lesssim \sqrt{n}\times\frac{n}{\underline{\lambda}}\sqrt{\frac{\kappa_{\mathcal{E}}^3\log n}{\underline{\lambda}L}}\left(\sqrt{\frac{\bar{\lambda}\log n}{L}}+\frac{\kappa_{\mathcal{E}}^2 n\log n}{L}\right)\lesssim\frac{\kappa_{\mathcal{E}}^2 n\sqrt{n}\log n}{\underline{\lambda}L}\left(n\sqrt{\frac{\kappa_{\mathcal{E}}^3\log n}{\underline{\lambda}L}}+1\right),\end{aligned}$$

and

$$\begin{aligned}\|I_2\|_2 &= \left\|\begin{pmatrix}\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})^\dagger & \frac{1}{n+1}\mathbf{1}\\ \frac{1}{n+1}\mathbf{1}^\top & 0\end{pmatrix}\begin{pmatrix}\nabla\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})-\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})+\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\\ 0\end{pmatrix}\right\|_2\\ &= \left\|\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})^\dagger\left(\nabla\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})-\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})+\nabla^2\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\right)\right\|_2\\ &\lesssim \sqrt{n}\left\|\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})^\dagger\right\|_2\|\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}};\boldsymbol{x})-\nabla\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})-\nabla^2\mathcal{L}(\boldsymbol{\theta}^*;\boldsymbol{x})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\|_\infty\\ &\lesssim \frac{\kappa_{\mathcal{E}}^2 n\log n}{L}\times\frac{\sqrt{n}}{\underline{\lambda}}=\frac{\kappa_{\mathcal{E}}^2 n\sqrt{n}\log n}{\underline{\lambda}L},\end{aligned}$$

where the second equality is due to the fact that $\mathbf{1}_{n+1}^\top \nabla \mathcal{L}(\boldsymbol{\theta};\boldsymbol{x}) = 0$ and $\mathbf{1}_{n+1}^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta};\boldsymbol{x}) = \mathbf{0}$ for any $\boldsymbol{\theta} \in \mathbb{R}^{n+1}$. Thus in turn we have that

$$\|\mathbf{R}_0\|_2 \le \|I_1\|_2 + \|I_2\|_2 \lesssim \frac{\kappa_{\mathcal{E}}^2 n\sqrt{n}\log n}{\underline{\lambda} L}\left(n\sqrt{\frac{\kappa_{\mathcal{E}}^3 \log n}{\underline{\lambda} L}} + 1\right).$$

EC.10. Complexity of Generating the Observed Offer Sets for Numerical Studies

PROPOSITION EC.1. *Denote by $B_{\mathcal{E}}$ the complexity of generating $\mathcal{E}$ with the procedure in Section 4. When $p \asymp n\log n/2^n$, we have $\mathbb{E}(B_{\mathcal{E}}) = O(n^2\log n)$. Furthermore, with probability at least $1 - O(n^{-1})$, we have $B_{\mathcal{E}} \le Cn^2\log n$ for some constant $C > 0$.*

Proof. Note that Step (1) only requires generating a scalar random variable, and its computational cost is dominated by Steps (2) and (3). We first show that $\mathbb{E}(B_{\mathcal{E}}) = O(n^2\log n)$. Since $M \sim \text{Binomial}(2^n - 1, p)$, we know that $\mathbb{E}(M) = (2^n - 1)p = O(n\log n)$. Conditional on M, for each $j = 1, \ldots, M$, denote by I_j the number of sampling attempts for the j-th iteration, and we have that I_j follows a geometric distribution with success probability equal to $(2^n - j)/2^n$, i.e., $I_j - 1 \sim \text{Geometric}(1, (2^n - j)/2^n)$ and $\mathbb{E}(I_j|M) = \mathbb{E}(I_j) = 1 + j/(2^n - j)$. Hence we have that

$$\begin{aligned}
\mathbb{E}(B_{\mathcal{E}}) &\lesssim n\mathbb{E}\Big(\sum_{j=1}^{M} I_j\Big) = n\mathbb{E}\Big(\mathbb{E}\Big(\sum_{j=1}^{M} I_j|M\Big)\Big) = n\mathbb{E}\Big(M + \sum_{j=1}^{M}\frac{j}{2^n - j}\Big) \\
&\lesssim n^2\log n + n\mathbb{E}\Big(\frac{M^2}{2^n - M}\Big) = n^2\log n + \frac{np}{1-p}\sum_{M=1}^{2^n-1}(M - 1 + 1)\binom{2^n - 1}{M - 1}p^{M-1}(1-p)^{2^n - M} \\
&\lesssim n^2\log n + np\mathbb{E}(M) \lesssim n^2\log n.
\end{aligned}$$

Next, we move on to show that there exists a constant $C > 0$ such that $B_{\mathcal{E}} \le Cn^2\log n$ with probability at least $1 - O(n^{-1})$. By Hoeffding's inequality (Hoeffding 1994), with probability at least $1 - n^{-1}$, we have that

$$M \le \mathbb{E}(M) + \sqrt{2^{n+1}p(1-p)\log n} \le Cn\log n, \tag{EC.28}$$

where $C > 0$ is a constant independent of n. Besides, conditional on the event $M \le Cn\log n$, the probability that $I_j = 1$ for all $j = 1, \ldots, M$ is

$$\prod_{j=2}^{M}\frac{2^n - j}{2^n} = 1 - O(n^2\log^2 n/2^n) = 1 - O(n^{-1}). \tag{EC.29}$$

Therefore, combining (EC.28) and (EC.29), with probability at least $1 - O(n^{-1})$, the total complexity is $B_{\mathcal{E}} \le Cn^2\log n$. □

References

Auriau V, Aouad A, Désir A, Malherbe E (2024) Choice-Learn: Large-scale choice modeling for operational contexts through the lens of machine learning. *Journal of Open Source Software* 9(101):6899.

Bodea T, Ferguson M, Garrow L (2009) Data set—Choice-based revenue management: Data from a major hotel chain. *Manufacturing & Service Operations Management* 11(2):356–361.

Bubeck S (2015) Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning* 8(3-4):231–357.

Chen Y, Fan J, Ma C, Wang K (2019) Spectral method and regularized MLE are both optimal for top-K ranking. *The Annals of Statistics* 47(4):2204–2235.

Chernozhukov V, Chetverikov D, Kato K (2013) Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics* 41(6):2786–2819.

ColumbiaX (2019) Choice-based conjoint analysis: Tablet product data. URL https://courses.edx.org/asset-v1%3AColumbiaX%2BBAMM.104x%2B3T2019%2Btype%40asset%2Bblock/choice_based_conjoint.nb.html, course material for ColumbiaX BAMM.104x Business Analytics.

Croissant Y (2023) *Ecdat: Data Sets for Econometrics*. URL https://vincentarelbundock.github.io/Rdatasets/, R package. Data sets Tuna, Yogurt, and Catsup.

Hoeffding W (1994) Probability inequalities for sums of bounded random variables. *The Collected Works of Wassily Hoeffding*, 409–426 (New York: Springer).

Kim BD, Blattberg RC, Rossi PE (1995) Modeling the distribution of price sensitivity and implications for optimal retail pricing. *Journal of Business & Economic Statistics* 13(3):291–303.

Kim C, Cho S, Im J (2022) *RMM: Revenue Management Modeling*. URL https://CRAN.R-project.org/package=RMM, R package version 0.1.0.

Tropp JA (2012) User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics* 12(4):389–434.

Tropp JA (2015) An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning* 8(1-2):1–230.