

# Globally-Optimal Greedy Experiment Selection for Active Sequential Estimation

Xiaoou Li

University of Minnesota

Hongru Zhao

University of Minnesota

## Abstract

Motivated by modern applications such as computerized adaptive testing, sequential rank aggregation, and heterogeneous data source selection, we study the problem of active sequential estimation, which involves adaptively selecting experiments for sequentially collected data. The goal is to design experiment selection rules for more accurate model estimation. Greedy information-based experiment selection methods, optimizing the information gain for one-step ahead, have been employed in practice thanks to their computational convenience, flexibility to context or task changes, and broad applicability. However, statistical analysis is restricted to one-dimensional cases due to the problem's combinatorial nature and the seemingly limited capacity of greedy algorithms, leaving the multidimensional problem open.

In this study, we close the gap for multidimensional problems. In particular, we propose adopting a class of greedy experiment selection methods and provide statistical analysis for the maximum likelihood estimator following these selection rules. This class encompasses both existing methods and introduces new methods with improved numerical efficiency. We prove that these methods produce consistent and asymptotically normal estimators. Additionally, within a decision theory framework, we establish that the proposed methods achieve asymptotic optimality when the risk measure aligns with the selection rule. We also conduct extensive numerical studies on both simulated and real data to illustrate the efficacy of the proposed methods.

From a technical perspective, we devise new analytical tools to address theoretical challenges. For instance, we demonstrate that functions of inverted Fisher information have a regularization effect when used in selection rules, thereby automatically exploring necessary experiments. Additionally, we show that a class of greedy and stochastic optimization methods converges to the minimum of a convex function over a simplex

almost surely. These analytical tools are of independent theoretical interest and may be reused in related problems involving stochastic approximation and sequential designs.

Keywords: Active sequential estimation, optimality theory, sequential analysis, computerized adaptive testing

## 1 Introduction

In many modern applications, data are collected sequentially and adaptively through varied experiments, with the distribution being influenced by both unknown model parameters and the experiments. Active sequential estimation, which involves the adaptive selection of the experiments, enables more efficient model estimation. It has received considerable attention across various disciplines recently. A few examples are provided below.

**Computerized Adaptive Testing (CAT)** CAT refers to a form of educational assessment where test items are administered adaptively and sequentially based on the test taker’s responses to previous items. For instance, if a test taker answers questions correctly, they may receive a more challenging item subsequently. Over the past decades, CAT has gained popularity due to its ability to achieve a more accurate assessment with fewer test items compared to traditional non-adaptive tests. To implement CAT, Item Response Theory (IRT) models are typically employed (Chen et al., 2024; Reckase, 2006). IRT models assume that a test-taker’s responses, whether correct or incorrect, are influenced by both their latent trait parameter and the selected item. A crucial aspect of CAT design involves developing effective item selection rules to estimate the latent trait parameter as accurately as possible. For a comprehensive review on this topic, see Bartroff et al. (2008); Chang and Ying (2009); Wang et al. (2017), and the references therein.

**Sequential rank aggregation** The rank aggregation problem involves inferring a global rank for a set of items by aggregating noisy pairwise comparison results. This problem finds applications across various domains such as social choice (Saaty and Vargas (2012)), sports (Elo (1978)), and search rankings (Page et al. (1999)). Statistical models such as the Bradley-Terry model (Bradley and Terry, 1952), which assigns a latent score parameter to each object, are often utilized to model the noisy pairwise comparison results. Subsequently, the global rank can be inferred from the estimated latent score parameters. Recently, the sequential rank aggregation problem has attracted increased interest. This approach involves sequentially and adaptively selecting the next pair to compare based on the comparison results of previously selected pairs (see, e.g., Chen et al. (2013, 2022, 2016)). A key question

of interest is the design of pair selection rules to enhance the efficiency of the rank aggregation process.

Besides the aforementioned applications, additional areas of application include active sampling in signal processing (Mukherjee et al., 2022), active contextual search (Chen et al., 2023), and dynamic pricing (Chen and Wang, 2023), among others.

In all of the above applications, the problem can be formulated as a sequential design-and-estimation problem where data  $X_1, \dots, X_n, \dots$  are collected sequentially. Each  $X_n$  has a density function  $f_{\boldsymbol{\theta}, a_n}(\cdot)$  relative to a baseline measure, with  $\boldsymbol{\theta} \in \mathbb{R}^p$  representing the underlying model parameter,  $a_n \in \mathcal{A}$  denoting the experiment selected at time  $n$ , and  $\mathcal{A}$  being a finite set encompassing all possible experiment choices. For example, in the context of CAT,  $\boldsymbol{\theta}$  corresponds to the latent proficiency level of a test-taker on  $p$  subjects or skills,  $a_n$  indicates the  $n$ -th test item,  $\mathcal{A}$  indicates the item bank which collects all the potential test items, and  $X_n \in \{0, 1\}$  indicates that whether the test-taker answers the  $n$ -th question correctly or not. At each time step  $n$ , a decision maker needs to select an experiment  $a_n$  based on the past observations  $X_1, a_1, X_2, a_2, \dots, X_{n-1}, a_{n-1}$ , sample  $X_n$  accordingly, and construct an estimator  $\hat{\boldsymbol{\theta}}_n$  for estimating  $\boldsymbol{\theta}$ . The goal is to find a good adaptive experiment selection rule and an estimator  $\hat{\boldsymbol{\theta}}_N$  so that  $\hat{\boldsymbol{\theta}}_N$  is as accurate as possible, where  $N$  could be a fixed sample size or a random stopping time depending on the application.

Greedy information-based experiment selection rules that maximize one-step-ahead information gain have been commonly adopted for item selection in CAT (see, e.g., Chang and Ying (1996); Cheng (2009); Van Der Linden (1999); Wang and Chang (2011)). For example, Wang and Chang (2011) and Tu et al. (2018) describe the following experiment selection rule:

$$a_{n+1} = \arg \min_{a \in \mathcal{A}} \text{tr} \left[ \left\{ \mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n, a) \right\}^{-1} \right], \quad (1)$$

where  $\mathbf{a}_n = (a_1, \dots, a_n)$  denotes the experiments selected up to time  $n$ ,

$$\hat{\boldsymbol{\theta}}_n^{\text{ML}} = \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^n \log f_{\boldsymbol{\theta}, a_i}(X_i)$$

denotes the maximum likelihood estimator (MLE) with  $n$  observations,

$$\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n, a) = \frac{1}{n+1} \left\{ \sum_{i=1}^n \mathcal{I}_{a_i}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}) + \mathcal{I}_a(\hat{\boldsymbol{\theta}}_n^{\text{ML}}) \right\}$$

represents the rescaled Fisher information matrix associated with the first  $n$  experiments and one extra experiment  $a$ , while  $\mathcal{I}_a(\boldsymbol{\theta}) = \mathbb{E}_{X \sim f_{\boldsymbol{\theta}, a}}[\nabla \log f_{\boldsymbol{\theta}, a}(X) \{\nabla \log f_{\boldsymbol{\theta}, a}(X)\}^T]$  denotes the Fisher information matrix associated with the experiment  $a$  at the parameter  $\boldsymbol{\theta}$ . Other

experiment selection rules in a similar form (e.g., substituting the trace function with other functions like  $\log \det(\cdot)$ ) are also explored in Wang et al. (2011).

These information-based experiment selection rules offer several benefits. First, the selection processes only require the calculation of the Fisher information and are easy to implement. Moreover, they are inherently parallelizable, offering scalability when  $|\mathcal{A}|$  is large. Second, they quantify the information gain associated with each experiment, thereby providing priority scores for them. This feature enables extension of these rules to various contexts and tasks (e.g.,  $\mathcal{A}$  varies over time). Additionally, given a parametric model, these rules can readily address problems in other applications.

Despite the computational advantages and wide applicability, the statistical analysis of greedy information-based experiment selection methods is limited to the one-dimensional case ( $p = 1$ ) in existing research. In this context, Chang and Ying (2009) established the consistency, asymptotic normality and optimality results for the MLE, and discussed the application in CAT. However, the multidimensional ( $p > 1$ ) case remains an open problem, partly due to the challenges regarding the combinatorial nature of the multidimensional problem and the seemingly limited capacity of greedy methods. The following example, which mimics the settings of an educational test measuring two latent traits, illustrates that one has to combine experiments carefully in order to obtain a consistent and/or risk-optimal estimator.

*Example 1.* Let  $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$  and  $\mathcal{A} = \{1, 2, 3\}$ . Let  $f_{\boldsymbol{\theta},a}$  be the probability mass function for Bernoulli variables with the probability parameter  $(1 + \exp(-\theta_1 + 0.1))^{-1}$ ,  $(1 + \exp(-\theta_2))^{-1}$ , and  $(1 + \exp(-\theta_1/2 - \theta_2))^{-1}$ , for  $a = 1, 2, 3$ , respectively. Let  $n_k$  be the number of times that experiment  $k$  is selected and  $\pi_k = n_k/n$  be its frequency ( $k = 1, 2, 3$ ) with  $n = \sum_k n_k$ . Then, a necessary condition for the existence of a consistent estimator  $\hat{\boldsymbol{\theta}}_n$  is  $\max\{\min(n_1, n_2), \min(n_1, n_3), \min(n_2, n_3)\} \rightarrow \infty$ . Moreover, in order to minimize the mean squared error  $\mathbb{E}\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\|^2$  asymptotically, a necessary condition is  $(\pi_1, \pi_2, \pi_3) \rightarrow \boldsymbol{\pi}^*(\boldsymbol{\theta})$  as the total sample size grows, where  $\boldsymbol{\pi}^*(\boldsymbol{\theta})$  is a vector-valued optimal proportion function depending on  $\boldsymbol{\theta}$ . See Figure 1 for an illustration of the function  $\boldsymbol{\pi}^*(\boldsymbol{\theta})$  and additional details in Section 5.2.

In this example, achieving consistent or asymptotically optimal estimators requires experiments to be combined carefully with a parameter-dependent frequency. However, information-based selection methods, being one-step-ahead greedy, do not consider the benefits of combining experiments or multi-step planning. Thus, it remains an open question whether these selection methods lead to consistent, asymptotically normal, or risk-optimal estimators.

In this study, we provide a definitive answer to the above question for a class of greedy-information-based experiment selection rules. In particular, we introduce two experiment

selection rules based on a pre-specified criterion function  $\mathbb{G}_{\boldsymbol{\theta}} : \mathbb{R}^{p \times p} \rightarrow \mathbb{R}$ ,

$$\text{GI0 : } a_{n+1} = \arg \min_{a \in \mathcal{A}} \mathbb{G}_{\hat{\boldsymbol{\theta}}_n^{\text{ML}}} \left[ \left\{ \mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n, a) \right\}^{-1} \right], \text{ and} \quad (2)$$

$$\text{GI1 : } a_{n+1} = \arg \max_{a \in \mathcal{A}} \text{tr} \left[ \nabla \mathbb{G}_{\hat{\boldsymbol{\theta}}_n^{\text{ML}}}(\hat{\boldsymbol{\Sigma}}_n) \hat{\boldsymbol{\Sigma}}_n \mathcal{I}_a(\hat{\boldsymbol{\theta}}_n^{\text{ML}}) \hat{\boldsymbol{\Sigma}}_n \right], \quad (3)$$

where  $\hat{\boldsymbol{\Sigma}}_n = \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1}$ ,  $\nabla \mathbb{G}_{\boldsymbol{\theta}}(\boldsymbol{\Sigma}) = \left( \frac{\partial \mathbb{G}_{\boldsymbol{\theta}}(\boldsymbol{\Sigma})}{\partial \Sigma_{ij}} \right)_{1 \leq i, j \leq p}$  denotes the gradient of  $\mathbb{G}_{\boldsymbol{\theta}}$  with respect to its matrix input and recall  $\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n) = \frac{1}{n} \sum_{i=1}^n \mathcal{I}_{a_i}(\hat{\boldsymbol{\theta}}_n^{\text{ML}})$ . We refer to the selection rule in (2) as the zero-order greedy information-based selection rule (GI0), and that in (3) as the first-order greedy information-based selection rule (GI1), because GI0 is minimizing a certain function of the Fisher information at the next time point, while GI1 is derived based on a first-order Taylor expansion of GI0; see Section 3 for more details. GI0 generalizes the selection rule in (1), accommodating more diverse settings. New methods can be obtained by specifying an appropriate function  $\mathbb{G}_{\boldsymbol{\theta}}$ . GI1 offers a class of new experiment selection rules which share similar asymptotic properties as GI0 but are computationally more efficient when both  $p$  and  $|\mathcal{A}|$  are large.

Our main theoretical contributions are as follows. First, we show that MLE is strongly consistent and asymptotically normal when using GI0 or GI1 as the experiment selection rule, under mild conditions. Second, we derive the asymptotic covariance matrix of the MLE as a function involving  $\mathbb{G}_{\boldsymbol{\theta}}(\cdot)$  and the Fisher information. Third, we prove that the empirical frequency of selected experiments converges to a limiting frequency. Fourth, we show that the experiment selection rule GI0 (or GI1) combined with the MLE is asymptotically optimal in minimizing certain risk measures related to  $\mathbb{G}_{\boldsymbol{\theta}}(\cdot)$ . In particular, if  $\mathbb{G}_{\boldsymbol{\theta}}(\cdot) = \text{tr}(\cdot)$ , then the MLE has the smallest asymptotic mean squared error (MSE), when compared with other experiment selection rules and estimators. Moreover, these results are valid not only for fixed sample sizes, but also for random stopping times, which is beneficial for applications that use early stopping criteria.

Beyond the methodological and theoretical contributions, we have developed new analytical tools for addressing technical challenges. For example, we show that the inverted Fisher information, through its directional derivatives in experiment selection rules, acts as a regularizer. This facilitates automatic exploration of necessary experiments, removing the need for additional exploration steps traditionally employed in stochastic control methods for related problems (e.g., two-stage design in sequential design for hypothesis tests (Chernoff, 1959; Naghshvar and Javidi, 2013)). Furthermore, we show that a class of greedy and stochastic optimization methods converges to the minimum of a convex function over a simplex almost surely. In addition, we refine and extend several classic results in stochastic

analysis, such as Anscombe’s theorem (Anscombe, 1952) and the Robbins-Siegmund theorem (Robbins and Siegmund, 1971). These theoretical results and technical tools are important in their own right and may be reused in other related problems. See Section 6 for more details of the technical challenges and our new analytical tools.

The rest of the paper is organized as follows. Section 2 formalizes the active sequential estimation problem. Section 3 introduces the greedy information-based experiment selection rules GI0 and GI1, elaborating on their implementation. Section 4 offers the main theoretical results regarding the MLE and the experiment selection rules. Section 5 details the methods and theory in applications including the item selection in CAT and sequential rank aggregation. Section 6 gives new analytical tools and a proof sketch. Section 7 presents two simulation studies, which illustrate the finite sample performance and the computational efficiency of the proposed methods. Section 8 showcases the performance of the proposed method on a real-data example. Section 9 summarizes the main results and provides discussions on future directions. All the technical proofs for the theoretical results and additional simulation results are given in the supplementary material.

## 1.1 Notations

In this paper, we use the following notations and mathematical conventions. Let  $\overline{C}$  and  $\underline{C}$  represent generic constants that are bounded from above and below, respectively. These generic constants are independent of  $\boldsymbol{\theta}$  and  $a \in \mathcal{A}$ , and their values may vary from place to place. Let  $|\mathcal{A}|$  denote the cardinality of a set  $\mathcal{A}$ . Let  $I(\cdot)$  denote the indicator function. Let  $I_p$  denote the  $p \times p$  identity matrix. The inner product between real matrices (or vectors)  $\mathbf{A}$  and  $\mathbf{B}$  of the same size is defined by  $\langle \mathbf{A}, \mathbf{B} \rangle = \text{tr}(\mathbf{A}^T \mathbf{B})$ . For a real matrix  $\mathbf{A}$ , define the operator norm  $\|\mathbf{A}\|_{op}$  as the maximum singular value of  $\mathbf{A}$ . For a vector  $\mathbf{x}$ , denote its Euclidean norm by  $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ . For a symmetric matrix  $\mathbf{A}$ ,  $\lambda_{max}(\mathbf{A})$ ,  $\lambda_{min}(\mathbf{A})$ , and  $\kappa(\mathbf{A})$  denote its maximum eigenvalue, minimum eigenvalue, and condition number, respectively. If  $\mathbf{A}$  is a positive definite matrix, then  $\kappa(\mathbf{A}) = \frac{\lambda_{max}(\mathbf{A})}{\lambda_{min}(\mathbf{A})}$ . For a differentiable matrix function  $\mathbb{G}(\boldsymbol{\Sigma})$ , its gradient is denoted by  $\nabla \mathbb{G}(\boldsymbol{\Sigma})$ , and is defined as the matrix such that  $\mathbb{G}(\boldsymbol{\Sigma} + \Delta \boldsymbol{\Sigma}) - \mathbb{G}(\boldsymbol{\Sigma}) = \langle \nabla \mathbb{G}(\boldsymbol{\Sigma}), \Delta \boldsymbol{\Sigma} \rangle + o(\|\Delta \boldsymbol{\Sigma}\|)$ . For symmetric matrices  $\mathbf{A}$  and  $\mathbf{B}$ , define the partial order  $\mathbf{A} \preceq \mathbf{B}$  if and only if  $\mathbf{B} - \mathbf{A}$  is a positive semidefinite matrix. Throughout the paper, all the vectors are column vectors, unless otherwise specified.

## 2 Problem Statement

Let  $X_1, \dots, X_n, \dots$  be data collected sequentially,  $\mathcal{A}$  be a finite set with cardinality  $k$ , and  $a_1, \dots, a_n, \dots \in \mathcal{A}$  be the experiments selected at different time points. Denote by  $\mathcal{F}_n = \sigma(a_1, X_1, \dots, a_n, X_n)$ , the sigma field that contains information of the observations and the selected experiments up to time  $n$ . At each time  $n$ , a decision maker needs to select the experiment  $a_{n+1}$  adaptively based on past information. That is,  $a_{n+1}$  is measurable with respect to  $\mathcal{F}_n$ . Throughout the study, we assume that the distribution of  $X_{n+1}$  satisfies

$$X_{n+1}|\mathcal{F}_n \sim f_{\theta, a_{n+1}}(\cdot) \text{ for } \theta \in \Theta \subset \mathbb{R}^p$$

where  $\theta$  is a  $p$ -dimensional model parameter,  $\Theta$  is a compact parameter space and  $f_{\theta, a_{n+1}}(\cdot)$  denotes the probability density of  $X_{n+1}$  with respect to a baseline measure. That is,  $X_{n+1}$  is assumed to follow a parametric model, and its distribution is determined by both the underlying model parameter  $\theta$  and the selected experiment  $a_{n+1}$ .

In an active sequential estimation problem, the goal is to design an experiment selection rule for  $\{a_n\}_{n \geq 1}$  and find an estimator  $\hat{\theta}_n$  that is measurable with respect to  $\mathcal{F}_n$ , so that  $\hat{\theta}_n$  is close to the true underlying parameter  $\theta^*$  with high probability. In some applications, the data collection process may be stopped early to save for the sampling cost. In these cases, we are also interested in  $\hat{\theta}_N$ , where  $N$  is a random stopping time.

## 3 Methods

For the estimation method, we focus on the MLE, although some of the methods and theoretical results may be extended to other estimators. The definition of MLE is given as follows. Let the selected experiments up to time  $n$  be  $\mathbf{a}_n = (a_1, \dots, a_n)$ . Then, the rescaled log-likelihood and the corresponding MLE are

$$l_n(\theta) = l_n(\theta; \mathbf{a}_n) = \frac{1}{n} \sum_{i=1}^n \log f_{\theta, a_i}(X_i), \text{ and} \quad (4)$$

$$\hat{\theta}_n^{\text{ML}} \in \arg \max_{\theta \in \Theta} l_n(\theta; \mathbf{a}_n). \quad (5)$$

We propose adopting two experimental selection rules, including the zero-order greedy information-based selection rule GI0 and the first-order greedy information-based selection rule GI1. The precise description of these methods are given in Algorithms 1 and 2.

We explain steps in Algorithms 1 and 2. First, we note that both algorithms require a pre-specified criterion function  $\mathbb{G}_\theta : \mathbb{R}^{p \times p} \rightarrow \mathbb{R}$ . Motivated by Kiefer (1974) on the design of

---

**Algorithm 1** GI0 Algorithm

---

- 1: **Input:**  $\hat{\theta}_0, a_1^0, \dots, a_{n_0}^0$ .
  - 2: **Require:**  $\hat{\theta}_0 \in \Theta, a_1^0, \dots, a_{n_0}^0 \in \mathcal{A}$  such that  $\sum_{i=1}^{n_0} \mathcal{I}_{a_i^0}(\hat{\theta}_0)$  is nonsingular.
  - 3: **Initialization:**  $a_1 = a_1^0, \dots, a_{n_0} = a_{n_0}^0$ , collecting responses  $X_1, X_2, \dots, X_{n_0}$  correspondingly.
  - 4: **for**  $n = n_0$  to  $N$  **do**
  - 5:   calculating the MLE  $\hat{\theta}_n^{\text{ML}}$  according to equation (5)
  - 6:   selecting experiment  $a_{n+1}$  according to equation (2)
  - 7:   collecting response  $X_{n+1}$  corresponding to the selected experiment  $a_{n+1}$
  - 8: **end for**
  - 9: **Output:**  $\hat{\theta}_N^{\text{ML}}$
- 

---

**Algorithm 2** GI1 Algorithm

---

- 1: **Input:**  $\hat{\theta}_0, a_1^0, \dots, a_{n_0}^0$ .
  - 2: **Require:**  $\hat{\theta}_0 \in \Theta, a_1^0, \dots, a_{n_0}^0 \in \mathcal{A}$  such that  $\sum_{i=1}^{n_0} \mathcal{I}_{a_i^0}(\hat{\theta}_0)$  is nonsingular.
  - 3: **Initialization:**  $a_1 = a_1^0, \dots, a_{n_0} = a_{n_0}^0$ , collecting responses  $X_1, X_2, \dots, X_{n_0}$  correspondingly.
  - 4: **for**  $n = n_0$  to  $N$  **do**
  - 5:   calculating the MLE  $\hat{\theta}_n^{\text{ML}}$  according to equation (5)
  - 6:   selecting experiment  $a_{n+1}$  according to equation (3)
  - 7:   collecting response  $X_{n+1}$  corresponding to the selected experiment  $a_{n+1}$
  - 8: **end for**
  - 9: **Output:**  $\hat{\theta}_N^{\text{ML}}$
- 

experiments, a reasonable choice is

$$\mathbb{G}_{\theta}(\Sigma) = \Phi_q(\Sigma) = \begin{cases} \log \det(\Sigma), & \text{if } q = 0; \\ \text{tr}(\Sigma^q), & \text{if } 0 < q < 1; \\ (\text{tr}(\Sigma^q))^{1/q}, & \text{if } q \geq 1, \end{cases} \quad (6)$$

for a prespecified  $q \geq 0$ . In the context of adaptive item selection in CAT, GI0 with  $\mathbb{G}_{\theta}(\Sigma) = \Phi_q(\Sigma)$  has been adopted in Van Der Linden (1999) and Wang and Chang (2011). In particular, the selection rule in (1) corresponds to GI0 with  $\mathbb{G}_{\theta}(\cdot) = \Phi_1(\cdot)$ . Both GI0 and GI1 are relatively new in other applications described in Section 1. Note that for  $\mathbb{G}_{\theta}(\Sigma) = \Phi_q(\Sigma)$ , it is a function independent with the input  $\theta$ . Another option is  $\mathbb{G}_{\theta}(\Sigma) = \text{tr}(\mathbf{H}_{\theta}\Sigma)$ , where  $\mathbf{H}_{\theta}$  is a positive definite matrix depending on  $\theta$ . This criterion function is useful in the cases where we would like to assign different weights to different values of  $\theta$ . For more details, please refer to Theorem 4.8.

Second, both algorithms require an initialization step where  $n_0$  experiments are selected so that the Fisher information matrix  $\mathcal{I}(\hat{\theta}_0; \mathbf{a}_{n_0})$  is nonsingular. This initialization step



ensures that  $\mathcal{I}(\hat{\boldsymbol{\theta}}_n; \mathbf{a}_n)$  is nonsingular and the experiment selection rules in (2) and (3) are well-defined for all  $n \geq n_0$ . In practice, it is usually straightforward to find such  $a_1^0, \dots, a_{n_0}^0$ . For instance, in Example 1, we could choose  $\hat{\boldsymbol{\theta}}_0 = (0, 0)$ ,  $(a_1^0, a_2^0) = (1, 2)$ , and  $n_0 = 2$ . Then, at each time point, the algorithm first calculates the MLE based on the available information, selects a new experiment according to (2) for GI0 (or (3) for GI1), and then samples a new observation according to the selected experiment.

We refer to the selection rule in Algorithm 1 as GI0 and that in Algorithm 2 as GI1, because GI0 tries to minimize the criterion function  $\mathbb{G}_{\hat{\boldsymbol{\theta}}_n} \left[ \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n, a)\}^{-1} \right]$  for one-step ahead, while GI1 tries to minimize its first-order approximation, i.e.,

$$\begin{aligned} & \mathbb{G}_{\hat{\boldsymbol{\theta}}_n^{\text{ML}}} \left[ \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n, a)\}^{-1} \right] - \mathbb{G}_{\hat{\boldsymbol{\theta}}_n^{\text{ML}}} \left[ \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1} \right] \\ & \approx \left\langle \bar{\boldsymbol{\pi}}_{n+1}^a - \bar{\boldsymbol{\pi}}_n, \frac{\partial}{\partial \boldsymbol{\pi}} \mathbb{G}_{\hat{\boldsymbol{\theta}}_n^{\text{ML}}} \left[ \left\{ \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\hat{\boldsymbol{\theta}}_n^{\text{ML}}) \right\}^{-1} \right] \Big|_{\boldsymbol{\pi} = \bar{\boldsymbol{\pi}}_n} \right\rangle \\ & = -\frac{1}{n+1} \text{tr} \left[ \nabla \mathbb{G}_{\hat{\boldsymbol{\theta}}_n^{\text{ML}}} \left( \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1} \right) \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1} \mathcal{I}_a(\hat{\boldsymbol{\theta}}_n^{\text{ML}}) \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1} \right] \\ & \quad + \frac{1}{n+1} \text{tr} \left[ \nabla \mathbb{G}_{\hat{\boldsymbol{\theta}}_n^{\text{ML}}} \left( \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1} \right) \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1} \right], \end{aligned} \tag{7}$$

where the empirical frequency vector  $\bar{\boldsymbol{\pi}}_n$  is defined as

$$\bar{\boldsymbol{\pi}}_n = \bar{\boldsymbol{\pi}}_n[\mathbf{a}_n] = \left( \frac{1}{n} |\{i; a_i = a, 1 \leq i \leq n\}| \right)_{a \in \mathcal{A}}, \tag{8}$$

with  $\mathbf{a}_n = (a_1, \dots, a_n)$  collects experiments selected up to time  $n$ , and  $\bar{\boldsymbol{\pi}}_{n+1}^a(a') = \frac{n}{n+1} \bar{\boldsymbol{\pi}}_n(a') + \frac{1}{n+1} I(a = a')$  is the empirical frequency at the time  $n+1$  if  $a'$  is selected at that time. Note that GI0 minimizes the first line of (7), GI1 minimizes the first term on the last equation of (7), and the second term on the last equation of (7) does not depend on the choice of experiment  $a$ . This suggests that GI0 and GI1 are asymptotically equivalent, although the rigorous theoretical justification is much more involved.

### 3.1 Improving Computational Efficiency

If  $k, p$  are large, and  $\mathcal{I}_a(\boldsymbol{\theta})$  has some low-dimensional representation, GI1 can be implemented with improved numerical efficiency. In particular, we consider two specific cases which are commonly seen in applications, including (1) low-rank information:  $\mathcal{I}_a(\boldsymbol{\theta}) = L_a(\boldsymbol{\theta}) L_a^T(\boldsymbol{\theta})$  where  $L_a(\boldsymbol{\theta}) \in \mathbb{R}^{p \times s}$  for all  $a$  and  $\boldsymbol{\theta}$  and  $s < p$ ; (2) sparse and low-rank information:  $L_a(\boldsymbol{\theta})$  has no more than  $s$  non-zero rows. For these cases, Algorithm 2 can be implemented using the following accelerated version.

---

**Algorithm 3** Accelerated GI1 Algorithm

---

We modify line 6 in Algorithms 2, while keeping the other lines of the algorithms unchanged.

6: selecting experiment  $a_{n+1}$  according to

$$\mathbf{M} = \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1} \nabla \mathbb{G}_{\hat{\boldsymbol{\theta}}_n^{\text{ML}}}(\{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1}) \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1},$$
$$a_{n+1} = \arg \max_{a \in \mathcal{A}} \text{tr} \left[ L_a^T(\hat{\boldsymbol{\theta}}_n^{\text{ML}}) \mathbf{M} L_a(\hat{\boldsymbol{\theta}}_n^{\text{ML}}) \right].$$

---

**Lemma 3.1.** *Assume the computational complexity of evaluating  $\mathbb{G}_{\boldsymbol{\theta}}(\boldsymbol{\Sigma})$  and  $\nabla \mathbb{G}_{\boldsymbol{\theta}}(\boldsymbol{\Sigma})$  is no more than  $O(p^3)$ . Given the MLE  $\hat{\boldsymbol{\theta}}_n^{\text{ML}}$  and  $\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)$ , we have*

1. *the computational complexity for each iteration in GI0 is of the order  $O(kp^3)$ ;*
2. *the computational complexity for each iteration in the accelerated GI1 Algorithm 3 is  $O(ksp^2 + p^3)$ , assuming that the information matrices are low-rank matrices with given  $L_a(\boldsymbol{\theta}) \in \mathbb{R}^{p \times s}$ . Moreover, the computational cost for the accelerated GI1 Algorithm 3 becomes  $O(ks^2p + p^3)$  if  $L_a(\boldsymbol{\theta})$  has no more than  $s$  non-zero rows.*

According to the above lemma, the accelerated GI1 algorithm is computationally much more efficient than GI0, when  $k$  and  $p$  are large and  $s$  is small. Numerical results supporting these findings can be found in Section 7.2.

### 3.2 Early Stopping

In many applications, the data collection process is stopped early when sufficient observations have been gathered to make accurate statistical inference. For instance, in the context of CAT, educational tests often have variable lengths determined by specific early stopping rules. These rules generally lead to less fatigue and a better experience for examinees. In this section, we introduce two early stopping rules suitable for active sequential estimation.

The first stopping rule  $\tau_c^{(1)}$  is concerned with the estimation of a differentiable function of the parameter  $h(\boldsymbol{\theta}) \in \mathbb{R}$ , and it is defined as

$$\tau_c^{(1)} = \min \{m \geq n_0; \widehat{\text{SE}}(h(\hat{\boldsymbol{\theta}}_m^{\text{ML}})) \leq c\}, \quad (9)$$

where  $\widehat{\text{SE}}^2(h(\hat{\boldsymbol{\theta}}_m^{\text{ML}})) = \frac{1}{m} \{\nabla h(\hat{\boldsymbol{\theta}}_m^{\text{ML}})\}^T \{\mathcal{I}(\hat{\boldsymbol{\theta}}_m^{\text{ML}}; \mathbf{a}_m)\}^{-1} \nabla h(\hat{\boldsymbol{\theta}}_m^{\text{ML}})$ . The second stopping rule  $\tau_c^{(2)}$  is concerned with the estimation of the vector  $\boldsymbol{\theta}$ , and is defined as

$$\tau_c^{(2)} = \min \left\{ m \geq n_0; \widehat{\text{MSE}}(\hat{\boldsymbol{\theta}}_m^{\text{ML}}) \leq c \right\}, \quad \text{where } \widehat{\text{MSE}}(\hat{\boldsymbol{\theta}}_m^{\text{ML}}) = \frac{1}{m} \text{tr} \left( \{\mathcal{I}(\hat{\boldsymbol{\theta}}_m^{\text{ML}}; \mathbf{a}_m)\}^{-1} \right). \quad (10)$$

Here,  $\widehat{\text{SE}}$  serves as an approximation of  $\text{sd}(h(\widehat{\boldsymbol{\theta}}_m^{\text{ML}}))$  and  $\widehat{\text{MSE}}(\widehat{\boldsymbol{\theta}}_m^{\text{ML}})$  serves as an approximation of  $\text{MSE}(\widehat{\boldsymbol{\theta}}_m^{\text{ML}}) = \mathbb{E}_{\boldsymbol{\theta}^*} \|\widehat{\boldsymbol{\theta}}_m^{\text{ML}} - \boldsymbol{\theta}^*\|^2$ . Both rules terminate the data collection process once a certain error estimator falls below a predetermined threshold  $c$ .

## 4 Theoretical Results

In this section, we first introduce the regularity conditions, and then present the main theoretical results regarding the consistency, asymptotic normality, and the optimality of the proposed method.

### 4.1 Regularity Conditions

Throughout Section 4, we make the following Assumptions 1–5, along with Assumptions 6A and 7A, and we will refer to this set of assumptions as the ‘regularity conditions’. All the theoretical results still hold when 6A and 7A are replaced with the more relaxed Assumptions 6B and 7B.

**Assumption 1.** The parameter space  $\boldsymbol{\Theta}$  is a non-empty compact and convex subset of  $\mathbb{R}^p$ . The true parameter  $\boldsymbol{\theta}^*$  is an interior point of  $\boldsymbol{\Theta}$ .

**Assumption 2.** The support of the probability density  $f_{\boldsymbol{\theta},a}$ , denoted as  $\text{supp}(f_{\boldsymbol{\theta},a})$ , depends only on  $a$  and does not depend on  $\boldsymbol{\theta}$ , where the support of a function is defined as

$$\text{supp}(f_{\boldsymbol{\theta},a}) = \text{cl}\{x^a; f_{\boldsymbol{\theta},a}(x^a) > 0\},$$

and  $\text{cl}(S)$  denotes the closure of a set  $S$ . Moreover, for all  $a \in \mathcal{A}$  and  $X^a \in \text{supp}(f_{\boldsymbol{\theta},a})$ , the gradient  $\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta},a}(X^a) = (\frac{\partial \log f_{\boldsymbol{\theta},a}(X^a)}{\partial \theta_i})_{1 \leq i \leq p}$  and the Hessian matrix  $\nabla_{\boldsymbol{\theta}}^2 \log f_{\boldsymbol{\theta},a}(X^a) = (\frac{\partial^2 \log f_{\boldsymbol{\theta},a}(X^a)}{\partial \theta_i \partial \theta_j})_{1 \leq i, j \leq p}$  exist, where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$ . Assume that there exist functions  $\Psi_1^a$  and  $\Psi_2^a$  satisfying  $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}_{X^a \sim f_{\boldsymbol{\theta},a}} \{\Psi_1^a(X^a)\}^2 < \infty$ ,  $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}_{X^a \sim f_{\boldsymbol{\theta},a}} \Psi_2^a(X^a) < \infty$ ,

$$\|\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}_1,a}(X^a) - \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}_2,a}(X^a)\| \leq \Psi_1^a(X^a) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|, \text{ and} \quad (11)$$

$$\|\nabla_{\boldsymbol{\theta}}^2 \log f_{\boldsymbol{\theta}_1,a}(X^a) - \nabla_{\boldsymbol{\theta}}^2 \log f_{\boldsymbol{\theta}_2,a}(X^a)\|_{op} \leq \Psi_2^a(X^a) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|, \quad (12)$$

for all  $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$  and  $a \in \mathcal{A}$ . Furthermore, for all  $a \in \mathcal{A}$ ,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}_{X \sim f_{\boldsymbol{\theta}^*,a}} \{\|\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta},a}(X)\|^2\} < \infty \text{ and } \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}_{X \sim f_{\boldsymbol{\theta}^*,a}} \{\|\nabla_{\boldsymbol{\theta}}^2 \log f_{\boldsymbol{\theta},a}(X)\|_{op}\} < \infty.$$

**Assumption 3.** The Fisher information matrices satisfy the following conditions:

$$\mathcal{I}_a(\boldsymbol{\theta}) = \mathbb{E}_{X \sim f_{\boldsymbol{\theta},a}} [\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta},a}(X) \{\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta},a}(X)\}^T] = -\mathbb{E}_{X \sim f_{\boldsymbol{\theta},a}} \{\nabla_{\boldsymbol{\theta}}^2 \log f_{\boldsymbol{\theta},a}(X)\},$$

and those Fisher information matrices are continuously differentiable with respect to  $\boldsymbol{\theta}$  for all  $a \in \mathcal{A}$ . Furthermore,  $\sum_{a \in \mathcal{A}} \mathcal{I}_a(\boldsymbol{\theta})$  is positive definite for every  $\boldsymbol{\theta} \in \Theta$ .

**Assumption 4.** Let  $M(\boldsymbol{\theta}; \boldsymbol{\pi}) = \sum_{a \in \mathcal{A}} \pi(a) \mathbb{E}_{X \sim f_{\boldsymbol{\theta},a}} \{\log f_{\boldsymbol{\theta},a}(X)\}$  for  $\boldsymbol{\pi} = (\pi(a))_{a \in \mathcal{A}}$ . Assume the following uniform law of large numbers holds for all sequence  $\mathbf{a}_n = (a_1, \dots, a_n)$  such that  $a_i$  is measurable with respect to  $\mathcal{F}_{i-1}$ , for all  $1 \leq i \leq n$ :

$$\mathbb{P} \left\{ \lim_{n \rightarrow \infty} \sup_{\boldsymbol{\theta} \in \Theta} |l_n(\boldsymbol{\theta}; \mathbf{a}_n) - M(\boldsymbol{\theta}; \bar{\boldsymbol{\pi}}_n[\mathbf{a}_n])| = 0 \right\} = 1, \quad (13)$$

where  $\bar{\boldsymbol{\pi}}_n[\mathbf{a}_n] = (\bar{\pi}_n(a; \mathbf{a}_n))_{a \in \mathcal{A}}$ , and  $\bar{\pi}_n(a; \mathbf{a}_n) = \frac{1}{n} |\{i; a_i = a, 1 \leq i \leq n\}|$  denotes the empirical frequency that the experiment  $a$  is selected up to time  $n$ .

**Assumption 5.** The criterion function  $\mathbb{G}_{\boldsymbol{\theta}}$  takes one of the following forms:

1.  $\mathbb{G}_{\boldsymbol{\theta}}(\cdot) = \Phi_q(\cdot)$  for some  $q \geq 0$  where  $\Phi_q(\cdot)$  is defined in (6), or
2. the function  $\mathbb{G}_{\boldsymbol{\theta}}(\cdot) : \mathbb{R}^{p \times p} \mapsto \mathbb{R}$  is convex, and it satisfies: for all positive definite matrix  $\boldsymbol{\Sigma}$ ,  $\nabla_{\boldsymbol{\theta}} \nabla \mathbb{G}_{\boldsymbol{\theta}}(\boldsymbol{\Sigma})$  and  $\nabla^2 \mathbb{G}_{\boldsymbol{\theta}}(\boldsymbol{\Sigma})$  are continuous in  $(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ ; and for all positive definite matrices satisfying  $\mathbf{A} \succeq \mathbf{B}$ , we have  $\mathbb{G}_{\boldsymbol{\theta}}(\mathbf{A}) \geq \mathbb{G}_{\boldsymbol{\theta}}(\mathbf{B})$ . Additionally,  $\sup_{\mathbf{A}} \kappa(\nabla \mathbb{G}_{\boldsymbol{\theta}}(\mathbf{A})) < \infty$  and  $\lim_{\lambda_{\max}(\mathbf{A}) \rightarrow \infty} \inf_{\boldsymbol{\theta} \in \Theta} \mathbb{G}_{\boldsymbol{\theta}}(\mathbf{A}) = \infty$ .

**Assumption 6A** (Reparametrization). There exist matrices  $\{\mathbf{Z}_a\}_{a \in \mathcal{A}}$  and probability density functions  $\{h_{\mathbf{Z}_a \boldsymbol{\theta}, a}(\cdot)\}_{a \in \mathcal{A}}$  satisfying the following requirements

1.  $\mathbf{Z}_a$  is a matrix of dimension  $p_a \times p$  with rank  $p_a$  and  $f_{\boldsymbol{\theta}, a}(\cdot) = h_{\mathbf{Z}_a \boldsymbol{\theta}, a}(\cdot)$  for all  $a \in \mathcal{A}$ .
2. Let  $\boldsymbol{\xi}_a = \mathbf{Z}_a \boldsymbol{\theta}$  be a reparametrization of  $\boldsymbol{\theta}$ . Assume that the Fisher information matrix of each experiment  $a$  is nonsingular with respect to  $\boldsymbol{\xi}_a$ . That is, the compressed Fisher information matrix

$$\begin{aligned} \mathcal{I}_{\boldsymbol{\xi}_a, a}(\boldsymbol{\xi}_a) &= \mathbb{E}_{X \sim h_{\boldsymbol{\xi}_a, a}} [\nabla_{\boldsymbol{\xi}_a} \log h_{\boldsymbol{\xi}_a, a}(X) \{\nabla_{\boldsymbol{\xi}_a} \log h_{\boldsymbol{\xi}_a, a}(X)\}^T] \\ &= -\mathbb{E}_{X \sim h_{\boldsymbol{\xi}_a, a}} \{\nabla_{\boldsymbol{\xi}_a}^2 \log h_{\boldsymbol{\xi}_a, a}(X)\} \end{aligned}$$

is nonsingular for all  $\boldsymbol{\theta} \in \Theta$ .

**Assumption 7A** (Identifiability). There exists a constant  $C > 0$  such that for all  $\boldsymbol{\theta} \in \Theta$ ,

$$D_{\text{KL}}(h_{\boldsymbol{\xi}_a^*, a} \| h_{\boldsymbol{\xi}_a, a}) \geq C \|\boldsymbol{\xi}_a^* - \boldsymbol{\xi}_a\|^2, \quad (14)$$

where  $\boldsymbol{\xi}_a^* = \mathbf{Z}_a \boldsymbol{\theta}^*$  is the compressed parameter after reparametrization, and  $D_{\text{KL}}(h_{\boldsymbol{\xi}_a^*, a} \| h_{\boldsymbol{\xi}_a, a})$  denotes the Kullback–Leibler divergence between the density functions  $h_{\boldsymbol{\xi}_a^*, a}$  and  $h_{\boldsymbol{\xi}_a, a}$ , and is defined as  $D_{\text{KL}}(h_{\boldsymbol{\xi}_a^*, a} \| h_{\boldsymbol{\xi}_a, a}) = \mathbb{E}_{X \sim h_{\boldsymbol{\xi}_a^*, a}} \log \left( \frac{h_{\boldsymbol{\xi}_a^*, a}(X)}{h_{\boldsymbol{\xi}_a, a}(X)} \right)$ .

We comment on the above regularity conditions. Assumptions 1, 2, 3 and 4 are extensions of standard regularity conditions for the consistency of the MLE based on independent and identically distributed (i.i.d.) observations (see, e.g., Chapter 5 of Van der Vaart (2000)). In particular, Assumption 1 ensures the existence of MLE. Assumption 2 requires that the gradient of log-density function associated with each experiment is stochastic Lipschitz and has a bounded second moment. Condition (12) can be replaced by a more relaxed condition:

$$\left\| \nabla_{\boldsymbol{\theta}}^2 \log f_{\boldsymbol{\theta}_1, a}(X^a) - \nabla_{\boldsymbol{\theta}}^2 \log f_{\boldsymbol{\theta}_2, a}(X^a) \right\|_{op} \leq \Psi_2^a(X^a) \psi(\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|), \quad (15)$$

where  $\psi : [0, \infty) \rightarrow [0, \infty)$  is a strictly increasing continuous function such that  $\psi(0) = 0$ . Assumption 3 requires that the Fisher information matrices are well-behaved. Under this assumption, each Fisher information matrix  $\mathcal{I}_a(\boldsymbol{\theta})$  may be singular, but their sum is nonsingular. In other words, if we combine all the experiments together, the Fisher information matrix is nonsingular. Assumption 4 requires that the log-likelihood follows the uniform law of large numbers. This assumption can be verified by uniform martingale laws of large numbers (see Rakhlin et al. (2015)) in most applications. Assumption 5 describes the requirement on the criterion function  $\mathbb{G}_{\boldsymbol{\theta}}(\cdot)$ . Assumptions 6A and 7A require that for each experiment  $a$ , we can reparameterize the model with a new parameter  $\boldsymbol{\xi}_a$  with possibly lower dimension  $p_a \leq p$  such that  $\boldsymbol{\xi}_a$  is locally identifiable around the true model parameter, and the Fisher information matrix with respect to  $\boldsymbol{\xi}_a$  is nonsingular. Note that Fisher information with respect to  $\boldsymbol{\theta}$  may be singular in this case.

All the regularity assumptions are easily satisfied in practical problems, including the item selection in CAT and the sequential rank aggregation problem described in Section 1. See Section 5 for detailed justifications of the assumptions in these applications. Note that 6A and 7A can be relaxed to a more general condition, allowing for non-linear model reparameterization. These relaxed conditions are provided below.

**Assumption 6B.** For  $Q \subset \mathcal{A}$ , define a vector space  $V_Q = V_Q(\boldsymbol{\theta}) = \sum_{a \in Q} \mathcal{R}(\mathcal{I}_a(\boldsymbol{\theta}))$ , where  $\mathcal{R}(\mathbf{A})$  represents the column space of a matrix  $\mathbf{A}$ . Assume that the dimension  $\dim(V_Q(\boldsymbol{\theta}))$  does not depend on  $\boldsymbol{\theta}$ , and there exist constants  $0 < \underline{c} \leq \bar{c} < \infty$ , which do not depend on  $Q$  and  $\boldsymbol{\theta}$ , such that for all  $Q \subset \mathcal{A}$  and  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$\underline{c} \cdot \mathbf{P}_{V_Q(\boldsymbol{\theta})} \preceq \sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}) \preceq \bar{c} \cdot \mathbf{P}_{V_Q(\boldsymbol{\theta})}, \quad (16)$$

where  $\mathbf{P}_{V_Q(\boldsymbol{\theta})}$  denotes the orthogonal projection matrix onto vector space  $V_Q(\boldsymbol{\theta})$ .

**Assumption 7B.** Let  $\mathcal{S}^{\mathcal{A}} = \{\boldsymbol{\pi} = (\pi(a))_{a \in \mathcal{A}} : \sum_{a \in \mathcal{A}} \pi(a) = 1 \text{ and } \pi(a) \geq 0 \text{ for all } a \in \mathcal{A}\}$  denote the simplex in  $\mathbb{R}^{\mathcal{A}}$ . Assume that there exists a positive constant  $C$  such that for all  $\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}$  and  $\boldsymbol{\theta} \in \Theta$ ,

$$\sum_{a \in \mathcal{A}} \pi(a) D_{\text{KL}}(f_{\boldsymbol{\theta}^*, a} \| f_{\boldsymbol{\theta}, a}) \geq C \sum_{a \in \mathcal{A}} \pi(a) (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathcal{I}_a(\boldsymbol{\theta}^*) (\boldsymbol{\theta} - \boldsymbol{\theta}^*), \quad (17)$$

where  $D_{\text{KL}}(f_{\boldsymbol{\theta}^*, a} \| f_{\boldsymbol{\theta}, a})$  is the Kullback–Leibler divergence between the density functions  $f_{\boldsymbol{\theta}^*, a}$  and  $f_{\boldsymbol{\theta}, a}$ .

## 4.2 Main Theoretical Results

In this section, we present the main theoretical results, including the consistency, asymptotic normality and the optimality of the proposed method. Recall that the regularity conditions (Assumptions 1 – 5, along with Assumptions 6A – 7A or 6B – 7B) are assumed throughout the section.

### 4.2.1 Strong Consistency

We start with the strong consistency of the MLE following GI0 or GI1.

**Theorem 4.1** (Strong consistency). *Let  $\hat{\boldsymbol{\theta}}_n^{ML}$  be the MLE following the experiment selection rule GI0 or GI1, as described in Algorithm 1 and Algorithm 2. Then,*

$$\lim_{n \rightarrow \infty} \hat{\boldsymbol{\theta}}_n^{ML} = \boldsymbol{\theta}^* \text{ a.s. } \mathbb{P}_*,$$

where  $\mathbb{P}_*$  denotes the data-generating probability distribution under the true model parameter  $\boldsymbol{\theta}^*$ .

Theorem 4.1 suggests that the MLE will be close to the true model parameter with a large sample size following GI0 or GI1.

### 4.2.2 Limiting Selection Frequency and Asymptotic Normality of MLE

Let

$$\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}) = \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}),$$

be the weighted Fisher information associated with a proportion vector  $\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}$ . The distribution of MLE depends on the empirical frequency vector  $\bar{\boldsymbol{\pi}}_n$ , which is defined by (8).

We first present an auxiliary asymptotic normality result for the MLE following a general active experiment selection rule that is not necessarily GI0 or GI1.

**Theorem 4.2** (Asymptotic normality following general experiment selection rules). *Let  $\hat{\boldsymbol{\theta}}_n^{ML}$  be the MLE calculated according to (5) following an active experiment selection rule that is not necessarily GI0 or GI1. Let  $\bar{\boldsymbol{\pi}}_n$  be the corresponding empirical frequency vector.*

*Assume that there exists  $\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}$  such that  $\bar{\boldsymbol{\pi}}_n$  converges to  $\boldsymbol{\pi}$  in probability  $\mathbb{P}_*$  as  $n \rightarrow \infty$ , and  $\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*)$  is nonsingular. Then,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n^{ML} - \boldsymbol{\theta}^*) \xrightarrow{d} N_p(\mathbf{0}_p, \{\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*)\}^{-1}) \text{ as } n \rightarrow \infty, \quad (18)$$

where ' $\xrightarrow{d}$ ' denotes the convergence in distribution.

The above Theorem 4.2 extends the classic asymptotic normality results for MLE to the sequential setting with active experiment selection. It roughly states that if the frequency of the selected experiment approximates a limiting proportion as the sample size grows, and the Fisher information weighted by the limiting proportion is nonsingular, then the MLE is asymptotically normal and the asymptotic covariance matrix is the inverted weighted Fisher information. Next, we will show that if we follow the experiment selection rule GI0 or GI1, then the frequency for the selected experiments is approaching a limiting proportion that is determined by the criterion function  $\mathbb{G}_{\boldsymbol{\theta}}$ . For this purpose, we first define a function  $\mathbb{F}_{\boldsymbol{\theta}} : \mathcal{S}^{\mathcal{A}} \rightarrow \mathbb{R}$ ,

$$\mathbb{F}_{\boldsymbol{\theta}}(\boldsymbol{\pi}) = \mathbb{G}_{\boldsymbol{\theta}} \left[ \left\{ \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}) \right\}^{-1} \right]. \quad (19)$$

**Theorem 4.3** (Limiting experiment selection frequency following GI0 or GI1). *Assume that  $\mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi})$  has a unique minimizer, denoted by  $\boldsymbol{\pi}^*$ . That is,  $\boldsymbol{\pi}^* = \arg \min_{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}} \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi})$ . Then, GI0 and GI1 both satisfy*

$$\lim_{n \rightarrow \infty} \bar{\boldsymbol{\pi}}_n = \boldsymbol{\pi}^* \text{ a.s. } \mathbb{P}_*, \quad (20)$$

where  $\bar{\boldsymbol{\pi}}_n$  is the corresponding empirical frequency vector. Moreover, for a general function  $\mathbb{F}_{\boldsymbol{\theta}^*}(\cdot)$  whose minimizer is not necessarily unique, we have

$$\lim_{n \rightarrow \infty} n^{\beta} \{ \mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_n) - \min_{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}} \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}) \} = 0 \text{ a.s. } \mathbb{P}_*. \quad (21)$$

for all  $0 \leq \beta < 1/2$ , given that GI0 or GI1 is used as the experiment selection rule.

The asymptotic normality of the MLE following GI0 or GI1 is proved by combining the above two theorems. We summarize this result in the next theorem.

**Theorem 4.4** (Asymptotic normality following GI0 or GI1). *Let  $\hat{\boldsymbol{\theta}}_n^{ML}$  be the MLE following the experiment selection rule GI0 or GI1, as described in Algorithm 1 and Algorithm 2. Assume  $\mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi})$  has a unique minimizer  $\boldsymbol{\pi}^*$ . Then,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n^{ML} - \boldsymbol{\theta}^*) \xrightarrow{d} N_p(\mathbf{0}_p, \{\mathcal{I}^{\boldsymbol{\pi}^*}(\boldsymbol{\theta}^*)\}^{-1}). \quad (22)$$

The covariance of the MLE can be approximated by the plug-in estimator  $n^{-1}\{\mathcal{I}^{\bar{\boldsymbol{\pi}}_n}(\hat{\boldsymbol{\theta}}_n^{ML})\}^{-1}$ . This is justified by the next theorem.

**Theorem 4.5** (Asymptotic covariance matrix of the MLE). *Under the settings of Theorem 4.4,*

$$\sqrt{n}\{\mathcal{I}^{\bar{\boldsymbol{\pi}}_n}(\hat{\boldsymbol{\theta}}_n^{ML})\}^{1/2}(\hat{\boldsymbol{\theta}}_n^{ML} - \boldsymbol{\theta}^*) \xrightarrow{d} N_p(\mathbf{0}_p, I_p). \quad (23)$$

*In addition, for any continuously differentiable function  $g : \boldsymbol{\Theta} \rightarrow \mathbb{R}$  such that  $\nabla g(\boldsymbol{\theta}^*) \neq \mathbf{0}_p$ ,*

$$\frac{\sqrt{n}(g(\hat{\boldsymbol{\theta}}_n^{ML}) - g(\boldsymbol{\theta}^*))}{\left\| \{\mathcal{I}^{\bar{\boldsymbol{\pi}}_n}(\hat{\boldsymbol{\theta}}_n^{ML})\}^{-1/2} \nabla g(\hat{\boldsymbol{\theta}}_n^{ML}) \right\|} \xrightarrow{d} N(0, 1). \quad (24)$$

The first part of the above theorem justifies the use of the plug-in estimator for the covariance matrix of the MLE. The second part of the theorem suggests that the approximate  $1 - \alpha$  confidence interval for  $g(\boldsymbol{\theta})$  can be constructed as  $g(\hat{\boldsymbol{\theta}}_n^{ML}) \pm z_{\alpha/2} \left\| \{\mathcal{I}^{\bar{\boldsymbol{\pi}}_n}(\hat{\boldsymbol{\theta}}_n^{ML})\}^{-1/2} \nabla g(\hat{\boldsymbol{\theta}}_n^{ML}) \right\|$  where  $z_{\alpha/2}$  is the  $1 - \alpha/2$  quantile of the standard normal distribution.

### 4.2.3 Asymptotic Optimality

In this section, we present results regarding the optimality of the proposed methods. We consider two notions of optimality, including the optimal design and asymptotic efficiency of the estimators under a decision theory framework. The former extends a similar concept in the literature on the design of experiments, and the latter builds upon the classic asymptotic efficiency results for MLE with i.i.d. observations. We start with the notion of optimality in terms of the optimal design.

**Definition 4.6** ( $\mathbb{G}_{\boldsymbol{\theta}^*}$ - optimality). A selection rule is said to be  $\mathbb{G}_{\boldsymbol{\theta}^*}$  a.s. optimal design if its corresponding selection frequency  $\{\bar{\boldsymbol{\pi}}_n\}_{n \in \mathbb{Z}_+}$  satisfies

$$\lim_{n \rightarrow \infty} \mathbb{G}_{\boldsymbol{\theta}^*}(\{\mathcal{I}^{\bar{\boldsymbol{\pi}}_n}(\boldsymbol{\theta}^*)\}^{-1}) = \min_{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}} \mathbb{G}_{\boldsymbol{\theta}^*}(\{\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*)\}^{-1}) \text{ a.s. } \mathbb{P}_*. \quad (25)$$

The above notion of  $\mathbb{G}_{\boldsymbol{\theta}^*}$ - optimal selection rules extends the classic concept of optimal designs adopted in the literature on the design of experiments (see, e.g., Kiefer (1974); Yang



et al. (2013)). It allows for general criteria functions and adaptive experiment selection rules. If an adaptive experiment selection rule is  $\mathbb{G}_{\theta^*}$ -optimal, it approximately minimizes the criterion function when the sample size is large. Theorem 4.3 implies the following result.

**Theorem 4.7** ( $\mathbb{G}_{\theta^*}$ -optimal selection). *Both GI0 and GI1 are  $\mathbb{G}_{\theta^*}$  a.s. optimal.*

The above theorem indicates that the proposed experiment selection rules have the best performance in some sense when compared with other experiment selection rules. Next, we consider the optimality property of the MLE when combined with GI0 or GI1 under the lens of a sequential decision theory framework for the design-and-estimation problem.

Consider a loss function  $L(\theta^*, \hat{\theta})$  for an estimator  $\hat{\theta}$  following an active experiment selection rule, and the corresponding risk  $\mathbb{E}_{\theta^*} L(\theta^*, \hat{\theta})$ . The next theorem first establishes a lower bound for the asymptotic risk for unbiased estimators and then shows that the MLE combined with the selection rule GI0 (or GI1) achieves this lower bound when the criterion function matches the loss function.

**Theorem 4.8** (Minimum risk for unbiased estimators). *Let  $L(\theta, \hat{\theta})$  be a loss function twice continuously differentiable in  $\hat{\theta}$  satisfying that  $L(\theta, \hat{\theta}) \geq 0$ ,  $L(\theta, \hat{\theta}) = 0$  if and only if  $\hat{\theta} = \theta$ , and  $\eta I_p \preceq \frac{1}{2} \nabla_{\hat{\theta}}^2 L(\theta^*, \hat{\theta}) \preceq \eta' I_p$  for some positive constants  $\eta$  and  $\eta'$ , and all  $\hat{\theta} \in \Theta$ . Let  $H_{\theta} = \frac{1}{2} \nabla_{\hat{\theta}}^2 L(\theta, \hat{\theta}) \Big|_{\hat{\theta}=\theta}$ . Then, the following results hold.*

1. *Assume regularity conditions (but without Assumption 5) hold. Consider an unbiased estimator  $\mathbf{T}_n$  of  $\theta$  following an arbitrary adaptive experiment selection rule. If the loss function does not satisfy  $L(\theta^*, \hat{\theta}) \equiv \langle H_{\theta^*}(\theta^* - \hat{\theta}), \theta^* - \hat{\theta} \rangle$ , we further assume for any  $\varepsilon > 0$ ,  $\limsup_{n \rightarrow \infty} \mathbb{E}_{\theta^*} n \|\mathbf{T}_n - \theta^*\|^2 I(\|\mathbf{T}_n - \theta^*\| > \varepsilon) = 0$ . Then,*

$$\liminf_{n \rightarrow \infty} \mathbb{E}_{\theta^*} \left[ n \cdot L(\theta^*, \mathbf{T}_n) \right] \geq \inf_{\pi \in \mathcal{S}^{\mathcal{A}}} \text{tr}(H_{\theta^*} \{\mathcal{I}^{\pi}(\theta^*)\}^{-1}). \quad (26)$$

*In particular, if the squared error loss  $L(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$  is used, then for any unbiased estimator  $\mathbf{T}_n$ ,  $\liminf_{n \rightarrow \infty} \left[ n \cdot \text{MSE}(\mathbf{T}_n) \right] \geq \inf_{\pi \in \mathcal{S}^{\mathcal{A}}} \text{tr}(\{\mathcal{I}^{\pi}(\theta^*)\}^{-1})$ .*

2. *Under Assumptions 1-4, 6A and 7A, and further assume that there exists  $\alpha > 0$ , such that for any  $\xi_a = \mathbf{Z}_a \theta$ ,  $\theta \in \Theta$  and  $x^a \in \text{supp}(f_{\theta,a})$ ,*

$$\lambda_{\min}(-\nabla_{\xi_a}^2 \log h_{\xi_a,a}(x^a)) \geq \alpha > 0. \quad (27)$$

*Assume there exists  $\delta > 0$  such that  $\mathbb{E}_{X^a \sim f_{\theta^*,a}} \|\nabla_{\theta} \log f_{\theta,a}(X^a)\|^{2+\delta} < \infty$ . Assume that  $\text{tr}(H_{\theta^*} \{\mathcal{I}^{\pi}(\theta^*)\}^{-1})$  has a unique minimizer, denoted by  $\pi^*$ . If we choose  $\mathbb{G}_{\theta}(\Sigma) =$*

$\text{tr}(H_{\boldsymbol{\theta}}\boldsymbol{\Sigma})$  and use the experiment selection rule *GI0* (or *GI1*) described in Algorithm 1 (or Algorithm 2), then the MLE achieves the lower bound in (26). That is,

$$\lim_{n \rightarrow \infty} \mathbb{E}_{\boldsymbol{\theta}^*} \{n \cdot L(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_n^{ML})\} = \min_{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}} \text{tr}(H_{\boldsymbol{\theta}^*} \{\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*)\}^{-1}). \quad (28)$$

In particular, if  $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2$ , the corresponding criterion function is  $\mathbb{G}_{\boldsymbol{\theta}}(\cdot) = \Phi_1(\cdot) = \text{tr}(\cdot)$ . MLE combined with *GI0* (or *GI1*) achieves the asymptotic lower bound for  $n \cdot \text{MSE}(\mathbf{T}_n)$  for unbiased estimator  $\mathbf{T}_n$ .

The first part of the above theorem provides a lower bound for the risk of any unbiased estimator combined with an arbitrary experiment selection rule. In particular, when  $p = |\mathcal{A}| = 1$ , it aligns with the classic Cramér - Rao lower bound for the variance of unbiased estimators with independent observations. The second part of the theorem suggests that the asymptotic risk of the MLE combined with the proposed *GI0* (or *GI1*) matches the lower bound, if the criterion function aligns with the loss function. When  $p = |\mathcal{A}| = 1$ , this matching risk gives an extension of the classic asymptotic efficiency result for MLE with i.i.d. data.

We note that Theorem 4.8 does not directly imply that the proposed method minimizes risk within a class of decision rules, since the MLE is not necessarily unbiased. This scenario is analogous to the classic asymptotic efficiency result for MLE with i.i.d. observations, where the MLE is shown to have the asymptotic variance matching the Cramér - Rao bound for unbiased estimators but the MLE itself is not unbiased. On the other hand, the asymptotic optimality of the MLE within a decision theory framework can be formalized using concepts such as local asymptotically normal (LAN) estimators and asymptotic concentration (see Chapter 8 of Van der Vaart (2000)) in classic asymptotic statistics. The next theorem suggests that MLE combined with the proposed experiment selection method is also asymptotically optimal in a similar sense. Here, we omit the definitions of notations and terminology such as “ $\rightsquigarrow$ ”, “ $\star$ ”, and “bowl-shaped functions”, and refer readers to Theorem 8.8 and 8.11 in Chapter 8 of Van der Vaart (2000), as the formal definitions of these notations are lengthy.

**Theorem 4.9** (Local asymptotic minimax risk). *Assume  $a_1, \dots, a_n, \dots$  are experiments selected following an active experiment selection rule such that  $a_{n+1}$  is measurable with respect to  $\mathcal{F}_n$  for all  $n$ . Assume that the sequence  $(\mathbf{T}_n(a_1, X_1, \dots, a_n, X_n), \bar{\boldsymbol{\pi}}_n)$  is regular at  $(\boldsymbol{\theta}, \boldsymbol{\pi}) \in \boldsymbol{\Theta} \times \mathcal{S}^{\mathcal{A}}$  for estimating parameter  $\boldsymbol{\theta}$ , which means that for every  $\mathbf{h} \in \mathbb{R}^p$ ,*

$$\sqrt{n} \left( \mathbf{T}_n - \left( \boldsymbol{\theta} + \frac{\mathbf{h}}{\sqrt{n}} \right) \right) \overset{\boldsymbol{\theta} + \frac{\mathbf{h}}{\sqrt{n}}}{\rightsquigarrow} L_{\boldsymbol{\theta}}^{\boldsymbol{\pi}} \text{ and } \bar{\boldsymbol{\pi}}_n \overset{\boldsymbol{\theta} + \frac{\mathbf{h}}{\sqrt{n}}}{\rightsquigarrow} \boldsymbol{\pi}, \quad (29)$$

for some distribution  $L_{\boldsymbol{\theta}}^{\pi}$ , and  $\mathcal{I}^{\pi}(\boldsymbol{\theta}) = \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta})$  is nonsingular. Then, the following statements hold.

1. (Convolution theorem) There exists a probability measure  $M_{\boldsymbol{\theta}}^{\pi}$  such that

$$L_{\boldsymbol{\theta}}^{\pi} = N_p(\mathbf{0}_p, \{\mathcal{I}^{\pi}(\boldsymbol{\theta})\}^{-1}) * M_{\boldsymbol{\theta}}^{\pi}. \quad (30)$$

In particular, if  $L_{\boldsymbol{\theta}}^{\pi}$  has the covariance matrix  $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{\pi}$ , then  $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{\pi} \succeq \{\mathcal{I}^{\pi}(\boldsymbol{\theta})\}^{-1}$ .

2. (Local asymptotic minimax theorem) For any bowl-shaped loss function  $\ell$ ,

$$\sup_{|F| < \infty, F \subset \mathbb{R}^p} \liminf_{n \rightarrow \infty} \sup_{h \in F} \mathbb{E}_{\boldsymbol{\theta} + \frac{h}{\sqrt{n}}} \ell \left( \sqrt{n} (\mathbf{T}_n - (\boldsymbol{\theta} + \frac{h}{\sqrt{n}})) \right) \geq \mathbb{E} \ell(V^{\pi}) \geq \min_{\pi} \mathbb{E} \ell(V^{\pi}), \quad (31)$$

where the first supremum is taken over all finite subsets  $F$  of  $\mathbb{R}^p$ , and  $V^{\pi} \sim N_p(\mathbf{0}_p, \{\mathcal{I}^{\pi}(\boldsymbol{\theta})\}^{-1})$ .

In the case where  $\ell(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|^2$  and  $\mathbb{G}_{\boldsymbol{\theta}}(\cdot) = \Phi_1(\cdot) = \text{tr}(\cdot)$ , the second part of Theorem 4.8 together with Theorem 4.9 imply that the MLE combined with both GI0 and GI1 selection achieves the local asymptotic minimax lower bound on the MSE of estimators.

### 4.3 Theoretical Results Regarding Early Stopping Rules

As discussed in Section 3.2, early stopping rules are adopted in many applications to reduce the expected sample size. In this section, we provide consistency and asymptotic normality results for the MLE obtained at a large random stopping time.

**Theorem 4.10** (Strong consistency at a random stopping time). *Let  $\widehat{\boldsymbol{\theta}}_n^{ML}$  be the MLE following the experiment selection rule GI0 or GI1, as described in Algorithm 1 and Algorithm 2, and let  $\tau_n \in \mathbb{N}$  be a sequence of stopping time with respect to the filtration  $\{\mathcal{F}_n\}_{n \in \mathbb{Z}_+}$  such that  $\lim_{n \rightarrow \infty} \tau_n = \infty$  a.s. and  $\tau_n < \infty$  a.s. for each  $n$ . Then,*

$$\lim_{n \rightarrow \infty} \widehat{\boldsymbol{\theta}}_{\tau_n}^{ML} = \boldsymbol{\theta}^* \text{ a.s. } \mathbb{P}_*.$$

The above theorem extends Theorem 4.1 to allow for random stopping times. It suggests that the MLE is close to the true model parameter at a large random sample size. Next, we present the result on asymptotic normality, which enables statistical inference at large stopping times.

**Theorem 4.11** (Asymptotic normality following GI0 or GI1 with an early stopping rule). *Let  $\widehat{\boldsymbol{\theta}}_n^{ML}$  be the MLE following the experiment selection rule GI0 or GI1, as described in*

*Algorithm 1 and Algorithm 2. Assume  $\mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi})$  has a unique minimizer  $\boldsymbol{\pi}^*$ . Let  $\{c_n\}_{n \geq 0}$  be a positive and decreasing sequence such that  $c_n \rightarrow 0$  as  $n \rightarrow \infty$ . Let  $h : \boldsymbol{\Theta} \rightarrow \mathbb{R}$  be a continuously differentiable function such that  $\nabla h(\boldsymbol{\theta}) \neq \mathbf{0}_p$  for all  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ . Consider stopping times  $\tau_{c_n}^{(1)}$  and  $\tau_{c_n}^{(2)}$  defined in (9) and (10), respectively. Then, for both stopping time  $\tau_n = \tau_{c_n}^{(1)}$  and  $\tau_n = \tau_{c_n}^{(2)}$ , we have*

$$\sqrt{\tau_n} \{ \mathcal{I}^{\pi_{\tau_n}}(\hat{\boldsymbol{\theta}}_{\tau_n}^{ML}) \}^{1/2} (\hat{\boldsymbol{\theta}}_{\tau_n}^{ML} - \boldsymbol{\theta}^*) \xrightarrow{d} N_p(\mathbf{0}_p, I_p). \quad (32)$$

*Furthermore, for any continuously differentiable function  $g : \boldsymbol{\Theta} \rightarrow \mathbb{R}$  such that  $\nabla g(\boldsymbol{\theta}^*) \neq \mathbf{0}_p$ ,*

$$\frac{\sqrt{\tau_n} (g(\hat{\boldsymbol{\theta}}_{\tau_n}^{ML}) - g(\boldsymbol{\theta}^*))}{\left\| \{ \mathcal{I}^{\pi_{\tau_n}}(\hat{\boldsymbol{\theta}}_{\tau_n}^{ML}) \}^{-1/2} \nabla g(\hat{\boldsymbol{\theta}}_{\tau_n}^{ML}) \right\|} \xrightarrow{d} N(0, 1). \quad (33)$$

## 5 Applications

In this section, we provide details on the methods and theoretical results to applications discussed in Section 1, including item selection in CAT and adaptive pairs selection in sequential rank aggregation problems. We also provide results regarding active estimation for generalized linear models (GLM), which encompass many useful models as its special cases.

### 5.1 Active Estimation for GLM

Consider the case where the distribution of the observations falls into an exponential family (see, e.g., McCullagh (2019)). Following the setting in Section 5 of Chaudhuri et al. (2015), we consider the density functions

$$f_{\boldsymbol{\theta},a}(x_a) = \zeta^a(x_a) \exp \{ x_a \mathbf{z}_a^T \boldsymbol{\theta} - B_a(\mathbf{z}_a^T \boldsymbol{\theta}) \}, \quad (34)$$

where  $x_a \in \mathbb{R}$ ,  $\mathbf{z}_a \in \mathbb{R}^p$ , and  $B_a(\cdot), a \in \mathcal{A}$ . Assume that the support of  $B_a$  is  $\mathbb{R}$ . Under this model,  $\boldsymbol{\theta}$  serves as the unknown linear coefficient in a GLM and we are interested in estimating it using the proposed Algorithm 1 and Algorithm 2. The Fisher information is given by

$$\mathcal{I}_a(\boldsymbol{\theta}) = B_a''(\mathbf{z}_a^T \boldsymbol{\theta}) \mathbf{z}_a \mathbf{z}_a^T \text{ and } \mathcal{I}(\boldsymbol{\theta}; \mathbf{a}_n) = \sum_{i=1}^n B_{a_i}''(\mathbf{z}_{a_i}^T \boldsymbol{\theta}) \mathbf{z}_{a_i} \mathbf{z}_{a_i}^T. \quad (35)$$

Based on the above equations, Algorithms 1 and 2 are simplified as follows.

---

**Algorithm 4** Simplified GI0/GI1 Algorithm for GLM

---

We modify the following lines in Algorithms 1 and 2, while keeping the other lines of the algorithms unchanged.

2: **Require:** choose  $a_1^0, \dots, a_{n_0}^0$  such that  $\dim(\text{span}\{\mathbf{z}_{a_n^0}; n = 1, 2, \dots, n_0\}) = p$ .

6: The Fisher information matrices used in line 6 of Algorithms 1 and 2 are calculated using the formula (35).

---

**Corollary 5.1.** *Assume the function  $B_a$  has the support  $\mathbb{R}$ ,  $\dim(\text{span}\{\mathbf{z}_a; a \in \mathcal{A}\}) = p$ , and Assumptions 1 and 5 hold. If the above Algorithm 4 for GI0 or GI1 is used, then all the theorems presented in Section 4.2 hold.*

Note that in the above corollary, the assumptions are greatly simplified compared to the regularity conditions described in Section 4.1, thanks to the nice form of GLMs. It only requires that the parameter space is compact, the true parameter is an interior point of the parameter space, and the parameter is identifiable when using all the experiments together. In practice, the parameter space  $\Theta$  may not be given in advance. In these cases, we may specify  $\Theta$  as a box (i.e.,  $\Theta = [-r, r]^p$ ) or ball (i.e.,  $\Theta = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\| \leq r\}$ ) for some large  $r$ . The theoretical results still apply, if the true parameter is an interior point of the parameter space.

## 5.2 Computerized Adaptive Testing (CAT)

CAT has gained prominence in recent decades as an innovative approach to educational assessment (Bartroff et al., 2008; Chang and Ying, 2009; Wainer et al., 2000). In CAT, test items are sequentially and adaptively chosen from an item pool based on the test-taker’s previous responses. This approach enhances test precision and shortens test length by selecting items tailored to the test-taker’s individual latent traits. Item Response Theory (IRT) and Multidimensional Item Response Theory (MIRT) models are commonly used to model a test-taker’s responses (See, e.g., Chen et al. (2024), Embretson and Reise (2013), and Reckase (2006) for reviews on IRT and MIRT models). In a binary MIRT model, a response to an item is coded as 0 or 1, where 1 indicates that the item was answered correctly and 0 indicates the it was answered incorrectly.

Let  $k$  be the total number of items in the item pool for an educational test, and let  $\mathcal{A} = \{1, \dots, k\}$  represent the indices of these items. Under a MIRT model, each item  $j \in \mathcal{A}$  is associated with a multidimensional item parameter  $(\mathbf{z}_j, b_j)$ , which quantifies item properties such as the item’s difficulty and the skills it measures. The test taker is associated with a latent trait parameter  $\boldsymbol{\theta} \in \mathbb{R}^p$ , typically interpreted as proficiency in  $p$  different skills.

Given the selected items and the test-taker's latent trait parameter, responses are assumed to be conditionally independent. The correct response probability  $P(\boldsymbol{\theta}; \mathbf{z}_j, b_j)$ , also known as the item response function (IRF) of item  $j$ , is a function of  $\boldsymbol{\theta}$  and depends on  $(\mathbf{z}_j, b_j)$ . For example, the commonly adopted multidimensional two-parameter logistic model (M2PL) assumes that the IRF takes the form

$$P(\boldsymbol{\theta}; \mathbf{z}_j, b_j) = \{1 + \exp(-\mathbf{z}_j^T \boldsymbol{\theta} - b_j)\}^{-1}, \quad (36)$$

where  $\mathbf{z}_j$  is the discrimination parameter, indicating the strength of each latent trait's influence on the response, and  $-b_j$  is the difficulty parameter of item  $j$ .

Item selection is critical for efficient CAT design. The objective is to accurately estimate the latent trait parameter  $\boldsymbol{\theta} \in \mathbb{R}^p$  by selecting the next item  $a_{n+1} \in \mathcal{A}$  based on previously selected items and responses  $a_1, X_1, \dots, a_n, X_n$ . Note that item parameters are typically pre-calibrated based on historical data and are assumed to be known in CAT. In the rest of the section, we provide details on applying the item selection rules GI0 and GI1 under the M2PL model. First, the density function is  $f_{\boldsymbol{\theta}, a}(x) = P(\boldsymbol{\theta}; \mathbf{z}_a, b_a)^x (1 - P(\boldsymbol{\theta}; \mathbf{z}_a, b_a))^{1-x}$ , and the corresponding Fisher information is

$$\mathcal{I}_a(\boldsymbol{\theta}) = P(\boldsymbol{\theta}; \mathbf{z}_a, b_a)(1 - P(\boldsymbol{\theta}; \mathbf{z}_a, b_a))\mathbf{z}_a\mathbf{z}_a^T \text{ and } \mathcal{I}(\boldsymbol{\theta}; \mathbf{a}_n) = \sum_{a \in \mathcal{A}} \bar{\pi}_n(a)\mathcal{I}_a(\boldsymbol{\theta}). \quad (37)$$

If the criterion function  $\mathbb{G}_{\boldsymbol{\theta}}(\cdot) = \Phi_q(\cdot)$ , then GI1 can be simplified as:

$$a_{n+1} = \arg \max_{a \in \mathcal{A}} P(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{z}_a, b_a)(1 - P(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{z}_a, b_a)) \cdot \mathbf{z}_a^T \left( \mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n) \right)^{-q-1} \mathbf{z}_a. \quad (38)$$

---

**Algorithm 5** Simplified GI0/GI1 Algorithm for M2PL model

---

We modify the following lines in Algorithms 1 and 2, while keeping the other lines of the algorithms unchanged.

2: **Require:**  $\dim(\text{span}\{\mathbf{z}_a; a \in \mathcal{A}\}) = p$  and choose  $a_1^0, \dots, a_{n_0}^0$  such that  $\dim(\text{span}\{\mathbf{z}_{a_n^0}; n = 1, 2, \dots, n_0\}) = p$ .

6: The Fisher information matrices used in line 6 of Algorithms 1 for GI0 are calculated using the formula (37). Selection in line 6 of Algorithms 2 for GI1 is replaced by (38).

---

**Corollary 5.2.** *Assume Assumption 1 holds,  $\dim(\text{span}\{\mathbf{z}_a; a \in \mathcal{A}\}) = p$ , and criterion function  $\mathbb{G}_{\boldsymbol{\theta}}(\cdot) = \Phi_q(\cdot)$ . If we consider simplified Algorithm 5 for GI0 and GI1 with M2PL model, then the conclusions for GI0 and GI1 from all the theorems presented in Section 4.2 hold.*

### 5.3 Sequential Rank Aggregation from Noisy Pairwise Comparison

Consider the problem of determining the global rank over  $p+1$  objects. Let  $\mathcal{A} \subset \{(j, l); j, l \in \{0, 1, 2, \dots, p\}\}$  be a subset of all possible pairs for comparison. At each time  $n$ , a pair  $a_n = (a_{n,1}, a_{n,2}) \in \mathcal{A}$  is chosen for comparison, yielding a random pairwise comparison outcome  $X_n \in \{0, 1\}$ . Here,  $X_n = 1$  indicates that the object  $a_{n,1}$  is preferred over  $a_{n,2}$  in the comparison, and  $X_n = 0$  indicates the opposite. To infer the global rank of objects, ranking models (e.g., Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952; Duncan, 1959) and the Thurstone model (Thurstone, 1927)) are usually assumed for the noisy pairwise comparison results. These models assume that each object  $i$  is associated with a latent score parameter  $\theta_i$ , the pairwise comparison result between object  $i$  and object  $j$  is depending on  $\theta_i$  and  $\theta_j$ , and the true global rank is the rank of the latent score parameters. For example, the BTL model assumes

$$f_{\boldsymbol{\theta},a}(x) = \left( \frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}} \right)^x \left( \frac{e^{\theta_j}}{e^{\theta_i} + e^{\theta_j}} \right)^{1-x} \quad (39)$$

for the pair  $a = (i, j)$ . For sequential rank aggregation, the goal is to design an active pair selection rule that determines the next pair  $a_{n+1}$  for comparison based on the prior pair comparison results  $(a_1, X_1, \dots, a_n, X_n)$ , so that the global rank can be inferred accurately. This problem boils down to the active sequential estimation of the latent score parameters.

In the rest of the section, we elaborate on the implementation and theoretical results for GI0 and GI1 for the sequential rank aggregation problem under a BTL model. Note that the distribution of the comparison results only depends on the differences  $\theta_i - \theta_j$  for  $0 \leq i, j \leq p$ . Thus, we fix  $\theta_0 = 0$  to ensure the identifiability of  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$ .

When  $a = (i, j)$ , we set  $\mathbf{z}_a = \mathbf{e}_j - \mathbf{e}_i$ , where  $\{\mathbf{e}_1, \dots, \mathbf{e}_p\}$  is the standard basis of  $\mathbb{R}^p$  and  $\mathbf{e}_0 = \mathbf{0}_p$ .

For  $a = (i, j)$ , the Fisher information and the weighted Fisher information are given by

$$\mathcal{I}_a(\boldsymbol{\theta}) = \frac{e^{\theta_i - \theta_j}}{(1 + e^{\theta_i - \theta_j})^2} \mathbf{z}_a \mathbf{z}_a^T, \text{ and } \mathcal{I}(\boldsymbol{\theta}; \mathbf{a}_n) = \sum_{a=(i,j) \in \mathcal{A}} \bar{\pi}_n(a) \frac{e^{\theta_i - \theta_j}}{(1 + e^{\theta_i - \theta_j})^2} \mathbf{z}_a \mathbf{z}_a^T. \quad (40)$$

If we take the criterion function  $\mathbb{G}_{\boldsymbol{\theta}}(\cdot) = \Phi_q(\cdot)$ , GI1 can be simplified as

$$a_{n+1} = \operatorname{argmax}_{a \in \mathcal{A}} \frac{e^{\mathbf{z}_a^T \hat{\boldsymbol{\theta}}_n^{\text{ML}}}}{(1 + e^{\mathbf{z}_a^T \hat{\boldsymbol{\theta}}_n^{\text{ML}}})^2} \mathbf{z}_a^T \left( \mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n) \right)^{-q-1} \mathbf{z}_a. \quad (41)$$

We treat  $V = \{0, 1, \dots, p\}$  as vertices and  $\mathcal{A}$  as the set of edges. Then,  $G = (V, \mathcal{A})$  is an

undirected graph. Assume that  $G$  is a connected graph. This condition ensures that  $\theta$  is identifiable when all the pairs in  $\mathcal{A}$  are compared. Under this condition, it is possible to select  $\mathcal{A}_0 = \{a_1^0, a_2^0, \dots, a_{n_0}^0\} \subset \mathcal{A}$  so that  $(V, \mathcal{A}_0)$  is a connected subgraph of  $G$ .

---

**Algorithm 6** Simplified GI0/GI1 Algorithm for BTL model

---

We modify the following lines in Algorithms 1 and 2, while keeping the other lines of the algorithms unchanged.

2: **Require:** The subgraph  $(V, \{a_1^0, \dots, a_{n_0}^0\})$  is a connected graph.

6: The Fisher information matrices used in line 6 of Algorithms 1 for GI0 are calculated using the formula (40). Selection in line 6 of Algorithms 2 for GI1 is replaced by (41).

---

**Corollary 5.3.** *Assume that Assumption 1 holds,  $G$  is a connected graph, and the criterion function  $\mathbb{G}_\theta(\cdot) = \Phi_q(\cdot)$ . If the above Algorithm 6 for GI0 or GI1 is used, then the conclusions for GI0 and GI1 from all the theorems presented in Section 4.2 hold.*

## 6 Technical Challenges, New Analytical Tools and a Proof Sketch for Theorem 4.3

In this section, we highlight the key technical challenges in proving Theorem 4.3 and introduce new analytical tools to address these challenges. The primary challenge lies in demonstrating that GI0/GI1 effectively balances the trade-off between exploration and exploitation, a well-known concept in the literature on sequential decision making involving unknown parameters. Exploration means sufficient sampling of all relevant experiments to ensure consistent parameter estimation. Exploitation means optimally sampling experiments once the parameter has been accurately estimated. Below, we discuss these two facets—exploration and exploitation—in the context of active sequential estimation.

### 6.1 Exploration

In order to have a consistent estimator, the selection rule needs to sample relevant experiments sufficiently often. This is formalized by the following condition,

$$n_I := \max_{S \subset \mathcal{A}: S \text{ is relevant}} \min_{a \in S} n_a \rightarrow \infty \text{ as } n \rightarrow \infty, \quad (42)$$

where  $n_a = |\{i; a_i = a, 1 \leq i \leq n\}|$  for  $a \in \mathcal{A}$ . Here, we say that a set of experiments  $S$  is relevant if  $\sum_{a \in S} \mathcal{I}_a(\theta)$  is nonsingular for any  $\theta \in \Theta$ . If  $S$  is relevant, then the model



parameter is identifiable when all the experiments in  $S$  are sampled. Equation (42) says that at least one of the relevant sets of experiments needs to be sampled infinitely often, in order to have a consistent estimator.

**Challenge 6.1.** *Show that  $n_I \rightarrow \infty$  as  $n \rightarrow \infty$  following GI0/GI1.*

We note that in related sequential design problems, exploration is usually achieved by incorporating an extra exploration step in the experiment selection rule. For example, in active sequential hypothesis testing problems (see, e.g., Chernoff (1959); Naghshvar and Javidi (2013)), a two-stage algorithm is often utilized, where the first stage is designed for exploration and the second stage is designed for exploitation. Another prevalent method for ensuring sufficient exploration is the use of the epsilon-greedy algorithm in reinforcement learning and multi-armed bandit (MAB) problems, where all available experiments are sampled with a minimum probability of  $\varepsilon$ . For methods that incorporate an explicit exploration component, verifying (42) is usually straightforward. However, for algorithms like GI0/GI1, which are greedy and lack an additional exploration component, proving (or disproving) Equation (42) is much more challenging.

Nevertheless, we tackle Challenge 6.1 and establish the following proposition concerning sufficient exploration for GI0 and GI1.

**Proposition 6.2.** *Under regularity conditions described in Section 4.1, both GI0 and GI1 satisfy that  $\liminf_{n \rightarrow \infty} \frac{n_I}{n} > 0$ .*

Below, we discuss the heuristic ideas for justifying the above proposition, while clarifying the rigorous proof is much more involved. Let  $\mathcal{A}_{\max} = \arg \max_{a \in \mathcal{A}} n_a$  be the set of experiments that are most frequently selected. A key observation is that the *inverted Fisher information, through its directional derivatives in experiment selection rules, acts as a regularizer*, which means that if  $n_{\max}/n_I$  is large enough, then we can show that

$$\partial_{\mathbf{l}_{a^m}} \mathbb{F}_{\hat{\boldsymbol{\theta}}_n}(\bar{\boldsymbol{\pi}}_n) > \partial_{\mathbf{l}_{a'}} \mathbb{F}_{\hat{\boldsymbol{\theta}}_n}(\bar{\boldsymbol{\pi}}_n) \quad (43)$$

for all  $a^m \in \mathcal{A}_{\max}$  and some  $a' \notin \mathcal{A}_{\max}$ , where

$$\partial_{\mathbf{l}_a} \mathbb{F}_{\hat{\boldsymbol{\theta}}_n} = \left\langle \bar{\boldsymbol{\pi}}_{n+1}^a - \bar{\boldsymbol{\pi}}_n, \frac{\partial}{\partial \boldsymbol{\pi}} \mathbb{G}_{\hat{\boldsymbol{\theta}}_n} \left[ \left\{ \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\hat{\boldsymbol{\theta}}_n) \right\}^{-1} \right] \Big|_{\boldsymbol{\pi} = \bar{\boldsymbol{\pi}}_n} \right\rangle$$

denotes the directional derivative along the direction  $\mathbf{l}_a = \bar{\boldsymbol{\pi}}_{n+1}^a - \bar{\boldsymbol{\pi}}_n$ . This implies that no action from  $\mathcal{A}_{\max}$  will be selected and the ratio  $n_I/n$  is bounded from below, if we follow the experiment selection rule  $a_{n+1} \in \arg \min_{a \in \mathcal{A}} \partial_{\mathbf{l}_a} \mathbb{F}_{\hat{\boldsymbol{\theta}}_n}(\bar{\boldsymbol{\pi}}_n)$ . According to Equation (7) and

additional asymptotic analysis, this experiment selection rule based on directional derivatives is asymptotically equivalent to GI0 and GI1, and, consequently, Proposition 6.2 holds.

Note that (43) itself is challenging to prove, for which we first prove that the following decomposition of the information holds:  $\hat{\Sigma}_n = \mathcal{I}(\hat{\theta}_n; \mathbf{a}_n) = \mathbf{A} + \mathbf{E}$ , and  $\mathbf{A}, \mathbf{E}$  satisfy:

1.  $\mathbf{A} = \sum_{a \in \mathcal{A}, \frac{n_a}{n} \geq U} \frac{n_a}{n} \mathcal{I}_a(\hat{\theta}_n)$  and  $\mathbf{E} = \sum_{a \in \mathcal{A}, \frac{n_a}{n} < U} \frac{n_a}{n} \mathcal{I}_a(\hat{\theta}_n)$ , for some  $U > 0$ .
2.  $\mathbf{A}$  is a singular and positive semidefinite matrix.
3.  $\mathbf{E}$  is a positive semidefinite matrix and the maximum eigenvalue of  $\mathbf{E}$  is much smaller than the smallest non-zero eigenvalue of  $\mathbf{A}$ .
4. There exists  $a' \in \mathcal{A}$  such that  $n_{a'} \leq n_I$  and  $\mathcal{I}_{a'}(\hat{\theta}_n) \notin \mathcal{R}(\mathbf{A})$ , where  $\mathcal{R}(\mathbf{A})$  denotes the column space of  $\mathbf{A}$ . This implies

$$\liminf_{\|\mathbf{E}\| \rightarrow 0} \text{tr} \left[ \nabla \mathbb{G}_{\hat{\theta}_n}(\hat{\Sigma}_n)(\mathbf{A} + \mathbf{E})^{-1} \mathcal{I}_{a'}(\hat{\theta}_n)(\mathbf{A} + \mathbf{E})^{-1} \right] = \infty.$$

5. For all  $a^m \in \mathcal{A}_{\max}$ ,  $\mathcal{I}_{a^m}(\hat{\theta}_n) \in \mathcal{R}(\mathbf{A})$ . This implies

$$\limsup_{\|\mathbf{E}\| \rightarrow 0} \text{tr} \left[ \nabla \mathbb{G}_{\hat{\theta}_n}(\hat{\Sigma}_n)(\mathbf{A} + \mathbf{E})^{-1} \mathcal{I}_{a^m}(\hat{\theta}_n)(\mathbf{A} + \mathbf{E})^{-1} \right] < \infty.$$

We treat  $\mathbf{A}$  as the dominating term and  $\mathbf{E}$  as a small perturbation when using the above matrix decomposition. With additional matrix perturbation analysis of  $\hat{\Sigma}_n^{-1} = (\mathbf{A} + \mathbf{E})^{-1}$  around its non-continuous point  $\mathbf{A}$ , a careful use of the Davis-Kahan sin  $\Theta$  theorem (Yu et al., 2015), and additional iterative analysis, we can show that  $\text{tr} [\nabla \mathbb{G}_{\hat{\theta}_n}(\hat{\Sigma}_n) \hat{\Sigma}_n^{-1} \mathcal{I}_{a'}(\hat{\theta}_n) \hat{\Sigma}_n^{-1}] > \text{tr} [\nabla \mathbb{G}_{\hat{\theta}_n}(\hat{\Sigma}_n) \hat{\Sigma}_n^{-1} \mathcal{I}_{a^m}(\hat{\theta}_n) \hat{\Sigma}_n^{-1}]$  for all  $a^m \in \mathcal{A}_{\max}$ . This, along with Equation (7) implies (43).

## 6.2 Exploitation

In active sequential estimation, optimal exploitation requires frequency of the selected experiments to approximate the optimal proportion  $\boldsymbol{\pi}^* = \arg \min_{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}} \mathbb{G}_{\boldsymbol{\theta}^*}(\{\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*)\}^{-1})$ , when the estimator is accurate enough. In related sequential decision problems, such as active sequential hypothesis testing, optimal exploitation is commonly attained through a ‘plug-in’ method. This method assumes the estimator is accurate and replaces  $\boldsymbol{\theta}^*$  with the estimator for calculating the proportion for the subsequent sampling. The ‘plug-in’ method’s theoretical analysis usually combines the consistency result with the optimization problem’s continuity. However, this approach does not work for GI0/GI1 algorithms, which optimize

one-step-ahead information gain over the *discrete set*  $\mathcal{A}$ , rather than the *probability simplex*  $\mathcal{S}^{\mathcal{A}}$ . Consequently, it is challenging to determine whether GI0/GI1 approximately solve the long-term optimization problem  $\arg \min \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi})$  over the probability simplex. This issue is divided into two specific challenges:

**Challenge 6.3** (Noiseless case). *For GI0 selection (2) and GI1 selection (3) with  $\widehat{\boldsymbol{\theta}}_1 = \widehat{\boldsymbol{\theta}}_2 = \dots = \boldsymbol{\theta}^*$ , do we have the convergence  $\lim_{n \rightarrow \infty} \mathbb{F}_{\boldsymbol{\theta}^*}(\overline{\boldsymbol{\pi}}_n) = \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*)$ ?*

**Challenge 6.4** (Noisy case). *How does the difference between  $\widehat{\boldsymbol{\theta}}_n$  and  $\boldsymbol{\theta}^*$  affect the convergence of the algorithms?*

Challenge 6.3 is roughly addressed using the following arguments. First, we can show that  $\mathbb{F}_{\boldsymbol{\theta}^*}(\cdot)$  is convex. By Jensen's inequality, we obtain that

$$\mathbb{F}_{\boldsymbol{\theta}^*}\left(\frac{n-1}{n}\overline{\boldsymbol{\pi}}_{n-1} + \frac{1}{n}\boldsymbol{\pi}^*\right) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*) \leq \left(1 - \frac{1}{n}\right)\{\mathbb{F}_{\boldsymbol{\theta}^*}(\overline{\boldsymbol{\pi}}_{n-1}) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*)\}.$$

Notice that by Taylor expansion, for any  $\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}$ ,

$$\mathbb{F}_{\boldsymbol{\theta}^*}\left(\frac{n-1}{n}\overline{\boldsymbol{\pi}}_{n-1} + \frac{1}{n}\boldsymbol{\pi}\right) - \mathbb{F}_{\boldsymbol{\theta}^*}(\overline{\boldsymbol{\pi}}_{n-1}) = \left\langle \nabla \mathbb{F}_{\boldsymbol{\theta}^*}(\overline{\boldsymbol{\pi}}_{n-1}), \frac{1}{n}\boldsymbol{\pi} - \frac{1}{n}\overline{\boldsymbol{\pi}}_{n-1} \right\rangle + O(1/n^2).$$

The first term on the right-hand side of the above equation is linear in  $\boldsymbol{\pi}$  over the simplex  $\mathcal{S}^{\mathcal{A}}$ . Thus, its minimum is achieved at a point mass at  $a'$  for some  $a' \in \mathcal{A}$ , i.e.,  $\boldsymbol{\pi} = \boldsymbol{\delta}_{a'} := (I(a = a'))_{a \in \mathcal{A}}$ . It can be shown that the solution to the optimization  $\arg \min_{a' \in \mathcal{A}} \left\langle \nabla \mathbb{F}_{\boldsymbol{\theta}^*}(\overline{\boldsymbol{\pi}}_{n-1}), \frac{1}{n}\boldsymbol{\delta}_{a'} - \frac{1}{n}\overline{\boldsymbol{\pi}}_{n-1} \right\rangle$  coincides with the selection rule GI1 if we replace the MLE with  $\boldsymbol{\theta}^*$  (see Equation (3)). Let  $a_n^0$  and  $a_n^1$  be the experiments selected by GI0 and GI1 (with the MLE replaced by the true parameter), respectively. Combining the above analysis with the definition of GI0, we obtain

$$\begin{aligned} \mathbb{F}_{\boldsymbol{\theta}^*}(\overline{\boldsymbol{\pi}}_n^{a_n^0}) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*) &\leq \mathbb{F}_{\boldsymbol{\theta}^*}(\overline{\boldsymbol{\pi}}_n^{a_n^1}) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*) = \min_{a' \in \mathcal{A}} \left\langle \nabla \mathbb{F}_{\boldsymbol{\theta}^*}(\overline{\boldsymbol{\pi}}_{n-1}), \frac{1}{n}\boldsymbol{\delta}_{a'} - \frac{1}{n}\overline{\boldsymbol{\pi}}_{n-1} \right\rangle + O(1/n^2) \\ &\leq \mathbb{F}_{\boldsymbol{\theta}^*}\left(\frac{n-1}{n}\overline{\boldsymbol{\pi}}_{n-1} + \frac{1}{n}\boldsymbol{\pi}^*\right) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*) + O(1/n^2) \leq \left(1 - \frac{1}{n}\right)(\mathbb{F}_{\boldsymbol{\theta}^*}(\overline{\boldsymbol{\pi}}_{n-1}) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*)) + O(1/n^2). \end{aligned}$$

The above display suggests that the distance  $\mathbb{F}_{\boldsymbol{\theta}^*}(\overline{\boldsymbol{\pi}}_{n-1}) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*)$  is reduced at the factor  $1 - 1/n$  for GI0 and GI1 under the noiseless case. With additional iterative analysis, we can further show that  $\mathbb{F}_{\boldsymbol{\theta}^*}(\overline{\boldsymbol{\pi}}_n) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*) \leq O(\log n/n)$ . Consequently, the frequency of the selected experiment converges to the optimal proportion.

On the other hand, the above heuristic analysis does not justify the convergence of the algorithm in the noisy case (Challenge 6.4), nor does it provide the convergence rate. We address these challenges by establishing and using a modified Robbins-Siegmund theorem,

which extends the classic result by Robbins and Siegmund (1971), to the stochastic process  $Z_n = \mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_n) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*)$ .

## 7 Simulation

In this section, we present two simulation studies. The first assesses the finite sample performance of the proposed methods under the setting of Example 1. The second is concerned with situations where  $p$  or  $|\mathcal{A}|$  is large. Throughout the section, we choose the criterion function  $\mathbb{G}_{\boldsymbol{\theta}}(\boldsymbol{\Sigma}) = \Phi_1(\boldsymbol{\Sigma}) = \text{tr}(\boldsymbol{\Sigma})$  for GI0 and GI1. Due to the page limit, we leave some detailed specifications and additional simulation results in the supplementary material.

### 7.1 Simulation Study 1

We first evaluate the performance of the proposed methods under the settings of Example 1. Specifically, let  $p = 2$ ,  $\mathcal{A} = \{1, 2, 3\}$ , and  $f_{\boldsymbol{\theta},1}(1) = 1/(1 + e^{-(0.1+\theta_1)})$ ,  $f_{\boldsymbol{\theta},2}(1) = 1/(1 + e^{-\theta_2})$  and  $f_{\boldsymbol{\theta},3}(1) = 1/(1 + e^{-(\theta_1/2+\theta_2)})$ . Also, let  $\boldsymbol{\Theta} = [-3, 3]^2$  in this section.

We start with illustrating the optimal proportion  $\boldsymbol{\pi}^*$ . According to Theorem 4.3, the optimal proportion for experiment selection is

$$\boldsymbol{\pi}^* = (\pi^*(1), \pi^*(2), \pi^*(3)) = \arg \min_{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}} \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}) = \arg \min_{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}} \text{tr} \{ \mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*)^{-1} \}. \quad (44)$$

Note that  $\boldsymbol{\pi}^*$  is dependent on the true model parameter  $\boldsymbol{\theta}^*$ . Figure 1 illustrates the dependency between  $\boldsymbol{\pi}^*$  and  $\theta_2$  while fixing  $\theta_1^* = 1$ . From the figure, we see that the optimal proportion varies as  $\theta_2$  changes. Additionally, for relatively small  $\theta_2$ , all three experiments have non-zero optimal proportions. However, for large  $\theta_2$ ,  $\pi^*(3)$  stays at zero, meaning that experiment  $a = 3$  is unnecessary in this case.

Next, we investigate the empirical proportion of selected experiments following the proposed methods. Recall that the empirical proportion is defined as  $\bar{\pi}_n(a) = \frac{1}{n} |\{i; a_i = a, 1 \leq i \leq n\}|$ , for  $a \in \{1, 2, 3\}$ . We generate data from the model with the true parameter  $\boldsymbol{\theta}^* = (1, 0)^T$  and plot the sample path of  $\bar{\pi}_n(a)$  against different sample size  $n$  following GI0 and GI1 in Figure 3. We clarify that the values of the empirical proportion in the figure are obtained without averaging. That is, they are based on a Monte Carlo simulation with only one replication. From Figure 3, we can see that the empirical proportions are approximating their respective optimal values as  $n$  increases, for both GI0 and GI1. This is consistent with Theorem 4.3, which states that the empirical proportion almost surely converges to the optimal proportion. We also observe that the selections made by GI0 and GI1

are almost identical. This may be due to the fact that they are asymptotically equivalent (see Equation (7)), and they are initialized with the same random seed.

Now, we evaluate the estimation accuracy of MLE following the proposed GI0 and GI1, and compare it with other experiment selection methods. The estimation accuracy is quantified using the estimated MSE, defined as  $\widehat{\text{MSE}}_n = \frac{1}{N} \sum_{j=1}^N \|\hat{\boldsymbol{\theta}}_{n,j}^{\text{ML}} - \boldsymbol{\theta}^*\|^2$ , where  $N = 20000$  is the number of Monte Carlo replications and  $\hat{\boldsymbol{\theta}}_{n,j}^{\text{ML}}$  is the MLE from the  $j$ -th Monte Carlo experiment with the sample size  $n$ . We compare GI0 and GI1 with the following experiment selection rules:

1. Uniform selection (Unif):  $a_{n+1}$  is uniformly sampled from  $\mathcal{A}$ .
2. Random optimal proportion selection (Opt\_random):  $a_{n+1}$  is sampled randomly from  $\mathcal{A}$  according to the optimal proportion  $\boldsymbol{\pi}^*$ . That is,  $\mathbb{P}(a_{n+1} = a | \mathcal{F}_n) = \pi^*(a)$  for  $a \in \mathcal{A}$ .
3. Deterministic optimal proportion selection (Opt\_deterministic):  

$$a_{n+1} = \arg \min_{a \in \mathcal{A}} \{\bar{\pi}_n(a) - \pi^*(a)\}.$$

We clarify that both Opt\_random and Opt\_deterministic methods require knowledge of the unknown parameter  $\boldsymbol{\theta}^*$ , so they are not implementable in practice. These methods serve as ‘oracle’ benchmarks allowing comparison with the proposed methods. Figure 2 depicts the estimated MSE as a function of the sample size  $n$  for different experiment selection rules. According to the figure, GI0, GI1, and Opt\_deterministic perform very similarly and outperform both Unif and Opt\_random. These findings are consistent with Theorem 4.8.

Finally, we check the finite sample validity of the normal approximation of the MLE. According to Theorem 4.5 and Theorem 4.11, for large  $n$  and small  $c$ ,

$$\mathbf{d}^T \hat{\boldsymbol{\theta}}_n^{\text{ML}} \pm \frac{1}{\sqrt{n}} Z_{\alpha/2} \left\| \left\{ \mathcal{I}^{\bar{\pi}_n}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}) \right\}^{-1/2} \mathbf{d} \right\| \text{ and } \mathbf{d}^T \hat{\boldsymbol{\theta}}_{\tau_c}^{\text{ML}} \pm Z_{\alpha/2} \cdot c \quad (45)$$

give approximate  $1 - \alpha$  confidence intervals (CIs) for  $\mathbf{d}^T \boldsymbol{\theta}$  where  $\mathbf{d} \in \mathbb{R}^2$  is nonzero and the stopping time  $\tau_c$  is defined in (9) with  $h(\boldsymbol{\theta}) = \mathbf{d}^T \boldsymbol{\theta}$ . Table 1 shows the coverage probability of the above CIs at different sample sizes, following GI0 or GI1, where we set  $\alpha = 0.05$ ,  $\mathbf{d} = (-0.5454216, -0.8381619)^T$  and  $\boldsymbol{\theta}^* = (1, 0)^T$ , based on a Monte Carlo simulation. From the table, we see that the coverage probability is close to the confidence level  $1 - \alpha$  for reasonably large  $n$  and the random stopping time  $\tau_c$ .

We have also performed additional simulation studies and produced histograms of the estimators. These additional simulation results are given in the supplementary material.

$n$	25	50	100	$\tau_{0.1}$
GI0	0.981	0.954	0.951	0.955
GI1	0.977	0.958	0.959	0.938

Table 1: Coverage probability for CIs based on (45), where the number of Monte Carlo replications is 1000. The Monte Carlo standard error for the values presented in the table is upper bounded by 0.007626.

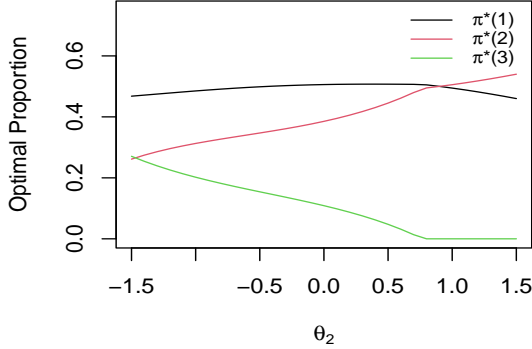


Figure 1: Optimal proportion  $\pi^*$  as a function of  $\theta_2$ , where the true parameter satisfies  $\theta^* = (1, \theta_2)^T$ .

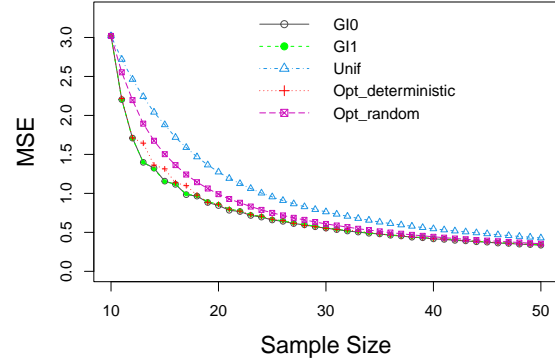


Figure 2: MSE of the MLE as sample size  $n$  varies.

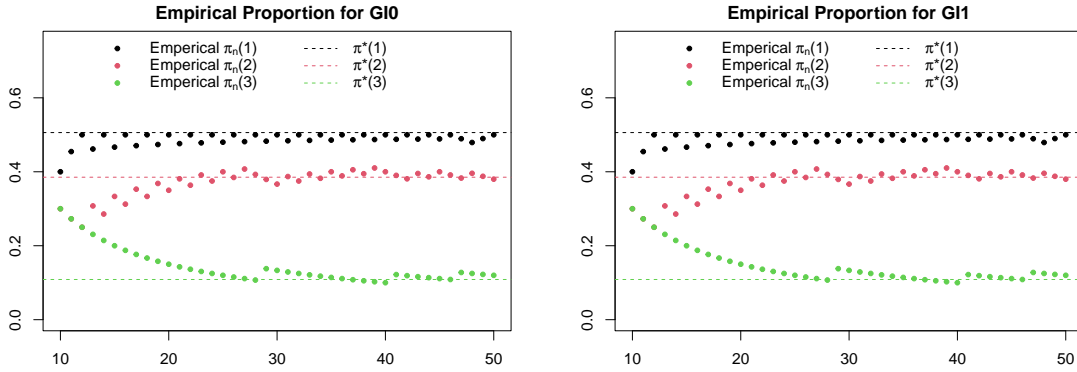


Figure 3: Empirical proportion  $\bar{\pi}_n(a)$  and the optimal proportion  $\pi(a)$  for  $a = 1, 2, 3$ .

## 7.2 Simulation Study 2

In our theoretical results, we assumed that  $|\mathcal{A}|$  and  $p$  are fixed and  $n$  grows to infinity. In this simulation study, we investigate the impact of large  $|\mathcal{A}|$  and  $p$  on the computational time and the performance of the proposed methods. Consider the sequential rank aggregation problem described in Section 5.3. We simulate the pairwise comparison results from a BTL model (see Equation (39)). Each coordinate of the true value of  $\theta \in \mathbb{R}^p$  are sampled independently from a uniform distribution  $\mathcal{U}(-2, 2)$ . We vary the value of  $p$  and  $|\mathcal{A}|$ , with  $p$  and  $|\mathcal{A}|$  ranging from 25 to 500 and from  $p$  to  $\frac{p(p+1)}{2}$ , respectively. The computation time is given by Table 2.

	$p = 25$			$p = 50$		
	$k = 25$	$k = 52$	$k = 325$	$k = 50$	$k = 102$	$k = 1275$
GI1 non-parallel	0.676 sec	0.430 sec	0.416 sec	0.963 sec	0.489 sec	1.046 secs
GI0 parallel	1.328 secs	1.070 secs	1.784 secs	1.639 secs	1.963 secs	10.296 secs
GI0 non-parallel	1.079 secs	1.325secs	4.790 secs	3.202 secs	5.030 secs	33.430 secs
	$p = 100$			$p = 500$		
	$k = 100$	$k = 202$	$k = 5050$	$k = 500$	$k = 1002$	$k = 125250$
GI1 non-parallel	2.028 secs	1.358 secs	10.758 secs	1.387 mins	57.900 secs	2.03 hours
GI0 parallel	6.736 secs	9.594 secs	3.240 mins	34.322 mins	1.168 hours	6 days
GI0 non-parallel	19.45 secs	35.065 secs	12.992 mins	1.925 hours	3.843 hours	about 20 days

Table 2: The computation time for solving the MLE and selecting a new experiment at a single time point, based on 100 Monte Carlo replications, is recorded for the non-paralleled GI1 algorithm as well as for the non-paralleled and paralleled versions of the GI0 algorithm. For each value of  $p$ ,  $k = |\mathcal{A}|$  takes values in  $p$ ,  $2(p + 1)$ , and  $\frac{p(p+1)}{2}$ . All computations are carried out on a MacBook Pro (13-inch, 2019) equipped with a 1.4 GHz Quad-Core Intel Core i5 processor.

Based on Table 2, the non-paralleled GI1 is much faster than both the non-paralleled and paralleled GI0 when both  $p$  and  $|\mathcal{A}|$  are large. This is consistent with Lemma 3.1.

We also perform additional Monte Carlo simulations to assess the estimation accuracy of the proposed methods, and to study how the choice of  $r$  in  $\Theta = [-r, r]^p$  affects the accuracy. Due to the page limit, details of these additional simulation studies are postponed to the supplementary material.

## 8 Real Data Example

We apply the proposed methods to a sushi preference dataset (Maystre and Grossglauser, 2017). This dataset contains feedback from 5,000 participants who ranked 10 different types of sushi, selected from a total of 100 types of sushi. Similar to the data pre-processing steps in Maystre and Grossglauser (2017), we first transform each 10-item ranking into pairwise comparison results, yielding  $\binom{10}{2} \times 5000 = 225000$  pairwise comparison results. We fit the BTL model described in Equation (39) with  $\Theta = [-3, 3]^p$  using all pairwise comparison data and treat the MLE of  $\theta$  as the ground truth. Under this setting,  $p = 99$ , and  $|\mathcal{A}| = 4809$ . We note that  $|\mathcal{A}| < \binom{100}{2}$  due to the absence of comparisons for some pairs in the dataset.

We vary the sample size  $n$  and compare the performance of the proposed GI0 and GI1 with two other experiment selection methods: uniform sampling and uncertainty sampling. Uncertainty sampling is a popular approach for active learning. In the context of sequential rank aggregation (see Maystre and Grossglauser (2017)), uncertainty sampling refers to

sampling the pair that is most difficult to distinguish. That is,

$$a_{n+1} = \arg \max_{a \in \mathcal{A}} [\min\{1 - f_{\hat{\theta}_n, a}(1), f_{\hat{\theta}_n, a}(0)\}] = \arg \min_{a=(i,j) \in \mathcal{A}} \{|\hat{\theta}_{n,i} - \hat{\theta}_{n,j}|\}. \quad (46)$$

The performance of the experiment selection rules is measured through the Kendall's  $\tau$  correlation, which is often used to measure the accuracy of ranking algorithms. Specifically, define Kendall's  $\tau$  correlation as

$$\tau(\hat{\theta}_n, \theta^*) = \binom{100}{2}^{-1} \sum_{1 \leq i < j \leq 100} \text{sign}(\hat{\theta}_{n,i} - \hat{\theta}_{n,j}) \cdot \text{sign}(\theta_i^* - \theta_j^*),$$

where  $\text{sign}$  denotes the sign function,  $\hat{\theta}_n$  denotes the MLE based on  $n$  observations, and  $\theta^*$  is the ground truth obtained using the MLE based on all 225000 comparisons.

Figure 4 illustrates the Kendall's  $\tau$  coefficient for GI0, GI1, uniform selection and uncertainty sampling for different number of comparisons  $n$ , based on a Monte Carlo simulation with 100 replications. From Figure 4, GI0 and GI1 behave similarly, and both outperform uniform selection and uncertainty sampling. Additional details of the Monte Carlo simulation are provided in the supplementary material.

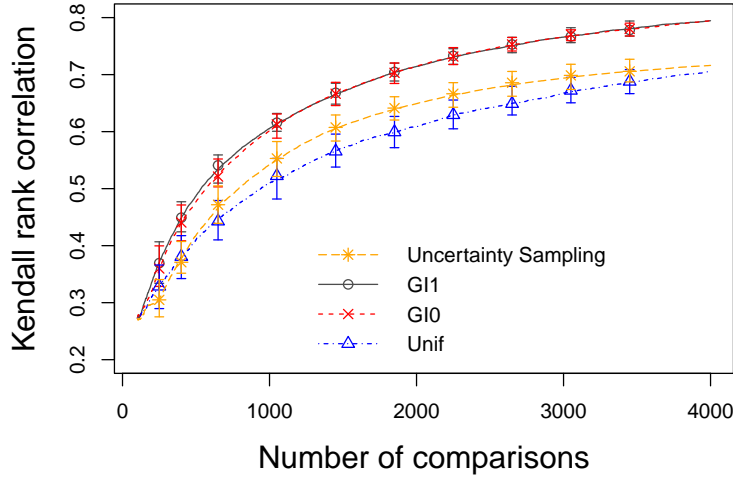


Figure 4: Comparison of different selection methods through Kendall's  $\tau$  coefficient. The averaged Kendall's  $\tau$  correlation between  $\theta^*$  and  $\hat{\theta}_n^{\text{ML}}$  versus the number of comparisons is plotted, along with the first and third quartiles, following different active experiment selection methods.



## 9 Conclusion and Further Discussion

In this study, we consider the problem of efficient sequential design for active sequential estimation. This problem has widespread applications across different fields; however, a systematic statistical analysis is lacking for the multidimensional case. We introduce a class of experiment selection rules that not only covers existing methods but also presents new approaches with improved numerical efficiency. Furthermore, we provide theoretical analysis including the consistency, asymptotic normality, and asymptotic optimality of the MLE following the proposed selection rule. These findings are also extended to scenarios involving early stopping rules, which are commonly used in practice. The theoretical results are highly non-trivial, and standard techniques in the literature of sequential decision making and stochastic control are not applicable. We have developed new analytical tools to tackle the theoretical challenges, which are important on their own and may be reused for other related problems.

The current study can be extended in several directions. First, in some applications, different experiments are associated with varying sampling cost. The current method may be extended to incorporate the sampling cost in the experiment selection rules. We expect similar analytical tools can be used in the theoretical analysis. Second, theoretical results can be extended to the case where  $p$  and  $k$  slowly grow to infinity as  $n$  grows. On the other hand, the consistency results do not hold under the high-dimensional setting where  $p \geq n$ . Some modifications to the estimation and experiment selection methods are necessary to ensure valid statistical inference in this case. Third, nuisance parameters may be present in some applications, where we are only interested in estimating part of the parameter efficiently. In this case, the proposed GI0 and GI1 still lead to a consistent and asymptotically normal MLE. However, the estimator may be asymptotically inefficient when there are redundant experiments measuring nuisance parameters. Of interest is how to design an experiment selection rule and an estimator to achieve asymptotic optimality. This is worth further investigation.

# Supplement to “Globally-Optimal Greedy Experiment Selection for Active Sequential Estimation”

This supplement contains additional simulation results, specifications for simulation and real data analysis, and technical proof for all the theoretical results.

## 10 Detailed Specifications for Simulation Studies

In this section, we provide detailed specifications for the simulation studies in Section 7. Recall that  $\mathbb{G}_\theta(\Sigma) = \text{tr}(\Sigma)$  throughout the simulation study.

### 10.1 Detailed Specifications for Section 7.1

#### 10.1.1 Algorithm for Solving the Optimal Optimal Selection Proportion

To solve the optimal selection proportion  $\pi^*$  numerically, we apply the projected gradient descent algorithm over the simplex  $\mathcal{S}^{\mathcal{A}}$  (see, e.g., Chen and Ye (2011)). Let  $P_{\mathcal{S}^{\mathcal{A}}}$  denote the projection operator onto the simplex  $\mathcal{S}^{\mathcal{A}}$ . Initializing  $\pi_0 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ , the iterative algorithm is given by

$$\pi_{n+1} = P_{\mathcal{S}^{\mathcal{A}}}(\pi_n - \eta \nabla \mathbb{F}_{\theta^*}(\pi_n)),$$

where the learning rate  $\eta$  is set to 0.001 and the maximum number of iterations is set to 10000.

#### 10.1.2 GI0 and GI1 Implementation

When implementing GI0 and GI1, the first two lines of the algorithm requires the input  $\hat{\theta}_0$  and  $a_1^0, \dots, a_{n_0}^0$ . Here we specify  $\hat{\theta}_0 = (0, 0)^T$ ,  $n_0 = 9$ , and  $a_1 = a_4 = a_7 = 1$ ,  $a_2 = a_5 = a_8 = 2$ , and  $a_3 = a_6 = a_9 = 3$ .

Additionally, the MLE is solved using the R function `glmnet` function from the R package `glmnet` with the constraint  $\Theta = [-3, 3]^2$ .

#### 10.1.3 Coverage Probability for CIs

The coverage probability of confidence intervals described (45) is estimated as follows:

$$\frac{1}{N} \sum_{j=1}^N I\left(|\mathbf{d}^T \hat{\theta}_n^j - \mathbf{d}^T \theta^*| \leq \frac{Z_{\alpha/2}}{\sqrt{n}} \left\| \left\{ \mathcal{I}^{\bar{\pi}_n}(\hat{\theta}_n^{\text{ML}}) \right\}^{-1/2} \mathbf{d} \right\| \right), \text{ for } n \in \{25, 50, 100\},$$

and

$$\frac{1}{N} \sum_{j=1}^N I(|\mathbf{d}^T \hat{\boldsymbol{\theta}}_{\tau_c}^j - \mathbf{d}^T \boldsymbol{\theta}^*| \leq Z_{\alpha/2} \cdot c),$$

where  $N = 1000$  is the number of Monte Carlo simulation,  $\hat{\boldsymbol{\theta}}^j$  and  $\hat{\boldsymbol{\theta}}_{\tau_c}^j$  are the MLE obtained in the  $j$ -th Monte Carlo replication with the sample size  $n$  and  $\tau_c$ , respectively.

## 10.2 Detailed Specifications for Section 7.2

Let  $\boldsymbol{\Theta} = [-3, 3]^p$ . To solve for the MLE (constrained in  $\boldsymbol{\Theta}$ ), we use the `glmnet` function from the R package `glmnet`. The computation time shown in Table 2 is determined using RStudio.

### 10.2.1 Sampling of $\mathcal{A}$

For a sequential rank aggregation problem under a BTL model, the graph  $G = (\{0, \dots, p\}, \mathcal{A})$  needs to be a connected graph for the identifiability of the model parameter (see Corollary 5.3). This implies that  $\mathcal{A}$  needs to satisfy some condition rather than being an arbitrary set of pairs to ensure the identifiability of the problem. Below we describe the random sampling scheme of  $\mathcal{A}$  used in the Monte Carlo simulation which ensures the connectivity of  $G$ . Note that  $|\mathcal{A}| \in \{p, 2(p+1), \frac{p(p+1)}{2}\}$  in the simulation study.

- If  $|\mathcal{A}| = \frac{p(p+1)}{2}$ ,  $G$  is a fully connected graph, meaning that  $\mathcal{A}$  collects all the pairs among the  $p+1$  objects. In this case,  $\mathcal{A}$  is fixed throughout the Monte Carlo simulation.
- If  $|\mathcal{A}| = p$ , a connected  $G$  is equivalent to that  $G$  is a minimal spanning tree for a fully connected graph. In this case, we sample  $G$  uniformly from all minimal spanning trees in the Monte Carlo simulation. This is implemented using the function `sample_spanning_tree` from the R package `igraph`.
- If  $|\mathcal{A}| = 2(p+1)$ , we restrict  $G$  to be 4-regular, which means that each node from  $\{0, 1, \dots, p\}$  has exactly 4 neighbors. In this case, we sample  $G$  uniformly from all 4-regular graphs. This is implemented using the function `sample_k_regular` from the R package `igraph`.

### 10.2.2 Initial estimator $\hat{\boldsymbol{\theta}}_0$ and experiments $a_1^0, \dots, a_{n_0}^0$

For implementing GI0 and GI1, we set  $\hat{\boldsymbol{\theta}}_0 = \mathbf{0}$  in Algorithms 6.

According to Corollary 5.3 in Section 5.3, the initial experiments needs to be selected so that  $G_0 = (\{0, \dots, p\}, \{a_1^0, \dots, a_{n_0}^0\})$  is a connected subgraph of  $G$ . Here, elements of

$\{a_1^0, \dots, a_{n_0}^0\}$  are not necessary to be distinct. Throughout the Monte Carlo simulation, we set  $n_0 = 4p$ , and sample  $G_0$  randomly using the following steps and collect the edges in  $G_0$  to form the set of initial experiments  $\{a_1^0, \dots, a_{n_0}^0\}$ .

Step 1: Sample uniformly from all the minimal spanning trees from  $G$ , which is implemented using the R function `sample_spanning_tree`. Let  $(\{0, \dots, p\}, \{a_1^{\text{tree}}, \dots, a_p^{\text{tree}}\})$  denote the sampled tree.

Step 2: Randomly sample  $3p$  pairs from  $\mathcal{A}$  without replacement. Let  $\{a'_{p+1}, \dots, a'_{4p}\}$  denote all sets of pairs (possibly repeated) sampled from this step.

Step 3:  $\{a_1^0, \dots, a_{n_0}^0\} = \{a_1^{\text{tree}}, \dots, a_p^{\text{tree}}, a'_{p+1}, \dots, a'_{4p}\}$  collects all the edges generated in the first and second steps.

Among the steps mentioned above, the first step yields a connected subgraph of  $G$  with  $p$  edges, and the second step expands this subgraph into another connected subgraph to have at most  $p + 3p = 4p$  edges.

### 10.2.3 Algorithm Acceleration

The accelerated GI1 algorithm (as described in Algorithm 3) is employed for GI1 selection, because in sequential rank aggregation problem the information matrix can be decomposed into a structure that is both sparse and of low rank with  $s = 2$  (see Lemma 3.1). To accelerate GI0 Algorithm 6, we parallel the calculation of (2) when  $|\mathcal{A}|$  is large.

## 11 Detailed Specifications for the Real Data Analysis in Section 8

### 11.1 Data Structure

The transformed dataset contains 225000 pairwise comparison results. We list these comparison results as the dataset  $\mathcal{D} = \{(a^{(i)}, X^{(i)})\}_{i=1}^T$ , where  $T = 225000$ ,  $a^{(i)}$  indicates the pairs to compare and  $X^{(i)}$  is binary, indicating the corresponding pairwise comparison result. We note that  $\mathcal{D}$  is a multiset, meaning that it may have repeated elements.

### 11.2 Sequential Sampling of the Pairwise Comparison Data

We note that, for the real data analysis, each element in the data set  $\mathcal{D}$  is sampled at most once, to prevent the redundancy of using the same data points multiple times. As a result,

when we implementing an active sampling scheme for the real data analysis, we will always sample elements from  $\mathcal{D}$  without replacement.

Specifically, for all the experiment selection methods compared in this section, we set  $n_0 = p = 99$ , and generate the initial experiments  $\{a_1^0, \dots, a_{n_0}^0\}$  randomly following the Step 1 procedure in Section 10.2.2. For each  $j \in \{1, 2, \dots, n_0\}$ , we sample the initial pairwise comparison results as follows: we sample  $s_j$  uniformly from  $\{i : a^{(i)} = a_j^0, 1 \leq i \leq T\}$ . Then, the initial pairs and comparison results are given by  $(a^{(s_1)}, X^{(s_1)}), \dots, (a^{(s_{n_0})}, X^{(s_{n_0})})$ . This gives the initial data  $(a_1^0, X_1), \dots, (a_{n_0}^0, X_{n_0})$ .

Let  $S_{n_0} = \{s_1, \dots, s_{n_0}\}$ ,  $[T] = \{1, 2, \dots, T\}$ . For each  $S \subset [T]$ , define

$$\mathcal{A}_S = \{a^{(i)} \in \mathcal{A} : i \in [T] \setminus S\}.$$

Next, we provide details of the implementation of different adaptive pair selection rules for  $n > n_0$ .

Uniform sampling: For  $n = n_0, \dots, T - 1$ , sample  $s_{n+1}$  uniformly from  $[T] \setminus S_n$ . Let  $S_{n+1} = S_n \cup \{s_{n+1}\}$ . The  $(n + 1)$ -th pair and comparison result  $(a_{n+1}, X_{n+1})$  is given by  $(a^{(s_{n+1})}, X^{(s_{n+1})})$ .

GI0 and GI1: For  $n = n_0, \dots, T - 1$ , calculate  $a_{n+1}$  according to (2) and (3) with  $\mathcal{A}$  replaced by  $\mathcal{A}_{S_n}$  for GI0 and GI1, respectively. Next, we uniformly sample the index  $s_{n+1}$  from  $\{i \in [T] \setminus S_n : a^{(i)} = a_{n+1}\}$ . Let  $S_{n+1} = S_n \cup \{s_{n+1}\}$ . The  $(n + 1)$ -th pair and comparison result  $(a_{n+1}, X_{n+1})$  is given by  $(a^{(s_{n+1})}, X^{(s_{n+1})})$ .

certainty sampling: For  $n = n_0, \dots, T - 1$ , calculate  $a_{n+1}$  according to (46) with  $\mathcal{A}$  replaced by  $\mathcal{A}_{S_n}$ . Next, we uniformly sample the index  $s_{n+1}$  from  $\{i \in [T] \setminus S_n : a^{(i)} = a_{n+1}\}$ . Let  $S_{n+1} = S_n \cup \{s_{n+1}\}$ . The  $(n + 1)$ -th pair and comparison result  $(a_{n+1}, X_{n+1})$  is given by  $(a^{(s_{n+1})}, X^{(s_{n+1})})$ .

### 11.3 Specifications for the Monte Carlo Experiments

For Figure 4, we perform a Monte Carlo simulation with 100 replications. For each replication, we sample 99 pairs of comparisons at random for initialization, and then perform sequential sampling for following different methods using the initialization and sampling method described in Section 11.2. We specify  $\mathbb{G}_\theta(\Sigma) = \text{tr}(\Sigma)$  for implementing GI0 and GI1 and  $\Theta = [-3, 3]^p$  to solve the MLE.

## 12 Additional Simulation Results

In this section, we present additional simulation results.

## 12.1 Additional Simulation Results for Simulation Study 1

Let the true value  $\boldsymbol{\theta}^* = (1, 0)^T$ . The initial estimator  $\hat{\boldsymbol{\theta}}_0$ ,  $n_0$ , and experiments  $\{a_1^0, \dots, a_{n_0}^0\}$  are selected according to Section 10.1.2. Let the sample size  $n = 50$ . Define

$$Z_1^j = \frac{\sqrt{N}(\hat{\theta}_1^j - \theta_1^*)}{(e_1^T \{\mathcal{I}^{\pi_n}(\hat{\boldsymbol{\theta}}_n^j)\}^{-1} e_1)^{1/2}} \text{ and } Z_2^j = \frac{\sqrt{N}(\hat{\theta}_2^j - \theta_2^*)}{(e_2^T \{\mathcal{I}^{\pi_n}(\hat{\boldsymbol{\theta}}_n^j)\}^{-1} e_2)^{1/2}},$$

where  $\hat{\boldsymbol{\theta}}_n^j = (\hat{\theta}_1^j, \hat{\theta}_2^j)^T$  represents the MLE of  $\boldsymbol{\theta}$  based on  $j$ -th Monte Carlo replication, and  $N = 1000$  is the number of Monte Carlo replications. That is,  $Z_1^j$  and  $Z_2^j$  are i.i.d. copies of  $Z_1 = \frac{\sqrt{N}(\hat{\theta}_1 - \theta_1^*)}{(e_1^T \{\mathcal{I}^{\pi_n}(\hat{\boldsymbol{\theta}}_n)\}^{-1} e_1)^{1/2}}$  and  $Z_2 = \frac{\sqrt{N}(\hat{\theta}_2 - \theta_2^*)}{(e_2^T \{\mathcal{I}^{\pi_n}(\hat{\boldsymbol{\theta}}_n)\}^{-1} e_2)^{1/2}}$ . In Figure 5, we plot the histogram for  $\{Z_1^j\}_{j=1}^N$  and  $\{Z_2^j\}_{j=1}^N$  following GI0 and GI1.

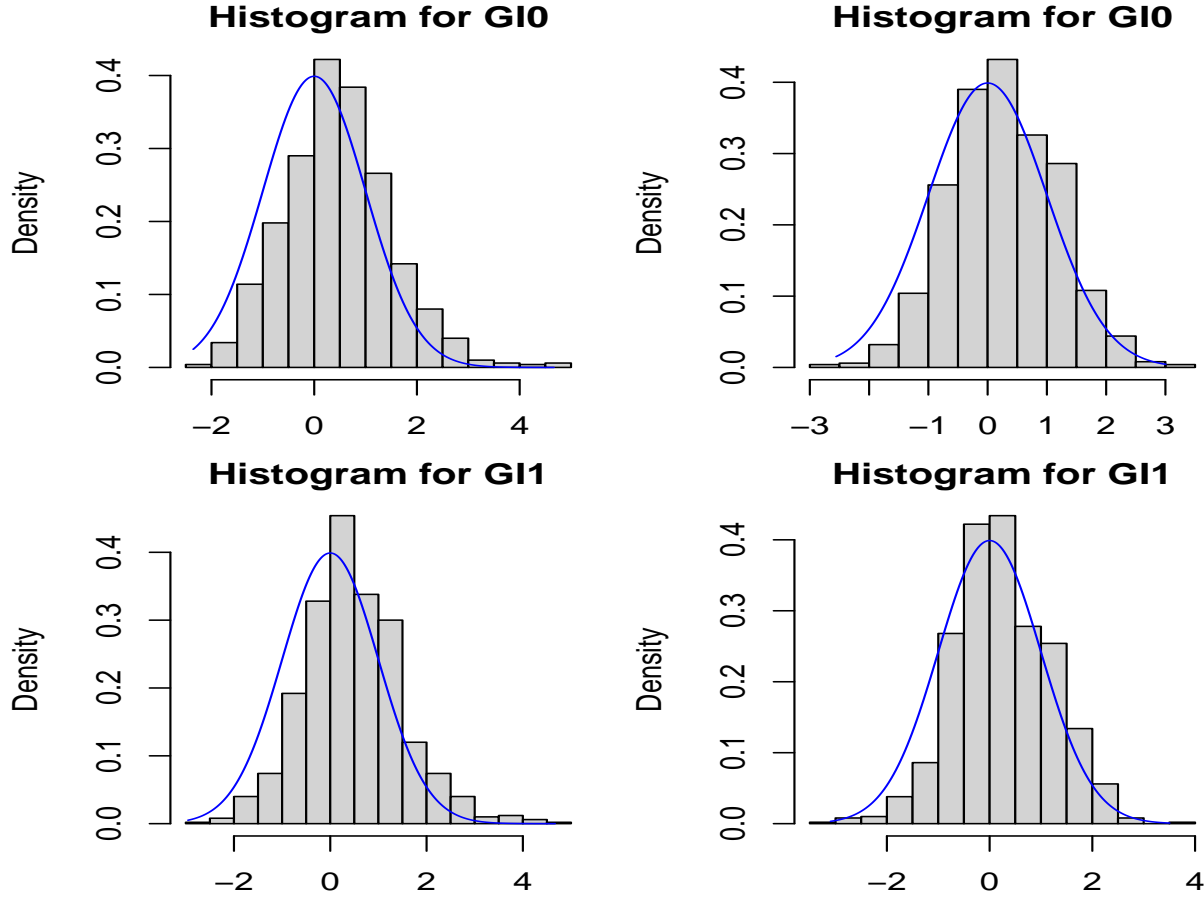


Figure 5: Histograms for  $\{Z_1^j\}_{j=1}^N$  and  $\{Z_2^j\}_{j=1}^N$  following GI0 and GI1, and the density curve for the standard normal distribution. The upper left and bottom left panels show the histogram of  $\{Z_1^j\}_{j=1}^N$  following GI0 and GI1, respectively. The upper right and bottom right panels show the histogram of  $\{Z_2^j\}_{j=1}^N$  following GI0 and GI1, respectively.

In Figure 5, the histogram closely approximates the standard normal density curve. This is consistent with Theorem 4.4.

## 12.2 Additional Simulation Results for Simulation Study 2

The theoretical results in the manuscript assume that  $p$  and  $|\mathcal{A}|$  are fixed while the sample size  $n$  grows large. In this section, we investigate the performance of the proposed method when  $p$  and  $|\mathcal{A}|$  are comparable with  $n$ , and this condition is violated. We investigate the performance of the proposed methods under a sequential rank aggregation problem assuming a BTL model.

We consider the following simulation settings. Set  $p = 10$  or  $50$ . Entries of  $\theta^*$  are i.i.d. and generated from  $\mathcal{U}(-2, 2)$ .  $|\mathcal{A}| = 2(p + 1)$ , and  $\mathcal{A}$  is sampled uniformly from all 4-regular graphs (see Section 10.2.1). The initial estimator and experiments are selected in the same way as those in Section 10.2.2 except that  $n_0$  is set as  $2p$  instead of  $4p$ .

### 12.2.1 Empirical and Optimal Frequency

We plot the expected value of  $F_n = \mathbb{F}_{\theta^*}(\bar{\pi}_n) - \mathbb{F}_{\theta^*}(\pi^*)$  for different methods in Figure 6 based on  $N = 1000$  Monte Carlo replications. Here, the optimal proportion  $\pi^*$  is computed according to Section 10.1.1. From the Figure 6, we see that  $F_n$  is approaching zero when  $n$  is large. This is consistent with Theorem 4.3. However, it is far from zero when  $p$  is comparable with  $n$  (e.g.,  $p = 50$  and  $n = 150$ ). This is expected as it becomes a high-dimensional problem under this setting.

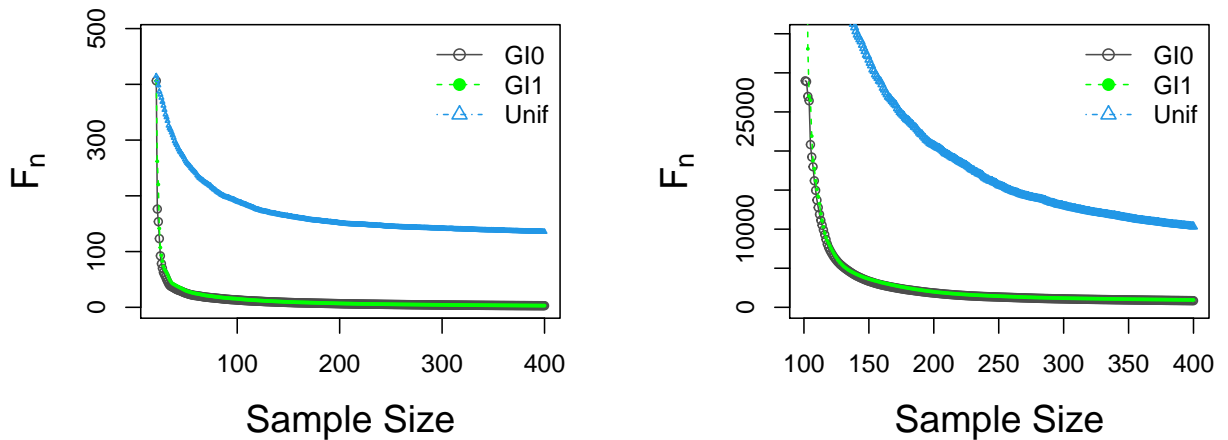


Figure 6: Comparison of  $F_n$  for different selection methods (GI0, GI1, Unif selections) and different sample size  $n$ . The left panel and the right panel show  $F_n$  with  $p = 10$  and  $p = 50$ , respectively.

### 12.2.2 Estimation Accuracy

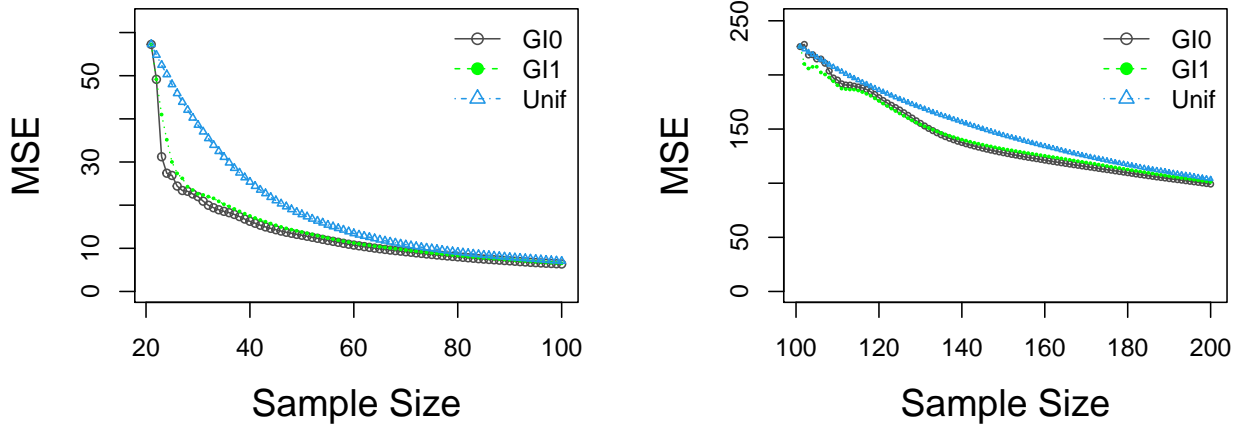


Figure 7: Comparison of performance of MSE for different selection methods (GI0, GI1, uniform selections) for the rank aggregation problem. The left panel and the right panel show MSE with  $p = 10$  and  $p = 50$ , respectively.

In Figure 7, we plot the MSE for the MLE at different sample size  $n$  following different experiment selection methods based on  $N = 10000$  Monte Carlo replications. The MSE is not close to zero when  $p$  is relatively large compared to  $n$ , which is expected. However, GI0 and GI1 still perform much better when compared with Unif.

### 12.2.3 Impact of the Choice of $\Theta$

In our theoretical results, we assume the true parameter  $\theta^*$  is an inner point of  $\Theta$ . In this section, we study the impact of the choice of  $\Theta$  on the estimation accuracy. We consider the following simulation setting:  $p = 50$ , each element of  $\theta^*$  is sampled i.i.d. from  $\mathcal{U}(-2, 2)$ ,  $\mathcal{A}$  is randomly sampled from all 4-regular graphs with  $N = 100$  Monte Carlo simulations. As a result,  $|\mathcal{A}| = 102$ . We consider 4 choices of  $\Theta$  when solving for the MLE:  $\Theta = [-1, 1]^p$ ,  $\Theta = [-2, 2]^p$ ,  $\Theta = [-3, 3]^p$  and  $\Theta = [-5, 5]^p$ . The initial sample size is set to  $n_0 = p$ .

In Figure 8, we compare the Kendall's correlation of the MLE following GI1 for different choices of  $\Theta$ , and obtain the following findings.

1. For cube 2 ( $\Theta = [-2, 2]^p$ ), it coincides with the data generation distribution  $\mathcal{U}(-2, 2)$ . The Kendall's  $\tau$  correlation is the largest among all cubes and sample sizes.
2. For cube 1 ( $\Theta = [-1, 1]^p$ ), it does not satisfy the condition  $\theta^* \in \Theta$  for the theoretical results. For small sample size ( $n < 500$ ), it performs similarly as cube 2. However, for larger  $n$ , it's performance becomes worse.



3. For cube 3 and cube 5 ( $\Theta = [-3, 3]^p$  and  $\Theta = [-5, 5]^p$ ), they cover the true model parameter, but are larger than the support of sampling distribution of  $\theta^*$ . For small sample size, the larger the cube, the poorer the performance is. However, as the sample size increases, the performance becomes better than cube 1.
4. Overall, the choice of  $r$  in  $\Theta = [-r, r]^p$  does not seem affect the overall trend between Kendall's correlation and sample size.

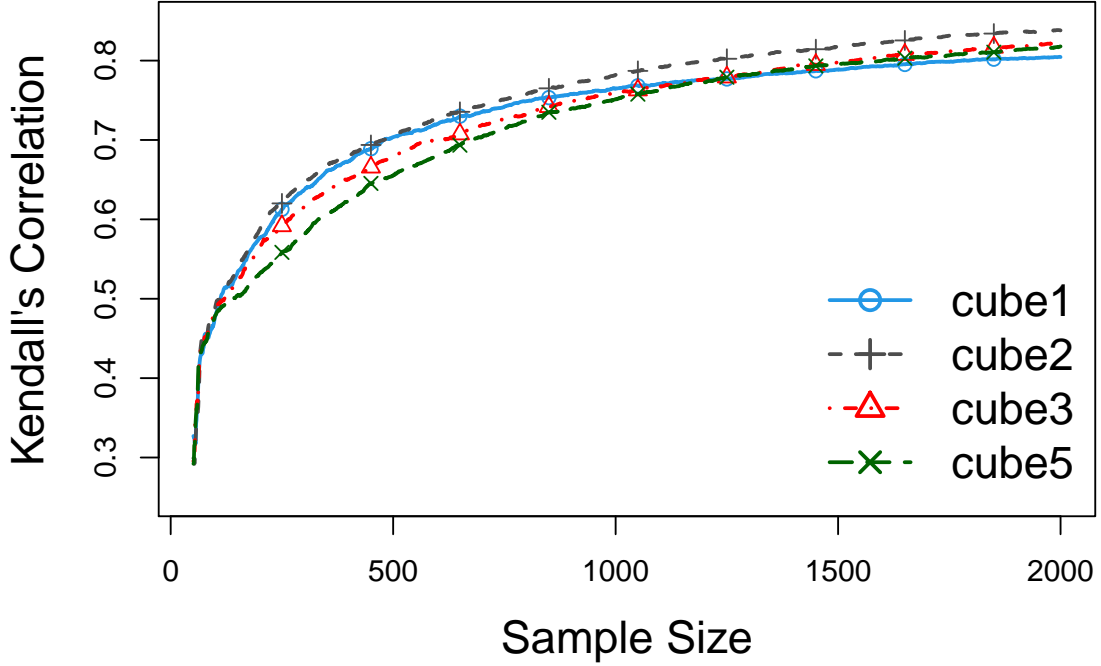


Figure 8: Comparison of Performance of Different Compact Cubes with GI1 Selection for the Rank Aggregation Problem. The curves for Cube1, Cube2, Cube3, and Cube5 represent the plot of Kendall's  $\tau$  correlation versus sample size over compact cubes  $\Theta = [-1, 1]^p$ ,  $\Theta = [-2, 2]^p$ ,  $\Theta = [-3, 3]^p$ , and  $\Theta = [-5, 5]^p$ , respectively.

## 13 Preliminary Theoretical Results and Supporting Lemmas

In this section, we present preliminary theoretical results and supporting lemmas which are useful for the rest of the theoretical analysis. Moreover, they may be useful for other problems involving the analysis of stochastic processes, functions of matrices, and linear algebra for spaces indexed by a parameter.

### 13.1 Useful Results for the Convergence of Stochastic Processes

The next lemma extends the classic Kolmogorov's three-series theorem with relaxed moments and independence conditions. It is useful for proving almost sure convergence results for dependent stochastic processes.

**Lemma 13.1** (Modified Kolmogorov's three-series theorem). *Consider nested  $\sigma$ -fields  $\mathcal{F}_n \subset \mathcal{F}_{n+1}, n \geq 0$ . Let  $\{X_n\}_{n=1}^\infty$  and  $\{\varepsilon_n\}_{n=1}^\infty$  be two sequences of random variables, adaptive to  $\{\mathcal{F}_n\}_{n=1}^\infty$ , respectively. Consider a sequence of events  $E_n$  such that*

$$\mathbb{P}(\liminf_n E_n) = \mathbb{P}\left(\bigcup_{n=1}^\infty \bigcap_{m=n}^\infty E_m\right) = 1.$$

*If there exists  $0 < \gamma \leq 1$  such that,*

$$\mathbb{E}[|X_n|^\gamma I_{E_n} \mid \mathcal{F}_{n-1}] \leq \varepsilon_{n-1} \text{ a.s., and, } \sum_{n=0}^\infty \mathbb{E}\varepsilon_n < \infty,$$

*then  $\sum_{n=1}^\infty X_n$  converges almost surely.*

*Proof of Lemma 13.1.* Let  $S_N = \sum_{n=1}^N X_n$ . It is sufficient to show that with probability 1,

$$\lim_{m \rightarrow \infty} \sup_{n, l \geq m} |S_n - S_l| = 0.$$

Applying  $C_r$  inequality (see 9.1.a in Lin (2010)), for any  $0 < \gamma \leq 1, k \geq 1$ , we have

$$\left| \sum_{i=1}^k X_{m+i} \right|^\gamma \leq \sum_{i=1}^k |X_{m+i}|^\gamma.$$

For any  $m \in \mathbb{N}$ , and  $\varepsilon > 0$ , applying  $C_r$  inequality (see 9.1.a in Lin (2010)) and Markov

inequality, we have

$$\begin{aligned}
& \mathbb{P} \left( \sup_{n,l \geq m} |S_n - S_l| \geq 2\varepsilon \right) \\
& \leq \mathbb{P} \left( 2 \sup_{k \in \mathbb{N}} \left| \sum_{i=1}^k X_{m+i} \right| \geq 2\varepsilon \right) \\
& = \mathbb{P} \left( \sup_{k \in \mathbb{N}} \left| \sum_{i=1}^k X_{m+i} \right|^\gamma \geq \varepsilon^\gamma \right) \\
& \leq \mathbb{P} \left( \sup_{k \in \mathbb{N}} \sum_{i=1}^k |X_{m+i}|^\gamma \geq \varepsilon^\gamma \right) \\
& \leq \mathbb{P} \left( \left\{ \sup_{k \in \mathbb{N}} \sum_{i=1}^k |X_{m+i}|^\gamma \geq \varepsilon^\gamma \right\} \cap \left( \bigcap_{n=m+1}^{\infty} E_n \right) \right) + \mathbb{P} \left( \overline{\bigcap_{n=m+1}^{\infty} E_n} \right) \\
& \leq \limsup_{k \rightarrow \infty} \mathbb{P} \left( \sum_{i=1}^k |X_{m+i}|^\gamma I \left( \bigcap_{n=m+1}^{\infty} E_n \right) \geq \varepsilon^\gamma \right) + \mathbb{P} \left( \overline{\bigcap_{n=m+1}^{\infty} E_n} \right) \\
& \leq \limsup_{k \rightarrow \infty} \mathbb{P} \left( \sum_{i=1}^k |X_{m+i}|^\gamma I(E_{m+i}) \geq \varepsilon^\gamma \right) + \mathbb{P} \left( \overline{\bigcap_{n=m+1}^{\infty} E_n} \right) \\
& \leq \limsup_{k \rightarrow \infty} \frac{1}{\varepsilon^\gamma} \sum_{i=1}^k \mathbb{E}[\mathbb{E} \{ |X_{m+i}|^\gamma I_{E_{m+i}} \mid \mathcal{F}_{m+i-1} \}] + \mathbb{P} \left( \overline{\bigcap_{n=m+1}^{\infty} E_n} \right) \\
& \leq \frac{1}{\varepsilon^\gamma} \sum_{i=1}^{\infty} \mathbb{E} \varepsilon_{m+i-1} + \mathbb{P} \left( \overline{\bigcap_{n=m+1}^{\infty} E_n} \right),
\end{aligned}$$

where we used the assumption  $\mathbb{E}[|X_n|^\gamma I_{E_n} \mid \mathcal{F}_{n-1}] \leq \varepsilon_{n-1}$  for all  $n$  for obtaining the last inequality. Notice that

$$\lim_{m \rightarrow \infty} \mathbb{P} \left( \overline{\bigcap_{n=m+1}^{\infty} E_n} \right) = 1 - \lim_{m \rightarrow \infty} \mathbb{P} \left( \bigcap_{n=m+1}^{\infty} E_n \right) = 1 - \mathbb{P} \left( \bigcup_{m=1}^{\infty} \bigcap_{n=m+1}^{\infty} E_n \right) = 0.$$

Let  $m \rightarrow \infty$ , we obtain that for all  $\varepsilon > 0$ ,

$$\begin{aligned}
& \mathbb{P} \left( \bigcap_{m \geq 1} \left\{ \sup_{n, l \geq m} |S_n - S_l| \geq 2\varepsilon \right\} \right) \\
&= \lim_{m \rightarrow \infty} \mathbb{P} \left( \left\{ \sup_{n, l \geq m} |S_n - S_l| \geq 2\varepsilon \right\} \right) \\
&\leq \frac{1}{\varepsilon^\gamma} \lim_{m \rightarrow \infty} \sum_{i=1}^{\infty} \mathbb{E} \varepsilon_{m+i-1} + \lim_{m \rightarrow \infty} \mathbb{P} \left( \bigcap_{n=m+1}^{\infty} E_n \right) \\
&= 0
\end{aligned}$$

This implies  $\mathbb{P} \left( \bigcap_{m \geq 1} \left\{ \sup_{n, l \geq m} |S_n - S_l| \geq 2\varepsilon \right\} \right) = 0$  and completes the proof.  $\square$

Next, we extends Theorem 2.19 in Hall and Heyde (1980) obtain a law of large number result for martingale differences which allows for adaptive experiment selection.

**Lemma 13.2** (Modified Theorem 2.19 in Hall and Heyde (1980)). *Let  $\{X_n\}_{n=1}^{\infty}$  be a sequence of random variables and  $\{\mathcal{F}_n\}_{n=1}^{\infty}$  be an increasing sequence of  $\sigma$ -fields with  $X_n$  measurable with respect to  $\mathcal{F}_n$  for all  $n$ . Let  $\{a_n\}_{n=1}^{\infty}$  denote a sequence of discrete random variables, where each variable takes values from the set  $\{1, 2, \dots, k\}$ . Let  $X^1, \dots, X^k$  be a sequence of random variables such that  $\max_{1 \leq a \leq k} \mathbb{E}|X^a| < \infty$ . If the conditional distribution function of  $X_n | \mathcal{F}_{n-1}, a_n = a$  is the same as the distribution function  $X^a$  with probability 1, then*

$$n^{-1} \sum_{i=1}^n \{X_i - \mathbb{E}(X_i | \mathcal{F}_{i-1})\} \xrightarrow{a.s.} 0. \quad (47)$$

*Proof of Lemma 13.2.* Let  $Y_n = X_n I_{\{|X_n| \leq n\}}$ ,  $n \geq 1$ .

Note that  $\mathbb{E}|X^a| < \infty$  for any  $1 \leq a \leq k$ , and for any  $x > 0$ ,

$$\begin{aligned}
\mathbb{P}(|X_n| > x) &= \mathbb{E} \mathbb{P}(|X_n| > x | \mathcal{F}_{n-1}) = \mathbb{E} \sum_{a=1}^k \mathbb{P}(|X_n| > x | \mathcal{F}_{n-1}, a_n = a) \mathbb{P}(a_n = a | \mathcal{F}_{n-1}) \\
&= \mathbb{E} \sum_{a=1}^k \mathbb{P}(|X^a| > x) \mathbb{P}(a_n = a | \mathcal{F}_{n-1}) \leq \sum_{a=1}^k \mathbb{P}(|X^a| > x) < \infty.
\end{aligned}$$

Similar to the proof of Theorem 2.19 in Hall and Heyde (1980), we obtain that

$$\begin{aligned}
& \sum_{n=1}^{\infty} \frac{1}{n^2} \mathbb{E}[\{Y_n - \mathbb{E}(Y_n | \mathcal{F}_{n-1})\}^2] \leq 2 \sum_{n=1}^{\infty} \frac{1}{n^2} \int_{0 < x \leq n} x \mathbb{P}(|X_n| > x) dx \\
&\leq 2 \sum_{a=1}^k \sum_{n=1}^{\infty} \frac{1}{n^2} \int_{0 < x \leq n} x \mathbb{P}(|X^a| > x) dx \leq 4 \sum_{a=1}^k \sum_{i=1}^{\infty} \mathbb{P}(|X^a| > i-1) < \infty,
\end{aligned}$$

$$n^{-1} \sum_{i=1}^n \{Y_i - \mathbb{E}(Y_i | \mathcal{F}_{i-1})\} \xrightarrow{\text{a.s.}} 0,$$

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n \neq Y_n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_n| > n) \leq \sum_{a=1}^k \sum_{n=1}^{\infty} \mathbb{P}(|X^a| > n) < \infty$$

and

$$n^{-1} \sum_{i=1}^n \{X_i - \mathbb{E}(Y_i | \mathcal{F}_{i-1})\} \xrightarrow{\text{a.s.}} 0. \quad (48)$$

Notice that with probability 1, as  $n \rightarrow \infty$ ,

$$\begin{aligned} & \mathbb{E}(|X_n| I(|X_n| > n) | \mathcal{F}_{n-1}) \\ &= \int_n^{\infty} \mathbb{P}(|X_n| > x | \mathcal{F}_{n-1}) dx \\ &= \int_n^{\infty} \sum_{a=1}^k \mathbb{P}(|X_n| > x | \mathcal{F}_{n-1}, a_n = a) \mathbb{P}(a_n = a | \mathcal{F}_{n-1}) dx \\ &\leq \int_n^{\infty} \sum_{a=1}^k \mathbb{P}(|X^a| > x) dx \\ &= \sum_{a=1}^k \mathbb{E}(|X^a| I(|X^a| > n)) \\ &\rightarrow 0. \end{aligned}$$

Thus, with probability 1,

$$\begin{aligned} & n^{-1} \sum_{i=1}^n |\mathbb{E}(X_i - Y_i | \mathcal{F}_{i-1})| \\ &\leq n^{-1} \sum_{i=1}^n \mathbb{E}(|X_i| I(|X_i| > i) | \mathcal{F}_{i-1}) \\ &\leq n^{-1} \sum_{i=1}^n \sum_{a=1}^k \mathbb{E}\{|X_i| I(|X_i| > i) | \mathcal{F}_{i-1}, a_i = a\} \mathbb{P}(a_i = a | \mathcal{F}_{i-1}) \\ &\leq \sum_{a=1}^k \frac{1}{n} \sum_{i=1}^n \mathbb{E}(|X^a| I(|X^a| > i)). \end{aligned} \quad (49)$$

Because  $\mathbb{E}|X^a| < \infty$ , we know that  $\lim_{n \rightarrow \infty} \mathbb{E}(|X^a| I(|X^a| > n)) = 0$ . Because the arithmetic mean of a sequence converges to the same limit as the sequence itself, we obtain that for all

$a \in \{1, \dots, k\}$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E}(|X^a| I(|X^a| > i)) = 0.$$

In conclusion, we obtain, with probability 1, that

$$\left| n^{-1} \sum_{i=1}^n \mathbb{E}(X_i - Y_i \mid \mathcal{F}_{i-1}) \right| \leq \sum_{a=1}^k \frac{1}{n} \sum_{i=1}^n \mathbb{E}(|X^a| I(|X^a| > i)),$$

and the expression on the right-hand side is a deterministic sequence converging to 0, which implies that

$$n^{-1} \sum_{i=1}^n \mathbb{E}(X_i - Y_i \mid \mathcal{F}_{i-1}) \xrightarrow{\text{a.s.}} 0. \quad (50)$$

Combining (48) and (50), we obtain (47).  $\square$

Anscombe's theorem (Anscombe, 1952) is a classic limit theorem for randomly indexed processes. We prove a multivariate version of Anscombe's theorem as follows, which generalizes the univariate Anscombe's theorem with Gaussian limit (see Mukhopadhyay and Chattopadhyay (2012)).

**Theorem 13.3** (Multivariate Anscombe's theorem). *Let  $\{T_n\}_{n \geq 1}$  be a sequence of column random vectors and  $\{W_n\}_{n \geq 1}$  be a sequence of positive definite matrices satisfying multivariate Anscombe's condition, namely, for every  $\varepsilon > 0$ ,  $0 < \gamma < 1$  there exists some  $\delta > 0$  such that*

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left( \max_{|n' - n| \leq \delta n} \|T_{n'} - T_n\| \geq \varepsilon \lambda_{\min}(W_n) \right) < \gamma$$

*hold. Moreover, we assume that*

$$\sup_{n \geq 1} \frac{\lambda_{\max}(W_n)}{\lambda_{\min}(W_n)} < \infty,$$

*and there exists positive sequence  $\rho_n \rightarrow \infty$  such that*

$$\rho_n W_n \xrightarrow{\mathbb{P}^*} W, \quad (51)$$

*where  $W$  is a real positive definite matrix.*

*Assume that there exists a real column vector  $\theta \in \mathbb{R}^p$  and as  $n \rightarrow \infty$*

$$W_n^{-1}(T_n - \theta) \xrightarrow{d} N_p(\mathbf{0}_p, I_p).$$

*Consider  $\{N_n\}_{n \geq 1}$ , a sequence of positive integer-valued stopping times defined on the same*

probability space where  $\{T_n\}_{n \geq 1}$  is defined. Let  $\{r_n\}_{n \geq 1}$  be an increasing sequence of positive integers such that  $\lim_{n \rightarrow \infty} r_n = \infty$ . If  $N_n/r_n \rightarrow 1$  in probability as  $n \rightarrow \infty$ , then as  $n \rightarrow \infty$

$$W_{r_n}^{-1}(T_{N_n} - \theta) \xrightarrow{d} N_p(\mathbf{0}_p, I_p).$$

*Proof of Theorem 13.3.* For any  $b \in \mathbb{R}^p$  such that  $\|b\| = 1$ , we have

$$\frac{b^T(T_n - \theta)}{\|W_n b\|} = \frac{(W_n b)^T}{\|W_n b\|} W_n^{-1}(T_n - \theta).$$

Let  $h_n = \frac{W_n b}{\|W_n b\|} = \frac{\rho_n W_n b}{\|\rho_n W_n b\|}$  and  $h = \frac{Wb}{\|Wb\|}$ . By the continuous mapping theorem, we know that  $h_n \rightarrow h$  in probability.

Recall that  $W_n^{-1}(T_n - \theta) \xrightarrow{d} N_p(\mathbf{0}_p, I_p)$ . Hence, by Slutsky theorem, as  $n \rightarrow \infty$

$$\frac{(W_n b)^T}{\|W_n b\|} W_n^{-1}(T_n - \theta) = h^T W_n^{-1}(T_n - \theta) + \left(h_n - h\right)^T W_n^{-1}(T_n - \theta) \xrightarrow{d} N(0, 1).$$

In conclusion, we know that

$$\frac{b^T T_n - b^T \theta}{\|W_n b\|} \xrightarrow{d} N(0, 1).$$

Note that  $N_n/r_n \rightarrow 1$  in probability and

$$\mathbb{P}\left(\max_{|n'-n| \leq \delta n} |b^T T_n - b^T \theta| \geq \varepsilon \|W_n b\|\right) \leq \mathbb{P}\left(\max_{|n'-n| \leq \delta n} \|T_n - \theta\| \geq \varepsilon \cdot \lambda_{\min}(W_n)\right),$$

and

$$\limsup_{n \rightarrow \infty} \mathbb{P}\left(\max_{|n'-n| \leq \delta n} \|T_n - \theta\| \geq \varepsilon \cdot \lambda_{\min}(W_n)\right) \leq \gamma.$$

Applying Theorem 3.1 in Mukhopadhyay and Chattopadhyay (2012), we obtain that for any  $b \neq 0$  and as  $n \rightarrow \infty$

$$\frac{b^T T_{N_n} - b^T \theta}{\|W_{r_n} b\|} \xrightarrow{d} N(0, 1). \quad (52)$$

Furthermore, by (51), we know that

$$\rho_{r_n} \|W_{r_n} b\| \rightarrow \|Wb\|. \quad (53)$$

Thus, we know that  $\rho_{r_n} b^T(T_{N_n} - \theta) = O_p(1)$  for any  $b \in \mathbb{R}^p$ , which implies

$$\rho_{r_n}(T_{N_n} - \theta) = O_p(1). \quad (54)$$

Note that

$$\begin{aligned}
& \|W_{r_n}^{-1}(T_{N_n} - \theta)\| \\
& \leq \frac{\|T_{N_n} - \theta\|}{\lambda_{\min}(W_{r_n})} \\
& \leq \sum_{i=1}^p \frac{|\mathbf{e}_i^T(T_{N_n} - \theta)|}{\|W_{r_n} \mathbf{e}_i\|} \sup_{n \geq 1} \kappa(W_n) \\
& = O_p(1),
\end{aligned}$$

where

$$\kappa(W_n) = \frac{\lambda_{\max}(W_n)}{\lambda_{\min}(W_n)}.$$

By Cramér–Wold theorem (see Billingsley (1999) p383), it is sufficient to show that for all  $h \in \mathbb{R}^p$  such that  $\|h\| = 1$ , we have

$$h^T W_{r_n}^{-1}(T_{N_n} - \theta) \xrightarrow{d} N(0, 1).$$

Set  $b_n = \frac{W_{r_n}^{-1}h}{\|W_{r_n}^{-1}h\|}$ , and  $b = \frac{W^{-1}h}{\|W^{-1}h\|}$ . We have  $h = \frac{W_{r_n} b_n}{\|W_{r_n} b_n\|}$ . By the continuous mapping theorem, we know that  $b_n \rightarrow b$  in probability. Notice that

$$\left| \frac{\|W_{r_n} b_n\|}{\|W_{r_n} b\|} - 1 \right| \leq \kappa(W_{r_n}) \|b_n - b\| \rightarrow 0,$$

which implies  $\frac{\|W_{r_n} b_n\|}{\|W_{r_n} b\|} \rightarrow 1$  in probability. Combine this with (52), we obtain that as  $n \rightarrow \infty$

$$\begin{aligned}
& h^T W_{r_n}^{-1}(T_{N_n} - \theta) \\
& = \frac{b_n^T(T_{N_n} - \theta)}{\|W_{r_n} b\|} \frac{\|W_{r_n} b\|}{\|W_{r_n} b_n\|} \\
& = \frac{b_n^T(T_{N_n} - \theta)}{\|W_{r_n} b\|} (1 + o_p(1)) \\
& = \left\{ \frac{b^T(T_{N_n} - \theta)}{\|W_{r_n} b\|} + \frac{(b_n - b)^T \rho_{r_n}(T_{N_n} - \theta)}{\rho_{r_n} \|W_{r_n} b\|} \right\} (1 + o_p(1)).
\end{aligned}$$

Combining (53), (54) and  $b_n \rightarrow b$  in probability, we know that

$$\frac{(b_n - b)^T \rho_{r_n}(T_{N_n} - \theta)}{\rho_{r_n} \|W_{r_n} b\|} = o_p(1),$$



which implies that

$$h^T W_{r_n}^{-1} (T_{N_n} - \theta) = \left\{ \frac{b^T (T_{N_n} - \theta)}{\|W_{r_n} b\|} + o_p(1) \right\} (1 + o_p(1)) \xrightarrow{d} N(0, 1).$$

Thus, we know that for any  $h \neq 0$ , as  $n \rightarrow \infty$

$$h^T W_{r_n}^{-1} (T_{N_n} - \theta) \xrightarrow{d} N(0, \|h\|^2).$$

By Cramér–Wold theorem (see Billingsley (1999) p383), we complete the proof of Theorem 13.3.  $\square$

## 13.2 Results regarding Functions of Matrices

In this section, we provide results on derivatives of functions of matrices, and properties on functions of a convex combination of matrices.

**Lemma 13.4.** *Let  $\mathcal{I}_a, a \in \mathcal{A}$  be a sequence of positive semidefinite matrix. Assume  $\pi_0(a) \geq 0, a \in \mathcal{A}$  (not necessary that  $\pi_0 \in \mathcal{S}^{\mathcal{A}}$ ) such that  $\sum_{a' \in \mathcal{A}} \pi_0(a') \mathcal{I}_{a'}$  is a real positive definite matrix. Then, for all  $a \in \mathcal{A}$  we have*

$$\left. \frac{\partial (\sum_{a' \in \mathcal{A}} \pi(a') \mathcal{I}_{a'})^{-1}}{\partial \pi(a)} \right|_{\pi = \pi_0} = - \left( \sum_{a' \in \mathcal{A}} \pi_0(a') \mathcal{I}_{a'} \right)^{-1} \mathcal{I}_a \left( \sum_{a' \in \mathcal{A}} \pi_0(a') \mathcal{I}_{a'} \right)^{-1}. \quad (55)$$

*Proof of Lemma 13.4.* By definition,

$$\left. \frac{\partial (\sum_{a' \in \mathcal{A}} \pi(a') \mathcal{I}_{a'})^{-1}}{\partial \pi(a)} \right|_{\pi = \pi_0} = \lim_{\varepsilon \rightarrow 0} \frac{(\sum_{a' \in \mathcal{A}} \pi_0(a') \mathcal{I}_{a'} + \varepsilon \mathcal{I}_a)^{-1} - (\sum_{a' \in \mathcal{A}} \pi_0(a') \mathcal{I}_{a'})^{-1}}{\varepsilon}.$$

Because  $\sum_{a' \in \mathcal{A}} \pi_0(a') \mathcal{I}_{a'}$  is a positive definite matrix, then for small enough  $\varepsilon$ , the inverse of  $\sum_{a' \in \mathcal{A}} \pi_0(a') \mathcal{I}_{a'} + \varepsilon \mathcal{I}_a$  exists. Furthermore,

$$\left( \sum_{a' \in \mathcal{A}} \pi_0(a') \mathcal{I}_{a'} + \varepsilon \mathcal{I}_a \right)^{-1} - \left( \sum_{a' \in \mathcal{A}} \pi_0(a') \mathcal{I}_{a'} \right)^{-1} = -\varepsilon \left( \sum_{a' \in \mathcal{A}} \pi_0(a') \mathcal{I}_{a'} + \varepsilon \mathcal{I}_a \right)^{-1} \mathcal{I}_a \left( \sum_{a' \in \mathcal{A}} \pi_0(a') \mathcal{I}_{a'} \right)^{-1},$$

and

$$\lim_{\varepsilon \rightarrow 0} \left( \sum_{a' \in \mathcal{A}} \pi_0(a') \mathcal{I}_{a'} + \varepsilon \mathcal{I}_a \right)^{-1} = \left( \sum_{a' \in \mathcal{A}} \pi_0(a') \mathcal{I}_{a'} \right)^{-1},$$

which implies (55).  $\square$

Next, we will define and derive the Gateaux derivative of the criteria function  $\Phi_q$ . For a

real positive definite matrix  $\Sigma$ , recall that

$$\Phi_0(\Sigma) = \log |\Sigma|, \quad \Phi_q(\Sigma) = \text{tr } \Sigma^q, \quad 0 < q < 1,$$

and

$$\Phi_q(\Sigma) = (\text{tr}(\Sigma^q))^{1/q}, \quad q \geq 1.$$

The Gateaux derivative  $\nabla_{\mathbf{H}}\Phi_q(\Sigma)$  of  $\Phi_q$  at  $\Sigma$  in direction  $\mathbf{H}$ , which is a symmetric matrix, is defined as

$$\nabla_{\mathbf{H}}\Phi_q(\Sigma) = \lim_{\varepsilon \rightarrow 0} \frac{\Phi_q(\Sigma + \varepsilon \mathbf{H}) - \Phi_q(\Sigma)}{\varepsilon} = \left. \frac{d}{d\varepsilon} \Phi_q(\Sigma + \varepsilon \mathbf{H}) \right|_{\varepsilon=0}. \quad (56)$$

If the limit specified in (56) exists for all symmetric matrices  $\mathbf{H}$ , we say that  $\Phi_q$  is Gateaux differentiable at  $\Sigma$ .

The next lemma provides the Gateaux derivative of  $\Phi_q$ . This lemma allows non-integer values for  $q$ , and is thus more general than a similar result in Yang et al. (2013).

**Lemma 13.5.**  *$\Phi_q$  is Gateaux differentiable at any real positive definite matrix  $\Sigma$  for any  $q \geq 0$ . Moreover, we have*

$$\nabla_{\mathbf{H}}\Phi_q(\Sigma) = \begin{cases} \text{tr}(\mathbf{H}\Sigma^{-1}), & \text{if } q = 0, \\ q \cdot \text{tr}(\Sigma^{q-1}\mathbf{H}), & \text{if } 0 < q < 1, \\ (\text{tr } \Sigma^q)^{1/q-1} \cdot \text{tr}(\Sigma^{q-1}\mathbf{H}), & \text{if } q \geq 1, \end{cases} \quad (57)$$

and

$$\frac{\partial \Phi_q(\mathcal{I}^{-\pi})}{\partial \pi(a)} = \begin{cases} -\text{tr}(\mathcal{I}^{-\pi}\mathcal{I}_a), & \text{if } q = 0, \\ -q \cdot \text{tr}\left((\mathcal{I}^{-\pi})^{q+1}\mathcal{I}_a\right), & \text{if } 0 < q < 1, \\ -\left[\text{tr}\left((\mathcal{I}^{-\pi})^q\right)\right]^{1/q-1} \cdot \text{tr}\left((\mathcal{I}^{-\pi})^{q+1}\mathcal{I}_a\right), & \text{if } q \geq 1, \end{cases} \quad (58)$$

where  $\mathcal{I}^\pi = \sum_{a \in \mathcal{A}} \pi(a)\mathcal{I}_a$  and  $\mathcal{I}^{-\pi} = \left\{ \sum_{a \in \mathcal{A}} \pi(a)\mathcal{I}_a \right\}^{-1}$ .

*Remark 13.6.* Based on (57), and the Riesz representation theorem over the Hilbert space of symmetric matrix, for any positive definite  $\Sigma$ , there exists unique symmetric matrix  $\nabla\Phi_q(\Sigma) = \left\{ \frac{\partial}{\partial \Sigma_{ij}} \Phi_q(\Sigma) \right\}_{1 \leq i, j \leq n}$ , such that for any symmetric matrix  $\mathbf{H}$  of comparable size,

$$\nabla_{\mathbf{H}}\Phi_q(\Sigma) = \langle \nabla\Phi_q(\Sigma), \mathbf{H} \rangle.$$

*Proof of Lemma 13.5.* Let  $q = 0$ . By the definition of the Gateaux derivative, for any

symmetric  $\mathbf{H}$  and positive definite matrix  $\mathbf{\Sigma}$ ,

$$\begin{aligned}\nabla_{\mathbf{H}}\Phi_0(\mathbf{\Sigma}) &= \lim_{\varepsilon \rightarrow 0} \frac{\Phi_0(\mathbf{\Sigma} + \varepsilon \mathbf{H}) - \Phi_0(\mathbf{\Sigma})}{\varepsilon} \\ &= \lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} \log |I + \varepsilon \mathbf{H} \mathbf{\Sigma}^{-1}| = \lim_{\varepsilon \rightarrow 0} \frac{\log(1 + \varepsilon \lambda_1) + \log(1 + \varepsilon \lambda_2) + \cdots + \log(1 + \varepsilon \lambda_n)}{\varepsilon} \\ &= (\lambda_1 + \lambda_2 + \cdots + \lambda_n) = \text{tr}(\mathbf{H} \mathbf{\Sigma}^{-1}).\end{aligned}$$

where  $\lambda_1, \lambda_2, \dots, \lambda_n$  denote all eigenvalues of  $\mathbf{H} \mathbf{\Sigma}^{-1}$  counting multiplicity.

When  $q$  is a positive integer, by expanding  $(\mathbf{\Sigma} + \varepsilon \mathbf{H})^q$ , we have

$$\text{tr}((\mathbf{\Sigma} + \varepsilon \mathbf{H})^q) - \text{tr}(\mathbf{\Sigma}^q) = \varepsilon \cdot q \cdot \text{tr}(\mathbf{\Sigma}^{q-1} \mathbf{H}) + o(\varepsilon).$$

Note that when  $q$  is a positive integer

$$\begin{aligned}& \frac{\Phi_q(\mathbf{\Sigma} + \varepsilon \mathbf{H}) - \Phi_q(\mathbf{\Sigma})}{\varepsilon} \\ &= \frac{1}{\varepsilon} (\text{tr} \mathbf{\Sigma}^q)^{1/q} \left[ \left( 1 + (\text{tr}(\mathbf{\Sigma} + \varepsilon \mathbf{H})^q - \text{tr}(\mathbf{\Sigma}^q)) / \text{tr}(\mathbf{\Sigma}^q) \right)^{1/q} - 1 \right] \\ &= \frac{1}{\varepsilon \cdot q} (\text{tr}(\mathbf{\Sigma}^q))^{1/q-1} \cdot \left[ \text{tr}((\mathbf{\Sigma} + \varepsilon \mathbf{H})^q) - \text{tr}(\mathbf{\Sigma}^q) \right] + o(1) \\ &= (\text{tr}(\mathbf{\Sigma}^q))^{1/q-1} \cdot \text{tr}(\mathbf{\Sigma}^{q-1} \mathbf{H}) + o(1).\end{aligned}\tag{59}$$

Thus, when  $q$  is a positive integer, (57) holds.

Now, consider the case when  $q > 0$  and  $q$  is not an integer. Because we can not expand  $(\mathbf{\Sigma} + \varepsilon \mathbf{H})^q$  and due to the lack of commutative between  $\mathbf{\Sigma}$  and  $\mathbf{H}$ , we need some more complicated techniques. Assume  $\mathbf{\Sigma}$  is of size  $n \times n$ . Let  $\lambda_1(\varepsilon) \geq \lambda_2(\varepsilon) \geq \cdots \geq \lambda_n(\varepsilon)$  be all the eigenvalues of  $\mathbf{\Sigma} + \varepsilon \mathbf{H}$ . Denote the corresponding eigenvectors by  $\mathbf{u}_1(\varepsilon), \dots, \mathbf{u}_n(\varepsilon)$ .

Let  $\lambda_i = \lambda_i(0)$ , and  $\mathbf{u}_i = \mathbf{u}_i(0)$  for  $1 \leq i \leq n$ . Set  $\lambda_0 = -\infty, \lambda_{n+1} = \infty$ .

Let  $\lambda$  be an eigenvalue of  $\mathbf{\Sigma} + \varepsilon \mathbf{H}$ , there exists  $0 \leq r < s \leq n+1$  such that

$$\lambda_{r-1} > \lambda = \lambda_r = \cdots = \lambda_s > \lambda_{s+1}.$$

Let  $d = s - r + 1$ ,  $\mathbf{U}_\lambda = [\mathbf{u}_r, \mathbf{u}_{r+1}, \dots, \mathbf{u}_s]$  and  $\mathbf{U}_\lambda(\varepsilon) = [\mathbf{u}_r(\varepsilon), \mathbf{u}_{r+1}(\varepsilon), \dots, \mathbf{u}_s(\varepsilon)]$ .

By Wely's inequality,  $|\lambda_i(\varepsilon) - \lambda| \leq |\varepsilon| \|\mathbf{H}\|_{op}$ .

Notice that when  $|\varepsilon|$  is small enough, we have

$$\begin{aligned}
& \text{tr}(\mathbf{U}_\lambda^T(\varepsilon)(\mathbf{\Sigma} + \varepsilon \mathbf{H})^q \mathbf{U}_\lambda(\varepsilon)) - \text{tr}(\mathbf{U}_\lambda^T \mathbf{\Sigma}^q \mathbf{U}_\lambda) \\
&= (\lambda_r^q(\varepsilon) - \lambda^q) + \cdots + (\lambda_s^q(\varepsilon) - \lambda^q) \\
&= \lambda^q \left[ \left( \left( 1 + \frac{\lambda_r(\varepsilon) - \lambda}{\lambda} \right)^q - 1 \right) + \cdots + \left( \left( 1 + \frac{\lambda_s(\varepsilon) - \lambda}{\lambda} \right)^q - 1 \right) \right] \\
&= q \lambda^{q-1} ((\lambda_r(\varepsilon) - \lambda) + \cdots + (\lambda_s(\varepsilon) - \lambda)) + o(\varepsilon),
\end{aligned} \tag{60}$$

and

$$\begin{aligned}
& ((\lambda_r(\varepsilon) - \lambda) + \cdots + (\lambda_s(\varepsilon) - \lambda)) \\
&= \text{tr}(\mathbf{U}_\lambda^T(\varepsilon)(\mathbf{\Sigma} + \varepsilon \mathbf{H}) \mathbf{U}_\lambda(\varepsilon)) - \text{tr}(\mathbf{U}_\lambda^T \mathbf{\Sigma} \mathbf{U}_\lambda) \\
&= \varepsilon \cdot \text{tr}(\mathbf{U}_\lambda^T(\varepsilon) \mathbf{H} \mathbf{U}_\lambda(\varepsilon)) + \text{tr}(\mathbf{U}_\lambda^T(\varepsilon) \mathbf{\Sigma} \mathbf{U}_\lambda(\varepsilon)) - \text{tr}(\mathbf{U}_\lambda^T \mathbf{\Sigma} \mathbf{U}_\lambda).
\end{aligned} \tag{61}$$

To proceed, we first prove the following equation,

$$\text{tr}((\mathbf{\Sigma} - \lambda I)(\mathbf{U}_\lambda(\varepsilon) - \mathbf{U}_\lambda)(\mathbf{U}_\lambda(\varepsilon) - \mathbf{U}_\lambda)^T) = \text{tr}(\mathbf{U}_\lambda^T(\varepsilon) \mathbf{\Sigma} \mathbf{U}_\lambda(\varepsilon)) - \text{tr}(\mathbf{U}_\lambda^T \mathbf{\Sigma} \mathbf{U}_\lambda). \tag{62}$$

(62) is justified by the following matrix calculation,

$$\begin{aligned}
& \text{tr}((\mathbf{\Sigma} - \lambda I)(\mathbf{U}_\lambda(\varepsilon) - \mathbf{U}_\lambda)(\mathbf{U}_\lambda(\varepsilon) - \mathbf{U}_\lambda)^T) - [\text{tr}(\mathbf{U}_\lambda^T(\varepsilon) \mathbf{\Sigma} \mathbf{U}_\lambda(\varepsilon)) - \text{tr}(\mathbf{U}_\lambda^T \mathbf{\Sigma} \mathbf{U}_\lambda)] \\
&= \text{tr}((\mathbf{U}_\lambda(\varepsilon) - \mathbf{U}_\lambda)^T (\mathbf{\Sigma} - \lambda I)(\mathbf{U}_\lambda(\varepsilon) - \mathbf{U}_\lambda)) - [\text{tr}(\mathbf{U}_\lambda^T(\varepsilon) (\mathbf{\Sigma} - \lambda I) \mathbf{U}_\lambda(\varepsilon)) - \text{tr}(\mathbf{U}_\lambda^T (\mathbf{\Sigma} - \lambda I) \mathbf{U}_\lambda)] \\
&= -2 \text{tr}(\mathbf{U}_\lambda^T(\varepsilon) (\mathbf{\Sigma} - \lambda I) \mathbf{U}_\lambda) - 2 \text{tr}(\mathbf{U}_\lambda^T (\mathbf{\Sigma} - \lambda I) \mathbf{U}_\lambda) = 0, \text{ because } (\mathbf{\Sigma} - \lambda I) \mathbf{U}_\lambda = 0.
\end{aligned}$$

Note that  $\mathbf{\Sigma} - \lambda I = (I - \mathbf{U}_\lambda \mathbf{U}_\lambda^T)(\mathbf{\Sigma} - \lambda I)(I - \mathbf{U}_\lambda \mathbf{U}_\lambda^T)$ ,  $-\|\mathbf{\Sigma} - \lambda I\|_{op} I \preceq \mathbf{\Sigma} - \lambda I \preceq \|\mathbf{\Sigma} - \lambda I\|_{op} I$ , and  $(\mathbf{U}_\lambda(\varepsilon) - \mathbf{U}_\lambda)(\mathbf{U}_\lambda(\varepsilon) - \mathbf{U}_\lambda)^T$  is positive semidefinite, we obtain

$$\begin{aligned}
& |\text{tr}((\mathbf{\Sigma} - \lambda I)(\mathbf{U}_\lambda(\varepsilon) - \mathbf{U}_\lambda)(\mathbf{U}_\lambda(\varepsilon) - \mathbf{U}_\lambda)^T)| \\
&\leq \|\mathbf{\Sigma} - \lambda I\|_{op} \cdot \text{tr}((I - \mathbf{U}_\lambda \mathbf{U}_\lambda^T)(\mathbf{U}_\lambda(\varepsilon) - \mathbf{U}_\lambda)(\mathbf{U}_\lambda(\varepsilon) - \mathbf{U}_\lambda)^T(I - \mathbf{U}_\lambda \mathbf{U}_\lambda^T)) \\
&= \|\mathbf{\Sigma} - \lambda I\|_{op} \cdot \text{tr}((I - \mathbf{U}_\lambda \mathbf{U}_\lambda^T) \mathbf{U}_\lambda(\varepsilon) \mathbf{U}_\lambda^T(\varepsilon)) \\
&= \|\mathbf{\Sigma} - \lambda I\|_{op} \cdot (d - \|\mathbf{U}_\lambda^T \mathbf{U}_\lambda(\varepsilon)\|_F^2) \\
&= \|\mathbf{\Sigma} - \lambda I\|_{op} \cdot (d - \|\cos \mathbf{\Theta}(\mathbf{U}_\lambda, \mathbf{U}_\lambda(\varepsilon))\|_F^2) \\
&= \|\mathbf{\Sigma} - \lambda I\|_{op} \cdot \|\sin \mathbf{\Theta}(\mathbf{U}_\lambda, \mathbf{U}_\lambda(\varepsilon))\|_F^2,
\end{aligned} \tag{63}$$

where  $\|\cos \mathbf{\Theta}(\mathbf{U}_\lambda, \mathbf{U}_\lambda(\varepsilon))\|_F = \|\mathbf{U}_\lambda^T \mathbf{U}_\lambda(\varepsilon)\|_F$  is due to the definition of principal angles between column spaces of  $\mathbf{U}_\lambda$  and  $\mathbf{U}_\lambda(\varepsilon)$  (see Yu et al. (2015)).

By Davis-Kahan theorem (see Yu et al. (2015)), we obtain

$$\|\sin \Theta(\mathbf{U}_\lambda, \mathbf{U}_\lambda(\varepsilon))\|_F \leq \frac{2\|(\Sigma + \varepsilon \mathbf{H}) - \Sigma\|_F}{\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})} = O(\varepsilon). \quad (64)$$

Combining (60),(61),(62),(63) and (64), we obtain

$$\begin{aligned} ((\lambda_r(\varepsilon) - \lambda) + \dots + (\lambda_s(\varepsilon) - \lambda)) &= \varepsilon \cdot \text{tr}(\mathbf{U}_\lambda^T(\varepsilon) \mathbf{H} \mathbf{U}_\lambda(\varepsilon)) + O(\varepsilon^2), \text{ and} \\ \text{tr}(\mathbf{U}_\lambda^T(\varepsilon)(\Sigma + \varepsilon \mathbf{H})^q \mathbf{U}_\lambda(\varepsilon)) - \text{tr}(\mathbf{U}_\lambda^T \Sigma^q \mathbf{U}_\lambda) &= \varepsilon \cdot q \lambda^{q-1} \text{tr}(\mathbf{U}_\lambda^T(\varepsilon) \mathbf{H} \mathbf{U}_\lambda(\varepsilon)) + o(\varepsilon). \end{aligned}$$

Let  $\mathbf{P}_\varepsilon = \mathbf{U}_\lambda(\varepsilon) \mathbf{U}_\lambda^T(\varepsilon)$  and  $\mathbf{P} = \mathbf{U}_\lambda \mathbf{U}_\lambda^T$ , we know that

$$\begin{aligned} \|\mathbf{P}_\varepsilon - \mathbf{P}\|_F^2 &= \text{tr}(\mathbf{P}_\varepsilon - 2\mathbf{P}_\varepsilon \mathbf{P} + \mathbf{P}) = 2(d - \text{tr}(\mathbf{P}_\varepsilon \mathbf{P})) = 2(d - \|\mathbf{U}_\lambda^T \mathbf{U}_\lambda(\varepsilon)\|_F^2) \\ &= 2(d - \|\cos \Theta(\mathbf{U}_\lambda, \mathbf{U}_\lambda(\varepsilon))\|_F^2) = 2 \|\sin \Theta(\mathbf{U}_\lambda, \mathbf{U}_\lambda(\varepsilon))\|_F^2, \end{aligned} \quad (65)$$

and  $\|\mathbf{P}_\varepsilon - \mathbf{P}\|_F = \sqrt{2} \|\sin \Theta(\mathbf{U}_\lambda, \mathbf{U}_\lambda(\varepsilon))\|_F = O(\varepsilon)$ . Due to  $\Sigma \mathbf{U}_\lambda = \lambda \mathbf{U}_\lambda$ , we know that  $\Sigma^{q-1} \mathbf{U}_\lambda = \lambda^{q-1} \mathbf{U}_\lambda$ , which means that

$$\begin{aligned} ((\lambda_r(\varepsilon) - \lambda) + \dots + (\lambda_s(\varepsilon) - \lambda)) &= \varepsilon \cdot q \lambda^{q-1} \text{tr}(\mathbf{H} \mathbf{U}_\lambda(\varepsilon) \mathbf{U}_\lambda^T(\varepsilon)) + o(\varepsilon) \\ = \varepsilon \cdot q \lambda^{q-1} \text{tr}(\mathbf{H} \mathbf{U}_\lambda \mathbf{U}_\lambda^T) + o(\varepsilon) &= \varepsilon \cdot q \lambda^{q-1} \text{tr}(\mathbf{U}_\lambda^T \mathbf{H} \mathbf{U}_\lambda) + o(\varepsilon) = \varepsilon \cdot q \cdot \text{tr}(\Sigma^{q-1} \mathbf{H} \mathbf{U}_\lambda \mathbf{U}_\lambda^T) + o(\varepsilon). \end{aligned}$$

This leads to

$$\begin{aligned} \text{tr}((\Sigma + \varepsilon \mathbf{H})^q) - \text{tr}(\Sigma^q) &= \sum_\lambda \text{tr}(\mathbf{U}_\lambda^T(\varepsilon)(\Sigma + \varepsilon \mathbf{H})^q \mathbf{U}_\lambda(\varepsilon)) - \text{tr}(\mathbf{U}_\lambda^T \Sigma^q \mathbf{U}_\lambda) \\ &= \sum_\lambda \varepsilon \cdot q \cdot \text{tr}(\Sigma^{q-1} \mathbf{H} \mathbf{U}_\lambda \mathbf{U}_\lambda^T) + o(\varepsilon) = \varepsilon \cdot q \cdot \text{tr}(\Sigma^{q-1} \mathbf{H}) + o(\varepsilon), \end{aligned} \quad (66)$$

where the last equation holds due to  $\sum_\lambda \mathbf{U}_\lambda \mathbf{U}_\lambda^T = \mathbf{I}$ .

Thus, when  $0 < q < 1$ , we know that

$$\nabla_{\mathbf{H}} \Phi_q(\Sigma) = q \cdot \text{tr}(\Sigma^{q-1} \mathbf{H}).$$

Combining the first two equations in (59) with the equation (66), if  $q \geq 1$ , we know that

$$\nabla_{\mathbf{H}} \Phi_q(\Sigma) = (\text{tr}(\Sigma^q))^{1/q-1} \cdot \text{tr}(\Sigma^{q-1} \mathbf{H}).$$

Applying (59) again, we complete the proof of (57) in Lemma 13.5.

By applying the chain rule and combining (55) in Lemma 13.4 with (57), we obtain that

$$\frac{\partial \Phi_q(\mathcal{I}^{-\pi})}{\partial \pi(a)} = \left\langle \nabla \Phi_q(\mathcal{I}^{-\pi}), \frac{\partial \mathcal{I}^{-\pi}}{\partial \pi(a)} \right\rangle = \langle \nabla \Phi_q(\mathcal{I}^{-\pi}), -\mathcal{I}^{-\pi} \mathcal{I}_a \mathcal{I}^{-\pi} \rangle = \nabla_{-\mathcal{I}^{-\pi} \mathcal{I}_a \mathcal{I}^{-\pi}} \Phi_q(\mathcal{I}^{-\pi}),$$

which completes the proof of (58) in Lemma 13.5.  $\square$

**Lemma 13.7.** *Assume  $\mathcal{I}_a, a \in \mathcal{S}^{\mathcal{A}}$  are positive semidefinite matrices. Consider a convex matrix function,  $g$ , such that for any pair of positive semidefinite matrices,  $\mathbf{A}$  and  $\mathbf{B}$ , if  $\mathbf{A} - \mathbf{B}$  is a positive semidefinite matrix (denoted as  $\mathbf{A} \succeq \mathbf{B}$ ), then  $g(\mathbf{A}) \geq g(\mathbf{B})$ . For any  $\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}$ , define*

$$F(\boldsymbol{\pi}) = g \left( \left\{ \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a \right\}^{-1} \right). \quad (67)$$

*Then,  $F(\boldsymbol{\pi})$  is a convex function over the set  $\{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}; \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a \text{ is nonsingular}\}$ .*

*In particular, under Assumption 5, the function  $\mathbb{F}_{\boldsymbol{\theta}}(\boldsymbol{\pi})$  is a convex function over the set  $\{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}; \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}) \text{ is nonsingular}\}$ .*

*Proof of Lemma 13.7.* Let  $\mathcal{C}_0 = \{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}; \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a \text{ is nonsingular}\}$ . Assume that  $\boldsymbol{\pi}, \boldsymbol{\pi}' \in \mathcal{C}_0$ . Then,  $\mathbf{A} = \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a$  and  $\mathbf{B} = \sum_{a \in \mathcal{A}} \pi'(a) \mathcal{I}_a$  are positive definite.

For any  $0 < t < 1$ ,  $t\mathbf{A} + (1-t)\mathbf{B}$  is positive definite. Note that for any vector  $v$ , applying Schur complement condition (see Theorem 1.12 (b) in Zhang (2006)), we obtain that

$$\begin{bmatrix} v^T \mathbf{A}^{-1} v & v^T \\ v & \mathbf{A} \end{bmatrix} \text{ and } \begin{bmatrix} v^T \mathbf{B}^{-1} v & v^T \\ v & \mathbf{B} \end{bmatrix}$$

are positive semi-definite matrices. Notice that the following matrix is positive semi-definite

$$t \begin{bmatrix} v^T \mathbf{A}^{-1} v & v^T \\ v & \mathbf{A} \end{bmatrix} + (1-t) \begin{bmatrix} v^T \mathbf{B}^{-1} v & v^T \\ v & \mathbf{B} \end{bmatrix} = \begin{bmatrix} tv^T \mathbf{A}^{-1} v + (1-t)v^T \mathbf{B}^{-1} v & v^T \\ v & t\mathbf{A} + (1-t)\mathbf{B} \end{bmatrix},$$

by Theorem 1.12 (b) in Zhang (2006), we obtain that

$$tv^T \mathbf{A}^{-1} v + (1-t)v^T \mathbf{B}^{-1} v \geq v^T (t\mathbf{A} + (1-t)\mathbf{B})^{-1} v.$$

Since  $v$  is arbitrary, we have

$$t\mathbf{A}^{-1} + (1-t)\mathbf{B}^{-1} \succeq (t\mathbf{A} + (1-t)\mathbf{B})^{-1}. \quad (68)$$

Now, we have

$$\begin{aligned} tF(\boldsymbol{\pi}) + (1-t)F(\boldsymbol{\pi}') &= tg(\mathbf{A}^{-1}) + (1-t)g(\mathbf{B}^{-1}) \\ &\geq g(t\mathbf{A}^{-1} + (1-t)\mathbf{B}^{-1}) \geq g(\{t\mathbf{A} + (1-t)\mathbf{B}\}^{-1}) = F(t\boldsymbol{\pi} + (1-t)\boldsymbol{\pi}'). \end{aligned}$$

This shows that  $F$  is convex.

We proceed to the proof of the ‘In particular’ part of the lemma. Note that under Assumption 5, there are two cases: Case 1:  $\mathbb{G}_{\boldsymbol{\theta}}(\boldsymbol{\Sigma}) = \Phi_q(\boldsymbol{\Sigma})$  for  $q \geq 0$ , and Case 2:  $\mathbb{G}_{\boldsymbol{\theta}}(\cdot)$  is a convex function satisfying  $\mathbb{G}_{\boldsymbol{\theta}}(\mathbf{A}) \geq \mathbb{G}_{\boldsymbol{\theta}}(\mathbf{B})$  whenever  $\mathbf{A} \succeq \mathbf{B}$ . For Case 2, we can apply our previous analysis directly for  $g(\cdot) = \mathbb{G}_{\boldsymbol{\theta}}(\cdot)$ , and obtain that  $\mathbb{F}_{\boldsymbol{\theta}}(\boldsymbol{\pi})$  is convex. Thus, we focus our analysis on Case 1 in the rest of the proof. By Courant-Fischer-Wely minimax principle (see Corollary III.1.2 in Bhatia (1997)), the  $i$ -th largest eigenvalue satisfies  $\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B})$  for any  $1 \leq i \leq n$ . Thus,  $\Phi_q(\mathbf{A}) \geq \Phi_q(\mathbf{B})$  for any  $q \geq 0$ .

If  $\mathbb{G}_{\boldsymbol{\theta}}(\boldsymbol{\Sigma}) = \Phi_q(\boldsymbol{\Sigma}) = (\text{tr}(\boldsymbol{\Sigma}^q))^{1/q}$  with  $q \geq 1$ , then  $\Phi_q(\boldsymbol{\Sigma})$  is the Schatten  $q$ -norm (see equation (IV.31) in Bhatia (1997)), which implies that  $\mathbb{G}_{\boldsymbol{\theta}}(\boldsymbol{\Sigma})$  is convex. More generally, if  $\mathbb{G}_{\boldsymbol{\theta}}(\boldsymbol{\Sigma})$  is convex in  $\boldsymbol{\Sigma}$ , then by (67), we obtain that

$$\mathbb{F}_{\boldsymbol{\theta}}(\boldsymbol{\pi}) = \mathbb{G}_{\boldsymbol{\theta}}(\{\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta})\}^{-1})$$

is convex in  $\boldsymbol{\pi}$  over

$$\{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}; \sum_{a \in \mathcal{A}} \pi(a)\mathcal{I}_a(\boldsymbol{\theta}) \text{ is nonsingular}\}.$$

If  $\mathbb{G}_{\boldsymbol{\theta}}(\boldsymbol{\Sigma}) = \Phi_0(\boldsymbol{\Sigma}) = \log \det \boldsymbol{\Sigma}$ , we know that

$$\Phi_0(\{\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta})\}^{-1}) = -\log \det \left( \sum_{a \in \mathcal{A}} \pi(a)\mathcal{I}_a(\boldsymbol{\theta}) \right).$$

We aim to show that  $-\log \det(\mathbf{A})$  is convex over positive definite matrices.

Notice that for any  $p \times p$  positive definite matrix  $\mathbf{A}$ ,

$$\int_{\mathbb{R}^p} e^{-1/2\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle} d\mathbf{x} = \frac{1}{(2\pi)^{p/2} \det(\mathbf{A})^{1/2}}.$$

By Hölder’s inequality, for any positive definite  $p \times p$  matrices  $\mathbf{A}$  and  $\mathbf{B}$ , we have

$$\int_{\mathbb{R}^p} e^{-1/2\langle (t\mathbf{A} + (1-t)\mathbf{B})\mathbf{x}, \mathbf{x} \rangle} d\mathbf{x} \leq \left( \int_{\mathbb{R}^p} e^{-1/2\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle} d\mathbf{x} \right)^t \left( \int_{\mathbb{R}^p} e^{-1/2\langle \mathbf{B}\mathbf{x}, \mathbf{x} \rangle} d\mathbf{x} \right)^{1-t}, \quad (69)$$

which implies that

$$-\log \det(t\mathbf{A} + (1-t)\mathbf{B}) \leq -t \log \det(\mathbf{A}) - (1-t) \log \det(\mathbf{B}).$$

This shows that  $-\log \det(\mathbf{A})$  is convex over positive definite matrices.

Thus,  $\mathbb{F}_\theta(\boldsymbol{\pi}) = -\log \det(\sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}))$  is convex in  $\boldsymbol{\pi}$  over

$$\{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}; \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}) \text{ is nonsingular}\}.$$

If  $\mathbb{G}_\theta(\boldsymbol{\Sigma}) = \Phi_q(\boldsymbol{\Sigma}) = \text{tr}(\boldsymbol{\Sigma}^q)$  with  $0 < q < 1$ , we know that

$$\Phi_q(\{\mathcal{I}^\pi(\boldsymbol{\theta})\}^{-1}) = \text{tr} \left( \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}) \right)^{-q}.$$

By Löwner-Heinz Theorem (see Theorem 2.6 in Carlen (2010)), we know that  $\text{tr}(\mathbf{A}^{-q})$  is operator convex, which means that for all positive definite matrices  $\mathbf{A}$  and  $\mathbf{B}$ ,

$$\left( t\mathbf{A} + (1-t)\mathbf{B} \right)^{-q} \preceq t\mathbf{A}^{-q} + (1-t)\mathbf{B}^{-q},$$

which implies that  $\text{tr}(\mathbf{A}^{-q})$  is a convex function over positive definite matrices. In conclusion, we obtain that  $\mathbb{F}_\theta(\boldsymbol{\pi}) = \Phi_q(\{\mathcal{I}^\pi(\boldsymbol{\theta})\}^{-1})$  is convex in  $\boldsymbol{\pi}$  over

$$\{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}; \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}) \text{ is nonsingular}\}.$$

□

### 13.3 Decoupling Active Sequential Sampling

The next lemma provides a decoupling result which makes it easier to analyze the likelihood for problems with adaptive experiment selection.

**Lemma 13.8.** *Consider deterministic sequential selection functions  $h_m(\cdot)$ ,*

$$h_m(a_1, Y_1, a_2, Y_2, \dots, a_{m-1}, Y_{m-1}) \in \mathcal{A}, \text{ for } m \geq 2,$$

*and  $h_1 \in \mathcal{A}$ . Consider two random vectors generated from the following procedures.*

1. (Decoupled Sampling) *Independently generate  $\{X_m^a\}_{m \geq 1, a \in \mathcal{A}}$ , where  $X_m^a \sim f_{\boldsymbol{\theta}^*, a}(\cdot)$ . Let*

$$a_1 = h_1, a_2 = h_2(a_1, X_1^{a_1}), \dots, a_{m+1} = h_{m+1}(a_1, X_1^{a_1}, a_2, X_2^{a_2}, \dots, a_m, X_m^{a_m}), \dots$$



2. (Iterative Sampling) Let  $a'_1 = h_1$ . Generate  $X_1 \sim f_{\theta^*, a'_1}(\cdot)$ . For  $m \geq 1$ , obtain  $a'_{m+1}$  and  $X_{m+1}$  iteratively as

$$a'_{m+1} = h_{m+1}(a'_1, X_1, \dots, a'_m, X_m),$$

then generate  $X_{m+1} | \mathcal{F}_m \sim f_{\theta, a'_{m+1}}$ , where the  $\sigma$ -algebra  $\mathcal{F}_m = \sigma(a'_1, X_1, a'_2, X_2, \dots, a'_m, X_m)$ .

Then, the random vectors  $(X_1^{a_1}, \dots, X_m^{a_m}, a_1, \dots, a_m)$  and  $(X_1, \dots, X_m, a'_1, \dots, a'_m)$  have the same distribution for all  $m \geq 1$ .

*Proof of Lemma 13.8.* We prove the lemma by induction. When  $m = 1$ , we know that  $a_1 = h_1 = a'_1$ , and  $X_1 | a'_1$  and  $X_1^{a_1} | a_1$  have the same distribution. Thus,  $(X_1^{a_1}, a_1)$  and  $(X_1, a'_1)$  have the same distribution.

By induction, assume that when  $m = n$ , random vectors  $(X_1^{a_1}, X_2^{a_2}, \dots, X_n^{a_n}, a_1, \dots, a_n)$  and  $(X_1, X_2, \dots, X_n, a'_1, \dots, a'_n)$  have the same distribution.

Let  $m = n + 1$ . Define  $\mathbf{X}_n^{\mathcal{A}} = \{X_i^a\}_{1 \leq i \leq n, a \in \mathcal{A}}$ , and  $\mathbf{a}^n = (a^1, \dots, a^n) \in \mathcal{A}^n$ . The density of  $\mathbf{X}_n^{\mathcal{A}}$  is given by

$$f_{\theta}(\mathbf{X}_n^{\mathcal{A}}) = \prod_{i=1}^n \prod_{a \in \mathcal{A}} f_{\theta, a}(X_i^a).$$

Let  $\mathbf{a}_m = (a_1, \dots, a_m)$ , where  $a_1, \dots, a_m$  are obtained from the decoupled sampling. Given  $\mathbf{X}_n^{\mathcal{A}}$ , the conditional probability mass function of  $\mathbf{a}_n$  and  $\mathbf{a}_{n+1}$  are

$$f_{\theta}(\mathbf{a}^n | \mathbf{X}_n^{\mathcal{A}}) = I(a^1 = a_1, \dots, a^n = a_n) \text{ and } f_{\theta}(\mathbf{a}^{n+1} | \mathbf{X}_n^{\mathcal{A}}) = I(a^1 = a_1, \dots, a^{n+1} = a_{n+1}),$$

where we used the fact that the  $\{a_m\}_{1 \leq m \leq n+1}$  is measurable with respect to  $\sigma(\mathbf{X}_n^{\mathcal{A}})$ . As a result, the joint density functions for  $(\mathbf{X}_n^{\mathcal{A}}, \mathbf{a}_n)$  and  $(\mathbf{X}_{n+1}^{\mathcal{A}}, \mathbf{a}_{n+1})$  are

$$f_{\theta}(\mathbf{X}_n^{\mathcal{A}}, \mathbf{a}^n) = \prod_{i=1}^n \prod_{a \in \mathcal{A}} f_{\theta, a}(X_i^a) I(a^1 = a_1, \dots, a^n = a_n)$$

and

$$f_{\theta}(\mathbf{X}_{n+1}^{\mathcal{A}}, \mathbf{a}^{n+1}) = f_{\theta}(\mathbf{X}_n^{\mathcal{A}}, \mathbf{a}^n) \prod_{a \in \mathcal{A}} f_{\theta, a}(X_{n+1}^a) I(a^{n+1} = a_{n+1}).$$

Thus, given  $\mathbf{X}_n^{\mathcal{A}}$  and  $\mathbf{a}_n$ , the condition density for  $\{X_{n+1}^a\}_{a \in \mathcal{A}}, a_{n+1}$  is

$$f_{\theta}(\{X_{n+1}^a\}_{a \in \mathcal{A}}, a^{n+1} | \mathbf{X}_n^{\mathcal{A}}, \mathbf{a}^n) = \prod_{a \in \mathcal{A}} f_{\theta, a}(X_{n+1}^a) I(a^{n+1} = a_{n+1}).$$

Note that  $a_{n+1} = h_{n+1}(a_1, X_1^{a_1}, a_2, X_2^{a_2}, \dots, a_n, X_n^{a_n})$ . So  $f_{\theta}(\{X_{n+1}^a\}_{a \in \mathcal{A}}, a^{n+1} | \mathbf{X}_n^{\mathcal{A}}, \mathbf{a}^n)$  de-

depends on  $\mathbf{X}_n^{\mathcal{A}}, \mathbf{a}^n$  only through  $a_1, X_1^{a_1}, \dots, a_n, X_n^{a_n}$ .

Define  $\sigma$ -algebra  $\mathcal{F}'_n = \sigma(a_1, X_1^{a_1}, \dots, a_n, X_n^{a_n})$ . Because  $f_{\boldsymbol{\theta}}(\{X_{n+1}^a\}_{a \in \mathcal{A}}, a^{n+1} | \mathbf{X}_n^{\mathcal{A}}, \mathbf{a}^n)$  is measurable in  $\mathcal{F}'_n$ , we have

$$f_{\boldsymbol{\theta}}(\{X_{n+1}^a\}_{a \in \mathcal{A}}, a^{n+1} | \mathcal{F}'_n) = f_{\boldsymbol{\theta}}(\{X_{n+1}^a\}_{a \in \mathcal{A}}, a^{n+1} | \mathbf{X}_n^{\mathcal{A}}, \mathbf{a}_n) = \prod_{a \in \mathcal{A}} f_{\boldsymbol{\theta}, a}(X_{n+1}^a) I(a^{n+1} = a_{n+1}) \quad (70)$$

We have  $X_{n+1}^{a_{n+1}} | \{a_1, X_1^{a_1}, \dots, a_n, X_n^{a_n}\} \sim f_{\boldsymbol{\theta}, a_{n+1}}(\cdot)$ , and  $a_{n+1} = h_n(a_1, X_1^{a_1}, \dots, a_n, X_n^{a_n})$ . Recall that  $X_{n+1} | \{a'_1, X_1, \dots, a'_n, X_n\} \sim f_{\boldsymbol{\theta}, a'_{n+1}}(\cdot)$ ,  $a'_{n+1} = h_n(a'_1, X_1, \dots, a'_n, X_n)$ , as well as the induction assumption that random vectors  $(X_1^{a_1}, X_2^{a_2}, \dots, X_n^{a_n}, a_1, \dots, a_n)$  and  $(X_1, X_2, \dots, X_n, a'_1, \dots, a'_n)$  have the same distribution. Consequently, random vectors  $(X_1^{a_1}, X_2^{a_2}, \dots, X_{n+1}^{a_{n+1}}, a_1, \dots, a_{n+1})$  and  $(X_1, X_2, \dots, X_{n+1}, a'_1, \dots, a'_{n+1})$  also have the same distribution. We complete the proof of Lemma 13.8 by induction.  $\square$

### 13.4 Results on Linear Spaces Indexed by a Parameter

Linear spaces spanned by the Fisher information play a crucial role in the proof of the theorems. Note that the Fisher information matrices are depending on the parameter  $\boldsymbol{\theta}$ . In this section, we present useful linear algebra results where the linear spaces are indexed by a parameter.

Recall  $V_Q(\boldsymbol{\theta}) = \sum_{a \in Q} \mathcal{R}(\mathcal{I}_a(\boldsymbol{\theta}))$ , and  $\mathcal{I}_a(\boldsymbol{\theta})$  is the Fisher information matrix at the parameter  $\boldsymbol{\theta}$  with the experiment  $a$ . Throughout the section, we only used the property that  $\mathcal{I}_a(\boldsymbol{\theta})$  is a positive semidefinite matrix and is continuous in  $\boldsymbol{\theta}$ , for all  $a \in \mathcal{A}$ , which is guaranteed under the regularity assumptions in Section 4.1. The results in this section still hold even when  $\mathcal{I}_a(\boldsymbol{\theta})$  is not the Fisher information matrix, as long as it is still positive semidefinite and continuous in  $\boldsymbol{\theta}$ , for all  $a \in \mathcal{A}$ . We do not require any further assumptions.

**Lemma 13.9.** *For all  $Q \subset \mathcal{A}$ , and  $x_a > 0, a \in Q$ , we have*

$$\dim(V_Q(\boldsymbol{\theta})) = \text{rank} \left( \sum_{a \in Q} x_a \mathcal{I}_a(\boldsymbol{\theta}) \right).$$

*Proof of Lemma 13.9.* It suffices to show that  $V_Q(\boldsymbol{\theta})^\perp = \ker(\sum_{a \in Q} x_a \mathcal{I}_a(\boldsymbol{\theta}))$ . This equation holds because  $\mathbf{u} \in V_Q(\boldsymbol{\theta})^\perp$  if and only if  $\langle \mathcal{I}_a(\boldsymbol{\theta}) \mathbf{y}_a, \mathbf{u} \rangle = 0$  for all  $a \in Q$  and  $\mathbf{y}_a \in \mathbb{R}^p$ , if and only if  $\mathcal{I}_a(\boldsymbol{\theta}) \mathbf{u} = 0$  for all  $a \in Q$ , if and only if  $\mathbf{u}^T \mathcal{I}_a(\boldsymbol{\theta}) \mathbf{u} = 0$  for all  $a \in Q$ , if and only if  $\mathbf{u}^T (\sum_{a \in Q} x_a \mathcal{I}_a(\boldsymbol{\theta})) \mathbf{u} = 0$ , if and only if  $\mathbf{u} \in \ker(\sum_{a \in Q} x_a \mathcal{I}_a(\boldsymbol{\theta}))$ .  $\square$

**Lemma 13.10.** *Assume  $\Theta$  is a path connected and compact set. The following statements are equivalent:*

1. for all  $Q \subset \mathcal{A}$ ,  $\dim(V_Q(\boldsymbol{\theta}))$  does not depend on  $\boldsymbol{\theta}$ ,
2. for all  $Q \subset \mathcal{A}$ ,  $\text{rank}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}))$  does not depend on  $\boldsymbol{\theta}$ ,
3. there exists  $0 < \underline{c} < \bar{c} < \infty$ , which does not depend on  $Q$ , such that

$$\underline{c} \cdot \mathbf{P}_{V_Q(\boldsymbol{\theta})} \preceq \sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}) \preceq \bar{c} \cdot \mathbf{P}_{V_Q(\boldsymbol{\theta})}, \forall Q \subset \mathcal{A},$$

where  $\mathbf{P}_V$  denotes the orthogonal projection matrix onto vector space  $V$ .

*Proof of Lemma 13.10.*

**1  $\iff$  2** This equivalency holds because  $\dim(V_Q(\boldsymbol{\theta})) = \text{rank}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}))$ , according to Lemma 13.9.

**3  $\implies$  2** For  $Q \subset \mathcal{A}$ , let  $r(\boldsymbol{\theta}) = \text{rank}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}))$ . Also, let

$$r = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \text{rank}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta})). \quad (71)$$

By the definition of supremum, there exists  $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$  such that

$$r - 1/2 \leq \text{rank}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}_0)) \leq r.$$

Because the rank of a matrix can only take integer values, we know that  $\text{rank}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}_0)) = r$ . Let  $\mu_1(\mathbf{A}) \geq \mu_2(\mathbf{A}) \geq \dots \geq \mu_p(\mathbf{A})$  be the eigenvalues of a positive semidefinite matrix  $\mathbf{A}$ . Applying Courant–Fischer–Weyl min-max principle (see Chapter I of Hilbert and Courant (1953) or Corollary III.1.2 in Bhatia (1997)) to  $\underline{c} \cdot \mathbf{P}_{V_Q(\boldsymbol{\theta})} \preceq \sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta})$ , and  $r(\boldsymbol{\theta}) = \dim(V_Q(\boldsymbol{\theta}))$  (see Lemma 13.9), we obtain

$$\mu_{r(\boldsymbol{\theta})}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta})) \geq \mu_{r(\boldsymbol{\theta})}(\underline{c} \cdot \mathbf{P}_{V_Q(\boldsymbol{\theta})}) = \underline{c} > 0, \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}. \quad (72)$$

Applying Courant–Fischer–Weyl min-max principle to  $\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}) \preceq \bar{c} \cdot \mathbf{P}_{V_Q(\boldsymbol{\theta})}$ , and  $r(\boldsymbol{\theta}) = \dim(V_Q(\boldsymbol{\theta}))$ , we obtain

$$\mu_{r(\boldsymbol{\theta})+1}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta})) \leq \mu_{r(\boldsymbol{\theta})+1}(\bar{c} \cdot \mathbf{P}_{V_Q(\boldsymbol{\theta})}) = 0, \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

We will prove  $r(\boldsymbol{\theta}) = r$  for all  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$  by contradiction. Assume, in contrast, that there exists  $r_1 = r(\boldsymbol{\theta}_1) < r(\boldsymbol{\theta}_0) = r$ . Then, there exists a continuous path  $h : [0, 1] \rightarrow \boldsymbol{\Theta}$  such that  $h(0) = \boldsymbol{\theta}_0$ , and  $h(1) = \boldsymbol{\theta}_1$ . Set

$$u(t) = \mu_r\left(\sum_{a \in Q} \mathcal{I}_a(h(t))\right).$$

Note that  $u(0) = \mu_r(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}_0)) \geq \underline{c}$  and  $u(1) = \mu_r(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}_1)) = 0$ . Because  $u(t)$  is a continuous function in  $t \in [0, 1]$ , by the intermediate value theorem, there exists  $t' \in (0, 1)$  such that  $u(t') = \underline{c}/2$ .

Let  $\boldsymbol{\theta}_2 = h(t')$ . Because  $\mu_r(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}_2)) = \underline{c}/2 > 0$ , we know that  $\text{rank}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}_2)) \geq r$ . By definition (71), we know that  $\text{rank}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}_2)) \leq r$ . Thus,  $r(\boldsymbol{\theta}_2) = r$ . However,  $\mu_r(\boldsymbol{\theta}_2)(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}_2)) = \underline{c}/2$  contradicts inequality (72). This completes the proof that  $r(\boldsymbol{\theta}) = r$  for all  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ .

**2  $\implies$  3** For  $Q \subset \mathcal{A}$ , define

$$c_{\min}(Q) = \min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \Lambda_{\min}\left(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta})\right), \text{ and } c_{\max}(Q) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \Lambda_{\max}\left(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta})\right),$$

where  $\Lambda_{\min}$  and  $\Lambda_{\max}$  represent the smallest and largest non-zero eigenvalue of a positive semidefinite matrix, respectively.

Because  $\text{rank}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}))$  does not depend on  $\boldsymbol{\theta}$ , let  $r = \text{rank}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}))$ . Let  $\lambda_{(s)}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}))$  denote the  $s$ -th largest eigenvalue of  $\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta})$ ,  $s = 1, 2, \dots, p$ . Note that  $\Lambda_{\min}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta})) = \lambda_{(r)}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}))$  and  $\Lambda_{\max}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta})) = \lambda_{(1)}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}))$ . Now, we know that  $\Lambda_{\min}$  and  $\Lambda_{\max}$  are continuous functions provided  $\text{rank}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}))$  does not depend on  $\boldsymbol{\theta}$ , and  $\mathcal{I}_a(\boldsymbol{\theta})$  is continuous over compact set  $\boldsymbol{\Theta}$ . Thus,  $0 < c_{\min}(Q) \leq c_{\max}(Q) < \infty$ .

Recall that  $V_Q(\boldsymbol{\theta})^\perp = \ker(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}))$  from Lemma 13.9. Because for the positive semidefinite matrix  $\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta})$ ,  $\ker(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}))^\perp = \mathcal{R}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}))$ , we obtain  $V_Q(\boldsymbol{\theta}) = \mathcal{R}(\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}))$ .

Applying eigendecomposition of  $\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta})$ , we have

$$c_{\min}(Q) \cdot \mathbf{P}_{V_Q(\boldsymbol{\theta})} \preceq \sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta}) \preceq c_{\max}(Q) \cdot \mathbf{P}_{V_Q(\boldsymbol{\theta})}.$$

Set  $\underline{c} = \min_{Q \subset \mathcal{A}} c_{\min}(Q)$  and  $\bar{c} = \max_{Q \subset \mathcal{A}} c_{\max}(Q)$ . Since  $\mathcal{A}$  is a finite set, we know that  $\underline{c} > 0$  and  $\bar{c} < \infty$ .  $\square$

## 13.5 Other Supporting Lemmas

**Lemma 13.11.** *Assume that positive semidefinite matrices  $\mathbf{A}, \mathbf{B}$  and  $\mathbf{C}$  have the same size. If  $\mathbf{A} \succeq \mathbf{C}$ , then*

$$\text{tr}(\mathbf{AB}) \geq \text{tr}(\mathbf{CB}). \quad (73)$$

*Proof of Lemma 13.11.* Because  $\mathbf{B}^{1/2}(\mathbf{A} - \mathbf{C})\mathbf{B}^{1/2}$  is positive semi-definite,

$$\text{tr}(\mathbf{AB}) - \text{tr}(\mathbf{CB}) = \text{tr}(\mathbf{B}^{1/2}(\mathbf{A} - \mathbf{C})\mathbf{B}^{1/2}) \geq 0. \quad (74)$$

This completes the proof.  $\square$

**Lemma 13.12** (Multivariate Cauchy-Schwartz Inequality). *For any random variable  $z$  and random vector  $\mathbf{y}$ , if  $\text{cov}(\mathbf{y}) = \Sigma_{\mathbf{y}}$  is positive definite matrix, then*

$$\text{var}(z) \geq \text{cov}(z, \mathbf{y}) \text{cov}(\mathbf{y})^{-1} \text{cov}(\mathbf{y}, z). \quad (75)$$

*Proof of Lemma 13.12.* Because

$$\begin{aligned} 0 &\leq \text{var}(z - \text{cov}(z, \mathbf{y})\Sigma_{\mathbf{y}}^{-1}\mathbf{y}) \\ &= \text{var}(z) - 2\text{cov}\left(\text{cov}(z, \mathbf{y})\Sigma_{\mathbf{y}}^{-1}\mathbf{y}, z\right) + \text{var}\left(\text{cov}(z, \mathbf{y})\Sigma_{\mathbf{y}}^{-1}\mathbf{y}\right) \\ &= \text{var}(z) - 2\text{cov}(z, \mathbf{y})\Sigma_{\mathbf{y}}^{-1}\text{cov}(\mathbf{y}, z) + \text{cov}(z, \mathbf{y})\Sigma_{\mathbf{y}}^{-1}\Sigma_{\mathbf{y}}\Sigma_{\mathbf{y}}^{-1}\text{cov}(\mathbf{y}, z) \\ &= \text{var}(z) - \text{cov}(z, \mathbf{y})\{\text{cov}(\mathbf{y})\}^{-1}\text{cov}(\mathbf{y}, z), \end{aligned}$$

we complete the proof of Lemma 13.12.  $\square$

**Lemma 13.13.** *Assumptions 6A and 7A imply Assumptions 6B and 7B.*

*Proof of Lemma 13.13.* Under Assumption 6A, we have

$$\mathcal{I}_a(\boldsymbol{\theta}) = \mathbf{Z}_a^T \mathcal{I}_{\xi_a, a}(\mathbf{Z}_a \boldsymbol{\theta}) \mathbf{Z}_a.$$

Let  $\mathbf{Z}_a^\dagger$  be the Moore-Penrose inverse of  $\mathbf{Z}_a$ . Because  $\mathbf{Z}_a$  has full row rank, we know that  $\mathbf{Z}_a \mathbf{Z}_a^\dagger = I_{p_a}$ , and

$$\mathcal{I}_{\xi_a, a}(\mathbf{Z}_a \boldsymbol{\theta}) = \{\mathbf{Z}_a^\dagger\}^T \mathcal{I}_a(\boldsymbol{\theta}) \mathbf{Z}_a^\dagger,$$

which implies that  $\mathcal{I}_{\xi_a, a}(\mathbf{Z}_a \boldsymbol{\theta})$  is continuous in  $\boldsymbol{\theta}$ . Thus, there exists  $0 < c_1 < c_2 < \infty$  such that

$$c_1 I_{p_a} \preceq \mathcal{I}_{\xi_a, a}(\mathbf{Z}_a \boldsymbol{\theta}) \preceq c_2 I_{p_a},$$

and for any  $Q \subset \mathcal{A}$ ,

$$V_Q(\boldsymbol{\theta}) = \sum_{a \in Q} \mathcal{R}(\mathcal{I}_a(\boldsymbol{\theta})) = \sum_{a \in Q} \mathcal{R}(\mathbf{Z}_a^T).$$

By Lemma 13.10, we know that

$$\dim(V_Q(\boldsymbol{\theta})) = \dim\left(\sum_{a \in Q} \mathcal{R}(\mathbf{Z}_a^T)\right)$$

does not depend on  $\boldsymbol{\theta}$ .

By Lemma 13.10, we obtain inequality (16). This proves that Assumption 6A implies Assumption 6B.

Next, we consider Assumption 7B. Under Assumptions 6A and 7A, we obtain

$$\begin{aligned} & D_{\text{KL}}(f_{\boldsymbol{\theta}^*, a} \| f_{\boldsymbol{\theta}, a}) \\ &= D_{\text{KL}}(h_{\boldsymbol{\xi}_a^*, a} \| h_{\boldsymbol{\xi}_a, a}) \\ &\geq C \|\boldsymbol{\xi}_a^* - \boldsymbol{\xi}_a\|^2 \\ &= C(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{Z}_a^T \mathbf{Z}_a (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\ &\geq \frac{C}{c_2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{Z}_a^T \mathcal{I}_{\boldsymbol{\xi}_a, a} (\mathbf{Z}_a \boldsymbol{\theta}^*) \mathbf{Z}_a (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\ &= \frac{C}{c_2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathcal{I}_a(\boldsymbol{\theta}^*) (\boldsymbol{\theta} - \boldsymbol{\theta}^*). \end{aligned}$$

Thus, for any  $\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}$  and  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ , we have

$$\sum_{a \in \mathcal{A}} \pi(a) D_{\text{KL}}(f_{\boldsymbol{\theta}^*, a} \| f_{\boldsymbol{\theta}, a}) \geq \frac{C}{c_2} \sum_{a \in \mathcal{A}} \pi(a) (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathcal{I}_a(\boldsymbol{\theta}^*) (\boldsymbol{\theta} - \boldsymbol{\theta}^*).$$

Replacing the constant  $\frac{C}{c_2}$  by  $C$ , we obtain inequality (17) in Assumption 7B. Thus, we obtain Assumptions 6B-7B. □

## 14 Proof of Theoretical Results

### 14.1 Proof of Lemma 3.1

*Proof of Lemma 3.1.* The standard computational complexity for both matrix multiplication and matrix inversion of a matrix of size  $p \times p$  is  $O(p^3)$ . Consequently, the computational complexity of evaluating  $\mathbb{G}_{\hat{\boldsymbol{\theta}}_n^{\text{ML}}} \left[ \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n, a)\}^{-1} \right]$  for each  $a \in \mathcal{A}$  is  $O(p^3)$ . Therefore, the computational complexity for the GI0 selection is  $O(kp^3)$ .

The GI1 selection rule (3) can be reformulated as

$$\begin{aligned}
& a_{n+1} \\
&= \arg \max_{a \in \mathcal{A}} \text{tr} \left[ \nabla \mathbb{G}_{\hat{\boldsymbol{\theta}}_n^{\text{ML}}} (\{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1}) \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1} \mathcal{I}_a(\hat{\boldsymbol{\theta}}_n^{\text{ML}}) \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1} \right] \\
&= \arg \max_{a \in \mathcal{A}} \text{tr} \left[ L_a^T(\hat{\boldsymbol{\theta}}_n^{\text{ML}}) \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1} \nabla \mathbb{G}_{\hat{\boldsymbol{\theta}}_n^{\text{ML}}} (\{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1}) \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1} L_a(\hat{\boldsymbol{\theta}}_n^{\text{ML}}) \right].
\end{aligned}$$

Thus, Algorithm 3 produce the same outcome as Algorithm 2.

Note that the matrix

$$\mathbf{M} = \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1} \nabla \mathbb{G}_{\hat{\boldsymbol{\theta}}_n^{\text{ML}}} (\{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1}) \{\mathcal{I}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1}$$

only needs to be computed once. Matrix multiplication involving matrices of sizes  $p \times p$  and  $p \times s_a$  is of order  $O(p^2 s)$ , given  $s_a \leq s$ . Under the assumption that  $L_a(\boldsymbol{\theta})$  has size  $p \times s_a$ , the computational complexity of the GI1 is bounded by  $O(p^3 + k s p^2)$ .

Furthermore, if the matrices  $\{L_a(\boldsymbol{\theta})\}$  are primarily supported on an  $s \times s$  submatrix, then the computational cost of the multiplication  $\boldsymbol{\Sigma} L_a(\hat{\boldsymbol{\theta}}_n^{\text{ML}})$  is at most equivalent to multiplying matrices of sizes  $p \times s$  and  $s \times s$  for any matrix  $\boldsymbol{\Sigma}$ , which has a complexity of  $O(s^2 p)$ . Therefore, the overall computational complexity of the GI1 selection is  $O(p^3 + k s^2 p)$ .  $\square$

## 14.2 Proof of Proposition 6.2

We prove the following Theorem 14.1 instead, which is a generalized version of Proposition 6.2 allowing for an arbitrary sequence of  $\boldsymbol{\theta}_n$  that is not necessarily the MLE.

**Theorem 14.1.** *Under the regularity conditions described in Section 4.1, and also assume that the initial experiments  $a_1, \dots, a_{n_0} \in \mathcal{A}$  are such that  $\mathcal{I}(\boldsymbol{\theta}; \mathbf{a}_{n_0})$  is nonsingular. For any sequence of (random or non-random) vectors  $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots$  in  $\boldsymbol{\Theta}$ , if we consider the following generalized GI0 or GI1 selection rules: for any  $n \geq n_0$*

$$\text{GI0: } a_{n+1} = \arg \min_{a \in \mathcal{A}} \mathbb{G}_{\boldsymbol{\theta}_n} [\{\mathcal{I}(\boldsymbol{\theta}_n; \mathbf{a}_n, a)\}^{-1}], \text{ and} \quad (76)$$

$$\text{GI1: } a_{n+1} = \arg \max_{a \in \mathcal{A}} \text{tr} \left[ \nabla \mathbb{G}_{\boldsymbol{\theta}_n} (\boldsymbol{\Sigma}_n) \boldsymbol{\Sigma}_n \mathcal{I}_a(\boldsymbol{\theta}_n) \boldsymbol{\Sigma}_n \right], \text{ where we define } \boldsymbol{\Sigma}_n = \{\mathcal{I}(\boldsymbol{\theta}_n; \mathbf{a}_n)\}^{-1}, \quad (77)$$

then there exists  $C > 0$  such that

$$\inf_{n \geq n_0} \frac{n_I}{n} \geq C,$$

and the lower bound is independent of the choice of  $\theta_n$ . Moreover, under the same settings,

$$\mathcal{I}^{\bar{\pi}_n}(\theta) \succeq \underline{c} \cdot C \cdot I_p. \quad (78)$$

for all  $\theta \in \Theta$ .

*Proof of Proposition 6.2.* Applying Theorem 14.1 with  $\theta_n = \hat{\theta}_n^{\text{ML}}$ , we complete the proof of Proposition 6.2.  $\square$

The proof of Theorem 14.1 is involved. We break it down to the following series of lemmas and steps.

**Step 1: Define the order statistics of experiments counts, permutations, and find their connections with  $n_{\max}$  and  $n_I$**  Let  $m_a^n = |\{i; a_i = a, 1 \leq i \leq n\}|$  be the number of times that the experiment  $a$  has been selected up to time  $n$ . Without loss of generality, let  $\mathcal{A} = [k] = \{1, 2, \dots, k\}$ . Let  $\mathcal{P}_k$  be the set of all permutations over  $[k]$ .

For each  $m^n = (m_a^n)_{a \in \mathcal{A}}$ , define  $\mathcal{P}_k^{m^n} \subset \mathcal{P}_k$ , which is described by the following statements: permutation  $\tau \in \mathcal{P}_k^{m^n}$  if and only if  $\tau \in \mathcal{P}_k$  and

$$m_{\tau(1)}^n \geq m_{\tau(2)}^n \geq \dots \geq m_{\tau(k)}^n.$$

The set  $\mathcal{P}_k^{m^n}$  is not empty, because order statistic exists.

For any permutation  $\tau \in \mathcal{P}_k^{m^n}$ , define the set  $Q_s(\tau) = \{\tau(1), \tau(2), \dots, \tau(s)\}$  for  $s \in [k]$ .  $Q_s(\tau)$  collects the indices of the top- $s$  most frequently selected experiments. Here,  $\tau$  is introduced to handle the case where there may be ties among  $m_a^n$  for  $a \in \mathcal{A}$ .

Define

$$t_n = t_n(m^n) = \min_{\tau \in \mathcal{P}_k^{m^n}} \{s \in [k]; \dim(V_{Q_s(\tau)}(\theta)) = p\}. \quad (79)$$

Note that  $\dim(V_{Q_k(\tau)}(\theta)) = \text{rank}(\sum_{a \in \mathcal{A}} \mathcal{I}_a(\theta)) = p$  for all  $\tau \in \mathcal{P}_k^{m^n}$ , according to Lemma 13.9 and Assumption 3. Also note that according to Lemma 13.10 and Assumption 6B,  $\dim(V_{Q_s(\tau)}(\theta))$  does not depend on  $\theta$ . Thus, the above  $t_n$  is well defined and does not depend on  $\theta$ . For the same reason, we will drop ' $\theta$ ' and write  $\dim(V_A)$  for  $\dim(V_A(\theta))$  for  $A \subset \mathcal{A}$  in the rest of the proof when the context is clear.

The next lemma specifies the permutations that we would like to focus on when there may be ties among  $m_a^n$ .

**Lemma 14.2.** *There exists  $\tau_n \in \mathcal{P}_k$  such that*

$$m_{\tau_n(1)}^n \geq m_{\tau_n(2)}^n \geq \dots \geq m_{\tau_n(k)}^n, \quad (80)$$



$\dim(V_{Q_{t_n}(\tau_n)}) = p$ , and for all  $\tau' \in \mathcal{P}_k^{m^n}$  and all  $s \leq t_n - 1$ ,  $\dim(V_{Q_s(\tau')}) < p$ .

*Proof of Lemma 14.2.* First, according to the definition of  $t_n$  in (79) and  $\mathcal{P}_k^{m^n} \neq \emptyset$ , we know that

$$S' = \arg \min_{\tau \in \mathcal{P}_k^{m^n}} \{s \in [k]; \dim(V_{Q_s(\tau)}(\boldsymbol{\theta})) = p\}$$

is not empty. Let  $\tau_n \in S'$ . We know that  $\tau_n$  satisfies (80), and  $\dim(V_{Q_{t_n}(\tau_n)}) = p$ .

Assume there exist  $\tau' \in \mathcal{P}_k^{m^n}$  and  $s \leq t_n - 1$ , such that  $\dim(V_{Q_s(\tau')}) = p$ . This leads to the following contradiction

$$t_n = \min_{\tau \in \mathcal{P}_k^{m^n}} \{s \in [k]; \dim(V_{Q_s(\tau)}(\boldsymbol{\theta})) = p\} \leq s \leq t_n - 1.$$

This completes the proof of Lemma 14.2.  $\square$

Recall that  $n_{\max} = \max_{a \in \mathcal{A}} n_a = \max_{a \in \mathcal{A}} m_a^n$  is defined in Section 6. We obtain that  $n_{\max} = m_{\tau_n(1)}^n$ . The following Lemma shows that  $n_I = m_{\tau_n(t_n)}^n$ .

**Lemma 14.3.** *Let  $\tau_n$  be a permutation satisfying the properties described in Lemma 14.2. Then,  $n_I = m_{\tau_n(t_n)}^n$ , where  $n_I$  is defined in (42).*

*Proof of Lemma 14.3.* Because  $\dim(V_{Q_{t_n}(\tau_n)}(\boldsymbol{\theta})) = p$ , we know that  $Q = Q_{t_n}(\tau_n)$  is relevant, which means that  $\sum_{a \in Q} \mathcal{I}_a(\boldsymbol{\theta})$  is non-singular for all  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ . By the definition of  $n_I$  in (42),

$$n_I \geq \min_{a \in Q} m_a^n = m_{\tau_n(t_n)}^n.$$

It suffices to prove  $n_I \leq m_{\tau_n(t_n)}^n$ . Assume, on the contrary, that  $n_I > m_{\tau_n(t_n)}^n$ . In the rest of the proof, we aim to find a contradiction.

For any  $S \subset \mathcal{A}$  such that  $S$  is relevant, define  $Q(S) = \{a \in \mathcal{A}; m_a^n \geq \min_{a \in S} m_a^n\}$ . Since  $S \subset Q(S)$ ,  $Q(S)$  is also relevant, and

$$\min_{a \in S} m_a^n = \min_{a \in Q(S)} m_a^n.$$

By the definition of  $n_I$  in (42), there exists a relevant  $S' \subset \mathcal{A}$  such that  $\min_{a \in S'} m_a^n = n_I$ . Thus,  $n_I = \min_{a \in S'} m_a^n = \min_{a \in Q(S')} m_a^n$ ,  $Q(S')$  is relevant and  $\min_{a \in Q(S')} m_a^n > m_{\tau_n(t_n)}^n$ . This implies that

$$\min_{a \in Q(S')} m_a^n \geq m_{\tau_n(t_n-1)}^n.$$

Thus,  $Q(S') \subset Q_{t_n-1}(\tau_n)$ , which implies that  $\dim(V_{Q(S')}) \leq \dim(Q_{t_n-1}(\tau_n)) < p$ . By Assumption 6B and Lemma 13.10, we know that  $Q(S')$  is not relevant, which contradicts the previous assumption that  $Q(S')$  is relevant.  $\square$

The next lemma compares the ratio between  $n_{\max} = m_{\tau_n(1)}^n$  and  $n_I = m_{\tau_n(t_n)}^n$  with the ratio between the maximum and minimum counts of experiments for a set of relevant experiments.

**Lemma 14.4.** *For any  $Q \subset \mathcal{A}$  such that  $\dim(V_Q) = p$ ,*

$$\frac{m_{\tau_n(1)}^n}{m_{\tau_n(t_n)}^n} \leq \frac{\max_{a \in \mathcal{A}} m_a^n}{\min_{a \in Q} m_a^n}.$$

*Proof of Lemma 14.4.* By the definition of  $\tau_n$ , we know that

$$m_{\tau_n(1)}^n \geq \cdots \geq m_{\tau_n(k)}^n.$$

Define  $m_{\tau_n(0)}^n = \infty$  and  $m_{\tau_n(k+1)}^n = -\infty$ .

Because  $(m_{\tau_n(1)}^n, \dots, m_{\tau_n(k)}^n)$  forms the order statistic of  $(m_a^n)_{a \in [k]}$  (with possibly ties), there exists  $s \in [k]$  such that  $Q \subset \{\tau_n(1), \dots, \tau_n(s)\} = Q_s(\tau_n)$ , and

$$m_{\tau_n(s)}^n = \min_{a \in Q} m_a^n \text{ and } m_{\tau_n(s+1)}^n < \min_{a \in Q} m_a^n.$$

Because  $V_Q(\boldsymbol{\theta}) \subset V_{Q_s(\tau_n)}(\boldsymbol{\theta})$ , we have  $p = \dim(V_Q) \leq \dim(V_{Q_s(\tau_n)})$ . By the definition of  $t_n$  in (79), we obtain that  $t_n \leq s$ . This implies  $m_{\tau_n(t_n)}^n \geq m_{\tau_n(s)}^n = \min_{a \in Q} m_a^n$ . We complete the proof by noting that  $m_{\tau_n(1)}^n = \max_{a \in \mathcal{A}} m_a^n$ . □

**Step 2: Unify the proof for generalized GI0 and GI1** To simplify the analysis, we use the next lemma to extract a key property shared by generalized GI0 and GI1.

**Lemma 14.5.** *Assume that  $\sum_{i=1}^{n_0} \mathcal{I}_{a_i}(\boldsymbol{\theta})$  is non-singular.*

*For a fixed (or random) sequence  $\boldsymbol{\theta}_n \in \boldsymbol{\Theta}$  and for any  $n \geq n_0$ , we consider the following generalized GI0 selection rule*

$$a_{n+1} = \arg \min_{a \in \mathcal{A}} \mathbb{F}_{\boldsymbol{\theta}_n} \left( \frac{n}{n+1} \bar{\boldsymbol{\pi}}_n + \frac{1}{n+1} \delta_a \right) = \arg \min_{a \in \mathcal{A}} \mathbb{G}_{\boldsymbol{\theta}_n} \left( \left\{ \frac{1}{n+1} \mathbf{A} + \frac{1}{n+1} \mathcal{I}_a(\boldsymbol{\theta}_n) \right\}^{-1} \right),$$

*and GI1 selection rule*

$$a_{n+1} = \arg \min_{a \in \mathcal{A}} \frac{\partial \mathbb{F}_{\boldsymbol{\theta}_n}(\bar{\boldsymbol{\pi}}_n)}{\partial \pi(a)} = \arg \max_{a \in \mathcal{A}} \langle \nabla \mathbb{G}_{\boldsymbol{\theta}_n}(\{\mathbf{A}/n\}^{-1}), \mathbf{A}^{-1} \mathcal{I}_a(\boldsymbol{\theta}_n) \mathbf{A}^{-1} \rangle, \quad (81)$$

where  $\mathbf{A} = \sum_{a \in \mathcal{A}} m_a^n \mathcal{I}_a(\boldsymbol{\theta}_n)$ ,  $\delta_a = (\delta_a(a'))_{a' \in \mathcal{A}}$ , and  $\delta_a(a') = I(a = a')$ .

Let  $A_2(t_1, t_2) = \sum_{a \in \mathcal{A}} \mathcal{I}_a(\boldsymbol{\theta}_n) + t_1 \mathcal{I}_{a'}(\boldsymbol{\theta}_n) + t_2 \mathcal{I}_{a''}(\boldsymbol{\theta}_n)$ , and  $\mathbf{S}_n = \mathbf{S}_n(t_1, t_2) =$

$\nabla \mathbb{G}_{\boldsymbol{\theta}_n}(\{A_2(t_1, t_2)/(n+1)\}^{-1})$ . Then, both generalized GI0 and GI1 satisfy the following property for all  $n \geq n_0$ :

If  $a', a'' \in \mathcal{A}$  are such that

$$\langle \mathbf{S}_n, \{A_2(t_1, t_2)\}^{-1} \mathcal{I}_{a'}(\boldsymbol{\theta}_n) \{A_2(t_1, t_2)\}^{-1} \rangle > \langle \mathbf{S}_n, \{A_2(t_1, t_2)\}^{-1} \mathcal{I}_{a''}(\boldsymbol{\theta}_n) \{A_2(t_1, t_2)\}^{-1} \rangle, \quad (82)$$

for all  $t_1, t_2 \in [0, 1]$ , then  $a_{n+1} \neq a''$ .

*Remark 14.6.* The generalized GI0 and GI1 defined in (76) and (77) are the same as GI0 and GI1 selections described in Lemma 14.5, respectively.

*Proof of Lemma 14.5.* Let  $a', a''$  satisfy (82). Assume, in the contrast, that  $a_{n+1} = a''$ . We will find contradictions for both GI0 and GI1 in the rest of the proof.

We start with GI0, which selects  $a_{n+1} = \arg \min_{a \in \mathcal{A}} \mathbb{F}_{\boldsymbol{\theta}_n}(\frac{n}{n+1} \bar{\boldsymbol{\pi}}_n + \frac{1}{n+1} \delta_a)$ . Thus,  $a'' = a_{n+1}$  satisfies

$$\mathbb{F}_{\boldsymbol{\theta}_n}(\frac{n}{n+1} \bar{\boldsymbol{\pi}}_n + \frac{1}{n+1} \delta_{a''}) \leq \mathbb{F}_{\boldsymbol{\theta}_n}(\frac{n}{n+1} \bar{\boldsymbol{\pi}}_n + \frac{1}{n+1} \delta_{a'}). \quad (83)$$

Define  $h(t) = \mathbb{F}_{\boldsymbol{\theta}_n}(\frac{n}{n+1} \bar{\boldsymbol{\pi}}_n + \frac{1}{n+1} \{(1-t)\delta_{a'} + t\delta_{a''}\})$ . Then, (83) is equivalent to that  $h(1) - h(0) \leq 0$ .

Let  $A_1(t) = \sum_{a \in \mathcal{A}} \mathcal{I}_a(\boldsymbol{\theta}_n) + (1-t)\mathcal{I}_{a'}(\boldsymbol{\theta}_n) + t\mathcal{I}_{a''}(\boldsymbol{\theta}_n)$ . By Lemma 13.4, we know that

$$h'(t) = \langle \nabla \mathbb{G}_{\boldsymbol{\theta}_n}(\{A_1(t)/(n+1)\}^{-1}), -\{A_1(t)/(n+1)\}^{-1} \{\mathcal{I}_{a''}(\boldsymbol{\theta}_n) - \mathcal{I}_{a'}(\boldsymbol{\theta}_n)\} \{A_1(t)/(n+1)\}^{-1} \rangle.$$

Note that  $A_1(t) = A_2(1-t, t)$ . Thus, (82) holds for all  $t_1, t_2 \in [0, 1]$  implies that it holds for  $(t_1, t_2) = (1-t, t)$ , which further implies  $h'(t) > 0$  for any  $t \in (0, 1)$ . This contradicts with  $h(1) - h(0) \leq 0$ . Thus,  $a_{n+1} \neq a''$ .

We proceed to the analysis of GI1. By the definition of the generalized GI1,  $a'' = a_{n+1}$  satisfies

$$\langle \nabla \mathbb{G}_{\boldsymbol{\theta}_n}(\{\mathbf{A}/n\}^{-1}), \mathbf{A}^{-1} \mathcal{I}_{a'}(\boldsymbol{\theta}_n) \mathbf{A}^{-1} \rangle \leq \langle \nabla \mathbb{G}_{\boldsymbol{\theta}_n}(\{\mathbf{A}/n\}^{-1}), \mathbf{A}^{-1} \mathcal{I}_{a''}(\boldsymbol{\theta}_n) \mathbf{A}^{-1} \rangle,$$

Note that  $\mathbf{A} = A_2(0, 0)$ . Thus, the above inequality contradicts with (82) with  $(t_1, t_2) = (0, 0)$ .  $\square$

**Step 3: Regularization effect of GI0 and GI1** In this step, we show that both GI0 and GI1 regularize the experiment selection process through the property established in Lemma 14.5. This is proved through the following Lemma 14.7, Lemma 14.8, and Lemma 14.10.

**Lemma 14.7.** Assume that  $\sum_{i=1}^{n_0} \mathcal{I}_{a_i}(\boldsymbol{\theta})$  is non-singular for some  $n_0$ .

Assume the condition number  $\max_{\boldsymbol{\theta} \in \Theta, \Sigma \geq 0} \kappa(\nabla \mathbb{G}_{\boldsymbol{\theta}}(\Sigma)) \leq K$  for some  $0 < K < \infty$ . Let  $(a^{(1)}, a^{(2)}, \dots, a^{(k)})$  be a permutation of  $\mathcal{A}$  such that  $m_{a^{(1)}}^n \geq \dots \geq m_{a^{(k)}}^n$  and  $\dim(Q_{t_n-1}) < \dim(Q_{t_n}) = p$ , where  $Q_s = \{a^{(1)}, a^{(2)}, \dots, a^{(s)}\}$ , and  $t_n = t_n(m^n)$ .

If for some  $n \geq n_0$  and  $1 \leq s \leq t_n - 1$

$$\left( \frac{m_{a^{(s)}}^n}{m_{a^{(s+1)}}^n} \right)^2 > \frac{8\bar{c}^3 p K}{\underline{c}^3} \left( 1 + 16p \frac{\bar{c}^2}{\underline{c}^2} \left( \frac{m_{a^{(s+1)}}^n}{m_{a^{(t_n)}}^n} \right)^2 \right), \quad (84)$$

then

$$a_{n+1} \in \mathcal{G}(Q_s) = \{a \in \mathcal{A} : \dim(V_{Q_s \cup \{a\}}) > \dim(V_{Q_s})\}$$

for GI0 and GI1, where

$$V_Q = V_Q(\boldsymbol{\theta}_n) = \sum_{a \in Q} \mathcal{R}(\mathcal{I}_a(\boldsymbol{\theta}_n)).$$

*Proof of Lemma 14.7.* Let  $t = t_n$ . (84) implies that  $t \geq 2$ . By Lemma 14.2, for any  $s \leq t-1$ , we have  $\dim(V_{Q_t}) = p$ . Because  $\dim(V_{Q_s}) < p$ , we know that  $|\mathcal{G}(Q_s)| > 0$ .

For any  $Q \subset \mathcal{A}$ , let  $\mathbf{P}_{V_Q}$  be the orthogonal projection matrix on  $V_Q(\boldsymbol{\theta}_n)$ . We will simplify the notation and write it as  $\mathbf{P}_Q$  for the ease of exposition when the context is clear. Then  $\mathbf{P}_{Q_s}$  denote the orthogonal projection matrix on  $V_{Q_s}$ .

According to Lemma 14.5, it is sufficient to show that, if (84) holds, then for all  $a'' \notin \mathcal{G}(Q_s)$ ,  $a' \in \mathcal{G}(Q_s)$  and all  $t_1, t_2 \in [0, 1]$ .

$$\langle \mathbf{S}_n, \{A_2(t_1, t_2)\}^{-1} \mathcal{I}_{a'}(\boldsymbol{\theta}_n) \{A_2(t_1, t_2)\}^{-1} \rangle > \langle \mathbf{S}_n, \{A_2(t_1, t_2)\}^{-1} \mathcal{I}_{a''}(\boldsymbol{\theta}_n) \{A_2(t_1, t_2)\}^{-1} \rangle, \quad (85)$$

where we recall that  $\mathbf{S}_n = \nabla \mathbb{G}_{\boldsymbol{\theta}_n}(\{A_2(t_1, t_2)/(n+1)\}^{-1})$ . In the rest of the proof, we abuse the notation a little and write  $\mathbf{A} = A_2(t_1, t_2)$  for the ease of exposition. Then, it is sufficient to show that for all  $a'' \notin \mathcal{G}(Q_s)$ ,  $a' \in \mathcal{G}(Q_s)$  and all  $t_1, t_2 \in [0, 1]$

$$\langle \mathbf{S}_n, \mathbf{A}^{-1} \mathcal{I}_{a'}(\boldsymbol{\theta}_n) \mathbf{A}^{-1} \rangle > \langle \mathbf{S}_n, \mathbf{A}^{-1} \mathcal{I}_{a''}(\boldsymbol{\theta}_n) \mathbf{A}^{-1} \rangle. \quad (86)$$

The rest proof of the consists of the following three steps:

**Step A: Connect (86) with  $\text{tr}(\mathbf{P}_{Q_s \cup \{a'\}} \mathbf{A}^{-2})$  and  $\text{tr}(\mathbf{P}_{Q_s} \mathbf{A}^{-2})$**  Let

$$\bar{m}_a^n = \begin{cases} m_{a'}^n + t_1 & \text{if } a = a' \\ m_{a'}^n + t_2 & \text{if } a = a'' \\ m_a^n & \text{otherwise} \end{cases}. \quad (87)$$

Then, for  $0 \leq t_1, t_2 \leq 1$ ,  $\bar{m}_a^n \leq m_a^n + 1$  for all  $a \in \mathcal{A}$ , and  $\mathbf{A} = \sum_{a \in \mathcal{A}} \bar{m}_a^n \mathcal{I}_a(\boldsymbol{\theta}_n)$ .

Note that  $\lambda_{\max}(\mathbf{S}_n) \mathbf{I}_p \succeq \mathbf{S}_n \succeq \lambda_{\min}(\mathbf{S}_n) \mathbf{I}_p$ . we have

$$\begin{aligned}
& \langle \mathbf{S}_n, \mathbf{A}^{-1} \mathcal{I}_{a'}(\boldsymbol{\theta}_n) \mathbf{A}^{-1} \rangle \\
&= \text{tr}(\mathbf{S}_n \mathbf{A}^{-1} \mathcal{I}_{a'}(\boldsymbol{\theta}_n) \mathbf{A}^{-1}) \\
&= \text{tr}(\mathbf{S}_n \mathbf{A}^{-1} \sum_{a \in Q_s \cup \{a'\}} \mathcal{I}_a(\boldsymbol{\theta}_n) \mathbf{A}^{-1}) - \text{tr}(\mathbf{S}_n \mathbf{A}^{-1} \sum_{a \in Q_s} \mathcal{I}_a(\boldsymbol{\theta}_n) \mathbf{A}^{-1}) \\
&\geq \underline{c} \cdot \lambda_{\min}(\mathbf{S}_n) \text{tr}(\mathbf{P}_{Q_s \cup \{a'\}} \mathbf{A}^{-2}) - \bar{c} \cdot \lambda_{\max}(\mathbf{S}_n) \text{tr}(\mathbf{A}^{-1} \mathbf{P}_{Q_s} \mathbf{A}^{-1}),
\end{aligned} \tag{88}$$

where the last inequality is due to Assumption 6B and Lemma 13.11.

Notice that  $a'' \notin \mathcal{G}(Q_s)$  implies

$$\dim(V_{Q_s \cup \{a''\}}) = \dim(V_{Q_s}).$$

Combined with  $V_{Q_s} \subset V_{Q_s \cup \{a''\}}$ , we know that  $V_{Q_s} = V_{Q_s \cup \{a''\}}$ . This implies

$$\mathcal{R}(\mathcal{I}_{a''}(\boldsymbol{\theta}_n)) \subset V_{Q_s}. \tag{89}$$

By Assumption 6B, we obtain

$$\mathcal{I}_{a''}(\boldsymbol{\theta}_n) \preceq \bar{c} \cdot \mathbf{P}_{\mathcal{R}(\mathcal{I}_{a''}(\boldsymbol{\theta}_n))} \preceq \bar{c} \cdot \mathbf{P}_{Q_s}.$$

Hence,

$$\langle \mathbf{S}_n, \mathbf{A}^{-1} \mathcal{I}_{a''}(\boldsymbol{\theta}_n) \mathbf{A}^{-1} \rangle = \text{tr}(\mathbf{S}_n \mathbf{A}^{-1} \mathcal{I}_{a''}(\boldsymbol{\theta}_n) \mathbf{A}^{-1}) \leq \bar{c} \cdot \lambda_{\max}(\mathbf{S}_n) \text{tr}(\mathbf{A}^{-1} \mathbf{P}_{Q_s} \mathbf{A}^{-1}). \tag{90}$$

Thus, to show (86), it is sufficient to show that (84) implies

$$\underline{c} \cdot \lambda_{\min}(\mathbf{S}_n) \text{tr}(\mathbf{A}^{-1} \mathbf{P}_{Q_s \cup \{a'\}} \mathbf{A}^{-1}) > 2\bar{c} \cdot \lambda_{\max}(\mathbf{S}_n) \text{tr}(\mathbf{A}^{-1} \mathbf{P}_{Q_s} \mathbf{A}^{-1}), \tag{91}$$

which is equivalent to

$$\frac{\text{tr}(\mathbf{P}_{Q_s \cup \{a'\}} \mathbf{A}^{-2})}{\text{tr}(\mathbf{P}_{Q_s} \mathbf{A}^{-2})} > \frac{2\bar{c} \cdot \lambda_{\max}(\mathbf{S}_n)}{\underline{c} \cdot \lambda_{\min}(\mathbf{S}_n)} = \frac{2\bar{c}}{\underline{c}} \kappa(\mathbf{S}_n). \tag{92}$$

We focus on proving the above inequality in the rest of the proof.

**Step B: Establish a lower bound for  $\text{tr}(\mathbf{P}_{Q_s \cup \{a'\}} \mathbf{A}^{-2})$**  Because  $V_{Q_s} \subset V_{Q_s \cup \{a'\}}$  and  $\dim(V_{Q_s}) < \dim(V_{Q_s \cup \{a'\}})$ , we know that  $(\mathbf{I} - \mathbf{P}_{Q_s}) \mathbf{P}_{Q_s \cup \{a'\}} = \mathbf{P}_{Q_s \cup \{a'\}} - \mathbf{P}_{Q_s} \neq \mathbf{0}$ . Thus,

there exists a unit vector  $\mathbf{u} \in \mathbb{R}^p$  such that  $\|(I - \mathbf{P}_Q)\mathbf{P}_{Q_s \cup \{a'\}}\mathbf{u}\| = 1$ .

Applying the Rayleigh–Ritz quotient for the largest eigenvalue, we know that

$$\begin{aligned} \text{tr}(\mathbf{P}_{Q_s \cup \{a'\}}\mathbf{A}^{-2}) &= \text{tr}(\mathbf{P}_{Q_s \cup \{a'\}}\mathbf{A}^{-2}\mathbf{P}_{Q_s \cup \{a'\}}) \\ &\geq \lambda_{\max}(\mathbf{P}_{Q_s \cup \{a'\}}\mathbf{A}^{-2}\mathbf{P}_{Q_s \cup \{a'\}}) \geq \mathbf{u}^T \mathbf{P}_{Q_s \cup \{a'\}}\mathbf{A}^{-2}\mathbf{P}_{Q_s \cup \{a'\}}\mathbf{u} = \|\mathbf{A}^{-1}\mathbf{P}_{Q_s \cup \{a'\}}\mathbf{u}\|^2. \end{aligned} \quad (93)$$

Set  $\mathbf{v} = \mathbf{A}^{-1}\mathbf{P}_{Q_s \cup \{a'\}}\mathbf{u}$ . Then,

$$\|(I - \mathbf{P}_{Q_s})\mathbf{A}\|_{op} \|\mathbf{v}\| \geq \|(I - \mathbf{P}_{Q_s})\mathbf{A}\mathbf{v}\| = \|(I - \mathbf{P}_{Q_s})\mathbf{P}_{Q_s \cup \{a'\}}\mathbf{u}\| = 1. \quad (94)$$

Notice that  $m_{a(s+1)}^n \geq m_{a(t_n)}^n \geq 1$  for any  $s \leq t - 1$ . Thus,  $m_{a(s+1)}^n + 1 \leq 2m_{a(s+1)}^n$ . Also note that by the definition of  $\mathcal{G}(Q_s)$ ,  $a_{s'} \notin \mathcal{G}(Q_s)$  for all  $s' \in [s]$ . Thus, for all  $a \in \mathcal{G}(Q_s)$ ,  $m_a \leq m_a^{(s+1)}$ . The above analysis, together with Assumption 6B, implies

$$\sum_{a \in \mathcal{G}(Q_s)} \bar{m}_a^n \mathcal{I}_a(\boldsymbol{\theta}_n) \preceq \bar{c} \cdot (m_a^{(s+1)} + 1) \cdot \mathbf{P}_{\mathcal{G}(Q_s)} \mathcal{I}_a(\boldsymbol{\theta}_n) \preceq 2\bar{c} \cdot m_a^{(s+1)} \cdot \mathbf{P}_{\mathcal{G}(Q_s)}. \quad (95)$$

Note that  $\mathbf{A} = \sum_{a \in \mathcal{G}(Q_s)} \bar{m}_a^n \mathcal{I}_a(\boldsymbol{\theta}_n) + \sum_{a \notin \mathcal{G}(Q_s)} \bar{m}_a^n \mathcal{I}_a(\boldsymbol{\theta}_n)$ . Also note that if  $a \notin \mathcal{G}(Q_s)$ , then  $\mathcal{I}_a(\boldsymbol{\theta}_n) \in V_{Q_s}$ , which implies  $(I - \mathbf{P}_{Q_s})\mathcal{I}_a(\boldsymbol{\theta}_n) = \mathbf{0}$ . Thus, (95) further implies

$$\begin{aligned} &\|(I - \mathbf{P}_{Q_s})\mathbf{A}\|_{op} \\ &= \left\| \sum_{a \in \mathcal{G}(Q_s)} \bar{m}_a^n (I - \mathbf{P}_{Q_s})\mathcal{I}_a(\boldsymbol{\theta}_n) \right\|_{op} \\ &\leq \|(I - \mathbf{P}_{Q_s})\|_{op} \left\| \sum_{a \in \mathcal{G}(Q_s)} \bar{m}_a^n \mathcal{I}_a(\boldsymbol{\theta}_n) \right\|_{op} \\ &\leq 2\bar{c} \cdot m_a^{(s+1)}. \end{aligned}$$

The above inequality and (94) implies  $\|\mathbf{v}\| \geq \frac{1}{2\bar{c} \cdot m_a^{(s+1)}}$ . This, along with (93), implies

$$\text{tr}(\mathbf{P}_{Q_s \cup \{a'\}}\mathbf{A}^{-2}) \geq \|\mathbf{v}\|^2 \geq \frac{1}{4\bar{c}^2 \cdot (m_{a(s+1)}^n)^2}. \quad (96)$$

**Step C: Establish an upper bound for  $\text{tr}(\mathbf{P}_{Q_s}\mathbf{A}^{-2})$**

$$\text{tr}(\mathbf{A}^{-1}\mathbf{P}_{Q_s}\mathbf{A}^{-1}) \leq p \cdot \lambda_{\max}(\mathbf{A}^{-1}\mathbf{P}_{Q_s}\mathbf{A}^{-1}). \quad (97)$$

Set  $\mathbf{A}_s = \sum_{a \notin \mathcal{G}(Q_s)} \bar{m}_a^n \mathcal{I}_a(\boldsymbol{\theta}_n)$ . Let  $r = \dim(V_{Q_s})$ . We first show that  $\text{rank}(\mathbf{A}_s) = r$  and

$\mathcal{R}(\mathbf{A}_s) = V_{Q_s}$ . By Lemma 13.9,

$$\text{rank}(\mathbf{A}_s) = \dim(V_{\mathcal{A} \setminus \mathcal{G}(Q_s) \cap \{a \in \mathcal{A}; m_a^n \geq 1\}}). \quad (98)$$

Because  $Q_s \subset \mathcal{A} \setminus \mathcal{G}(Q_s) \cap \{a \in \mathcal{A}; m_a^n \geq 1\}$ , we know that

$$\dim(V_{Q_s}) \leq \dim(V_{\mathcal{A} \setminus \mathcal{G}(Q_s) \cap \{a \in \mathcal{A}; m_a^n \geq 1\}}) \leq \dim(V_{\mathcal{A} \setminus \mathcal{G}(Q_s)}). \quad (99)$$

By (89) and  $V_{Q_s} \subset V_{\mathcal{A} \setminus \mathcal{G}(Q_s)}$ , we know that

$$V_{\mathcal{A} \setminus \mathcal{G}(Q_s)} = \sum_{a \notin \mathcal{G}(Q_s)} \mathcal{R}(\mathcal{I}_a(\boldsymbol{\theta}_n)) \subset V_{Q_s} \subset V_{\mathcal{A} \setminus \mathcal{G}(Q_s)}. \quad (100)$$

Hence,  $V_{\mathcal{A} \setminus \mathcal{G}(Q_s)} = V_{Q_s}$ , and  $\dim(V_{\mathcal{A} \setminus \mathcal{G}(Q_s)}) = \dim(V_{Q_s})$ . Combined with (98) and (99), we know that

$$\text{rank}(\mathbf{A}_s) = \dim(V_{Q_s}) = \dim(V_{\mathcal{A} \setminus \mathcal{G}(Q_s) \cap \{a \in \mathcal{A}; m_a^n \geq 1\}}) = \dim(V_{\mathcal{A} \setminus \mathcal{G}(Q_s)}) = r.$$

The above analysis and

$$\mathcal{R}(\mathbf{A}_s) \subset \sum_{a \notin \mathcal{G}(Q_s)} \mathcal{R}(\mathcal{I}_a(\boldsymbol{\theta}_n)) = V_{\mathcal{A} \setminus \mathcal{G}(Q_s)} = V_{Q_s}$$

together imply that  $\mathcal{R}(\mathbf{A}_s) = V_{Q_s}$ .

Assume the eigendecomposition  $\mathbf{A}_s \hat{\mathbf{u}}_i = \hat{\lambda}_i \hat{\mathbf{u}}_i$ , and  $\mathbf{A} \mathbf{u}_i = \lambda_i \mathbf{u}_i$ ,  $1 \leq i \leq p$ , where  $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_r > \hat{\lambda}_{r+1} = \dots = \hat{\lambda}_p = 0$ ,  $\lambda_1 \geq \dots \geq \lambda_p$  with  $\hat{\mathbf{U}}_s = [\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \dots, \hat{\mathbf{u}}_r]$ ,  $\hat{\mathbf{U}}_{-s} = [\hat{\mathbf{u}}_{r+1}, \hat{\mathbf{u}}_{r+2}, \dots, \hat{\mathbf{u}}_p]$ ,  $\hat{\mathbf{U}} = [\hat{\mathbf{U}}_s, \hat{\mathbf{U}}_{-s}]$ ,  $\mathbf{U}_s = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r]$ ,  $\mathbf{U}_{-s} = [\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \dots, \mathbf{u}_p]$ , and  $\mathbf{U} = [\mathbf{U}_s, \mathbf{U}_{-s}]$ . Based on the previous notation, we know that  $\mathbf{P}_{Q_s}$  is the orthogonal projection on  $\mathcal{R}(\mathbf{A}_s)$ , and thus, it equals  $\hat{\mathbf{U}}_s \hat{\mathbf{U}}_s^T$ .

Let  $\boldsymbol{\Theta}(\hat{\mathbf{U}}_s, \mathbf{U}_s)$  denote the  $r \times r$  diagonal matrix whose  $j$ -th diagonal entry is the  $j$ -th principal angle  $\cos^{-1}(\sigma_j)$ , where  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$  are the singular values of  $\hat{\mathbf{U}}_s^T \mathbf{U}_s$ .

Applying a variant of the Davis–Kahan theorem (Theorem 2 in Yu et al. (2015)), we have

$$\left\| \sin \boldsymbol{\Theta}(\hat{\mathbf{U}}_s, \mathbf{U}_s) \right\|_F \leq \frac{2 \|\mathbf{A} - \mathbf{A}_s\|_F}{\hat{\lambda}_r - \hat{\lambda}_{r+1}} = \frac{2 \|\mathbf{A} - \mathbf{A}_s\|_F}{\hat{\lambda}_r}. \quad (101)$$

Note that

$$\mathbf{A} - \mathbf{A}_s = \sum_{a \in \mathcal{G}(Q_s)} \bar{m}_a^n \mathcal{I}_a(\boldsymbol{\theta}_n) \preceq 2\bar{c}m_{a(s+1)}^n \mathbf{P}_{\mathcal{G}(Q_s)}, \quad (102)$$

and

$$\|\mathbf{A} - \mathbf{A}_s\|_F^2 \leq 4p \cdot \bar{c}^2 (m_{a(s+1)}^n)^2. \quad (103)$$

Note that  $Q_s \subset \mathcal{A} \setminus \mathcal{G}(Q_s)$ . Because  $\mathbf{A} \succeq \mathbf{A}_s \succeq \sum_{a \in Q_s} \bar{m}_a^n \mathcal{I}_a(\boldsymbol{\theta}_n) \succeq \underline{c} m_{a(s)}^n \cdot \mathbf{P}_{Q_s}$  and  $\mathbf{A} \succeq \sum_{a \in Q_t} m_a^n \mathcal{I}_a(\boldsymbol{\theta}_n) \succeq \underline{c} m_{a(t)}^n I_p$ , by Courant–Fischer–Weyl min-max principle (see Chapter I of Hilbert and Courant (1953) or Corollary III.1.2 in Bhatia (1997)), we have

$$\lambda_r \geq \hat{\lambda}_r \geq \underline{c} m_{a(s)}^n \text{ and } \lambda_p \geq \underline{c} m_{a(t)}^n. \quad (104)$$

Combining (101), (103), and (104), we obtain

$$\left\| \sin \boldsymbol{\Theta}(\hat{\mathbf{U}}_s, \mathbf{U}_s) \right\|_F^2 \leq \frac{16p \cdot \bar{c}^2 (m_{a(s+1)}^n)^2}{\underline{c}^2 (m_{a(s)}^n)^2}. \quad (105)$$

By definition, we obtain

$$\left\| \sin \boldsymbol{\Theta}(\hat{\mathbf{U}}_s, \mathbf{U}_s) \right\|_F^2 = r - (\cos^2(\sigma_1) + \cos^2(\sigma_2) + \cdots + \cos^2(\sigma_r)) = r - \left\| \hat{\mathbf{U}}_s^T \mathbf{U}_s \right\|_F^2,$$

and

$$r = \left\| \hat{\mathbf{U}}_s^T [\mathbf{U}_s, \mathbf{U}_{-s}] \right\|_F^2 = \left\| \hat{\mathbf{U}}_s^T \mathbf{U}_s \right\|_F^2 + \left\| \hat{\mathbf{U}}_s^T \mathbf{U}_{-s} \right\|_F^2.$$

Thus,

$$\left\| \sin \boldsymbol{\Theta}(\hat{\mathbf{U}}_s, \mathbf{U}_s) \right\|_F^2 = \left\| \hat{\mathbf{U}}_s^T \mathbf{U}_{-s} \right\|_F^2. \quad (106)$$

Combining (105) and (106), we have

$$\begin{aligned} & \lambda_{\max}(\mathbf{A}^{-1} \mathbf{P}_{Q_s} \mathbf{A}^{-1}) \\ &= \lambda_{\max}(\hat{\mathbf{U}}_s^T \mathbf{A}^{-2} \hat{\mathbf{U}}_s) \\ &= \lambda_{\max}(\hat{\mathbf{U}}_s^T [\mathbf{U}_s, \mathbf{U}_{-s}] \text{diag}(\lambda_1^{-2}, \lambda_2^{-2}, \dots, \lambda_p^{-2}) [\mathbf{U}_s, \mathbf{U}_{-s}]^T \hat{\mathbf{U}}_s) \\ &= \lambda_{\max}(\hat{\mathbf{U}}_s^T \mathbf{U}_s \text{diag}(\lambda_1^{-2}, \dots, \lambda_r^{-2}) \mathbf{U}_s^T \hat{\mathbf{U}}_s + \hat{\mathbf{U}}_s^T \mathbf{U}_{-s} \text{diag}(\lambda_{r+1}^{-2}, \dots, \lambda_p^{-2}) \mathbf{U}_{-s}^T \hat{\mathbf{U}}_s) \\ &\leq \lambda_{\max}(\hat{\mathbf{U}}_s^T \mathbf{U}_s \text{diag}(\lambda_1^{-2}, \dots, \lambda_r^{-2}) \mathbf{U}_s^T \hat{\mathbf{U}}_s) + \lambda_{\max}(\hat{\mathbf{U}}_s^T \mathbf{U}_{-s} \text{diag}(\lambda_{r+1}^{-2}, \dots, \lambda_p^{-2}) \mathbf{U}_{-s}^T \hat{\mathbf{U}}_s) \quad (107) \\ &\leq \lambda_r^{-2} \left\| \hat{\mathbf{U}}_s^T \mathbf{U}_s \right\|_{op}^2 + \lambda_p^{-2} \left\| \mathbf{U}_{-s}^T \hat{\mathbf{U}}_s \right\|_{op}^2 \\ &\leq \lambda_r^{-2} + \lambda_p^{-2} \left\| \sin \boldsymbol{\Theta}(\hat{\mathbf{U}}_s, \mathbf{U}_s) \right\|_F^2 \\ &\leq \frac{1}{(\underline{c} m_{a(s)}^n)^2} \left( 1 + \frac{16p(\bar{c} m_{a(s+1)}^n)^2}{(\underline{c} m_{a(t)}^n)^2} \right). \end{aligned}$$



The above display and (97) implies

$$\text{tr}(\mathbf{A}^{-1} \mathbf{P}_{Q_s} \mathbf{A}^{-1}) \leq p \frac{1}{(\underline{c} m_{a(s)}^n)^2} \left( 1 + \frac{16p(\bar{c} m_{a(s+1)}^n)^2}{(\underline{c} m_{a(t)}^n)^2} \right). \quad (108)$$

Combining (96), (97) and (108), we have

$$\frac{\text{tr}(\mathbf{P}_{Q_s \cup \{a'\}} \mathbf{A}^{-2})}{\text{tr}(\mathbf{A}^{-1} \mathbf{P}_{Q_s} \mathbf{A}^{-1})} \geq \left\{ \frac{4\bar{c}^2 p}{\underline{c}^2} \left( 1 + \frac{16p(\bar{c} m_{a(s+1)}^n)^2}{(\underline{c} m_{a(t)}^n)^2} \right) \right\}^{-1} \left( \frac{m_{a(s)}^n}{m_{a(s+1)}^n} \right)^2 \quad (109)$$

Thus, (84) implies (92). □

**Lemma 14.8.** Assume that  $\sum_{i=1}^{n_0} \mathcal{I}_{a_i}(\boldsymbol{\theta})$  is nonsingular for some  $n_0$ .

Consider the pre-specified criteria function  $\mathbb{G}_{\boldsymbol{\theta}}(\boldsymbol{\Sigma}) = \Phi_q(\boldsymbol{\Sigma})$ . Let  $(a^{(1)}, a^{(2)}, \dots, a^{(k)})$  be a permutation of  $\mathcal{A}$  such that  $m_{a^{(1)}}^n \geq \dots \geq m_{a^{(k)}}^n$  and  $\dim(Q_{t_n-1}) < \dim(Q_{t_n}) = p$ , where  $Q_s = \{a^{(1)}, a^{(2)}, \dots, a^{(s)}\}$ , and  $t_n = t_n(m^n)$ .

For a fixed (or random) sequence  $\boldsymbol{\theta}_n \in \boldsymbol{\Theta}$  and for any  $n \geq n_0$ , we consider the generalized GI0 selection rule

$$a_{n+1} = \arg \min_{a \in \mathcal{A}} \Phi_q \left( \left\{ \frac{1}{n+1} \mathbf{A} + \frac{1}{n+1} \mathcal{I}_a(\boldsymbol{\theta}_n) \right\}^{-1} \right),$$

and GI1 selection rule

$$a_{n+1} = \arg \max_{a \in \mathcal{A}} \text{tr} \left( \mathbf{A}^{-(q+1)} \mathcal{I}_a(\boldsymbol{\theta}_n) \right), \quad (110)$$

where  $\mathbf{A} = \sum_{a \in \mathcal{A}} m_a^n \mathcal{I}_a(\boldsymbol{\theta}_n)$ . The generalized GI1 selection based on (110) coincides with the selection (81).

If for some  $1 \leq s \leq t_n - 1$

$$\left( \frac{m_{a(s)}^n}{m_{a(s+1)}^n} \right)^{q+1} \left( 1 - \frac{16p\bar{c}^2}{\underline{c}^2} \left( \frac{m_{a(s+1)}^n}{m_{a(s)}^n} \right)^2 \right) > \left( \frac{2\bar{c}}{\underline{c}} \right)^{q+2} p \left( 1 + \frac{16p\bar{c}^2}{\underline{c}^2} \left( \frac{m_{a(s+1)}^n}{m_{a(t_n)}^n} \right)^{q+1} \left( \frac{m_{a(s+1)}^n}{m_{a(s)}^n} \right)^{1-q} \right), \quad (111)$$

then

$$a_{n+1} \in \mathcal{G}(Q_s) = \{a \in \mathcal{A} : \dim(V_{Q_s \cup \{a\}}) > \dim(V_{Q_s})\}$$

for generalized GI0 and GI1, where

$$V_Q = \sum_{a \in Q} \mathcal{R}(\mathcal{I}_a(\boldsymbol{\theta}_n)).$$

*Proof of Lemma 14.8.* To prove the lemma, we follow similar steps as those in the proof of

Lemma 14.7. We will omit the repetitive details and only state the main differences.

By assumption  $\mathbb{G}_{\boldsymbol{\theta}}(\boldsymbol{\Sigma}) = \Phi_q(\boldsymbol{\Sigma})$ , Lemma 13.4 and 13.5, we know that for any  $a, a' \in \mathcal{A}$  and positive definite matrix  $\mathbf{A} \in \mathbb{R}^{p \times p}$ ,

$$\begin{aligned} \langle \nabla \Phi_q(\{\mathbf{A}/n\}^{-1}), \mathbf{A}^{-1} \mathcal{I}_a(\boldsymbol{\theta}_n) \mathbf{A}^{-1} \rangle &> \langle \nabla \Phi_q(\{\mathbf{A}/n\}^{-1}), \mathbf{A}^{-1} \mathcal{I}_{a'}(\boldsymbol{\theta}_n) \mathbf{A}^{-1} \rangle \\ \text{if and only if } \text{tr}(\mathbf{A}^{-(q+1)} \mathcal{I}_a(\boldsymbol{\theta}_n)) &> \text{tr}(\mathbf{A}^{-(q+1)} \mathcal{I}_{a'}(\boldsymbol{\theta}_n)). \end{aligned} \quad (112)$$

Thus, the generalized GI1 selection based on (110) coincides with the selection (81).

Similar to the arguments for (86), to prove the lemma, it is sufficient to show that (111) implies that for all  $a'' \notin \mathcal{G}(Q_s)$ ,  $a' \in \mathcal{G}(Q_s)$  and all  $t_1, t_2 \in [0, 1]$ ,

$$\langle \mathbf{S}_n, \mathbf{A}^{-1} \mathcal{I}_{a'}(\boldsymbol{\theta}_n) \mathbf{A}^{-1} \rangle > \langle \mathbf{S}_n, \mathbf{A}^{-1} \mathcal{I}_{a''}(\boldsymbol{\theta}_n) \mathbf{A}^{-1} \rangle, \quad (113)$$

where  $\mathbf{A}$  is redefined as  $A_2(t_1, t_2)$  and  $\mathbf{S}_n = \nabla \mathbb{G}_{\boldsymbol{\theta}_n}(\{A_2(t_1, t_2)/(n+1)\}^{-1}) = \nabla \Phi_q(\{\mathbf{A}/(n+1)\}^{-1})$ . Applying (112), we know that (113) is equivalent to

$$\text{tr}(\mathbf{A}^{-(q+1)} \mathcal{I}_{a'}(\boldsymbol{\theta}_n)) > \text{tr}(\mathbf{A}^{-(q+1)} \mathcal{I}_{a''}(\boldsymbol{\theta}_n)). \quad (114)$$

It is sufficient to show that (111) implies (114) for all  $a'' \notin \mathcal{G}(Q_s)$ ,  $a' \in \mathcal{G}(Q_s)$  and all  $t_1, t_2 \in [0, 1]$ . Similar to the proof of Lemma 14.7, this is proved using the following 3 Steps.

**Step A: Connect (114) with  $\text{tr}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s \cup \{a'\}})$  and  $\text{tr}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s})$**  Similar to the derivation leading to (88), we have

$$\text{tr}(\mathbf{A}^{-(q+1)} \mathcal{I}_{a'}(\boldsymbol{\theta}_n)) \geq \underline{c} \cdot \text{tr}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s \cup \{a'\}}) - \bar{c} \cdot \text{tr}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s}).$$

Similar to (90), we have

$$\text{tr}(\mathbf{A}^{-(q+1)} \mathcal{I}_{a''}(\boldsymbol{\theta}_n)) \leq \bar{c} \cdot \text{tr}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s}). \quad (115)$$

Thus, to prove (114), it is sufficient to show (111) implies that

$$\underline{c} \cdot \text{tr}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s \cup \{a'\}}) > 2\bar{c} \cdot \text{tr}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s}), \quad (116)$$

which is equivalent to

$$\frac{\text{tr}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s \cup \{a'\}})}{\text{tr}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s})} > \frac{2\bar{c}}{\underline{c}}. \quad (117)$$

**Step B: Establish a lower bound for  $\text{tr}(\mathbf{A}^{-(q+1)}\mathbf{P}_{Q_s \cup \{a'\}})$**  Similar to the proof of Lemma 14.7, we define  $\mathbf{A}_s = \sum_{a \in \mathcal{G}(Q_s)} \overline{m}_a^n \mathcal{I}_a(\boldsymbol{\theta}_n)$ , then  $r = \dim(V_{Q_s}) = \text{rank}(\mathbf{A}_s)$  and  $\mathcal{R}(\mathbf{A}_s) = V_{Q_s}$ . Assume the eigendecomposition  $\mathbf{A}_s \hat{\mathbf{u}}_i = \hat{\lambda}_i \hat{\mathbf{u}}_i$ , and  $\mathbf{A} \mathbf{u}_i = \lambda_i \mathbf{u}_i$ ,  $1 \leq i \leq p$ , where  $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_r > \hat{\lambda}_{r+1} = \dots = \hat{\lambda}_p = 0$ ,  $\lambda_1 \geq \dots \geq \lambda_p$  with  $\hat{\mathbf{U}}_s = [\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \dots, \hat{\mathbf{u}}_r]$ ,  $\hat{\mathbf{U}}_{-s} = [\hat{\mathbf{u}}_{r+1}, \hat{\mathbf{u}}_{r+2}, \dots, \hat{\mathbf{u}}_p]$ ,  $\hat{\mathbf{U}} = [\hat{\mathbf{U}}_s, \hat{\mathbf{U}}_{-s}]$ ,  $\mathbf{U}_s = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r]$ ,  $\mathbf{U}_{-s} = [\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \dots, \mathbf{u}_p]$ , and  $\mathbf{U} = [\mathbf{U}_s, \mathbf{U}_{-s}]$ .

Because  $a' \notin V_{Q_s}$ , there exists a unit vector  $\mathbf{u} \in \mathbb{R}^p$  such that  $\mathbf{P}_{Q_s} \mathbf{u} = \mathbf{0}$ , and  $\mathbf{P}_{Q_s \cup \{a'\}} \mathbf{u} = \mathbf{u}$ . Then,

$$\lambda_{\max}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s \cup \{a'\}}) = \lambda_{\max}(\mathbf{P}_{Q_s \cup \{a'\}} \mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s \cup \{a'\}}) \geq \mathbf{u}^T \mathbf{A}^{-(q+1)} \mathbf{u}.$$

Assume  $\mathbf{u} = \sum_{i=1}^p b_i \mathbf{u}_i = \sum_{i=1}^p \hat{b}_i \hat{\mathbf{u}}_i$ . Because  $\mathbf{P}_{Q_s} = \sum_{i=1}^r \hat{\mathbf{u}}_i^T \hat{\mathbf{u}}_i$  and

$$\mathbf{0} = \mathbf{P}_{Q_s} \mathbf{u} = \sum_{i=1}^p \hat{b}_i \mathbf{P}_{Q_s} \hat{\mathbf{u}}_i = \sum_{i=1}^r \hat{b}_i \hat{\mathbf{u}}_i,$$

we obtain that  $\hat{b}_1 = \hat{b}_2 = \dots = \hat{b}_r = 0$ . Thus, we can rewrite  $\mathbf{u}$  as  $\hat{\mathbf{U}}_{-s} \hat{\boldsymbol{\beta}}_1$  with  $\|\hat{\boldsymbol{\beta}}_1\| = 1$ .

Note that

$$\mathbf{u}^T \mathbf{A}^{-(q+1)} \mathbf{u} = \sum_{i=1}^p \lambda_i^{-(q+1)} b_i^2 \geq \lambda_{r+1}^{-(q+1)} \sum_{i=r+1}^p b_i^2 = \lambda_{r+1}^{-(q+1)} \|\mathbf{U}_{-s}^T \mathbf{u}\|^2.$$

Because  $\hat{\boldsymbol{\beta}}_1 \in \mathbb{R}^{p-r}$  and  $\|\hat{\boldsymbol{\beta}}_1\| = 1$ , we know that there exist unit vectors  $\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_3, \dots, \hat{\boldsymbol{\beta}}_{p-r}$  such that  $\hat{\boldsymbol{\beta}} = [\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \dots, \hat{\boldsymbol{\beta}}_{p-r}]$  is an orthogonal matrix. Thus, we know that

$$\begin{aligned} \|\mathbf{U}_{-s}^T \mathbf{u}\|^2 &= \|\mathbf{U}_{-s}^T \hat{\mathbf{U}}_{-s} \hat{\boldsymbol{\beta}}_1\|^2 = \|\mathbf{U}_{-s}^T \hat{\mathbf{U}}_{-s} \hat{\boldsymbol{\beta}}\|^2 - \sum_{i=2}^{p-r} \|\mathbf{U}_{-s}^T \hat{\mathbf{U}}_{-s} \hat{\boldsymbol{\beta}}_i\|^2 \\ &\geq \|\mathbf{U}_{-s}^T \hat{\mathbf{U}}_{-s}\|_F^2 - (p-r-1) = 1 - \|\sin \boldsymbol{\Theta}(\hat{\mathbf{U}}_{-s}, \mathbf{U}_{-s})\|_F^2, \end{aligned}$$

where the last equation holds because of the definition of  $\sin \boldsymbol{\Theta}(\hat{\mathbf{U}}_{-s}, \mathbf{U}_{-s})$ .

Combining the above inequalities, we obtain that

$$\lambda_{\max}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s \cup \{a'\}}) \geq \lambda_{r+1}^{-(q+1)} \cdot \left(1 - \|\sin \boldsymbol{\Theta}(\hat{\mathbf{U}}_{-s}, \mathbf{U}_{-s})\|_F^2\right). \quad (118)$$

By Weyl's inequality and (102), we know that

$$\lambda_{r+1} = |\lambda_{r+1} - \hat{\lambda}_{r+1}| \leq \|\mathbf{A} - \mathbf{A}_s\|_{op} \leq 2\bar{c}m_{a^{(s+1)}}^n.$$

Similar to how we show (105), we also have that

$$\left\| \sin \Theta(\widehat{\mathbf{U}}_{-s}, \mathbf{U}_{-s}) \right\|_F^2 \leq \frac{16p \cdot \bar{c}^2 (m_{a(s+1)}^n)^2}{\underline{c}^2 (m_{a(s)}^n)^2}. \quad (119)$$

Thus, we obtain

$$\text{tr}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s \cup \{a'\}}) \geq \lambda_{\max}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s \cup \{a'\}}) \geq (2\bar{c}m_{a(s+1)}^n)^{-q-1} \left\{ 1 - \frac{16p \cdot \bar{c}^2 (m_{a(s+1)}^n)^2}{\underline{c}^2 (m_{a(s)}^n)^2} \right\}. \quad (120)$$

**Step C: Establish an upper bound for  $\text{tr}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s})$**  Similar to (107) and according to (105), we have

$$\begin{aligned} & \lambda_{\max}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s}) \\ &= \lambda_{\max}(\widehat{\mathbf{U}}_s^T \mathbf{A}^{-(q+1)} \widehat{\mathbf{U}}_s) \\ &= \lambda_{\max}(\widehat{\mathbf{U}}_s^T [\mathbf{U}_s, \mathbf{U}_{-s}] \text{diag}(\lambda_1^{-(q+1)}, \lambda_2^{-(q+1)}, \dots, \lambda_p^{-(q+1)}) [\mathbf{U}_s, \mathbf{U}_{-s}]^T \widehat{\mathbf{U}}_s) \\ &= \lambda_{\max}(\widehat{\mathbf{U}}_s^T \mathbf{U}_s \text{diag}(\lambda_1^{-(q+1)}, \dots, \lambda_r^{-(q+1)}) \mathbf{U}_s^T \widehat{\mathbf{U}}_s + \widehat{\mathbf{U}}_s^T \mathbf{U}_{-s} \text{diag}(\lambda_{r+1}^{-(q+1)}, \dots, \lambda_p^{-(q+1)}) \mathbf{U}_{-s}^T \widehat{\mathbf{U}}_s) \\ &\leq \lambda_{\max}(\widehat{\mathbf{U}}_s^T \mathbf{U}_s \text{diag}(\lambda_1^{-(q+1)}, \dots, \lambda_r^{-(q+1)}) \mathbf{U}_s^T \widehat{\mathbf{U}}_s) + \lambda_{\max}(\widehat{\mathbf{U}}_s^T \mathbf{U}_{-s} \text{diag}(\lambda_{r+1}^{-(q+1)}, \dots, \lambda_p^{-(q+1)}) \mathbf{U}_{-s}^T \widehat{\mathbf{U}}_s) \\ &\leq \lambda_r^{-(q+1)} \left\| \widehat{\mathbf{U}}_s^T \mathbf{U}_s \right\|_{op}^2 + \lambda_p^{-(q+1)} \left\| \mathbf{U}_{-s}^T \widehat{\mathbf{U}}_s \right\|_{op}^2 \\ &\leq \lambda_r^{-(q+1)} + \lambda_p^{-(q+1)} \left\| \sin \Theta(\widehat{\mathbf{U}}_s, \mathbf{U}_s) \right\|_F^2 \\ &\leq \frac{1}{(\underline{c}m_{a(s)}^n)^{q+1}} \left( 1 + \frac{16p(\bar{c})^2 (m_{a(s+1)}^n)^2}{(\underline{c})^2 (m_{a(t_n)}^n)^{q+1} (m_{a(s)}^n)^{1-q}} \right), \end{aligned}$$

where we used  $\lambda_p \geq \underline{c}m_{a(t)}^n$  in the last inequality. Thus,

$$\text{tr}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s}) \leq p \lambda_{\max}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s}) \leq \frac{p}{(\underline{c}m_{a(s)}^n)^{q+1}} \left( 1 + \frac{16p(\bar{c})^2 (m_{a(s+1)}^n)^2}{(\underline{c})^2 (m_{a(t_n)}^n)^{q+1} (m_{a(s)}^n)^{1-q}} \right). \quad (121)$$

Combining (120) and (121), we obtain

$$\frac{\text{tr}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s \cup \{a'\}})}{\text{tr}(\mathbf{A}^{-(q+1)} \mathbf{P}_{Q_s})} \geq \left( \frac{\underline{c}}{2\bar{c}} \right)^{q+1} \frac{(m_{a(s)}^n)^{q+1}}{p(m_{a(s+1)}^n)^{q+1}} \frac{\left( 1 - \frac{16p \cdot \bar{c}^2 (m_{a(s+1)}^n)^2}{\underline{c}^2 (m_{a(s)}^n)^2} \right)}{\left( 1 + \frac{16p(\bar{c})^2 (m_{a(s+1)}^n)^2}{(\underline{c})^2 (m_{a(t_n)}^n)^{q+1} (m_{a(s)}^n)^{1-q}} \right)}. \quad (122)$$

Hence, if

$$\left(\frac{m_{a^{(s)}}^n}{m_{a^{(s+1)}}^n}\right)^{q+1} \left(1 - \frac{16p\bar{c}^2}{\underline{c}^2} \left(\frac{m_{a^{(s+1)}}^n}{m_{a^{(s)}}^n}\right)^2\right) > \left(\frac{2\bar{c}}{\underline{c}}\right)^{q+2} p \left(1 + \frac{16p\bar{c}^2}{\underline{c}^2} \left(\frac{m_{a^{(s+1)}}^n}{m_{a^{(t_n)}}^n}\right)^{q+1} \left(\frac{m_{a^{(s+1)}}^n}{m_{a^{(s)}}^n}\right)^{1-q}\right),$$

then (117) holds.  $\square$

The next lemma is useful for controlling a sequence based on some iterative inequality.

**Lemma 14.9.** *Given  $M_0 > 0, A' > 0, B', C' \geq 0$ , and  $q \geq 0$ , there exists  $M_1, M_2, \dots$  such that*

$$x^{q+1} \leq A'x^{q-1} + B'(1 + C'M_j^{q+1}x^{q-1}),$$

*then  $x \leq M_{j+1}/M_j$ . Here, each  $M_j$  depends only on  $M_0, \dots, M_{j-1}$ ,  $q$ , and  $A', B', C'$ .*

*Proof of Lemma 14.9.* Let

$$M_{j+1} := M_j \cdot \sup\{x; x^{q+1} \leq A'x^{q-1} + B'(1 + C'M_j^{q+1}x^{q-1})\}. \quad (123)$$

By induction, assume  $M_j < \infty$ . Due to

$$\lim_{x \rightarrow \infty} \frac{1}{x^{q+1}} \left( A'x^{q-1} + B'(1 + C'M_j^{q+1}x^{q-1}) \right) = 0,$$

we know that the set  $\{x; x^{q+1} \leq A'x^{q-1} + B'(1 + C'M_j^{q+1}x^{q-1})\}$  is bounded from above.

Thus,  $M_{j+1}$  defined by (123) is bounded from above. By induction, we complete the proof.  $\square$

**Lemma 14.10.** *Assume that  $\sum_{i=1}^{n_0} \mathcal{I}_{a_i}(\boldsymbol{\theta})$  is nonsingular.*

*Let  $(a_n^{(1)}, a_n^{(2)}, \dots, a_n^{(k)})$  be a permutation of  $\mathcal{A}$  such that  $m_{a_n^{(1)}}^n \geq \dots \geq m_{a_n^{(k)}}^n$  and  $\dim(Q_{t_n-1}) < \dim(Q_{t_n}) = p$ , where  $Q_{s,n} = \{a_n^{(1)}, a_n^{(2)}, \dots, a_n^{(s)}\}$ , and  $t_n = t_n(m^n)$ . To simplify the notation, let  $(a^{(1)}, a^{(2)}, \dots, a^{(k)}) = (a_n^{(1)}, a_n^{(2)}, \dots, a_n^{(k)})$  and  $Q_s = Q_{s,n}$ .*

*Also assume that the experiment selection rule satisfy the following property:*

*There exists constants  $A' \geq 0, B' > 0, C' > 0$  such that for all  $n \geq n_0$ , if for some  $1 \leq s \leq t_n - 1$ ,*

$$\left(\frac{m_{a^{(s)}}^n}{m_{a^{(s+1)}}^n}\right)^{q+1} \left(1 - A' \left(\frac{m_{a^{(s+1)}}^n}{m_{a^{(s)}}^n}\right)^2\right) > B' \left(1 + C' \left(\frac{m_{a^{(s+1)}}^n}{m_{a^{(t_n)}}^n}\right)^{q+1} \left(\frac{m_{a^{(s+1)}}^n}{m_{a^{(s)}}^n}\right)^{1-q}\right),$$

*then  $a_{n+1} \in \mathcal{G}(Q_s)$ .*

Then, this experiment selection rule also satisfies that  $\sup_{n \geq n_0} \frac{m_{a^{(1)}}^n}{m_{a^{(t_n)}}^n} < \infty$  and  $\inf_{n \geq n_0} \frac{n_I}{n_{max}} \geq C > 0$ , where  $C$  depending only on  $A', B', C', k$  and  $\frac{m_{a^{(1)}}^{n_0}}{m_{a^{(t_{n_0})}}^{n_0}}$ .

*Proof of Lemma 14.10.* Recall the definition of  $t_n = t_n(m^n)$ , there exists a permutation  $\tau_n \in \mathcal{P}$  over  $\mathcal{A}$  such that

$$\begin{aligned} m_{\tau_n(1)}^n &\geq m_{\tau_n(2)}^n \geq \cdots \geq m_{\tau_n(k)}^n, \\ a^{(1)} &= \tau_n(1), \cdots, a^{(k)} = \tau_n(k), \end{aligned}$$

and  $\dim(V_{Q_{t_n-1}}) < \dim(V_{Q_{t_n}}) = p$ . Define

$$Ind(n) = \frac{m_{a^{(1)}}^n}{m_{a^{(t_n)}}^n}.$$

To show  $\sup_{n \geq n_0} \frac{m_{a^{(1)}}^n}{m_{a^{(t_n)}}^n} < \infty$ , it is sufficient to show that if  $n \geq n_0$ ,

$$\sup_{n \geq n_0} Ind(n) < \infty.$$

Let  $M_0 = \frac{m_{a^{(1)}}^{n_0}}{m_{a^{(t_{n_0})}}^{n_0}}$ . According to the definition of  $a^{(1)}$  and  $a^{(t_{n_0})}$ , we know that  $M_0 \geq 1$ . Next, we use induction to prove that for all  $n \geq n_0$

$$Ind(n) \leq 2 \max_{0 \leq i \leq k-1} M_i, \quad (124)$$

where the sequence  $\{M_i\}_{i=1}^{k-1}$  is defined in Lemma 14.9.

For the base case, when  $n = n_0$ , we know that  $\sum_{i=1}^{n_0} \mathcal{I}_{a_i}(\boldsymbol{\theta})$  is nonsingular, and thus  $m_{a^{(t_{n_0})}}^{n_0} \geq 1$ . This implies,

$$Ind(n_0) \leq M_0 \leq m_{a^{(1)}}^{n_0} < \infty.$$

For the induction step  $n > n_0$ , assume that  $Ind(n) \leq 2 \max_{0 \leq i \leq k-1} M_i$ . We discuss two cases

Case 1: if  $a_{n+1} \in \arg \max_a m_a^n$ , then we will show that  $Ind(n) \leq \max_{0 \leq i \leq k-1} M_i$  and  $Ind(n+1) \leq 2 \max_{0 \leq i \leq k-1} M_i$ ,

Case 2: if  $a_{n+1} \notin \arg \max_a m_a^n$ , then we will show that  $Ind(n+1) \leq Ind(n)$ .

Below are the detailed analysis for these two cases.

**Case 1:**  $a_{n+1} \in \arg \max_a m_a^n$  Without loss of generality, we assume that  $\tau_n(1) = a^{(1)} = a_{n+1}$ . If  $t_n = 1$ , then  $Ind(n) = 1 \leq \max_{0 \leq i \leq k-1} M_i$ . Now, we focus on the case where  $t_n \geq 2$ .

Note that  $\tau_n(1)$  has been selected as  $a_{n+1}$ , and  $\tau_n(1) \notin \mathcal{G}(Q_1)$ . According to the lemma's assumption, we know that for all  $1 \leq s \leq t_n - 1$ ,

$$\left( \frac{m_{a(s)}^n}{m_{a(s+1)}^n} \right)^{q+1} \left( 1 - A' \left( \frac{m_{a(s+1)}^n}{m_{a(s)}^n} \right)^2 \right) \leq B' \left( 1 + C' \left( \frac{m_{a(s+1)}^n}{m_{a(t_n)}^n} \right)^{q+1} \left( \frac{m_{a(s+1)}^n}{m_{a(s)}^n} \right)^{1-q} \right). \quad (125)$$

Next, we use this inequality iteratively for  $s = t_n - 1, t_n - 2, \dots, 1$  to show that  $Ind(n) \leq \max_{0 \leq i \leq k-1} M_i$ . We start with setting  $s = t_n - 1$  in (125), we obtain that

$$x^{q+1} \leq A' x^{q-1} + B' (1 + C' M_0^{q+1} x^{q-1}),$$

where  $x = \frac{m_{a(t_n-1)}^n}{m_{a(t_n)}^n}$ . According to Lemma 14.9, this implies  $x = \frac{m_{a(t_n-1)}^n}{m_{a(t_n)}^n} \leq M_1$ .

Set  $s = t_n - 2$  in (125), and combine it with  $\frac{m_{a(t_n-1)}^n}{m_{a(t_n)}^n} \leq M_1$ , we have

$$x^{q+1} \leq A' x^{q-1} + B' (1 + C' M_1^{q+1} x^{q-1}),$$

where  $x = \frac{m_{a(t_n-2)}^n}{m_{a(t_n-1)}^n}$ . Apply Lemma 14.9 again, we obtain that  $\frac{m_{a(t_n-2)}^n}{m_{a(t_n-1)}^n} \leq M_2/M_1$ , which further implies  $\frac{m_{a(t_n-2)}^n}{m_{a(t_n)}^n} \leq M_2$ . By similar arguments, set  $s = t_n - 3, t_n - 4, \dots, 1$ , we obtain that

$$\frac{m_{a(1)}^n}{m_{a(t_n)}^n} \leq M_{t_n-1} \leq \max_{0 \leq i \leq k-1} M_i.$$

That is,  $Ind(n) \leq \max_{0 \leq i \leq k-1} M_i$ .

Note that in this case  $m_{\tau_n(1)}^{n+1} = m_{\tau_n(1)}^n + 1$ , and  $m_{\tau_n(s)}^{n+1} = m_{\tau_n(s)}^n$ , for any  $s \geq 2$ . Set  $Q = \{\tau_n(1), \tau_n(2), \dots, \tau_n(t_n)\}$ . We know that  $\dim(V_Q) = p$ . Hence,

$$\frac{\max_a m_a^{n+1}}{\min_{a \in Q} m_a^{n+1}} = \frac{m_{\tau_n(1)}^{n+1}}{m_{\tau_n(t_n)}^{n+1}} = \frac{m_{\tau_n(1)}^n + 1}{m_{\tau_n(t_n)}^n} \leq 2 \frac{m_{\tau_n(1)}^n}{m_{\tau_n(t_n)}^n} \leq 2 \max_{0 \leq i \leq k-1} M_i.$$

By Lemma 14.4, we know that

$$Ind(n+1) = \frac{m_{\tau_{n+1}(1)}^{n+1}}{m_{\tau_{n+1}(t_{n+1})}^{n+1}} \leq \frac{\max_a m_a^{n+1}}{\min_{a \in Q} m_a^{n+1}} \leq 2 \max_{0 \leq i \leq k-1} M_i.$$

**Case 2:**  $a_{n+1} \notin \arg \max_a m_a^n$  In this case,  $\max_a m_a^{n+1} = m_{\tau_n(1)}^n$  and  $m_{\tau_n(t_n)}^{n+1} \geq m_{\tau_n(t_n)}^n = \min_{a \in Q} m_a^n$ , where we let  $Q = \{\tau_n(1), \tau_n(2), \dots, \tau_n(t_n)\}$ .

Applying Lemma 14.4, we have

$$Ind(n+1) \leq \frac{\max_a m_a^{n+1}}{\min_{a \in Q} m_a^{n+1}} \leq \frac{m_{\tau_n(1)}^n}{m_{\tau_n(t_n)}^n} = Ind(n) \leq 2 \max_{0 \leq i \leq k-1} M_i,$$

where the last inequality in the above display is due to the induction assumption.

Combine the results from both cases. By induction, we have

$$\frac{m_{a(1)}^n}{m_{a(t_n)}^n} \leq 2 \max_{0 \leq i \leq k-1} M_i,$$

for all  $n \geq n_0$ . Combined with Lemma 14.3, we know that

$$\frac{n_I}{n} = \frac{m_{a(t_n)}^n}{n} \geq \frac{m_{a(t_n)}^n}{k \cdot m_{a(1)}^n} \geq \frac{1}{2k \cdot \max_{0 \leq i \leq k-1} M_i} > 0,$$

where  $k \cdot \max_{0 \leq i \leq k-1} M_{k-1}$  only depend on  $A', B', C', k$  and  $\frac{m_{a(1)}^{n_0}}{m_{a(t_{n_0})}^{n_0}}$ . □

*Proof of Theorem 14.1.* Combining Lemmas 14.7, 14.8 and 14.10, we complete the proof of

$$\inf_{n \geq n_0} \frac{n_I}{n} \geq C > 0.$$

By Assumption 6B, we know that

$$\mathcal{I}^{\bar{\pi}_n}(\boldsymbol{\theta}) \succeq \sum_{a; m_a^n \geq n_I} \frac{m_a^n}{n} \mathcal{I}_a(\boldsymbol{\theta}) \succeq \frac{n_I}{n} \sum_{a; m_a^n \geq n_I} \mathcal{I}_a(\boldsymbol{\theta}) \succeq \underline{c} \cdot C \cdot I_p. \quad (126)$$

for all  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ . □

### 14.3 Proof of Theorem 4.1

To show Theorem 4.1, we prove the following more general Theorem 14.11 instead, which applies to general experiment selection rules that are not necessarily GI0 and GI1.

**Theorem 14.11.** *Let  $U \in (0, 1)$ . Assume the experiment selection rule satisfies that  $\bar{\pi}_n(\mathbf{a}_n) \in K_U$  for large enough  $n$ , where*

$$K_U = \left\{ \boldsymbol{\pi} \in \mathcal{S}^A : \max_{S \subset A: S \text{ is relevant}} \min_{a \in S} \pi(a) \geq U \right\}. \quad (127)$$

*Here, we say that a set of experiments  $S$  is relevant if  $\sum_{a \in A} \mathcal{I}_a(\boldsymbol{\theta})$  is nonsingular for any  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ . Given Assumptions 1-4 along with either Assumptions 6A-7A or 6B-7B, the  $\hat{\boldsymbol{\theta}}_n^{ML}$  converges to  $\boldsymbol{\theta}^*$  almost surely.*

*Proof of Theorem 4.1.* According to Proposition 6.2, there exists  $U > 0$  such that (127) holds for  $n$  large enough, following GI0 or GI1. Theorem 4.1 then follows by applying



Theorem 14.11.

□

*Proof of Theorem 14.11.* Let  $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n^{\text{ML}}$  for the ease of exposition. According to (5), we know that  $l_n(\hat{\boldsymbol{\theta}}_n; \mathbf{a}_n) \geq l_n(\boldsymbol{\theta}^*; \mathbf{a}_n)$ . According to (13) in Assumption 4, with probability 1, for any  $\eta > 0$ , there exists  $N$  such that for  $n > N$ ,  $\bar{\boldsymbol{\pi}}_n \in K_U$

$$|l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) - M(\boldsymbol{\theta}^*; \bar{\boldsymbol{\pi}}_n)| \leq \frac{\eta}{3} \text{ and } |l_n(\hat{\boldsymbol{\theta}}_n; \mathbf{a}_n) - M(\hat{\boldsymbol{\theta}}_n; \bar{\boldsymbol{\pi}}_n)| \leq \frac{\eta}{3}.$$

It follows that

$$l_n(\hat{\boldsymbol{\theta}}_n; \mathbf{a}_n) \geq M(\boldsymbol{\theta}^*; \bar{\boldsymbol{\pi}}_n) - \frac{\eta}{3}.$$

Also, we have

$$M(\boldsymbol{\theta}^*; \bar{\boldsymbol{\pi}}_n) - M(\hat{\boldsymbol{\theta}}_n; \bar{\boldsymbol{\pi}}_n) \leq l_n(\hat{\boldsymbol{\theta}}_n; \mathbf{a}_n) - M(\hat{\boldsymbol{\theta}}_n; \bar{\boldsymbol{\pi}}_n) + \frac{\eta}{3} \leq \frac{2\eta}{3}.$$

That is, for  $\eta > 0$ ,

$$\mathbb{P} \left\{ \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \{M(\hat{\boldsymbol{\theta}}_n; \bar{\boldsymbol{\pi}}_n) - M(\boldsymbol{\theta}^*; \bar{\boldsymbol{\pi}}_n) \geq -\frac{2}{3}\eta\} \right\} = 1.$$

It follows that

$$\mathbb{P} \left\{ \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} \{M(\hat{\boldsymbol{\theta}}_n; \bar{\boldsymbol{\pi}}_n) - M(\boldsymbol{\theta}^*; \bar{\boldsymbol{\pi}}_n) \leq -\eta\} \right\} = 0.$$

Notice that

$$M(\boldsymbol{\theta}^*; \boldsymbol{\pi}) - M(\boldsymbol{\theta}; \boldsymbol{\pi}) = \sum_{a \in \mathcal{A}} \pi(a) D_{KL}(f_{\boldsymbol{\theta}^*, a} \| f_{\boldsymbol{\theta}, a})$$

By Assumption 7B, we can show that for any  $\boldsymbol{\pi} \in K_U$ , and any  $\varepsilon > 0$ , there exists a finite positive number  $\eta = \eta(U, \varepsilon)$ , such that

$$\sup_{\boldsymbol{\theta}: \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \geq \varepsilon} M(\boldsymbol{\theta}; \boldsymbol{\pi}) \leq M(\boldsymbol{\theta}^*; \boldsymbol{\pi}) - \eta.$$

This means that for large enough  $n$ ,

$$\left\{ \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| \geq \varepsilon \right\} \subset \left\{ M(\hat{\boldsymbol{\theta}}_n; \bar{\boldsymbol{\pi}}_n) \leq M(\boldsymbol{\theta}^*; \bar{\boldsymbol{\pi}}_n) - \eta \right\}.$$

It follows that for any  $\varepsilon > 0$ ,

$$\mathbb{P} \left( \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} \left\{ \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\| \geq \varepsilon \right\} \right) = 0.$$

Thus,

$$\begin{aligned} \mathbb{P} \left( \lim_{n \rightarrow \infty} \hat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}^* \right) &= \mathbb{P} \left( \bigcap_{l=1}^{\infty} \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \left\{ \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\| < \frac{1}{l} \right\} \right) \\ &= \lim_{l \rightarrow \infty} \mathbb{P} \left( \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \left\{ \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\| < \frac{1}{l} \right\} \right) = 1. \end{aligned}$$

□

## 14.4 Proof of Theorem 4.2

The proof of Theorem 4.2 follows the similar strategy as the proof of the classic asymptotic normality result for MLE with i.i.d. observations, which involves the asymptotic analysis of the Taylor expansion of the score equation. However, the proof for Theorem 4.2 requires the analysis of dependent stochastic processes and is more delicate.

In the following series of lemmas, we first justify the use of the score equation in Lemma 14.12. Then, we provide (almost surely) asymptotic bounds for the Hessian of the log-likelihood and the score statistic in Lemma 14.13. Lemma 14.14 provides a Taylor expansion for the score function around the true parameter and the MLE, and gives an upper bound for the remaining terms. Finally, these lemmas are combined together to obtain the proof of Theorem 4.2.

**Lemma 14.12.** *Under the setting of Theorem 14.11, if  $\boldsymbol{\theta}^* \in \text{int}(\boldsymbol{\Theta})$ , we have*

$$\mathbb{P} \left( \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \left\{ \nabla_{\boldsymbol{\theta}} l_n(\hat{\boldsymbol{\theta}}_n; \mathbf{a}_n) = \mathbf{0} \right\} \right) = 1.$$

*Proof of Lemma 14.12.* Let  $B(\boldsymbol{\theta}^*, \delta)$  denote the open ball with the center  $\boldsymbol{\theta}^*$  and radius  $\delta > 0$  such that  $B(\boldsymbol{\theta}^*, \delta) \subset \text{int}(\boldsymbol{\Theta})$ .

Because  $l_n(\boldsymbol{\theta}; \mathbf{a}_n)$  is differentiable in  $\boldsymbol{\theta}$ , we know that

$$\left\{ \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\| < \delta \right\} \subset \left\{ \nabla_{\boldsymbol{\theta}} l_n(\hat{\boldsymbol{\theta}}_n; \mathbf{a}_n) = \mathbf{0} \right\}.$$

Thus,

$$1 \geq \mathbb{P} \left( \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \{\nabla_{\boldsymbol{\theta}} l_n(\hat{\boldsymbol{\theta}}_n; \mathbf{a}_n) = \mathbf{0}\} \right) \geq \mathbb{P} \left( \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \{\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| < \delta\} \right) = 1,$$

where the last equation is due to the almost sure convergence of  $\hat{\boldsymbol{\theta}}_n$  obtained from Theorem 14.11.  $\square$

**Lemma 14.13.** *Under Assumptions 1-4, if  $\bar{\pi}_n \in K_U$  for large enough  $n$ , i.e.,*

$$\mathbb{P} \left( \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \{\bar{\pi}_n \in K_U\} \right) = 1,$$

*Also assume that the estimator  $\hat{\boldsymbol{\theta}}_n \rightarrow \boldsymbol{\theta}^*$  a.s.  $\mathbb{P}_*$ . Then, with probability 1,*

$$\limsup_{n \rightarrow \infty} \frac{1}{n-1} \sum_{j=1}^{n-1} \Psi_2^{a_j}(X_j) \leq \sum_{a=1}^k \mathbb{E}_{X \sim f_{\boldsymbol{\theta}^*, a}} \Psi_2^a(X) =: \mu_Y < \infty,$$

$$\lim_{n \rightarrow \infty} \left\| -\nabla_{\boldsymbol{\theta}}^2 l_{n-1}(\boldsymbol{\theta}^*; \mathbf{a}_{n-1}) - \mathcal{I}^{\bar{\pi}_{n-1}}(\boldsymbol{\theta}^*) \right\|_{op} = 0,$$

$$\left\| -\nabla_{\boldsymbol{\theta}}^2 l_{n-1}(\boldsymbol{\theta}; \mathbf{a}_{n-1}) + \nabla_{\boldsymbol{\theta}}^2 l_{n-1}(\boldsymbol{\theta}^*; \mathbf{a}_{n-1}) \right\|_{op} \leq \frac{1}{n-1} \sum_{i=1}^{n-1} \Psi_2^{a_i}(X_i) \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|,$$

$$\limsup_{n \rightarrow \infty} \left\| (-\nabla_{\boldsymbol{\theta}}^2 l_{n-1}(\hat{\boldsymbol{\theta}}_{n-2}; \mathbf{a}_{n-1}))^{-1} \right\|_{op} \leq \frac{1}{\min_{\boldsymbol{\pi} \in K_U} \lambda_{\min}(\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*))} < \infty,$$

$$\limsup_{n \rightarrow \infty} \left\| (-\nabla_{\boldsymbol{\theta}}^2 l_{n-2}(\hat{\boldsymbol{\theta}}_{n-2}; \mathbf{a}_{n-2}))^{-1} \right\|_{op} \leq \frac{1}{\min_{\boldsymbol{\pi} \in K_U} \lambda_{\min}(\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*))} < \infty,$$

$$\limsup_{n \rightarrow \infty} \left\| (-\nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*; \mathbf{a}_n))^{-1} \right\|_{op} \leq \frac{1}{\min_{\boldsymbol{\pi} \in K_U} \lambda_{\min}(\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*))} < \infty.$$

*Proof of Lemma 14.13.* Let the information filtration be

$$\mathcal{F}_n = \sigma(\{a_1, X_1, \dots, a_n, X_n\}).$$

In the rest of the proof, we restrict the analysis to the event  $\bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \{\bar{\pi}_n \in K_U\}$ , which has probability 1 by the assumption. Applying Lemma 13.2 on each entry of  $-\nabla_{\boldsymbol{\theta}}^2 \log f_{\boldsymbol{\theta}^*, a_i}(X_i)$ , and note that  $\mathbb{E}(\nabla_{\boldsymbol{\theta}}^2 \log f_{\boldsymbol{\theta}^*, a_i}(X_i) | \mathcal{F}_{i-1}) = -\mathcal{I}_{a_i}(\boldsymbol{\theta}^*)$ , we obtain

$$-\nabla_{\boldsymbol{\theta}}^2 l_{n-1}(\boldsymbol{\theta}^*; \mathbf{a}_{n-1}) - \mathcal{I}^{\bar{\pi}_{n-1}}(\boldsymbol{\theta}^*) = \frac{1}{n-1} \sum_{i=1}^{n-1} (-\nabla_{\boldsymbol{\theta}}^2 \log f_{\boldsymbol{\theta}^*, a_i}(X_i) - \mathcal{I}_{a_i}(\boldsymbol{\theta}^*)) \xrightarrow{\text{a.s.}} 0. \quad (128)$$

By Assumption 2 and the relaxed condition (15), we know that

$$\left\| -\nabla_{\boldsymbol{\theta}}^2 l_{n-1}(\boldsymbol{\theta}; \mathbf{a}_{n-1}) + \nabla_{\boldsymbol{\theta}}^2 l_{n-1}(\boldsymbol{\theta}^*; \mathbf{a}_{n-1}) \right\|_{op} \leq \frac{1}{n-1} \sum_{i=1}^{n-1} \Psi_2^{a_i}(X_i) \psi(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|).$$

Let  $\{X_j^a\}_{1 \leq j \leq n, a \in \mathcal{A}}$  be a sequence of independent random variables such that  $X_j^a \sim f_{\boldsymbol{\theta}^*, a}$  for all  $j \geq 1$  and  $a \in \mathcal{A}$ . By Lemma 13.8, we can replace  $(X_1, X_2, \dots, X_n)$  with  $(X_1^{a_1}, X_2^{a_2}, \dots, X_n^{a_n})$  without changing the joint distribution for all  $n$ .

Let  $Y_i = \sum_{a \in \mathcal{A}} \Psi_2^a(X_i^a)$ . We know that  $\{Y_i\}_{i=1}^\infty$  are i.i.d. and by Assumption 2,  $\mu_Y := \mathbb{E}_{\boldsymbol{\theta}^*} Y_1 < \infty$ .

The strong law of large numbers (see Theorem 2.1 in Ross (2014)) implies that with probability 1,

$$\frac{1}{n} \sum_{j=1}^n Y_j \rightarrow \mu_Y.$$

Thus, with probability 1

$$\limsup_{n \rightarrow \infty} \frac{1}{n-1} \sum_{j=1}^{n-1} \Psi_2^{a_j}(X_j^{a_j}) \leq \limsup_{n \rightarrow \infty} \frac{1}{n-1} \sum_{j=1}^{n-1} \sum_{a \in \mathcal{A}} \Psi_2^a(X_j^a) = \sum_{a=1}^k \mathbb{E}_{X \sim f_{\boldsymbol{\theta}^*, a}} \Psi_2^a(X) = \mu_Y < \infty.$$

Set  $A_n = \mathcal{I}^{\bar{\pi}^{n-1}}(\boldsymbol{\theta}^*)$ , and  $\Delta A_n = -\nabla^2 l_{n-1}(\hat{\boldsymbol{\theta}}_{n-2}; \mathbf{a}_{n-1}) - A_n$  for all  $n \geq 1$ . Notice that

$$A_n^{-1} - (A_n + \Delta A_n)^{-1} \Delta A_n A_n^{-1} = (A_n + \Delta A_n)^{-1} (A_n + \Delta A_n - \Delta A_n) A_n^{-1} = (A_n + \Delta A_n)^{-1},$$

$$\|(A_n + \Delta A_n)^{-1} - A_n^{-1}\|_{op} = \|(A_n + \Delta A_n)^{-1} \Delta A_n A_n^{-1}\|_{op} \leq \|(A_n + \Delta A_n)^{-1}\|_{op} \|\Delta A_n\|_{op} \|A_n^{-1}\|_{op},$$

as well as

$$\|A_n^{-1}\|_{op} \leq \frac{1}{\min_{\boldsymbol{\pi} \in K_U} \lambda_{\min}(\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*))} < \infty,$$

for any  $n$ . Furthermore,

$$\|(A_n + \Delta A_n)^{-1}\|_{op} \leq \|A_n^{-1}\|_{op} + \|(A_n + \Delta A_n)^{-1}\|_{op} \|A_n^{-1}\|_{op} \|\Delta A_n\|_{op},$$

which implies

$$\|(A_n + \Delta A_n)^{-1}\|_{op} \leq \frac{\|A_n^{-1}\|_{op}}{1 - \|A_n^{-1}\|_{op} \|\Delta A_n\|_{op}}$$

given that  $\|A_n^{-1}\|_{op} \|\Delta A_n\|_{op} < 1$ . Note that

$$\|\Delta A_n\|_{op} \leq \frac{1}{n-1} \sum_{j=1}^{n-1} \Psi_2^{a_j}(X_j) \left\| \hat{\boldsymbol{\theta}}_{n-2} - \boldsymbol{\theta}^* \right\| + \left\| -\nabla^2 l_{n-1}(\boldsymbol{\theta}^*) - \mathcal{I}^{\pi_{n-1}}(\boldsymbol{\theta}^*) \right\|_{op}.$$

The first term on the right-hand side of the above inequality converges to 0 a.s., because of the almost sure convergence assumption on  $\hat{\boldsymbol{\theta}}_n$ , and the second term converges to 0 a.s. because of (128). Consequently,  $\|\Delta A_n\|_{op} \xrightarrow{\text{a.s.}} 0$ . This further implies that, for  $n$  large enough,  $\|\Delta A_n\|_{op} \leq \frac{1}{2} \min_{\boldsymbol{\pi} \in K_U} \lambda_{\min}(\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*))$ . For such  $n$ , we have

$$\left\| (A_n + \Delta A_n)^{-1} - A_n^{-1} \right\|_{op} \leq \frac{\|\Delta A_n\|_{op} \|A_n^{-1}\|_{op}^2}{1 - \|\Delta A_n\|_{op} \|A_n^{-1}\|_{op}} \xrightarrow{\text{a.s.}} 0.$$

Note that  $A_n + \Delta A_n = -\nabla^2 l_{n-1}(\hat{\boldsymbol{\theta}}_{n-2}; \mathbf{a}_{n-1})$ . Thus, with probability 1,

$$\limsup_{n \rightarrow \infty} \left\| (-\nabla_{\boldsymbol{\theta}}^2 l_{n-1}(\hat{\boldsymbol{\theta}}_{n-2}; \mathbf{a}_{n-1}))^{-1} \right\|_{op} \leq \limsup_{n \rightarrow \infty} \|A_n^{-1}\|_{op} \leq \frac{1}{\min_{\boldsymbol{\pi} \in K_U} \lambda_{\min}(\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*))} < \infty.$$

Similarly, we can also show that with probability 1,

$$\limsup_{n \rightarrow \infty} \left\| (-\nabla_{\boldsymbol{\theta}}^2 l_{n-2}(\hat{\boldsymbol{\theta}}_{n-2}; \mathbf{a}_{n-2}))^{-1} \right\|_{op} \leq \frac{1}{\min_{\boldsymbol{\pi} \in K_U} \lambda_{\min}(\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*))} < \infty,$$

as well as

$$\limsup_{n \rightarrow \infty} \left\| (-\nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*; \mathbf{a}_n))^{-1} \right\|_{op} \leq \frac{1}{\min_{\boldsymbol{\pi} \in K_U} \lambda_{\min}(\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*))} < \infty.$$

□

**Lemma 14.14.** *Under Assumptions 1-4, the Taylor expansion for the score function  $\nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}; \mathbf{a}_n)$  is given by*

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^* + \mathbf{W}_n/\sqrt{n}; \mathbf{a}_n) &= \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) + \nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) \mathbf{W}_n/\sqrt{n} + R(\boldsymbol{\theta}^*, \mathbf{W}_n), \\ \nabla_{\boldsymbol{\theta}} l_n(\hat{\boldsymbol{\theta}}_{n'}; \mathbf{a}_n) &= \nabla_{\boldsymbol{\theta}} l_n(\hat{\boldsymbol{\theta}}_n; \mathbf{a}_n) + \nabla_{\boldsymbol{\theta}}^2 l_n(\hat{\boldsymbol{\theta}}_n; \mathbf{a}_n) (\hat{\boldsymbol{\theta}}_{n'} - \hat{\boldsymbol{\theta}}_n) + R'(\hat{\boldsymbol{\theta}}_{n'}, \hat{\boldsymbol{\theta}}_n), \end{aligned} \quad (129)$$

where  $\hat{\boldsymbol{\theta}}_n$  and  $\hat{\boldsymbol{\theta}}_{n'}$  are the MLE based on  $l_n(\boldsymbol{\theta}; \mathbf{a}_n)$  and  $l_{n'}(\boldsymbol{\theta}; \mathbf{a}_{n'})$ , respectively,  $\mathbf{W}_n = \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)$ ,

$$\|R(\boldsymbol{\theta}^*, \mathbf{W}_n)\| \leq \frac{1}{n} \sum_{j=1}^n \Psi_2^{a_j}(X_j) \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\| \psi\left(\left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\|\right), \text{ and}$$

$$\left\| R'(\hat{\boldsymbol{\theta}}_{n'}, \hat{\boldsymbol{\theta}}_n) \right\| \leq \frac{1}{n} \sum_{j=1}^n \Psi_2^{a_j}(X_j) \left\| \hat{\boldsymbol{\theta}}_{n'} - \hat{\boldsymbol{\theta}}_n \right\| \psi\left(\left\| \hat{\boldsymbol{\theta}}_{n'} - \hat{\boldsymbol{\theta}}_n \right\|\right).$$

*Proof of Lemma 14.14.* Set  $g(t) = \langle \mathbf{b}, \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^* + t(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*); \mathbf{a}_n) \rangle$ . By the Lagrange mean value theorem, there exists  $0 < t^* < 1$  such that

$$g(1) - g(0) = g'(t^*),$$

i.e.,

$$\langle \mathbf{b}, \nabla_{\boldsymbol{\theta}} l_n(\widehat{\boldsymbol{\theta}}_n; \mathbf{a}_n) - \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) \rangle = \langle \mathbf{b}, \nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^* + t^*(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*); \mathbf{a}_n)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \rangle.$$

Then,

$$\langle \mathbf{b}, R(\boldsymbol{\theta}^*, \mathbf{W}_n) \rangle = \langle \mathbf{b}, \left\{ \nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^* + t^*(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*); \mathbf{a}_n) - \nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) \right\} (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \rangle. \quad (130)$$

Under Assumption 2, we have

$$\begin{aligned} \|R(\boldsymbol{\theta}^*, \mathbf{W}_n)\| &= \sup_{\|\mathbf{b}\| \leq 1} \langle \mathbf{b}, R(\boldsymbol{\theta}^*, \mathbf{W}_n) \rangle \\ &\leq \max_{0 \leq t^* \leq 1} \left\| \nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^* + t^*(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*); \mathbf{a}_n) - \nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) \right\|_{op} \left\| \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\| \\ &\leq \frac{1}{n} \sum_{j=1}^n \Psi_2^{a_j}(X_j) \left\| \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\| \psi \left( \left\| \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\| \right). \end{aligned}$$

Similarly, we can show that

$$\langle \mathbf{b}, R'(\widehat{\boldsymbol{\theta}}_{n'}, \widehat{\boldsymbol{\theta}}_n) \rangle = \langle \mathbf{b}, \left\{ \nabla_{\boldsymbol{\theta}}^2 l_n(\widehat{\boldsymbol{\theta}}_n + t^*(\widehat{\boldsymbol{\theta}}_{n'} - \widehat{\boldsymbol{\theta}}_n); \mathbf{a}_n) - \nabla_{\boldsymbol{\theta}}^2 l_n(\widehat{\boldsymbol{\theta}}_n; \mathbf{a}_n) \right\} (\widehat{\boldsymbol{\theta}}_{n'} - \widehat{\boldsymbol{\theta}}_n) \rangle.$$

Under Assumption 2, we have

$$\begin{aligned} &\left\| R'(\widehat{\boldsymbol{\theta}}_{n'}, \widehat{\boldsymbol{\theta}}_n) \right\| \\ &= \sup_{\|\mathbf{b}\| \leq 1} \langle \mathbf{b}, R'(\widehat{\boldsymbol{\theta}}_{n'}, \widehat{\boldsymbol{\theta}}_n) \rangle \\ &\leq \max_{0 \leq t^* \leq 1} \left\| \nabla_{\boldsymbol{\theta}}^2 l_n(\widehat{\boldsymbol{\theta}}_n + t^*(\widehat{\boldsymbol{\theta}}_{n'} - \widehat{\boldsymbol{\theta}}_n); \mathbf{a}_n) - \nabla_{\boldsymbol{\theta}}^2 l_n(\widehat{\boldsymbol{\theta}}_n; \mathbf{a}_n) \right\|_{op} \left\| \widehat{\boldsymbol{\theta}}_{n'} - \widehat{\boldsymbol{\theta}}_n \right\| \\ &\leq \frac{1}{n} \sum_{j=1}^n \Psi_2^{a_j}(X_j) \left\| \widehat{\boldsymbol{\theta}}_{n'} - \widehat{\boldsymbol{\theta}}_n \right\| \psi \left( \left\| \widehat{\boldsymbol{\theta}}_{n'} - \widehat{\boldsymbol{\theta}}_n \right\| \right). \end{aligned}$$

□

*Proof of Theorem 4.2.* Write  $\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n^{\text{ML}}$  for the ease of exposition. By Theorem 4.1, we know that  $\widehat{\boldsymbol{\theta}}_n$  converges to  $\boldsymbol{\theta}^*$  almost surely. By Lemma 14.12, with probability 1, there exists

random integer  $N < \infty$  such that for any  $n \geq N$ ,  $\nabla_{\boldsymbol{\theta}} l_n(\hat{\boldsymbol{\theta}}_n; \mathbf{a}_n) = \mathbf{0}$ . Let  $\mathbf{W}_n = \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)$ .

Recall the remainder function defined in Lemma 14.14,

$$R(\boldsymbol{\theta}^*, \mathbf{W}_n) := \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^* + \mathbf{W}_n/\sqrt{n}; \mathbf{a}_n) - \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) - \nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) \mathbf{W}_n/\sqrt{n}.$$

With  $\nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^* + \mathbf{W}_n/\sqrt{n}; \mathbf{a}_n) = \mathbf{0}$  provided  $n \geq N$  in mind, we can write  $\mathbf{W}_n$

$$\mathbf{W}_n = -\{\nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*; \mathbf{a}_n)\}^{-1} \{\sqrt{n} \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) + \sqrt{n} R(\boldsymbol{\theta}^*, \mathbf{W}_n)\}. \quad (131)$$

The rest of the proof consists of three parts: in Part I, we show that  $\sqrt{n} \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) \xrightarrow{d} N(0, \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}^*))$ ; in Part II, we show that  $\nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) \xrightarrow{\mathbb{P}^*} -\sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}^*)$ ; and in Part III, we show that  $\sqrt{n} R(\boldsymbol{\theta}^*, \mathbf{W}_n) = o_p(1)$ .

**Part I: Show that  $\sqrt{n} \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) \xrightarrow{d} N(0, \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}^*))$  as  $n \rightarrow \infty$**  Let  $\mathbf{b}$  be any constant vector in  $\mathbb{R}^p$  with  $\|\mathbf{b}\| = 1$ . For  $i = 1, \dots, n$ , let

$$\xi_{n,i} := \frac{1}{\sqrt{n}} \mathbf{b}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*, a_i}(X_i).$$

Set  $\mathcal{F}_i = \sigma\{a_1, X_1, a_2, X_2, \dots, a_i, X_i\}$  for any  $i \geq 1$  and  $\mathcal{F}_0$  denote the trivial  $\sigma$ -algebra. Applying the Dominated Convergence Theorem, coupled with the classical proof of differentiation under the integral sign, we arrive at the conclusion that  $\mathbb{E}(\xi_{n,i} | \mathcal{F}_{i-1}) = 0$ , which implies that  $\mathbb{E}(\xi_{n,i}) = 0$ . Denote  $\sigma_{n,i}^2 := \mathbb{E}(\xi_{n,i}^2 | \mathcal{F}_{i-1}) = \frac{1}{n} \mathbf{b}^T \mathcal{I}_{a_i}(\boldsymbol{\theta}^*) \mathbf{b}$ .

Let  $S_n := \sum_{i=1}^n \xi_{n,i}$ . Note that  $\mathbb{E}(S_n) = 0$  and  $\mathbb{E}(S_n^2) < \infty$ , since  $\mathbb{E}(\xi_{n,i}^2) < \infty$  for all  $i$ . Then,  $\{S_n, \mathcal{F}_n\}_{n \geq 1}$  is a martingale array with mean 0 and finite variance. We will apply the martingale central limit theorem to  $S_n$ . We check the conditions first.

We first check the conditional variance condition. We write

$$\sum_{i=1}^n \sigma_{n,i}^2 = \mathbf{b}^T \left\{ \frac{1}{n} \sum_{i=1}^n \mathcal{I}_{a_i}(\boldsymbol{\theta}^*) \right\} \mathbf{b} = \mathbf{b}^T \left\{ \sum_{a \in \mathcal{A}} \bar{\pi}_n(a) \mathcal{I}_a(\boldsymbol{\theta}^*) \right\} \mathbf{b}.$$

Due to the convergence assumption of  $\bar{\pi}_n$ , we have

$$\mathbf{b}^T \left\{ \sum_{a \in \mathcal{A}} \bar{\pi}_n(a) \mathcal{I}_a(\boldsymbol{\theta}^*) \right\} \mathbf{b} \xrightarrow{\mathbb{P}^*} \mathbf{b}^T \left\{ \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}^*) \right\} \mathbf{b}.$$

Then the conditional variance condition holds.

We then check the conditional Lindeberg's condition. Assume random variables  $\{X^a\}_{a \in \mathcal{A}}$

have densities  $\{f_{\boldsymbol{\theta}^*,a}\}_{a \in \mathcal{A}}$ , respectively. For any  $\varepsilon > 0$ , with probability 1,

$$\sum_{i=1}^n \mathbb{E} \left\{ \xi_{n,i}^2 I(|\xi_{n,i}| > \varepsilon) | \mathcal{F}_{i-1} \right\} \leq \sum_{a=1}^k \mathbb{E} \left\{ \|\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*,a}(X^a)\|^2 I(\|\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*,a}(X^a)\| > \sqrt{n}\varepsilon) \right\}.$$

By Assumption 2,

$$\lim_{n \rightarrow \infty} \mathbb{E} \left\{ \|\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*,a}(X^a)\|^2 I(\|\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*,a}(X^a)\| > \sqrt{n}\varepsilon) \right\} = 0.$$

Thus, the conditional Lindeberg condition holds.

By the Martingale Central Limit Theorem (Corollary 3.1 in Hall and Heyde (1980)), we have

$$\sum_{i=1}^n \xi_{n,i} \xrightarrow{d} N \left( 0, \mathbf{b}^T \left\{ \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}^*) \right\} \mathbf{b} \right).$$

It follows by Cramér–Wold theorem (see Billingsley (1999) p383) that

$$\sqrt{n} \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) \xrightarrow{d} N \left( 0, \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}^*) \right). \quad (132)$$

**Part II: Show that**  $\nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) \xrightarrow{\mathbb{P}^*} -\sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}^*)$  For each  $i = 1, \dots, n$ , by Assumption A3, we have

$$\mathbb{E} \left\{ \nabla_{\boldsymbol{\theta}}^2 \log f_{\boldsymbol{\theta}^*,a_i}(X_i) | \mathcal{F}_{i-1} \right\} = -\mathcal{I}_{a_i}(\boldsymbol{\theta}^*).$$

Also, the conditional expectation has

$$\frac{1}{n} \sum_{i=1}^n \mathcal{I}_{a_i}(\boldsymbol{\theta}^*) = \mathcal{I}^{\bar{\pi}_n}(\boldsymbol{\theta}^*) \xrightarrow{\mathbb{P}^*} \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}^*),$$

due to the convergence assumption of  $\bar{\pi}_n$ .

For each  $i, l = 1, \dots, p$ , define

$$G_{i,l} = \sum_{a \in \mathcal{A}, X^a \sim f_{\boldsymbol{\theta}^*,a}(\cdot)} \left| (\nabla_{\boldsymbol{\theta}}^2 \log f_{\boldsymbol{\theta}^*,a}(X^a))_{i,l} \right|.$$

Then, for all  $x \geq 0$  and  $i, l \geq 1$ ,

$$\mathbb{P} \left\{ \left| (\nabla_{\boldsymbol{\theta}}^2 \log f_{\boldsymbol{\theta}^*,a}(X^a))_{i,l} \right| > x \right\} \leq \mathbb{P} \{ G_{i,l} > x \}$$



This implies that  $\sum_{a \in \mathcal{A}} \mathbb{E} \left( \left| (\nabla_{\boldsymbol{\theta}}^2 \log f_{\boldsymbol{\theta}^*, a}(X^a))_{i,l} \right| \right) \leq \sum_{i,l} \mathbb{E}(G_{i,l}) < \infty$  under Assumption 2.

By Lemma 13.2 and the Slutsky's theorem, we arrive at

$$\frac{1}{n} \sum_{i=1}^n \nabla_{\boldsymbol{\theta}}^2 \log f_{\boldsymbol{\theta}^*, a_i}(X_i) = \nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) \xrightarrow{\mathbb{P}_*} - \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}^*). \quad (133)$$

**Part III: Show that  $\sqrt{n}R(\boldsymbol{\theta}^*, \mathbf{W}_n) = o_p(1)$**  According to Lemma 14.13 and Lemma 14.14, we know that

$$\|\sqrt{n}R(\boldsymbol{\theta}^*, \mathbf{W}_n)\| \leq \frac{1}{n} \sum_{j=1}^n \Psi_2^{a_j}(X_j) \|\mathbf{W}_n\| \psi \left( \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\| \right). \quad (134)$$

By (131), we have

$$\|\mathbf{W}_n\| \leq \left\| \{ \nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) \}^{-1} \right\|_{op} \left\{ \|\sqrt{n} \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^*; \mathbf{a}_n)\| + \|\sqrt{n}R(\boldsymbol{\theta}^*, \mathbf{W}_n)\| \right\}. \quad (135)$$

The above two inequalities together implies

$$\|\sqrt{n}R(\boldsymbol{\theta}^*, \mathbf{W}_n)\| \leq \frac{\frac{1}{n} \sum_{j=1}^n \Psi_2^{a_j}(X_j) \psi \left( \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\| \right) \left\| \{ \nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*) \}^{-1} \right\|_{op} \|\sqrt{n} \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^*)\|}{1 - \frac{1}{n} \sum_{j=1}^n \Psi_2^{a_j}(X_j) \left\| \{ \nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*) \}^{-1} \right\|_{op} \psi \left( \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\| \right)} \quad (136)$$

given that  $\frac{1}{n} \sum_{j=1}^n \Psi_2^{a_j}(X_j) \left\| \{ \nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*) \}^{-1} \right\|_{op} \psi \left( \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\| \right) < 1$ . We have shown in (133) in Part II that

$$\nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) = - \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}^*) + o_p(1),$$

and thus

$$\left\| \{ \nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) \}^{-1} \right\|_{op} = \lambda_{\min}^{-1} \left( \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}^*) \right) + o_p(1). \quad (137)$$

Furthermore, under Assumption 5, by the consistency result in Theorem 4.1 and Lemma 14.13, we have

$$\left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\| = o_p(1),$$

which implies that

$$\frac{1}{n} \sum_{j=1}^n \Psi_2^{a_j}(X_j) \left\| \{ \nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*) \}^{-1} \right\|_{op} \psi \left( \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\| \right) = o_p(1). \quad (138)$$

Let the event  $D_n := \left\{1 - \frac{1}{n} \sum_{j=1}^n \Psi_2^{a_j}(X_j) \left\| \{\nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*)\}^{-1} \right\|_{op} \psi \left( \left\| \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\| \right) > \frac{1}{2} \right\}$ . We have

$$\mathbb{P}(D_n) \rightarrow 1, \text{ as } n \rightarrow \infty.$$

On the event  $D_n$ , according to (132), we have

$$\left\| \sqrt{n} \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^*) \right\| = O_p(1),$$

which together with (136), (138) and Lemma 14.13 yields  $\|\mathbf{W}_n\| = O_p(1)$ . It follows from (134) that

$$\left\| \sqrt{n} R(\boldsymbol{\theta}^*, \mathbf{W}_n) \right\| = o_p(1).$$

Therefore, applying Slutsky's Theorem and the continuous mapping Theorem to (131), we have

$$\mathbf{W}_n \xrightarrow{d} N \left( 0, \left( \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}^*) \right)^{-1} \right),$$

which concludes the proof.  $\square$

## 14.5 Proof of Theorem 4.3

We first present an extension of the classic convergence theorem by Robbins and Siegmund (1971), which is frequently employed to prove convergence of stochastic processes within the fields of stochastic approximation and reinforcement learning. It provides conditions on a stochastic process  $\{Z_n\}$  for it to converge almost surely. The following modified version of the Robbins-Siegmund Theorem allows us to obtain a better estimate of the convergence rate of  $\{Z_n\}$ . Later in this section, we will apply this result to  $Z_n = \mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_n) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*)$  for proving Theorem 4.3.

**Lemma 14.15** (Modified Robbins-Siegmund Theorem). *Let  $a_n, c_n$  be integrable random variables and  $Z_n$  be a non-negative integrable random variable adaptive to filtration  $\mathcal{F}_n$  for all  $n \geq 1$ , and  $\mathcal{F}_1 \subset \mathcal{F}_2 \cdots$ . Assume that*

$$\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] \leq (1 - a_n) Z_n + c_n, \text{ for all } n \geq 1. \quad (139)$$

*Set  $a_n^- = \max\{0, -a_n\}$  and  $a_n^+ = \max\{0, a_n\}$ . Assume*

$$\sum_{n=1}^{\infty} a_n^- < \infty.$$

Then, the following statements hold.

1. If

$$\sum_{n=1}^{\infty} c_n \text{ exists with probability 1,} \quad (140)$$

then there exists non-negative random variable  $Z_{\infty}$  such that  $\lim_{n \rightarrow \infty} Z_n = Z_{\infty}$  with probability 1.

2. If we assume (140) holds and further require

$$\{a_n^+ Z_n\}_{n=1}^{\infty} \text{ are all intergrable, and } \sum_{n=1}^{\infty} a_n = +\infty \text{ with probability 1,} \quad (141)$$

then  $\lim_{n \rightarrow \infty} Z_n = 0$  with probability 1.

3. Assume (141) holds. If there exists  $0 < \beta < c$  such that  $a_n \geq \frac{c}{n}$  and the limit  $\sum_{n=1}^{\infty} n^{\beta} c_n$  exists with probability 1, then  $\lim_{n \rightarrow \infty} n^{\beta} Z_n = 0$  with probability 1.

*Proof of Lemma 14.15.*

**Part 1** First of all, (139) implies

$$\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] \leq (1 + a_n^-) Z_n + c_n.$$

Set  $Z'_n = \frac{Z_n}{\prod_{i=1}^{n-1} (1 + a_i^-)}$ , and  $c'_n = \frac{c_n}{\prod_{i=1}^n (1 + a_i^-)}$ . Notice that  $\sum_{n=1}^{\infty} a_n^- < \infty$  implies  $\prod_{n=1}^{\infty} (1 + a_n^-) < \infty$ . By Abel's test for series (see Exercise 9.15 in Ghorpade and Limaye (2006)), we know that

$$\mathbb{P}\left(\sum_{n=1}^{\infty} c'_n \text{ exists}\right) = 1, \quad (142)$$

Because  $|c'_n| \leq |c_n|$ ,  $0 \leq Z'_n \leq Z_n$  as well as  $c_n$  and  $Z_n$  are integrable, we know that  $c'_n$  and  $Z'_n$  are also integrable. Note that

$$\mathbb{E}[Z'_{n+1} \mid \mathcal{F}_n] \leq Z'_n + c'_n. \quad (143)$$

Let  $Y_1 = Z'_1$  and  $Y_n = Z'_n - (c'_1 + \cdots + c'_{n-1})$ , which are integrable for all  $n \geq 2$ . We know that  $Y_n$  is integrable for all  $n \geq 1$ . By (143), we obtain

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] \leq Y_n. \quad (144)$$

Let  $\tau_T = \inf\{n; \sum_{k=1}^n c'_k > T\}$ , for any  $T \geq 0$ . Note that  $\{\tau_T > n\} = \{\sum_{i=1}^n c'_i \leq T\} \in \mathcal{F}_n$ . Because

$$Y_{n \wedge \tau_T} = \sum_{l=1}^n Y_l \cdot I(\tau_T = l) + Y_n \cdot I(\tau_T > n), \text{ and } |Y_{n \wedge \tau_T}| \leq \sum_{i=1}^n |Y_i|,$$

we obtain that  $Y_{n \wedge \tau_T}$  is integrable for all  $n \geq 1$ .

By the definition of  $\tau_T$ , we know that  $i \leq \tau_T - 1 \implies \sum_{k=1}^i c'_k \leq T$ , which implies that for any  $n \geq 1$

$$Y_{n \wedge \tau_T} \geq - \sum_{k=1}^{n \wedge \tau_T - 1} c'_k \geq -T.$$

By (144), we know that for any  $n \geq 1$ ,

$$\begin{aligned} \mathbb{E}[Y_{(n+1) \wedge \tau_T} | \mathcal{F}_n] &= Y_{\tau_T} I(\tau_T \leq n) + \mathbb{E}[Y_{n+1} | \mathcal{F}_n] I(\tau_T > n) \\ &\leq \sum_{l=1}^n Y_{\tau_T} I(\tau_T = l) + Y_n I(\tau_T > n) = Y_{n \wedge \tau_T}, \text{ and} \end{aligned}$$

$$0 \leq \mathbb{E}[Y_{n \wedge \tau_T} + T] \leq \dots \leq \mathbb{E}[Y_{1 \wedge \tau_T} + T] = \mathbb{E}Z'_1 + T < \infty.$$

This concludes that  $Y_{n \wedge \tau_T} + T$  is a non-negative supermartingale (see Section 1.1 in Hall and Heyde (1980)). Applying Doob's convergence theorem (see Theorem 2.5 in Hall and Heyde (1980)) to  $L^1$  uniformly bounded submartingale  $-(Y_{n \wedge \tau_T} + T)$ , we know that  $\lim_{n \rightarrow \infty} Y_{n \wedge \tau_T}$  exists and is finite for any  $T \geq 0$ .

In conclusion,  $\lim_{n \rightarrow \infty} Y_n$  exists and is finite almost surely on event

$$\{\tau_T = \infty\} = \left\{ \sum_{i=1}^n c'_i \leq T \text{ for any } n \geq 1 \right\} \text{ for any } T \geq 0.$$

Combining this with (142), we know that  $\lim_{n \rightarrow \infty} Y_n$  exists and is finite almost surely. Hence, with probability 1, we have

$$Z_\infty = \lim_{n \rightarrow \infty} Z_n = \prod_{n=1}^{\infty} (1 + a_n^-) \left( \lim_{n \rightarrow \infty} Y_n - \sum_{k=1}^{\infty} c'_k \right).$$

**Part 2** Because  $a_n^+ = \max\{0, a_n\}$ , we have  $a'_n = \frac{a_n^+}{1+a_n^-} \geq 0$ . Similar to the arguments in Part 1, we have

$$\mathbb{E}[Z'_{n+1} | \mathcal{F}_n] \leq (1 - a'_n) Z'_n + c'_n.$$

Because we assume that  $\sum_{n=1}^{\infty} a_n = +\infty$  with probability 1, we have

$$\sum_{n=1}^N a'_n \geq \frac{1}{1 + \sup_n a_n^-} \sum_{n=1}^N (a_n - a_n^-) \rightarrow +\infty,$$

as  $N \rightarrow \infty$  with probability 1. Let  $Y'_1 = Z'_1$  and for any  $n \geq 2$

$$Y'_n = Z'_n + \sum_{k=1}^{n-1} a'_k Z'_k - \sum_{k=1}^{n-1} c'_k.$$

Since  $|a'_n| \leq |a_n|$ ,  $|a'_n Z'_n| \leq a_n^+ Z_n$ ,  $|a_n|$  and  $a_n^+ Z_n$  are intergrable, we know that  $Y'_n$  is intergrable for any  $n \geq 1$ . Similar to the arguments in Part 1, we obtain that  $Y'_{n \wedge \tau_T}$  is intergrable for any  $T \geq 0$ ,

$$\begin{aligned} \mathbb{E}[Y'_{n+1} | \mathcal{F}_n] &\leq Y'_n \text{ and} \\ Y'_{n \wedge \tau_T} &\geq - \sum_{k=0}^{n \wedge \tau_T - 1} c'_k \geq -T, \text{ and} \end{aligned}$$

$$\begin{aligned} \mathbb{E}[Y'_{(n+1) \wedge \tau_T} | \mathcal{F}_n] &= Y'_{\tau_T} I(\tau_T \leq n) + \mathbb{E}[Y'_{n+1} | \mathcal{F}_n] I(\tau_T > n) \\ &\leq \sum_{l=1}^n Y'_{\tau_T} I(\tau_T = l) + Y'_n I(\tau_T > n) = Y'_{n \wedge \tau_T}. \end{aligned}$$

In conclusion, we obtain that  $Y'_{n \wedge \tau_T}$  is a super-martingale, such that

$$0 \leq \mathbb{E}[Y'_{n \wedge \tau_T} + T] \leq \cdots \leq \mathbb{E}[Y'_{1 \wedge \tau_T} + T] = \mathbb{E}Z'_1 + T < \infty.$$

This concludes that  $\{Y'_{n \wedge \tau_T} + T\}_{n=1}^{\infty}$  is a  $L^1$  uniformly bounded supermartingale (see Section 1.1 in Hall and Heyde (1980)). Applying Doob's convergence theorem (see Theorem 2.5 in Hall and Heyde (1980)) to  $L^1$  uniformly bounded submartingale  $-(Y'_{n \wedge \tau_T} + T)$ , we know that  $\lim_{n \rightarrow \infty} Y'_{n \wedge \tau_T}$  exists and is finite for any  $T \geq 0$ . Similar to the arguments in Part 1, we know that with probability 1,  $\lim_{n \rightarrow \infty} Y'_n$  exists and is finite.

Notice that with probability 1,

$$0 \leq \sum_{k=1}^{n-1} a'_k Z'_k = Y'_n - Z'_n + \sum_{k=1}^{n-1} c'_k \leq Y'_n + \sum_{k=1}^{n-1} c'_k < \infty.$$

Combined with (142), we obtain that  $\sum_{k=1}^{\infty} a'_k Z'_k$  exists with probability 1.

Because with probability 1,

$$\sum_{n=1}^{\infty} a'_n = +\infty, \quad \sum_{k=1}^{\infty} a'_k Z'_k < \infty, \quad \text{and} \quad \lim_{n \rightarrow \infty} Z'_n \text{ exists,}$$

we obtain that with probability 1

$$\lim_{n \rightarrow \infty} Z'_n = 0.$$

**Part 3** We define  $g(t) = (1 - ct)(1 + t)^\beta, t \geq 0$ . Notice that  $\lim_{t \rightarrow 0+} \frac{g(t) - g(0)}{t} = g'(0) = -(c - \beta) < 0$ . Thus, there exists  $N > 0$  such that  $g(\frac{1}{n}) \leq 1 - \frac{c - \beta}{2n}$ , for all  $n \geq N$ . Define  $C_n = (n + 1)^\beta c_n$  and  $A_n = 1 - A'_n$ , where

$$A'_n = \begin{cases} g(\frac{1}{n}), & n < N \\ 1 - \frac{c - \beta}{2n}, & n \geq N. \end{cases}$$

Note that

$$\frac{(n + 1)^\beta}{n^\beta} (1 - a_n) \leq (1 + \frac{1}{n})^\beta (1 - \frac{c}{n}) = g(\frac{1}{n}) \leq A'_n = 1 - A_n, n \geq 1.$$

This implies that

$$\mathbb{E} [(n + 1)^\beta Z_{n+1} \mid \mathcal{F}_n] \leq (1 - A_n) n^\beta Z_n + C_n.$$

Because the limit  $\sum_{n=1}^{\infty} n^\beta c_n$  exists and  $\{(\frac{n+1}{n})^\beta\}_{n=1}^{\infty}$  is a monotone and bounded sequence with probability 1, by Abel's test for series (see Exercise 9.15 in Ghorpade and Limaye (2006)), the limit  $\sum_{n=1}^{\infty} C_n$  exists with probability 1.

It is straightforward to check that

$$\sum_{n=1}^{\infty} A_n = \infty, \quad \text{and} \quad \sum_{n=1}^{\infty} A_n^- < \infty.$$

Applying the second conclusion in Lemma 14.15, we obtain that with probability 1

$$\lim_{n \rightarrow \infty} n^\beta Z_n = 0.$$

□

In the rest of the section, let  $Z_n = \mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_n) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*)$ . Applying Assumption 5 and Lemma 13.7, we know that  $\mathbb{F}_{\boldsymbol{\theta}}(\boldsymbol{\pi})$  is convex in  $\boldsymbol{\pi}$ . Notice

$$\mathbb{E}[Z_n \mid \mathcal{F}_{n-1}] = \mathbb{F}_{\boldsymbol{\theta}^*} \left( \frac{n-1}{n} \bar{\boldsymbol{\pi}}_{n-1} + \frac{1}{n} \delta_{a_n} \right) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*),$$

and

$$Z_{n-1} = \mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_{n-1}) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*).$$

**Lemma 14.16.**  $K_U$  defined in (127) satisfies that

- $\boldsymbol{\pi}, \boldsymbol{\pi}' \in K_U, t \in (0, 1) \implies t\boldsymbol{\pi} + (1-t)\boldsymbol{\pi}' \in K_{U/2}$ , and
- $\boldsymbol{\pi} \in K_U \implies \lambda_{\max}(\{\mathcal{I}^\pi(\boldsymbol{\theta})\}^{-1}) \leq \frac{1}{\underline{c}U}$ .

Moreover, there exists  $U_0 > 0$  such that

$$\bigcup_{\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}} \arg \min_{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}} \mathbb{F}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\pi}) \subset K_{U_0}. \quad (145)$$

and for both generalized GI0 and GI1 defined in (76) and (77), and for all  $n \geq n_0$ , we have

$$\bar{\boldsymbol{\pi}}_n \in K_{U_0}, \forall n \geq n_0,$$

where  $n_0$  satisfies that  $\sum_{i=1}^{n_0} \mathcal{I}_{a_i}(\hat{\boldsymbol{\theta}}_0)$  is non-singular for some  $\hat{\boldsymbol{\theta}}_0 \in \boldsymbol{\Theta}$ .

*Proof of Lemma 14.16.* For any  $\boldsymbol{\pi}, \boldsymbol{\pi}' \in K_U$ , and  $t \in (0, 1/2]$ ,

$$\max_{S \subset \mathcal{A}: S \text{ is relevant}} \min_{a \in S} t\pi(a) + (1-t)\pi'(a) \geq \frac{1}{2} \max_{S \subset \mathcal{A}: S \text{ is relevant}} \min_{a \in S} \pi'(a) \geq \frac{U}{2}.$$

When  $t \in [1/2, 1)$ , we can obtain the same lower bound, which means that  $t\boldsymbol{\pi} + (1-t)\boldsymbol{\pi}' \in K_{U/2}$  for any  $t \in (0, 1)$ . For any  $\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}$ , define

$$\boldsymbol{\pi}_I := \max_{S \subset \mathcal{A}: S \text{ is relevant}} \min_{a \in S} \pi(a). \quad (146)$$

By Assumption 6B,

$$\underline{c}\boldsymbol{\pi}_I \cdot I_p \preceq \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\boldsymbol{\theta}) \preceq \sum_{a: \pi(a) > 0} \mathcal{I}_a(\boldsymbol{\theta}) \preceq \bar{c} \cdot \boldsymbol{P}_{V_{\{a: \pi(a) > 0\}}}(\boldsymbol{\theta}). \quad (147)$$

Notice that for any  $\boldsymbol{\theta}$ ,

$$\boldsymbol{\pi}_I > 0 \implies \mathcal{I}^\pi(\boldsymbol{\theta}) \succ 0 \implies \dim \left( V_{\{a: \pi(a) > 0\}}(\boldsymbol{\theta}) \right) = p \implies \boldsymbol{\pi}_I \geq \min_{a: \pi(a) > 0} \pi(a) > 0.$$

Thus, we know that  $\pi_I > 0$ , if and only if  $\mathcal{I}^\pi(\theta^*)$  is nonsingular, if and only if  $\mathcal{I}^\pi(\theta)$  is nonsingular for any  $\theta \in \Theta$ .

Let  $Q = \{a \in \mathcal{A}; \pi(a) > \pi_I\}$ . We will show by contradiction that if  $Q$  is not empty, then  $\dim(V_Q) < p$ . If  $\dim(V_Q) = p$ , then by Assumption 6B, we know that  $\mathbf{P}_{V_Q(\theta)} = I_p$ . By (16), we know that  $\mathcal{I}^\pi(\theta^*)$  is nonsingular, which means that  $Q \subset \mathcal{A}$  is relevant. However,  $\min_{a \in Q} \pi(a) > \pi_I$ , which contradicts the definition of  $\pi_I$  in (146).

Thus,  $\dim(V_Q) < p$  and  $\mathbf{P}_{V_Q(\theta)} \neq I_p$ . By Assumption 6B, we obtain

$$\underline{c}\pi_I \cdot I_p \preceq \sum_{a \in \mathcal{A}} \pi(a) \mathcal{I}_a(\theta) \preceq \sum_{a \in Q} \pi(a) \mathcal{I}_a(\theta) + \sum_{a \notin Q} \pi_I \mathcal{I}_a(\theta) \preceq \bar{c} \cdot \mathbf{P}_{V_Q(\theta)} + \bar{c}\pi_I \cdot I_p. \quad (148)$$

Applying Courant–Fischer–Weyl min-max principle (see Chapter I of Hilbert and Courant (1953) or Corollary III.1.2 in Bhatia (1997)) for Rayleigh quotient on (148), we obtain that

$$\lambda_{\min}(\mathcal{I}^\pi(\theta)) \in [\underline{c}\pi_I, \bar{c}\pi_I]. \quad (149)$$

Applying Theorem 14.1,  $\pi \in K_U$  implies  $\lambda_{\min}(\mathcal{I}^\pi(\theta)) \geq \underline{c}U$ , which further implies  $\lambda_{\max}(\{\mathcal{I}^\pi(\theta)\}^{-1}) \leq \frac{1}{\underline{c}U}$ .

Also, we have

$$\lambda_{\max}(\{\mathcal{I}^\pi(\theta^*)\}^{-1}) \rightarrow \infty \iff \lambda_{\max}(\{\mathcal{I}^\pi(\theta)\}^{-1}) \rightarrow \infty, \forall \theta \in \Theta \iff \pi_I \rightarrow 0. \quad (150)$$

We will show (145) by contradiction. Set  $U_n = \frac{1}{n}$ . Assume, in contrast to (145), that there exists  $\hat{\theta}^n \in \Theta$  and  $\pi^n \in \mathcal{S}^{\mathcal{A}}$ , such that

$$\mathbb{F}_{\hat{\theta}^n}(\pi^n) = \min_{\pi \in \mathcal{S}^{\mathcal{A}}} \mathbb{F}_{\hat{\theta}^n}(\pi), \text{ and } \pi_I^n \leq U_n = \frac{1}{n}.$$

Then,

$$\limsup_{n \rightarrow \infty} \mathbb{F}_{\hat{\theta}^n}(\pi^n) = \limsup_{n \rightarrow \infty} \min_{\pi \in \mathcal{S}^{\mathcal{A}}} \mathbb{F}_{\hat{\theta}^n}(\pi) \leq \max_{\theta \in \Theta} \min_{\pi \in \mathcal{S}^{\mathcal{A}}} \mathbb{F}_\theta(\pi) < \infty \quad (151)$$

Set  $\mathbf{A}_n = \mathcal{I}^{\pi^n}(\theta)$ . We know that  $\|\mathbf{A}_n\|_{op} \leq \bar{c}$  and by (149)  $\lambda_{\min}(\mathbf{A}_n) \leq \bar{c}\pi_I^n \leq \bar{c}/n$ . Let  $\mathbb{G}_\theta(\mathbf{A}_n^{-1}) = \Phi_q(\mathbf{A}_n^{-1})$ . When  $q = 0$ , we know that

$$\lim_{n \rightarrow \infty} \min_{\theta \in \Theta} \Phi_0(\mathbf{A}_n^{-1}) \geq \lim_{n \rightarrow \infty} \log(n/\bar{c}) + (p-1) \log(1/\bar{c}) = \infty. \quad (152)$$

When  $q > 0$ , we know that

$$\lim_{n \rightarrow \infty} \min_{\theta \in \Theta} \Phi_q(\mathbf{A}_n^{-1}) \geq \lim_{n \rightarrow \infty} \min_{\theta \in \Theta} \lambda_{\max}(\mathbf{A}_n^{-1}) \geq \lim_{n \rightarrow \infty} n/\bar{c} = \infty. \quad (153)$$



Combining (152) and (153) with Assumption 5 and (150), we know that

$$\lim_{\pi_I^n \rightarrow 0} \min_{\theta \in \Theta} \mathbb{F}_\theta(\pi^n) = \lim_{\pi_I^n \rightarrow 0} \min_{\theta \in \Theta} \mathbb{G}_\theta(\{\mathcal{I}^{\pi^n}(\theta)\}^{-1}) = \infty. \quad (154)$$

By equivalence result (150) and limit result (154), since  $\pi_I^n \rightarrow 0$  as  $n \rightarrow \infty$ , we obtain

$$\liminf_{n \rightarrow \infty} \mathbb{F}_{\hat{\theta}^n}(\pi^n) \geq \liminf_{\pi_I^n \rightarrow 0} \min_{\theta \in \Theta} \mathbb{F}_\theta(\pi^n) \rightarrow \infty,$$

which contradicts (151). Thus, (145) holds.

By Theorem 14.1, we know that there exists  $U_0 > 0$  such that

$$\bar{\pi}_n \in K_{U_0}, n \geq n_0.$$

Combined with (149), we completes the proof.  $\square$

**Lemma 14.17.** *Under Assumptions 1-5, there exists  $L_U < \infty$  such that*

$$\|\nabla \nabla_\theta \mathbb{F}_\theta(\pi)\|_{op} \leq L_U, \text{ and } \|\nabla^2 \mathbb{F}_\theta(\pi)\|_{op} \leq L_U,$$

for any  $\theta \in \Theta$  and  $\pi \in K_U$ .

*Proof of Lemma 14.17.* Define  $u_\theta(\mathbf{A}) = \mathbb{G}_\theta(\mathbf{A}^{-1})$ .

For any positive definite matrix  $\mathbf{A}$ , each element of  $\mathbf{A}^{-1}$  is a well-defined composition of elementary functions of  $\mathbf{A}$ . Therefore, each element of  $\mathbf{A}^{-1}$  is infinitely differentiable.

By Assumptions 5 and Lemma 13.5,  $\nabla_\theta \mathbb{G}_\theta(\Sigma)$  and  $\nabla^2 \mathbb{G}_\theta(\Sigma)$  are continuous in  $(\theta, \Sigma)$  for any  $\theta \in \Theta$  and positive definite matrix  $\Sigma$ . Set  $\mathbf{A} = \mathcal{I}^\pi(\theta)$ . We have  $\mathbb{F}_\theta(\pi) = u_\theta(\mathcal{I}^\pi(\theta))$ .

By the chain rule, we know that

$$\frac{\partial}{\partial \theta_i} \mathbb{F}_\theta(\pi) = \left\langle \frac{\partial}{\partial \mathbf{A}} u_\theta(\mathbf{A}) \Big|_{\mathbf{A}=\mathcal{I}^\pi(\theta)}, \frac{\partial \mathcal{I}^\pi(\theta)}{\partial \theta_i} \right\rangle + \frac{\partial}{\partial \theta_i} \mathbb{G}_\theta(\mathbf{A}^{-1}) \Big|_{\mathbf{A}=\mathcal{I}^\pi(\theta)}. \quad (155)$$

Notice that each element of  $\frac{\partial}{\partial \mathbf{A}} u_\theta(\mathbf{A}) \Big|_{\mathbf{A}=\mathcal{I}^\pi(\theta)}$  is continuously differentiable in  $\mathbf{A}$ . Thus

$$\frac{\partial}{\partial \pi(a)} \frac{\partial}{\partial \mathbf{A}} u_\theta(\mathcal{I}^\pi(\theta)) \quad (156)$$

exists and is continuous in  $(\pi, \theta) \in K_U \times \Theta$ .

Furthermore, we know that

$$\frac{\partial}{\partial \pi(a)} \frac{\partial \mathcal{I}^\pi(\theta)}{\partial \theta_i} = \frac{\partial \mathcal{I}_a(\theta)}{\partial \theta_i}, \text{ and} \quad (157)$$

$$\frac{\partial}{\partial \pi(a)} \left( \frac{\partial}{\partial \theta_i} \mathbb{G}_{\boldsymbol{\theta}}(\mathbf{A}^{-1}) \Big|_{\mathbf{A}=\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta})} \right) = \left\langle \frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \mathbf{A}} \mathbb{G}_{\boldsymbol{\theta}}(\mathbf{A}^{-1}) \Big|_{\mathbf{A}=\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta})}, \mathcal{I}_a(\boldsymbol{\theta}) \right\rangle. \quad (158)$$

Combining (155), (156), (157) and (158), we obtain that  $\|\nabla \nabla_{\boldsymbol{\theta}} \mathbb{F}_{\boldsymbol{\theta}}(\boldsymbol{\pi})\|$  is continuous over  $\boldsymbol{\Theta} \times K_U$  for any  $U > 0$ . By Lemma 14.16 and the definition of  $K_U$  in (127), we know that  $K_U$  is a close subset of  $\mathcal{S}^{\mathcal{A}}$ . Thus,  $\boldsymbol{\Theta} \times K_U$  is compact.

By chain rule, we know that

$$\frac{\partial}{\partial \pi(a)} \mathbb{F}_{\boldsymbol{\theta}}(\boldsymbol{\pi}) = \left\langle \frac{\partial}{\partial \mathbf{A}} u_{\boldsymbol{\theta}}(\mathbf{A}) \Big|_{\mathbf{A}=\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta})}, \frac{\partial \mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta})}{\partial \pi(a)} \right\rangle = \left\langle \frac{\partial}{\partial \mathbf{A}} u_{\boldsymbol{\theta}}(\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta})), \mathcal{I}_a(\boldsymbol{\theta}) \right\rangle. \quad (159)$$

Because  $u_{\boldsymbol{\theta}}(\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}))$  is twice continuously differentiable in  $\mathbf{A}$ , we know that  $\nabla^2 \mathbb{F}_{\boldsymbol{\theta}}(\boldsymbol{\pi})$  is continuous over compact set  $\boldsymbol{\Theta} \times K_U$  for any  $U$ .

In conclusion, there exists  $L_U < \infty$  such that

$$\|\nabla \nabla_{\boldsymbol{\theta}} \mathbb{F}_{\boldsymbol{\theta}}(\boldsymbol{\pi})\|_{op} \leq L_U, \text{ and } \|\nabla^2 \mathbb{F}_{\boldsymbol{\theta}}(\boldsymbol{\pi})\|_{op} \leq L_U,$$

for any  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$  and  $\boldsymbol{\pi} \in K_U$ . □

**Lemma 14.18.** *Under Assumptions 1-5 as well as 6A-7A (or 6B-7B), the generalized GI0 and (76) and GI1, defined in (77), satisfy that there exists a constant  $L > 0$  such that,*

$$\mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_n) - \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\boldsymbol{\pi}_n^*) \leq (1 - \frac{1}{n})(\mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_{n-1}) - \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\boldsymbol{\pi}_n^*)) + \frac{L}{n^2}, n \geq n_0. \quad (160)$$

*Proof of Lemma 14.18.* By Theorem 14.1 and Lemma 14.16, there exists  $0 < U < \infty$  such that  $\bar{\boldsymbol{\pi}}_n, \boldsymbol{\pi}_n^* \in K_U$ , where we define

$$\boldsymbol{\pi}_n^* \in \arg \min_{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}} \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\boldsymbol{\pi}).$$

By Lemma 13.7,  $\mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\boldsymbol{\pi})$  is convex in  $\boldsymbol{\pi}$ . According to Jensen's inequality,

$$\mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\frac{n-1}{n}\bar{\boldsymbol{\pi}}_{n-1} + \frac{1}{n}\boldsymbol{\pi}_n^*) \leq \frac{n-1}{n}\mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_{n-1}) + \frac{1}{n}\mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\boldsymbol{\pi}_n^*). \quad (161)$$

Thus,

$$\mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\frac{n-1}{n}\bar{\boldsymbol{\pi}}_{n-1} + \frac{1}{n}\boldsymbol{\pi}_n^*) - \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\boldsymbol{\pi}_n^*) \leq (1 - \frac{1}{n})(\mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_{n-1}) - \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\boldsymbol{\pi}_n^*)). \quad (162)$$

Notice that

$$\mathbb{F}_{\boldsymbol{\theta}_{n-1}}\left(\frac{n-1}{n}\bar{\boldsymbol{\pi}}_{n-1} + \frac{1}{n}\boldsymbol{\pi}_n^*\right) - \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_{n-1}) = \left\langle \nabla \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_{n-1}), \frac{1}{n}\boldsymbol{\pi}_n^* - \frac{1}{n}\bar{\boldsymbol{\pi}}_{n-1} \right\rangle + \bar{R}(\bar{\boldsymbol{\pi}}_{n-1}, \boldsymbol{\pi}_n^*), \quad (163)$$

where

$$\bar{R}(\bar{\boldsymbol{\pi}}_{n-1}, \boldsymbol{\pi}_n^*) = \left\langle \nabla \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\boldsymbol{\pi}') - \nabla \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_{n-1}), \frac{1}{n}\boldsymbol{\pi}_n^* - \frac{1}{n}\bar{\boldsymbol{\pi}}_{n-1} \right\rangle,$$

for some  $\boldsymbol{\pi}'$  between  $\bar{\boldsymbol{\pi}}_{n-1}$  and  $\frac{n-1}{n}\bar{\boldsymbol{\pi}}_{n-1} + \frac{1}{n}\boldsymbol{\pi}_n^*$ . By Lemma 14.16, we know that  $\boldsymbol{\pi}' \in K_{U/2}$ .

By Assumptions 1-5 and Lemma 14.17, there exists a constant  $C' < \infty$  such that

$$\begin{aligned} & |\bar{R}(\bar{\boldsymbol{\pi}}_{n-1}, \boldsymbol{\pi}_n^*)| \\ & \leq \left\| \frac{1}{n}\boldsymbol{\pi}_n^* - \frac{1}{n}\bar{\boldsymbol{\pi}}_{n-1} \right\|^2 \sup_{\boldsymbol{\pi} \in K_{U/2}, \boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \nabla^2 \mathbb{F}_{\boldsymbol{\theta}}(\boldsymbol{\pi}) \right\|_{op} \\ & \leq \frac{C'}{n^2} \sup_{\boldsymbol{\pi} \in K_{U/2}, \boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \nabla^2 \mathbb{F}_{\boldsymbol{\theta}}(\boldsymbol{\pi}) \right\|_{op} \\ & = \frac{L}{n^2}, \end{aligned} \quad (164)$$

where

$$L = C' \sup_{\boldsymbol{\pi} \in K_{U/2}, \boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \nabla^2 \mathbb{F}_{\boldsymbol{\theta}}(\boldsymbol{\pi}) \right\|_{op} < \infty.$$

By Lemma 13.4, we know that

$$\langle \nabla \mathbb{F}_{\boldsymbol{\theta}}(\boldsymbol{\pi}), \delta_a \rangle = \frac{\partial}{\partial \pi(a)} \mathbb{F}_{\boldsymbol{\theta}}(\boldsymbol{\pi}) = - \langle \nabla \mathbb{G}_{\boldsymbol{\theta}}(\{\mathcal{I}^\pi(\boldsymbol{\theta})\}^{-1}), \{\mathcal{I}^\pi(\boldsymbol{\theta})\}^{-1} \mathcal{I}_a(\boldsymbol{\theta}) \{\mathcal{I}^\pi(\boldsymbol{\theta})\}^{-1} \rangle.$$

Let  $a_n^{(1)}$  be the experiment selected following the generalized GI1. Then, according to the definition of GI1, it minimizes the following function over  $\mathcal{S}^{\mathcal{A}}$  with respect to  $a$ :

$$\left\langle \nabla \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_{n-1}), \frac{1}{n}\boldsymbol{\pi} - \frac{1}{n}\bar{\boldsymbol{\pi}}_{n-1} \right\rangle = \sum_{a=1}^k \boldsymbol{\pi}(a) \left\langle \nabla \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_{n-1}), \frac{1}{n}\delta_a - \frac{1}{n}\bar{\boldsymbol{\pi}}_{n-1} \right\rangle.$$

By similar Taylor expansion arguments as those for (164), we have for all  $n \geq n_0$ ,

$$\left| \mathbb{F}_{\boldsymbol{\theta}_{n-1}}\left(\frac{n-1}{n}\bar{\boldsymbol{\pi}}_{n-1} + \frac{1}{n}\delta_{a_n^{(1)}}\right) - \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_{n-1}) - \left\langle \nabla \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_{n-1}), \frac{1}{n}\delta_{a_n^{(1)}} - \frac{1}{n}\bar{\boldsymbol{\pi}}_{n-1} \right\rangle \right| \leq \frac{L}{n^2}.$$

The above inequality implies that for all  $n \geq n_0$ , GI1 satisfies

$$\mathbb{F}_{\hat{\boldsymbol{\theta}}_{n-1}}(\bar{\boldsymbol{\pi}}_n) = \mathbb{F}_{\hat{\boldsymbol{\theta}}_{n-1}}\left(\frac{n-1}{n}\bar{\boldsymbol{\pi}}_{n-1} + \frac{1}{n}\delta_{a_n^{(1)}}\right) \leq \mathbb{F}_{\hat{\boldsymbol{\theta}}_{n-1}}\left(\frac{n-1}{n}\bar{\boldsymbol{\pi}}_{n-1} + \frac{1}{n}\boldsymbol{\pi}_n^*\right) + \frac{L}{n^2}. \quad (165)$$

For GI0, let  $a_n^{(0)}$  be the experiment selected at time  $n$ . Then, according to its definition we have

$$\mathbb{F}_{\hat{\theta}_{n-1}}\left(\frac{n-1}{n}\bar{\pi}_{n-1} + \frac{1}{n}\delta_{a_n^{(0)}}\right) \leq \mathbb{F}_{\hat{\theta}_{n-1}}\left(\frac{n-1}{n}\bar{\pi}_{n-1} + \frac{1}{n}\delta_{a_n^{(1)}}\right). \quad (166)$$

Therefore, the proof of Lemma 14.18 is concluded by combining inequalities (162) – (166).  $\square$

**Lemma 14.19.** *Under Assumptions 1-5 as well as 6A-7A (or 6B-7B), the generalized GI0 selection (76) and GI1 selection (77) satisfy that there exists  $0 < C < \infty$  such that*

$$\mathbb{F}_{\theta^*}(\bar{\pi}_n) - \mathbb{F}_{\theta^*}(\pi^*) \leq \left(1 - \frac{1}{n}\right)(\mathbb{F}_{\theta^*}(\bar{\pi}_{n-1}) - \mathbb{F}_{\theta^*}(\pi^*)) + c_{n-1}, n \geq n_0, \quad (167)$$

where  $c_{n-1} = \frac{C}{n^2} + \frac{C}{n} \|\theta_{n-1} - \theta^*\|$ .

*Proof of Lemma 14.19.* We can rewrite (160) as

$$\mathbb{F}_{\theta_{n-1}}(\bar{\pi}_n) - \mathbb{F}_{\theta_{n-1}}(\bar{\pi}_{n-1}) + \frac{1}{n}(\mathbb{F}_{\theta_{n-1}}(\bar{\pi}_{n-1}) - \mathbb{F}_{\theta_{n-1}}(\pi_n^*)) \leq \frac{L}{n^2}. \quad (168)$$

We first show that for all  $\pi_0, \pi_1 \in K_U$ , and  $\theta_1, \theta_2 \in \Theta$ ,

$$|\mathbb{F}_{\theta_1}(\pi_1) - \mathbb{F}_{\theta_1}(\pi_0) - \{\mathbb{F}_{\theta_2}(\pi_1) - \mathbb{F}_{\theta_2}(\pi_0)\}| \leq C_1 \|\pi_0 - \pi_1\| \|\theta_1 - \theta_2\|, \quad (169)$$

where  $C_1 = \sup_{\pi \in K_{U/2}, \theta \in \Theta} \|\nabla_{\theta} \nabla \mathbb{F}_{\theta}(\pi)\|_{op}$  is a positive constant. To show this, set  $g(t) = \mathbb{F}_{\theta_1}(\pi(t)) - \mathbb{F}_{\theta_2}(\pi(t))$ , where  $\pi(t) = t\pi_1 + (1-t)\pi_0, t \in [0, 1]$  and  $\pi_0, \pi_1 \in K_U$ , where  $K_U$  is chosen according to the proof of Lemma 14.18. By Lagrange mean value theorem, there exists  $0 < t < 1$  such that

$$g(1) - g(0) = g'(t) = \langle \nabla \mathbb{F}_{\theta_1}(\pi(t)) - \nabla \mathbb{F}_{\theta_2}(\pi(t)), \pi_1 - \pi_0 \rangle.$$

By Assumptions 1-5, Lemma 14.16 and Lemma 14.17, we know that

$$\frac{\|\nabla \mathbb{F}_{\theta_1}(\pi(t)) - \nabla \mathbb{F}_{\theta_2}(\pi(t))\|}{\|\theta_1 - \theta_2\|} \leq \sup_{\pi \in K_{U/2}, \theta \in \Theta} \|\nabla_{\theta} \nabla \mathbb{F}_{\theta}(\pi)\|_{op} < \infty.$$

Set

$$C_1 = \sup_{\pi \in K_{U/2}, \theta \in \Theta} \|\nabla_{\theta} \nabla \mathbb{F}_{\theta}(\pi)\|_{op}.$$

Then, the above inequality implies (169). Note that

$$\|\bar{\pi}_n - \bar{\pi}_{n-1}\| \leq \frac{2}{n}.$$

The above inequality together with (169) implies

$$\begin{aligned} |(\mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_n) - \mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_{n-1})) - (\mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_n) - \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_{n-1}))| &\leq \frac{2C_1}{n} \|\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}^*\|, \text{ and} \\ |(\mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_{n-1}) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*)) - (\mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_{n-1}) - \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\boldsymbol{\pi}^*))| &\leq 2C_1 \|\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}^*\|. \end{aligned} \quad (170)$$

Because  $\mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\boldsymbol{\pi}^*) \geq \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\boldsymbol{\pi}_n^*)$ , we have

$$\begin{aligned} &(\mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_{n-1}) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*)) - (\mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_{n-1}) - \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\boldsymbol{\pi}_n^*)) \\ &\leq (\mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_{n-1}) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*)) - (\mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_{n-1}) - \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\boldsymbol{\pi}^*)) \\ &\leq 2C_1 \|\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}^*\|. \end{aligned} \quad (171)$$

By triangular inequality, inequalities (168), (170) and (171), we obtain

$$\begin{aligned} &\mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_n) - \mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_{n-1}) + \frac{1}{n}(\mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_{n-1}) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*)) \\ &\leq |(\mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_n) - \mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_{n-1})) - (\mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_n) - \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_{n-1}))| \\ &\quad + (\mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_n) - \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_{n-1})) + \frac{1}{n}(\mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\bar{\boldsymbol{\pi}}_{n-1}) - \mathbb{F}_{\boldsymbol{\theta}_{n-1}}(\boldsymbol{\pi}_n^*)) + \frac{2C_1 \|\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}^*\|}{n} \\ &\leq \frac{4C_1}{n} \|\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}^*\| + \frac{L}{n^2}. \end{aligned} \quad (172)$$

In conclusion, we know that

$$\mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_n) - \mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_{n-1}) + \frac{1}{n}(\mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_{n-1}) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*)) \leq c_{n-1},$$

where  $c_{n-1} = \frac{C}{n^2} + \frac{C}{n} \|\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}^*\|$ , and  $C = 4C_1 + L$ . □

As a corollary of Theorem 4.2, we establish the following:

**Corollary 14.20.** *Under Assumptions 1-5 as well as Assumptions 6A-7A (or Assumptions 6B-7B), if there exists  $U > 0$  such that  $n \geq n_0 \implies \bar{\boldsymbol{\pi}}_n \in K_U$ , then with probability 1,*

$$\sum_{n=1}^{\infty} n^{-s} \left\| \sqrt{n}(\hat{\boldsymbol{\theta}}_n^{ML} - \boldsymbol{\theta}^*) \right\|^t < \infty.$$

provided  $s > 1, 0 < t \leq 2$ .

*Proof of Corollary 14.20.* Let  $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n^{ML}$ . We first assume  $t = 2$ . Applying Lemma 14.13,

$$\limsup_{n \rightarrow \infty} \left\| \{-\nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*)\}^{-1} \right\|_{op} \leq \frac{1}{\min_{\boldsymbol{\pi} \in K_U} \lambda_{\min}(I^{\pi}(\boldsymbol{\theta}^*))}.$$

Set  $D_n := \left\{ 1 - \frac{1}{n} \sum_{j=1}^n \Psi_2^{aj}(X_j) \left\| \{\nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*)\}^{-1} \right\|_{op} \psi \left( \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\| \right) > \frac{1}{2}, \left\| \{-\nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*)\}^{-1} \right\|_{op} \leq \frac{2}{\min_{\pi \in K} \lambda_{\min}(I^\pi(\boldsymbol{\theta}^*))} \right\}$ . By Lemma 4.1, we know that

$$\mathbb{P} \left( \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} D_m \right) = 1. \quad (173)$$

Note that (134) and (135) yield

$$\|\mathbf{W}_n\|_{I_{D_n}} \leq I_{D_n} \frac{\left\| \{\nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*)\}^{-1} \right\|_{op} \|\sqrt{n} \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^*)\|}{1 - \frac{1}{n} \sum_{j=1}^n \Psi_2^{aj}(X_j) \left\| \{\nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*)\}^{-1} \right\|_{op} \psi \left( \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right\| \right)} \leq C' \|\sqrt{n} \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^*)\|,$$

where  $C' < \infty$  and  $\mathbf{W}_n = \sqrt{n}(\hat{\boldsymbol{\theta}}_n^{\text{ML}} - \boldsymbol{\theta}^*)$ . Set  $S_n = \sum_{i=1}^n \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*, a_i}(X_i)$ . By Assumption 2, we know that

$$\sigma^2 := \max_{a \in \mathcal{A}} \mathbb{E}_{X \sim f_{\boldsymbol{\theta}^*, a}} \left\{ \|\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*, a}(X)\|^2 \right\} < \infty.$$

By induction, we obtain that

$$\begin{aligned} & \mathbb{E} \left\| \sqrt{n} \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^*) \right\|^2 \\ &= \frac{1}{n} \mathbb{E} \|S_n\|^2 \\ &= \frac{1}{n} \mathbb{E} \left[ \mathbb{E} \left\{ \|\log f_{\boldsymbol{\theta}^*, a_n}(X_n) + S_{n-1}\|^2 \mid \mathcal{F}_{n-1} \right\} \right] \\ &= \frac{1}{n} \mathbb{E} \left[ \|S_{n-1}\|^2 + 2 \langle S_{n-1}, \mathbb{E} \{ \log f_{\boldsymbol{\theta}^*, a_n}(X_n) \mid \mathcal{F}_{n-1} \} \rangle + \mathbb{E} \left\{ \|\log f_{\boldsymbol{\theta}^*, a_n}(X_n)\|^2 \mid \mathcal{F}_{n-1} \right\} \right] \\ &= \frac{1}{n} \mathbb{E} \left[ \|S_{n-1}\|^2 + \mathbb{E} \left\{ \|\log f_{\boldsymbol{\theta}^*, a_n}(X_n)\|^2 \mid \mathcal{F}_{n-1} \right\} \right] \\ &\leq \frac{1}{n} (\mathbb{E} \|S_{n-1}\|^2 + \sigma^2) \leq \dots \leq \sigma^2. \end{aligned}$$

Apply Lemma 13.1 with  $X_n = \frac{1}{n^s} \left\| \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \right\|^2$ ,  $E_n = D_n$ ,  $\gamma = 1$ , and  $\varepsilon_{n-1} = \frac{1}{n^s} (C')^2 \mathbb{E}[\|\sqrt{n} \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^*)\|^2 \mid \mathcal{F}_{n-1}]$ , because

$$\sum_{n=0}^{\infty} \mathbb{E} \varepsilon_n \leq \sum_{n=1}^{\infty} \frac{\sigma^2 (C')^2}{n^s} < \infty,$$

we obtain that with probability 1,

$$\sum_{n=1}^{\infty} n^{-s} \cdot \mathbb{E} \left\{ \left\| \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \right\|^2 I_{D_n} \right\} < \infty.$$

Combined with (173), we obtain that

$$\begin{aligned}
& \mathbb{P}\left(\sum_{n=1}^{\infty} n^{-s} \left\| \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \right\|^2 = \infty\right) \\
& \leq \mathbb{P}\left(\sum_{n=1}^{\infty} n^{-s} \left\| \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \right\|^2 I_{D_n} = \infty\right) + \mathbb{P}\left(\sum_{n=1}^{\infty} n^{-s} \left\| \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \right\|^2 I_{D_n^c} = \infty\right) \\
& \leq 0 + \mathbb{P}\left(\sum_{n=1}^{\infty} I_{D_n} = \infty\right) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} D_m^c\right) = 0,
\end{aligned}$$

that is,  $\sum_{n=1}^{\infty} n^{-s} \left\| \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \right\|^2 < \infty$  with probability 1.

If  $0 < t < 2$ , set  $s_1 = 1 - \frac{t}{2}$ ,  $s_2 = \frac{t}{2}$ ,  $s_0 = \frac{s-1}{2}$ ,  $p = \frac{1}{s_1} > 1$ , and  $q = \frac{1}{s_2} > 1$ . We have  $1/p + 1/q = 1$ , and  $s_1 + s_2 + 2s_0 = s$ . Notice that

$$\left(\sum_{n=1}^{\infty} n^{-(s_1+s_0)p}\right)^{1/p} = \left(\sum_{n=1}^{\infty} n^{-(1+s_0p)}\right)^{1/p} < \infty,$$

and with probability 1,

$$\sum_{n=1}^{\infty} n^{-(s_2+s_0)q} \left\| \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \right\|^{tq} = \sum_{n=1}^{\infty} n^{-1-s_0q} \left\| \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \right\|^2 < \infty.$$

By Hölder's inequality, with probability 1

$$\sum_{n=1}^{\infty} n^{-s} \left\| \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \right\|^t \leq \left(\sum_{n=1}^{\infty} n^{-(s_1+s_0)p}\right)^{1/p} \left(\sum_{n=1}^{\infty} n^{-(s_2+s_0)q} \left\| \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \right\|^{tq}\right)^{1/q} < \infty.$$

□

**Lemma 14.21.** *Under Assumptions 1-5 as well as Assumptions 6A-7A (or Assumptions 6B-7B), if the sequence of estimators  $\hat{\boldsymbol{\theta}}_n$  satisfies that for  $0 \leq \beta < \frac{1}{2}$ ,*

$$\sum_{n \geq n_0} n^{\beta-1} \left\| \hat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^* \right\| < \infty \text{ a.s.} \quad (174)$$

*then the generalized GIO and GI1 (with  $\boldsymbol{\theta}_n$  replaced by  $\hat{\boldsymbol{\theta}}_n$ ) satisfy*

$$n^\beta Z_n \xrightarrow{\text{a.s.}} 0,$$

*where  $Z_n = \mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_n) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*)$ .*

*Proof of Lemma 14.21.* Based on Lemma 14.19, the GI0 and GI1 selection rules satisfy that there exists  $C < \infty$  such that for any  $n \geq n_0$ ,

$$\mathbb{E}[Z_n | \mathcal{F}_{n-1}] \leq \left(1 - \frac{1}{n}\right) Z_{n-1} + c_{n-1},$$

where  $c_{n-1} = C\left(\frac{1}{n^2} + \frac{\|\hat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*\|}{n}\right)$ . Notice that with probability 1, we have

$$\begin{aligned} \sum_{n=n_0+1}^{\infty} (n-1)^\beta c_{n-1} &\leq \sum_{n=n_0}^{\infty} C \cdot n^\beta \left( \frac{1}{n^2} + \frac{\|\hat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*\|}{n} \right) \\ &= \sum_{n=n_0+1}^{\infty} C \cdot \frac{1}{n^{2-\beta}} + \sum_{n=n_0+1}^{\infty} \frac{C}{n^{1-\beta}} \|\hat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*\| < \infty. \end{aligned}$$

Applying the third part of Lemma 14.15 to  $Z_n$  with  $a_n = 1/(n+1)$ ,  $c = 1$ , we obtain

$$n^\beta Z_n \xrightarrow{\text{a.s.}} 0.$$

□

*Proof of Theorem 4.3.* By Corollary 14.20, we know that with probability 1, if  $0 \leq \beta < \frac{1}{2}$ , then with probability 1,

$$\sum_{n=1}^{\infty} n^{\beta-1} \|\hat{\boldsymbol{\theta}}_n^{\text{ML}} - \boldsymbol{\theta}^*\| = \sum_{n=1}^{\infty} n^{\beta-3/2} \left\| \sqrt{n}(\hat{\boldsymbol{\theta}}_n^{\text{ML}} - \boldsymbol{\theta}^*) \right\| < \infty.$$

By Lemma 14.21, we obtain that  $n^\beta Z_n \rightarrow 0$  a.s.  $\mathbb{P}_*$ . That is,  $\lim_{n \rightarrow \infty} n^\beta \{\mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_n) - \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*)\}$ , a.s.

Next, we prove by contradiction that, when  $\mathbb{F}_{\boldsymbol{\theta}^*}(\cdot)$  has a unique minimizer, we also have  $\lim_{n \rightarrow \infty} \bar{\boldsymbol{\pi}}_n = \boldsymbol{\pi}^*$  a.s. Assume, on the contrary, that there exists a sub-sequence such that  $\bar{\boldsymbol{\pi}}_{n_l} \rightarrow \boldsymbol{\pi}_1 \neq \boldsymbol{\pi}^*$ , as  $l \rightarrow \infty$ . Then, by the continuity of  $\mathbb{F}_{\boldsymbol{\theta}^*}(\cdot)$ , we have  $\mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_{n_l}) \rightarrow \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}_1)$ .

Set  $\beta = 0$ . We obtain that

$$\mathbb{F}_{\boldsymbol{\theta}^*}(\bar{\boldsymbol{\pi}}_n) \rightarrow \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*) \text{ a.s. } \mathbb{P}_*.$$

Given that  $\mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi})$  has a unique global minimizer, it must be the case that  $\mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}_1) \neq \mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi}^*)$ . This contradicts with the above display. □



## 14.6 Proof of Theorem 4.4

*Proof of Theorem 4.4.* By applying Theorem 4.2 and Theorem 4.3, we conclude the proof of Theorem 4.4.  $\square$

## 14.7 Proof of Theorem 4.5

*Proof of Theorem 4.5.* Under the assumptions of Theorem 4.4, the conclusions from Theorem 4.1 and Theorem 4.3 still apply. Hence, we have

$$\lim_{n \rightarrow \infty} \mathcal{I}^{\bar{\pi}_n}(\hat{\boldsymbol{\theta}}_n^{\text{ML}}) = \lim_{n \rightarrow \infty} \sum_{a \in \mathcal{A}} \bar{\pi}_n(a) \mathcal{I}_a(\hat{\boldsymbol{\theta}}_n^{\text{ML}}) = \mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*) \text{ a.s. },$$

and

$$\lim_{n \rightarrow \infty} \left\| \{\mathcal{I}^{\bar{\pi}_n}(\hat{\boldsymbol{\theta}}_n^{\text{ML}})\}^{-1/2} \nabla g(\hat{\boldsymbol{\theta}}_n^{\text{ML}}) \right\| = \left\| \{\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*)\}^{-1/2} \nabla g(\boldsymbol{\theta}^*) \right\| \text{ a.s.}$$

By Slutsky's theorem and Theorem 4.4, we derive the limit result as in (23).

Moreover, through the Delta method, we find

$$\frac{\sqrt{n}(g(\hat{\boldsymbol{\theta}}_n^{\text{ML}}) - g(\boldsymbol{\theta}^*))}{\left\| \{\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*)\}^{-1/2} \nabla g(\boldsymbol{\theta}^*) \right\|} \xrightarrow{d} N(0, 1).$$

Once again, by Slutsky's theorem, we establish the limit result in (24).  $\square$

## 14.8 Proof of Theorem 4.8

We first provide an extension of the Cramér-Rao lower bound for unbiased estimators based on sequential observations following an active experiment selection rule.

**Lemma 14.22** (Cramér-Rao lower bound for sequential data). *Assume that for some initial values  $a_1^0, \dots, a_{n_0}^0 \in \mathcal{A}$ , we consider initial selections  $a_i = a_i^0$  for  $i = 1, \dots, n_0$ , such that the sum  $\sum_{i=1}^{n_0} \mathcal{I}_{a_i^0}(\boldsymbol{\theta})$  is nonsingular for all  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ . Given any deterministic selection function  $h_n$ , we consider the selections  $a_n = h_n(a_1, X_1, \dots, a_{n-1}, X_{n-1}) \in \mathcal{A}, \forall n > n_0$ . Let  $\mathbf{T}_n = T(X_1, X_2, \dots, X_n, \mathbf{a}_n)$  be an unbiased estimator of vector  $\mathbf{h}(\boldsymbol{\theta})$  with a finite second moment, for all  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ , that is  $\mathbf{h}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{T}_n]$  and  $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}_{\boldsymbol{\theta}} \|\mathbf{T}_n\|^2 < \infty$ . Then, under Assumptions 1-4, we have*

$$\text{cov}_{\boldsymbol{\theta}}(\mathbf{T}_n) \succeq \frac{1}{n} \left\{ \nabla_{\boldsymbol{\theta}} \mathbf{h}(\boldsymbol{\theta}) \right\}^T \{ \mathcal{I}^{\mathbb{E}_{\boldsymbol{\theta}} \bar{\pi}_n}(\boldsymbol{\theta}) \}^{-1} \nabla_{\boldsymbol{\theta}} \mathbf{h}(\boldsymbol{\theta}).$$

Specifically, if  $\mathbf{h}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ , then

$$\mathbb{G}_{\boldsymbol{\theta}}(n \text{cov}_{\boldsymbol{\theta}}(\mathbf{T}_n)) \geq \inf_{\pi \in \mathcal{S}^{\mathcal{A}}} \mathbb{G}_{\boldsymbol{\theta}}(\{\mathcal{I}^{\pi}(\boldsymbol{\theta})\}^{-1}).$$

*Proof of Lemma 14.22.* Assume  $\mathbf{h}(\boldsymbol{\theta}) \in \mathbb{R}^l$ . For any  $\mathbf{b} \in \mathbb{R}^l$ , define  $h_{\mathbf{b}}(\boldsymbol{\theta}) = \mathbf{b}^T \mathbb{E}_{\boldsymbol{\theta}}[T_n]$ .

Let  $\{X_i^a\}_{a \in \mathcal{A}, i \geq 1}$  be a sequence of independent random elements, such that  $X_i^a \sim f_{\boldsymbol{\theta}, a}(\cdot)$ . According to Lemma 13.8, we can assume that the observations and experiments are  $a_1, X_1^{a_1}, \dots, a_n, X_n^{a_n}$  in the rest of the proof, where  $a_{n+1} = h_{n+1}(a_1, X_1^{a_1}, \dots, a_n, X_n^{a_n})$ , for all  $n \geq n_0$ .

The joint density for  $\mathbf{X}_n^{\mathcal{A}} = \{X_i^a\}_{1 \leq i \leq n, a \in \mathcal{A}}$  and  $\mathbf{a}_n = (a_1, \dots, a_n)$  is given by

$$f_{\boldsymbol{\theta}}(\mathbf{X}_n^{\mathcal{A}}, \mathbf{a}^n) = \prod_{i=1}^n \prod_{a \in \mathcal{A}} f_{\boldsymbol{\theta}, a}(X_i^a) I(a_1 = a^1, \dots, a_{n_0} = a^{n_0}, a_{n_0+1} = a^{n_0+1}, \dots, a_n = a^n).$$

Notice that

$$\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\mathbf{X}_n^{\mathcal{A}}, \mathbf{a}^n) = f_{\boldsymbol{\theta}}(\mathbf{X}_n^{\mathcal{A}}, \mathbf{a}^n) \sum_{i=1}^n \sum_{a \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(X_i^a).$$

Assume that probability density  $f_{\boldsymbol{\theta}, a}(\cdot)$  is with respect to baseline measure  $\mu_a(\cdot)$ . By Assumption 2, denote the support of probability density  $f_{\boldsymbol{\theta}, a}(\cdot)$  by  $\Omega_a = \text{supp}(f_{\boldsymbol{\theta}, a})$ , which does not depend on  $\boldsymbol{\theta}$ . Let product measure  $d\boldsymbol{\mu}^n(\mathbf{X}_n^{\mathcal{A}}) = \prod_{1 \leq i \leq n, a \in \mathcal{A}} d\mu_a(X_i^a)$ , and product space  $\Omega^1 = \times_{a \in \mathcal{A}} \Omega_a$ ,  $\Omega^n = \times_{a \in \mathcal{A}} \Omega_a \times \Omega^{n-1}$ .

Set  $\mathbf{T}_n = T_n(a_1, X_1^{a_1}, \dots, a_n, X_n^{a_n})$ . Because

$$\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{b}^T T_n] = \sum_{\mathbf{a}^n \in \mathcal{A}^n} \int_{\Omega^n} \mathbf{b}^T T_n(a_1, X_1^{a_1}, \dots, a_n, X_n^{a_n}) f_{\boldsymbol{\theta}}(\mathbf{X}_n^{\mathcal{A}}, \mathbf{a}^n) d\boldsymbol{\mu}^n(\mathbf{X}_n^{\mathcal{A}}),$$

we know that

$$\nabla_{\boldsymbol{\theta}} h_{\mathbf{b}}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{b}^T T_n] = \sum_{\mathbf{a}^n \in \mathcal{A}^n} \nabla_{\boldsymbol{\theta}} \int_{\Omega^n} \mathbf{b}^T T_n(a_1, X_1^{a_1}, \dots, a_n, X_n^{a_n}) f_{\boldsymbol{\theta}}(\mathbf{X}_n^{\mathcal{A}}, \mathbf{a}^n) d\boldsymbol{\mu}^n(\mathbf{X}_n^{\mathcal{A}}).$$

By Assumption 2, we know that for any  $a \in \mathcal{A}$

$$\|\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(X^a)\| \leq \|\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*, a}(X^a)\| + \Psi_1^a(X^a) \max_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| =: F_a(X^a), \forall X^a \in \Omega_a,$$

where the dominate function  $F_a$  satisfies that

$$\sup_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{X^a \sim f_{\boldsymbol{\theta}, a}} \{F_a(X^a)\}^2 < \infty.$$

Notice that

$$\begin{aligned}
& \int_{\Omega^n} \mathbf{b}^T T_n(a_1, X_1^{a_1}, \dots, a_n, X_n^{a_n}) \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\mathbf{X}_n^{\mathcal{A}}, \mathbf{a}^n) d\boldsymbol{\mu}^n(\mathbf{X}_n^{\mathcal{A}}) \\
&= \int_{\Omega^n} \mathbf{b}^T T_n(a_1, X_1^{a_1}, \dots, a_n, X_n^{a_n}) \sum_{i=1}^n \sum_{a \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta},a}(X_i^a) f_{\boldsymbol{\theta}}(\mathbf{X}_n^{\mathcal{A}}, \mathbf{a}^n) d\boldsymbol{\mu}^n(\mathbf{X}_n^{\mathcal{A}}), \\
& \left\| \mathbf{b}^T \mathbf{T}_n \sum_{i=1}^n \sum_{a \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta},a}(X_i^a) \right\| \leq |\mathbf{b}^T \mathbf{T}_n| \cdot \sum_{i=1}^n \sum_{a \in \mathcal{A}} F_a(X_i^a),
\end{aligned}$$

and by Hölder's inequality,

$$\mathbb{E}_{\boldsymbol{\theta}}[|\mathbf{b}^T \mathbf{T}_n| \cdot \sum_{i=1}^n \sum_{a \in \mathcal{A}} F_a(X_i^a)] \leq \sum_{i=1}^n \sum_{a \in \mathcal{A}} \left( \mathbb{E}_{\boldsymbol{\theta}}[(\mathbf{b}^T \mathbf{T}_n)^2] \cdot \mathbb{E}_{\boldsymbol{\theta}}[\{F_a(X_i^a)\}^2] \right)^{1/2} < \infty$$

Taking into account that  $\Omega^n$  is independent of  $\boldsymbol{\theta}$ , and by applying the Dominated Convergence Theorem together with the classical proof of differentiation under the integral sign, we arrive at

$$\begin{aligned}
& \nabla_{\boldsymbol{\theta}} \int_{\Omega^n} \mathbf{b}^T T_n(a_1, X_1^{a_1}, \dots, a_n, X_n^{a_n}) f_{\boldsymbol{\theta}}(\mathbf{X}_n^{\mathcal{A}}, \mathbf{a}^n) d\boldsymbol{\mu}^n(\mathbf{X}_n^{\mathcal{A}}) \\
&= \int_{\Omega^n} \mathbf{b}^T T_n(a_1, X_1^{a_1}, \dots, a_n, X_n^{a_n}) \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\mathbf{X}_n^{\mathcal{A}}, \mathbf{a}^n) d\boldsymbol{\mu}^n(\mathbf{X}_n^{\mathcal{A}}).
\end{aligned}$$

In conclusion, we know that

$$\nabla_{\boldsymbol{\theta}} h_{\mathbf{b}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \left[ \mathbf{b}^T \mathbf{T}_n \sum_{i=1}^n \sum_{a \in \mathcal{A}} \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta},a}(X_i^a) \right].$$

Next, we show that

$$\mathbb{E}_{\boldsymbol{\theta}} \left[ \mathbf{b}^T \mathbf{T}_n \sum_{i=1}^n \sum_{a \in \mathcal{A}, a \neq a_i} \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta},a}(X_i^a) \right] = 0. \tag{175}$$

First of all, let  $\mathcal{F}_{i+1} = \sigma\{a_1, X_1^{a_1}, a_2, X_2^{a_2}, \dots, a_i, X_i^{a_i}\}$ . Note that  $a_{i+1}$  is measurable with respect to  $\mathcal{F}_i$  for all  $i$ . Notice that  $\{X_n^a\}_{a \in \mathcal{A}}$  are independent of  $\mathcal{F}_{n-1}$ , as well as  $\{X_n^a\}_{a \in \mathcal{A}, a \neq a_n}$

and  $X_n^{a_n}$  are independent, given  $\mathcal{F}_{n-1}$ . Also recall that  $a_n$  is measurable in  $\mathcal{F}_{n-1}$ . Thus,

$$\begin{aligned} & \mathbb{E}_\theta \left[ \mathbf{b}^T \mathbf{T}_n \sum_{a \in \mathcal{A}, a \neq a_n} \nabla_\theta \log f_{\theta,a}(X_n^a) \middle| \mathcal{F}_{n-1}, X_n^{a_n} \right] \\ &= \mathbf{b}^T \mathbf{T}_n \sum_{a \in \mathcal{A}, a \neq a_n} \mathbb{E}_\theta \left[ \nabla_\theta \log f_{\theta,a}(X_n^a) \middle| \mathcal{F}_{n-1}, X_n^{a_n} \right] \\ &= \mathbf{b}^T \mathbf{T}_n \sum_{a \in \mathcal{A}, a \neq a_n} \mathbb{E}_\theta \left[ \nabla_\theta \log f_{\theta,a}(X_n^a) \right] = 0. \end{aligned}$$

Note that for fixed  $1 \leq i < n$ ,  $\{X_j^a\}_{j \geq i, a \in \mathcal{A}}$  and  $\mathcal{F}_{i-1}$  are independent. Define another  $\sigma$ -algebra,  $\mathcal{G}_{i-1} = \sigma(\mathcal{F}_{i-1}, \{X_j^a\}_{i+1 \leq j \leq n, a \in \mathcal{A}})$ . Note that  $a_i, a_{i+1}, \dots, a_n$  are measurable in  $\sigma(\mathcal{G}_{i-1}, X_i^{a_i})$ . Furthermore,  $\{X_i^a\}_{a \in \mathcal{A}, a \neq a_i}$  and  $X_i^{a_i}$  are independent, given  $\mathcal{G}_{i-1}$ . Thus, for any  $1 \leq i < n$

$$\begin{aligned} & \mathbb{E}_\theta \left[ \mathbf{b}^T \mathbf{T}_n \sum_{a \in \mathcal{A}, a \neq a_i} \nabla_\theta \log f_{\theta,a}(X_i^a) \middle| \mathcal{G}_{i-1}, X_i^{a_i} \right] \\ &= \mathbf{b}^T \mathbf{T}_n \sum_{a \in \mathcal{A}, a \neq a_i} \mathbb{E}_\theta \left[ \nabla_\theta \log f_{\theta,a}(X_i^a) \middle| \mathcal{G}_{i-1}, X_i^{a_i} \right] \\ &= \mathbf{b}^T \mathbf{T}_n \sum_{a \in \mathcal{A}, a \neq a_i} \mathbb{E}_\theta \left[ \nabla_\theta \log f_{\theta,a}(X_i^a) \right] = 0. \end{aligned}$$

By the law of iterated expectation, we have proved (175). Hence, we know that

$$\nabla_\theta h_{\mathbf{b}}(\theta) = \mathbb{E}_\theta \left[ \mathbf{b}^T \mathbf{T}_n \nabla_\theta \sum_{i=1}^n \log f_{\theta,a_i}(X_i^{a_i}) \right] = 0.$$

Set  $\mathbf{Y}_n = \nabla_\theta \sum_{i=1}^n \log f_{\theta,a_i}(X_i^{a_i})$ , and we have

$$\nabla_\theta h_{\mathbf{b}}(\theta) = \mathbb{E}_\theta[\mathbf{Y}_n \mathbf{T}_n^T \mathbf{b}] = \text{cov}_\theta(\mathbf{Y}_n, \mathbf{b}^T \mathbf{T}_n).$$

By multivariate Cauchy-Schwartz inequality (75), for any  $\mathbf{b} \in \mathbb{R}^l$ ,

$$\begin{aligned} & \mathbf{b}^T \text{cov}_\theta(\mathbf{T}_n) \mathbf{b} \\ &= \text{var}_\theta(\mathbf{b}^T \mathbf{T}_n) \\ &\geq \text{cov}_\theta \left( \mathbf{b}^T \mathbf{T}_n, \mathbf{Y}_n \right) \{ \text{cov}_\theta(\mathbf{Y}_n) \}^{-1} \text{cov}_\theta \left( \mathbf{Y}_n, \mathbf{b}^T \mathbf{T}_n \right) \\ &= \mathbf{b}^T \{ \nabla_\theta h(\theta) \}^T \{ \text{cov}_\theta(\mathbf{Y}_n) \}^{-1} \nabla_\theta h(\theta) \mathbf{b}. \end{aligned}$$

Note that

$$\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{Y}_{i-1}\{\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta},a_i}(X_i^{a_i})\}^T] = \mathbb{E}_{\boldsymbol{\theta}}\left\{\mathbf{Y}_{i-1} \cdot \mathbb{E}_{\boldsymbol{\theta}}[\{\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta},a_i}(X_i^{a_i})\}^T | \mathcal{F}_{i-1}]\right\} = 0.$$

Thus

$$\text{cov}_{\boldsymbol{\theta}}(\mathbf{Y}_n) = \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{Y}_n \mathbf{Y}_n^T] = \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{Y}_{n-1} \mathbf{Y}_{n-1}^T] + \mathbb{E}_{\boldsymbol{\theta}}[\mathcal{I}_{a_i}(\boldsymbol{\theta})] = n \cdot \mathcal{I}^{\mathbb{E}_{\boldsymbol{\theta}} \bar{\pi}_n}(\boldsymbol{\theta}).$$

In conclusion, for any  $\mathbf{b} \in \mathbb{R}^l$ , we obtain that

$$\mathbf{b}^T \text{cov}_{\boldsymbol{\theta}}(\mathbf{T}_n) \mathbf{b} = \text{var}_{\boldsymbol{\theta}}(\mathbf{b}^T \mathbf{T}_n) \geq \mathbf{b}^T \left[ \frac{1}{n} \left\{ \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}) \right\}^T \left\{ \mathcal{I}^{\mathbb{E}_{\boldsymbol{\theta}} \bar{\pi}_n}(\boldsymbol{\theta}) \right\}^{-1} \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}) \right] \mathbf{b}.$$

This implies that

$$\text{cov}_{\boldsymbol{\theta}}(\mathbf{T}_n) \succeq \frac{1}{n} \left\{ \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}) \right\}^T \left\{ \mathcal{I}^{\mathbb{E}_{\boldsymbol{\theta}} \bar{\pi}_n}(\boldsymbol{\theta}) \right\}^{-1} \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}).$$

If  $h(\boldsymbol{\theta}) = \boldsymbol{\theta}$ , we know that

$$n \text{cov}_{\boldsymbol{\theta}}(\mathbf{T}_n) \succeq \left\{ \mathcal{I}^{\mathbb{E}_{\boldsymbol{\theta}} \bar{\pi}_n}(\boldsymbol{\theta}) \right\}^{-1}.$$

By assumption 5, we obtain

$$\mathbb{G}_{\boldsymbol{\theta}}(n \text{cov}_{\boldsymbol{\theta}}(\mathbf{T}_n)) \geq \mathbb{G}_{\boldsymbol{\theta}}(\left\{ \mathcal{I}^{\mathbb{E}_{\boldsymbol{\theta}} [\bar{\pi}_n]}(\boldsymbol{\theta}) \right\}^{-1}) \geq \inf_{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}} \mathbb{G}_{\boldsymbol{\theta}}(\left\{ \mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}) \right\}^{-1}).$$

□

*Proof of Theorem 4.8.*

**Part 1** Notice that  $L(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$  is a loss function, which means that  $L(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = 0$ . Due to  $L(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$  is differentiable in  $\hat{\boldsymbol{\theta}}$ , we know that  $\nabla_{\hat{\boldsymbol{\theta}}} L(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = \mathbf{0}$ .

Applying first order Taylor expansion to  $L(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$  with respect to  $\hat{\boldsymbol{\theta}}$ , we obtain that

$$L(\boldsymbol{\theta}^*, \mathbf{T}_n) = \frac{1}{2} \left\langle \nabla_{\hat{\boldsymbol{\theta}}}^2 L(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) \Big|_{\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}_n} (\boldsymbol{\theta}^* - \mathbf{T}_n), \boldsymbol{\theta}^* - \mathbf{T}_n \right\rangle \geq \eta \|\boldsymbol{\theta}^* - \mathbf{T}_n\|^2, \quad (176)$$

where  $\tilde{\boldsymbol{\theta}}_n = t_n \boldsymbol{\theta}^* + (1 - t_n) \mathbf{T}_n$  for some  $t_n \in (0, 1)$ . Thus,

$$\mathbb{E}_{\boldsymbol{\theta}^*} n \cdot L(\boldsymbol{\theta}^*, \mathbf{T}_n) \geq \eta \mathbb{E}_{\boldsymbol{\theta}^*} \left\| \sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}^*) \right\|^2.$$

To show (26), without loss of generality, we assume that

$$\limsup_{n \rightarrow \infty} \mathbb{E}_{\boldsymbol{\theta}^*} \left\| \sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}^*) \right\|^2 < \infty.$$

This implies  $\mathbf{T}_n \xrightarrow{\mathbb{P}_*} \boldsymbol{\theta}^*$ . It also implies that  $\mathbf{T}_n$  has finite second moment, and, thus, conditions of Lemma 14.22 are satisfied. By Lemma 14.22, we have  $\text{cov}_{\boldsymbol{\theta}}(\mathbf{T}_n) \succeq \frac{1}{n} \{\mathcal{I}^{\mathbb{E}\pi_n}(\boldsymbol{\theta})\}^{-1}$ .

If  $L(\boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}) \equiv \langle H_{\boldsymbol{\theta}^*}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}), \boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}} \rangle$ , we obtain

$$\mathbb{E}_{\boldsymbol{\theta}^*}[nL(\boldsymbol{\theta}^*, \mathbf{T}_n)] = n\langle H_{\boldsymbol{\theta}^*}, \text{cov}_{\boldsymbol{\theta}^*}(\mathbf{T}_n) \rangle \geq \text{tr}(H_{\boldsymbol{\theta}^*} \{\mathcal{I}^{\mathbb{E}\pi_n}(\boldsymbol{\theta}^*)\}^{-1}).$$

If  $L(\boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}) \neq \langle H_{\boldsymbol{\theta}^*}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}), \boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}} \rangle$ , under the theorem's assumption

$$\limsup_{n \rightarrow \infty} \mathbb{E}_{\boldsymbol{\theta}^*} n \|\mathbf{T}_n - \boldsymbol{\theta}^*\|^2 I(\|\mathbf{T}_n - \boldsymbol{\theta}^*\| > \varepsilon) = 0.$$

Define  $\mathbf{V}_n = \sqrt{n}(\boldsymbol{\theta}^* - \mathbf{T}_n)$ , and its truncation  $\mathbf{V}_n^M = \mathbf{V}_n I(\|\mathbf{V}_n\| \leq M)$ . Define

$$H(\boldsymbol{\theta}^*, \mathbf{T}_n) = \frac{1}{2} \nabla_{\widehat{\boldsymbol{\theta}}}^2 L(\boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}) \Big|_{\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_n},$$

where  $\widehat{\boldsymbol{\theta}}_n = t_n \boldsymbol{\theta}^* + (1 - t_n) \mathbf{T}_n$ . According to (176),  $L(\boldsymbol{\theta}^*, \mathbf{T}_n) = \langle H(\boldsymbol{\theta}^*, \mathbf{T}_n)(\boldsymbol{\theta}^* - \mathbf{T}_n), \boldsymbol{\theta}^* - \mathbf{T}_n \rangle$ . Furthermore, for any  $\varepsilon > 0$ ,

$$\begin{aligned} & \mathbb{E}_{\boldsymbol{\theta}^*} \left| n \cdot L(\boldsymbol{\theta}^*, \mathbf{T}_n) - \langle H_{\boldsymbol{\theta}^*} \mathbf{V}_n, \mathbf{V}_n \rangle \right| I(\|\mathbf{T}_n - \boldsymbol{\theta}^*\| \leq \varepsilon) \\ & \leq \max_{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \leq \varepsilon} \left\| H(\boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}) - H_{\boldsymbol{\theta}^*} \right\|_{op} \mathbb{E}_{\boldsymbol{\theta}^*} \|\mathbf{V}_n\|^2 \\ & = o(1). \end{aligned}$$

Now, we obtain that

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}^*} [n \cdot L(\boldsymbol{\theta}^*, \mathbf{T}_n)] & \geq \mathbb{E}_{\boldsymbol{\theta}^*} \left[ n \cdot L(\boldsymbol{\theta}^*, \mathbf{T}_n) \cdot I(\|\mathbf{V}_n\| \leq \varepsilon \sqrt{n}) \right] \\ & \geq \mathbb{E}_{\boldsymbol{\theta}^*} \left\langle H_{\boldsymbol{\theta}^*} \mathbf{V}_n^{\varepsilon \sqrt{n}}, \mathbf{V}_n^{\varepsilon \sqrt{n}} \right\rangle - \max_{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \leq \varepsilon} \left\| H(\boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}) - H_{\boldsymbol{\theta}^*} \right\|_{op} \mathbb{E}_{\boldsymbol{\theta}^*} \|\mathbf{V}_n\|^2 \\ & = \mathbb{E}_{\boldsymbol{\theta}^*} \langle H_{\boldsymbol{\theta}^*} \mathbf{V}_n, \mathbf{V}_n \rangle - \mathbb{E}_{\boldsymbol{\theta}^*} n \|\mathbf{T}_n - \boldsymbol{\theta}^*\|^2 I(\|\mathbf{T}_n - \boldsymbol{\theta}^*\| > \varepsilon) - o(1) \\ & \geq \min_{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}} \text{tr}(H_{\boldsymbol{\theta}^*} \{\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*)\}^{-1}) - \mathbb{E}_{\boldsymbol{\theta}^*} n \|\mathbf{T}_n - \boldsymbol{\theta}^*\|^2 I(\|\mathbf{T}_n - \boldsymbol{\theta}^*\| > \varepsilon) - o(1), \end{aligned}$$

where the last inequality is due to Lemma 14.22. Taking the inferior limit as  $n \rightarrow \infty$  and then taking the inferior limit as  $\varepsilon \rightarrow 0^+$ , we obtain

$$\liminf_{n \rightarrow \infty} \mathbb{E}_{\boldsymbol{\theta}^*} [n \cdot L(\boldsymbol{\theta}^*, \mathbf{T}_n)] \geq \min_{\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}} \text{tr}(H_{\boldsymbol{\theta}^*} \{\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*)\}^{-1}).$$

The ‘in particular’ part is proved by noting that  $H_{\boldsymbol{\theta}^*} = I_p$  in this case.

**Part 2** Recall the log-likelihood defined in 4. Because  $f_{\theta,a}(\cdot) = h_{\xi_{a,a}}(\cdot)$ , we obtain that

$$-\nabla_{\theta}^2 l_n(\theta; \mathbf{a}_n) = -\frac{1}{n} \sum_{i=1}^n \mathbf{Z}_{a_i}^T \nabla_{\xi_{a_i}}^2 \log h_{\xi_{a_i,a_i}}(X_i) \mathbf{Z}_{a_i} \succeq \alpha \sum_{a \in \mathcal{A}} \bar{\pi}_n(a) \mathbf{Z}_a^T \mathbf{Z}_a, \quad (177)$$

and  $\mathcal{I}_{\xi_{a,a}}(\xi_a) = -\mathbb{E}_{X^a \sim h_{\xi_{a,a}}} \nabla_{\xi_a}^2 \log h_{\xi_{a,a}}(X^a) \succeq \alpha I$ , where  $\xi_a = \mathbf{Z}_a \theta$ .

Under Assumption 6A, we obtain  $\mathcal{I}_a(\theta) = \mathbf{Z}_a^T \mathcal{I}_{\xi_{a,a}}(\xi_a) \mathbf{Z}_a$  and

$$\nabla_{\theta}^2 \log f_{\theta,a}(X^a) = \mathbf{Z}_a^T \nabla_{\xi_a}^2 \log h_{\xi_{a,a}}(X^a) \mathbf{Z}_a,$$

and

$$\mathcal{I}^{\pi}(\theta^*) = \sum_{a \in \mathcal{A}} \pi(a) \mathbf{Z}_a^T \mathcal{I}_{\xi_{a,a}}(\xi_a^*) \mathbf{Z}_a \succeq \alpha \sum_{a \in \mathcal{A}} \pi(a) \mathbf{Z}_a^T \mathbf{Z}_a$$

is a positive definite matrix.

Applying the first order Taylor expansion of  $L(\theta^*, \hat{\theta})$  over  $\hat{\theta}$ , we obtain that

$$L(\theta^*, \hat{\theta}_n) = \frac{1}{2} \left\langle \nabla_{\hat{\theta}}^2 L(\theta^*, \hat{\theta}) \Big|_{\hat{\theta}=\tilde{\theta}_n} (\theta^* - \hat{\theta}_n), \theta^* - \hat{\theta}_n \right\rangle \leq \eta' \left\| \theta^* - \hat{\theta}_n \right\|^2, \quad (178)$$

where  $\tilde{\theta}_n = t_n \theta^* + (1 - t_n) \hat{\theta}_n$  for some  $t_n \in (0, 1)$ . Recall that  $\mathbb{G}_{\theta}(\Sigma) = \text{tr}(H_{\theta} \Sigma)$ . Note that  $\nabla \mathbb{G}_{\theta}(\Sigma) = H_{\theta}$  and  $\kappa(H_{\theta}) \leq \frac{\eta'}{\eta} < \infty$ , which implies that  $\mathbb{G}_{\theta}(\Sigma)$  satisfies Assumption 5.

By Theorem 4.3,  $\bar{\pi}_n \xrightarrow{\mathbb{P}^*} \pi^* = \arg \min_{\pi \in \mathcal{S}^{\mathcal{A}}} \mathbb{F}_{\theta^*}(\pi)$  and  $\mathcal{I}^{\pi^*}(\theta)$  is nonsingular for any  $\theta \in \Theta$ , applying Theorem 4.2, we obtain

$$\sqrt{n}(\hat{\theta}_n^{\text{ML}} - \theta^*) \xrightarrow{d} N_p\left(0, \{\mathcal{I}^{\pi^*}(\theta^*)\}^{-1}\right) \text{ as } n \rightarrow \infty. \quad (179)$$

Notice that for any  $n \geq n_0$ , by Lemma 14.1 and Assumption 6B, there exists  $\underline{C} > 0$  such that

$$-\nabla_{\theta}^2 l_n(\theta) \succeq \alpha \sum_{a \in \mathcal{A}} \bar{\pi}_n(a) \mathbf{Z}_a^T \mathbf{Z}_a \succeq \alpha \inf_{n \geq n_0} \lambda_{\min}(\bar{\pi}_n(a) \mathbf{Z}_a^T \mathbf{Z}_a) I_p \succeq 2\underline{C} I_p.$$

By Taylor expansion, we obtain

$$0 \leq l_n(\hat{\theta}_n^{\text{ML}}; \mathbf{a}_n) - l_n(\theta^*; \mathbf{a}_n) \leq \left\langle \nabla_{\theta} l_n(\theta^*; \mathbf{a}_n), \hat{\theta}_n^{\text{ML}} - \theta^* \right\rangle - \underline{C} \left\| \hat{\theta}_n^{\text{ML}} - \theta^* \right\|^2.$$

Thus,

$$\left\| \hat{\theta}_n^{\text{ML}} - \theta^* \right\| \leq \frac{1}{\underline{C}} \left\| \nabla_{\theta} l_n(\theta^*; \mathbf{a}_n) \right\|.$$

By Theorem 6.2 in DasGupta (2008), to show that

$$\mathbb{E}_{\boldsymbol{\theta}^*} \left[ n \left\| \widehat{\boldsymbol{\theta}}_n^{\text{ML}} - \boldsymbol{\theta}^* \right\|^2 \right] \rightarrow \text{tr}(\{\mathcal{I}^\pi(\boldsymbol{\theta}^*)\}^{-1}),$$

combined with (179), it suffices to show that

$$\limsup_{n \rightarrow \infty} \mathbb{E}_{\boldsymbol{\theta}^*} \left( \sqrt{n} \left\| \widehat{\boldsymbol{\theta}}_n^{\text{ML}} - \boldsymbol{\theta}^* \right\| \right)^{2+\delta} < \infty.$$

Note that

$$\mathbb{E}_{\boldsymbol{\theta}^*} \left( \sqrt{n} \left\| \widehat{\boldsymbol{\theta}}_n^{\text{ML}} - \boldsymbol{\theta}^* \right\| \right)^{2+\delta} \leq \frac{1}{(C)^{1+\delta}} \mathbb{E}_{\boldsymbol{\theta}^*} \left\| \sqrt{n} \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) \right\|^{2+\delta}.$$

By classical  $c_r$ -inequality (see Chapter 9 of Lin (2010)), we have

$$\mathbb{E}_{\boldsymbol{\theta}^*} \left\| \sqrt{n} \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) \right\|^{2+\delta} \leq p^{\delta/2} \sum_{j=1}^p \mathbb{E}_{\boldsymbol{\theta}^*} \left| \sum_{i=1}^n \frac{1}{\sqrt{n}} \mathbf{e}_j^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*, a_i}(X_i) \right|^{2+\delta}.$$

Since  $\sum_{i=1}^n \mathbf{e}_j^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*, a_i}(X_i)$  is a martingale, applying inequality (45) in Lin (2010), we obtain

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}^*} \left| \sum_{i=1}^n \frac{1}{\sqrt{n}} \mathbf{e}_j^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*, a_i}(X_i) \right|^{2+\delta} &\leq C_{2+\delta} \cdot n^{\delta/2} \sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta}^*} \left| \frac{1}{\sqrt{n}} \mathbf{e}_j^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*, a_i}(X_i) \right|^{2+\delta} \\ &\leq C_{2+\delta} \sum_{a \in \mathcal{A}} \mathbb{E}_{X^a \sim f_{\boldsymbol{\theta}^*, a}} \left\| \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*, a}(X^a) \right\|^{2+\delta} \\ &\leq C_{2+\delta} \sum_{a \in \mathcal{A}} \mathbb{E}_{X^a \sim f_{\boldsymbol{\theta}^*, a}} \left\| \nabla_{\boldsymbol{\theta}} \log h_{\boldsymbol{\xi}_a^*, a}(X^a) \right\|^{2+\delta} \left\| \mathbf{Z}_a \right\|_{op}^{2+\delta}. \end{aligned}$$

In conclusion, we obtain

$$\sup_{n \geq n_0} \mathbb{E}_{\boldsymbol{\theta}^*} \left( \sqrt{n} \left\| \widehat{\boldsymbol{\theta}}_n^{\text{ML}} - \boldsymbol{\theta}^* \right\| \right)^{2+\delta} < \infty. \quad (180)$$

Notice that as  $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathbb{P}_*} \boldsymbol{\theta}^*$ , we know that

$$\frac{1}{2} \nabla_{\boldsymbol{\theta}}^2 L(\boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}) \Big|_{\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_n} \xrightarrow{\mathbb{P}_*} H_{\boldsymbol{\theta}^*}.$$

Thus, we obtain that

$$nL(\boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}_n^{\text{ML}}) = n \left\langle H_{\boldsymbol{\theta}^*}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_n^{\text{ML}}), \boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_n^{\text{ML}} \right\rangle + o_p(1). \quad (181)$$



By (180) and (178), we obtain that

$$\sup_{n \geq n_0} \mathbb{E}_{\boldsymbol{\theta}^*} \left[ nL(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_n^{\text{ML}}) \right]^{1+\delta/2} \leq (\eta')^{1+\delta/2} \sup_{n \geq n_0} \mathbb{E}_{\boldsymbol{\theta}^*} \left( \sqrt{n} \left\| \hat{\boldsymbol{\theta}}_n^{\text{ML}} - \boldsymbol{\theta}^* \right\| \right)^{2+\delta} < \infty.$$

Applying Theorem 6.2 in DasGupta (2008), we obtain that as  $n \rightarrow \infty$ ,

$$\mathbb{E}_{\boldsymbol{\theta}^*} \left[ nL(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_n^{\text{ML}}) \right] \rightarrow \mathbb{E} \langle H_{\boldsymbol{\theta}^*} \mathbf{V}, \mathbf{V} \rangle = \text{tr}(H_{\boldsymbol{\theta}^*} \{\mathcal{I}^\pi(\boldsymbol{\theta}^*)\}^{-1}), \mathbf{V} \sim N_p(\mathbf{0}_p, \{\mathcal{I}^\pi(\boldsymbol{\theta}^*)\}^{-1}).$$

Applying Theorem 4.7, the proof of the second part of Theorem 4.8 is completed.  $\square$

## 14.9 Proof of Theorem 4.9

*Proof of Theorem 4.9.* The proof of Theorem 4.9 is similar to that of Theorem 8.8 and Theorem 8.11 in Van der Vaart (2000). Thus, we will only state the main differences and omit the repetitive details.

For proving the first part of the theorem, we follow the proof of Theorem 8.8 in Van der Vaart (2000). We need to verify Theorem 8.3, Theorem 7.10, as well as Proposition 8.4, as presented in Van der Vaart (2000), under our sequential setting.

For proving the second part of the theorem, we follow the proof of Theorem 8.11 in Van der Vaart (2000). It is sufficient to modify and prove Theorem 7.2 and Proposition 8.6, as presented in Van der Vaart (2000), under our sequential setting.

Below we verify the above mentioned results in our context.

**Differentiable in quadratic mean** We need to show that densities  $\{f_{\boldsymbol{\theta},a}(\cdot)\}_{a \in \mathcal{A}}$  are differentiable in quadratic mean at  $\boldsymbol{\theta}$ , which means that

$$\int \left[ \sqrt{f_{\boldsymbol{\theta}+\mathbf{h},a}(x)} - \sqrt{f_{\boldsymbol{\theta},a}(x)} - \frac{1}{2} \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta},a}(x) \sqrt{f_{\boldsymbol{\theta},a}(x)} \right]^2 d\mu(x) = o(\|\mathbf{h}\|^2). \quad (182)$$

By applying the regularity conditions and using Lemma 7.6 from Van der Vaart (2000), we have completed the proof of (182) for any  $a \in \mathcal{A}$  and  $\boldsymbol{\theta}$ , where  $\boldsymbol{\theta}$  is an interior point of  $\boldsymbol{\Theta}$ .

**Modified Theorem 7.2 in Van der Vaart (2000)** We modified Theorem 7.2 in Van der Vaart (2000) in our context as follows. Let  $P_{n,\boldsymbol{\theta}}$  denote the joint distribution of  $(a_1, X_1, \dots, a_n, X_n)$  following some experiment selection rule with the empirical selection

proportion  $\bar{\pi}_n$ . Then, given that  $\mathbf{h}_n = \mathbf{h} + o(1)$ ,

$$\begin{aligned}
& \log \frac{P_{n, \boldsymbol{\theta} + \mathbf{h}_n / \sqrt{n}}(a_1, X_1, \dots, a_n, X_n)}{P_{n, \boldsymbol{\theta}}} \\
& \sim \log \prod_{j=1}^n \frac{f_{\boldsymbol{\theta} + \mathbf{h}_n / \sqrt{n}, a_j}(X_j^{a_j})}{f_{\boldsymbol{\theta}, a_j}(X_j^{a_j})} \\
& = \frac{1}{\sqrt{n}} \sum_{j=1}^n \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a_j}(X_j^{a_j}) - \frac{1}{2} \mathbf{h}^T \mathcal{I}^{\pi}(\boldsymbol{\theta}) \mathbf{h} + o_p(1)
\end{aligned} \tag{183}$$

where  $\{X_j^a\}_{a \in \mathcal{A}, j \geq 1}$ , where  $X_j^a \sim f_{\boldsymbol{\theta}, a}(\cdot)$  are independent random variables, ‘ $\sim$ ’ means that random variables on both sides share the same distribution, the second line is due to Lemma 13.8, and the last line is obtained following a similar proof as that of Theorem 7.2 in Van der Vaart (2000), which is detailed below.

By Assumptions 1-4, the Dominated Convergence Theorem, and the proof of the classical differentiation under the integral sign, we arrive at

$$\mathbb{E}[\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a_j}(X_j^{a_j}) | \mathcal{F}_{j-1}] = \mathbf{0} \text{ and } \mathbb{E}\left[\sum_{j=1}^n \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a_j}(X_j^{a_j})\right] = \mathbf{0}.$$

The proof of the first Equation (7.3) in Van der Vaart (2000) needs to be modified as follows.

Let  $W_{nj} = 2\left(\sqrt{\frac{f_{\boldsymbol{\theta} + \mathbf{h}_n / \sqrt{n}, a_j}(X_j^{a_j})}{f_{\boldsymbol{\theta}, a_j}(X_j^{a_j})}} - 1\right)$  and  $V_n = \sum_{j=1}^n W_{nj} - \frac{1}{\sqrt{n}} \mathbf{h}^T \sum_{j=1}^n \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a_j}(X_j^{a_j})$ . We know that

$$\begin{aligned}
& \text{var}(V_n) \\
& = \text{var}(V_{n-1}) + \text{var}\left(W_{nn} - \frac{1}{\sqrt{n}} \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a_n}(X_n^{a_n})\right) + 2 \text{cov}\left(V_{n-1}, W_{nn} - \frac{1}{\sqrt{n}} \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a_n}(X_n^{a_n})\right)
\end{aligned}$$

and

$$\begin{aligned}
& \text{cov}\left(V_{n-1}, W_{nn} - \frac{1}{\sqrt{n}} \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a_n}(X_n^{a_n})\right) \\
& = \mathbb{E}\left[(V_{n-1} - \mathbb{E}V_{n-1})\left(W_{nn} - \frac{1}{\sqrt{n}} \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a_n}(X_n^{a_n}) - \mathbb{E}\left(W_{nn} - \frac{1}{\sqrt{n}} \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a_n}(X_n^{a_n})\right)\right)\right] \\
& = \mathbb{E}\left\{(V_{n-1} - \mathbb{E}V_{n-1})\mathbb{E}\left[\left(W_{nn} - \frac{1}{\sqrt{n}} \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a_n}(X_n^{a_n}) - \mathbb{E}\left(W_{nn} - \frac{1}{\sqrt{n}} \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a_n}(X_n^{a_n})\right)\right) \middle| \mathcal{F}_{n-1}\right]\right\} \\
& = 0.
\end{aligned}$$

By induction and (182), we obtain that as  $n \rightarrow \infty$ ,

$$\begin{aligned}
\text{var} \left( V_n \right) &= \sum_{j=1}^n \text{var} \left( W_{nj} - \frac{1}{\sqrt{n}} \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a_j}(X_j^{a_j}) \right) \\
&\leq \sum_{j=1}^n \mathbb{E} \left( W_{nj} - \frac{1}{\sqrt{n}} \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a_j}(X_j^{a_j}) \right)^2 \\
&\leq \sum_{j=1}^n \sum_{a \in \mathcal{A}} \mathbb{E} \left( 2 \left( \sqrt{\frac{f_{\boldsymbol{\theta} + \mathbf{h}_n / \sqrt{n}, a}(X_j^a)}{f_{\boldsymbol{\theta}, a}(X_j^a)}} - 1 \right) - \frac{1}{\sqrt{n}} \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(X_j^a) \right)^2 \\
&\leq 8n \sum_{a \in \mathcal{A}} \int \left[ \sqrt{f_{\boldsymbol{\theta} + \mathbf{h}_n / \sqrt{n}, a}(x)} - \sqrt{f_{\boldsymbol{\theta}, a}(x)} - \frac{1}{2\sqrt{n}} \mathbf{h}_n^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(x) \sqrt{f_{\boldsymbol{\theta}, a}(x)} \right]^2 d\mu(x) \\
&\quad + 2(\mathbf{h} - \mathbf{h}_n)^T \sum_{a \in \mathcal{A}} \mathcal{I}_a(\boldsymbol{\theta})(\mathbf{h} - \mathbf{h}_n) \\
&= 8o(\|\mathbf{h}\|^2) + 2(\mathbf{h} - \mathbf{h}_n)^T \sum_{a \in \mathcal{A}} \mathcal{I}_a(\boldsymbol{\theta})(\mathbf{h} - \mathbf{h}_n) \rightarrow 0.
\end{aligned}$$

Because of (182), we obtain that

$$\begin{aligned}
&\left\| \left\| \sqrt{f_{\boldsymbol{\theta} + \mathbf{h}_n / \sqrt{n}, a}(x)} - \sqrt{f_{\boldsymbol{\theta}, a}(x)} \right\|_{L^2(\mu)} - \left\| \frac{1}{2\sqrt{n}} \mathbf{h}_n^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(x) \sqrt{f_{\boldsymbol{\theta}, a}(x)} \right\|_{L^2(\mu)} \right\| \\
&\leq \left( \int \left[ \sqrt{f_{\boldsymbol{\theta} + \mathbf{h}_n / \sqrt{n}, a}(x)} - \sqrt{f_{\boldsymbol{\theta}, a}(x)} - \frac{1}{2\sqrt{n}} \mathbf{h}_n^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(x) \sqrt{f_{\boldsymbol{\theta}, a}(x)} \right]^2 d\mu(x) \right)^{1/2} = o\left(\frac{\|\mathbf{h}\|}{\sqrt{n}}\right).
\end{aligned}$$

Note that

$$\left\| \frac{1}{2\sqrt{n}} \mathbf{h}_n^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(x) \sqrt{f_{\boldsymbol{\theta}, a}(x)} \right\|_{L^2(\mu)}^2 = \frac{1}{4n} \mathbf{h}_n^T \mathcal{I}_a(\boldsymbol{\theta}) \mathbf{h}_n = O\left(\frac{\|\mathbf{h}\|^2}{n}\right).$$

By inequality  $|x^2 - y^2| \leq |x - y|(|x| + |y|) \leq |x - y|(2|x| + |x - y|)$ , we obtain

$$\begin{aligned}
&\left| \left\| \sqrt{f_{\boldsymbol{\theta} + \mathbf{h}_n / \sqrt{n}, a}(x)} - \sqrt{f_{\boldsymbol{\theta}, a}(x)} \right\|_{L^2(\mu)}^2 - \frac{1}{4n} \mathbf{h}_n^T \mathcal{I}_a(\boldsymbol{\theta}) \mathbf{h}_n \right| \\
&\leq o\left(\frac{\|\mathbf{h}\|}{\sqrt{n}}\right) \left| 2 \cdot O\left(\frac{\|\mathbf{h}\|}{\sqrt{n}}\right) + o\left(\frac{\|\mathbf{h}\|}{\sqrt{n}}\right) \right| = o\left(\frac{\|\mathbf{h}\|^2}{n}\right).
\end{aligned}$$

Thus, the second Equation (7.3) in Van der Vaart (2000) is modified by

$$\begin{aligned}
\mathbb{E}[W_{nj}|\mathcal{F}_{j-1}] &= 2\left(\int \sqrt{f_{\boldsymbol{\theta}+\mathbf{h}_n/\sqrt{n},a_j}(x)f_{\boldsymbol{\theta},a_j}(x)}d\mu(x) - 1\right) \\
&= -\int \left[\sqrt{f_{\boldsymbol{\theta}+\mathbf{h}_n/\sqrt{n},a_j}(x)} - \sqrt{f_{\boldsymbol{\theta},a_j}(x)}\right]^2 d\mu(x) = -\frac{1}{4n}\mathbf{h}_n^T \mathcal{I}_{a_j}(\boldsymbol{\theta})\mathbf{h}_n + o\left(\frac{1}{n}\right) \\
&= -\frac{1}{4n}\mathbf{h}^T \mathcal{I}_{a_j}(\boldsymbol{\theta})\mathbf{h} + \frac{1}{n}o(1),
\end{aligned} \tag{184}$$

where the  $o(1)$  converges uniformly to 0 as  $n \rightarrow \infty$ . Now, we obtain that

$$\mathbb{E}\bar{\boldsymbol{\pi}}_n = \boldsymbol{\pi} + o(1), \text{ and } \mathbb{E} \sum_{j=1}^n W_{nj} = -\frac{1}{4}\mathbf{h}^T \mathcal{I}^{\mathbb{E}\bar{\boldsymbol{\pi}}_n}(\boldsymbol{\theta})\mathbf{h} + o(1) = -\frac{1}{4}\mathbf{h}^T \mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta})\mathbf{h} + o(1).$$

We define

$$A_{ni} = nW_{ni}^2 - \left(\mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta},a_i}(X_i^{a_i})\right)^2$$

and

$$A'_{ni} = \sum_{a \in \mathcal{A}} \left| 4n \left( \sqrt{\frac{f_{\boldsymbol{\theta}+\mathbf{h}_n/\sqrt{n},a}(X_i^a)}{f_{\boldsymbol{\theta},a}(X_i^a)}} - 1 \right)^2 - \left( \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta},a}(X_i^a) \right)^2 \right|.$$

Notice that

$$\begin{aligned}
A'_{ni} &= \sum_{a \in \mathcal{A}} \left| 4n \left( \sqrt{\frac{f_{\boldsymbol{\theta}+\mathbf{h}_n/\sqrt{n},a}(X_i^a)}{f_{\boldsymbol{\theta},a}(X_i^a)}} - 1 \right)^2 - \left( \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta},a}(X_i^a) \right)^2 \right| \\
&\leq \sum_{a \in \mathcal{A}} \left| 2\sqrt{n} \left( \sqrt{\frac{f_{\boldsymbol{\theta}+\mathbf{h}_n/\sqrt{n},a}(X_i^a)}{f_{\boldsymbol{\theta},a}(X_i^a)}} - 1 \right) - \left( \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta},a}(X_i^a) \right) \right| \\
&\quad \left| 2\sqrt{n} \left( \sqrt{\frac{f_{\boldsymbol{\theta}+\mathbf{h}_n/\sqrt{n},a}(X_i^a)}{f_{\boldsymbol{\theta},a}(X_i^a)}} - 1 \right) + \left( \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta},a}(X_i^a) \right) \right|
\end{aligned}$$

By Hölder's inequality and the definition of differentiable in quadratic mean at  $\boldsymbol{\theta}$ , we obtain

that

$$\begin{aligned}
\mathbb{E}|A'_{ni}| &\leq \sum_{a \in \mathcal{A}} \left( \mathbb{E} \left| 2\sqrt{n} \left( \sqrt{\frac{f_{\boldsymbol{\theta} + \mathbf{h}_n/\sqrt{n}, a}(X_i^a)}{f_{\boldsymbol{\theta}, a}(X_i^a)}} - 1 \right) - \left( \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(X_i^a) \right) \right|^2 \right)^{1/2} \\
&\quad \left( \mathbb{E} \left| 2\sqrt{n} \left( \sqrt{\frac{f_{\boldsymbol{\theta} + \mathbf{h}_n/\sqrt{n}, a}(X_i^a)}{f_{\boldsymbol{\theta}, a}(X_i^a)}} - 1 \right) + \left( \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(X_i^a) \right) \right|^2 \right)^{1/2} \\
&\leq \sum_{a \in \mathcal{A}} \left( \mathbb{E} \left| 2\sqrt{n} \left( \sqrt{\frac{f_{\boldsymbol{\theta} + \mathbf{h}_n/\sqrt{n}, a}(X_i^a)}{f_{\boldsymbol{\theta}, a}(X_i^a)}} - 1 \right) - \left( \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(X_i^a) \right) \right|^2 \right)^{1/2} \\
&\quad \left[ \left( \mathbb{E} \left| 2\sqrt{n} \left( \sqrt{\frac{f_{\boldsymbol{\theta} + \mathbf{h}_n/\sqrt{n}, a}(X_i^a)}{f_{\boldsymbol{\theta}, a}(X_i^a)}} - 1 \right) - \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(X_i^a) \right|^2 \right)^{1/2} \right. \\
&\quad \left. + 2 \left( \mathbb{E} \left| \left( \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(X_i^a) \right) \right|^2 \right)^{1/2} \right].
\end{aligned}$$

Due to

$$\begin{aligned}
&\left( \mathbb{E} \left| 2\sqrt{n} \left( \sqrt{\frac{f_{\boldsymbol{\theta} + \mathbf{h}_n/\sqrt{n}, a}(X_i^a)}{f_{\boldsymbol{\theta}, a}(X_i^a)}} - 1 \right) - \left( \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(X_i^a) \right) \right|^2 \right)^{1/2} \\
&\leq \left( \mathbb{E} \left| 2\sqrt{n} \left( \sqrt{\frac{f_{\boldsymbol{\theta} + \mathbf{h}_n/\sqrt{n}, a}(X_i^a)}{f_{\boldsymbol{\theta}, a}(X_i^a)}} - 1 \right) - \left( \mathbf{h}_n^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(X_i^a) \right) \right|^2 \right)^{1/2} \\
&\quad + \left( \mathbb{E} \left| \left( \mathbf{h}_n - \mathbf{h} \right)^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(X_i^a) \right|^2 \right)^{1/2} \\
&= o(\|\mathbf{h}_n\|) + \left( (\mathbf{h} - \mathbf{h}_n)^T \mathcal{I}_a(\boldsymbol{\theta}) (\mathbf{h} - \mathbf{h}_n) \right)^{1/2} = o(1),
\end{aligned}$$

we obtain that

$$\mathbb{E}|A'_{ni}| = \sum_{a \in \mathcal{A}} o(1) \left( o(1) + 2(\mathbf{h}^T \mathcal{I}_a(\boldsymbol{\theta}) \mathbf{h})^{1/2} \right) = o(1).$$

Because  $|A_{ni}| \leq A'_{ni}$ , we know that  $\mathbb{E}|A_{ni}| \rightarrow 0$  and  $\mathbb{E} \frac{1}{n} \sum_{i=1}^n |A_{ni}| \rightarrow 0$  as  $n \rightarrow \infty$ . By Lemma 13.2 and (29), we know that

$$\sum_{i=1}^n W_{ni}^2 = \frac{1}{n} \sum_{i=1}^n \left( \mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a_i}(X_i^{a_i}) \right)^2 + \frac{1}{n} \sum_{i=1}^n A_{ni} \xrightarrow{\mathbb{P}_g} \mathbf{h}^T \mathcal{I}^\pi(\boldsymbol{\theta}) \mathbf{h}.$$

By triangle inequality and Markov's inequality, as  $n \rightarrow \infty$ ,

$$\begin{aligned}
& \mathbb{P}(\max_{1 \leq i \leq n} |W_{ni}| > \varepsilon \sqrt{2}) \leq n\mathbb{P}(|W_{ni}| > \varepsilon \sqrt{2}) \\
& \leq n\mathbb{P}\left((\mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a_i}(X_i^{a_i}))^2 > n\varepsilon^2\right) + n\mathbb{P}\left(|A_{ni}| > n\varepsilon^2\right) \\
& \leq n\mathbb{P}\left(\sum_{a \in \mathcal{A}} (\mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(X_i^a))^2 > n\varepsilon^2\right) + n\mathbb{P}\left(|A'_{ni}| > n\varepsilon^2\right) \\
& \leq \frac{1}{\varepsilon^2} \mathbb{E} \sum_{a \in \mathcal{A}} (\mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(X_i^a))^2 I\left(\sum_{a \in \mathcal{A}} (\mathbf{h}^T \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(X_i^a))^2 > n\varepsilon^2\right) + \frac{\mathbb{E} A'_{ni}}{\varepsilon^2} \\
& \rightarrow 0.
\end{aligned}$$

Based on the rest of the proof of Theorem 7.2 in Van der Vaart (2000), we complete the proof of modified Theorem 7.2.

**Modified Theorem 7.10 in Van der Vaart (2000)** . The modified theorem is as follows: if statistics  $\mathbf{T}_n = \mathbf{T}_n(a_1, X_1^{a_1}, \dots, a_n, X_n^{a_n})$  satisfies the limit results in (29) under every  $\mathbf{h}$ , then there exists a randomized statistic  $\mathbf{T}$  in the experiment  $\{N_p(\mathbf{h}, \{\mathcal{I}^\pi(\boldsymbol{\theta})\}^{-1}) : \mathbf{h} \in \mathbb{R}^p\}$  such that  $\mathbf{T}_n \xrightarrow{h} \mathbf{T}$  for every  $\mathbf{h}$ .

The proof mostly follows that of Theorem 7.10 in Van der Vaart (2000) with the following modifications. Without loss of generality, let

$$P_{n, \mathbf{h}} = P_{n, \boldsymbol{\theta} + \mathbf{h}/\sqrt{n}}(a_1, X_1^{a_1}, \dots, a_n, X_n^{a_n}), \mathbf{J} = \mathcal{I}^\pi(\boldsymbol{\theta}), \Delta_n = \frac{1}{\sqrt{n}} \sum_{j=1}^n \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a_j}(X_j^{a_j}).$$

There exists random vector  $(\mathbf{S}, \Delta)$  such that

$$(\mathbf{T}_n, \Delta_n) \xrightarrow{0} (\mathbf{S}, \Delta).$$

Applying the modified Theorem 7.2 and follow similar arguments as those in the proof of Theorem 7.10 in Van der Vaart (2000), we obtain

$$\left(\mathbf{T}_n, \log \frac{dP_{n, \mathbf{h}}}{dP_{n, \mathbf{0}}}\right) \xrightarrow{0} \left(\mathbf{S}, \mathbf{h}^T \Delta - \frac{1}{2} \mathbf{h}^T \mathbf{J} \mathbf{h}\right).$$

The rest of the proof remains unchanged.

**Modified Theorem 8.3 in Van der Vaart (2000)** With a similar proof, the conclusion in Theorem 8.3 in Van der Vaart (2000) is modified as follows. If the limit results in (29) hold, then there exists a randomized statistic  $\mathbf{T}$  in  $\{N_p(\mathbf{h}, \{\mathcal{I}^\pi(\boldsymbol{\theta})\}^{-1}) : \mathbf{h} \in \mathbb{R}^p\}$  such that

$T - \mathbf{h}$  has the distribution  $L_{\boldsymbol{\theta}}^{\pi}$  for every  $\mathbf{h}$ .

**Proposition 8.4 in Van der Vaart (2000)** This proposition directly apply to our setting and does not required to be changed.

With the above modifications, we follow a similar proof as that for Theorem 8.8 in Van der Vaart (2000), we obtain (30) as well as the first part of the theorem.

**Proposition 8.6 in Van der Vaart (2000)** This proposition directly applies to our problem and does not need to be modified.

Following the proof of Theorem 8.11 Van der Vaart (2000) with the above modifications, we complete the proof of (31) and the second part of the theorem.  $\square$

## 14.10 Proof of Theorem 4.10

*Proof of Theorem 4.10.* By Theorem 4.1, we know that

$$\lim_{n \rightarrow \infty} \widehat{\boldsymbol{\theta}}_n^{\text{ML}} = \boldsymbol{\theta}^*, a.s.$$

Because  $\lim_{n \rightarrow \infty} \tau_n = \infty$  a.s., we obtain that

$$\lim_{n \rightarrow \infty} \widehat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}} = \boldsymbol{\theta}^*, a.s.$$

$\square$

## 14.11 Proof of Theorem 4.11

We prove the theorem for a class more general stopping rules instead. We first define a deterministic stopping rule

$$\tau(\Gamma_{\boldsymbol{\theta}^*}, c, \boldsymbol{\theta}, \boldsymbol{\pi}) = \min \left\{ m \geq n_0; \frac{1}{m} \Gamma_{\boldsymbol{\theta}^*}(\{\mathcal{I}^{\pi}(\boldsymbol{\theta})\}^{-1}) \leq c \right\},$$

where  $\Gamma_{\boldsymbol{\theta}}$  is a continuous function that maps a positive definite matrix to a positive number,  $c$  is a positive number,  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ , and  $\boldsymbol{\pi} \in \mathcal{S}^{\mathcal{A}}$ . Note that  $\tau(\Gamma, c, \boldsymbol{\theta}, \boldsymbol{\pi}) = \max\{\lceil \frac{\Gamma(\{\mathcal{I}^{\pi}(\boldsymbol{\theta})\}^{-1})}{c} \rceil, n_0\}$ , where  $\lceil \cdot \rceil$  is the ceiling function.

Consider a class of functions  $\{\Gamma_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}}$ , such that for any  $0 < u_1 < u_2 < \infty$ ,

$$\lim_{\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}^*} \max_{u_1 I \preceq \boldsymbol{\Sigma} \preceq u_2 I} |\Gamma_{\boldsymbol{\theta}}(\boldsymbol{\Sigma}) - \Gamma_{\boldsymbol{\theta}^*}(\boldsymbol{\Sigma})| = 0, \text{ and } \min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \min_{u_1 I \preceq \boldsymbol{\Sigma} \preceq u_2 I} \Gamma_{\boldsymbol{\theta}}(\boldsymbol{\Sigma}) > 0. \quad (185)$$

Define a random stopping time

$$\tau_c = \min \left\{ m \geq n_0; \frac{1}{m} \Gamma_{\hat{\boldsymbol{\theta}}_m} (\{\mathcal{I}^{\bar{\pi}_m}(\hat{\boldsymbol{\theta}}_m)\}^{-1}) \leq c \right\}, \quad (186)$$

where  $\hat{\boldsymbol{\theta}}_m$  is an estimator of  $\boldsymbol{\theta}$  based on  $m$  observations. Later, we will show that the stopping rules considered in Theorem 4.11 are special cases of the general stopping rule defined above.

The following theorem generalizes Theorem 4.11.

**Theorem 14.23** (General result for Asymptotic normality with stopping time). *Let  $\hat{\boldsymbol{\theta}}_n^{ML}$  be the MLE following the experiment selection rule GI0 or GI1, as described in Algorithm 1 and Algorithm 2. Assume that  $\mathbb{F}_{\boldsymbol{\theta}^*}(\boldsymbol{\pi})$  has a unique minimizer  $\boldsymbol{\pi}^*$ . Let  $\{c_n\}_{n \geq 0}$  be a positive decreasing sequence such that  $c_n \rightarrow 0$  as  $n \rightarrow \infty$ . Consider the stopping time  $\tau_{c_n}$  given by (186), where  $\Gamma_{\boldsymbol{\theta}}$  satisfies (185). Then,*

$$\sqrt{\tau_n} \left\{ \mathcal{I}^{\bar{\pi}_{\tau_{c_n}}}(\hat{\boldsymbol{\theta}}_{\tau_{c_n}}^{ML}) \right\}^{1/2} (\hat{\boldsymbol{\theta}}_{\tau_{c_n}}^{ML} - \boldsymbol{\theta}^*) \xrightarrow{d} N_p(\mathbf{0}_p, I_p). \quad (187)$$

Furthermore, for any continuously differentiable function  $g : \boldsymbol{\Theta} \rightarrow \mathbb{R}$  such that  $\nabla g(\boldsymbol{\theta}^*) \neq \mathbf{0}$ ,

$$\frac{\sqrt{\tau_{c_n}}(g(\hat{\boldsymbol{\theta}}_{\tau_{c_n}}^{ML}) - g(\boldsymbol{\theta}^*))}{\left\| \left\{ \mathcal{I}^{\bar{\pi}_{\tau_{c_n}}}(\hat{\boldsymbol{\theta}}_{\tau_{c_n}}^{ML}) \right\}^{-1/2} \nabla g(\hat{\boldsymbol{\theta}}_{\tau_{c_n}}^{ML}) \right\|} \xrightarrow{d} N(0, 1). \quad (188)$$

Given the above generalized theorem, the proof of 4.11 is provided below. The proof of Theorem 14.23 is provided later in this section.

*Proof of Theorem 4.11.* Note that  $\tau_c^{(1)}$  and  $\tau_c^{(2)}$  can be rewritten as

$$\begin{aligned} \tau_c^{(1)} &= \min \{ m \geq n_0; \frac{1}{m} \Gamma_{\boldsymbol{\theta}}^{(1)} (\{\mathcal{I}(\hat{\boldsymbol{\theta}}_{\tau_n}^{ML}; \mathbf{a}_m)\}^{-1}) \leq c^2 \} \\ \tau_c^{(2)} &= \min \{ m \geq n_0; \frac{1}{m} \Gamma_{\boldsymbol{\theta}}^{(1)} (\{\mathcal{I}(\hat{\boldsymbol{\theta}}_{\tau_n}^{ML}; \mathbf{a}_m)\}^{-1}) \leq c \}, \end{aligned}$$

where

$$\Gamma_{\boldsymbol{\theta}}^{(1)}(\boldsymbol{\Sigma}) = \text{tr} \left( \{\nabla h(\boldsymbol{\theta})\}^T \boldsymbol{\Sigma} h(\boldsymbol{\theta}) \right), \Gamma_{\boldsymbol{\theta}}^{(2)}(\boldsymbol{\Sigma}) = \text{tr}(\boldsymbol{\Sigma}),$$

Both  $\Gamma_{\boldsymbol{\theta}}^{(l)}(\boldsymbol{\Sigma})$  ( $l = 1, 2$ ) are continuously differentiable in  $\boldsymbol{\theta}$  and  $\boldsymbol{\Sigma}$  so the first part of (185) is satisfied. The second part of (185) is satisfied for  $\Gamma^{(2)}$  is straightforward. For  $\Gamma_{\boldsymbol{\theta}}^{(1)}(\boldsymbol{\Sigma})$ , the second part of (185) is satisfied due to the assumption that  $\nabla h(\boldsymbol{\theta}) \neq \mathbf{0}$  for all  $\boldsymbol{\theta}$ . Thus, conditions of Theorem 14.23 are satisfied, and the proof is completed by applying this theorem.  $\square$



In the rest of the section, we present the proof of Theorem 14.23. Roughly, Theorem 14.23 is proved by combining the following lemma, compares the random and deterministic stopping times, with the multivariate Anscombe's theorem (Lemma 13.3).

**Lemma 14.24.** *Assume a family of function  $\Gamma_{\boldsymbol{\theta}}$  satisfies (185). Assume there exists constants  $0 < u_1 < u_2 < \infty$  such that for any  $n \geq n_0$  and  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ ,*

$$u_1 I \preceq \mathcal{I}^{\bar{\pi}_n}(\boldsymbol{\theta}) \preceq u_2 I. \quad (189)$$

*If as  $n \rightarrow \infty$ ,*

$$\begin{aligned} c_n &\rightarrow 0, c_n \geq c_{n+1} > 0, \\ \hat{\boldsymbol{\theta}}_n &\rightarrow \boldsymbol{\theta}^* \text{ a.s. } \mathbb{P}_*, \text{ and} \\ \bar{\pi}_n &\rightarrow \pi \text{ a.s. } \mathbb{P}_*, \end{aligned} \quad (190)$$

*where  $\mathcal{I}^{\pi}(\boldsymbol{\theta}^*)$  is nonsingular. Then,  $\tau_{c_n} < \infty$  a.s.  $\mathbb{P}_*$  and as  $n \rightarrow \infty$ ,*

$$\tau(\Gamma_{\boldsymbol{\theta}^*, c_n, \boldsymbol{\theta}^*, \pi}) \rightarrow \infty, \quad \tau_{c_n} \rightarrow \infty, \quad \text{and} \quad \frac{\tau_{c_n}}{\tau(\Gamma_{\boldsymbol{\theta}^*, c_n, \boldsymbol{\theta}^*, \pi})} \rightarrow 1, \text{ a.s. } \mathbb{P}_*. \quad (191)$$

*Proof of Lemma 14.24.* By Theorem (14.1) and equation (78), we know that there exists  $0 < \underline{c} < \bar{c} < \infty$  such that

$$\underline{c} I_p \preceq \mathcal{I}^{\bar{\pi}_m}(\hat{\boldsymbol{\theta}}_m) \preceq \bar{c} I_p,$$

for all  $m \geq n_0$ . By assumption (185), there exists  $0 < v_1 < v_2 < \infty$  such that for any  $n \geq n_0$ ,

$$v_1 \leq \Gamma_{\hat{\boldsymbol{\theta}}_m}(\{\mathcal{I}^{\bar{\pi}_m}(\hat{\boldsymbol{\theta}}_m)\}^{-1}) \leq v_2.$$

Note that  $\tau(\Gamma_{\boldsymbol{\theta}^*, c_n, \boldsymbol{\theta}^*, \pi}) = \max\{\lceil \frac{\Gamma_{\boldsymbol{\theta}^*}(\{\mathcal{I}^{\pi}(\boldsymbol{\theta}^*)\}^{-1})}{c_n} \rceil, n_0\} \rightarrow \infty$ , as  $n \rightarrow \infty$ . Also note that

$$\left\{m \geq n_0; \frac{v_2}{m} \leq c_n\right\} \subset \left\{m \geq n_0; \frac{1}{m} \Gamma_{\hat{\boldsymbol{\theta}}_m}(\{\mathcal{I}^{\bar{\pi}_m}(\hat{\boldsymbol{\theta}}_m)\}^{-1}) \leq c_n\right\} \subset \left\{m \geq n_0; \frac{v_1}{m} \leq c_n\right\}.$$

Thus, for any fixed  $n$ ,

$$\tau_{c_n} \leq \min \left\{m \geq n_0; \frac{v_2}{m} \leq c_n\right\} \leq \left\lceil \frac{v_2}{c_n} \right\rceil + n_0 < \infty, \text{ a.s. } \mathbb{P}_*,$$

and as  $n \rightarrow \infty$ ,

$$\tau_{c_n} \geq \min \left\{m \geq n_0; \frac{v_1}{m} \leq c_n\right\} \geq \left\lceil \frac{v_1}{c_n} \right\rceil \rightarrow \infty.$$

By assumption (190), we know that

$$\lim_{n \rightarrow \infty} \{\mathcal{I}^{\bar{\pi}_n}(\hat{\boldsymbol{\theta}}_n)\}^{-1} = \{\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*)\}^{-1} \text{ a.s. } \mathbb{P}_*.$$

Combining the compact convergence assumption (185) and (189), we obtain that

$$\lim_{n \rightarrow \infty} \Gamma_{\hat{\boldsymbol{\theta}}_n}(\{\mathcal{I}^{\bar{\pi}_n}(\hat{\boldsymbol{\theta}}_n)\}^{-1}) = \Gamma_{\boldsymbol{\theta}^*}(\{\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*)\}^{-1}) \text{ a.s. } \mathbb{P}_*.$$

That is, with probability 1, for any  $\eta > 0$ , there exists  $N_\eta \geq n_0$  such that for any  $n \geq N_\eta$ ,

$$|\Gamma_{\hat{\boldsymbol{\theta}}_n}(\{\mathcal{I}^{\bar{\pi}_n}(\hat{\boldsymbol{\theta}}_n)\}^{-1}) - \Gamma_{\boldsymbol{\theta}^*}(\{\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*)\}^{-1})| < \eta.$$

Set  $N_2 = \min \left\{ n \geq N_\eta; \lceil \frac{v_1}{c_n} \rceil \geq N_\eta \right\} < \infty$ . For any  $n \geq N_2$ , we obtain that,

$$\begin{aligned} & \left\{ m \geq n_0; \frac{1}{m} \{ \Gamma_{\boldsymbol{\theta}^*}(\{\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*)\}^{-1}) + \eta \} \leq c_n \right\} \\ & \subset \left\{ m \geq n_0; \frac{1}{m} \Gamma_{\hat{\boldsymbol{\theta}}_m}(\{\mathcal{I}^{\bar{\pi}_m}(\hat{\boldsymbol{\theta}}_m)\}^{-1}) \leq c_n \right\} \\ & \subset \left\{ m \geq n_0; \frac{1}{m} \{ \Gamma_{\boldsymbol{\theta}^*}(\{\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*)\}^{-1}) - \eta \} \leq c_n \right\}, \end{aligned}$$

which implies that

$$\left\lceil \frac{\Gamma_{\boldsymbol{\theta}^*}(\{\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*)\}^{-1}) - \eta}{c_n} \right\rceil \leq \tau_{c_n} \leq \left\lceil \frac{\Gamma_{\boldsymbol{\theta}^*}(\{\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*)\}^{-1}) + \eta}{c_n} \right\rceil.$$

Set  $\eta = \xi \cdot \Gamma_{\boldsymbol{\theta}^*}(\{\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*)\}^{-1})$  and  $d_n = \Gamma_{\boldsymbol{\theta}^*}(\{\mathcal{I}^{\boldsymbol{\pi}}(\boldsymbol{\theta}^*)\}^{-1})/c_n \rightarrow \infty$ . Note that

$$\frac{\lceil (1 - \xi)d_n \rceil}{\lceil d_n \rceil} \leq \frac{\tau_{c_n}}{\tau(\Gamma_{\boldsymbol{\theta}^*}, c_n, \boldsymbol{\theta}^*, \boldsymbol{\pi})} \leq \frac{\lceil (1 + \xi)d_n \rceil}{\lceil d_n \rceil}. \quad (192)$$

Taking the infimum limit and supremum limit over both sides of inequalities (192), we obtain that for any  $\xi > 0$ ,

$$(1 - \xi) \leq \liminf_{n \rightarrow \infty} \frac{\tau_{c_n}}{\tau(\Gamma_{\boldsymbol{\theta}^*}, c_n, \boldsymbol{\theta}^*, \boldsymbol{\pi})} \leq \limsup_{n \rightarrow \infty} \frac{\tau_{c_n}}{\tau(\Gamma_{\boldsymbol{\theta}^*}, c_n, \boldsymbol{\theta}^*, \boldsymbol{\pi})} \leq (1 + \xi).$$

In conclusion,

$$\lim_{n \rightarrow \infty} \frac{\tau_{c_n}}{\tau(\Gamma_{\boldsymbol{\theta}^*}, c_n, \boldsymbol{\theta}^*, \boldsymbol{\pi})} = 1, \text{ a.s. } \mathbb{P}_*.$$

□

*Proof of Theorem 14.23.* Let  $v_n = \tau(\Gamma_{\boldsymbol{\theta}^*}, c_n, \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)$ . According to Theorem 4.1 and Theorem

4.3, the conditions in (190) are satisfied. By Lemma 14.24, we obtain (191), which implies that

$$\sqrt{\tau_n} \{\mathcal{I}^{\pi_n}(\hat{\boldsymbol{\theta}}_n^{\text{ML}})\}^{1/2} \frac{1}{\sqrt{v_n}} \{\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*)\}^{-1/2} \rightarrow I_p, \text{ a.s. } \mathbb{P}_*. \quad (193)$$

For the ease of exposition, we write  $\tau_n = \tau_{c_n}$ . To show the limit result (187), by (193) and Slutsky's theorem, it suffices to show that

$$\sqrt{v_n} \{\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*)\}^{1/2} (\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}} - \boldsymbol{\theta}^*) \xrightarrow{d} N_p(\mathbf{0}_p, I_p). \quad (194)$$

According to Theorem 13.3 with  $T_n = \hat{\boldsymbol{\theta}}_n^{\text{ML}}$ ,  $\theta = \boldsymbol{\theta}^*$ ,  $N_n = \tau_n$ ,  $r_n = v_n$ , and  $W_n = n^{-1/2} \{\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*)\}^{-1/2}$ , (194) we only need to verify the following conditions for Theorem 13.3: for all  $\gamma > 0$ ,  $\varepsilon > 0$ , there exists  $0 < \delta < 1$  such that

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left( \max_{|n' - n| \leq \delta n} \left\| \hat{\boldsymbol{\theta}}_{n'}^{\text{ML}} - \hat{\boldsymbol{\theta}}_n^{\text{ML}} \right\| \geq \frac{\varepsilon}{\sqrt{n}} \lambda_{\min}(\{\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*)\}^{-1/2}) \right) \leq \gamma.$$

To show the above inequality, it suffices to show that for any  $\gamma > 0$ ,  $\varepsilon > 0$ , there exists  $0 < \delta < 1$  such that

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left( \max_{n', n'', n''' \in [n, (1+\delta)n]} \sqrt{n'''} \left\| \hat{\boldsymbol{\theta}}_{n'}^{\text{ML}} - \hat{\boldsymbol{\theta}}_{n'''}^{\text{ML}} \right\| \geq \varepsilon \lambda_{\min}(\{\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*)\}^{-1/2}) \right) \leq \gamma.$$

Note that

$$\begin{aligned} & \mathbb{P} \left( \max_{n', n'', n''' \in [n, (1+\delta)n]} \sqrt{n'''} \left\| \hat{\boldsymbol{\theta}}_{n'}^{\text{ML}} - \hat{\boldsymbol{\theta}}_{n'''}^{\text{ML}} \right\| \geq \varepsilon \lambda_{\min}(\{\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*)\}^{-1/2}) \right) \\ & \leq \mathbb{P} \left( \max_{n', n'' \in [n, (1+\delta)n]} \sqrt{n} \left( \left\| \hat{\boldsymbol{\theta}}_{n'}^{\text{ML}} - \hat{\boldsymbol{\theta}}_n^{\text{ML}} \right\| + \left\| \hat{\boldsymbol{\theta}}_n^{\text{ML}} - \hat{\boldsymbol{\theta}}_{n''}^{\text{ML}} \right\| \right) \geq \frac{\varepsilon}{\sqrt{1+\delta}} \lambda_{\min}(\{\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*)\}^{-1/2}) \right) \\ & \leq 2 \cdot \mathbb{P} \left( \max_{n' \in [n, (1+\delta)n]} \sqrt{n} \left\| \hat{\boldsymbol{\theta}}_{n'}^{\text{ML}} - \hat{\boldsymbol{\theta}}_n^{\text{ML}} \right\| \geq \frac{\varepsilon}{2\sqrt{2}} \lambda_{\min}(\{\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*)\}^{-1/2}) \right). \end{aligned}$$

Thus, we only need to show that for all  $\varepsilon > 0$ ,

$$\lim_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} \mathbb{P} \left( \max_{n' \in [n, (1+\delta)n]} \sqrt{n} \left\| \hat{\boldsymbol{\theta}}_{n'}^{\text{ML}} - \hat{\boldsymbol{\theta}}_n^{\text{ML}} \right\| \geq \varepsilon \lambda_{\min}(\{\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*)\}^{-1/2}) \right) = 0. \quad (195)$$

Let

$$\begin{aligned}
D_n = & \left\{ \nabla_{\boldsymbol{\theta}} l_n(\widehat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n) = 0 \right\} \\
& \cap \left\{ \frac{1}{n} \sum_{j=1}^n \Psi_2^{a_j}(X_j) \left\| \{\nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*; \mathbf{a}_n)\}^{-1} \right\|_{op} \sup_{n' \geq n} \psi \left( \left\| \widehat{\boldsymbol{\theta}}_n^{\text{ML}} - \widehat{\boldsymbol{\theta}}_{n'}^{\text{ML}} \right\| \right) \leq \frac{1}{2} \right\} \\
& \cap \left\{ \frac{1}{n} \sum_{j=1}^n \Psi_2^{a_j}(X_j) \left\| \{\nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*; \mathbf{a}_n)\}^{-1} \right\|_{op} \psi \left( \left\| \widehat{\boldsymbol{\theta}}_n^{\text{ML}} - \boldsymbol{\theta}^* \right\| \right) \leq \frac{1}{2} \right\} \\
& \cap \left\{ \left\| \{\nabla_{\boldsymbol{\theta}}^2 l_n(\widehat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1} \right\|_{op} \leq \frac{2}{\lambda_{\min}(\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*))} \right\} \\
& \cap \left\{ \left\| \{\nabla_{\boldsymbol{\theta}}^2 l_n(\boldsymbol{\theta}^*; \mathbf{a}_n)\}^{-1} \right\|_{op} \leq \frac{2}{\lambda_{\min}(\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*))} \right\}.
\end{aligned}$$

By Theorem 4.1, Theorem 14.1, Corollary 14.12, Lemma 14.13, and with probability 1,

$$\limsup_{n \rightarrow \infty} \left\| \{\nabla_{\boldsymbol{\theta}}^2 l_n(\widehat{\boldsymbol{\theta}}_n^{\text{ML}}; \mathbf{a}_n)\}^{-1} \right\|_{op} \leq \frac{1}{\lambda_{\min}(\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*))},$$

we know that

$$\mathbb{P} \left( \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} D_n \right) = 1.$$

By Lemma 14.14, we obtain that

$$\bigcap_{m \geq n} D_m \subset \left\{ \sup_{n': n \leq n' \leq (1+\delta)n} \left\| \widehat{\boldsymbol{\theta}}_{n'}^{\text{ML}} - \widehat{\boldsymbol{\theta}}_n^{\text{ML}} \right\| \leq \frac{4}{\lambda_{\min}(\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*))} \sup_{n': n \leq n' \leq (1+\delta)n} \left\| \nabla_{\boldsymbol{\theta}} l_n(\widehat{\boldsymbol{\theta}}_{n'}^{\text{ML}}; \mathbf{a}_n) \right\| \right\},$$

and

$$D_n \subset \left\{ \sqrt{n} \left\| \widehat{\boldsymbol{\theta}}_n^{\text{ML}} - \boldsymbol{\theta}^* \right\| \leq \frac{4}{\lambda_{\min}(\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*))} \sqrt{n} \left\| \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^*; \mathbf{a}_n) \right\| \right\}.$$

Thus

$$\begin{aligned}
& \mathbb{P} \left( \sup_{n': n \leq n' \leq (1+\delta)n} \left\| \widehat{\boldsymbol{\theta}}_{n'}^{\text{ML}} - \widehat{\boldsymbol{\theta}}_n^{\text{ML}} \right\| \geq \frac{\varepsilon}{\sqrt{n}} \lambda_{\min}(\{\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*)\}^{-1/2}) \right) \\
& \leq \mathbb{P} \left( \left\{ \bigcap_{m \geq n} D_m \right\}^c \right) \\
& \quad + \mathbb{P} \left( \left\{ \sup_{n': n \leq n' \leq (1+\delta)n} \sqrt{n} \left\| \nabla_{\boldsymbol{\theta}} l_n(\widehat{\boldsymbol{\theta}}_{n'}^{\text{ML}}; \mathbf{a}_n) \right\| \geq C(\varepsilon) \right\} \cap \bigcap_{m \geq n} D_m \right),
\end{aligned} \tag{196}$$

where  $C(\varepsilon) = \frac{\varepsilon \lambda_{\min}(\{\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*)\}^{-1/2})}{4 \{\lambda_{\min}(\mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*))\}^{-1}}$ . Let  $n' \in [n+1, (1+\delta)n]$ . With  $\nabla_{\boldsymbol{\theta}} l_{n'}(\widehat{\boldsymbol{\theta}}_{n'}^{\text{ML}}; \mathbf{a}_{n'}) = 0$ , we

know that that over  $\cap_{m \geq n} D_m$

$$\begin{aligned}
\sqrt{n} \left\| \nabla_{\boldsymbol{\theta}} l_n(\hat{\boldsymbol{\theta}}_{n'}^{\text{ML}}; \mathbf{a}_n) \right\| &= \frac{1}{\sqrt{n}} \left\| \nabla_{\boldsymbol{\theta}} \sum_{j=n+1}^{n'} \log f_{\hat{\boldsymbol{\theta}}_{n'}^{\text{ML}}, a_j}(X_j) \right\| \\
&\leq \frac{1}{\sqrt{n}} \left\| \sum_{j=n+1}^{n'} \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*, a_j}(X_j) \right\| + \delta \frac{\sum_{j=n+1}^{n'} \Psi_1^{a_j}(X_j)}{\delta n} \sqrt{n} \left\| \hat{\boldsymbol{\theta}}_{n'}^{\text{ML}} - \boldsymbol{\theta}^* \right\| \\
&\leq \frac{1}{\sqrt{n}} \left\| \sum_{j=n+1}^{n'} \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*, a_j}(X_j) \right\| \\
&\quad + \frac{4\delta}{\lambda_{\min}(\mathcal{I}\pi^*(\boldsymbol{\theta}^*))} \frac{\sum_{j=n+1}^{(1+\delta)n} \Psi_1^{a_j}(X_j)}{\delta n} \frac{1}{\sqrt{n}} \left\| \sum_{j=1}^{n'} \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*, a_j}(X_j) \right\|.
\end{aligned} \tag{197}$$

By Markov inequality, we obtain that for any  $M > 0$

$$\mathbb{P}\left(\frac{\sum_{n+1 \leq j \leq (1+\delta)n} \Psi_1^{a_j}(X_j)}{\delta n} \geq M\right) \leq \frac{\mu_Y}{M},$$

where  $\mu_Y = \sum_{a \in \mathcal{A}} \mathbb{E}_{X^a \sim f_{\boldsymbol{\theta}^*, a}} \Psi_1^a(X^a) < \infty$ . Thus,

$$\frac{\sup_{n': n+1 \leq n' \leq (1+\delta)n} \sum_{j=n+1}^{n'} \Psi_1^{a_j}(X_j)}{\delta n} = \frac{\sum_{j=n+1}^{(1+\delta)n} \Psi_1^{a_j}(X_j)}{\delta n} = O_p(1).$$

Note that  $S_m^n = \sum_{j=n+1}^{n+m} \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*, a_j}(X_j)$  is martingale sequence with respect to filtration  $\mathcal{F}_m^n = \mathcal{F}_{n+m}$ .

By Assumption 2,  $C_1 := \max_{a \in \mathcal{A}} \max_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{X \sim f_{\boldsymbol{\theta}^*, a}} \{\|\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}, a}(X)\|^2\} < \infty$ . Since  $\|\cdot\|$  is convex, by Jensen's inequality,  $\|S_m^n\|$  is a submartingale. Applying Doob's inequality (see Theorem 6.5.d. in Lin (2010)), we obtain that

$$\mathbb{P}\left(\max_{1 \leq m \leq l} \|S_m^n\| \geq M\right) \leq \frac{\mathbb{E} \|S_l^n\|^2}{M^2} \leq \frac{l \cdot C_1}{M^2}.$$

Hence, we obtain

$$\mathbb{P}\left(\max_{n+1 \leq n' \leq (1+\delta)n} \frac{1}{\sqrt{n\delta}} \left\| \sum_{j=n+1}^{n'} \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*, a_j}(X_j) \right\| \geq M\right) \leq \frac{C_1}{M^2}, \tag{198}$$

and

$$\mathbb{P}\left(\max_{n+1 \leq n' \leq (1+\delta)n} \frac{1}{\sqrt{n}} \left\| \sum_{j=1}^{n'} \nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*, a_j}(X_j) \right\| \geq M\right) \leq \frac{(1+\delta)n \cdot C_1}{n \cdot M^2} \leq \frac{2C_1}{M^2}. \tag{199}$$

Combining (197), (198) and (199), we obtain

$$\max_{n': n+1 \leq n' \leq (1+\delta)n} \sqrt{n} \left\| \nabla_{\boldsymbol{\theta}} l_n(\hat{\boldsymbol{\theta}}_{n'}^{\text{ML}}; \mathbf{a}_n) \right\| = \sqrt{\delta} O_p(1) + \delta O_P(1).$$

Thus,

$$\begin{aligned} & \limsup_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} \mathbb{P} \left( \left\{ \sup_{n': n \leq n' \leq (1+\delta)n} \sqrt{n} \left\| \nabla_{\boldsymbol{\theta}} l_n(\hat{\boldsymbol{\theta}}_{n'}^{\text{ML}}; \mathbf{a}_n) \right\| \geq C(\varepsilon) \right\} \bigcap_{m \geq n} D_m \right) \\ & \leq \limsup_{\delta \rightarrow 0} \mathbb{P} \left( \sqrt{\delta} O_p(1) \geq C(\varepsilon) \right) = 0. \end{aligned}$$

This completes the proof of (187).

We proceed to the proof of the ‘Furthermore’ part of the theorem. Note that

$$g(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) - g(\boldsymbol{\theta}^*) = \{\nabla g(\tilde{\boldsymbol{\theta}}_n)\}^T (\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}} - \boldsymbol{\theta}^*),$$

where  $\tilde{\boldsymbol{\theta}}_n \rightarrow \boldsymbol{\theta}^*$  and  $\nabla g(\tilde{\boldsymbol{\theta}}_n) \rightarrow \nabla g(\boldsymbol{\theta}^*)$  a.s.  $\mathbb{P}_*$  as  $n \rightarrow \infty$ . Then,

$$\begin{aligned} & \frac{\sqrt{\tau_n} (g(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) - g(\boldsymbol{\theta}^*))}{\left\| \left\{ \mathcal{I}^{\bar{\pi}_{\tau_n}}(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\}^{-1/2} \nabla g(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\|} = \frac{\sqrt{\tau_n} \{\nabla g(\tilde{\boldsymbol{\theta}}_n)\}^T (\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}} - \boldsymbol{\theta}^*)}{\left\| \left\{ \mathcal{I}^{\bar{\pi}_{\tau_n}}(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\}^{-1/2} \nabla g(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\|} \\ & = \frac{\left[ \left\{ \mathcal{I}^{\bar{\pi}_{\tau_n}}(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\}^{-1/2} \nabla g(\tilde{\boldsymbol{\theta}}_n) \right]^T}{\left\| \left\{ \mathcal{I}^{\bar{\pi}_{\tau_n}}(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\}^{-1/2} \nabla g(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\|} \sqrt{\tau_n} \left\{ \mathcal{I}^{\bar{\pi}_{\tau_n}}(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\}^{1/2} (\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}} - \boldsymbol{\theta}^*). \end{aligned}$$

By Theorem 14.1 and Assumption 6B, we know that there exists  $U > 0$  such that for any  $n \geq n_0$

$$\underline{c} U I_p \preceq \mathcal{I}^{\bar{\pi}_n}(\boldsymbol{\theta}) \preceq \bar{c} I_p, \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

Thus, the condition number of  $\mathcal{I}^{\bar{\pi}_n}(\boldsymbol{\theta})$ ,  $\kappa(\mathcal{I}^{\bar{\pi}_n}(\boldsymbol{\theta})) \leq \frac{\bar{c}}{\underline{c}U} < \infty$ , for any  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$  and  $n \geq n_0$ . Moreover, we know that

$$\begin{aligned} & \left\| \frac{\left\{ \mathcal{I}^{\bar{\pi}_{\tau_n}}(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\}^{-1/2} (\nabla g(\tilde{\boldsymbol{\theta}}_n) - \nabla g(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}))}{\left\| \left\{ \mathcal{I}^{\bar{\pi}_{\tau_n}}(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\}^{-1/2} \nabla g(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\|} \right\| \\ & \leq \kappa \left( \left\{ \mathcal{I}^{\bar{\pi}_{\tau_n}}(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\}^{-1/2} \right) \frac{\left\| \nabla g(\tilde{\boldsymbol{\theta}}_n) - \nabla g(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\|}{\left\| \nabla g(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\|} = o_p(1). \end{aligned}$$

Let  $h_n = \frac{\left\{ \mathcal{I}^{\bar{\pi}_{\tau_n}}(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\}^{-1/2} \nabla g(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}})}{\left\| \left\{ \mathcal{I}^{\bar{\pi}_{\tau_n}}(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\}^{-1/2} \nabla g(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\|}$ , then  $\|h_n\| = 1$ . By continuous mapping theorem,

$$h_n \rightarrow h := \frac{\left\{ \mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*) \right\}^{-1/2} \nabla g(\boldsymbol{\theta}^*)}{\left\| \left\{ \mathcal{I}^{\pi^*}(\boldsymbol{\theta}^*) \right\}^{-1/2} \nabla g(\boldsymbol{\theta}^*) \right\|} \text{ a.s. } \mathbb{P}_*.$$

As  $n \rightarrow \infty$ ,

$$\frac{\sqrt{\tau_n}(g(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) - g(\boldsymbol{\theta}^*))}{\left\| \left\{ \mathcal{I}^{\bar{\pi}_{\tau_n}}(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\}^{-1/2} \nabla g(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\|} = h^T \sqrt{\tau_n} \left\{ \mathcal{I}^{\bar{\pi}_{\tau_n}}(\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}}) \right\}^{1/2} (\hat{\boldsymbol{\theta}}_{\tau_n}^{\text{ML}} - \boldsymbol{\theta}^*) + o_p(1) \xrightarrow{d} N(0, 1).$$

This completes the proof of (188). □

## 14.12 Proof of Corollaries 5.1, 5.2, and 5.3

In this section, we will verify the regularity conditions for the applications presented in Sections 5.1, 5.2, and 5.3, thereby proving Corollaries 5.1, 5.2, and 5.3. First, according to Lemma 13.13, Assumptions 6A and 7A imply Assumptions 6B and 7B. The next lemma is useful for verifying Assumption 4.

**Lemma 14.25.** *Let  $\mathcal{F}^a = \{\log f_{\boldsymbol{\theta},a}(\cdot) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ ,  $a \in \mathcal{A}$  be collections of measurable functions with a  $\mathbb{P}_*$  integrable envelope functions. That is,  $F_a$  satisfies that for all  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ ,*

$$|\log f_{\boldsymbol{\theta},a}(x^a)| \leq F_a(x^a), \text{ a.s. } \mathbb{P}_* \text{ and } \mathbb{E}_{X^a \sim f_{\boldsymbol{\theta}^*,a}} F_a(X^a) < \infty.$$

*If  $\boldsymbol{\Theta}$  is compact and mapping  $\boldsymbol{\theta} \mapsto \log f_{\boldsymbol{\theta},a}(x^a)$  is continuous for every  $x^a$  and  $a \in \mathcal{A}$ , then*

$$\mathbb{P}_* \left\{ \lim_{n \rightarrow \infty} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |l_n(\boldsymbol{\theta}; \mathbf{a}_n) - M(\boldsymbol{\theta}; \bar{\pi}_n)| = 0 \right\} = 1.$$

*Proof of Lemma 14.25.* The proof is similar to the proof of Theorem 2.4.1 in Vaart and Wellner (1997). First, we show that the bracketing numbers  $N_{[\cdot]}(\varepsilon, \mathcal{F}^a, L_1(\mathbb{P}_*)) < \infty$ , for every  $a \in \mathcal{A}$  and  $\varepsilon > 0$ , where the definition of bracketing number  $N_{[\cdot]}(\varepsilon, \mathcal{F}, \|\cdot\|)$  is given by Definition 2.1.6 in Vaart and Wellner (1997).

Let  $B(\boldsymbol{\theta}, \delta) = \{\boldsymbol{\theta}' \in \boldsymbol{\Theta}; \|\boldsymbol{\theta}' - \boldsymbol{\theta}\| < \delta\}$ . Define

$$u_{B(\boldsymbol{\theta}', \delta)}^a(x^a) = \sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < \delta} \log f_{\boldsymbol{\theta}, a}(x^a), \text{ and, } l_{B(\boldsymbol{\theta}', \delta)}^a(x^a) = \inf_{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < \delta} \log f_{\boldsymbol{\theta}, a}(x^a).$$

Because the envelope function  $F_a$  is integrable with respect to  $\mathbb{P}_*$  and  $\log f_{\boldsymbol{\theta}, a}(\cdot)$  is continuous in  $\boldsymbol{\theta}$ , the Dominated Convergence Theorem ensures that for any  $\boldsymbol{\theta}'$  and  $\varepsilon > 0$ , there exists  $\delta > 0$  such that

$$\mathbb{E}_{\boldsymbol{\theta}^*} (u_{B(\boldsymbol{\theta}', \delta)}^a(X^a) - l_{B(\boldsymbol{\theta}', \delta)}^a(X^a)) < \varepsilon.$$

By compactness of  $\boldsymbol{\Theta}$ , there exists  $(\boldsymbol{\theta}_1, \delta_1), (\boldsymbol{\theta}_2, \delta_2), \dots, (\boldsymbol{\theta}_m, \delta_m)$ , such that for any  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ , there exists  $1 \leq j \leq m$ ,

$$l_j(x^a) \leq \log f_{\boldsymbol{\theta}, a}(x^a) \leq u_j(x^a),$$

where  $u_j = u_{B(\boldsymbol{\theta}_j, \delta_j)}^a$  and  $l_j = l_{B(\boldsymbol{\theta}_j, \delta_j)}^a$ . Thus, the bracketing numbers  $N_{[\cdot]}(\varepsilon, \mathcal{F}^a, L_1(\mathbb{P}_*)) < \infty$ , for all  $\varepsilon > 0$  and  $a \in \mathcal{A}$ . That is, we can choose finitely many  $\varepsilon$ -brackets  $[l_j^a, u_j^a]$ , whose union contains  $\mathcal{F}^a$  and such that  $\mathbb{E}_{\boldsymbol{\theta}^*}(u_j^a(X^a) - l_j^a(X^a)) < \varepsilon$ , for every  $j$ . Hence, for every  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ ,  $a \in \mathcal{A}$ , there exists  $j_a$  such that

$$l_{j_a}(x^a) \leq \log f_{\boldsymbol{\theta}, a}(x^a) \leq u_{j_a}(x^a).$$

The above inequality also implies that

$$\mathbb{E}_{X \sim f_{\boldsymbol{\theta}, a}} l_{j_a}(X) \leq \mathbb{E}_{X \sim f_{\boldsymbol{\theta}, a}} \log f_{\boldsymbol{\theta}, a}(x^a) \leq \mathbb{E}_{X \sim f_{\boldsymbol{\theta}, a}} u_{j_a}(X) \quad (200)$$

Note that, if the functions  $f_{\boldsymbol{\theta}, a}(\cdot)$  are inside the brackets  $[l^a, u^a]$  for all  $a$ , then

$$\begin{aligned} & l_n(\boldsymbol{\theta}; \mathbf{a}_n) - M(\boldsymbol{\theta}; \bar{\boldsymbol{\pi}}_n) \\ & \leq \frac{1}{n} \sum_{i=1}^n (u^{a_i}(X_i) - \mathbb{E}[l^{a_i}(X_i) | \mathcal{F}_{i-1}]) \\ & \leq \frac{1}{n} \sum_{i=1}^n (u^{a_i}(X_i) - \mathbb{E}[u^{a_i}(X_i) | \mathcal{F}_{i-1}]) + \varepsilon. \end{aligned}$$

Thus,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} (l_n(\boldsymbol{\theta}; \mathbf{a}_n) - M(\boldsymbol{\theta}; \bar{\boldsymbol{\pi}}_n)) \leq \max_{\substack{a \in \mathcal{A} \\ u^a \in \{u_j^a; 1 \leq j \leq m\}}} \frac{1}{n} \sum_{i=1}^n (u^{a_i}(X_i) - \mathbb{E}[u^{a_i}(X_i) | \mathcal{F}_{i-1}]) + \varepsilon.$$



By Lemma 13.2,

$$\frac{1}{n} \sum_{i=1}^n (u^{a_i}(X_i) - \mathbb{E}[u^{a_i}(X_i) | \mathcal{F}_{i-1}]) \xrightarrow{\text{a.s.}} 0.$$

Consequently,

$$\limsup_{n \rightarrow \infty} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} (l_n(\boldsymbol{\theta}; \mathbf{a}_n) - M(\boldsymbol{\theta}; \bar{\boldsymbol{\pi}}_n)) \leq \varepsilon, \text{ a.s. } \mathbb{P}_*.$$

A similar argument yields that

$$\liminf_{n \rightarrow \infty} \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} (l_n(\boldsymbol{\theta}; \mathbf{a}_n) - M(\boldsymbol{\theta}; \bar{\boldsymbol{\pi}}_n)) \geq -\varepsilon, \text{ a.s. } \mathbb{P}_*.$$

Taking  $\varepsilon \rightarrow 0$ , we obtain that

$$\mathbb{P} \left\{ \lim_{n \rightarrow \infty} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |l_n(\boldsymbol{\theta}; \mathbf{a}_n) - M(\boldsymbol{\theta}; \bar{\boldsymbol{\pi}}_n)| = 0 \right\} = 1.$$

□

*Remark 14.26.* Under Assumptions 1 and 2, if we assume  $\nabla_{\boldsymbol{\theta}}^2 f_{\boldsymbol{\theta},a}(x^a)$  is continuous in  $(\boldsymbol{\theta}, x^a)$  and

$$L := \max_{a \in \mathcal{A}} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, x^a \in \text{supp}(f_{\boldsymbol{\theta},a})} \|\nabla_{\boldsymbol{\theta}}^2 \log f_{\boldsymbol{\theta},a}(x^a)\|_{op} < \infty,$$

then by the first order Taylor expansion with Lagrange remainder, we can choose the envelop function  $F_a$  as

$$F_a(x^a) = |\log f_{\boldsymbol{\theta}^*,a}(x^a)| + \|\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}^*,a}(x^a)\| \cdot \text{diameter}(\boldsymbol{\Theta}) + \frac{L}{2} \cdot \text{diameter}(\boldsymbol{\Theta})^2.$$

*Proof of Corollary 5.1.* Let  $\boldsymbol{\xi}_a = \mathbf{z}_a^T \boldsymbol{\theta}$  and  $h_{\boldsymbol{\xi}_a,a}(x^a) = \zeta_a(x^a) \exp\{x^a \boldsymbol{\xi}_a - B_a(\boldsymbol{\xi}_a)\}$ . Let  $X^a \sim f_{\boldsymbol{\theta}^*,a}$ . Note that

$$\nabla_{\boldsymbol{\xi}_a} \log h_{\boldsymbol{\xi}_a,a}(X^a) = X^a - B'_a(\boldsymbol{\xi}_a) \text{ and } -\nabla_{\boldsymbol{\xi}_a}^2 \log h_{\boldsymbol{\xi}_a,a}(X^a) = B''_a(\boldsymbol{\xi}_a) \geq \min_{\boldsymbol{\xi}_a = \mathbf{z}_a^T \boldsymbol{\theta}, \boldsymbol{\theta} \in \boldsymbol{\Theta}} B''_a(\boldsymbol{\xi}_a) > 0.$$

Assumption 1,  $\boldsymbol{\theta}^*$  is in the interior of  $\boldsymbol{\Theta}$ ,  $\boldsymbol{\xi}_a^* = \mathbf{z}_a^T \boldsymbol{\theta}^*$  is in the interior of  $\{\mathbf{z}_a^T \boldsymbol{\theta}; \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ . Applying Theorem 5.8 in Lehmann and Casella (2006), we know that all moments for  $\nabla_{\boldsymbol{\xi}_a} \log h_{\boldsymbol{\xi}_a^*,a}(X^a) = X^a - B'_a(\boldsymbol{\xi}_a^*)$  exist. Also note that  $B''_{a_i}(\mathbf{z}_{a_i}^T \boldsymbol{\theta}) > 0$  and  $\mathcal{I}_{\boldsymbol{\xi}_a,a}(\boldsymbol{\xi}_a) = B''(\boldsymbol{\xi}_a)$  is nonsingular. Thus, Assumptions 2 and 6A hold. Note that from the above derivations,  $\|\nabla_{\boldsymbol{\xi}_a}^2 \log f_{\boldsymbol{\theta},a}(x^a)\|_{op}$  does not depend on  $x^a$ . This, together with Lemma 14.25 and the accompanying Remark 14.26, implies that Assumption 4 is satisfied.

Note that

$$D_{KL}(h_{\xi_a^*, a} \| h_{\xi_a, a}) = B_a(\xi_a) - B_a(\xi_a^*) + (\xi_a^* - \xi_a^*) B'_a(\xi_a^*) \geq \frac{1}{2} \|\xi_a^* - \xi_a^*\|^2 \min_{\tilde{\xi}_a = z_a^T \theta, \theta \in \Theta} B''_a(\tilde{\xi}_a).$$

Thus, Assumption 7A is satisfied with  $C = \min_{\tilde{\xi}_a = z_a^T \theta, \theta \in \Theta} B''_a(\tilde{\xi}_a)/2$ .

Thus, to prove the corollary, it is sufficient to verify Assumption 3, which will be the focus of the rest of the proof.

Because the Fisher information is  $\mathcal{I}_{\xi_a, a}(\xi_a) = B''(\xi_a)$ , the first part of conditions of Assumption 3 on the smoothness of the Fisher information in  $\theta$  holds. We proceed to verify that  $\sum_a \mathcal{I}_a(\theta)$  is positive definite.

Note that

$$\mathcal{I}_a(\theta) = B''(z_a^T \theta) z_a z_a^T.$$

Thus,

$$\underline{c} \sum_{a \in \mathcal{A}} z_a z_a^T \leq \sum_{a \in \mathcal{A}} \mathcal{I}_a(\theta) \leq \bar{c} \sum_{a \in \mathcal{A}} z_a z_a^T,$$

where  $\underline{c} = \inf_{\theta \in \Theta, a \in \mathcal{A}} B''_{a_i}(z_a^T \theta) > 0$  and  $\bar{c} = \sup_{\theta \in \Theta, a \in \mathcal{A}} B''_{a_i}(z_a^T \theta) < \infty$ .

So, it is sufficient to show that  $\sum_{a \in \mathcal{A}} z_a z_a^T$  is non-singular. In the rest of the proof, we show that  $\sum_{a \in \mathcal{A}} z_a z_a^T$  is non-singular by proving the following result in linear algebra:

$$\{z_a; a \in \mathcal{A}\}^\perp = \ker \left( \sum_{a \in \mathcal{A}} z_a z_a^T \right). \quad (201)$$

**Proof of (201)** Because  $\sum_{a \in \mathcal{A}} z_a z_a^T$  is positive semidefinite,

$$\begin{aligned} z \in \ker \left( \sum_{a \in \mathcal{A}} z_a z_a^T \right) &\iff z^T \left( \sum_{a \in \mathcal{A}} z_a z_a^T \right) z = 0 \iff \sum_{a \in \mathcal{A}} |z_a^T z|^2 = 0 \\ &\iff \langle z, z_a \rangle = 0, \forall a \in \mathcal{A} \iff z \in \{z_a; a \in \mathcal{A}\}^\perp. \end{aligned}$$

Since  $\sum_{a \in \mathcal{A}} z_a z_a^T$  is symmetric,

$$\mathcal{R} \left( \sum_{a \in \mathcal{A}} z_a z_a^T \right) = \ker \left( \sum_{a \in \mathcal{A}} z_a z_a^T \right)^\perp = \left( \{z_a; a \in \mathcal{A}\}^\perp \right)^\perp = \text{span}\{z_a; a \in \mathcal{A}\} = \mathbb{R}^p.$$

This completes the proof of Corollary 5.1. □

*Proof of Corollary 5.2.* Because

$$f_{\boldsymbol{\theta},a}(x_a) = \exp(b_a x_a) \exp \{ \mathbf{z}_a^T \boldsymbol{\theta} - \log(1 + \exp(\mathbf{z}_a^T \boldsymbol{\theta} + b_a)) \}, x_a \in \{0, 1\},$$

the M2PL model described in (36) is a special case of the GLM described in (34), with  $B_a(\boldsymbol{\xi}_a) = \log\{1 + \exp(\boldsymbol{\xi}_a + b_a)\}$ , and  $\zeta^a(x_a) = \exp(b_a x_a)$ . Because the support of  $B_a(\cdot)$  is  $\mathbb{R}$ , conditions of Corollary 5.1 are satisfied. As a result, Corollary 5.2 follows by directly applying Corollary 5.2.  $\square$

*Proof of Corollary 5.3.* Note that the BTL model described in (39) is a special case of the M2PL model described in (36) with the following  $\mathbf{z}_a$  and  $b_a$  for  $a = (i, j)$ , and  $0 \leq i < j \leq p$ ,

$$(\mathbf{z}_a, b_a) = \begin{cases} (\mathbf{e}_j - \mathbf{e}_i, 0) & \text{if } i \neq 0 \\ (\mathbf{e}_j, 0) & \text{if } i = 0 \end{cases}. \quad (202)$$

Thus, Corollary 5.3 is implied by Corollary 5.2 as long as we can verify that a connected graph  $G$  ensures that  $\dim(\text{span}\{\mathbf{z}_a; a \in \mathcal{A}\}) = p$ . In the rest of the proof, we prove a slightly stronger result: for all  $\mathcal{G} = \{a^1, \dots, a^s\} \subset \mathcal{A}$ , if the graph  $(V, \mathcal{G})$  is connected, where  $V = \{0, 1, 2, \dots, p\}$ , then  $\dim(\text{span}\{\mathbf{z}_a; a \in \mathcal{A}\}) = p$ .

First of all, if the graph  $G = (V, E)$  is connected, then it implies that  $s \geq p$ . Let  $\mathbf{Z}_{\mathcal{G}} = [\mathbf{z}_{a^1}, \dots, \mathbf{z}_{a^s}]$ . It suffices to demonstrate that  $\text{rank}(\mathbf{Z}_{\mathcal{G}}) = p$ . To proceed, we construct a matrix that possesses the same rank as  $\mathbf{Z}_{\mathcal{G}}$ , as described below. For  $a = (i, j)$ ,  $0 \leq i < j \leq p$ , let

$$\mathbf{z}_a^+ = \begin{cases} (-1, \mathbf{z}_a^T)^T & \text{if } i = 0 \\ (0, \mathbf{z}_a^T)^T & \text{if } i > 0. \end{cases} \quad (203)$$

Then, define  $\mathbf{Z}_{\mathcal{G}}^+ = [\mathbf{z}_{a^1}^+, \dots, \mathbf{z}_{a^s}^+] \in \mathbb{R}^{(p+1) \times s}$ . Note that  $\mathbf{z}_a^+ = (-\mathbf{z}_a^T \mathbf{1}_p, \mathbf{z}_a^T)^T$  for all  $a$ . Consequently,  $\text{rank}(\mathbf{Z}_{\mathcal{G}}) = \text{rank}(\mathbf{Z}_{\mathcal{G}}^+)$ .

Let  $\mathbf{e}_0^+ = \mathbf{e}_1, \dots, \mathbf{e}_p^+ = \mathbf{e}_{p+1}$ , where  $\mathbf{e}_1, \dots, \mathbf{e}_{p+1}$  is the standard basis for  $\mathbb{R}^{p+1}$ . It is easy to check that  $\mathbf{z}_a^+ = \mathbf{e}_{a_1}^+ - \mathbf{e}_{a_2}^+$ , where  $a = (a_1, a_2)$ . Let  $v_1 = a_1^1$ ,  $v_2 = a_2^1$ ,  $S_2 = \{v_1, v_2\}$  and  $\mathcal{G}_{-1} = \{a \in \mathcal{G}; a \neq a^1\}$ , where  $a^1 = (a_1^1, a_2^1)$ . Set  $\hat{a}_1 = a^1$ . Now we know that  $\text{rank}(\mathbf{z}_{\hat{a}_1}^+) = 1$ .

Since  $(V, \mathcal{G})$  is a connected graph, there exists  $a' \in \mathcal{G}_{-1}$  such that  $a' = (a'_1, a'_2)$   $a'_1 \in S_2$  and  $a'_2 \notin S_2$  (or  $a'_2 \in S_2$  and  $a'_1 \notin S_2$ ). Set  $v_3 = a'_2$  (or  $v_3 = a'_1$ ),  $S_3 = S_2 \cup \{v_3\}$ ,  $\hat{a}_2 = a'$ , and  $\mathcal{G}_{-2} = \{a \in \mathcal{G}_{-1}; a \neq \hat{a}_2\}$ . By our construction, we know that  $\mathbf{z}_{\hat{a}_2}^+ = \mathbf{e}_{a'_1}^+ - \mathbf{e}_{a'_2}^+ \notin \text{span}\{\mathbf{z}_{\hat{a}_1}^+\}$ . Thus,  $\text{rank}([\mathbf{z}_{\hat{a}_1}^+, \mathbf{z}_{\hat{a}_2}^+]) > \text{rank}(\mathbf{z}_{\hat{a}_1}^+)$ .

Because  $(V, \mathcal{G})$  is a connected graph, we can always repeat the above process, until

$S_{p+1} = V$ . By this process, we obtain a sequence  $\hat{a}_1, \dots, \hat{a}_p$ , such that

$$1 = \text{rank}(\mathbf{z}_{\hat{a}_1}^+) < \dots < \text{rank}([\mathbf{z}_{\hat{a}_1}^+, \dots, \mathbf{z}_{\hat{a}_{p-1}}^+]) < \text{rank}([\mathbf{z}_{\hat{a}_1}^+, \dots, \mathbf{z}_{\hat{a}_p}^+]) = p.$$

Notice that  $p = \text{rank}([\mathbf{z}_{\hat{a}_1}^+, \dots, \mathbf{z}_{\hat{a}_p}^+]) \leq \text{rank}(\mathbf{Z}_G^+) = \text{rank}(\mathbf{Z}_G) \leq p$ . This implies that  $\text{rank}(\mathbf{Z}_G) = p$ .

□

## References

- Anscombe, F. J. (1952). Large-sample theory of sequential estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 48, pages 600–607. Cambridge University Press.
- Bartroff, J., Finkelman, M., and Lai, T. L. (2008). Modern sequential analysis and its applications to computerized adaptive testing. *Psychometrika*, 73:473–486.
- Bhatia, R. (1997). *Matrix analysis*, volume 169. Springer Science & Business Media.
- Billingsley, P. (1999). *Convergence of probability measures*. John Wiley & Sons.
- Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.
- Carlen, E. (2010). Trace inequalities and quantum entropy: an introductory course. *Entropy and the quantum*, 529:73–140.
- Chang, H.-H. and Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20(3):213–229.
- Chang, H.-H. and Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics*, pages 1466–1488.
- Chaudhuri, K., Kakade, S. M., Netrapalli, P., and Sanghavi, S. (2015). Convergence rates of active learning for maximum likelihood estimation. *Advances in Neural Information Processing Systems*, 28.
- Chen, X., Bennett, P. N., Collins-Thompson, K., and Horvitz, E. (2013). Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 193–202.

- Chen, X., Chen, Y., and Li, X. (2022). Asymptotically optimal sequential design for rank aggregation. *Mathematics of Operations Research*, 47(3):2310–2332.
- Chen, X., Jiao, K., and Lin, Q. (2016). Bayesian decision process for cost-efficient dynamic ranking via crowdsourcing. *The Journal of Machine Learning Research*, 17(1):7617–7656.
- Chen, X., Liu, Q., and Wang, Y. (2023). Active learning for contextual search with binary feedback. *Management Science*, 69(4):2165–2181.
- Chen, X. and Wang, Y. (2023). Robust dynamic pricing with demand learning in the presence of outlier customers. *Operations Research*, 71(4):1362–1386.
- Chen, Y., Li, X., Liu, J., and Ying, Z. (2024). Item response theory—a statistical framework for educational and psychological measurement. *Statistical Science*.
- Chen, Y. and Ye, X. (2011). Projection onto a simplex. *arXiv preprint arXiv:1101.6081*.
- Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, 74:619–632.
- Chernoff, H. (1959). Sequential Design of Experiments. *The Annals of Mathematical Statistics*, 30(3):755 – 770.
- DasGupta, A. (2008). *Asymptotic theory of statistics and probability*, volume 180. Springer.
- Duncan, L. R. (1959). Individual choice behavior: A theoretical analysis. *Courier Corporation*.
- Elo, A. (1978). The rating of chessplayers past and present. arco pub (1978). *Glickman, ME, Paired comparison models with time-varying parameters, Tech.*
- Embretson, S. E. and Reise, S. P. (2013). *Item response theory*. Psychology Press.
- Ghorpade, S. R. and Limaye, B. V. (2006). *A course in calculus and real analysis*. Springer.
- Hall, P. and Heyde, C. C. (1980). *Martingale limit theory and its application*. Academic press.
- Hilbert, D. and Courant, R. (1953). *Methods of Mathematical Physics*, volume 1. Interscience, New York.
- Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *The Annals of Statistics*, pages 849–879.

- Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.
- Lin, Z. (2010). *Probability inequalities*. Springer.
- Maystre, L. and Grossglauser, M. (2017). Just sort it! a simple and effective approach to active preference learning. In *International Conference on Machine Learning*, pages 2344–2353. PMLR.
- McCullagh, P. (2019). *Generalized linear models*. Routledge.
- Mukherjee, A., Tajer, A., Chen, P.-Y., and Das, P. (2022). Active Sampling of Multiple Sources for Sequential Estimation. *IEEE Transactions on Signal Processing*, 70:4571–4585.
- Mukhopadhyay, N. and Chattopadhyay, B. (2012). A tribute to Frank Anscombe and random central limit theorem from 1952. *Sequential Analysis*, 31(3):265–277.
- Naghshvar, M. and Javidi, T. (2013). Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703 – 2738.
- Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The Pagerank citation ranking: Bringing order to the web. Technical report, Stanford infolab.
- Rakhlin, A., Sridharan, K., and Tewari, A. (2015). Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161:111–153.
- Reckase, M. D. (2006). Multidimensional item response theory. *Handbook of statistics*, 26:607–642.
- Robbins, H. and Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Elsevier.
- Ross, S. M. (2014). *Introduction to probability models*. Academic press.
- Saaty, T. L. and Vargas, L. G. (2012). The possibility of group choice: pairwise comparisons and merging functions. *Social Choice and Welfare*, 38(3):481–496.
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychology Review*, 34:273–286.

- Tu, D., Han, Y., Cai, Y., and Gao, X. (2018). Item selection methods in multidimensional computerized adaptive testing with polytomously scored items. *Applied Psychological Measurement*, 42(8):677–694.
- Vaart, A. v. d. and Wellner, J. A. (1997). Weak convergence and empirical processes with applications to statistics. *Journal of the Royal Statistical Society-Series A Statistics in Society*, 160(3):596–608.
- Van Der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*, 24(4):398–412.
- Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., and Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.
- Wang, C. and Chang, H.-H. (2011). Item selection in multidimensional computerized adaptive testing—Gaining information from different angles. *Psychometrika*, 76(3):363–384.
- Wang, C., Chang, H.-H., and Boughton, K. A. (2011). Kullback–Leibler information and its applications in multi-dimensional adaptive testing. *Psychometrika*, 76:13–39.
- Wang, S., Fellouris, G., and Chang, H.-H. (2017). Computerized adaptive testing that allows for response revision: Design and asymptotic theory. *Statistica Sinica*, pages 1987–2010.
- Yang, M., Biedermann, S., and Tang, E. (2013). On optimal designs for nonlinear models: a general and efficient algorithm. *Journal of the American Statistical Association*, 108(504):1411–1420.
- Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323.
- Zhang, F. (2006). *The Schur complement and its applications*, volume 4. Springer Science & Business Media.