

Online Bernstein–von Mises theorem

Jeyong Lee

JYLEE1024@POSTECH.AC.KR

*Department of Industrial and Management Engineering
Pohang University of Science and Technology
Pohang, Gyeongbuk 37673, South Korea*

Junhyeok Choi

CJUNH4810@POSTECH.AC.KR

*Department of Industrial and Management Engineering
Pohang University of Science and Technology
Pohang, Gyeongbuk 37673, South Korea*

Minwoo Chae*

MCHAE@POSTECH.AC.KR

*Department of Industrial and Management Engineering
Pohang University of Science and Technology
Pohang, Gyeongbuk 37673, South Korea*

Editor: Daniel Roy

Abstract

Online learning is an inferential paradigm in which parameters are updated incrementally from sequentially available data, in contrast to batch learning, where the entire dataset is processed at once. In this paper, we assume that mini-batches from the full dataset become available sequentially. The Bayesian framework, which updates beliefs about unknown parameters after observing each mini-batch, is naturally suited for online learning. At each step, we update the posterior distribution using the current prior and new observations, with the updated posterior serving as the prior for the next step. However, this recursive Bayesian updating is rarely computationally tractable unless the model and prior are conjugate. When the model is regular, the updated posterior can be approximated by a normal distribution, as justified by the Bernstein–von Mises theorem. We adopt a variational approximation at each step and investigate the frequentist properties of the final posterior obtained through this sequential procedure. Under mild assumptions, we show that the accumulated approximation error becomes negligible once the mini-batch size exceeds a threshold depending on the parameter dimension. As a result, the sequentially updated posterior is asymptotically indistinguishable from the full posterior.

Keywords: Bernstein–von Mises theorem, Laplace approximation, Bayesian online learning, variational approximation

1. Introduction

Online learning is an inferential paradigm in which parameters are updated sequentially as new data arrive. Unlike batch learning, which processes a fixed dataset all at once, online learning incrementally adjusts parameters with each new observation. This approach is particularly well-suited for analyzing streaming data, where data become available sequentially. Moreover, even for fixed datasets, online learning can be an advantageous approach, as its

*. Corresponding author.

algorithms are often significantly more computationally efficient than batch learning methods. Over the past few decades, substantial progress has been made in the development of online learning techniques for various statistical models, such as topic models (Hoffman et al., 2010; Wang et al., 2011; Kim et al., 2016), matrix factorization (Mairal et al., 2010), survival analysis (Xue et al., 2020; Wu et al., 2021; Choi et al., 2025) and quantile regression (Chen et al., 2019; Lee et al., 2024). We provide a comprehensive review of online learning methodologies in Section 1.1.

The Bayesian philosophy, which updates beliefs about an unknown parameter after observing data, aligns well with the online learning paradigm. As new data arrive, one can update the posterior distribution using the current prior and the new data, with the updated posterior serving as the new prior. However, unless the model-prior pair is conjugate, this straightforward Bayesian approach is rarely computationally tractable, particularly when dealing with complex hierarchical models. Consequently, Bayesian approaches to online learning typically include an additional step to approximate the updated posterior with a simpler one, ensuring computational feasibility for subsequent updates (Opper and Winther, 1999; Solla and Winther, 1999). A common choice for this additional step is variational approximation (Broderick et al., 2013; Lin, 2013; Nguyen et al., 2018).

In the Bayesian online learning framework described above, the posterior distribution must be repeatedly approximated as new data arrive. While the approximation error at each step may be negligible, the cumulative error over multiple updates may not be. Thus, a central concern is whether the accumulated approximation error remains small. In this paper, we provide a rigorous theoretical analysis of this issue by investigating the frequentist properties of the sequentially updated and approximated posterior distributions.

To set the scene, let $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ be a statistical model indexed by $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^p$ is the parameter space. Let $\mathbf{D} = (Y_1, \dots, Y_N)$ be the complete set of observations, and let $\Pi \in \mathcal{Q}$ denote the (initial) prior on Θ , where \mathcal{Q} is a class of priors. The corresponding posterior distribution is $\Pi(\cdot | \mathbf{D})$, which we refer to as the full posterior. We assume that the data become available sequentially in the order $\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_T$, where $\mathbf{D}_t = (Y_{N_{t-1}+1}, Y_{N_{t-1}+2}, \dots, Y_{N_t})$ represents the t -th mini-batch, with $N_0 = 0$. For convenience, we assume that each mini-batch has the same size, denoted by n , so that $N_t = nt$ for every t .

With $\Pi_0 = \Pi$, the Bayesian online learning framework considered in this paper consists of the following inductive steps. For $t \geq 1$, let $\tilde{\Pi}_t(\cdot | \mathbf{D}_t)$ be the posterior distribution obtained by updating the prior Π_{t-1} with the data \mathbf{D}_t using Bayes' formula:

$$\tilde{\Pi}_t(\mathcal{A} | \mathbf{D}_t) = \frac{\int_{\mathcal{A}} \exp\{L_t(\theta)\} d\Pi_{t-1}(\theta)}{\int_{\Theta} \exp\{L_t(\theta)\} d\Pi_{t-1}(\theta)}, \quad \text{for any measurable } \mathcal{A} \subset \Theta, \quad (1)$$

where $L_t(\cdot)$ is the log-likelihood corresponding to the t -th mini-batch \mathbf{D}_t . Next, we approximate the updated posterior $\tilde{\Pi}_t(\cdot | \mathbf{D}_t)$ by projecting it onto \mathcal{Q} , the space of prior distributions. Since we use a variational approximation, the approximated posterior is given by

$$\Pi_t = \operatorname{argmin}_{Q \in \mathcal{Q}} K(Q; \tilde{\Pi}_t(\cdot | \mathbf{D}_t)), \quad (2)$$

where $K(P; Q)$ denotes the Kullback–Leibler (KL) divergence, defined as

$$K(P; Q) = \int \log \left(\frac{dP}{dQ} \right) dP. \quad (3)$$

Thus, we obtain the sequence $(\Pi_t)_{t \leq T}$ of approximated posterior distributions, where $\Pi_t \in \mathcal{Q}$ for every t .

The main goal of this paper is to provide sufficient conditions under which the accumulated approximation error in the online learning process remains negligible, ensuring that the sequentially updated posterior distribution $\Pi_T(\cdot)$ is nearly identical to the full posterior distribution $\Pi(\cdot | \mathbf{D})$. Although this is intuitively obvious when \mathcal{Q} is sufficiently large, a mathematically rigorous analysis is challenging even for $T = 2$.

We assume that Y_1, \dots, Y_N are independent but not necessarily identically distributed, and there is the true parameter $\theta_0 \in \Theta$ generating data. This is a frequentist assumption that is commonly adopted in the Bayesian asymptotics literature. We also assume that the parametric model \mathcal{P} is regular in the sense that the log-likelihood function is locally approximately quadratic around θ_0 . Under this regularity condition, the full posterior distribution $\Pi(\cdot | \mathbf{D})$ is approximately normal, centered around the linear efficient estimator, with variance given by the inverse Fisher information matrix, as stated by the celebrated Bernstein–von Mises (BvM) theorem. Accordingly, the class of normal distributions serves as a natural choice for \mathcal{Q} .

Our main theorem guarantees, under suitable assumptions, that the total variation distance between the full posterior $\Pi(\cdot | \mathbf{D})$ and the sequentially updated posterior $\Pi_T(\cdot)$ is sufficiently small with high probability. When the dimension p of θ_0 is fixed, the required condition boils down to $n \gg (\log N)^4$. The main results are formulated in a non-asymptotic framework and also cover the case where p diverges at a polynomial rate with respect to the sample size N . In particular, when $\log p \asymp \log N$, $n \gg p^3$ is sufficient; see Theorem 11 for detailed statements. We refer to this result as the *online BvM theorem*. It ensures the frequentist validity and asymptotic efficiency of statistical inferences, such as point estimation and uncertainty quantification, based on Π_T .

To prove the main results, sharp non-asymptotic results on the quadratic approximation of the log-likelihood function are crucial, for which substantial progress has been made in recent studies; see Spokoiny (2012), Spokoiny (2017) and Katsevich and Rigollet (2024). In addition to these techniques, we develop several novel techniques, including bounding the KL divergence between the updated posterior and the approximated variational posterior, as well as handling accumulated errors. It is also important to note that, while the prior considered in the standard BvM theorem is typically a flat prior, the priors used in the intermediate steps of the online learning process are highly informative.

To the best of our knowledge, no theoretical work has studied the theory of Bayesian online learning with the level of rigor presented in this paper. In contrast, rigorous theoretical frameworks have recently been developed for frequentist online learning based on empirical risk minimization using one-pass stochastic gradient descent (SGD) algorithms. Specifically, Toulis and Airoldi (2017) demonstrated that the *implicit SGD* estimator is asymptotically efficient, admitting a limiting distribution with optimal variance. In addition, Chen et al. (2020) proposed the *batch-means SGD* estimator, which yields asymptotically valid confidence intervals.

Since the one-pass SGD approaches described above process a single datum at each step, it is natural to ask whether similar results can be achieved with $T = N$. We conjecture that the online BvM theorem does not hold in this case. Although we do not have a formal proof, our numerical experiments in Section 9 provide supporting evidence. Specifically, we observe that the relative efficiency of the point estimator—the posterior mean obtained from the sequentially updated Π_t —when $T = N$, compared to a batch estimator, is significantly greater than 1, indicating that it is not asymptotically efficient. We believe the online BvM theorem fails to hold because our results rely on the quadratic approximation of the log-likelihood, which is valid only when the mini-batch sample size n is much larger than the parameter dimension p . Given the positive results of one-pass algorithms in Toulis and Airolidi (2017) and Chen et al. (2020), it would be interesting to develop an online Bayesian procedure with appropriate algorithmic modifications that ensure the validity of the online BvM theorem.

The remainder of this paper is organized as follows. In the next subsections, we provide a comprehensive review of online learning methods and introduce elementary notations. Section 2 describes the basic setup and key definitions related to Bayesian online learning. Section 3 provides a high-level overview of the online BvM theorem and presents a motivating example illustrating where our theory applies. Sections 4 and 5 present the variational approximation of the sequentially updated posterior and the penalized M-estimation, respectively. Section 6 provides a non-asymptotic analysis of several regularity quantities. Section 7 establishes the BvM theorem for the full posterior. Our main results concerning the online BvM theorem are presented in Section 8. Numerical results supporting our theory are provided in Section 9. Concluding remarks follow in Section 10, and all proofs and additional technical details are deferred to the Appendix.

1.1 Related works

1.1.1 FREQUENTIST METHODS

In this sub-section, we provide a brief introduction to the recent advancements in online statistical inference by surveying theoretical investigations of SGD-type estimators. Due to its computational advantages, the SGD estimator (Robbins and Monro, 1951) and its variants have been extensively studied in the frequentist online learning literature. As pioneering works, Ruppert (1988) and Polyak and Juditsky (1992) independently proposed averaging of SGD iterates, with Polyak and Juditsky (1992) establishing the asymptotic normality of the averaged SGD estimator. More recently, Toulis and Airolidi (2017) introduced the implicit SGD estimator, which exhibits stable performance in finite samples and is asymptotically normal under a suitably specified learning rate.

However, in the online learning setting, the asymptotic normality alone does not guarantee that confidence intervals can be constructed. This is because the asymptotic covariance matrix is typically computed using the entire dataset, which is not available when data arrives sequentially. For example, if our interest lies in the Fisher information matrix evaluated at an estimator $\hat{\theta}_T$, we would need to evaluate $\sum_{t=1}^T \partial^2 L_t(\theta) / \partial \theta \partial \theta^\top$ at $\hat{\theta}_T$, where $L_t(\theta)$ denotes the log-likelihood function for the t -th mini-batch \mathbf{D}_t . In a batch learning setup, the computation of this matrix is straightforward. However, in an online learning setting, the early mini-batches $\mathbf{D}_1, \dots, \mathbf{D}_{T-1}$ are discarded once \mathbf{D}_T arrives, while $\hat{\theta}_T$ is only

available at time T . This limitation hinders the applicability of the conventional batch learning approach.

As a result, constructing tractable and asymptotically valid confidence intervals has become an important topic in online learning literature. To address this challenge, Chen et al. (2020) proposed a *batch-means* method that aims to estimate the limiting covariance matrix, and Zhu et al. (2023) improved upon this method by eliminating the need for a priori knowledge of the total sample size N . In another line of work, Lee et al. (2022) developed computationally efficient confidence intervals by applying a functional central limit theorem to SGD iterates, which further extended to quantile regression (Lee et al., 2024). Alternatively, instead of directly estimating the limiting covariance matrix, Fang et al. (2018) constructed tractable confidence intervals by employing a bootstrap resampling procedure based on randomly perturbed SGD updates.

1.1.2 BAYESIAN METHODS

In recent years, significant methodological advancements have been made in Bayesian online learning. Several studies have proposed various approximation methods for updating the posterior distribution sequentially. For example, Broderick et al. (2013) introduced a general framework for large-scale and streaming data, and Nguyen et al. (2018) adapted the online VB algorithm for neural networks. In the realm of nonparametric models, Lin (2013) developed a VB algorithm for Dirichlet process mixture models, while Jeong et al. (2023) employed an assumed density filtering (ADF) approach for similar tasks. More recently, Lambert et al. (2022)¹ and Lambert et al. (2023) proposed computationally efficient online VB approximations using the Gaussian variational family.

These works demonstrate that projecting onto a tractable class \mathcal{Q} of distributions is a popular strategy due to its computational feasibility. Among the proposed methods, both ADF and VB are prominent. However, ADF requires a moment-matching step for each mini-batch, which can be computationally costly for certain hierarchical models (e.g., topic models), whereas VB can avoid this step (Broderick et al., 2013). Consequently, VB and its variants have emerged as the preferred approaches in recent Bayesian online learning literature (Lin, 2013; Bui et al., 2017; Nguyen et al., 2018; Lambert et al., 2022; Choi et al., 2025).

1.2 Notations

For two real numbers a and b , $a \vee b$ and $a \wedge b$ denote the maximum and minimum of a and b , respectively. For two positive sequences (a_n) and (b_n) , $a_n \lesssim b_n$ (or $a_n = O(b_n)$) means that $a_n \leq Cb_n$ for some constant $C \in (0, \infty)$. Also, $a_n \asymp b_n$ indicates that $a_n \lesssim b_n$ and $b_n \lesssim a_n$. The notation $a_n \ll b_n$ (or $a_n = o(b_n)$) implies that $a_n/b_n \rightarrow 0$ as $n \rightarrow \infty$. We use the standard O_P and o_p notation for stochastic order symbols; see van der Vaart (2000). With a slight abuse of notation, for $\nu_n \in \mathbb{R}^p$ satisfying $\|\nu_n\|_2 = o(1)$, we write $\nu_n = o(1)$. For $1 \leq q \leq \infty$, $\|\cdot\|_q$ indicates the ℓ_q -norm of a vector. For $m \in \mathbb{N}$, let $[m] = \{1, 2, \dots, m\}$.

1. The procedure proposed in Lambert et al. (2022) coincides with the setup that will be introduced in Section 2.1.

Table 1: Important notations

Notation	Location	Notation	Location
$\Theta(\theta, \mathbf{F}, r), \Theta(\mathbf{F}, r)$	(6), (7)	$\mathbf{F}_{t,\theta}, \widetilde{\mathbf{F}}_{t,\theta}$	(5)
$r_{\text{eff},t}, \widetilde{r}_{\text{eff},1:t}, r_{\text{eff},1:t}$	(28), (43)	$\mathcal{E}_{\text{est},1}, \mathcal{E}_{\text{est},2}$	(32), (44)
$p_{\text{eff},t}, \lambda_t$	(28)	$r_{\text{LA}}, \widehat{\tau}_{3,t}, \widehat{\tau}_{4,t}$	(19)
M_n	(27), (42)	$\widehat{\tau}_{3,t,r}$	(24)
p_*	(EX)	$\tau_{3,t}^*$	(29)
$d_V(\cdot, \cdot), K(\cdot; \cdot)$	(20), (3)	$\epsilon_{n,t,\text{TV}}, \epsilon_{n,t,\text{KL}}$	Theorem 2, 3
$\widehat{\theta}_t, \theta_t^*$	(9)	K_{\min}, K_{\max}	(34), (35), (36)
$\widehat{\theta}_{1:t}, \widehat{\theta}_{1:t}^{\text{ML}}, \theta_{1:t}^*$	(40)	$K_{\text{low}}, K_{\text{up}}$	Proposition 6

For $z = (z_j)_{j \in [p]} \in \mathbb{R}^p$ and $k \geq 2$, let

$$z^{\otimes k} = (z_{i_1} \times \dots \times z_{i_k})_{i_1, \dots, i_k \in [p]} \in \mathbb{R}^{p^k}.$$

For two k -order tensors $\mathbf{A} = (A_{i_1, \dots, i_k})_{i_1, \dots, i_k \in [p]} \in \mathbb{R}^{p^k}$ and $\mathbf{B} = (B_{i_1, \dots, i_k})_{i_1, \dots, i_k \in [p]} \in \mathbb{R}^{p^k}$, let

$$\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i_1, \dots, i_k \in [p]} A_{i_1, \dots, i_k} B_{i_1, \dots, i_k}.$$

Let \mathcal{S}_{++}^p denote the set of all $p \times p$ -dimensional symmetric positive definite matrices. Let $\mathbf{I}_p \in \mathbb{R}^{p \times p}$ denote the identity matrix. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n \times p}$, let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote the smallest and largest singular values of \mathbf{A} , respectively. For simplicity, $\|\mathbf{A}\|_2$ will often be used interchangeably with $\lambda_{\max}(\mathbf{A})$. Let $\|\mathbf{A}\|_F = (\sum_{ij} a_{ij}^2)^{1/2}$ be the Frobenius norm. For two distinct matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$, $\mathbf{A} \succeq \mathbf{B}$ means $\mathbf{A} - \mathbf{B}$ is positive semi-definite matrix. For a k -th order tensor $\mathbf{A} = (A_{i_1, \dots, i_k})_{i_1, \dots, i_k \in [p]} \in \mathbb{R}^{p^k}$, define the operator norm of \mathbf{A} by

$$\|\mathbf{A}\|_{\text{op}} = \sup_{u_1, \dots, u_k \in \mathcal{U}} |\langle \mathbf{A}, u_1 \otimes \dots \otimes u_k \rangle|,$$

where $\mathcal{U} = \{u \in \mathbb{R}^p : \|u\|_2 = 1\}$.

For $\sigma^2 \geq 0$, a random vector $X \in \mathbb{R}^p$ is said to be SubG(σ^2) if

$$\log \mathbb{E} \exp(\alpha^\top (X - \mathbb{E}X)) \leq \sigma^2 \|\alpha\|_2^2 / 2, \quad \forall \alpha \in \mathbb{R}^p$$

With the convention, we consider that $\inf \emptyset = \infty$ in this paper. Throughout our paper, the constants c_1, c_2, \dots may vary depending on the context.

2. Preliminaries

2.1 Setup for online learning

In this subsection, we precisely formulate the Bayesian online learning procedure briefly introduced in the introduction. Suppose the entire dataset $\mathbf{D} = (Y_i)_{i \in [N]}$ consists of in-

dependent (not necessarily identically distributed) observations Y_1, \dots, Y_N . The dataset is partitioned into T mini-batches of equal size n , so that the total sample size up to the t -th mini-batch is given by $N_t = nt$ for all $t \in \{0, 1, 2, \dots, T\}$. The t -th mini-batch and the collection of all samples up to the t -th mini-batch are denoted as

$$\mathbf{D}_t = (Y_{N_{t-1}+1}, Y_{N_{t-1}+2}, \dots, Y_{N_t}), \quad \text{and} \quad \mathbf{D}_{1:t} = (\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_t) = (Y_i)_{i \in [N_t]}.$$

For $i \in [N]$, let $p_{\theta,i}(\cdot)$ be the probability density function for Y_i parametrized by $\theta \in \Theta \subset \mathbb{R}^p$, and let $\ell_{\theta,i}(y) = \log p_{\theta,i}(y)$ be the log density. For $t \in [T]$, let

$$L_t(\theta) = L_t(\theta; \mathbf{D}_t) = \sum_{i=N_{t-1}+1}^{N_t} \ell_{\theta,i}(Y_i).$$

Let $\mathbb{P}_\theta^{(N)}$ denote the joint probability measure corresponding to the product density function $(y_1, y_2, \dots, y_N) \mapsto \prod_{i=1}^N p_{\theta,i}(y_i)$. We assume that the model is well-specified; that is, \mathbf{D} is generated from $\mathbb{P}_{\theta_0}^{(N)}$ for some *true parameter* $\theta_0 \in \Theta$. Let $\mathbb{P}_{0,t}(\cdot)$ denote the joint probability measure corresponding to the product density

$$(y_{N_{t-1}+1}, y_{N_{t-1}+2}, \dots, y_{N_t}) \mapsto \prod_{i=N_{t-1}+1}^{N_t} p_{\theta_0,i}(y_i),$$

and let \mathbb{E}_t denote the expectation under $\mathbb{P}_{0,t}$. Since the model is assumed to be well-specified, we have $\theta_0 = \operatorname{argmax}_{\theta \in \Theta} \mathbb{E}_t L_t(\theta)$ for all $t \in [T]$.

Now, we introduce our online learning procedure. Let \mathcal{Q} be the collection of all Gaussian measures with nonsingular covariance matrices. Given an initial prior Π_0 and the log-likelihood $L_t(\theta)$ for the t -th mini-batch data \mathbf{D}_t , we iteratively define the posterior distribution $\tilde{\Pi}_t(\cdot | \mathbf{D}_t)$ and the corresponding variational approximation $\tilde{\Pi}_t(\cdot)$ as described in (1) and (2). We denote by $\tilde{\pi}_t(\cdot | \mathbf{D}_t)$ and $\pi_t(\cdot)$ the density functions of $\tilde{\Pi}_t(\cdot | \mathbf{D}_t)$ and $\Pi_t(\cdot)$, respectively.

In particular, we consider a normal distribution $\mathcal{N}(\mu_0, \mathbf{\Omega}_0^{-1})$ as the initial prior Π_0 , where $\mu_0 \in \mathbb{R}^p$ and $\mathbf{\Omega}_0 \in \mathcal{S}_{++}^p$. For $t \in \{0, 1, \dots, T\}$, we denote

$$\Pi_t = \mathcal{N}(\mu_t, \mathbf{\Omega}_t^{-1}) \tag{4}$$

for $\mu_t \in \mathbb{R}^p$ and $\mathbf{\Omega}_t \in \mathcal{S}_{++}^p$.

2.2 Definitions

For a four times differentiable function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ with $\theta \mapsto f(\theta)$, let

$$\begin{aligned} \nabla f(\theta) &= \left(\frac{\partial}{\partial \theta_{i_1}} f(\theta) \right)_{i_1 \in [p]} \in \mathbb{R}^p, \quad \nabla^2 f(\theta) = \left(\frac{\partial^2}{\partial \theta_{i_1} \partial \theta_{i_2}} f(\theta) \right)_{i_1, i_2 \in [p]} \in \mathbb{R}^{p \times p}, \\ \nabla^3 f(\theta) &= \left(\frac{\partial^3}{\partial \theta_{i_1} \partial \theta_{i_2} \partial \theta_{i_3}} f(\theta) \right)_{i_1, i_2, i_3 \in [p]} \in \mathbb{R}^{p \times p \times p}, \\ \nabla^4 f(\theta) &= \left(\frac{\partial^4}{\partial \theta_{i_1} \partial \theta_{i_2} \partial \theta_{i_3} \partial \theta_{i_4}} f(\theta) \right)_{i_1, i_2, i_3, i_4 \in [p]} \in \mathbb{R}^{p \times p \times p \times p}. \end{aligned}$$

To ensure an accurate approximation of the posterior, we will impose several smoothness conditions on the log-likelihood L_t , requiring that it is at least four times continuously differentiable with probability 1. For such a differentiable $L_t(\cdot)$, we introduce some notations used for posterior analysis. For $t \in [T]$ and $\theta \in \Theta$, define

$$\mathbf{F}_{t,\theta} = -\nabla^2 L_t(\theta) \in \mathbb{R}^{p \times p}, \quad \tilde{\mathbf{F}}_{t,\theta} = \boldsymbol{\Omega}_{t-1} + \mathbf{F}_{t,\theta}, \quad (5)$$

where $\boldsymbol{\Omega}_{t-1}$ is defined in (4). For $\theta_c \in \mathbb{R}^p$, $r \geq 0$ and $\mathbf{F} \in \mathcal{S}_{++}^p$, let

$$\Theta(\theta_c, \mathbf{F}, r) = \left\{ \theta \in \Theta : \|\mathbf{F}^{1/2}(\theta - \theta_c)\|_2 \leq r \right\} \quad (6)$$

denote the local elliptical vicinity of θ_c . For notational simplicity, let

$$\Theta(\mathbf{F}, r) = \Theta(0, \mathbf{F}, r). \quad (7)$$

Given the prior distribution Π_{t-1} , which reflects the information in $\mathbf{D}_{1:t-1}$, we define the penalized log-likelihood function as

$$\tilde{L}_t(\theta) = L_t(\theta) - \frac{1}{2} \left\| \boldsymbol{\Omega}_{t-1}^{1/2}(\theta - \mu_{t-1}) \right\|_2^2. \quad (8)$$

For each $t \in [T]$, the penalized maximum likelihood estimator (pMLE) and its population version are defined as

$$\hat{\theta}_t = \operatorname{argmax}_{\theta \in \Theta} \tilde{L}_t(\theta), \quad \theta_t^* = \operatorname{argmax}_{\theta \in \Theta} \mathbb{E}_t \tilde{L}_t(\theta). \quad (9)$$

From the standard M-estimation theory, it is expected that $\hat{\theta}_t$ is close to θ_t^* ; this will be addressed in Section 5. It is noteworthy that $\hat{\theta}_t$ need not converge to θ_0 because, in general, $\theta_t^* \neq \theta_0$ due to the penalization term. Therefore, for $\hat{\theta}_t$ to be a reliable estimator of θ_0 , the bias $\|\theta_t^* - \theta_0\|_2$ should be sufficiently small. Roughly speaking, if $\|\boldsymbol{\Omega}_{t-1}^{1/2}(\theta_0 - \mu_{t-1})\|_2$ is not too large, one can expect that $\|\theta_t^* - \theta_0\|_2$ will also be small. The magnitude of this bias will be addressed in Proposition 6, followed by a refined analysis in (55).

2.3 Smoothness condition

In this subsection, we introduce an essential tool for characterizing the smoothness structure of (concave) functions—namely, *self-concordance*-type condition. For a simple illustration, consider a 3-times differentiable function $f : \mathbb{R} \rightarrow \mathbb{R}$ satisfying

$$|f'''(\theta)| \leq 2f''(\theta)^{3/2}, \quad \forall \theta \in \mathbb{R},$$

where f'' and f''' denote the second and third derivative of $f(\cdot)$, respectively. Then, one can prove that

$$(1 - \delta(\theta, \eta))f''(\theta) \leq f''(\eta) \leq 1/(1 - \delta(\theta, \eta))^2 f''(\theta), \quad \forall \theta, \eta \in \mathbb{R} \text{ with } \delta(\theta, \eta) \leq 1$$

where $\delta(\theta, \eta) = |f''(\theta)^{1/2}(\eta - \theta)|$; see Theorem 5.1.7 in Nesterov et al. (2018) for a general statement. Intuitively, this condition ensures that the second derivative of f does not change too abruptly, which in turn facilitates the analysis of local quadratic approximations.

Recently, the self-concordance condition has been invoked in the statistical literature. Originally introduced in the context of convex optimization problems (Nesterov and Nemirovskii, 1994), it has since been adapted for statistical applications (Bach, 2010; Ostrovskii and Bach, 2021; Spokoiny, 2025). These conditions facilitate non-asymptotic quadratic approximation theory, leading to sharp theoretical analyses. For this purpose, we rely heavily on the smoothness structures induced by the self-concordance condition.

Now, we formalize the notion of smoothness. Let $f : \Theta \rightarrow \mathbb{R}$ be a four-times differentiable function. For some $\mathbf{F} \in \mathcal{S}_{++}^p$ and $\tau_3, r \geq 0$, we say that f satisfies the *third order smoothness* at $\theta \in \Theta$ with parameters (τ_3, \mathbf{F}, r) if

$$\sup_{u \in \Theta(\mathbf{F}, r)} \sup_{z \in \mathbb{R}^p} \frac{|\langle \nabla^3 f(\theta + u), z^{\otimes 3} \rangle|}{\|\mathbf{F}^{1/2} z\|_2^3} \leq \tau_3. \quad (10)$$

Similarly, we say that f satisfies the *fourth order smoothness* at $\theta \in \Theta$ with parameters (τ_4, \mathbf{F}, r) if

$$\sup_{u \in \Theta(\mathbf{F}, r)} \sup_{z \in \mathbb{R}^p} \frac{|\langle \nabla^4 f(\theta + u), z^{\otimes 4} \rangle|}{\|\mathbf{F}^{1/2} z\|_2^4} \leq \tau_4. \quad (11)$$

By Theorem 2.1 in Zhang et al. (2012), each left-hand side in (10) and (11) is equal to the following expression:

$$\sup_{u \in \Theta(\mathbf{F}, r)} \sup_{z_1, \dots, z_k \in \mathbb{R}^p} \frac{|\langle \nabla^k f(\theta + u), z_1 \otimes \dots \otimes z_k \rangle|}{\|\mathbf{F}^{1/2} z_1\|_2 \times \dots \times \|\mathbf{F}^{1/2} z_k\|_2}, \quad k \in \{3, 4\}. \quad (12)$$

It is noteworthy that τ_3 and τ_4 can typically be chosen to be sufficiently small. More specifically, for $\theta \in \Theta$, $\mathbf{F} \in \mathcal{S}_{++}^p$ and $r \geq 0$, we prove in Lemma 42 that (10) and (11) hold with

$$\tau_3 = \lambda_{\min}^{-3/2}(\mathbf{F}) \sup_{u \in \Theta(\mathbf{F}, r)} \|\nabla^3 f(\theta + u)\|_{\text{op}} \quad \text{and} \quad \tau_4 = \lambda_{\min}^{-2}(\mathbf{F}) \sup_{u \in \Theta(\mathbf{F}, r)} \|\nabla^4 f(\theta + u)\|_{\text{op}}, \quad (13)$$

respectively. To aid understanding, suppose that $f = L_t$, $\mathbf{F} = -\nabla^2 L_t(\theta)$ with

$$\lambda_{\min}(\mathbf{F}) \asymp n, \quad \sup_{u \in \Theta(\mathbf{F}, r)} (\|\nabla^3 L_t(\theta + u)\|_{\text{op}} \vee \|\nabla^4 L_t(\theta + u)\|_{\text{op}}) \asymp n.$$

Then it follows that $\tau_3 \asymp n^{-1/2}$ and $\tau_4 \asymp n^{-1}$. Many of the subsequent analyses in this paper rely on the assumption that τ_3 and τ_4 are small enough. This can be achieved when $\lambda_{\min}(\mathbf{F})$ is large enough relative to the local third and fourth operator norms of L_t as described in (13). In this sense, $\lambda_{\min}(\mathbf{F})$ behaves as an *effective sample size* in our analysis. A detailed discussion of this point is deferred to Section 6.

3. High-level overview of the online BvM theorem

In this section, we provide a high-level overview of the online BvM theorem and provide a motivating example illustrating the applicability of our theory.

3.1 Proof framework and our contributions

For $t \in [T]$, we define the t -th *full posterior* by

$$\Pi(\mathcal{A} \mid \mathbf{D}_{1:t}) = \frac{\int_{\mathcal{A}} \exp\{L_{1:t}(\theta)\} d\Pi_0(\theta)}{\int_{\Theta} \exp\{L_{1:t}(\theta)\} d\Pi_0(\theta)} \quad \text{for any measurable } \mathcal{A} \subset \Theta, \quad (14)$$

where $L_{1:t}(\theta) = \sum_{s=1}^t L_s(\theta)$. As mentioned in Section 1, we aim to show that

$$d_V\left(\Pi_t, \Pi(\cdot \mid \mathbf{D}_{1:t})\right) = o_P(1), \quad \forall t \in [T], \quad (15)$$

which we call the *online BuM theorem*, stated in Theorem 11. To this end, we introduce a normal distribution \widehat{Q}_t satisfying

$$d_V\left(\widehat{Q}_t, \Pi(\cdot \mid \mathbf{D}_{1:t})\right) = o_P(1) \quad (16)$$

for all $t \in [T]$. In Section 7, \widehat{Q}_t will be specified in Theorem 7.

For regular parametric models, it is well known that the posterior distribution can be approximated by a normal distribution centered at an efficient estimator with variance equal to the inverse Fisher information matrix (van der Vaart, 2000). However, this normal distribution involves the Fisher information matrix, which depends on the unknown true parameter θ_0 . This motivates approximating the posterior by fully data-dependent distributions \widehat{Q}_t , a topic of significant interest in the Bayesian literature.

Substantial theoretical progress has been made in this direction, including Laplace approximation (LA) (Spokoiny, 2023, 2025; Katsevich, 2023) and VB methods (Katsevich and Rigollet, 2024). Accordingly, one may take such data-dependent distributions as \widehat{Q}_t . In Section 7, we will provide rigorous justifications for these arguments. In the remainder of this section, we focus on conveying the high-level ideas.

Now we present our high-level proof strategy. For $t \in [T]$, note that

$$\begin{aligned} d_V(\Pi_t, \Pi(\cdot \mid \mathbf{D}_{1:t})) &\leq d_V(\Pi_t, \widehat{Q}_t) + d_V(\widehat{Q}_t, \Pi(\cdot \mid \mathbf{D}_{1:t})) \\ &= d_V(\Pi_t, \widehat{Q}_t) + o_P(1). \end{aligned}$$

Therefore, it suffices to bound the first term $d_V(\Pi_t, \widehat{Q}_t)$.

To this end, we introduce an important formula. For two normal distributions $Q_1 = \mathcal{N}(m_1, \mathbf{V}_1^{-1})$ and $Q_2 = \mathcal{N}(m_2, \mathbf{V}_2^{-1})$ with sufficiently small $\|\mathbf{V}_2^{-1/2} \mathbf{V}_1 \mathbf{V}_2^{-1/2} - \mathbf{I}_p\|_2$, it holds that

$$d_V(Q_1, Q_2) \lesssim \left\| \mathbf{V}_1^{1/2} (m_1 - m_2) \right\|_2 \vee \left\| \mathbf{V}_2^{-1/2} \mathbf{V}_1 \mathbf{V}_2^{-1/2} - \mathbf{I}_p \right\|_{\mathbb{F}}. \quad (17)$$

Suppose that $\widehat{Q}_t = \mathcal{N}(\theta_t, \mathbf{F}_t^{-1})$ for some $\theta_t \in \Theta$ and $\mathbf{F}_t \in \mathcal{S}_{++}^p$.² By (17), the proof of (15) boils down to the problem of obtaining sharp upper bounds for the following quantities:

$$\left\| \boldsymbol{\Omega}_t^{1/2} (\theta_t - \mu_t) \right\|_2 \quad \text{and} \quad \left\| \mathbf{F}_t^{-1/2} \boldsymbol{\Omega}_t \mathbf{F}_t^{-1/2} - \mathbf{I}_p \right\|_{\mathbb{F}}. \quad (18)$$

2. Here, θ_t and \mathbf{F}_t are written in simplified form. More precise notations and a rigorous construction of these quantities will be introduced in Section 7.

Our main technical contribution lies in deriving explicit non-asymptotic bounds for these two terms in (18). To the best of our knowledge, no prior theoretical work has provided a rigorous non-asymptotic analysis of (18). Specifically, Propositions 9 and 10 in Section 8 address the first and second terms, respectively. We will revisit this issue in Section 8.

3.2 Example: logistic regression under random design

This subsection summarizes our main results in the context of the logistic regression model. Let $\mathbf{Y} = (Y_i)_{i \in [N]} \in \mathbb{R}^N$ be the response vector and $\mathbf{X} = (X_{ij})_{i \in [N], j \in [p]} \in \mathbb{R}^{N \times p}$ be the design matrix.

To illustrate that our theory applies to logistic regression with a “well-posed” random design, we consider the random matrix setup where each entry of the design matrix \mathbf{X} is an i.i.d. standard normal, i.e., $X_{ij} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$. For simplicity, we take the covariance matrix to be the identity matrix \mathbf{I}_p ; the analysis can be easily extended to a general covariance Σ satisfying

$$C^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C$$

for some constant $C > 0$. With slight abuse of notation, hereafter, let \mathbb{P} and \mathbb{E} denote the joint probability measure and expectation corresponding to (\mathbf{X}, \mathbf{Y}) , respectively. For a detailed description, see Appendix G.1.

To prove the online BvM theorem for the logistic regression model, we impose the following conditions.

(EX) The true parameter θ_0 and the initial prior parameters μ_0 and Ω_0 satisfy

$$\|\theta_0\|_2 \leq K_1, \quad \left\| \Omega_0^{1/2}(\theta_0 - \mu_0) \right\|_2 \leq K_2 p_*^{1/2}, \quad \|\Omega_0\|_2 \leq K_3 p_*,$$

where $p_* = p \vee \log n \vee \log T$ and $K_1, K_2, K_3 > 0$ are some universal constants. Furthermore, for a large enough constant $C = C(K_1, K_2, K_3) > 0$,

$$n \geq C \left[\left(p \log^6(T \vee n) \log^6(2n/p) \right) \vee \left(p^2 \log^4 T \right) \right].$$

We now state that the online BvM theorem holds for the logistic regression model.

Proposition 1 *Suppose that (EX) holds. Then, with \mathbb{P} -probability at least $1 - 5n^{-1} - 10e^{-n/72} - 4(Np)^{-1}$, the following inequality holds uniformly for all $t \in [T]$:*

$$d_V\left(\Pi_t, \Pi(\cdot \mid \mathbf{D}_{1:t})\right) \leq C \left(\frac{p_*^3}{n} \right)^{1/2},$$

where $C = C(K_1)$.

Technical statements and proofs are deferred to Appendix G; see Proposition 31 for the precise statements.

4. Variational approximation

In this section, we demonstrate that the posterior distribution $\tilde{\Pi}_t(\cdot \mid \mathbf{D}_t)$ is well approximated by its variational approximation Π_t . Before establishing the precise relation between Π_t and $\tilde{\Pi}_t(\cdot \mid \mathbf{D}_t)$, we first show that the posterior can be accurately approximated by its Laplace approximation, defined as

$$\Pi_t^{\text{LA}} = \mathcal{N}\left(\hat{\theta}_t, \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1}\right), \quad \forall t \in [T],$$

which admits a density function $\pi_t^{\text{LA}}(\cdot)$. Although our primary focus is on Π_t rather than Π_t^{LA} , understanding the concentration behavior of the pair $(\hat{\theta}_t, \tilde{\mathbf{F}}_{t, \hat{\theta}_t})$ allows us to facilitate the theoretical analysis needed to establish $\Pi_T \approx \Pi(\cdot \mid \mathbf{D})$. We will revisit this issue in Section 8.

Our main results and the corresponding proofs in this section are largely inspired by those in Spokoiny (2023). Notably, for the total variation (TV) metric, we refine the dimension dependency of Theorem 2.4 in Spokoiny (2023) by leveraging advanced analysis in Spokoiny (2024). Moreover, for the KL divergence, we significantly improve the convergence rate presented in Theorem 2.6 of Spokoiny (2023); see the discussion following Theorem 3. Since the variational posterior Π_t minimizes the KL divergence, obtaining a sharp upper bound for the KL divergence between Π_t^{LA} and $\tilde{\Pi}_t(\cdot \mid \mathbf{D}_t)$ is a crucial step in bounding $K(\Pi_t; \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t))$. A detailed proof of this section is provided in Appendix B, where we present self-contained non-asymptotic proofs adapted to the online learning framework.

As discussed in Section 2.3, to accurately approximate $\tilde{\Pi}_t(\cdot \mid \mathbf{D}_t)$, we adopt self-concordance-type smoothness conditions. To this end, we impose the following assumption:

- (A0) For every $t \in [T]$, assume that the map $\theta \mapsto L_t(\theta)$ is concave on Θ and at least four times continuously differentiable with probability 1.

For the remainder of this paper, every event is treated as the intersection with the event under which assumption (A0) holds, without further explicit restatement.

Let

$$\begin{aligned} r_{\text{LA}} &= 2\sqrt{p} + \sqrt{2 \log N} \\ \hat{\tau}_{3,t} &= \inf \left\{ \tau \in \mathbb{R}_+ : \sup_{u \in \Theta(\tilde{\mathbf{F}}_{t, \hat{\theta}_t}, 4r_{\text{LA}})} \sup_{z \in \mathbb{R}^p} \frac{|\langle \nabla^3 \tilde{L}_t(\hat{\theta}_t + u), z^{\otimes 3} \rangle|}{\left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} z \right\|_2^3} \leq \tau \right\}, \\ \hat{\tau}_{4,t} &= \inf \left\{ \tau \in \mathbb{R}_+ : \sup_{u \in \Theta(\tilde{\mathbf{F}}_{t, \hat{\theta}_t}, 4r_{\text{LA}})} \sup_{z \in \mathbb{R}^p} \frac{|\langle \nabla^4 \tilde{L}_t(\hat{\theta}_t + u), z^{\otimes 4} \rangle|}{\left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} z \right\|_2^4} \leq \tau \right\}. \end{aligned} \quad (19)$$

Then, with probability 1, $\tilde{L}_t(\cdot)$ satisfies the third and fourth order smoothness at $\hat{\theta}_t$ with parameters $(\hat{\tau}_{3,t}, \tilde{\mathbf{F}}_{t, \hat{\theta}_t}, 4r_{\text{LA}})$ and $(\hat{\tau}_{4,t}, \tilde{\mathbf{F}}_{t, \hat{\theta}_t}, 4r_{\text{LA}})$, respectively. Note that $\hat{\tau}_{3,t}$ and $\hat{\tau}_{4,t}$ are finite almost surely under (A0) because \tilde{L}_t is four times continuously differentiable and $\tilde{\mathbf{F}}_{t, \hat{\theta}_t} \in \mathcal{S}_{++}^p$.

For two probability distributions P and Q , the total variation (TV) distance is defined as

$$d_V(P, Q) = \sup_{\mathcal{A}} |P(\mathcal{A}) - Q(\mathcal{A})|, \quad (20)$$

where the supremum is taken over all measurable sets \mathcal{A} . The following theorem provides an upper bound on the total variation between Π_t^{LA} and $\tilde{\Pi}_t(\cdot \mid \mathbf{D}_t)$ in terms of $\hat{\tau}_{3,t}$ and $\hat{\tau}_{4,t}$. We also note that Theorem 2 can be obtained from Theorem 2.2 in Katsevich (2025). Although the form of (2.1) in that paper differs slightly from (19), the leading asymptotic orders are the same, and the proof techniques are similar.

Theorem 2 (Laplace approximation: TV distance) *Suppose that (A0) holds. Then, with probability 1, we have*

$$d_V\left(\Pi_t^{\text{LA}}(\cdot), \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t)\right) \leq K \epsilon_{n,t,\text{TV}}, \quad \forall t \in [T], \quad (21)$$

where

$$\epsilon_{n,t,\text{TV}} = (\hat{\tau}_{4,t} + \hat{\tau}_{3,t}^2) p^2 + \hat{\tau}_{3,t} p + \hat{\tau}_{3,t}^3 \log^3 N + e^{-8 \log N - 8p}$$

and $K > 0$ is a universal constant.

Theorem 2 implies that an accurate approximation in total variation can be achieved provided that $(\hat{\tau}_{3,t} r_{\text{LA}}^2) \vee (\hat{\tau}_{4,t} p^2)$ is sufficiently small. In particular, this condition guarantees that the quadratic approximation is accurate on the local region $\Theta(\hat{\theta}_t, \mathbf{F}_{t,\hat{\theta}_t}, 4r_{\text{LA}})$. In Proposition 6, we provide sufficient conditions under which the following asymptotic bounds hold on a given event:

$$\hat{\tau}_{3,t} = O(t^{-3/2} n^{-1/2}), \quad \hat{\tau}_{4,t} = O(t^{-2} n^{-1}). \quad (22)$$

Based on these bounds, $nt^2 \gg (p + \log N)^2$ is sufficient to ensure $(\hat{\tau}_{3,t} r_{\text{LA}}^2) \vee (\hat{\tau}_{4,t} p^2) = o(1)$. Furthermore, combining (22) with conditions $nt^3 \gg p^{-1} \log^3 N$ and $nt \gg p^2$, one can simplify the leading order of $\epsilon_{n,t,\text{TV}}$ as

$$\epsilon_{n,t,\text{TV}} = O(\hat{\tau}_{3,t} p) = O\left(t^{-3/2} (p^2/n)^{1/2}\right) \quad (23)$$

For the batch learning setup ($t = 1$), this bound coincides with the sharp bounds established in recent studies (Katsevich, 2024; Katsevich and Rigollet, 2024; Spokoiny, 2024). In particular, Katsevich (2023, Section 2.4) showed that the total variation error is lower bounded by a constant multiple of $(p^2/n)^{1/2}$. Therefore, the rate derived in Theorem 2 is optimal, at least in terms of its dependence on the sample size and parameter dimension. In this sense, for a fixed sample size n , the order $p = o(n^{1/2})$ is regarded as the *critical dimension* for the validity of the Laplace approximation.

Note that the critical dimension of order $p = o(n^{1/2})$ established in the above literature is derived for a certain class of models, including logistic regression. However, it is important to emphasize that the critical dimension for the Laplace approximation—or more broadly, the asymptotic normality of the posterior distribution (i.e., the BvM assertion)—depends on

the specific statistical model under consideration. For instance, Panov and Spokoiny (2015, Section 4.1) demonstrated that the condition $p = o(n^{1/3})$ cannot be relaxed for the BvM assertion in a particular model. Similarly, Chae (2023) established the asymptotic normality of the posterior distribution in the current status model under the assumption $p = o(n^{1/3})$. Although there is no formal proof that $p = o(n^{1/3})$ is the critical dimension in this setting, the bound appears to be tight based on the corresponding frequentist theory (Tang et al., 2012) for the same model. On the other hand, Yano and Kato (2020) showed that asymptotic normality of the posterior holds under the condition $p^2 \log^3 n = o(n)$ in linear regression with an unknown error distribution. These examples illustrate that the critical dimension for the Laplace approximation generally depends on the underlying statistical model.

To approximate the posterior in terms of the KL divergence, an additional step is required to carefully control the behavior of the posterior density in the tail region. In particular, the log-likelihood ratio $\theta \mapsto \log(\pi_t^{\text{LA}}(\theta)/\tilde{\pi}_t(\theta \mid \mathbf{D}_t))$ on $\Theta^c(\hat{\theta}_t, \tilde{\mathbf{F}}_{t, \hat{\theta}_t}, 4r_{\text{LA}})$ should not grow too rapidly. To quantify this growth rate, we introduce the following quantity:

$$\hat{\tau}_{3,t,r} = \inf \left\{ \tau \in \mathbb{R}_+ : \sup_{u \in \Theta(\tilde{\mathbf{F}}_{t, \hat{\theta}_t}, r)} \sup_{z \in \mathbb{R}^p} \frac{|\langle \nabla^3 \tilde{L}_t(\hat{\theta}_t + u), z^{\otimes 3} \rangle|}{\|\tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} z\|_2^3} \leq \tau \right\}, \quad r > 4r_{\text{LA}}. \quad (24)$$

This quantity plays a key role in bounding the KL divergence $K(\Pi_t^{\text{LA}}(\cdot); \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t))$. To ensure that this divergence remains sufficiently small, we impose the following regularity condition.

(KL) Suppose that $N \geq 2$. Assume also that, on an event \mathcal{E}_1 the following inequalities hold uniformly for all $t \in [T]$:

$$\begin{aligned} (\hat{\tau}_{3,t} r_{\text{LA}}^2) \vee (\hat{\tau}_{4,t} p^2) &\leq \frac{1}{8}, \\ \hat{\tau}_{3,t,r} &\leq N e^{8p} \exp \left(\left[\sqrt{p} + \sqrt{2 \log N} - 3 \right] r \right), \quad \forall r > 4r_{\text{LA}}. \end{aligned} \quad (25)$$

Assumption **(KL)** is very mild. In the proof of Proposition 6, we show that (25) holds under mild regularity conditions. Here, we provide a simple sufficient condition for (25). According to the discussion following (22), the condition $nt^2 \gg (p + \log N)^2$ is sufficient to ensure the first part of (25). Furthermore, by Lemma 42, the second condition in (25) holds when

$$\sup_{\theta \in \Theta(\hat{\theta}_t, \tilde{\mathbf{F}}_{t, \hat{\theta}_t}, r)} \|\nabla^3 L_t(\theta)\|_{\text{op}} \leq \lambda_{\min}^{3/2}(\tilde{\mathbf{F}}_{t, \hat{\theta}_t}) N e^{8p} \exp \left(\left[\sqrt{p} + \sqrt{2 \log N} - 3 \right] r \right).$$

Note that the last display is rather mild and holds in many examples. For instance, under the logistic regression with a simple random design setup, one can prove that

$$\lambda_{\min}(\tilde{\mathbf{F}}_{t, \hat{\theta}_t}) \asymp nt, \quad \max_{t \in [T]} \sup_{\theta \in \Theta} \|\nabla^k L_t(\theta)\|_{\text{op}} \lesssim n, \quad \forall k \in \{3, 4\}$$

provided that $p^2 \log^{12}(T \vee n) = o(n)$ (see Propositions 36 and 39 for the precise statements). Combining the last two displays, one easily checks that $n \gg p^2 \log^{12}(T \vee n)$ is sufficient for (25).

Theorem 3 (Laplace approximation: KL divergence) *Suppose that (A0) and (KL) hold on some event \mathcal{E}_1 . Then, on \mathcal{E}_1 , the following inequality holds uniformly for all $t \in [T]$:*

$$K \left(\Pi_t^{\text{LA}}(\cdot); \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right) \leq K \epsilon_{n,t,\text{KL}}^2,$$

where $K > 0$ is a universal constant and

$$\epsilon_{n,t,\text{KL}} = \left([\hat{\tau}_{4,t} + \hat{\tau}_{3,t}^2] p^2 + \hat{\tau}_{3,t}^3 \log^3 N + e^{-7 \log N} \right)^{1/2}.$$

According to the bounds established in (22), the leading order term of $\epsilon_{n,t,\text{KL}}$ satisfies

$$\epsilon_{n,t,\text{KL}}^2 \lesssim (\hat{\tau}_{4,t} p^2) \vee (\hat{\tau}_{3,t}^3 \log^3 N) \lesssim \left(\frac{p^2}{nt^2} \right) \vee \left(\frac{\log^2 N}{nt^3} \right)^{3/2},$$

which simplifies to p^2/n when $n \gg p^{-4} \log^6 N$ and $t = 1$. As a comparable result for $t = 1$, Spokoiny (2023) demonstrated, under certain conditions, that $K(\Pi_t^{\text{LA}}(\cdot); \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t))$ is bounded by $(p_{\text{eff}}^3/n)^{1/2}$ up to a constant factor. Here, p_{eff} denotes an effective dimension satisfying $p_{\text{eff}} \leq p$, and we have $p_{\text{eff}} \asymp p$ unless $\lambda_{\min}(\mathbf{\Omega}_0) \gg 1$. In this sense, our result represents a substantial improvement over existing ones because $p^2/n \ll (p_{\text{eff}}^3/n)^{1/2}$ provided that $n \gg p$ and $p_{\text{eff}} \asymp p$.

Extending the results in Theorem 3 to the VB case is straightforward. By the definition of Π_t specified in (2), we have

$$K \left(\Pi_t(\cdot); \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right) \leq K \left(\Pi_t^{\text{LA}}(\cdot); \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right) \lesssim \epsilon_{n,t,\text{KL}}^2,$$

which, by Pinsker's inequality, further implies that

$$d_V \left(\Pi_t(\cdot), \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right) \leq \sqrt{\frac{1}{2} K \left(\Pi_t^{\text{LA}}(\cdot); \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right)} \lesssim \epsilon_{n,t,\text{KL}}.$$

For a simplified comparison, if we take $\epsilon_{n,1,\text{KL}} = O(\sqrt{p^2/n})$, our result for the TV metric aligns with the rate demonstrated in Katsevich and Rigollet (2024, Corollary 2.1).

Theorem 4 (Variational approximation) *Suppose that (A0) and (KL) hold on some event \mathcal{E}_1 . Then, on \mathcal{E}_1 , the following inequalities hold uniformly for all $t \in [T]$:*

$$K \left(\Pi_t(\cdot); \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right) \leq K \epsilon_{n,t,\text{KL}}^2, \quad d_V \left(\Pi_t(\cdot), \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right) \leq K \epsilon_{n,t,\text{KL}},$$

where $K > 0$ is a universal constant.

In the online setting, the sequential updates quickly yield a highly *informative prior*. Specifically, we typically have $\lambda_{\min}(\mathbf{\Omega}_t) \gg 1$ and $\|\mu_t - \theta_0\|_2 \ll 1$ for $t > 1$, and both quantities naturally scale with the sample size $N_t = nt$. To quantify the strength of such priors, we adopt the notion of *effective sample size*, originally introduced by Spokoiny (2017), defined as $n_{\text{eff},t} = \lambda_{\min}(-\nabla^2 \tilde{L}_t(\theta_t^*))$. A strong prior tends to inflate the effective sample size. We will show in Section 6 that $n_{\text{eff},t} \asymp N_t$ holds under suitable conditions.

A subtle but important phenomenon emerges in the rate of $\epsilon_{n,t,\text{TV}}$. One might initially expect the TV error based on N_t observations to scale as $(p^2/N_t)^{1/2}$, matching the order $(p^2/n)^{1/2}$ at $t = 1$. In contrast, $\epsilon_{n,t,\text{TV}} = O(t^{-1}[p^2/N_t]^{1/2})$ in (23) is strictly sharper in terms of its polynomial dependence on t . Notably, this improvement arises from the use of informative normal priors.

To provide more intuition behind this improvement, note that, by (13),

$$\widehat{\tau}_{k,t} \leq \lambda_{\min}^{-k/2}(\widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}) \sup_{\theta \in \Theta(\widehat{\theta}_t, \widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}, 4r_{\text{LA}})} \|\nabla^k \widetilde{L}_t(\theta)\|_{\text{op}}, \quad k \in \{3, 4\}.$$

For a simple illustration, we assume that

$$\lambda_{\min}(\widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}) \asymp nt \quad \text{and} \quad \sup_{\theta \in \Theta(\widehat{\theta}_t, \widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}, 4r_{\text{LA}})} \|\nabla^k L_t(\theta)\|_{\text{op}} \asymp n.$$

Then, the asymptotic bound (22) holds, which leads to the improved result (23). Note that the normal prior affects $\widetilde{L}_t(\cdot)$ only through (up to) second-order terms, whereas the key regularity of our model is governed by the ratio between the third (or fourth) and second derivatives. By using informative normal priors, we can effectively increase the effective sample size (as $\lambda_{\min}(\widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}) \asymp nt$) without altering the higher-order properties of the log-likelihood, since $\nabla^k \widetilde{L}_t = \nabla^k L_t$ for $k > 2$. In this sense, the normal prior effectively leverages the available prior information.

5. Penalized M-estimation

This section presents theoretical results for the penalized M-estimator $\widehat{\theta}_t$, which has been well-studied in the literature. Recently, Spokoiny (2012, 2017, 2024) developed advanced techniques for analyzing statistical behaviors of M-estimators with finite-sample guarantees. Accordingly, we adopt these techniques with modifications to adapt them to our framework.

We say the model is *stochastically linear* if the map $\theta \mapsto \zeta_t(\theta) = L_t(\theta) - \mathbb{E}_t L_t(\theta)$ is linear for every t ; that is, $\zeta_t(\theta) = a^\top \theta + b$ for some (random) quantities $a \in \mathbb{R}^p$ and $b \in \mathbb{R}$. Note that stochastic linearity implies that $\nabla^k L_t(\theta)$ is non-random for all $t \in [T]$ and $k \in \{2, 3, 4\}$. Also, since $\nabla \zeta_t(\theta)$ does not depend on θ , we hereafter denote this random vector by $\nabla \zeta_t$. The stochastically linear framework encompasses many important statistical models, such as the logistic regression, Poisson regression, nonparametric regression (Spokoiny and Panov, 2025), nonlinear inverse problem (Spokoiny, 2019), and covariance estimation (Puchkin et al., 2025). See Section 1.3 in Spokoiny (2024) for a detailed discussion.

We now introduce the assumptions and notations used in the estimation procedure.

- (A1) The model is stochastically linear. Also, $\mathbf{F}_{t,\theta}$ is nonsingular for every $t \in [T]$ and $\theta \in \Theta(\theta_0, \mathbf{I}_p, 1/2)$. Furthermore, there exist $\{\mathbf{V}_t : t \in [T]\} \subset \mathcal{S}_{++}^p$ and $M_n > 1$ such that

$$\mathbb{P}_{0,t} \left(\left\| \widetilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \zeta_t \right\|_2 \geq r_{\text{eff},t} \right) \leq e^{-(\log n + \log T)}, \quad (26)$$

$$\max_{t \in [T]} \sup_{\theta \in \Theta(\theta_0, \mathbf{I}_p, 1/2)} \left\| \mathbf{F}_{t,\theta}^{-1} \mathbf{V}_t \right\|_2 \leq \frac{M_n^2}{9}, \quad (27)$$

where

$$\begin{aligned} r_{\text{eff},t} &= p_{\text{eff},t}^{1/2} + \sqrt{2\lambda_t(\log n + \log T)}, \\ p_{\text{eff},t} &= \text{tr} \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \mathbf{V}_t \right), \quad \lambda_t = \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \mathbf{V}_t \right\|_2. \end{aligned} \quad (28)$$

As discussed in Section 2.3, we employ the self-concordance condition on $\mathbb{E}_t \tilde{L}_t(\cdot)$ over the local set $\Theta(\theta_t^*, \tilde{\mathbf{F}}_{t,\theta_t^*}, 4r_{\text{eff},t})$. For this purpose, we define the following quantity:

$$\tau_{3,t}^* = \inf \left\{ \tau_3 \in \mathbb{R}_+ : \sup_{u \in \Theta(\tilde{\mathbf{F}}_{t,\theta_t^*}, 4r_{\text{eff},t})} \sup_{z \in \mathbb{R}^p} \frac{|\langle \nabla^3 \mathbb{E}_t \tilde{L}_t(\theta_t^* + u), z^{\otimes 3} \rangle|}{\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} z \right\|_2^3} \leq \tau_3 \right\}. \quad (29)$$

Note that $\tau_{3,t}^*$ is random because both θ_t^* and $\mathbf{\Omega}_{t-1}$ depend on $(\mathbf{D}_1, \dots, \mathbf{D}_{t-1})$. To prove the convergence of the pMLE $\hat{\theta}_t$, we further impose the following assumption.

(Est) On an event \mathcal{E}_2 , the following inequality holds uniformly for all $t \in [T]$:

$$\tau_{3,t}^* r_{\text{eff},t} \leq 1/16. \quad (30)$$

Assumptions **(A1)** and **(Est)** ensure that the pMLE $\hat{\theta}_t$ converges to θ_t^* at an appropriate rate. More specifically, Theorem 5 shows that

$$\hat{\theta}_t \in \Theta(\theta_t^*, \tilde{\mathbf{F}}_{t,\theta_t^*}, 4r_{\text{eff},t}), \quad \forall t \in [T]$$

with high probability. Before stating the theorem, we provide a detailed discussion of the above assumptions.

In many examples, one can choose $\mathbf{V}_t = c \text{Var}(\nabla \zeta_t)$ for a sufficiently large constant $c > 0$. Other choices are also possible. For instance, if $\nabla \zeta_t \sim \text{subG}(\sigma_n^2)$ for some $\sigma_n^2 \geq 0$, then (26) holds with the choice $\mathbf{V}_t = \sigma_n^2 \mathbf{I}_p$. In this case, since $\nabla \zeta_t$ is the sum of n independent random variables, we roughly have $\sigma_n^2 \asymp n$. Note that $\nabla \zeta_t$ can be sub-exponential rather than sub-Gaussian in certain important examples, such as the Poisson regression model. Even in sub-exponential cases, however, one can still verify (26) under suitable conditions; see Appendix A in Spokoiny (2017) and Lemma B.2 in Lee et al. (2025).

The condition (27) is used to derive an appropriate bound for the radius $r_{\text{eff},t}$. In the batch learning setting, one may choose $r_{\text{eff},1} \asymp \sqrt{p} + \sqrt{\log n}$ so that $\hat{\theta}_1 \in \Theta(\theta_1^*, \tilde{\mathbf{F}}_{1,\theta_1^*}, 4r_{\text{eff},1})$ with high probability. In contrast, in the online learning setting, the order of $r_{\text{eff},t}$ exhibits an interesting behavior. In particular, note that the effective dimension $p_{\text{eff},t}$ decreases with t because $\tilde{\mathbf{F}}_{t,\theta_t^*}$ and \mathbf{V}_t are of orders proportional to the sample sizes nt and n , respectively. By combining (27) with additional conditions, one can derive the following bounds:

$$p_{\text{eff},t} = O(M_n^2 t^{-1} p), \quad r_{\text{eff},t} = O(M_n \sqrt{(p \vee \log n \vee \log T)/t}). \quad (31)$$

The results in (31) follow directly from Lemma 18 and Proposition 6.

The assumptions in **(A1)** are mild and hold in many examples (e.g., when the score function is sub-Gaussian). Statistical models with the sub-Gaussian score function encompass a

wide range of examples; a logistic regression model is one of the popular cases. In Appendix G, we present a theoretical verification of **(A1)** under a simple random design setting. For the logistic regression model, one can show (26) and (27) are satisfied with $\mathbf{V}_t = \mathbf{X}_t^\top \mathbf{X}_t/4$ and $M_n = O_P(1)$, where $\mathbf{X}_t \in \mathbb{R}^{n \times p}$ is the design matrix for the t -th mini-batch.

The nonsingularity of $\mathbf{F}_{t,\theta}$ on $\Theta(\theta_0, \mathbf{I}, 1/2)$ can be verified when $L_t(\cdot)$ is strictly concave; this includes, for example, the logistic regression model. The specific choice of radius $1/2$ in (27) does not carry inherent meaning and can be replaced by a decaying sequence r_n , with some additional technical effort.

Assumption (30) ensures that the effective sample size $n_{\text{eff},t}$ is sufficiently large relative to the effective dimension $p_{\text{eff},t}$, thereby guaranteeing the convergence of $\widehat{\theta}_t$ to θ_t^* . One can show that $\tau_{3,t}^*$ is of the order given in (22). Combining this with (31), we find that for a fixed $t \in [T]$, a sufficient condition for (30) is

$$M_n^2(p \vee \log T) = o(nt^4).$$

In the batch learning setup where $T = t = 1$, this condition reduces to the well-known requirement $p = o(n)$, provided that $M_n = O(1)$.

Note that the additional term $\log T$ in the definition of $r_{\text{eff},t}$ can be interpreted as the cost of requiring uniformity over $\{1, 2, \dots, T\}$. From (26), we have

$$\begin{aligned} \mathbb{P}_0 \left(\left\| \widetilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \zeta_t \right\|_2 \geq r_{\text{eff},t} \text{ for some } t \in [T] \right) &\leq T \cdot \max_{t \in [T]} \mathbb{P}_{0,t} \left(\left\| \widetilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \zeta_t \right\|_2 \geq r_{\text{eff},t} \right) \\ &\leq e^{-\log n} = n^{-1}. \end{aligned}$$

To state the main theorem in this section, let $\mathcal{E}_{\text{est},1}$ be the event on which the following inequality holds:

$$\left\| \widetilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \zeta_t \right\|_2 \leq r_{\text{eff},t} \quad \text{for all } t \in [T]. \quad (32)$$

Then, $\mathbb{P}_0^{(N)}(\mathcal{E}_{\text{est},1}) \geq 1 - n^{-1}$ under **(A1)**. The event $\mathcal{E}_{\text{est},1}$ plays a crucial role throughout this paper, as it ensures the statistical validity of the online estimation procedure.

Theorem 5 *Suppose that **(A0)**, **(A1)** and **(Est)** hold. Then, on $\mathcal{E}_{\text{est},1} \cap \mathcal{E}_2$, the following inequality holds uniformly for all $t \in [T]$:*

$$\left\| \widetilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\widehat{\theta}_t - \theta_t^*) \right\|_2 \leq 4r_{\text{eff},t}, \quad \forall t \in [T].$$

Theorem 5 provides a sharp upper bound for $\|\widehat{\theta}_t - \theta_t^*\|_2$, which corresponds to the variance term in the well-known bias–variance decomposition:

$$\|\widehat{\theta}_t - \theta_0\|_2 \leq \|\widehat{\theta}_t - \theta_t^*\|_2 + \|\theta_t^* - \theta_0\|_2. \quad (33)$$

By (31) and Theorem 5, we obtain

$$\|\widehat{\theta}_t - \theta_t^*\|_2 = O \left(M_n [p \vee \log n \vee \log T]^{1/2} t^{-1} n^{-1/2} \right)$$

provided that $\lambda_{\min}(\tilde{\mathbf{F}}_{t,\theta_t^*}) \gtrsim nt$. Ignoring the logarithmic factor, the quantity $\|\hat{\theta}_t - \theta_t^*\|_2$ decreases at the rate $t^{-1}n^{-1/2}$, rather than the standard rate $(nt)^{-1/2}$. As discussed at the end of Section 4, this improvement arises from the use of informative quadratic penalization.

Note that Theorem 5 does not address $\|\theta_t^* - \theta_0\|_2$, which corresponds to the bias term in the decomposition (33). In Sections 6 and 8, we show that $\|\theta_t^* - \theta_0\|_2$ decreases at the rate $(nt)^{-1/2}$. Given that this convergence rate cannot be improved in general, our result may be seen as a natural consequence.

6. Analysis of eigenvalues and remainder terms

This section presents informative bounds on important quantities such as $\hat{\tau}_{3,t}$, $\hat{\tau}_{4,t}$, and $\tau_{3,t}^*$. A crucial step in analyzing these quantities is to ensure that $\hat{\theta}_t$ and θ_t^* lie within a local neighborhood of θ_0 , say $\Theta(\theta_0, \mathbf{I}_p, 1/2)$. This localization guarantees an accurate quadratic approximation. Specifically, the quality of the quadratic approximation relies heavily on the magnitude of the smallest eigenvalues, such as $\lambda_{\min}(\tilde{\mathbf{F}}_{t,\theta_t^*})$ and $\lambda_{\min}(\tilde{\mathbf{F}}_{t,\hat{\theta}_t})$. Intuitively, these values serve as the *effective sample size*. Therefore, we need to show that the effective sample size is proportional to the accumulated (actual) sample size nt . Once this step is established, it is not too difficult to bound the other quantities (e.g., $\hat{\tau}_{3,t}$, $\hat{\tau}_{4,t}$ and $\tau_{3,t}^*$).

However, unless the log-likelihood $L_t(\cdot)$ is strongly concave in the sense that $\lambda_{\min}(\mathbf{F}_{t,\theta})$ is uniformly bounded below by a multiple of n for all $\theta \in \Theta$, it is not straightforward to ensure that $\lambda_{\min}(\tilde{\mathbf{F}}_{t,\theta_t^*}) \wedge \lambda_{\min}(\tilde{\mathbf{F}}_{t,\hat{\theta}_t}) \gtrsim nt$. If we assume the strong concavity of $L_t(\cdot)$, the analysis can be significantly simplified because the *localization* step (e.g., $\theta_t^* \in \Theta(\theta_0, \mathbf{I}_p, 1/2)$) can be omitted, and it is possible to obtain $\lambda_{\min}(\tilde{\mathbf{F}}_{t,\theta_t^*}) \gtrsim nt$ directly. For ease of analysis, this strong concavity assumption has been often adopted in the existing online learning literature (Chen et al., 2020; Zhu et al., 2023). However, many important statistical models (e.g., GLMs) may not exhibit strong concavity when $\Theta = \mathbb{R}^p$. For instance, the logistic regression model—which is our main statistical application—is not strongly concave but strictly concave. To accommodate these models, we impose some regularity conditions in a local vicinity of θ_0 . The precise statements of the local regularity conditions are as follows:

(A2) There exist constants $K_{\min} > 0$ and $K_{\max} \geq 1$ such that

$$\begin{aligned} \min_{t \in [T]} \inf_{\theta \in \Theta(\theta_0, \mathbf{I}_p, 1/2)} \lambda_{\min}(\mathbf{F}_{t,\theta}) &\geq K_{\min}n, \\ \max_{t \in [T]} \sup_{\theta \in \Theta(\theta_0, \mathbf{I}_p, 1/2)} \lambda_{\max}(\mathbf{F}_{t,\theta}) &\leq K_{\max}n, \end{aligned} \quad (34)$$

$$\begin{aligned} \max_{t \in [T]} \sup_{\theta \in \Theta(\theta_0, \mathbf{I}_p, 1/2)} \|\nabla^3 L_t(\theta)\|_{\text{op}} &\leq K_{\max}n, \\ \max_{t \in [T]} \sup_{\theta \in \Theta(\theta_0, \mathbf{I}_p, 1/2)} \|\nabla^4 L_t(\theta)\|_{\text{op}} &\leq K_{\max}n, \end{aligned} \quad (35)$$

$$\max_{t \in [T]} \sup_{\theta \in \Theta(\theta_0, \mathbf{I}_p, 1/2+r)} \|\nabla^3 L_t(\theta)\|_{\text{op}} \leq K_{\max}N e^{8p} e^{(\sqrt{p} + \sqrt{2 \log N} - 3)r}, \quad \forall r > 0. \quad (36)$$

In Appendix G, we show that assumption **(A2)** holds with high probability under the logistic regression model with random design. For precise statements, see Propositions 36 and 39.

For high-dimensional or nonparametric models, the effect of the prior may remain non-negligible even as the sample size increases (Cox, 1993; Freedman, 1999; Johnstone, 2010). Since we allow the dimension to diverge, i.e., $p = p_n \rightarrow \infty$ as $n \rightarrow \infty$, we impose non-asymptotic conditions on the initial prior parameters μ_0 and $\mathbf{\Omega}_0$. Before we state specific conditions, we recall the following notation:

$$p_* = p \vee \log n \vee \log T.$$

(P) The initial prior parameters μ_0 and $\mathbf{\Omega}_0$ satisfy

$$\left\| \mathbf{\Omega}_0^{1/2}(\theta_0 - \mu_0) \right\|_2 \leq \delta n^{1/2}, \quad \|\mathbf{\Omega}_0\|_2 \leq K_{\max} p_*$$

for a small enough constant $\delta = \delta(K_{\min}, K_{\max}) > 0$.

(S) For a large enough constant $C = C(K_{\min}, K_{\max}) > 0$,

$$n \geq C(\log^2 T \vee M_n^2) p_*^2.$$

We now present the main results of this section.

Proposition 6 *Suppose that **(A0)**, **(A1)**, **(A2)**, **(S)**, and **(P)** hold. Then, on $\mathcal{E}_{\text{est},1}$, the following inequalities hold uniformly for all $t \in [T]$:*

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2}(\theta_0 - \theta_t^*) \right\|_2 &\leq K_{\text{up}} M_n \sqrt{t p_*}, & \lambda_{\min}(\tilde{\mathbf{F}}_{t,\theta_t^*}) \wedge \lambda_{\min}(\tilde{\mathbf{F}}_{t,\hat{\theta}_t}) &\geq K_{\text{low}} n t, \\ \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2}(\hat{\theta}_t - \theta_t^*) \right\|_2 &\leq K_{\text{up}} M_n \sqrt{t^{-1} p_*}, & \lambda_{\max}(\tilde{\mathbf{F}}_{t,\theta_t^*}) \vee \lambda_{\max}(\tilde{\mathbf{F}}_{t,\hat{\theta}_t}) &\leq K_{\text{up}} n t, \end{aligned}$$

and

$$\hat{\tau}_{3,t} \vee \tau_{3,t}^* \leq K_{\text{up}} t^{-3/2} n^{-1/2}, \quad \hat{\tau}_{4,t} \leq K_{\text{up}} t^{-2} n^{-1}, \quad \epsilon_{n,t,\text{KL}} \leq K_{\text{up}} t^{-1} n^{-1/2} p_*,$$

where $K_{\text{up}} = K_{\text{up}}(K_{\min}, K_{\max})$ and $K_{\text{low}} = K_{\text{low}}(K_{\min})$.

Proof See the proof of Proposition 23; Proposition 6 is a special case of Proposition 23. \blacksquare

From the results in Proposition 6, we can quantify the relation between pMLE $(\hat{\theta}_t, \tilde{\mathbf{F}}_{t,\hat{\theta}_t})$ and variational parameters $(\mu_t, \mathbf{\Omega}_t)$. Recall that

$$d_V\left(\Pi_t(\cdot), \Pi_t^{\text{LA}}(\cdot)\right) \leq d_V\left(\Pi_t(\cdot), \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t)\right) + d_V\left(\tilde{\Pi}_t(\cdot \mid \mathbf{D}_t), \Pi_t^{\text{LA}}(\cdot)\right) \lesssim \epsilon_{n,t,\text{KL}},$$

where the inequality holds by Theorem 4 and Pinsker's inequality. If two normal distributions $Q_1 = \mathcal{N}(m_1, \mathbf{V}_1^{-1})$ and $Q_2 = \mathcal{N}(m_2, \mathbf{V}_2^{-1})$ with $d_V(Q_1, Q_2) \leq \epsilon$ for sufficiently small $\epsilon > 0$, Arbas et al. (2023) showed in Theorem 1.8 that

$$\left\| \mathbf{V}_1^{1/2}(m_1 - m_2) \right\|_2 \vee \left\| \mathbf{V}_2^{-1/2} \mathbf{V}_1 \mathbf{V}_2^{-1/2} - \mathbf{I}_p \right\|_{\text{F}} = O(\epsilon).$$

Based on this fact, it follows from Proposition 6 that on $\mathcal{E}_{\text{est},1}$

$$\left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2}(\mu_t - \hat{\theta}_t) \right\|_2 \vee \left\| \boldsymbol{\Omega}_t^{-1/2} \tilde{\mathbf{F}}_{t,\hat{\theta}_t} \boldsymbol{\Omega}_t^{-1/2} - \mathbf{I}_p \right\|_F \lesssim \epsilon_{n,t,\text{KL}} = O(t^{-1}n^{-1/2}p_*). \quad (37)$$

Note that, due to the symmetry of the total variation distance, the above bound remains invariant under interchange of $(\hat{\theta}_t, \tilde{\mathbf{F}}_{t,\hat{\theta}_t})$ and $(\mu_t, \boldsymbol{\Omega}_t)$. Importantly, (37) plays a crucial role in establishing the online BvM assertion, namely that $d_V(\Pi_T, \Pi(\cdot | \mathbf{D})) \rightarrow 0$ in probability.

We conclude this section by discussing the bias term $\|\theta_t^* - \theta_0\|_2$. By Proposition 6, we have

$$\|\theta_t^* - \theta_0\|_2 \lesssim \lambda_{\min}^{-1/2}(\tilde{\mathbf{F}}_{t,\theta_t^*}) M_n \sqrt{tp_*} \lesssim \left(\frac{p_*}{n}\right)^{1/2}. \quad (38)$$

Although this bound is sufficient to ensure that $\theta_t^* \in \Theta(\theta_0, \mathbf{I}_p, 1/2)$, the corresponding rate is not sharp enough to guarantee that $\|\hat{\theta}_t - \theta_0\|_2$ decreases at the optimal rate $(nt)^{-1/2}$ under the standard decomposition analysis in (33). Instead, we will obtain an appropriate rate of $\|\hat{\theta}_t - \theta_0\|_2$ via a batch learning estimator. We revisit this issue in Proposition 9.

7. Bernstein–von Mises theorem for full posterior

This section presents the BvM theorem of full posterior under the batch learning setting. Recall the definition of *full posterior* in (14). Also, we introduce some notations for analyzing t -th full posterior $\Pi(\cdot | \mathbf{D}_{1:t})$. For $t \in [T]$ and $\theta \in \Theta$, let

$$\begin{aligned} \tilde{L}_{1:t}(\theta) &= -\frac{1}{2} \left\| \boldsymbol{\Omega}_0^{1/2}(\theta - \mu_0) \right\|_2^2 + L_{1:t}(\theta), & \mathbf{F}_{1:t,\theta} &= -\nabla^2 L_{1:t}(\theta) \\ \zeta_{1:t}(\theta) &= L_{1:t}(\theta) - \mathbb{E}L_{1:t}(\theta), & \tilde{\mathbf{F}}_{1:t,\theta} &= \boldsymbol{\Omega}_0 + \mathbf{F}_{1:t,\theta}. \end{aligned} \quad (39)$$

and

$$\hat{\theta}_{1:t} = \operatorname{argmax}_{\theta \in \Theta} \tilde{L}_{1:t}(\theta), \quad \hat{\theta}_{1:t}^{\text{ML}} = \operatorname{argmax}_{\theta \in \Theta} L_{1:t}(\theta), \quad \theta_{1:t}^* = \operatorname{argmax}_{\theta \in \Theta} \mathbb{E} \tilde{L}_{1:t}(\theta). \quad (40)$$

By applying similar approximation techniques in Section 4, one can prove that the LA would provide an accurate approximation of $\Pi(\cdot | \mathbf{D}_{1:t})$ in TV sense. Specifically, since we impose several regularity conditions on the local neighborhood of θ_0 in **(A2)**, we need to establish the concentration behaviors of $\hat{\theta}_{1:t}$ and $\hat{\theta}_{1:t}^{\text{ML}}$ for suitable localization steps (e.g., $\hat{\theta}_{1:t}, \hat{\theta}_{1:t}^{\text{ML}} \in \Theta(\theta_0, \mathbf{I}_p, 1/2)$). For this purpose, we impose the following conditions, which are a version of **(A1)** adapted to the batch learning setup.

(A1*) Assume that **(A1)** holds. Also, there exists $\{\mathbf{V}_{1:t} : t \in [T]\} \subset \mathcal{S}_{++}^p$ such that

$$\begin{aligned} \mathbb{P}_0^{(N)} \left(\left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{-1/2} \nabla \zeta_{1:t} \right\|_2 \geq \tilde{r}_{\text{eff},1:t} \text{ for some } t \in [T] \right) &\leq n^{-1}, \\ \mathbb{P}_0^{(N)} \left(\left\| \mathbf{F}_{1:t,\theta_0}^{-1/2} \nabla \zeta_{1:t} \right\|_2 \geq r_{\text{eff},1:t} \text{ for some } t \in [T] \right) &\leq n^{-1}, \end{aligned} \quad (41)$$

and

$$\max_{t \in [T]} \sup_{\theta \in \Theta(\theta_0, \mathbf{I}_{p,1/2})} \left\| \mathbf{F}_{1:t,\theta}^{-1} \mathbf{V}_{1:t} \right\|_2 \leq \frac{M_n^2}{9}, \quad (42)$$

where

$$\tilde{\lambda}_{1:t} = \left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{-1} \mathbf{V}_{1:t} \right\|_2, \quad \lambda_{1:t} = \left\| \mathbf{F}_{1:t,\theta_0}^{-1} \mathbf{V}_{1:t} \right\|_2,$$

and

$$\begin{aligned} \tilde{r}_{\text{eff},1:t} &= \tilde{p}_{\text{eff},1:t}^{1/2} + \sqrt{2\tilde{\lambda}_{1:t}(\log n + \log T)}, & \tilde{p}_{\text{eff},1:t} &= \text{tr} \left(\tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{-1} \mathbf{V}_{1:t} \right), \\ r_{\text{eff},1:t} &= p_{\text{eff},1:t}^{1/2} + \sqrt{2\lambda_{1:t}(\log n + \log T)}, & p_{\text{eff},1:t} &= \text{tr} \left(\mathbf{F}_{1:t,\theta_0}^{-1} \mathbf{V}_{1:t} \right), \end{aligned} \quad (43)$$

and M_n is the quantity specified in **(A1)**.

Let $\mathcal{E}_{\text{est},2}$ be the event on which the following inequalities hold:

$$\left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{-1/2} \nabla \zeta_{1:t} \right\|_2 \leq \tilde{r}_{\text{eff},1:t}, \quad \left\| \mathbf{F}_{1:t,\theta_0}^{-1/2} \nabla \zeta_{1:t} \right\|_2 \leq r_{\text{eff},1:t} \quad \text{for all } t \in [T]. \quad (44)$$

Then, by (41), $\mathbb{P}_0^{(N)}(\mathcal{E}_{\text{est},2}) \geq 1 - 2n^{-1}$ under **(A2)**. On $\mathcal{E}_{\text{est},2}$, we can prove that the following inequalities hold uniformly for all $t \in [T]$:

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{1/2} (\hat{\theta}_{1:t} - \theta_0) \right\|_2 &\lesssim \left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{-1/2} \boldsymbol{\Omega}_0(\theta_0 - \mu_0) \right\|_2 + M_n p_*^{1/2} \\ \left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \theta_0) \right\|_2 &\lesssim M_n p_*^{1/2}. \end{aligned} \quad (45)$$

More detailed statements for the above inequalities are deferred to Appendix E. Here, $\left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{-1/2} \boldsymbol{\Omega}_0(\theta_0 - \mu_0) \right\|_2$ denotes the bias term arising from the quadratic penalization. Based on these results, we can prove Theorem 7. Note that the following theorem is the full posterior version of Theorem 1 in Section 3; hence, Theorem 2.2 in Katsevich (2025) yields an identical result.

Theorem 7 *Suppose that **(A0)**, **(A1*)**, **(A2)**, **(S)** and **(P)** hold. Then, on $\mathcal{E}_{\text{est},2}$, the following inequality holds uniformly for all $t \in [T]$:*

$$d_V \left(\mathcal{N} \left(\hat{\theta}_{1:t}, \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{-1} \right), \Pi(\cdot \mid \mathbf{D}_{1:t}) \right) \leq K \left(\frac{p_*^2}{nt} \right)^{1/2},$$

where $K = K(K_{\min}, K_{\max})$.

Although we conclude from Theorem 7 that $\mathcal{N}(\hat{\theta}_{1:t}, \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{-1}) \approx \Pi(\cdot \mid \mathbf{D}_{1:t})$, the relationship with $\mathcal{N}(\hat{\theta}_{1:t}^{\text{ML}}, \mathbf{F}_{1:t,\theta_0}^{-1})$ remains unresolved. A standard BvM theorem states that $\mathcal{N}(\hat{\theta}_{1:t}^{\text{ML}}, \mathbf{F}_{1:t,\theta_0}^{-1}) \approx \Pi(\cdot \mid \mathbf{D}_{1:t})$ in the TV sense. Importantly, $\mathcal{N}(\hat{\theta}_{1:t}^{\text{ML}}, \mathbf{F}_{1:t,\theta_0}^{-1})$ does not exhibit the prior effects. Hence, we impose an additional assumption **(P*)** for $(\mu_0, \boldsymbol{\Omega}_0)$ so that the prior effects become asymptotically negligible.

(**P***) Assume that (**P**) holds. Also, the initial prior parameters μ_0 and Ω_0 satisfy

$$\left\| \Omega_0^{1/2} (\theta_0 - \mu_0) \right\|_2 \leq K_{\max} M_n p_*^{1/2}.$$

Theorem 8 *Suppose that (**A0**), (**A1***), (**A2**), (**S**) and (**P***) hold. Then, on $\mathcal{E}_{\text{est},2}$, the following inequalities hold uniformly for all $t \in [T]$:*

$$\begin{aligned} d_V \left(\mathcal{N} \left(\hat{\theta}_{1:t}^{\text{ML}}, \mathbf{F}_{1:t, \theta_0}^{-1} \right), \mathcal{N} \left(\hat{\theta}_{1:t}, \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1} \right) \right) &\leq K M_n \left(\frac{p_*^2}{nt} \right)^{1/2}, \\ d_V \left(\mathcal{N} \left(\hat{\theta}_{1:t}^{\text{ML}}, \mathbf{F}_{1:t, \theta_0}^{-1} \right), \Pi(\cdot \mid \mathbf{D}_{1:t}) \right) &\leq K M_n \left(\frac{p_*^2}{nt} \right)^{1/2}, \end{aligned} \quad (46)$$

where $K = K(K_{\min}, K_{\max})$.

We briefly introduce the key idea in our proof of Theorem 8. Our proof strategy is similar to that in Spokoiny (2025) and Katsevich (2025) whose key idea is based on

$$\begin{aligned} &d_V \left(\mathcal{N} \left(\hat{\theta}_{1:t}^{\text{ML}}, \mathbf{F}_{1:t, \theta_0}^{-1} \right), \Pi(\cdot \mid \mathbf{D}_{1:t}) \right) \\ &\leq d_V \left(\mathcal{N} \left(\hat{\theta}_{1:t}, \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1} \right), \Pi(\cdot \mid \mathbf{D}_{1:t}) \right) + d_V \left(\mathcal{N} \left(\hat{\theta}_{1:t}^{\text{ML}}, \mathbf{F}_{1:t, \theta_0}^{-1} \right), \mathcal{N} \left(\hat{\theta}_{1:t}, \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1} \right) \right). \end{aligned} \quad (47)$$

Recall the formula (17) in Section 3. If the initial prior assigns sufficient probability mass to a neighborhood of θ_0 as described in (**P***), it can be shown that

$$\left\| \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{1/2} \left(\hat{\theta}_{1:t}^{\text{ML}} - \hat{\theta}_{1:t} \right) \right\|_2 \vee \left\| \mathbf{F}_{1:t, \theta_0}^{-1/2} \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}} \mathbf{F}_{1:t, \theta_0}^{-1/2} - \mathbf{I}_p \right\|_{\text{F}} = O \left(M_n \left(\frac{p_*^2}{nt} \right)^{1/2} \right), \quad (48)$$

which yields the first inequality in (46). Combining this with the result in Theorem 7, one can easily check that the second inequality in (46) holds via (47).

Since we assume the *stochastic linear model* in Section 5, $\mathbf{F}_{1:t, \theta_0}$ is non-random; indeed, it coincides with the Fisher information matrix. In more general cases, however, $\mathbf{F}_{1:t, \theta_0}$ may exhibit randomness. Logistic regression with a random design serves as a representative example.

To analyze such cases, it is natural to compare $\mathbf{F}_{1:t, \theta_0}$ with its population counterpart. For simplicity, suppose that $N_t^{-1} \mathbb{E}_{\mathbf{X}}[\mathbf{F}_{1:t, \theta_0}]$ is the same for all $t \in [T]$, and define $I_{\theta_0} = N_t^{-1} \mathbb{E}_{\mathbf{X}}[\mathbf{F}_{1:t, \theta_0}]$. Here, $\mathbb{E}_{\mathbf{X}}$ denotes the expectation over the randomness of \mathbf{X} . Under suitable regularity conditions (e.g. sub-Gaussian random design), we can obtain, for any $x > 0$,

$$\left\| \frac{1}{N_t} \mathbf{F}_{1:t, \theta_0} - I_{\theta_0} \right\|_2 \leq K \sqrt{\frac{p+x}{N_t}}, \quad \left\| \frac{1}{N_t} \mathbf{F}_{1:t, \theta_0} - I_{\theta_0} \right\|_{\text{F}} \leq K \sqrt{\frac{p(p+x)}{N_t}} \quad (49)$$

with probability at least $1 - e^{-x}$, where $K > 0$ is a quantity depending on some regularity conditions (e.g. sub-Gaussian parameter of each row of \mathbf{X}). These bounds are standard; see (Tropp, 2012; Vershynin, 2018).

This discrepancy is negligible in terms of the final TV convergence rate. Specifically, combining (49) with (17), it is not difficult to see that with probability at least $1 - e^{-x}$

$$d_V \left(\mathcal{N}(\widehat{\theta}_{1:t}^{\text{ML}}, \mathbf{F}_{1:t, \theta_0}^{-1}), \mathcal{N}(\widehat{\theta}_{1:t}^{\text{ML}}, [N_t I_{\theta_0}]^{-1}) \right) \lesssim K \sqrt{\frac{p(p+x)}{N_t}}$$

provided that $\lambda_{\min}(I_{\theta_0}) \geq C$ for some constant $C > 0$.

8. Online Bernstein–von Mises theorem

In this section, we present the main result of this paper, namely the *online BvM theorem*. We elaborate on the high-level proof strategy in Section 3. For clarity, we now introduce explicit definition for \widehat{Q}_t appearing in (16). Let $Q_{\text{LA}, t} = \mathcal{N}(\widehat{\theta}_{1:t}, \widetilde{\mathbf{F}}_{1:t, \widehat{\theta}_{1:t}}^{-1})$. By (18), our claim reduces to obtaining upper bounds for the following two quantities:

$$\left\| \boldsymbol{\Omega}_t^{1/2} (\widehat{\theta}_{1:t} - \mu_t) \right\|_2 \quad \text{and} \quad \left\| \widetilde{\mathbf{F}}_{1:t, \widehat{\theta}_{1:t}}^{-1/2} \boldsymbol{\Omega}_t \widetilde{\mathbf{F}}_{1:t, \widehat{\theta}_{1:t}}^{-1/2} - \mathbf{I}_p \right\|_{\mathbf{F}}. \quad (50)$$

Propositions 9 and 10 address the first and second terms, respectively.

To bound the first term in (50), we need to ensure that the accumulated approximation errors are asymptotically ignorable. To quantify these errors, we introduce some quantity. For $\theta \in \Theta$ and $t \in [T]$, define

$$\eta_t(\theta) = \widetilde{L}_t(\theta) + \frac{1}{2} \left\| \boldsymbol{\Omega}_t^{1/2} (\theta - \mu_t) \right\|_2^2, \quad \nabla \eta_t(\theta) = \nabla \widetilde{L}_t(\theta) + \boldsymbol{\Omega}_t (\theta - \mu_t). \quad (51)$$

Here, $\eta_t(\cdot)$ serves as the error term induced by the t -th variational approximation. Specifically, for $\theta \in \Theta$ and $t \in [T]$, it is not difficult to see that

$$\widetilde{L}_{1:t}(\theta) = \widetilde{L}_t(\theta) + \sum_{s=1}^{t-1} \eta_s(\theta). \quad (52)$$

Based on (52), we briefly present key ideas to bound the first term in (50). Recall that, the definitions of $\widehat{\theta}_{1:t}$ and $\widehat{\theta}_t$ imply the following equations

$$\nabla \widetilde{L}_{1:t}(\widehat{\theta}_{1:t}) = 0, \quad \nabla \widetilde{L}_t(\widehat{\theta}_t) = 0, \quad \forall t \in [T].$$

By these two score equations and (52), we have

$$\begin{aligned} 0 &= \nabla \widetilde{L}_{1:t}(\widehat{\theta}_{1:t}) = \nabla \widetilde{L}_t(\widehat{\theta}_{1:t}) + \sum_{s=1}^{t-1} \nabla \eta_s(\widehat{\theta}_{1:t}) = \nabla \widetilde{L}_t(\widehat{\theta}_{1:t}) - \nabla \widetilde{L}_t(\widehat{\theta}_t) + \sum_{s=1}^{t-1} \nabla \eta_s(\widehat{\theta}_{1:t}) \\ &= -\widetilde{\mathbf{F}}_t(\widehat{\theta}_{1:t}, \widehat{\theta}_t) (\widehat{\theta}_{1:t} - \widehat{\theta}_t) + \sum_{s=1}^{t-1} \nabla \eta_s(\widehat{\theta}_{1:t}) \end{aligned}$$

for some $\widetilde{\mathbf{F}}_t(\widehat{\theta}_{1:t}, \widehat{\theta}_t)$, depending on $\widehat{\theta}_{1:t}$ and $\widehat{\theta}_t$, by Taylor's theorem. If $\widehat{\theta}_{1:t}$ and $\widehat{\theta}_t$ are sufficiently close, one can replace $\widetilde{\mathbf{F}}_t(\widehat{\theta}_{1:t}, \widehat{\theta}_t)$ with $\widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}$ in the last display. Hence, we can obtain the following relation:

$$\widetilde{\mathbf{F}}_{t, \widehat{\theta}_t} (\widehat{\theta}_{1:t} - \widehat{\theta}_t) \approx \widetilde{\mathbf{F}}_t(\widehat{\theta}_{1:t}, \widehat{\theta}_t) (\widehat{\theta}_{1:t} - \widehat{\theta}_t) = \sum_{s=1}^{t-1} \nabla \eta_s(\widehat{\theta}_{1:t}). \quad (53)$$

From this approximation, we have

$$\left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} (\hat{\theta}_{1:t} - \hat{\theta}_t) \right\|_2 \approx \left\| \sum_{s=1}^{t-1} \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} \nabla \eta_s (\hat{\theta}_{1:t}) \right\|_2 \stackrel{\text{def}}{=} \varrho_{n,t}.$$

Also, recall from (37) that on $\mathcal{E}_{\text{est},1}$

$$\left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} (\hat{\theta}_t - \mu_t) \right\|_2 \vee \left\| \boldsymbol{\Omega}_t^{-1/2} \tilde{\mathbf{F}}_{t, \hat{\theta}_t} \boldsymbol{\Omega}_t^{-1/2} - \mathbf{I}_p \right\|_F \lesssim \epsilon_{n,t,\text{KL}} = o(1)$$

Combining these two facts, we can obtain the following rough relations:

$$\left\| \boldsymbol{\Omega}_t^{1/2} (\hat{\theta}_{1:t} - \mu_t) \right\|_2 \approx \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} (\hat{\theta}_{1:t} - \mu_t) \right\|_2 \approx \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} (\hat{\theta}_{1:t} - \hat{\theta}_t) \right\|_2 \approx \left\| \sum_{s=1}^{t-1} \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} \nabla \eta_s (\hat{\theta}_{1:t}) \right\|_2.$$

The main challenge is establishing that $\varrho_{n,t} = o_P(1)$ for all $t \in [T]$. In Proposition 9, we prove that on $\mathcal{E}_{\text{est},1} \cap \mathcal{E}_{\text{est},2}$

$$\varrho_{n,t} = \left\| \sum_{s=1}^{t-1} \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} \nabla \eta_s (\hat{\theta}_{1:t}) \right\|_2 = O \left(M_n^2 \sqrt{\frac{p_*^3}{n}} \right), \quad \forall t \in [T].$$

We now formally state the results addressing the first term in (50).

Proposition 9 *Suppose that (A0), (A1*), (A2), (S) and (P) hold. Then, on $\mathcal{E}_{\text{est},1} \cap \mathcal{E}_{\text{est},2}$, the following inequalities hold uniformly for all $t \in [T]$:*

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} (\hat{\theta}_{1:t} - \hat{\theta}_t) \right\|_2 \vee \left\| \boldsymbol{\Omega}_t^{1/2} (\hat{\theta}_{1:t} - \mu_t) \right\|_2 &\leq K M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2}, \\ \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} (\theta_0 - \hat{\theta}_t) \right\|_2 &\leq K M_n \sqrt{p_*}, \end{aligned} \quad (54)$$

where $K = K(K_{\min}, K_{\max})$.

From Proposition 9, we have $\|\hat{\theta}_t - \theta_0\|_2 \lesssim M_n (p_*/N_t)^{1/2}$, where $N_t = nt$. Since Proposition 6 implies that $\|\hat{\theta}_t - \theta_t^*\|_2 \lesssim M_n t^{-1/2} (p_*/N_t)^{1/2}$, we have

$$\|\theta_t^* - \theta_0\|_2 \lesssim \|\hat{\theta}_t - \theta_t^*\|_2 \vee \|\hat{\theta}_t - \theta_0\|_2 = O \left(M_n (p_*/N_t)^{1/2} \right), \quad (55)$$

This significantly improves the bound in (38).

Under the setting in Proposition 9, one can argue the efficiency of the estimators μ_t and $\hat{\theta}_t$. More precisely, we can show that

$$\begin{aligned} \mathbf{F}_{1:t, \theta_0}^{1/2} (\hat{\theta}_t - \theta_0) &= \mathbf{F}_{1:t, \theta_0}^{-1/2} \nabla L_{1:t}(\theta_0) + o(1), \\ \mathbf{F}_{1:t, \theta_0}^{1/2} (\mu_t - \theta_0) &= \mathbf{F}_{1:t, \theta_0}^{-1/2} \nabla L_{1:t}(\theta_0) + o(1) \end{aligned}$$

with high probability. Therefore, $\hat{\theta}_t$ and μ_t serve as asymptotically efficient estimators in the sense of van der Vaart (2000, Section 8). See Corollary 27 for detailed statements.

Proposition 10 *Suppose that (A0), (A1*), (A2), (S) and (P) hold. Then, on $\mathcal{E}_{\text{est},1} \cap \mathcal{E}_{\text{est},2}$, the following inequality holds uniformly for all $t \in [T]$:*

$$\left\| \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1/2} \tilde{\Omega}_t \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1/2} - \mathbf{I}_p \right\|_{\mathbb{F}} \leq K M_n \left(\frac{p_*^2}{n} \right)^{1/2}, \quad (56)$$

where $K = K(K_{\min}, K_{\max})$.

We present a sketch of the proof of Proposition 10. Provided that $\lambda_{\min}(\tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}) \asymp \lambda_{\max}(\tilde{\mathbf{F}}_{t, \hat{\theta}_t})$, we can obtain the following bound:

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1/2} \tilde{\Omega}_t \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1/2} - \mathbf{I}_p \right\|_{\mathbb{F}} &\lesssim \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} \tilde{\Omega}_t \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_{\mathbb{F}} + \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}} \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_{\mathbb{F}} \\ &\lesssim t^{-1} n^{-1/2} p_* + \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}} \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_{\mathbb{F}}, \end{aligned}$$

where the second inequality holds by (37). Hence, we only need to verify $\tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}} \approx \tilde{\mathbf{F}}_{t, \hat{\theta}_t}$ in a certain sense. From the definitions in (5) and (39), one can obtain the following equality:

$$\tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}} - \tilde{\mathbf{F}}_{t, \hat{\theta}_t} = \sum_{s=1}^{t-1} (\tilde{\mathbf{F}}_{s, \hat{\theta}_s} - \Omega_s) + \sum_{s=1}^t (\mathbf{F}_{s, \hat{\theta}_{1:t}} - \mathbf{F}_{s, \hat{\theta}_s}). \quad (57)$$

Then, under some conditions, we can show that

$$\left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}} \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_{\mathbb{F}} \lesssim t^{-1} \sum_{s=1}^{t-1} (s \cdot \epsilon_{s,1}) + t^{-1} \sum_{s=1}^t \epsilon_{s,2} \lesssim \left(\frac{p_*^2}{n} \right)^{1/2} + \max_{s \in [t]} \epsilon_{s,2},$$

where the second inequality holds by (37), and

$$\epsilon_{s,1} = \left\| \tilde{\mathbf{F}}_{s, \hat{\theta}_s}^{-1/2} \tilde{\Omega}_s \tilde{\mathbf{F}}_{s, \hat{\theta}_s}^{-1/2} - \mathbf{I}_p \right\|_{\mathbb{F}}, \quad \epsilon_{s,2} = \left\| \mathbf{F}_{s, \hat{\theta}_s}^{-1/2} \mathbf{F}_{s, \hat{\theta}_{1:t}} \mathbf{F}_{s, \hat{\theta}_s}^{-1/2} - \mathbf{I}_p \right\|_{\mathbb{F}}.$$

Bounding $\epsilon_{s,2}$ requires that the map $\theta \mapsto \mathbf{F}_{s, \theta}$ is sufficiently smooth. If $\hat{\theta}_{1:t}$ is sufficiently close to $\hat{\theta}_s$, then we can expect that $\mathbf{F}_{s, \hat{\theta}_{1:t}} \approx \mathbf{F}_{s, \hat{\theta}_s}$; see Lemma 40 for a precise statement. Indeed, under some conditions, we can derive from (45) and (54) that

$$\|\hat{\theta}_{1:t} - \hat{\theta}_s\|_2 \leq \|\hat{\theta}_{1:t} - \theta_0\|_2 + \|\hat{\theta}_s - \theta_0\|_2 = O(M_n(p_*/N_s)^{1/2}).$$

By applying Lemma 40, therefore, we can obtain $\max_{s \in [t]} \epsilon_{s,2} = O(M_n p_* n^{-1/2})$, which results in (56).

Theorem 11 (Online Bernstein–von Mises Theorem) *Suppose that (A0), (A1*), (A2), (S) and (P) hold. Then,*

$$\mathbb{P}_0^{(N)} \left(d_V \left(\Pi_t(\cdot), \Pi(\cdot | \mathbf{D}_{1:t}) \right) \leq K M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2} \text{ for all } t \in [T] \right) \geq 1 - 3n^{-1}.$$

If, in addition, (P*) holds, then

$$\mathbb{P}_0^{(N)} \left(d_V \left(\Pi_t, \mathcal{N} \left(\hat{\theta}_{1:t}^{\text{ML}}, \mathbf{F}_{1:t, \theta_0}^{-1} \right) \right) \leq K M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2} \text{ for all } t \in [T] \right) \geq 1 - 3n^{-1}.$$

Here, $K = K(K_{\min}, K_{\max})$.

When p is fixed, Theorem 11 guarantees that the online BvM theorem holds even with a very small mini-batch size n . Specifically, if $n \gg (\log N)^4$ (as required by condition **(S)**), the online BvM theorem holds. Moreover, Theorem 11 further ensures that credible sets based on Π_T yield asymptotically valid frequentist confidence sets. To be specific, for $\alpha \in (0, 1)$, consider the following Wald-type confidence and credible sets:

$$\begin{aligned}\widehat{C}_N(\alpha) &= \left\{ \theta \in \Theta : \left\| \mathbf{F}_{1:T, \theta_0}^{1/2} (\theta - \widehat{\theta}_{1:T}^{\text{ML}}) \right\|_2^2 \leq \chi_{p, \alpha}^2 \right\}, \\ \widehat{C}_{n, T}(\alpha) &= \left\{ \theta \in \Theta : \left\| \boldsymbol{\Omega}_T^{1/2} (\theta - \mu_T) \right\|_2^2 \leq \chi_{p, \alpha}^2 \right\},\end{aligned}$$

where $\chi_{p, \alpha}^2$ denotes the $(1 - \alpha)$ quantile of the χ_p^2 distribution. Here, $\widehat{C}_N(\alpha)$ represents the standard frequentist confidence set based on batch MLE, while $\widehat{C}_{n, T}(\alpha)$ is the credible set based on the sequentially updated posterior Π_T . Under the setting in Theorem 11, we have on $\mathcal{E}_{\text{est}, 1} \cap \mathcal{E}_{\text{est}, 2}$

$$\begin{aligned}N \left\| \boldsymbol{\Omega}_T^{-1} - \mathbf{F}_{1:t, \theta_0}^{-1} \right\|_{\text{F}} &= O \left(M_n (p_*^2/n)^{1/2} \right), \\ \left\| \mathbf{F}_{1:T, \theta_0}^{1/2} (\theta_0 - \widehat{\theta}_{1:T}^{\text{ML}}) \right\|_2 &= \left\| \boldsymbol{\Omega}_T^{1/2} (\theta_0 - \mu_T) \right\|_2 + O \left(M_n (p_*^3/n)^{1/2} \right).\end{aligned}$$

See Corollaries 29 and 30 for precise statements. Therefore, we can conclude that on $\mathcal{E}_{\text{est}, 1} \cap \mathcal{E}_{\text{est}, 2}$ the following inclusions hold:

$$\widehat{C}_N(\alpha - \epsilon_n) \subset \widehat{C}_{n, T}(\alpha) \subset \widehat{C}_N(\alpha + \epsilon_n),$$

where $\epsilon_n = CM_n (p_*^3/n)^{1/2}$ for some constant $C > 0$. Hence, as long as \widehat{C}_N is a valid confidence set, the credible set $\widehat{C}_{n, T}$ also provides valid frequentist coverage.

Our final bound in the online BvM theorem requires the condition $n \gg M_n^4 p_*^3$, which is stronger than the requirement $n \gg (M_n^2 \vee \log^2 T) p_*^2$ specified in the condition **(S)**. This apparent discrepancy arises from a technical artifact in our proof. Specifically, in Proposition 9, we bound the spectral norm by the Frobenius norm:

$$\left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{-1/2} \boldsymbol{\Omega}_t \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_2 \leq \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{-1/2} \boldsymbol{\Omega}_t \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_{\text{F}} \lesssim \left(\frac{p_*^2}{nt^2} \right)^{1/2},$$

which introduces an additional factor of p_* in the final bound.

A sharper analysis is possible by considering LA instead of VB. Under LA, we have $\mu_t = \widehat{\theta}_t$ and $\boldsymbol{\Omega}_t = \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}$, and thus

$$\left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{-1/2} \boldsymbol{\Omega}_t \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_2 = 0.$$

Consequently, the bounds in Proposition 9 and Theorem 11 can be improved from $M_n^2 (p_*^3/n)^{1/2}$ to $M_n^2 (p_*^2/n)^{1/2}$.

However, it should not be interpreted as implying that LA-based online updates are strictly superior to VB-based ones. Importantly, this discrepancy does not reflect an inherent limitation of VB. Rather, the weaker dimension dependence arises from our technical

analysis. Although we do not provide a formal proof, we believe that a refined analysis could also improve the dimension dependence under VB. We plan to investigate this direction in future work, especially in connection with one-pass algorithms and sharper dimension requirements.

9. Numerical experiments

In this section, we present small-scale numerical experiments to complement our theoretical findings. In particular, we observe that the Bayesian online estimator performs comparably to the batch estimator when the mini-batch size n exceeds a certain threshold, while its performance deteriorates when n falls below this threshold. Moreover, the performance gap between the online and batch estimators becomes larger as the number of mini-batches T increases when n is small, whereas the gap remains negligible for sufficiently large n regardless of T .

For the simulation study, we consider a sequence of observations $Y_i \stackrel{i.i.d.}{\sim} \text{Bernoulli}(1/2)$ for all $i \in [N]$, which corresponds to one-dimensional logistic regression with the true parameter $\theta_0 = 0$. We generate $N = 1000$ samples and partition them into mini-batches of size $n = 1, 2, 4, 6, 8, 10, 20, 50, 200$, and 1000.

For the online Bayesian updates, we solve the minimization problem (2). It is well known that minimizing the KL divergence in (2) is equivalent to maximizing the evidence lower bound (ELBO), defined as

$$\max_{Q \in \mathcal{Q}} \{ \mathbb{E}_Q [L_t(\theta)] - K(Q; \Pi_{t-1}) \}, \quad (58)$$

where \mathbb{E}_Q denotes the expectation with respect to the normal distribution $Q = \mathcal{N}(\mu, \mathbf{\Omega}^{-1})$. Although the ELBO for logistic regression is concave in $(\mu, \mathbf{\Omega})$, the expectation in (58) is not available in closed form. We approximate it via Monte Carlo integration using 10^3 samples, and then optimize (58) using a gradient ascent algorithm with a suitably chosen learning rate. The initial prior is set as $\Pi_0 = \mathcal{N}(0, 3^2)$.

We first compare online and batch learning in terms of point estimation performance. Specifically, we compare the online estimator μ_t with the batch estimator $\hat{\theta}_{1:t}^{\text{ML}}$ using the mean squared error (MSE) as the evaluation metric. The experiment was repeated $M = 500$ times, and for each replicate, we obtained an online estimator $\mu_{t,m}$ and a batch estimator $\hat{\theta}_{1:t,m}^{\text{ML}}$ for $m = 1, \dots, M$. In all cases (even when $n = 1$), we observed that $|\mu_t - \theta_0| \rightarrow 0$ as $t \rightarrow \infty$, indicating that the point estimators are consistent regardless of the mini-batch size n . For further comparison, we also computed the *relative efficiency*, defined as

$$\text{RE}_t = \frac{\sum_{m=1}^M |\mu_{t,m} - \theta_0|^2}{\sum_{m=1}^M |\hat{\theta}_{1:t,m}^{\text{ML}} - \theta_0|^2}.$$

Figure 1 shows the relative efficiency. For small mini-batch size (e.g., $n = 1$), the performance of the Bayesian online learning method deteriorates as the number of mini-batches T increases. In this case, the accumulated error becomes non-negligible and the online estimator suffers a significant loss of statistical efficiency relative to the MLE. A similar, though less severe, pattern is observed for $n = 2$. However, as n exceeds a certain

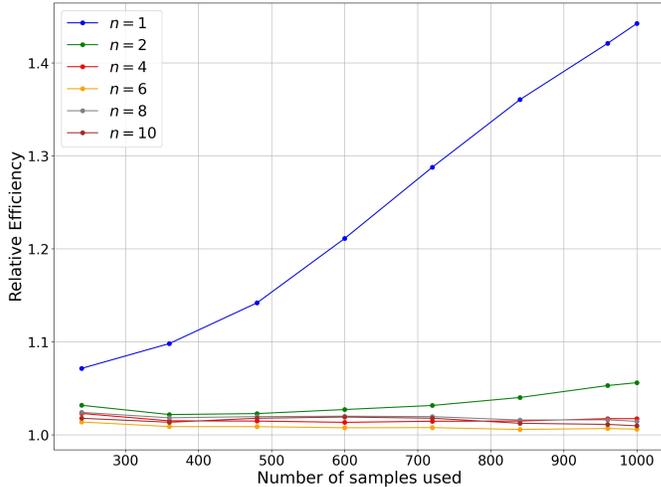


Figure 1: Relative efficiency in terms of MSE across varying mini-batch size.

Table 2: Coverage probability (CP) and length of 95% CIs over $M = 500$ repetitions.

Method	CP	Length
$\hat{\theta}_{1:t}^{\text{ML}}$	0.944	0.248
$\mu_t, n = 1$	0.986	0.398
$\mu_t, n = 2$	0.974	0.301
$\mu_t, n = 4$	0.956	0.269
$\mu_t, n = 6$	0.950	0.259
$\mu_t, n = 8$	0.948	0.255
$\mu_t, n = 10$	0.942	0.253
$\mu_t, n = 20$	0.942	0.250
$\mu_t, n = 50$	0.938	0.248
$\mu_t, n = 200$	0.938	0.248
$\mu_t, n = 1000$	0.942	0.248

threshold, the relative efficiency approaches 1, indicating that the online estimators perform comparably to the MLE.

We also assess the coverage probability of the credible/confidence intervals (CI) for θ_0 . Table 2 summarizes the coverage probability and average length of 95% CIs over 500 repetitions for various mini-batch sizes, alongside those based on the MLE. As the mini-batch size increases, both the coverage probability and CI length tend to align with those of the MLE.

When the mini-batch size is small, however, our numerical results reveal an intriguing phenomenon: the online variational posterior exhibits conservative coverage. This behavior is not immediately apparent from standard intuition. The variance underestimation of mean-field VB is well known in the literature (Blei et al., 2017). Even when a full Gaussian variational family is employed, variance underestimation may still occur in small-sample regimes due to the directionality of the KL divergence, $\min_Q K(Q, \Pi_t)$, although this effect vanishes asymptotically (see Section 4). Furthermore, the posterior mean exhibits a non-negligible bias when the mini-batch size is small (e.g., $n = 1, 2$), as illustrated in Figure 1. Such bias would ordinarily lead to under-coverage of credible sets.

Contrary to this expectation, the empirical coverage becomes conservative in this small- n regime. Although a rigorous explanation is currently lacking, these findings suggest that online VB may implicitly incorporate a form of safety effect. We conjecture that variance inflation induced by sequential updates offsets the usual variance underestimation associated with KL-based variational approximations. A theoretical justification of this implicit regularization effect is deferred to future work.

10. Discussion

There are many papers on Bayesian online learning, but a large portion of them are based on heuristic ideas. Although our theoretical analysis is limited to regular parametric models, the online BvM theorem established in this paper provides strong theoretical justification for online Bayesian procedures and offers a useful mathematical tool for studying further theoretical properties in more complex models.

Remarkably, the online BvM theorem holds with a very small mini-batch size; specifically, $n \gg (\log N)^4$ is sufficient when p is fixed. On the other hand, our simulation results indicate that if the mini-batch size is too small—say, $n = 1$ —then although the online procedure still yields a consistent estimator, it loses statistical efficiency. We believe that this is due to the particular choice of approximation in our online learning procedure, which is based on variational approximation. Given that frequentist one-pass algorithms can yield efficient estimators, we believe that it is possible to develop an online Bayesian procedure with some algorithmic modifications that also achieves frequentist efficiency. We leave this as a direction for future work.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00240861), a Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (RS-2024-00409092, 2024 HRD Program for Industrial Innovation), and Institute of Information & Communications Technology Planning & Evaluation(IITP)-Global Data-X Leader HRD program grant funded by the Korea government (MSIT) (IITP-2024-RS-2024-00441244).

References

- Radosław Adamczak, Alexander Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *J. Amer. Math. Soc.*, 23(2):535–561, 2010.
- Jamil Arbas, Hassan Ashtiani, and Christopher Liaw. Polynomial time and private learning of unbounded Gaussian mixture models. In *Proc. International Conference on Machine Learning*, pages 1018–1040. PMLR, 2023.
- Francis Bach. Self-concordant analysis for logistic regression. *Electron. J. Stat.*, 4:384–414, 2010.
- David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.*, 112(518):859–877, 2017.
- Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C Wilson, and Michael I Jordan. Streaming variational Bayes. In *Proc. Neural Information Processing Systems*, 2013.
- Thang D Bui, Cuong Nguyen, and Richard E Turner. Streaming sparse Gaussian process approximations. In *Proc. Neural Information Processing Systems*, 2017.

- Minwoo Chae. Adaptive Bayesian inference for current status data on a grid. *Bernoulli*, 29(1):403–427, 2023.
- Xi Chen, Weidong Liu, and Yichen Zhang. Quantile regression under memory constraint. *Ann. Statist.*, 47(6):3244–3273, 2019.
- Xi Chen, Jason D. Lee, Xin T. Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *Ann. Statist.*, 48(1):251 – 273, 2020.
- Junhyeok Choi, Jeyong Lee, Yongdai Kim, and Minwoo Chae. Online bayesian inference for cox proportional hazards model. *J. Comput. Graph. Statist.*, pages 1–17, 2025.
- Dennis D Cox. An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.*, pages 903–923, 1993.
- Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *J. Mach. Learn. Res.*, 19(78):1–21, 2018.
- David Freedman. Wald lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Statist.*, 27(4):1119–1141, 1999.
- Matthew Hoffman, Francis Bach, and David Blei. Online learning for latent Dirichlet allocation. In *Proc. Neural Information Processing Systems*, 2010.
- Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of sub-Gaussian random vectors. *Electron. Commun. Probab.*, 17:1–6, 2012.
- Kuhwan Jeong, Minwoo Chae, and Yongdai Kim. Online learning for the Dirichlet process mixture model via weakly conjugate approximation. *Comput. Statist. Data Anal.*, 179:107626, 2023.
- Iain M Johnstone. High dimensional Bernstein-von Mises: Simple examples. In *Borrowing Strength: Theory Powering Applications—A Festschrift for Lawrence D. Brown*, volume 6, pages 87–99. Institute of Mathematical Statistics, 2010.
- Anya Katsevich. The Laplace approximation accuracy in high dimensions: a refined analysis and new skew adjustment. *ArXiv:2306.07262*, 2023.
- Anya Katsevich. The Laplace asymptotic expansion in high dimensions. *ArXiv:2406.12706*, 2024.
- Anya Katsevich. Improved dimension dependence in the Bernstein–von Mises theorem via a new Laplace approximation bound. *Inf. Inference*, 14(3):iaaf020, 2025.
- Anya Katsevich and Philippe Rigollet. On the approximation accuracy of Gaussian variational inference. *Ann. Statist.*, 52(4):1384–1409, 2024.
- Yongdai Kim, Minwoo Chae, Kuhwan Jeong, Byungyup Kang, and Hyoju Chung. An online Gibbs sampler algorithm for hierarchical Dirichlet processes prior. In *Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 509–523, 2016.

- Marc Lambert, Silvère Bonnabel, and Francis Bach. The recursive variational Gaussian approximation (R-VGA). *Stat. Comput.*, 32(1):10, 2022.
- Marc Lambert, Silvère Bonnabel, and Francis Bach. The limited-memory recursive variational Gaussian approximation (L-RVGA). *Stat. Comput.*, 33(3):70, 2023.
- Jeyong Lee, Minwoo Chae, and Ryan Martin. Advances in Bayesian model selection consistency for high-dimensional generalized linear models. *Ann. Statist.*, 53(5):2012–2039, 2025.
- Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast and robust online inference with stochastic gradient descent via random scaling. In *Proc. AAAI Conference on Artificial Intelligence*, pages 7381–7389, 2022.
- Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast inference for quantile regression with tens of millions of observations. *J. Econometrics*, page 105673, 2024.
- Dahua Lin. Online learning of nonparametric mixture models via sequential variational approximation. In *Proc. Neural Information Processing Systems*, 2013.
- George G Lorentz, Manfred von Golitschek, and Yuly Makovoz. *Constructive Approximation: Advanced Problems*, volume 304. Citeseer, 1996.
- Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, 11(1), 2010.
- Yurii Nesterov and Arkadii Nemirovskii. *Interior-point Polynomial Algorithms in Convex Programming*. Society for Industrial Mathematics, 1994.
- Yurii Nesterov et al. *Lectures on Convex Optimization*, volume 137. Springer, 2018.
- Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. In *Proc. International Conference on Learning Representations*, 2018.
- Manfred Opper and Ole Winther. A Bayesian approach to on-line learning. In *On-line Learning in Neural Networks*, pages 363–378. Cambridge University Press, 1999.
- Dmitrii M Ostrovskii and Francis Bach. Finite-sample analysis of M-estimators using self-concordance. *Electron. J. Stat.*, 15(1):326–391, 2021.
- Maxim Panov and Vladimir Spokoiny. Finite sample Bernstein–von Mises theorem for semiparametric problems. *Bayesian Anal.*, 10(3):665–710, 2015.
- Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.
- Nikita Puchkin, Fedor Noskov, and Vladimir Spokoiny. Sharper dimension-free bounds on the Frobenius distance between sample covariance and its expectation. *Bernoulli*, 31(2):1664–1691, 2025.

- Philippe Rigollet. Kullback–Leibler aggregation and misspecified generalized linear models. *Ann. Statist.*, 40(2):639–665, 2012.
- Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statist.*, pages 400–407, 1951.
- David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.
- Sara A. Solla and Ole Winther. Optimal online learning: a Bayesian approach. *Computer Physics Communications*, 121-122:94–97, 1999. Proceedings of the Europhysics Conference on Computational Physics CCP 1998.
- Vladimir Spokoiny. Parametric estimation. Finite sample theory. *Ann. Statist.*, 40(6):2877–2909, 2012.
- Vladimir Spokoiny. Penalized maximum likelihood estimation and effective dimension. *Ann. Inst. Henri Poincaré Probab. Stat.*, 53(1):389–429, 2017.
- Vladimir Spokoiny. Bayesian inference for nonlinear inverse problems. *ArXiv:1912.12694*, 2019.
- Vladimir Spokoiny. Dimension free nonasymptotic bounds on the accuracy of high-dimensional Laplace approximation. *SIAM/ASA J. Uncertain. Quantif.*, 11(3):1044–1068, 2023.
- Vladimir Spokoiny. Estimation for SLS models: Finite sample guarantees. *ArXiv:2404.14227*, 2024.
- Vladimir Spokoiny. Inexact Laplace approximation and the use of posterior mean in Bayesian inference. *Bayesian Anal.*, 20(1):1303–1330, 2025.
- Vladimir Spokoiny and Maxim Panov. Accuracy of Gaussian approximation for high-dimensional posterior distributions. *Bernoulli*, 31(2):843–867, 2025.
- Runlong Tang, Moulinath Banerjee, and Michael R Kosorok. Likelihood based inference for current status data on a grid: A boundary phenomenon and an adaptive inference procedure. *Ann. Statist.*, 40(1):45–72, 2012.
- Panos Toulis and Edoardo M. Airoidi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *Ann. Statist.*, 45(4):1694 – 1727, 2017.
- Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.
- Aad van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.
- Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

- Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Trans. on Inform. Theory*, 55(5):2183–2202, 2009.
- Chong Wang, John Paisley, and David M Blei. Online variational inference for the hierarchical Dirichlet process. In *Proc. International Conference on Artificial Intelligence and Statistics*, 2011.
- Jing Wu, Ming-Hui Chen, Elizabeth D Schifano, and Jun Yan. Online updating of survival analysis. *J. Comput. Graph. Statist.*, 30(4):1209–1223, 2021.
- Yishu Xue, HaiYing Wang, Jun Yan, and Elizabeth D Schifano. An online updating approach for testing the proportional hazards assumption with streams of survival data. *Biometrics*, 76(1):171–182, 2020.
- Keisuke Yano and Kengo Kato. On frequentist coverage errors of Bayesian credible sets in moderately high dimensions. *Bernoulli*, 26(1), 2020.
- Xinzhen Zhang, Chen Ling, and Liqun Qi. The best rank-1 approximation of a symmetric tensor and related spherical optimization problems. *SIAM J. Matrix Anal. Appl.*, 33(3):806–821, 2012.
- Wanrong Zhu, Xi Chen, and Wei Biao Wu. Online covariance matrix estimation in stochastic gradient descent. *J. Amer. Statist. Assoc.*, 118(541):393–404, 2023.

Appendix A. Notations

Table 3: Table of Notations

Notation	Location
$D_t, N_{t,1}(\cdot), N_{t,2}(\cdot), N_{t,3}(\cdot)$	(62)
$\Delta_{\text{local},t}(\cdot), \Delta_{\text{tail},\tilde{\Pi},t}(\cdot), \Delta_{\text{tail,LA},t}(\cdot)$	(61)
$b_{n,t}$	(104)
$\tau_{3,t,\text{bias}}, \tau_{4,t,\text{bias}}$	(104)
Δ_t	(100)

For $k \geq 2$, we say that a k -th order tensor $\mathbf{A} = (A_{i_1, \dots, i_k})_{i_1, \dots, i_k \in [p]} \in \mathbb{R}^{p^k}$ is symmetric if $A_{i_1, \dots, i_k} = A_{\sigma(i_1, \dots, i_k)}$ for any permutation map σ of the tuple (i_1, \dots, i_k) . For a symmetric 3-order tensor $\mathbf{A} = (A_{ijk})_{i,j,k \in [p]} \in \mathbb{R}^{p \times p \times p}$, $\mathbf{B} = (B_{jk})_{j,k \in [p]} \in \mathcal{S}_{++}^p$ and $x \in \mathbb{R}^p$, let

$$\langle \mathbf{A}, \mathbf{B} \rangle = \left(\sum_{j,k \in [p]} A_{ijk} B_{jk} \right)_{i \in [p]} \in \mathbb{R}^p, \quad \langle \mathbf{A}, x \rangle = \left(\sum_{k \in [p]} A_{ijk} x_k \right)_{i,j \in [p]} \in \mathbb{R}^{p \times p}.$$

For a function $f : \Theta \rightarrow \mathbb{R}$, let $\|f\|_\infty = \sup_{\theta \in \Theta} |f(\theta)|$. For $A \subseteq \Theta$, let $\mathbf{1}_A(\theta) : \Theta \rightarrow \{0, 1\}$ be the indicator function, defined as 1 if $\theta \in A$ and 0 otherwise.

Appendix B. Proofs for Section 4

Throughout this section, let $\Theta_{n,t} = \Theta(\tilde{\mathbf{F}}_{t,\hat{\theta}_t}, 4r_{\text{LA}})$. (Note that the notation $\Theta_{n,t}$ is reserved for defining other local sets in subsequent sections.)

B.1 Laplace approximation

For a function $g : \Theta \rightarrow \mathbb{R}$, let

$$\begin{aligned} \mathcal{I}_{\tilde{\Pi},t}(g) &:= \frac{\int g(\theta) \exp[\tilde{L}_t(\theta)] \, d\theta}{\int \exp[\tilde{L}_t(\theta')] \, d\theta'} = \frac{\int g(u) e^{f_t(u;\hat{\theta}_t)} \, du}{\int e^{f_t(u';\hat{\theta}_t)} \, du'}, \\ \mathcal{I}_{\text{LA},t}(g) &:= \frac{\int g(\theta) \exp\left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} (\theta - \hat{\theta}_t) \right\|_2^2\right] \, d\theta}{\int \exp\left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} (\theta' - \hat{\theta}_t) \right\|_2^2\right] \, d\theta'} \\ &= \frac{\int g(\theta) \exp\left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2\right] \, du}{\int \exp\left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u' \right\|_2^2\right] \, du'}, \end{aligned} \quad (59)$$

where $\int f(u)du = \int_{\Theta} f(u)du$ and

$$f_t(u; \hat{\theta}_t) = \tilde{L}_t(\hat{\theta}_t + u) - \tilde{L}_t(\hat{\theta}_t) = \tilde{L}_t(\hat{\theta}_t + u) - \tilde{L}_t(\hat{\theta}_t) - \langle \nabla \tilde{L}_t(\hat{\theta}_t), u \rangle. \quad (60)$$

For $t \in [T]$, the total variation distance between Π_t^{LA} and $\tilde{\Pi}_t(\cdot \mid \mathbf{D}_t)$ can be represented by

$$d_V\left(\Pi_t^{\text{LA}}(\cdot), \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t)\right) = \sup_{\|g\|_{\infty} \leq 1} \left| \mathcal{I}_{\tilde{\Pi},t}(g) - \mathcal{I}_{\text{LA},t}(g) \right|,$$

where the supremum is taken over every measurable function g with $\|g\|_{\infty} \leq 1$. Define the following quantities:

$$\begin{aligned} \Delta_{\text{local},t}(g) &= \left| \frac{N_{t,1}(g)}{D_t} \right|, \quad \Delta_{\text{tail},\tilde{\Pi},t}(g) = \left| \frac{N_{t,2}(g)}{D_t} \right|, \quad \Delta_{\text{tail,LA},t}(g) = \left| \frac{N_{t,3}(g)}{D_t} \right|, \\ \Delta_{\text{tail},t}(g) &= \Delta_{\text{tail},\tilde{\Pi},t}(g) \vee \Delta_{\text{tail,LA},t}(g), \end{aligned} \quad (61)$$

where

$$\begin{aligned} D_t &= \int_{\Theta} \exp\left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2\right) \, du, \\ N_{t,1}(g) &= \int_{\Theta_{n,t}} g(u) \left[\exp\left(f_t(u, \hat{\theta}_t)\right) - \exp\left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2\right) \right] \, du, \\ N_{t,2}(g) &= \int_{\Theta_{n,t}^c} |g(u)| \exp\left(f_t(u, \hat{\theta}_t)\right) \, du, \\ N_{t,3}(g) &= \int_{\Theta_{n,t}^c} |g(u)| \exp\left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2\right) \, du. \end{aligned} \quad (62)$$

Lemma 12 *Suppose that (A0) holds. Also, assume that $\Delta_{\text{local},t}(\mathbf{1}_\Theta) + \Delta_{\text{tail},t}(\mathbf{1}_\Theta) \leq 1/2$ for all $t \in [T]$ on an event \mathcal{E} . Let $g : \Theta \rightarrow \mathbb{R}$ be a function satisfying $\|g\|_\infty \leq 1$. Then, for all $t \in [T]$,*

$$\left| \mathcal{I}_{\tilde{\Pi},t}(g) - \mathcal{I}_{\text{LA},t}(g) \right| \leq 2 \left(\Delta_{\text{local},t}(g) + \Delta_{\text{local},t}(\mathbf{1}_\Theta) + 2\Delta_{\text{tail},t}(g) + 2\Delta_{\text{tail},t}(\mathbf{1}_\Theta) \right)$$

on \mathcal{E} .

Proof In this proof, we work on the event \mathcal{E} without explicitly referring to it. Let

$$\begin{aligned} \text{(i)} &= \frac{\int g(u) e^{f_t(u; \hat{\theta}_t)} du}{\int e^{f_t(u; \hat{\theta}_t)} du} - \frac{\int g(u) \exp \left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \right] du}{\int e^{f_t(u; \hat{\theta}_t)} du}, \\ \text{(ii)} &= \frac{\int g(u) \exp \left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \right] du}{\int e^{f_t(u; \hat{\theta}_t)} du} - \frac{\int g(u) \exp \left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \right] du}{\int \exp \left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u' \right\|_2^2 \right] du'}. \end{aligned}$$

Note that

$$\left| \mathcal{I}_{\tilde{\Pi},t}(g) - \mathcal{I}_{\text{LA},t}(g) \right| \leq |\text{(i)}| + |\text{(ii)}|. \quad (63)$$

Firstly, we will obtain an upper bound of (i). Note that

$$|\text{(i)}| \leq \left| \int e^{f_t(u; \hat{\theta}_t)} du \right|^{-1} \times \left| \int g(u) \left[e^{f_t(u; \hat{\theta}_t)} - \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \right) \right] du \right|.$$

By the definitions in (61), note that

$$\begin{aligned} D_t \Delta_{\text{local},t}(\mathbf{1}_\Theta) &= |N_{t,1}(\mathbf{1}_\Theta)| = \left| \int_{\Theta_{n,t}} \left[\exp \left(f_t(u, \hat{\theta}_t) \right) - \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \right) \right] du \right| \\ &\geq \int_{\Theta_{n,t}} \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \right) du - \int_{\Theta_{n,t}} \exp \left(f_t(u, \hat{\theta}_t) \right) du, \end{aligned}$$

which implies that

$$\begin{aligned} &\int_{\Theta_{n,t}} \exp \left(f_t(u, \hat{\theta}_t) \right) du \\ &\geq \int_{\Theta_{n,t}} \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \right) du - D_t \Delta_{\text{local},t}(\mathbf{1}_\Theta) \\ &= \int_{\Theta_{n,t}} \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \right) du - \Delta_{\text{local},t}(\mathbf{1}_\Theta) \int \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \right) du. \end{aligned}$$

It follows that

$$\begin{aligned}
 & \int e^{f_t(u; \hat{\theta}_t)} du \geq \int_{\Theta_{n,t}} e^{f_t(u; \hat{\theta}_t)} du \\
 & \geq \int_{\Theta_{n,t}} \exp\left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2\right) du - \Delta_{\text{local}, t}(\mathbf{1}_\Theta) \int \exp\left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2\right) du \\
 & = \left(1 - \Delta_{\text{tail}, \text{LA}, t}(\mathbf{1}_\Theta) - \Delta_{\text{local}, t}(\mathbf{1}_\Theta)\right) \int \exp\left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2\right) du
 \end{aligned} \tag{64}$$

and

$$\begin{aligned}
 & \left| \int g(u) \left[e^{f_t(u; \hat{\theta}_t)} - \exp\left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2\right) \right] du \right| \\
 & \leq \left| \int_{\Theta_{n,t}} g(u) \left[e^{f_t(u; \hat{\theta}_t)} - \exp\left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2\right) \right] du \right| \\
 & \quad + \left| \int_{\Theta_{n,t}^c} g(u) \left[e^{f_t(u; \hat{\theta}_t)} - \exp\left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2\right) \right] du \right| \\
 & \leq \left(\Delta_{\text{local}, t}(g) + \Delta_{\text{tail}, \tilde{\Pi}, t}(g) + \Delta_{\text{tail}, \text{LA}, t}(g) \right) \int \exp\left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2\right) du.
 \end{aligned} \tag{65}$$

It follows that

$$\begin{aligned}
 |\text{(i)}| & \leq \left(1 - \Delta_{\text{tail}, \text{LA}, t}(\mathbf{1}_\Theta) - \Delta_{\text{local}, t}(\mathbf{1}_\Theta)\right)^{-1} \\
 & \quad \times \left(\Delta_{\text{local}, t}(g) + \Delta_{\text{tail}, \tilde{\Pi}, t}(g) + \Delta_{\text{tail}, \text{LA}, t}(g) \right).
 \end{aligned} \tag{66}$$

Next, we will obtain an upper bound of (ii). Let $\tilde{Z} \sim \mathcal{N}(0, \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1})$ and $\mathbb{E}_{\tilde{Z}}$ denote the expectation with respect to the law of \tilde{Z} . Note that

$$\begin{aligned}
 |\text{(ii)}| & \leq \left| \frac{\int g(u) \exp\left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2\right] du}{\int \exp\left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u' \right\|_2^2\right] du'} \right| \times \left| \frac{\int \exp\left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2\right] du}{\int e^{f_t(u'; \hat{\theta}_t)} du'} - 1 \right| \\
 & = \left| \mathbb{E}_{\tilde{Z}} g(\tilde{Z}) \right| \times \left| \frac{\int \exp\left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2\right] du}{\int e^{f_t(u'; \hat{\theta}_t)} du'} - 1 \right| \leq \left| \frac{\int \exp\left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2\right] du}{\int e^{f_t(u'; \hat{\theta}_t)} du'} - 1 \right| \\
 & \leq \left(1 - \Delta_{\text{tail}, \text{LA}, t}(\mathbf{1}_\Theta) - \Delta_{\text{local}, t}(\mathbf{1}_\Theta)\right)^{-1} \left(\Delta_{\text{local}, t}(\mathbf{1}_\Theta) + \Delta_{\text{tail}, \tilde{\Pi}, t}(\mathbf{1}_\Theta) + \Delta_{\text{tail}, \text{LA}, t}(\mathbf{1}_\Theta) \right),
 \end{aligned}$$

where the last inequality holds by (64) and (65). Combining the last display with (66), the right hand side of (63) is bounded by

$$\begin{aligned} & \frac{\left[\Delta_{\text{local},t}(g) + \Delta_{\text{tail},\tilde{\Pi},t}(g) + \Delta_{\text{tail,LA},t}(g) \right] + \left[\Delta_{\text{local},t}(\mathbf{1}_\Theta) + \Delta_{\text{tail},\tilde{\Pi},t}(\mathbf{1}_\Theta) + \Delta_{\text{tail,LA},t}(\mathbf{1}_\Theta) \right]}{\left(1 - \Delta_{\text{tail,LA},t}(\mathbf{1}_\Theta) - \Delta_{\text{local},t}(\mathbf{1}_\Theta) \right)} \\ & \leq 2 \left[\Delta_{\text{local},t}(g) + \Delta_{\text{local},t}(\mathbf{1}_\Theta) + 2\Delta_{\text{tail},t}(g) + 2\Delta_{\text{tail},t}(\mathbf{1}_\Theta) \right], \end{aligned}$$

where the inequality holds by the assumption $\Delta_{\text{local},t}(\mathbf{1}_\Theta) + \Delta_{\text{tail},t}(\mathbf{1}_\Theta) \leq 1/2$. This completes the proof. \blacksquare

Lemma 13 *Suppose that (A0) holds, and $N \geq 2$. Also, assume that $\widehat{\tau}_{3,t} r_{\text{LA}} \leq 1/8$ for all $t \in [T]$ on an event \mathcal{E} . Then, on \mathcal{E} ,*

$$\sup_{\|g\|_\infty \leq 1} \Delta_{\text{tail},\tilde{\Pi},t}(g) \leq 2e^{-8 \log N - 8p} \quad \text{and} \quad (67)$$

$$\sup_{\|g\|_\infty \leq 1} \Delta_{\text{tail,LA},t}(g) \leq e^{-16 \log N - 49p/2} \quad \forall t \in [T]. \quad (68)$$

Proof In this proof, we work on the event \mathcal{E} without explicitly referring to it. Note that

$$\sup_{\|g\|_\infty \leq 1} \Delta_{\text{tail},\tilde{\Pi},t}(g) = \Delta_{\text{tail},\tilde{\Pi},t}(\mathbf{1}_\Theta) = \frac{\int_{\Theta_{n,t}^c} e^{f_t(u; \widehat{\theta}_t)} du}{\int \exp\left(-\frac{1}{2} \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{1/2} u \right\|_2^2\right) du}. \quad (69)$$

To obtain an upper bound of the right-hand side in the last display, we will obtain an upper bound of $f_t(u; \widehat{\theta}_t)$ on $\Theta_{n,t}^c$. Let $u \in \mathbb{R}^p$ with $u \notin \Theta_{n,t}$. Then, we have

$$u^\circ \stackrel{\text{def}}{=} 4r_{\text{LA}} \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{1/2} u \right\|_2^{-1} u \in \partial\Theta_{n,t} \stackrel{\text{def}}{=} \left\{ \theta \in \Theta : \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{1/2} \theta \right\|_2 = 4r_{\text{LA}} \right\}. \quad (70)$$

By Taylor's theorem, note that

$$\begin{aligned} & \widetilde{L}_t(\widehat{\theta}_t + u^\circ) - \widetilde{L}_t(\widehat{\theta}_t) - \langle \nabla \widetilde{L}_t(\widehat{\theta}_t), u^\circ \rangle \leq \sup_{\bar{u} \in \Theta_{n,t}} \left[-\frac{1}{2} \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t + \bar{u}}^{1/2} u^\circ \right\|_2^2 \right] \\ & \leq -\frac{1}{2} (1 - 4\widehat{\tau}_{3,t} r_{\text{LA}}) \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{1/2} u^\circ \right\|_2^2, \end{aligned} \quad (71)$$

where the second inequality holds by Lemma 40. Also, Taylor's theorem gives

$$\begin{aligned} & \langle \nabla \widetilde{L}_t(\widehat{\theta}_t + u^\circ) - \nabla \widetilde{L}_t(\widehat{\theta}_t), u - u^\circ \rangle \leq \sup_{\bar{u} \in \Theta_{n,t}} \left[-\langle \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t + \bar{u}}^{1/2} u^\circ, u - u^\circ \rangle \right] \\ & \stackrel{(70)}{=} - \left([4r_{\text{LA}}]^{-1} \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{1/2} u \right\|_2 - 1 \right) \inf_{\bar{u} \in \Theta_{n,t}} \langle \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t + \bar{u}}^{1/2} u^\circ, u^\circ \rangle \\ & \leq - \left([4r_{\text{LA}}]^{-1} \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{1/2} u \right\|_2 - 1 \right) (1 - 4\widehat{\tau}_{3,t} r_{\text{LA}}) \langle \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{1/2} u^\circ, u^\circ \rangle \\ & = -(1 - 4\widehat{\tau}_{3,t} r_{\text{LA}}) \langle \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{1/2} u^\circ, u - u^\circ \rangle, \end{aligned} \quad (72)$$

where the second inequality holds by Lemma 40.

The concavity of $\tilde{L}_t(\cdot)$ implies that:

$$\tilde{L}_t(\hat{\theta}_t + u) \leq \tilde{L}_t(\hat{\theta}_t + u^\circ) + \langle \nabla \tilde{L}_t(\hat{\theta}_t + u^\circ), u - u^\circ \rangle. \quad (73)$$

Also,

$$\begin{aligned} f_t(u; \hat{\theta}_t) &= \tilde{L}_t(\hat{\theta}_t + u) - \tilde{L}_t(\hat{\theta}_t) - \langle \nabla \tilde{L}_t(\hat{\theta}_t), u \rangle \\ &= \left[\tilde{L}_t(\hat{\theta}_t + u) - \tilde{L}_t(\hat{\theta}_t + u^\circ) - \langle \nabla \tilde{L}_t(\hat{\theta}_t + u^\circ), u - u^\circ \rangle \right] \\ &\quad + \tilde{L}_t(\hat{\theta}_t + u^\circ) - \tilde{L}_t(\hat{\theta}_t) - \langle \nabla \tilde{L}_t(\hat{\theta}_t), u^\circ \rangle + \langle \nabla \tilde{L}_t(\hat{\theta}_t + u^\circ) - \nabla \tilde{L}_t(\hat{\theta}_t), u - u^\circ \rangle \\ &\stackrel{(73)}{\leq} \tilde{L}_t(\hat{\theta}_t + u^\circ) - \tilde{L}_t(\hat{\theta}_t) - \langle \nabla \tilde{L}_t(\hat{\theta}_t), u^\circ \rangle + \langle \nabla \tilde{L}_t(\hat{\theta}_t + u^\circ) - \nabla \tilde{L}_t(\hat{\theta}_t), u - u^\circ \rangle \\ &\stackrel{(71)}{\leq} -\frac{1}{2}(1 - 4\hat{\tau}_{3,t}r_{\text{LA}}) \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u^\circ \right\|_2^2 - (1 - 4\hat{\tau}_{3,t}r_{\text{LA}}) \langle \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u^\circ, u - u^\circ \rangle \\ &\stackrel{(70)}{=} -\frac{1}{2}(1 - 4\hat{\tau}_{3,t}r_{\text{LA}}) \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u^\circ \right\|_2^2 - (1 - 4\hat{\tau}_{3,t}r_{\text{LA}}) \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u^\circ \right\|_2 \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2 \\ &\quad + (1 - 4\hat{\tau}_{3,t}r_{\text{LA}}) \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u^\circ \right\|_2^2 \\ &= \frac{1}{2}(1 - 4\hat{\tau}_{3,t}r_{\text{LA}}) \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u^\circ \right\|_2^2 - (1 - 4\hat{\tau}_{3,t}r_{\text{LA}}) \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u^\circ \right\|_2 \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2 \\ &\leq -\frac{1}{2}(1 - 4\hat{\tau}_{3,t}r_{\text{LA}}) \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u^\circ \right\|_2 \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2 = -2(1 - 4\hat{\tau}_{3,t}r_{\text{LA}})r_{\text{LA}} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2, \end{aligned}$$

where the last line holds because $\left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2 > \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u^\circ \right\|_2 = 4r_{\text{LA}}$. It follows that

$$\begin{aligned} \int_{\Theta_{n,t}^c} e^{f_t(u; \hat{\theta}_t)} du &\leq \int_{\Theta_{n,t}^c} \exp \left[-2(1 - 4\hat{\tau}_{3,t}r_{\text{LA}})r_{\text{LA}} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2 \right] du \\ &= \int_{\Theta_{n,t}^c} \exp \left[-2(1 - 4\hat{\tau}_{3,t}r_{\text{LA}})r_{\text{LA}} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2 + \frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 \right] \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 \right) du. \end{aligned}$$

Combining with the last display, the right-hand side of (69) is bounded by

$$\begin{aligned} &\int_{\Theta_{n,t}^c} \exp \left[-2(1 - 4\hat{\tau}_{3,t}r_{\text{LA}})r_{\text{LA}} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2 + \frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 \right] \frac{\exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 \right)}{\int \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u' \right\|_2^2 \right) du'} du \\ &= \mathbb{E}_Z \left(\exp \left[-2(1 - 4\hat{\tau}_{3,t}r_{\text{LA}})r_{\text{LA}} \|Z\|_2 + \frac{1}{2} \|Z\|_2^2 \right] \mathbf{1} \{ \|Z\|_2 \geq 4r_{\text{LA}} \} \right) \\ &\leq \mathbb{E}_Z \left(\exp \left[-r_{\text{LA}} \|Z\|_2 + \frac{1}{2} \|Z\|_2^2 \right] \mathbf{1} \{ \|Z\|_2 \geq 4r_{\text{LA}} \} \right), \end{aligned}$$

where $Z \sim \mathcal{N}(0, \mathbf{I}_p)$ and the inequality holds by the assumption $\hat{\tau}_{3,t}r_{\text{LA}} \leq 1/8$. The right-hand side of the last display is equal to

$$\int_{4r_{\text{LA}}}^{\infty} \exp \left[-r_{\text{LA}}\omega + \frac{1}{2}\omega^2 \right] p(\omega) d\omega \quad (74)$$

where $p(\cdot)$ denotes the density function of $\|Z\|_2$. Note that $p(\cdot)$ is the derivative of the map $\omega \mapsto S(\omega) = -\mathbb{P}(\|Z\|_2 > \omega)$ and

$$\lim_{\omega \rightarrow \infty} \exp \left[-r_{\text{LA}}\omega + \frac{1}{2}\omega^2 \right] S(\omega) = 0$$

because we have, for any $\omega \geq \sqrt{p}$,

$$\begin{aligned} \exp \left[-r_{\text{LA}}\omega + \frac{1}{2}\omega^2 \right] \mathbb{P}(\|Z\|_2 > \omega) &\leq \exp \left[-r_{\text{LA}}\omega + \frac{1}{2}\omega^2 \right] \exp \left[-\frac{(\omega - \sqrt{p})^2}{2} \right] \\ &= \exp \left[(\sqrt{p} - r_{\text{LA}})\omega - \frac{p}{2} \right] = \exp \left[-(\sqrt{p} + \sqrt{2 \log N})\omega - \frac{p}{2} \right], \end{aligned}$$

where the inequality holds by the application of Lemma 49 with $\mathbf{A} = \mathbf{B} = \mathbf{I}_p$. Hence, integration by parts implies that (74) is equal to

$$\exp(4r_{\text{LA}}^2)\mathbb{P}(\|Z\|_2 > 4r_{\text{LA}}) + \int_{4r_{\text{LA}}}^{\infty} (\omega - r_{\text{LA}}) \exp \left(-r_{\text{LA}}\omega + \frac{1}{2}\omega^2 \right) \mathbb{P}(\|Z\|_2 > \omega) d\omega.$$

By Lemma 49 with $\mathbf{A} = \mathbf{B} = \mathbf{I}_p$,

$$\mathbb{P}(\|Z\|_2 > 4r_{\text{LA}}) \leq \exp \left(-\frac{1}{2} \left[7\sqrt{p} + 4\sqrt{2 \log N} \right]^2 \right),$$

which implies that

$$\begin{aligned} &\exp(4r_{\text{LA}}^2)\mathbb{P}(\|Z\|_2 > 4r_{\text{LA}}) \\ &\leq \exp \left(-\frac{1}{2} \left[7\sqrt{p} + 4\sqrt{2 \log N} \right]^2 + 4 \left[2\sqrt{p} + \sqrt{2 \log N} \right]^2 \right) \\ &\leq \exp \left(-\frac{17}{2}p - 8 \log N \right). \end{aligned} \tag{75}$$

At the end of this proof, we will show that

$$\int_{4r_{\text{LA}}}^{\infty} (\omega - r_{\text{LA}}) \exp \left(-r_{\text{LA}}\omega + \frac{1}{2}\omega^2 \right) \mathbb{P}(\|Z\|_2 > \omega) d\omega \leq \exp(-8p - 8 \log N). \tag{76}$$

By (75) and (76), we have

$$\sup_{\|g\|_{\infty} \leq 1} \Delta_{\text{tail}, \tilde{\Pi}, t}(g) \leq 2 \exp(-8p - 8 \log N),$$

which completes the proof of (67).

Similarly, one can bound the left hand side of (68) as

$$\begin{aligned} \sup_{\|g\|_{\infty} \leq 1} \Delta_{\text{tail}, \text{LA}, t}(g) &= \Delta_{\text{tail}, \text{LA}, t}(\mathbf{1}_{\Theta}) = \int_{\Theta_{n,t}^c} \frac{\exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \right)}{\int \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u' \right\|_2^2 \right) du'} du \\ &= \mathbb{P}(\|Z\|_2 > 4r_{\text{LA}}) \leq \exp \left(-\frac{1}{2} \left[7\sqrt{p} + 4\sqrt{2 \log N} \right]^2 \right) \leq \exp \left(-\frac{49}{2}p - 16 \log N \right). \end{aligned}$$

To complete the proof, we only need to prove (76). Note that

$$\begin{aligned}
 & \int_{4r_{\text{LA}}}^{\infty} (\omega - r_{\text{LA}}) \exp\left(-r_{\text{LA}}\omega + \frac{1}{2}\omega^2\right) \mathbb{P}(\|Z\|_2 > \omega) \, d\omega \\
 \stackrel{\text{Lemma 49}}{\leq} & \int_{4r_{\text{LA}}}^{\infty} (\omega - r_{\text{LA}}) \exp\left[-r_{\text{LA}}\omega + \frac{1}{2}\omega^2 - \frac{(\omega - \sqrt{p})^2}{2}\right] \, d\omega \\
 = & \int_0^{\infty} (\omega + 3r_{\text{LA}}) \exp\left[-r_{\text{LA}}(\omega + 4r_{\text{LA}}) + \frac{(\omega + 4r_{\text{LA}})^2}{2} - \frac{(\omega + 4r_{\text{LA}} - \sqrt{p})^2}{2}\right] \, d\omega \\
 = & \exp\left[-4r_{\text{LA}}(r_{\text{LA}} - \sqrt{p}) - \frac{p}{2}\right] \int_0^{\infty} (\omega + 3r_{\text{LA}}) \exp[-(r_{\text{LA}} - \sqrt{p})\omega] \, d\omega \\
 = & \exp\left[-4r_{\text{LA}}(r_{\text{LA}} - \sqrt{p}) - \frac{p}{2}\right] (r_{\text{LA}} - \sqrt{p})^{-1} (\mathbb{E}(W) + 3r_{\text{LA}}),
 \end{aligned}$$

where $W \sim \exp(r_{\text{LA}} - \sqrt{p})$ and density function of $\exp(\lambda)$ is given by $x \mapsto \lambda e^{-\lambda x}$. By $\mathbb{E}W = (r_{\text{LA}} - \sqrt{p})^{-1}$, the right-hand side of the last display is equal to

$$\begin{aligned}
 & \exp\left[-4r_{\text{LA}}(r_{\text{LA}} - \sqrt{p}) - \frac{p}{2}\right] \left[\frac{1}{(r_{\text{LA}} - \sqrt{p})^2} + \frac{3r_{\text{LA}}}{r_{\text{LA}} - \sqrt{p}}\right] \\
 \stackrel{(19)}{=} & \exp\left[-4\left(2\sqrt{p} + \sqrt{2\log N}\right)\left(\sqrt{p} + \sqrt{2\log N}\right) - \frac{p}{2}\right] \\
 & \times \left[\frac{1}{(\sqrt{p} + \sqrt{2\log N})^2} + \frac{6\sqrt{p} + 3\sqrt{2\log N}}{\sqrt{p} + \sqrt{2\log N}}\right] \\
 \leq & \exp\left[-4\left(2\sqrt{p} + \sqrt{2\log N}\right)\left(\sqrt{p} + \sqrt{2\log N}\right)\right] e^{-p/2} (1 + 6) \\
 = & \exp(-8p - 8\log N) \exp(-12\sqrt{2p\log N} - p/2 + \log 7) \leq \exp(-8p - 8\log N),
 \end{aligned}$$

where the last inequality holds by the assumption $N \geq 2$. This completes the proof. \blacksquare

Lemma 14 *Suppose that conditions in Lemma 13 hold. Then, for $0 \leq A_n \leq \sqrt{p} + \sqrt{2\log N}$,*

$$\int_{\Theta_{n,t}^c} \exp\left(A_n \left\|\tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u\right\|_2\right) \frac{\exp\left(-\frac{1}{2}\left\|\tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u\right\|_2^2\right)}{\int \exp\left(-\frac{1}{2}\left\|\tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u'\right\|_2^2\right) \, du'} \, du \leq 2e^{-8p-8\log N} \quad (77)$$

on \mathcal{E} , where \mathcal{E} is the event specified in Lemma 13.

Proof In this proof, we work on the event \mathcal{E} without explicitly referring to it. Let $\tilde{Z} = \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} Z$, where $Z \sim \mathcal{N}(0, \mathbf{I}_p)$. With slight abuse of notations, let \mathbb{E} be the corresponding

expectation for Z . Note that

$$\begin{aligned} & \int_{\Theta_{n,t}^c} \exp \left[A_n \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2 \right] \frac{\exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 \right)}{\int \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u' \right\|_2^2 \right) du'} du \\ &= \mathbb{E} \left(\exp \left[A_n \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} \tilde{Z} \right\|_2 \right] \mathbf{1}_{\{\|Z\|_2 \geq 4r_{\text{LA}}\}} \right) = \mathbb{E} \left(e^{A_n \|Z\|_2} \mathbf{1}_{\{\|Z\|_2 \geq 4r_{\text{LA}}\}} \right). \end{aligned}$$

As in the proof of Lemma 13, let $p(\cdot)$ be the density of $\|Z\|_2$, which is the derivative of the map $\omega \mapsto S(\omega) = -\mathbb{P}(\|Z\|_2 > \omega)$. Then, the last display equals

$$\int_{4r_{\text{LA}}}^{\infty} e^{A_n \omega} p(\omega) d\omega = \exp(4\delta_n r_{\text{LA}}) \mathbb{P}(\|Z\|_2 > 4r_{\text{LA}}) + \int_{4r_{\text{LA}}}^{\infty} A_n \exp(A_n \omega) \mathbb{P}(\|Z\|_2 > \omega) d\omega$$

by the integration by parts. Note that

$$\begin{aligned} \mathbb{P}(\|Z\|_2 > 4r_{\text{LA}}) &= \mathbb{P}(\|Z\|_2 > 8\sqrt{p} + 4\sqrt{2\log N}) \\ &\leq \mathbb{P}(\|Z\|_2 > \sqrt{p} + \sqrt{[7\sqrt{p} + 4\sqrt{2\log N}]^2}) \\ &\stackrel{\text{Lemma 49}}{\leq} \exp\left(-\frac{1}{2} [7\sqrt{p} + 4\sqrt{2\log N}]^2\right), \end{aligned}$$

which implies that

$$\begin{aligned} & \exp(4A_n r_{\text{LA}}) \mathbb{P}(\|Z\|_2 > 4r_{\text{LA}}) \\ & \leq \exp\left(-\frac{1}{2} [7\sqrt{p} + 4\sqrt{2\log N}]^2 + 4A_n [2\sqrt{p} + \sqrt{2\log N}]\right) \\ & \leq \exp\left(-\frac{33}{2}p - 8\log N\right), \end{aligned} \tag{78}$$

where the last inequality holds by the assumption $A_n \leq \sqrt{p} + \sqrt{2\log N}$. At the end of this proof, we will show that

$$\int_{4r_{\text{LA}}}^{\infty} A_n \exp(A_n \omega) \mathbb{P}(\|Z\|_2 > \omega) d\omega \leq \exp(-8p - 8\log N). \tag{79}$$

Therefore, (77) holds by (78) and (79).

To complete the proof, we only need to prove (79). Note that

$$\begin{aligned}
 & \int_{4r_{\text{LA}}}^{\infty} A_n \exp(A_n \omega) \mathbb{P}(\|Z\|_2 > \omega) d\omega \\
 \stackrel{\text{Lemma 49}}{\leq} & \int_{4r_{\text{LA}}}^{\infty} A_n \exp\left[A_n \omega - \frac{(\omega - \sqrt{p})^2}{2}\right] d\omega = \int_{4r_{\text{LA}}}^{\infty} A_n \exp\left[A_n \omega - \frac{\omega^2 - 2\sqrt{p}\omega + p}{2}\right] d\omega \\
 & = e^{A_n^2/2 + A_n \sqrt{p}} A_n \int_{4r_{\text{LA}}}^{\infty} \exp\left[-\frac{1}{2}\left(\omega - [\sqrt{p} + A_n]\right)^2\right] d\omega \\
 & \leq e^{A_n^2/2 + 2A_n \sqrt{p}} \int_{4r_{\text{LA}}}^{\infty} \exp\left[-\frac{1}{2}\left(\omega - [\sqrt{p} + A_n]\right)^2\right] d\omega \\
 & = e^{A_n^2/2 + 2A_n \sqrt{p} + \log(\sqrt{2\pi})} \int_{6\sqrt{p} + 3\sqrt{2\log N}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\omega^2/2} d\omega \\
 & \leq \exp\left[\frac{A_n^2}{2} + 2A_n \sqrt{p} + \log(\sqrt{2\pi}) - \frac{1}{2}\left(6\sqrt{p} + 3\sqrt{2\log N}\right)^2\right] \\
 & \leq \exp\left[-\frac{31}{2}p - 8\log N - 15\sqrt{2p\log N} + \log(\sqrt{2\pi})\right] \\
 & \leq \exp(-8p - 8\log N),
 \end{aligned}$$

where the fourth inequality holds by $A_n \leq \sqrt{p} + \sqrt{2\log N}$. This completes the proof. \blacksquare

For $\theta, u \in \Theta$, let

$$\begin{aligned}
 \mathcal{R}_{t,3}(\theta, u) &= \tilde{L}_t(\theta + u) - \tilde{L}_t(\theta) - \langle \nabla \tilde{L}_t(\theta), u \rangle - \frac{\langle \nabla^2 \tilde{L}_t(\theta), u^{\otimes 2} \rangle}{2}, \\
 \mathcal{R}_{t,4}(\theta, u) &= \mathcal{R}_{t,3}(\theta, u) - \frac{\langle \nabla^3 \tilde{L}_t(\theta), u^{\otimes 3} \rangle}{6}.
 \end{aligned} \tag{80}$$

For simplicity, we often use the notations $\mathcal{R}_{t,3}(u) = \mathcal{R}_{t,3}(\hat{\theta}_t, u)$ and $\mathcal{R}_{t,4}(u) = \mathcal{R}_{t,4}(\hat{\theta}_t, u)$.

Lemma 15 *Suppose that (A0) holds. Also, assume that*

$$(\hat{\tau}_{3,t} r_{\text{LA}}^2) \vee (\hat{\tau}_{4,t} p^2) \leq 1, \quad \forall t \in [T] \tag{81}$$

on an event \mathcal{E} . Then, on \mathcal{E} ,

$$\sup_{\|g\|_{\infty} \leq 1} \Delta_{\text{local},t}(g) \leq K \left((\hat{\tau}_{4,t} + \hat{\tau}_{3,t}^2) p^2 + \hat{\tau}_{3,t} p + \hat{\tau}_{3,t}^3 r_{\text{LA}}^6 \right) \tag{82}$$

$$\Delta_{\text{local},t}(\mathbf{1}_{\Theta}) \leq K \left((\hat{\tau}_{4,t} + \hat{\tau}_{3,t}^2) p^2 + \hat{\tau}_{3,t}^3 r_{\text{LA}}^6 \right) \tag{83}$$

for all $t \in [T]$, where $K > 1$ is a universal constant.

Proof In this proof, we work on the event \mathcal{E} without explicitly referring to it. By the definitions,

$$f_t(u; \hat{\theta}_t) = \mathcal{R}_{t,3}(u) - \frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 = \mathcal{R}_{t,4}(u) + \frac{1}{6} \langle \nabla^3 \tilde{L}_t(\hat{\theta}_t), u^{\otimes 3} \rangle - \frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2.$$

It follows that

$$f_t(u; \hat{\theta}_t) + \frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 = \mathcal{R}_{t,3}(u) = \mathcal{R}_{t,4}(u) + \frac{1}{6} \langle \nabla^3 \tilde{L}_t(\hat{\theta}_t), u^{\otimes 3} \rangle. \quad (84)$$

Hence,

$$\begin{aligned} \Delta_{\text{local},t}(g) &= \left[\int \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \right) du \right]^{-1} \left| \int_{\Theta_{n,t}} g(u) \left[e^{f_t(\hat{\theta}_t, u)} - \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \right) \right] du \right| \\ &= \left| \int_{\Theta_{n,t}} g(u) \left[e^{\mathcal{R}_{t,3}(u)} - 1 \right] \frac{\exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \right)}{\int \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u' \right\|_2^2 \right) du'} du \right|. \end{aligned}$$

Let $\tilde{Z} = \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} Z$, where $Z \sim \mathcal{N}(0, \mathbf{I}_p)$. Let $\mathbb{E}_{\tilde{Z}}$ be the corresponding expectation for \tilde{Z} . With slight abuse of notations, to simplify the notations, we further denote $\mathbb{E}_A(\cdot) = \mathbb{E}(\cdot \mathbf{1}\{\tilde{Z} \in A\})$ for a measurable set $A \subseteq \Theta$. Then, the right-hand side of the last display is equal to

$$\begin{aligned} & \left| \mathbb{E}_{\Theta_{n,t}} \left[g(\tilde{Z}) \left(e^{\mathcal{R}_{t,3}(\tilde{Z})} - 1 \right) \right] \right| \\ &= \left| e^{\mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,3}(\tilde{Z})} \mathbb{E}_{\Theta_{n,t}} \left[g(\tilde{Z}) e^{\mathcal{R}_{t,3}(\tilde{Z}) - \mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,3}(\tilde{Z})} \right] - \mathbb{E}_{\Theta_{n,t}} \left[g(\tilde{Z}) \right] \right| \\ &\leq \left| e^{\mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,3}(\tilde{Z})} \mathbb{E}_{\Theta_{n,t}} \left[g(\tilde{Z}) \left(e^{\mathcal{R}_{t,3}(\tilde{Z}) - \mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,3}(\tilde{Z})} - 1 \right) \right] \right| \\ &\quad + \left| \left(e^{\mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,3}(\tilde{Z})} - 1 \right) \mathbb{E}_{\Theta_{n,t}} \left[g(\tilde{Z}) \right] \right| \\ &= \left| e^{\mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,3}(\tilde{Z})} \mathbb{E}_{\Theta_{n,t}} \left[g(\tilde{Z}) \left(e^{\bar{\mathcal{R}}_{t,3}(\tilde{Z})} - 1 \right) \right] \right| \\ &\quad + \left| \left(e^{\mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,3}(\tilde{Z})} - 1 \right) \mathbb{E}_{\Theta_{n,t}} \left[g(\tilde{Z}) \right] \right| \\ &\leq \left| e^{\mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,3}(\tilde{Z})} \mathbb{E}_{\Theta_{n,t}} \left[g(\tilde{Z}) \left(e^{\bar{\mathcal{R}}_{t,3}(\tilde{Z})} - 1 \right) \right] \right| + \left| e^{\mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,3}(\tilde{Z})} - 1 \right|, \end{aligned} \quad (85)$$

where $\bar{\mathcal{R}}_{t,3}(\cdot) = \mathcal{R}_{t,3}(\cdot) - \mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,3}(\tilde{Z})$, and the last inequality holds by $\|g\|_\infty \leq 1$. To prove (82), therefore, we only need to bound the following quantities:

$$(i) = \mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,3}(\tilde{Z}), \quad (ii) = \mathbb{E}_{\Theta_{n,t}} \left[g(\tilde{Z}) \left(e^{\bar{\mathcal{R}}_{t,3}(\tilde{Z})} - 1 \right) \right].$$

Firstly, we will obtain an upper bound of (i). Note that

$$\mathbb{E}_{\Theta_{n,t}} \left[\langle \nabla^3 \tilde{L}_t(\hat{\theta}_t), \tilde{Z}^{\otimes 3} \rangle \right] = 0 \quad (86)$$

by the symmetry of $\Theta_{n,t}$. It follows that

$$\begin{aligned} (i) &= \mathbb{E}_{\Theta_{n,t}} \left[\mathcal{R}_{3,t}(\tilde{Z}) \right] \stackrel{(80)}{=} \mathbb{E}_{\Theta_{n,t}} \left[\frac{1}{6} \langle \nabla^3 \tilde{L}_t(\hat{\theta}_t), \tilde{Z}^{\otimes 3} \rangle + \mathcal{R}_{4,t}(\tilde{Z}) \right] \\ &\stackrel{(86)}{=} \mathbb{E}_{\Theta_{n,t}} \left[\mathcal{R}_{4,t}(\tilde{Z}) \right] \leq \sqrt{\mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{4,t}^2(\tilde{Z})}. \end{aligned} \quad (87)$$

By Taylor's theorem, for $u \in \Theta_{n,t}$, there exists $\tilde{u} \in \Theta_{n,t}$ such that

$$\mathcal{R}_{4,t}(u) = \frac{1}{24} \langle \nabla^4 \tilde{L}_t(\hat{\theta}_t + \tilde{u}), u^{\otimes 4} \rangle \stackrel{(19)}{\leq} \frac{\hat{\tau}_{4,t}}{24} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^4.$$

Hence,

$$\begin{aligned} \mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{4,t}^2(\tilde{Z}) &\leq \mathbb{E}_{\Theta_{n,t}} \left[\frac{\hat{\tau}_{4,t}^2}{24^2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} \tilde{Z} \right\|_2^8 \right] = \mathbb{E}_{\Theta_{n,t}} \left[\frac{\hat{\tau}_{4,t}^2}{24^2} \|\tilde{Z}\|_2^8 \right] \\ &\leq \frac{\hat{\tau}_{4,t}^2}{24^2} \left[\text{tr}(\mathbf{I}_p) + 3\|\mathbf{I}_p\|_2 \right]^4 = \frac{\hat{\tau}_{4,t}^2}{24^2} (p+3)^4, \end{aligned} \quad (88)$$

where the second inequality holds by Lemma 51. Combining with (87), the last display implies that

$$(i) \leq \sqrt{\mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{4,t}^2(\tilde{Z})} \leq \frac{\hat{\tau}_{4,t}}{24} (p+3)^2. \quad (89)$$

Next, we will obtain an upper bound of |(ii)|. By Lemma 55 with $X = \bar{\mathcal{R}}_{t,3}(\tilde{Z})$ and $\epsilon = 8\hat{\tau}_{3,t}r_{\text{LA}}^2$, we have

$$\begin{aligned} \left| \mathbb{E}_{\Theta_{n,t}} \left[\left(e^{\bar{\mathcal{R}}_{t,3}(\tilde{Z})} - 1 - \bar{\mathcal{R}}_{t,3}(\tilde{Z}) - \frac{\bar{\mathcal{R}}_{t,3}^2(u)}{2} \right) g(\tilde{Z}) \right] \right| &\leq \frac{5}{3} (8\hat{\tau}_{3,t}r_{\text{LA}}^2)^3 e^{64\hat{\tau}_{3,t}^2 r_{\text{LA}}^4} \\ &\stackrel{(81)}{\leq} c_1 \hat{\tau}_{3,t}^3 r_{\text{LA}}^6, \end{aligned}$$

for some universal constant $c_1 > 0$. It follows that

$$\begin{aligned} &\left| \mathbb{E}_{\Theta_{n,t}} \left[g(\tilde{Z}) \left(e^{\bar{\mathcal{R}}_{t,3}(\tilde{Z})} - 1 \right) \right] \right| \\ &\leq \left| \mathbb{E}_{\Theta_{n,t}} \left[\left(\bar{\mathcal{R}}_{t,3}(\tilde{Z}) + \frac{\bar{\mathcal{R}}_{t,3}^2(\tilde{Z})}{2} \right) g(\tilde{Z}) \right] \right| + c_1 \hat{\tau}_{3,t}^3 r_{\text{LA}}^6 \\ &\leq \mathbb{E}_{\Theta_{n,t}} \left| \bar{\mathcal{R}}_{t,3}(\tilde{Z}) g(\tilde{Z}) \right| + \frac{1}{2} \mathbb{E}_{\Theta_{n,t}} \left| \bar{\mathcal{R}}_{t,3}^2(\tilde{Z}) g(\tilde{Z}) \right| + c_1 \hat{\tau}_{3,t}^3 r_{\text{LA}}^6 \\ &\leq \mathbb{E}_{\Theta_{n,t}} \left| \bar{\mathcal{R}}_{t,3}(\tilde{Z}) \right| + \frac{1}{2} \mathbb{E}_{\Theta_{n,t}} \left[\bar{\mathcal{R}}_{t,3}^2(\tilde{Z}) \right] + c_1 \hat{\tau}_{3,t}^3 r_{\text{LA}}^6, \end{aligned}$$

where the last inequality holds by $\|g\|_\infty \leq 1$. Let

$$(iii) = \mathbb{E}_{\Theta_{n,t}} \left| \bar{\mathcal{R}}_{t,3}(\tilde{Z}) \right|, \quad (iv) = \mathbb{E}_{\Theta_{n,t}} \left[\bar{\mathcal{R}}_{t,3}^2(\tilde{Z}) \right].$$

To bound |(ii)|, we need to obtain upper bounds for (iii) and (iv). Firstly, we bound (iii). Let $\bar{\mathcal{R}}_{t,4}(\cdot) = \mathcal{R}_{t,4}(\cdot) - \mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,4}(\tilde{Z})$. By (84) and (86), we have

$$\mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,4}(\tilde{Z}) = \mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,3}(\tilde{Z}),$$

which implies that, for every $u \in \Theta$,

$$\overline{\mathcal{R}}_{t,3}(u) - \frac{1}{6} \langle \nabla^3 \tilde{L}_t(\hat{\theta}_t), u^{\otimes 3} \rangle = \overline{\mathcal{R}}_{t,4}(u). \quad (90)$$

It follows that

$$\mathbb{E}_{\Theta_{n,t}} \left| \overline{\mathcal{R}}_{t,3}(\tilde{Z}) - \frac{1}{6} \langle \nabla^3 \tilde{L}_t(\hat{\theta}_t), \tilde{Z}^{\otimes 3} \rangle \right| = \mathbb{E}_{\Theta_{n,t}} \left| \overline{\mathcal{R}}_{t,4}(\tilde{Z}) \right|,$$

which further implies that

$$(iii) \leq \mathbb{E}_{\Theta_{n,t}} \left| \overline{\mathcal{R}}_{t,4}(\tilde{Z}) \right| + \mathbb{E}_{\Theta_{n,t}} \left| \frac{1}{6} \langle \nabla^3 \tilde{L}_t(\hat{\theta}_t), \tilde{Z}^{\otimes 3} \rangle \right|.$$

Note that

$$\mathbb{E}_{\Theta_{n,t}} \left| \overline{\mathcal{R}}_{t,4}(\tilde{Z}) \right| \leq \sqrt{\mathbb{E}_{\Theta_{n,t}} \overline{\mathcal{R}}_{t,4}^2(\tilde{Z})} \leq \sqrt{\mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,4}^2(\tilde{Z})} \stackrel{(88)}{\leq} \frac{\hat{\tau}_{4,t}}{24} (p+3)^2.$$

Also,

$$\begin{aligned} \mathbb{E}_{\Theta_{n,t}} \left| \frac{1}{6} \langle \nabla^3 \tilde{L}_t(\hat{\theta}_t), \tilde{Z}^{\otimes 3} \rangle \right| &\leq \frac{1}{6} \sqrt{\mathbb{E}_{\Theta_{n,t}} \left| \langle \nabla^3 \tilde{L}_t(\hat{\theta}_t), \tilde{Z}^{\otimes 3} \rangle \right|^2} \\ &\stackrel{\text{Lemma 54}}{\leq} \frac{1}{6} \sqrt{15 \hat{\tau}_{3,t}^2 \|\mathbf{I}_p\|_2 \text{tr}^2(\mathbf{I}_p)} = \frac{1}{6} \sqrt{15 \hat{\tau}_{3,t}^2 p^2} = \frac{\sqrt{15}}{6} \hat{\tau}_{3,t} p. \end{aligned} \quad (91)$$

Combining the last three displays, we have

$$(iii) \leq \frac{\hat{\tau}_{4,t}}{24} (p+3)^2 + \frac{\sqrt{15}}{6} \hat{\tau}_{3,t} p.$$

To bound (iv), note that

$$\begin{aligned} (iv) &= \mathbb{E}_{\Theta_{n,t}} \left[\overline{\mathcal{R}}_{t,3}^2(\tilde{Z}) \right] \stackrel{(90)}{=} \mathbb{E}_{\Theta_{n,t}} \left[\left(\frac{1}{6} \langle \nabla^3 \tilde{L}_t(\hat{\theta}_t), \tilde{Z}^{\otimes 3} \rangle + \overline{\mathcal{R}}_{t,4}(\tilde{Z}) \right)^2 \right] \\ &\leq \frac{1}{18} \mathbb{E}_{\Theta_{n,t}} \left[\langle \nabla^3 \tilde{L}_t(\hat{\theta}_t), \tilde{Z}^{\otimes 3} \rangle^2 \right] + 2 \mathbb{E}_{\Theta_{n,t}} \left[\overline{\mathcal{R}}_{t,4}^2(\tilde{Z}) \right] \\ &\stackrel{(88)}{\leq} \frac{15}{18} \hat{\tau}_{3,t}^2 p^2 + 2 \frac{\hat{\tau}_{4,t}^2}{24^2} (p+3)^4 \leq \frac{5}{6} \hat{\tau}_{3,t}^2 p^2 + \frac{8}{9} \hat{\tau}_{4,t}^2 p^4, \end{aligned} \quad (92)$$

where the last inequality holds by $(p+3)^4 \leq 2^8 p^4$. Combining the last two displays, we have

$$\begin{aligned} |(ii)| &\leq |(iii)| + \frac{1}{2} |(iv)| + c_1 \hat{\tau}_{3,t}^3 r_{\text{LA}}^6 \\ &\leq \frac{\hat{\tau}_{4,t}}{24} (p+3)^2 + \frac{\sqrt{15}}{6} \hat{\tau}_{3,t} p + \frac{5}{12} \hat{\tau}_{3,t}^2 p^2 + \frac{4}{9} \hat{\tau}_{4,t}^2 p^4 + c_1 \hat{\tau}_{3,t}^3 r_{\text{LA}}^6 \\ &\leq c_2 \left[(\hat{\tau}_{3,t}^2 + \hat{\tau}_{4,t}) p^2 + \hat{\tau}_{3,t} p + \hat{\tau}_{3,t}^3 r_{\text{LA}}^6 \right] \end{aligned} \quad (93)$$

for some universal constant $c_2 = c_2(c_1) > 0$, where the last inequality holds by $\widehat{\tau}_{4,t}p^2 \leq 1$.

Now, we are ready to prove (82). By (89) and (93), the right-hand side of (85) is bounded by

$$\begin{aligned}
 & \left| e^{\widehat{\tau}_{4,t}(p+3)^2/24} \mathbb{E}_{\Theta_{n,t}} \left[g(\tilde{Z}) \left(e^{\overline{\mathcal{R}}_{t,3}(\tilde{Z})} - 1 \right) \right] \right| + \left| e^{\widehat{\tau}_{4,t}(p+3)^2/24} - 1 \right| \\
 & \leq \left| e^{2\widehat{\tau}_{4,t}p^2/3} \mathbb{E}_{\Theta_{n,t}} \left[g(\tilde{Z}) \left(e^{\overline{\mathcal{R}}_{t,3}(\tilde{Z})} - 1 \right) \right] \right| + \left| e^{2\widehat{\tau}_{4,t}p^2/3} - 1 \right| \\
 & \lesssim \left| \mathbb{E}_{\Theta_{n,t}} \left[g(\tilde{Z}) \left(e^{\overline{\mathcal{R}}_{t,3}(\tilde{Z})} - 1 \right) \right] \right| + \widehat{\tau}_{4,t}p^2 \\
 & \stackrel{(93)}{\leq} c_2 \left[(\widehat{\tau}_{3,t}^2 + \widehat{\tau}_{4,t}) p^2 + \widehat{\tau}_{3,t}p + \widehat{\tau}_{3,t}^3 r_{\text{LA}}^6 \right] + \widehat{\tau}_{4,t}p^2 \\
 & \leq c_3 \left[(\widehat{\tau}_{3,t}^2 + \widehat{\tau}_{4,t}) p^2 + \widehat{\tau}_{3,t}p + \widehat{\tau}_{3,t}^3 r_{\text{LA}}^6 \right]
 \end{aligned}$$

for some positive constants $c_3 = c_3(c_2)$, where the second inequality holds because $\widehat{\tau}_{4,t}p^2 \leq 1$ and $e^x \leq 1 + 2x$ for $x \leq 1.256$. Therefore, we have

$$\sup_{\|g\|_\infty \leq 1} \Delta_{\text{local},t}(g) \leq c_3 \left[\widehat{\tau}_{4,t}p^4 + \widehat{\tau}_{3,t}\widehat{\tau}_{4,t}p^3 + (\widehat{\tau}_{3,t}^2 + \widehat{\tau}_{4,t}) p^2 + \widehat{\tau}_{3,t}p + \widehat{\tau}_{3,t}^3 r_{\text{LA}}^6 \right],$$

which completes the proof of (82).

The proof of (83) is similar. However, there are some differences in obtaining the upper bound of (ii). Consider the case where $g(\cdot)$ is symmetric, meaning $g(u) = g(-u)$ for any $u \in \Theta_{n,t}$. Note that

$$\begin{aligned}
 & \left| \mathbb{E}_{\Theta_{n,t}} \left[g(\tilde{Z}) \left(e^{\overline{\mathcal{R}}_{t,3}(\tilde{Z})} - 1 \right) \right] \right| \\
 & \leq \left| \mathbb{E}_{\Theta_{n,t}} \left[\left(\overline{\mathcal{R}}_{t,3}(\tilde{Z}) + \frac{\overline{\mathcal{R}}_{t,3}^2(\tilde{Z})}{2} \right) g(\tilde{Z}) \right] \right| + c_1 \widehat{\tau}_{3,t}^3 r_{\text{LA}}^6 \\
 & \stackrel{(90)}{=} \left| \mathbb{E}_{\Theta_{n,t}} \left[\frac{1}{6} \langle \nabla^3 \tilde{L}_t(\hat{\theta}_t), \tilde{Z}^{\otimes 3} \rangle g(\tilde{Z}) + \overline{\mathcal{R}}_{t,4}(\tilde{Z}) g(\tilde{Z}) + \frac{\overline{\mathcal{R}}_{t,3}^2(\tilde{Z})}{2} g(\tilde{Z}) \right] \right| + c_1 \widehat{\tau}_{3,t}^3 r_{\text{LA}}^6 \\
 & = \left| \mathbb{E}_{\Theta_{n,t}} \left[\overline{\mathcal{R}}_{t,4}(\tilde{Z}) g(\tilde{Z}) + \frac{\overline{\mathcal{R}}_{t,3}^2(\tilde{Z})}{2} g(\tilde{Z}) \right] \right| + c_1 \widehat{\tau}_{3,t}^3 r_{\text{LA}}^6 \tag{94} \\
 & \leq \mathbb{E}_{\Theta_{n,t}} \left| \overline{\mathcal{R}}_{t,4}(\tilde{Z}) \right| + \frac{1}{2} \mathbb{E}_{\Theta_{n,t}} \overline{\mathcal{R}}_{t,3}^2(\tilde{Z}) + c_1 \widehat{\tau}_{3,t}^3 r_{\text{LA}}^6 \\
 & \leq \sqrt{\mathbb{E}_{\Theta_{n,t}} \overline{\mathcal{R}}_{t,4}^2(\tilde{Z})} + \frac{1}{2} \mathbb{E}_{\Theta_{n,t}} \overline{\mathcal{R}}_{t,3}^2(\tilde{Z}) + c_1 \widehat{\tau}_{3,t}^3 r_{\text{LA}}^6 \\
 & \stackrel{(88)}{\lesssim} \widehat{\tau}_{4,t}p^2 + \widehat{\tau}_{3,t}^2 p^2 + c_1 \widehat{\tau}_{3,t}^3 r_{\text{LA}}^6 \leq c_4 \left[(\widehat{\tau}_{3,t}^2 + \widehat{\tau}_{4,t}) p^2 + \widehat{\tau}_{3,t}^3 r_{\text{LA}}^6 \right] \tag{92}
 \end{aligned}$$

for some universal constant $c_4 = c_4(c_1) > 0$, where the second equality holds by the symmetry of $\Theta_{n,t}$ and $g(\cdot)$. For any symmetric $g(\cdot)$, therefore, we have

$$\begin{aligned}
 \Delta_{\text{local},t}(g) &\stackrel{(85)}{\leq} \left| e^{\mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,3}(\tilde{Z})} \left| \mathbb{E}_{\Theta_{n,t}} \left[g(\tilde{Z}) \left(e^{\bar{\mathcal{R}}_{t,3}(\tilde{Z})} - 1 \right) \right] \right| \right| + \left| e^{\mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,3}(\tilde{Z})} - 1 \right| \\
 &\stackrel{(89)}{\leq} \left| e^{2\hat{\tau}_{4,t}p^2/3} \left| \mathbb{E}_{\Theta_{n,t}} \left[g(\tilde{Z}) \left(e^{\bar{\mathcal{R}}_{t,3}(\tilde{Z})} - 1 \right) \right] \right| \right| + \left| e^{2\hat{\tau}_{4,t}p^2/3} - 1 \right| \\
 &\stackrel{(94)}{\lesssim} c_4 \left[(\hat{\tau}_{3,t}^2 + \hat{\tau}_{4,t}) p^2 + \hat{\tau}_{3,t}^3 r_{\text{LA}}^6 \right] + \hat{\tau}_{4,t} p^2 \leq c_5 \left[(\hat{\tau}_{3,t}^2 + \hat{\tau}_{4,t}) p^2 + \hat{\tau}_{3,t}^3 r_{\text{LA}}^6 \right]
 \end{aligned}$$

for some universal constants $c_5 = c_5(c_4) > 0$. This completes the proof of (83). \blacksquare

B.2 Proofs of Theorem 2, 3

Proof [Proof of Theorem 2] Since the total variation distance is bounded by 2, note that the inequality in (21) always holds provided that $(\hat{\tau}_{3,t} r_{\text{LA}}^2) \vee (\hat{\tau}_{4,t} p^2) \geq C$ for some universal constant $C > 0$ and $K = K(C)$ is large enough. Hence, we may assume that

$$(\hat{\tau}_{3,t} r_{\text{LA}}^2) \vee (\hat{\tau}_{4,t} p^2) \leq \delta, \quad \forall t \in [T] \quad (95)$$

for some small enough constant δ on an event \mathcal{E}_1 . In this proof, we work on the event \mathcal{E}_1 without explicitly referring to it. By (95) and $r_{\text{LA}} \geq 1$, we have

$$\hat{\tau}_{3,t} r_{\text{LA}} \leq \hat{\tau}_{3,t} r_{\text{LA}}^2 \leq \delta \leq 1/8, \quad N \geq 2,$$

which allows us to utilize Lemma 13. By Lemma 13, we have

$$\begin{aligned}
 \sup_{\|g\|_\infty \leq 1} \Delta_{\text{tail}, \tilde{\Pi}, t}(g) &\leq 2e^{-8 \log N - 8p}, \\
 \sup_{\|g\|_\infty \leq 1} \Delta_{\text{tail}, \text{LA}, t}(g) &\leq e^{-16 \log N - 49p/2},
 \end{aligned}$$

which, combining with $p \geq 1$, further implies that

$$\Delta_{\text{tail}, \tilde{\Pi}, t}(\mathbf{1}_\Theta) \vee \Delta_{\text{tail}, \text{LA}, t}(\mathbf{1}_\Theta) \leq 1/4.$$

Also, by (95), Lemma 15 implies that

$$\Delta_{\text{local},t}(\mathbf{1}_\Theta) \leq K \left([\hat{\tau}_{4,t} + \hat{\tau}_{3,t}^2] p^2 + \hat{\tau}_{3,t}^3 r_{\text{LA}}^6 \right) \stackrel{(95)}{\leq} 5K_1 \delta \leq \frac{1}{4},$$

where K_1 denotes the constant K in Lemma 15 and $\delta = \delta(K_1)$ is small enough constant. By Lemma 12, it follows that

$$\begin{aligned}
 d_V \left(\Pi_t^{\text{LA}}(\cdot), \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right) &= \sup_{\|g\|_\infty \leq 1} \left| \mathcal{I}_{\tilde{\Pi}_t}(g) - \mathcal{I}_{\text{LA},t}(g) \right| \\
 &\stackrel{\text{Lemma 12}}{\leq} 2 \sup_{\|g\|_\infty \leq 1} \left\{ \Delta_{\text{local},t}(g) + \Delta_{\text{local},t}(\mathbf{1}_\Theta) + 2\Delta_{\text{tail},t}(g) + 2\Delta_{\text{tail},t}(\mathbf{1}_\Theta) \right\} \\
 &\stackrel{\text{Lemma 13}}{\leq} \sup_{\|g\|_\infty \leq 1} \left\{ 2\Delta_{\text{local},t}(g) \right\} + 2\Delta_{\text{local},t}(\mathbf{1}_\Theta) + 16e^{-8 \log N - 8p} \\
 &\stackrel{\text{Lemma 15}}{\leq} 4K_1 \left([\hat{\tau}_{4,t} + \hat{\tau}_{3,t}^2] p^2 + \hat{\tau}_{3,t} p + \hat{\tau}_{3,t}^3 r_{\text{LA}}^6 \right) + 16e^{-8 \log N - 8p} \\
 &\leq (4K_1 \vee 16) \left([\hat{\tau}_{4,t} + \hat{\tau}_{3,t}^2] p^2 + \hat{\tau}_{3,t} p + \hat{\tau}_{3,t}^3 r_{\text{LA}}^6 + e^{-8 \log N - 8p} \right) \\
 &\lesssim (4K_1 \vee 16) \left([\hat{\tau}_{4,t} + \hat{\tau}_{3,t}^2] p^2 + \hat{\tau}_{3,t} p + \hat{\tau}_{3,t}^3 \log^3 N + e^{-8 \log N - 8p} \right),
 \end{aligned}$$

which completes the proof. \blacksquare

Remark 16 *The indicator function $\mathbf{1}_\Theta$ in (83) can be replaced by any symmetric function g*

Proof [Proof of Theorem 3] In this proof, we will work on the event \mathcal{E}_1 without explicitly referring to it. Let $\tilde{Z} = \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} Z$, where $Z \sim \mathcal{N}(0, \mathbf{I}_p)$, and $\mathbb{E}_{\tilde{Z}}$ be the corresponding expectation. As in the proof of Lemma 15, we also use the notation $\mathbb{E}_A(\cdot) = \mathbb{E}_{\tilde{Z}}(\cdot \mathbf{1}\{\tilde{Z} \in A\})$ for a measurable set $\mathbf{A} \subseteq \Theta$.

For $u \in \Theta$, note that

$$\begin{aligned}
 &\log(\pi_t^{\text{LA}}(\hat{\theta}_t + u) / \tilde{\pi}_t(\hat{\theta}_t + u \mid \mathbf{D}_t)) \\
 &= -\log \left[\frac{e^{\tilde{L}_t(\hat{\theta}_t + u)}}{\int e^{\tilde{L}_t(\hat{\theta}_t + u')} \mathrm{d}u'} \right] - \frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} (\hat{\theta}_t + u - \hat{\theta}_t) \right\|_2^2 \\
 &\quad - \log \left[\int \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} (\hat{\theta}_t + u - \hat{\theta}_t) \right\|_2^2 \right) \mathrm{d}u \right] \\
 &\stackrel{(60)}{=} -\log \left[\frac{e^{f_t(u; \hat{\theta}_t)}}{\int e^{f_t(u'; \hat{\theta}_t)} \mathrm{d}u'} \right] - \frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 - \log \left[\int \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 \right) \mathrm{d}u \right] \\
 &= -f_t(u; \hat{\theta}_t) - \frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 + \log \left[\int e^{f_t(u; \hat{\theta}_t)} \mathrm{d}u \right] - \log \left[\int \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 \right) \mathrm{d}u \right] \\
 &\stackrel{(80)}{=} -\mathcal{R}_{t,3}(\hat{\theta}_t, u) + \log \left[\int e^{f_t(u; \hat{\theta}_t)} \mathrm{d}u \right] - \log \left[\int \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 \right) \mathrm{d}u \right] \\
 &= -\mathcal{R}_{t,3}(\hat{\theta}_t, u) + W_{n,t},
 \end{aligned}$$

where

$$W_{n,t} = \log \left[\int e^{f_t(u; \hat{\theta}_t)} du \right] - \log \left[\int \exp \left(-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \right) du \right].$$

Hence,

$$\begin{aligned} K \left(\Pi_t^{\text{LA}}(\cdot); \tilde{\Pi}_t(\cdot | \mathbf{D}_t) \right) &= \int \log \left[\frac{\pi_t^{\text{LA}}(\hat{\theta}_t + u)}{\tilde{\pi}_t(\hat{\theta}_t + u | \mathbf{D}_t)} \right] \pi_t^{\text{LA}}(\hat{\theta}_t + u) du \\ &= \mathbb{E}_{\tilde{Z}} \left[-\mathcal{R}_{t,3}(\tilde{Z}) \right] + W_{n,t}. \end{aligned}$$

We will obtain upper bounds for the following quantities:

$$(i) = \mathbb{E}_{\tilde{Z}} \left[-\mathcal{R}_{t,3}(\tilde{Z}) \right], \quad (ii) = W_{n,t}.$$

Firstly, we will obtain an upper bound of (i). Note that

$$\begin{aligned} \mathbb{E}_{\tilde{Z}} \left[-\mathcal{R}_{t,3}(\tilde{Z}) \right] &= \mathbb{E}_{\Theta_{n,t}} \left[-\mathcal{R}_{t,3}(\tilde{Z}) \right] + \mathbb{E}_{\Theta_{n,t}^c} \left[-\mathcal{R}_{t,3}(\tilde{Z}) \right] \\ &\stackrel{(84)}{=} \mathbb{E}_{\Theta_{n,t}} \left[-\mathcal{R}_{t,4}(\tilde{Z}) \right] + \mathbb{E}_{\Theta_{n,t}^c} \left[-\mathcal{R}_{t,3}(\tilde{Z}) \right] \end{aligned}$$

For the first term in the right-hand side of the last display, note that

$$\mathbb{E}_{\Theta_{n,t}} \left[-\mathcal{R}_{t,4}(\tilde{Z}) \right] \leq \mathbb{E}_{\Theta_{n,t}} \left| \mathcal{R}_{t,4}(\tilde{Z}) \right| \leq \sqrt{\mathbb{E}_{\Theta_{n,t}} \mathcal{R}_{t,4}^2(\tilde{Z})} \stackrel{(89)}{\leq} \frac{1}{24} \hat{\tau}_{4,t} (p+3)^2. \quad (96)$$

Also, for $u \in \Theta_{n,t}^c$, we have

$$\begin{aligned} -\mathcal{R}_{t,3}(u) &= - \left[\tilde{L}_t(\hat{\theta}_t + u) - \tilde{L}_t(\hat{\theta}_t) - \langle \nabla \tilde{L}_t(\hat{\theta}_t), u \rangle + \frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \right] \\ &= \frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \theta_t^\circ}^{1/2} u \right\|_2^2 - \frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \leq \frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \theta_t^\circ}^{1/2} u \right\|_2^2 \end{aligned}$$

for some $\theta_t^\circ = \theta_t^\circ(\hat{\theta}_t, u) \in \Theta(\hat{\theta}_t, \tilde{\mathbf{F}}_{t, \hat{\theta}_t}, \|\tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u\|_2)$ by Taylor's theorem. Note that θ_t° may not be located in $\Theta(\hat{\theta}_t, \tilde{\mathbf{F}}_{t, \hat{\theta}_t}, 4r_{\text{LA}})$. Let $r = \|\tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u\|_2 > 4r_{\text{LA}} > 1$. Then,

$$\begin{aligned} \frac{1}{2} \left\| \tilde{\mathbf{F}}_{t, \theta_t^\circ}^{1/2} u \right\|_2^2 &= \frac{1}{2} u^\top \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} \left(\mathbf{I}_p + \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} \tilde{\mathbf{F}}_{t, \theta_t^\circ} \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right) \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \\ &\leq \frac{1}{2} \left(1 + \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} \tilde{\mathbf{F}}_{t, \theta_t^\circ} \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_2 \right) \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \\ &\stackrel{\text{Lemma 40}}{\leq} \frac{1}{2} \left(1 + \hat{\tau}_{3,t,r} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2 \right) \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^2 \\ &\leq (1 \vee \hat{\tau}_{3,t,r}) \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2^3 = (1 \vee \hat{\tau}_{3,t,r}) \exp \left(3 \log \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2 \right) \\ &\leq (1 \vee \hat{\tau}_{3,t,r}) \exp \left(3 \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2 \right) \\ &\leq N e^{8p} \exp \left(\left(\sqrt{p} + \sqrt{2 \log N} - 3 \right) \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2 \right) \exp \left(3 \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} u \right\|_2 \right) \end{aligned}$$

where the last inequality holds by **(KL)**. It follows that

$$\begin{aligned}
 \mathbb{E}_{\Theta_{n,t}^c} \left[-\mathcal{R}_{t,3}(\tilde{Z}) \right] &\leq \mathbb{E}_{\Theta_{n,t}^c} \left[\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} \tilde{Z} \right\|_2^2 \right] \\
 &\leq N e^{8p} \mathbb{E}_{\Theta_{n,t}^c} \left[\exp \left\{ \left(\sqrt{p} + \sqrt{2 \log N} - 3 \right) \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} \tilde{Z} \right\|_2 \right\} \exp \left(3 \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} \tilde{Z} \right\|_2 \right) \right] \\
 &= N e^{8p} \mathbb{E}_{\Theta_{n,t}^c} \left[\exp \left(\left[\sqrt{p} + \sqrt{2 \log N} \right] \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} \tilde{Z} \right\|_2 \right) \right] \\
 &\stackrel{\text{Lemma 14}}{\leq} N e^{8p} e^{-8p-8 \log N} = e^{-7 \log N}.
 \end{aligned} \tag{97}$$

Therefore, (96) and (97) imply that

$$\text{(i)} = \mathbb{E}_{\Theta_{n,t}} \left[-\mathcal{R}_{t,4}(\tilde{Z}) \right] + \mathbb{E}_{\Theta_{n,t}^c} \left[-\mathcal{R}_{t,3}(\tilde{Z}) \right] \leq \frac{1}{24} \hat{\tau}_{4,t} (p+3)^2 + e^{-7 \log N} \tag{98}$$

Next, we will obtain an upper bound of (ii). Note that

$$\begin{aligned}
 |e^{W_{n,t}} - 1| &= \left| \frac{\int \left(e^{f_t(u;\hat{\theta}_t)} - \exp \left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 \right] \right) du}{\int \exp \left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 \right] du} \right| \\
 &\leq \left| \frac{\int_{\Theta_{n,t}} \left(e^{f_t(u;\hat{\theta}_t)} - \exp \left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 \right] \right) du}{\int \exp \left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 \right] du} \right| + \left| \frac{\int_{\Theta_{n,t}^c} e^{f_t(u;\hat{\theta}_t)} du}{\int \exp \left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 \right] du} \right| \\
 &\quad + \left| \frac{\int_{\Theta_{n,t}^c} \exp \left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 \right] du}{\int \exp \left[-\frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} u \right\|_2^2 \right] du} \right| \\
 &\leq \Delta_{\text{local},t}(\mathbf{1}_\Theta) + \Delta_{\text{tail},\tilde{\Pi},t}(\mathbf{1}_\Theta) + \Delta_{\text{tail,LA},t}(\mathbf{1}_\Theta) \\
 &\leq K_1 \left([\hat{\tau}_{4,t} + \hat{\tau}_{3,t}^2] p^2 + \hat{\tau}_{3,t}^3 r_{\text{LA}}^6 \right) + 3e^{-8 \log N - 8p},
 \end{aligned}$$

where the last inequality holds by Lemmas 13 and 15, and K_1 denotes the constant K in Lemma 15. By $1 + x \leq e^x$ for $x \in \mathbb{R}$, it follows that

$$\text{(ii)} = W_{n,t} \leq K_1 \left([\hat{\tau}_{4,t} + \hat{\tau}_{3,t}^2] p^2 + \hat{\tau}_{3,t}^3 r_{\text{LA}}^6 \right) + 3e^{-8 \log N - 8p}. \tag{99}$$

By (98) and (99), therefore, we have

$$\begin{aligned}
 & K \left(\Pi_t^{\text{LA}}(\cdot); \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right) \\
 & \leq \frac{1}{24} \hat{\tau}_{4,t} (p+3)^2 + e^{-7 \log N} + K_1 \left([\hat{\tau}_{4,t} + \hat{\tau}_{3,t}^2] p^2 + \hat{\tau}_{3,t}^3 r_{\text{LA}}^6 \right) + 3e^{-8 \log N - 8p} \\
 & \leq K_2 \left([\hat{\tau}_{4,t} + \hat{\tau}_{3,t}^2] p^2 + \hat{\tau}_{3,t}^3 r_{\text{LA}}^6 + e^{-7 \log N} \right) \\
 & \lesssim K_2 \left([\hat{\tau}_{4,t} + \hat{\tau}_{3,t}^2] p^2 + \hat{\tau}_{3,t}^3 \log^3 N + e^{-7 \log N} \right),
 \end{aligned}$$

for some universal constant $K_2 = K_2(K_1) > 0$. This completes the proof. \blacksquare

B.3 Proof of Theorem 4

Proof [Proof of Theorem 4] In this proof, we will work on the event \mathcal{E}_1 without explicitly referring to it. By Theorems 2 and 3, we have

$$d_V \left(\Pi_t^{\text{LA}}(\cdot), \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right) \leq K_1 \epsilon_{n,t,\text{TV}}, \quad K \left(\Pi_t^{\text{LA}}(\cdot), \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right) \leq K_2 \epsilon_{n,t,\text{KL}}^2,$$

where K_1 and K_2 denote the constant K in Theorems 2 and 3, respectively. By the definition of $\Pi_t(\cdot)$, it follows that

$$K \left(\Pi_t(\cdot); \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right) \leq K \left(\Pi_t^{\text{LA}}(\cdot); \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right) \leq K_2 \epsilon_{n,t,\text{KL}}^2.$$

By Pinsker's inequality, we have

$$\begin{aligned}
 d_V \left(\Pi_t(\cdot), \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right) & \leq \sqrt{\frac{1}{2} K \left(\Pi_t(\cdot), \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right)} \leq \sqrt{\frac{K_2}{2}} \epsilon_{n,t,\text{KL}}, \\
 d_V \left(\Pi_t^{\text{LA}}(\cdot), \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right) & \leq \sqrt{\frac{1}{2} K \left(\Pi_t^{\text{LA}}(\cdot), \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right)} \leq \sqrt{\frac{K_2}{2}} \epsilon_{n,t,\text{KL}}.
 \end{aligned}$$

By taking the constant K in Theorem 4 as $K = K_2 \vee \sqrt{K_2/2}$, we complete the proof. \blacksquare

For $t \in [T]$, let

$$\begin{aligned}
 \Delta_t = & \left\| \mathbf{\Omega}_t^{1/2} (\mu_t - \hat{\theta}_t) \right\|_2 \vee \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} (\mu_t - \hat{\theta}_t) \right\|_2 \\
 & \vee \left\| \mathbf{\Omega}_t^{-1/2} \tilde{\mathbf{F}}_{t,\hat{\theta}_t} \mathbf{\Omega}_t^{-1/2} - \mathbf{I}_p \right\|_{\text{F}} \vee \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \mathbf{\Omega}_t \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_{\text{F}}.
 \end{aligned} \tag{100}$$

Corollary 17 *Suppose that conditions in Theorem 4 hold. Also, assume that*

$$\epsilon_{n,t,\text{KL}} \leq (1200K)^{-1}, \quad \forall t \in [T]$$

on \mathcal{E}_1 defined in Theorem 4, where K is the universal constant specified in Theorem 4. Then, on \mathcal{E}_1 ,

$$\Delta_t \leq 400K \epsilon_{n,t,\text{KL}}, \quad \forall t \in [T].$$

Proof From the proof of Theorem 4, we have

$$\begin{aligned} d_V \left(\Pi_t(\cdot), \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right) &\leq K_1 \epsilon_{n,t,\text{KL}}, \\ d_V \left(\Pi_t^{\text{LA}}(\cdot), \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right) &\leq K_1 \epsilon_{n,t,\text{KL}}, \end{aligned}$$

where K_1 denotes the constant K in Theorem 4. It follows that

$$\begin{aligned} d_V \left(\Pi_t(\cdot), \Pi_t^{\text{LA}}(\cdot) \right) &\leq d_V \left(\Pi_t(\cdot), \tilde{\Pi}_t(\cdot \mid \mathbf{D}_t) \right) + d_V \left(\tilde{\Pi}_t(\cdot \mid \mathbf{D}_t), \Pi_t^{\text{LA}}(\cdot) \right) \\ &\leq K_1 \epsilon_{n,t,\text{KL}} + K_1 \epsilon_{n,t,\text{KL}} \leq 2K_1 \epsilon_{n,t,\text{KL}}. \end{aligned}$$

Since $2K_1 \epsilon_{n,t,\text{KL}} \leq 1/600$, we can apply Lemma 44. By Lemma 44 and symmetry of $d_V(\cdot, \cdot)$, therefore, we have

$$\frac{\Delta_t}{200} \leq d_V \left(\Pi_t(\cdot), \Pi_t^{\text{LA}}(\cdot) \right) \leq 2K_1 \epsilon_{n,t,\text{KL}},$$

which completes the proof. ■

Appendix C. Proofs for Section 5

Proof [Proof of Theorem 5] In this proof, we work on the event $\mathcal{E}_2 \cap \mathcal{E}_{\text{est},1}$ without explicitly referring to it. Let $t \in [T]$. For simplicity in notations, let $\Theta_{n,t} = \Theta(\theta_t^*, \tilde{\mathbf{F}}_{t,\theta_t^*}, 4r_{\text{eff},t})$, $\bar{\Theta}_{n,t} = \Theta_{n,t}(\tilde{\mathbf{F}}_{t,\theta_t^*}, 4r_{\text{eff},t})$ in this proof, and

$$\partial\Theta_{n,t} = \left\{ \theta \in \Theta : \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2}(\theta - \theta_t^*) \right\|_2 = 4r_{\text{eff},t} \right\}.$$

For any $\theta \in \Theta$, let

$$g_t(\theta) = \mathbb{E}_t \tilde{L}_t(\theta) + \langle \nabla \zeta_t, \theta \rangle = \mathbb{E}_t \tilde{L}_t(\theta) + \langle \nabla \tilde{L}_t(\hat{\theta}_t) - \nabla \mathbb{E}_t \tilde{L}_t(\hat{\theta}_t), \theta \rangle. \quad (101)$$

By the right-hand side of the last display and the strong concavity of $\theta \mapsto \mathbb{E}_t \tilde{L}_t(\theta)$, note that $\nabla g_t(\hat{\theta}_t) = 0$ and $\theta \mapsto g_t(\theta)$ is concave. Hence, we only need to prove that the first-order stationary point $\hat{\theta}_t$ of $g_t(\cdot)$ is located in $\Theta_{n,t}$.

By the concavity of $g_t(\cdot)$, for any $\theta \in \Theta_{n,t}^c$, we have

$$g_t(\bar{\theta}) \geq \omega g_t(\theta) + (1 - \omega) g_t(\theta_t^*), \quad (102)$$

where $\bar{\theta} = \omega\theta + (1 - \omega)\theta_t^*$ and $\omega = 4r_{\text{eff},t} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2}(\theta - \theta_t^*) \right\|_2^{-1} \in (0, 1)$. At the end of this proof, we will show that

$$\sup_{\theta^\circ \in \partial\Theta_{n,t}} g_t(\theta^\circ) - g_t(\theta_t^*) \leq -2r_{\text{eff},t}^2 < 0. \quad (103)$$

It follows that, for any $\theta \in \Theta_{n,t}^c$,

$$0 > -2r_{\text{eff},t}^2 \geq \sup_{\theta^\circ \in \partial\Theta_{n,t}} g_t(\theta^\circ) - g_t(\theta_t^*) \stackrel{(102)}{\geq} \omega \left[g_t(\theta) - g_t(\theta_t^*) \right] \geq g_t(\theta) - g_t(\theta_t^*),$$

which implies that $\hat{\theta}_t \in \Theta_{n,t}$.

To complete the proof, we only need to prove (103). Let $\theta^\circ \in \partial\Theta_{n,t}$ and $u = \theta^\circ - \theta_t^*$. By Taylor's theorem, there exists some $\tilde{u} \in \Theta_{n,t}$ such that, on $\mathcal{E}_{\text{est},1}$,

$$\begin{aligned}
 g_t(\theta^\circ) - g_t(\theta_t^*) &= \nabla g_t(\theta_t^*)^\top u + \frac{1}{2} \langle \nabla^2 g_t(\theta_t^* + \tilde{u}), u^{\otimes 2} \rangle \\
 &= \left[\nabla \mathbb{E}_t \tilde{L}_t(\theta_t^*) + \nabla \zeta_t \right]^\top u + \frac{1}{2} \langle \nabla^2 \mathbb{E}_t \tilde{L}_t(\theta_t^* + \tilde{u}), u^{\otimes 2} \rangle \\
 &= \nabla \zeta_t^\top u + \frac{1}{2} \langle \nabla^2 \mathbb{E}_t \tilde{L}_t(\theta_t^* + \tilde{u}), u^{\otimes 2} \rangle \\
 &= \left[\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \zeta_t \right]^\top \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} u - \frac{1}{2} \langle \tilde{\mathbf{F}}_{t,\theta_t^* + \tilde{u}}, u^{\otimes 2} \rangle \\
 &\stackrel{\text{Lemma 40}}{\leq} \left(\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \zeta_t \right\|_2 - \frac{1}{2} (1 - 4\tau_{3,t}^* r_{\text{eff},t}) \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} u \right\|_2 \right) \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} u \right\|_2 \\
 &\stackrel{\text{(A1)}}{\leq} [r_{\text{eff},t} - 2(1 - 4\tau_{3,t}^* r_{\text{eff},t}) r_{\text{eff},t}] \times 4r_{\text{eff},t}, \quad (\because \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} u \right\|_2 = 4r_{\text{eff},t}) \\
 &\leq -2r_{\text{eff},t}^2,
 \end{aligned}$$

where the last inequality holds by $\tau_{3,t}^* r_{\text{eff},t} \leq 1/16$. This completes the proof. \blacksquare

Lemma 18 *Suppose that (A0), (A1) hold. Also, assume that $\|\theta_t^* - \theta_0\|_2 \leq 1/2$ for all $t \in [T]$ on an event \mathcal{E} . Then,*

$$r_{\text{eff},t} \leq M_n \left[\left\{ \frac{\lambda_{\max}(\mathbf{F}_{t,\theta_t^*})}{\lambda_{\min}(\tilde{\mathbf{F}}_{t,\theta_t^*})} \wedge 1 \right\} p_* \right]^{1/2}, \quad \forall t \in [T]$$

on \mathcal{E} , where $p_* = p \vee \log n \vee \log T$.

Proof Note that

$$\begin{aligned}
 r_{\text{eff},t} &= p_{\text{eff},t}^{1/2} + \sqrt{2\lambda_t(\log n + \log T)} = \sqrt{\text{tr}(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \mathbf{V}_t)} + \sqrt{2\lambda_t(\log n + \log T)} \\
 &\leq \sqrt{\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \mathbf{V}_t \right\|_2} p + \sqrt{2\lambda_t(\log n + \log T)} = \lambda_t^{1/2} \left(p^{1/2} + \sqrt{2(\log n + \log T)} \right) \\
 &\leq 3\lambda_t^{1/2} p_*^{1/2}.
 \end{aligned}$$

By the assumption $\|\theta_t^* - \theta_0\|_2 \leq 1/2$ on \mathcal{E} , we have

$$\begin{aligned}
 \lambda_t &= \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \mathbf{V}_t \right\|_2 = \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \mathbf{F}_{t,\theta_t^*} \mathbf{F}_{t,\theta_t^*}^{-1} \mathbf{V}_t \right\|_2 \leq \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \mathbf{F}_{t,\theta_t^*} \right\|_2 \left\| \mathbf{F}_{t,\theta_t^*}^{-1} \mathbf{V}_t \right\|_2 \\
 &\stackrel{\text{(A1)}}{\leq} \frac{M_n^2}{9} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \mathbf{F}_{t,\theta_t^*} \right\|_2 \leq \frac{M_n^2}{9} \left[1 \wedge \frac{\lambda_{\max}(\mathbf{F}_{t,\theta_t^*})}{\lambda_{\min}(\tilde{\mathbf{F}}_{t,\theta_t^*})} \right].
 \end{aligned}$$

Combining the last two displays, we complete the proof. \blacksquare

Appendix D. Proofs for Section 6

For $t \in [T]$, let

$$\begin{aligned}
 b_{n,t} &= \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \boldsymbol{\Omega}_{t-1} (\theta_0 - \mu_{t-1}) \right\|_2, \\
 \tau_{3,t,\text{bias}} &= \inf \left\{ \tau_3 \in \mathbb{R}_+ : \sup_{u \in \Theta(\tilde{\mathbf{F}}_{t,\theta_t^*}, 4b_{n,t})} \sup_{z \in \mathbb{R}^p} \frac{|\langle \nabla^3 \mathbb{E}_t \tilde{L}_t(\theta_t^* + u), z^{\otimes 3} \rangle|}{\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} z \right\|_2^3} \leq \tau_3 \right\}, \\
 \tau_{4,t,\text{bias}} &= \inf \left\{ \tau_4 \in \mathbb{R}_+ : \sup_{u \in \Theta(\tilde{\mathbf{F}}_{t,\theta_t^*}, 4b_{n,t})} \sup_{z \in \mathbb{R}^p} \frac{|\langle \nabla^4 \mathbb{E}_t \tilde{L}_t(\theta_t^* + u), z^{\otimes 4} \rangle|}{\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} z \right\|_2^4} \leq \tau_4 \right\}.
 \end{aligned} \tag{104}$$

Lemma 19 *Suppose that (A0) and (A1) hold. Also, assume that $\tau_{3,t,\text{bias}} b_{n,t} \leq 1/16$ on an event \mathcal{E} . Then, on \mathcal{E} ,*

$$\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\theta_0 - \theta_t^*) \right\|_2 \leq (a) + (b) + (c), \tag{105}$$

where

$$\begin{aligned}
 (a) &= \frac{1}{6} \tau_{4,t,\text{bias}} \left(1 - 4\tau_{3,t,\text{bias}} b_{n,t} \right)^{-1} \left(1 + \frac{1}{2} \tau_{3,t,\text{bias}} b_{n,t} \right)^3 b_{n,t}^3, \\
 (b) &= \frac{1}{2} \tau_{3,t,\text{bias}}^2 \left(1 - 4\tau_{3,t,\text{bias}} b_{n,t} \right)^{-1} \left(1 + \frac{1}{2} \tau_{3,t,\text{bias}} b_{n,t} \right) b_{n,t}^3, \\
 (c) &= \left(1 + \frac{1}{2} \tau_{3,t,\text{bias}} b_{n,t} \right) b_{n,t}.
 \end{aligned}$$

Proof In this proof, we work on the event \mathcal{E} without explicitly referring to it. Let $t \in [T]$. For simplicity in notations, let $\Theta_{n,t} = \Theta(\theta_t^*, \tilde{\mathbf{F}}_{t,\theta_t^*}, 4b_{n,t})$, $\bar{\Theta}_{n,t} = \Theta(\tilde{\mathbf{F}}_{t,\theta_t^*}, 4b_{n,t})$ in this proof. For $\theta \in \Theta$, let

$$g_t(\theta) = \mathbb{E}_t \tilde{L}_t(\theta) + \langle \boldsymbol{\Omega}_{t-1} (\theta_0 - \mu_{t-1}), \theta \rangle. \tag{106}$$

It follows that

$$\nabla g_t(\theta_0) \stackrel{(106)}{=} \nabla \mathbb{E}_t \tilde{L}_t(\theta_0) + \boldsymbol{\Omega}_{t-1} (\theta_0 - \mu_{t-1}) = \nabla \mathbb{E}_t L_t(\theta_0) = 0.$$

Let

$$\varphi_t = \boldsymbol{\Omega}_{t-1} (\theta_0 - \mu_{t-1}), \quad \phi_t = \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \left[\varphi_t + \frac{1}{2} \left\langle \nabla^3 \tilde{L}_t(\theta_t^*), \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right)^{\otimes 2} \right\rangle \right]. \tag{107}$$

By Taylor's theorem, we have

$$\begin{aligned}
 &\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} [\nabla g_t(\theta_t^* + \phi_t) - \nabla g_t(\theta_0)] \\
 &= \left[\int_0^1 \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla^2 g_t(\theta_0 + s \{\theta_t^* + \phi_t - \theta_0\}) \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} ds \right] \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\theta_t^* + \phi_t - \theta_0) \\
 &\stackrel{(106)}{=} \left[\int_0^1 \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla^2 \mathbb{E}_t \tilde{L}_t(\theta_0 + s \{\theta_t^* + \phi_t - \theta_0\}) \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} ds \right] \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\theta_t^* + \phi_t - \theta_0).
 \end{aligned} \tag{108}$$

Later, we will prove that

$$\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\theta_0 - \theta_t^*) \right\|_2 \vee \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \phi_t \right\|_2 \leq 4b_{n,t}, \quad (109)$$

which implies that $\theta_0, \theta_t^* + \phi_t \in \Theta_{n,t}$. By (108) and Lemma 40 with $f(\cdot) = \mathbb{E}_t \tilde{L}_t(\cdot)$ and $x = \theta_t^*$, we have

$$\begin{aligned} - \int_0^1 \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla^2 \mathbb{E}_t \tilde{L}_t(\theta_0 + s \{\theta_t^* + \phi_t - \theta_0\}) \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} ds &\succeq \int_0^1 \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \left[(1 - 4\tau_{3,t,\text{bias}} b_{n,t}) \tilde{\mathbf{F}}_{t,\theta_t^*} \right] \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} ds \\ &= (1 - 4\tau_{3,t,\text{bias}} b_{n,t}) \mathbf{I}_p, \end{aligned}$$

which implies that

$$\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} [\nabla g_t(\theta_t^* + \phi_t) - \nabla g_t(\theta_0)] \right\|_2 \geq (1 - 4\tau_{3,t,\text{bias}} b_{n,t}) \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\theta_t^* + \phi_t - \theta_0) \right\|_2.$$

It follows that

$$\begin{aligned} &\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\theta_t^* - \theta_0) \right\|_2 \\ &\leq (1 - 4\tau_{3,t,\text{bias}} b_{n,t})^{-1} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} [\nabla g_t(\theta_t^* + \phi_t) - \nabla g_t(\theta_0)] \right\|_2 + \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \phi_t \right\|_2 \\ &= (1 - 4\tau_{3,t,\text{bias}} b_{n,t})^{-1} \text{(ii)} + \text{(i)}, \end{aligned} \quad (110)$$

where

$$\text{(i)} = \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \phi_t \right\|_2, \quad \text{(ii)} = \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} [\nabla g_t(\theta_t^* + \phi_t) - \nabla g_t(\theta_0)] \right\|_2.$$

In the remainder of this proof, we will prove (109), and obtain upper bounds of (i) and (ii). Firstly, we will prove (109), which encompasses an upper bound of (i). By Lemma 41 with

$$\tau_3 = \tau_{3,t,\text{bias}}, \quad f(\cdot) = \mathbb{E}_t \tilde{L}_t(\cdot), \quad \theta = \theta_t^*, \quad \tilde{\theta} = \theta_0, \quad \beta = \varphi_t, \quad r = b_{n,t},$$

the condition $\tau_{3,t,\text{bias}} b_{n,t} \leq 1/16$ implies that

$$\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\theta_0 - \theta_t^*) \right\|_2 \leq 4b_{n,t}.$$

Also,

$$\begin{aligned} &\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \phi_t - \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \varphi_t \right\|_2 \stackrel{(107)}{=} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \times \frac{1}{2} \left\langle \nabla^3 \tilde{L}_t(\theta_t^*), \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right)^{\otimes 2} \right\rangle \right\|_2 \\ &= \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \frac{1}{2} \left| \left\langle \nabla^3 \tilde{L}_t(\theta_t^*), \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right)^{\otimes 2} \otimes \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} u \right) \right\rangle \right| \\ &\stackrel{(12)}{\leq} \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \frac{1}{2} \tau_{3,t,\text{bias}} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right\|_2^2 \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} u \right\|_2 = \frac{1}{2} \tau_{3,t,\text{bias}} b_{n,t}^2 \end{aligned} \quad (111)$$

It follows that

$$(i) = \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \phi_t \right\|_2 \leq \left(1 + \frac{1}{2} \tau_{3,t,\text{bias}} b_{n,t} \right) b_{n,t} \leq 4b_{n,t}, \quad (112)$$

which completes the proof of (109).

Next, we will obtain an upper bound of (ii). Since $\nabla g_t(\theta_0) = 0$, it suffices to obtain an upper bound of $\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla g_t(\theta_t^* + \phi_t) \right\|_2$. By (107), we have

$$\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \varphi_t = \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \phi_t - \frac{1}{2} \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \left\langle \nabla^3 \tilde{L}_t(\theta_t^*), \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right)^{\otimes 2} \right\rangle. \quad (113)$$

Also,

$$\begin{aligned} & \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla g_t(\theta_t^* + \phi_t) \right\|_2 \stackrel{(106)}{=} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \mathbb{E}_t \tilde{L}_t(\theta_t^* + \phi_t) + \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \varphi_t \right\|_2 \\ & \stackrel{(113)}{=} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \mathbb{E}_t \tilde{L}_t(\theta_t^* + \phi_t) + \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \phi_t - \frac{1}{2} \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \left\langle \nabla^3 \tilde{L}_t(\theta_t^*), \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right)^{\otimes 2} \right\rangle \right\|_2 \\ & \leq \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \mathbb{E}_t \tilde{L}_t(\theta_t^* + \phi_t) + \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \phi_t - \frac{1}{2} \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \left\langle \nabla^3 \tilde{L}_t(\theta_t^*), (\phi_t)^{\otimes 2} \right\rangle \right\|_2 \\ & \quad + \left\| \frac{1}{2} \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \left\langle \nabla^3 \tilde{L}_t(\theta_t^*), (\phi_t)^{\otimes 2} \right\rangle - \frac{1}{2} \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \left\langle \nabla^3 \tilde{L}_t(\theta_t^*), \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right)^{\otimes 2} \right\rangle \right\|_2. \end{aligned}$$

To bound (ii), we need to obtain an upper bounds of the following quantities:

$$\begin{aligned} (iii) &= \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \mathbb{E}_t \tilde{L}_t(\theta_t^* + \phi_t) + \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \phi_t - \frac{1}{2} \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \left\langle \nabla^3 \tilde{L}_t(\theta_t^*), (\phi_t)^{\otimes 2} \right\rangle \right\|_2, \\ (iv) &= \left\| \frac{1}{2} \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \left\langle \nabla^3 \tilde{L}_t(\theta_t^*), (\phi_t)^{\otimes 2} \right\rangle - \frac{1}{2} \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \left\langle \nabla^3 \tilde{L}_t(\theta_t^*), \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right)^{\otimes 2} \right\rangle \right\|_2. \end{aligned}$$

At the end of this proof, we will prove the following inequalities:

$$\begin{aligned} (iii) &\leq \frac{1}{6} \tau_{4,t,\text{bias}} \left(1 + \frac{1}{2} \tau_{3,t,\text{bias}} b_{n,t} \right)^3 b_{n,t}^3 \\ (iv) &\leq \frac{1}{2} \tau_{3,t,\text{bias}}^2 \left(1 + \frac{1}{2} \tau_{3,t,\text{bias}} b_{n,t} \right) b_{n,t}^3. \end{aligned} \quad (114)$$

By (110), we have

$$\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\theta_t^* - \theta_0) \right\|_2 \leq (1 - 4\tau_{3,t,\text{bias}} b_{n,t})^{-1} [(iii) + (iv)] + (i).$$

By (112) and (114), we complete the proof of (105).

To complete the proof, we only need to prove (114). First, we will prove the first inequality in (114). By (112), note that $\theta_t^* + \phi_t \in \Theta_{n,t}$. Also, the stochastic linearity in **(A1)** and the definition of θ_t^* imply that for $k \in \{2, 3, 4\}$

$$\nabla^k \mathbb{E}_t \tilde{L}_t(\cdot) = \nabla^k \tilde{L}_t(\cdot), \quad \nabla \mathbb{E}_t \tilde{L}_t(\theta_t^*) = 0.$$

By Taylor's theorem, we have

$$\begin{aligned}
 & \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \mathbb{E}_t \tilde{L}_t(\theta_t^* + \phi_t) + \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \phi_t - \frac{1}{2} \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \left\langle \nabla^3 \tilde{L}_t(\theta_t^*), (\phi_t)^{\otimes 2} \right\rangle \right\|_2 \\
 &= \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \left[\nabla \mathbb{E}_t \tilde{L}_t(\theta_t^* + \phi_t) - \nabla \mathbb{E}_t \tilde{L}_t(\theta_t^*) - \langle \nabla^2 \mathbb{E}_t \tilde{L}_t(\theta_t^*), \phi_t \rangle - \frac{1}{2} \left\langle \nabla^3 \mathbb{E}_t \tilde{L}_t(\theta_t^*), (\phi_t)^{\otimes 2} \right\rangle \right] \right\|_2 \\
 &\leq \sup_{\tilde{u} \in \bar{\Theta}_{n,t}} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \frac{1}{6} \left\langle \nabla^4 \mathbb{E}_t \tilde{L}_t(\theta_t^* + \tilde{u}), (\phi_t)^{\otimes 3} \right\rangle \right\|_2 \\
 &= \sup_{\tilde{u} \in \bar{\Theta}_{n,t}} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \frac{1}{6} \left\langle \nabla^4 \tilde{L}_t(\theta_t^* + \tilde{u}), (\phi_t)^{\otimes 3} \right\rangle \right\|_2 \\
 &= \sup_{\tilde{u} \in \bar{\Theta}_{n,t}} \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \frac{1}{6} \left\langle \nabla^4 \tilde{L}_t(\theta_t^* + \tilde{u}), (\phi_t)^{\otimes 3} \otimes \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} u \right) \right\rangle \\
 &\stackrel{(12)}{\leq} \frac{1}{6} \tau_{4,t,\text{bias}} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \phi_t \right\|_2^3 \stackrel{(112)}{\leq} \frac{1}{6} \tau_{4,t,\text{bias}} \left(1 + \frac{1}{2} \tau_{3,t,\text{bias}} b_{n,t} \right)^3 b_{n,t}^3,
 \end{aligned}$$

which completes the proof of the first inequality in (114).

Next, we will prove the second inequality in (114). Let

$$\mathcal{T} = (\mathcal{T}_{ijk})_{i,j,k \in [p]} = \nabla^3 \tilde{L}_t(\theta_t^*) / 6 \in \mathbb{R}^{p \times p \times p}.$$

With a slight abuse of notation, let $\mathcal{T} : \mathbb{R}^p \rightarrow \mathbb{R}$ be the function defined as

$$\mathcal{T}(u) = \langle \mathcal{T}, u^{\otimes 3} \rangle.$$

Then, we have

$$\nabla \mathcal{T}(u) = \left(3 \sum_{j,k} \mathcal{T}_{ijk} u_j u_k \right)_{i \in [p]} = 3 \left(\langle \mathcal{T}_i, u^{\otimes 2} \rangle \right)_{i \in [p]}, \quad \nabla^2 \mathcal{T}(u) = 6 \sum_{i=1}^p u_i \mathcal{T}_i,$$

where $\mathcal{T}_i = (\mathcal{T}_{ijk})_{j,k \in [p]} \in \mathbb{R}^{p \times p}$. By using the above definitions, note that

$$\begin{aligned}
 & \frac{1}{2} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \left\langle \nabla^3 \tilde{L}_t(\theta_t^*), (\phi_t)^{\otimes 2} \right\rangle - \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \left\langle \nabla^3 \tilde{L}_t(\theta_t^*), \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right)^{\otimes 2} \right\rangle \right\|_2 \\
 &= \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \mathcal{T}(\phi_t) - \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \mathcal{T} \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right) \right\|_2 \\
 &\leq \sup_{s \in [0,1]} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla^2 \mathcal{T} \left(s \phi_t + (1-s) \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right) \left(\phi_t - \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right) \right\|_2 \\
 &\leq \sup_{s \in [0,1]} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla^2 \mathcal{T} \left(s \phi_t + (1-s) \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right) \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \right\|_2 \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \left(\phi_t - \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right) \right\|_2.
 \end{aligned}$$

Therefore, we only need to bound the following quantities:

$$\sup_{s \in [0,1]} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla^2 \mathcal{T} \left(s \phi_t + (1-s) \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right) \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \right\|_2, \quad \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \left(\phi_t - \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right) \right\|_2.$$

For $u \in \Theta$, note that

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla^2 \mathcal{T}(u) \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \right\|_2 &= \sup_{\tilde{u} \in \mathbb{R}^p: \|\tilde{u}\|_2=1} 6 \left| \left\langle \mathcal{T}, \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \tilde{u} \right)^{\otimes 2} \otimes u \right\rangle \right| \\ &= \sup_{\tilde{u} \in \mathbb{R}^p: \|\tilde{u}\|_2=1} \left| \left\langle \nabla^3 \tilde{L}_t(\theta_t^*), \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \tilde{u} \right)^{\otimes 2} \otimes u \right\rangle \right| \stackrel{(12)}{\leq} \tau_{3,t,\text{bias}} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} u \right\|_2. \end{aligned}$$

Also, for $u = (u_i)_{i \in [p]} \in \Theta$,

$$u \mapsto \nabla^2 \mathcal{T}(u) = 6 \sum_{i=1}^p u_i \mathcal{T}_i$$

is a linear map. By the last two displays, we have

$$\begin{aligned} &\sup_{s \in [0,1]} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla^2 \mathcal{T} \left(s\phi_t + (1-s)\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right) \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \right\|_2 \\ &= \sup_{s \in [0,1]} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla^2 \mathcal{T} \left((1-s)\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right) \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} + \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla^2 \mathcal{T} (s\phi_t) \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \right\|_2 \\ &= \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla^2 \mathcal{T} \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right) \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \right\|_2 \vee \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla^2 \mathcal{T} (\phi_t) \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \right\|_2 \\ &\leq \tau_{3,t,\text{bias}} \left[\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \varphi_t \right\|_2 \vee \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \phi_t \right\|_2 \right] = \tau_{3,t,\text{bias}} \left[b_{n,t} \vee \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \phi_t \right\|_2 \right] \\ &\stackrel{(112)}{\leq} \tau_{3,t,\text{bias}} \left(1 + \frac{1}{2} \tau_{3,t,\text{bias}} b_{n,t} \right) b_{n,t}. \end{aligned}$$

Also, (111) implies that

$$\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} \left(\phi_t - \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \varphi_t \right) \right\|_2 \leq \frac{1}{2} \tau_{3,t,\text{bias}} b_{n,t}^2.$$

By the last two displays, we complete the proof of the second inequality in (114). \blacksquare

The quantities $(\hat{\tau}_{3,t}, \hat{\tau}_{4,t})$, $\hat{\tau}_{3,t,r}$, $\tau_{3,t}^*$ and $(\tau_{3,t,\text{bias}}, \tau_{4,t,\text{bias}})$, which appear in the following lemma, are defined in (19), (24), (28) and (104), respectively. By Lemma 42, these quantities can be characterized by the smallest eigenvalue of $\tilde{\mathbf{F}}_{t,\theta_t^*}$ and $\tilde{\mathbf{F}}_{t,\hat{\theta}_t}$.

Lemma 20 *Suppose that (A0) and (A1) hold, and $N \geq 2$. Let $\alpha \in (0, 1]$ and $t \in \{1, 2, \dots, T-1\}$. Assume that*

$$r_{\text{eff},t} \leq C_1 M_n \sqrt{t^{-1} p_*}, \quad \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\theta_0 - \theta_t^*) \right\|_2 \leq C_2 M_n t^\alpha \sqrt{p_*} \quad (115)$$

on an event \mathcal{E} , where $C_1, C_2 > 0$ are constants. Also, assume that

$$\left[\hat{\tau}_{3,t,r}^2 \right] \vee \left[\hat{\tau}_{4,t} p^2 \right] \vee \left[\tau_{3,t,r}^* r_{\text{eff},t} \right] \vee \left[\epsilon_{n,t,\text{KL}} \right] \leq \delta, \quad (116)$$

$$\begin{aligned} \tau_{3,t+1,\text{bias}} (M_n t^{1/2+2\alpha} \sqrt{p_*}) &\leq \delta, \\ (\epsilon_{n,t,\text{KL}} \vee [\tau_{3,t,r}^* r_{\text{eff},t}]) t^{1/2+\alpha} &\leq \delta, \end{aligned} \quad (117)$$

$$(\tau_{4,t+1,\text{bias}} \vee \tau_{3,t+1,\text{bias}}^2) M_n^2 t^{1/2+3\alpha} p_* \leq \delta,$$

on \mathcal{E} , where $\delta = \delta(C_1, C_2, \alpha) > 0$ is a small enough constant. Assume further that

$$\widehat{\tau}_{3,t,r} \leq N e^{\delta p} \exp\left([\sqrt{p} + \sqrt{2 \log N} - 3]r\right), \quad \forall r > 4r_{\text{LA}} \quad (118)$$

on \mathcal{E} . Then, on $\mathcal{E}_{\text{est},1} \cap \mathcal{E}$,

$$\left\| \widetilde{\mathbf{F}}_{t+1, \theta_{t+1}^*}^{1/2} (\theta_0 - \theta_{t+1}^*) \right\|_2 \leq K M_n (t^{1/2-\alpha} + t)^\alpha \sqrt{p_*}, \quad (119)$$

where $K = K(C_1, C_2, \alpha)$.

Proof In this proof, we work on the event $\mathcal{E} \cap \mathcal{E}_{\text{est},1}$ without explicitly referring to it, and assume that $\delta = \delta(C_1, C_2, \alpha)$ is small enough constant.

By Lemma 19, if $\tau_{3,t+1, \text{bias}} b_{n,t+1} \leq 1/16$, then

$$\left\| \widetilde{\mathbf{F}}_{t+1, \theta_{t+1}^*}^{1/2} (\theta_0 - \theta_{t+1}^*) \right\|_2 \leq (a) + (b) + (c), \quad (120)$$

where

$$\begin{aligned} (a) &= \frac{1}{6} \tau_{4,t+1, \text{bias}} \left(1 - 4\tau_{3,t+1, \text{bias}} b_{n,t+1}\right)^{-1} \left(1 + \frac{1}{2} \tau_{3,t+1, \text{bias}} b_{n,t+1}\right)^3 b_{n,t+1}^3, \\ (b) &= \frac{1}{2} \tau_{3,t+1, \text{bias}}^2 \left(1 - 4\tau_{3,t+1, \text{bias}} b_{n,t+1}\right)^{-1} \left(1 + \frac{1}{2} \tau_{3,t+1, \text{bias}} b_{n,t+1}\right) b_{n,t+1}^3, \\ (c) &= \left(1 + \frac{1}{2} \tau_{3,t+1, \text{bias}} b_{n,t+1}\right) b_{n,t+1}. \end{aligned}$$

We will obtain an upper bound of $b_{n,t+1}$ first, and then bound (a) + (b) + (c).

By $N \geq 2$, (118) and (116) with sufficiently small δ , we can utilize the results in Theorem 5 and Corollary 17. Hence, we have

$$\begin{aligned} \left\| \widetilde{\mathbf{F}}_{t, \theta_t^*}^{1/2} (\theta_t^* - \widehat{\theta}_t) \right\|_2 &\stackrel{\text{Theorem 5}}{\leq} 4r_{\text{eff},t}, \\ \left\| \widetilde{\mathbf{F}}_{t, \theta_t^*}^{-1/2} \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t} \widetilde{\mathbf{F}}_{t, \theta_t^*}^{-1/2} - \mathbf{I}_p \right\|_2 &\stackrel{\text{Lemma 40}}{\leq} 4\tau_{3,t}^* r_{\text{eff},t}, \\ \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{-1/2} \widetilde{\mathbf{\Omega}}_t \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_2 &\leq \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{-1/2} \widetilde{\mathbf{\Omega}}_t \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_{\mathbb{F}} \stackrel{\text{Corollary 17}}{\leq} c_1 \epsilon_{n,t, \text{KL}}, \\ \left\| \widetilde{\mathbf{\Omega}}_t^{1/2} (\widehat{\theta}_t - \mu_t) \right\|_2 &\stackrel{\text{Corollary 17}}{\leq} c_1 \epsilon_{n,t, \text{KL}}, \end{aligned} \quad (121)$$

for some universal constant $c_1 > 0$. Note that

$$\begin{aligned}
 b_{n,t+1} &= \left\| \tilde{\mathbf{F}}_{t+1,\theta_{t+1}^*}^{-1/2} \boldsymbol{\Omega}_t(\theta_0 - \mu_t) \right\|_2 \leq \left\| \boldsymbol{\Omega}_t^{-1/2} \boldsymbol{\Omega}_t(\theta_0 - \mu_t) \right\|_2 = \left\| \boldsymbol{\Omega}_t^{1/2}(\theta_0 - \mu_t) \right\|_2 \\
 &\leq \left\| \boldsymbol{\Omega}_t^{1/2}(\theta_0 - \theta_t^*) \right\|_2 + \left\| \boldsymbol{\Omega}_t^{1/2}(\theta_t^* - \hat{\theta}_t) \right\|_2 + \left\| \boldsymbol{\Omega}_t^{1/2}(\hat{\theta}_t - \mu_t) \right\|_2 \\
 &\leq \left\| \boldsymbol{\Omega}_t^{1/2} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \right\|_2 \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \right\|_2 \left(\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2}(\theta_0 - \theta_t^*) \right\|_2 + \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2}(\theta_t^* - \hat{\theta}_t) \right\|_2 \right) \\
 &\quad + \left\| \boldsymbol{\Omega}_t^{1/2}(\hat{\theta}_t - \mu_t) \right\|_2 \\
 &\leq \left(1 + \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \boldsymbol{\Omega}_t \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_2 \right)^{1/2} \left(1 + \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_2 \right)^{1/2} \\
 &\quad \times \left(\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2}(\theta_0 - \theta_t^*) \right\|_2 + \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2}(\theta_t^* - \hat{\theta}_t) \right\|_2 \right) + \left\| \boldsymbol{\Omega}_t^{1/2}(\hat{\theta}_t - \mu_t) \right\|_2 \\
 &\stackrel{(115)}{\leq} \sqrt{(1 + c_1 \epsilon_{n,t,\text{KL}}) \left(1 + 4\tau_{3,t}^* r_{\text{eff},t} \right) \left(C_2 M_n t^\alpha \sqrt{p_*} + 4C_1 M_n \sqrt{t^{-1} p_*} \right) + c_1 \epsilon_{n,t,\text{KL}}} \\
 &\stackrel{(121)}{\leq} \sqrt{(1 + c_1 \epsilon_{n,t,\text{KL}}) \left(1 + 4\tau_{3,t}^* r_{\text{eff},t} \right) \left(C_2 M_n t^\alpha \sqrt{p_*} + 4C_1 M_n \sqrt{t^{-1} p_*} + c_1 \epsilon_{n,t,\text{KL}} \right)}.
 \end{aligned}$$

It follows that

$$\begin{aligned}
 b_{n,t+1} &\leq (1 + c_1 \epsilon_{n,t,\text{KL}}) \left(1 + 4\tau_{3,t}^* r_{\text{eff},t} \right) \left(C_2 M_n t^\alpha \sqrt{p_*} + 4C_1 M_n \sqrt{t^{-1} p_*} + c_1 \epsilon_{n,t,\text{KL}} \right) \\
 &\leq (1 + c_1 \delta) (1 + 4\delta) \left(C_2 M_n t^\alpha \sqrt{p_*} + 4C_1 M_n \sqrt{t^{-1} p_*} + c_1 \delta \right) \\
 &\lesssim M_n t^\alpha \sqrt{p_*},
 \end{aligned}$$

where the second inequality holds by (116). Consequently, we have

$$\tau_{3,t+1,\text{bias}} b_{n,t+1} \stackrel{(117)}{\leq} 1/16, \tag{122}$$

which implies that (120) holds.

Next, we will obtain an upper bound of (a) + (b) + (c). Since $(1 - x)^{-1} \leq 1 + 2x$ for $x \leq 1/2$ and $(1 + x)^3 \leq 1 + 4x$ for $x \in [0, 0.3]$,

$$\begin{aligned}
 \text{(a)} &= \frac{1}{6} \tau_{4,t+1,\text{bias}} \left(1 - 4\tau_{3,t+1,\text{bias}} b_{n,t+1} \right)^{-1} \left(1 + \frac{1}{2} \tau_{3,t+1,\text{bias}} b_{n,t+1} \right)^3 b_{n,t+1}^3 \\
 &\leq \frac{1}{6} \tau_{4,t+1,\text{bias}} \left(1 + 8\tau_{3,t+1,\text{bias}} b_{n,t+1} \right) \left(1 + 2\tau_{3,t+1,\text{bias}} b_{n,t+1} \right) b_{n,t+1}^3 \\
 &\stackrel{(122)}{\leq} \frac{1}{6} \tau_{4,t+1,\text{bias}} \left(1 + \frac{1}{2} \right) \left(1 + \frac{1}{8} \right) b_{n,t+1}^3 \leq \tau_{4,t+1,\text{bias}} b_{n,t+1}^3.
 \end{aligned}$$

Similarly, we have

$$\begin{aligned}
 \text{(b)} &= \frac{1}{2} \tau_{3,t+1,\text{bias}}^2 \left(1 - 4\tau_{3,t+1,\text{bias}} b_{n,t+1} \right)^{-1} \left(1 + \frac{1}{2} \tau_{3,t+1,\text{bias}} b_{n,t+1} \right) b_{n,t+1}^3 \\
 &\leq \frac{1}{2} \tau_{3,t+1,\text{bias}}^2 \left(1 + 8\tau_{3,t+1,\text{bias}} b_{n,t+1} \right) \left(1 + \frac{1}{2} \tau_{3,t+1,\text{bias}} b_{n,t+1} \right) b_{n,t+1}^3 \\
 &\stackrel{(122)}{\leq} \frac{1}{2} \tau_{3,t+1,\text{bias}}^2 \left(1 + \frac{1}{2} \right) \left(1 + \frac{1}{32} \right) b_{n,t+1}^3 \leq \tau_{3,t+1,\text{bias}}^2 b_{n,t+1}^3.
 \end{aligned}$$

By (116), (117), and the inequality $b_{n,t+1} \lesssim M_n t^\alpha \sqrt{p_*}$, we have

$$\begin{aligned} (\tau_{3,t+1,\text{bias}} b_{n,t+1}) \vee (2\tau_{4,t+1,\text{bias}} b_{n,t+1}^2) \vee (2\tau_{3,t+1,\text{bias}}^2 b_{n,t+1}^2) &< 1/6, \\ (4c_1 \epsilon_{n,t,\text{KL}}) \vee (8\tau_{3,t}^* r_{\text{eff},t}) \vee (8C_1 K^{-1}) &< 1/6, \end{aligned} \quad (123)$$

for some large enough constant $K = K(C_1, C_2, \alpha) > 0$. (Here, K can be chosen depending only on C_1 , but later, it will be taken as a larger constant that depends on C_2 and α .) Also, we can further bound $b_{n,t+1}$ as follows:

$$\begin{aligned} b_{n,t+1} &\leq (1 + c_1 \epsilon_{n,t,\text{KL}}) (1 + 4\tau_{3,t}^* r_{\text{eff},t}) (C_2 M_n t^\alpha \sqrt{p_*} + 4C_1 M_n \sqrt{t^{-1} p_*} + c_1 \epsilon_{n,t,\text{KL}}) \\ &\leq (1 + c_1 \epsilon_{n,t,\text{KL}}) (1 + 4\tau_{3,t}^* r_{\text{eff},t}) (K M_n t^\alpha \sqrt{p_*} + 4C_1 M_n \sqrt{t^{-1} p_*} + c_1 \epsilon_{n,t,\text{KL}}) \\ &= (1 + c_1 \epsilon_{n,t,\text{KL}}) (1 + 4\tau_{3,t}^* r_{\text{eff},t}) \\ &\quad \times (1 + 4C_1 K^{-1} t^{-1/2-\alpha} + K^{-1} M_n^{-1} t^{-\alpha} p_*^{-1/2} c_1 \epsilon_{n,t,\text{KL}}) K M_n t^\alpha \sqrt{p_*} \\ &\leq (1 + c_1 \epsilon_{n,t,\text{KL}}) (1 + 4\tau_{3,t}^* r_{\text{eff},t}) (1 + 4C_1 K^{-1} t^{-1/2-\alpha} + c_1 \epsilon_{n,t,\text{KL}}) K M_n t^\alpha \sqrt{p_*}, \end{aligned} \quad (124)$$

where the second and last inequalities hold by taking sufficiently large K . Let

$$\begin{aligned} \rho_{n,t} &= \tau_{3,t+1,\text{bias}} b_{n,t+1} + 2\tau_{4,t+1,\text{bias}} b_{n,t+1}^2 + 2\tau_{3,t+1,\text{bias}}^2 b_{n,t+1}^2 \\ &\quad + 4c_1 \epsilon_{n,t,\text{KL}} + 8\tau_{3,t}^* r_{\text{eff},t} + 8C_1 K^{-1} t^{-1/2-\alpha} \stackrel{(123)}{<} 1. \end{aligned}$$

Therefore, from (120) and the bounds for (a) and (b) above, we have

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{t+1, \theta_{t+1}^*}^{1/2} (\theta_0 - \theta_{t+1}^*) \right\|_2 &\leq \tau_{4,t+1,\text{bias}} b_{n,t+1}^3 + \tau_{3,t+1,\text{bias}}^2 b_{n,t+1}^3 + \left(1 + \frac{1}{2} \tau_{3,t+1,\text{bias}} b_{n,t+1} \right) b_{n,t+1} \\ &= \left(1 + \frac{1}{2} \tau_{3,t+1,\text{bias}} b_{n,t+1} + \tau_{4,t+1,\text{bias}} b_{n,t+1}^2 + \tau_{3,t+1,\text{bias}}^2 b_{n,t+1}^2 \right) b_{n,t+1} \\ &\stackrel{(124)}{\leq} \left(1 + \frac{1}{2} \tau_{3,t+1,\text{bias}} b_{n,t+1} + \tau_{4,t+1,\text{bias}} b_{n,t+1}^2 + \tau_{3,t+1,\text{bias}}^2 b_{n,t+1}^2 \right) (1 + c_1 \epsilon_{n,t,\text{KL}}) (1 + 4\tau_{3,t}^* r_{\text{eff},t}) \\ &\quad \times (1 + 4C_1 K^{-1} t^{-1/2-\alpha} + c_1 \epsilon_{n,t,\text{KL}}) K M_n t^\alpha \sqrt{p_*} \\ &\leq (1 + \tau_{3,t+1,\text{bias}} b_{n,t+1} + 2\tau_{4,t+1,\text{bias}} b_{n,t+1}^2 + 2\tau_{3,t+1,\text{bias}}^2 b_{n,t+1}^2 + 4c_1 \epsilon_{n,t,\text{KL}} + 8\tau_{3,t}^* r_{\text{eff},t} + 8C_1 K^{-1} t^{-1/2-\alpha}) \\ &\quad \times K M_n t^\alpha \sqrt{p_*}, \\ &= (1 + \rho_{n,t}) K M_n t^\alpha \sqrt{p_*}, \end{aligned}$$

where the third inequality holds because

$$(1 + x_1)(1 + x_2)(1 + x_3)(1 + x_4) \leq e^{x_1 + x_2 + x_3 + x_4} \leq 1 + 2x_1 + 2x_2 + 2x_3 + 2x_4$$

for $x_1, x_2, x_3, x_4 \in \mathbb{R}_+$ with $x_1 + x_2 + x_3 + x_4 \in (0, 1)$.

Suppose that the following inequality holds:

$$2^{1/\alpha} \cdot \rho_{n,t} \cdot t^{1/2+\alpha} \leq 1. \quad (125)$$

Since

$$(1+x)^\omega \leq 1 + 2^\omega x$$

for $x \in (0, 1)$ and $\omega \geq 1$, we have

$$(1 + \rho_{n,t})^{1/\alpha} \leq 1 + 2^{1/\alpha} \rho_{n,t} \stackrel{(125)}{\leq} 1 + t^{-1/2-\alpha}.$$

It follows that

$$\begin{aligned} (1 + \rho_{n,t}) K M_n t^\alpha \sqrt{p_*} &\leq (1 + t^{-1/2-\alpha})^\alpha K M_n t^\alpha \sqrt{p_*} \\ &= \left(t [1 + t^{-1/2-\alpha}] \right)^\alpha K M_n \sqrt{p_*} = (t + t^{1/2-\alpha})^\alpha K M_n \sqrt{p_*}, \end{aligned}$$

which completes the proof of (119). Therefore, we only need to prove (125).

By (124) and (116), with a large enough K and small enough δ , we have

$$b_{n,t+1} \leq 2 K M_n t^\alpha \sqrt{p_*}.$$

It follows that

$$\begin{aligned} 2^{1/\alpha} (\tau_{3,t+1,\text{bias}} b_{n,t+1}) t^{1/2+\alpha} &\leq c_2 (M_n t^{1/2+2\alpha} \sqrt{p_*}) \tau_{3,t+1,\text{bias}} \stackrel{(117)}{\leq} 1/6, \\ 2^{1/\alpha} (2\tau_{4,t+1,\text{bias}} b_{n,t+1}^2) t^{1/2+\alpha} &\leq c_2 (M_n^2 t^{1/2+3\alpha} p_*) \tau_{4,t+1,\text{bias}} \stackrel{(117)}{\leq} 1/6, \\ 2^{1/\alpha} (2\tau_{3,t+1,\text{bias}}^2 b_{n,t+1}^2) t^{1/2+\alpha} &\leq c_2 (M_n^2 t^{1/2+3\alpha} p_*^2) \tau_{3,t+1,\text{bias}}^2 \stackrel{(117)}{\leq} 1/6, \end{aligned}$$

where $c_2 = c_2(\alpha, K) > 0$. Also,

$$\begin{aligned} 2^{1/\alpha} (4c_1 \epsilon_{n,t,\text{KL}}) t^{1/2+\alpha} &\leq c_3 t^{1/2+\alpha} \epsilon_{n,t,\text{KL}} \stackrel{(117)}{\leq} 1/6, \\ 2^{1/\alpha} (8\tau_{3,t}^* r_{\text{eff},t}) t^{1/2+\alpha} &\leq c_3 t^{1/2+\alpha} (\tau_{3,t}^* r_{\text{eff},t}) \stackrel{(117)}{\leq} 1/6, \\ 2^{1/\alpha} (8C_1 K^{-1} t^{-1/2-\alpha}) t^{1/2+\alpha} &= 2^{1/\alpha} (8C_1 K^{-1}) \leq 1/6, \end{aligned}$$

where $c_3 = c_3(\alpha, K)$, and the last inequality holds by taking a sufficiently large K . By the last two displays, we complete the proof of (125). \blacksquare

Remark 21 Note that the constant K in (119) can be chosen as

$$K = [(2^{1/\alpha}) 48 C_1] \vee C_2 \vee 1.$$

Lemma 22 Suppose that **(A0)**-**(A2)** hold. Let $\alpha \in [1/2, 1]$ and $t \in \{1, 2, \dots, T-1\}$. Also, on an event \mathcal{E} , assume that

$$\begin{aligned} \lambda_{\min} \left(\tilde{\mathbf{F}}_{t,\theta_t^*} \right) \wedge \lambda_{\min} \left(\tilde{\mathbf{F}}_{t,\hat{\theta}_t} \right) &\geq C_3 n t, \\ \lambda_{\max} \left(\tilde{\mathbf{F}}_{t,\theta_t^*} \right) \vee \lambda_{\max} \left(\tilde{\mathbf{F}}_{t,\hat{\theta}_t} \right) &\leq C_4 n t, \\ \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\theta_0 - \theta_t^*) \right\|_2 &\leq C_5 M_n t^\alpha \sqrt{p_*} \end{aligned} \tag{126}$$

for some constants C_3, C_4 and C_5 with

$$C_3 \in [K_{\min}/2, K_{\min}], \quad C_4 \in [K_{\max}, 2K_{\max}], \quad C_5 \geq (192C_3^{-1/2}K_{\max}^{1/2}) \vee 1.$$

Assume further that

$$n \geq CM_n^{2t} t^{4\alpha-2} p_*^2 \quad (127)$$

for a large enough constant $C = C(K_{\min}, K_{\max}, \alpha, C_5)$. Then, on $\mathcal{E}_{\text{est},1} \cap \mathcal{E}$,

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\hat{\theta}_t - \theta_t^*) \right\|_2 &\leq KM_n \sqrt{t^{-1} p_*}, \\ \tau_{3,t}^* \vee \hat{\tau}_{3,t} &\leq Kt^{-3/2} n^{-1/2}, \\ \hat{\tau}_{4,t} &\leq Kt^{-2} n^{-1}, \\ \Delta_t \vee \epsilon_{n,t,\text{KL}} &\leq Kt^{-1} n^{-1/2} p_*, \\ \hat{\theta}_t, \theta_t^* &\in \Theta(\theta_0, \mathbf{I}_p, 1/2), \end{aligned} \quad (128)$$

where $K = K(K_{\min}, K_{\max})$ is a large enough constant, and

$$\begin{aligned} \lambda_{\min} \left(\tilde{\mathbf{F}}_{t+1,\theta_{t+1}^*} \right) \wedge \lambda_{\min} \left(\tilde{\mathbf{F}}_{t+1,\hat{\theta}_{t+1}} \right) &\geq C_3(1 - \Delta_t)n(t+1), \\ \lambda_{\max} \left(\tilde{\mathbf{F}}_{t+1,\theta_{t+1}^*} \right) \vee \lambda_{\max} \left(\tilde{\mathbf{F}}_{t+1,\hat{\theta}_{t+1}} \right) &\leq C_4(1 + \Delta_t)n(t+1), \\ \hat{\theta}_{t+1}, \theta_{t+1}^* &\in \Theta(\theta_0, \mathbf{I}_p, 1/2), \\ \left\| \tilde{\mathbf{F}}_{t+1,\theta_{t+1}^*}^{1/2} (\theta_0 - \theta_{t+1}^*) \right\|_2 &\leq C_5 M_n (t+1)^\alpha \sqrt{p_*}. \end{aligned} \quad (129)$$

Furthermore, if $C_3(1 - \Delta_t) \geq K_{\min}/2$ and $C_4(1 + \Delta_t) \leq 2K_{\max}$, then

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{t+1,\theta_{t+1}^*}^{1/2} (\hat{\theta}_{t+1} - \theta_{t+1}^*) \right\|_2 &\leq KM_n \sqrt{(t+1)^{-1} p_*}, \\ \tau_{3,t+1}^* \vee \hat{\tau}_{3,t+1} &\leq K(t+1)^{-3/2} n^{-1/2}, \\ \hat{\tau}_{4,t+1} &\leq K(t+1)^{-2} n^{-1}, \\ \Delta_{t+1} \vee \epsilon_{n,t+1,\text{KL}} &\leq K(t+1)^{-1} n^{-1/2} p_* \end{aligned} \quad (130)$$

on $\mathcal{E}_{\text{est},1} \cap \mathcal{E}$.

Proof In this proof, we work on the event $\mathcal{E} \cap \mathcal{E}_{\text{est},1}$ without explicitly referring to it. Also, the dependency on C_3 and C_4 is not explicitly expressed, as it can be replaced by the dependence on K_{\min} and K_{\max} . The constant $C = C(K_{\min}, K_{\max}, \alpha, C_5)$ in (127) will be assumed to be large enough if necessary. The proof is divided into several steps.

Step 1: $\tau_{3,t}^*, r_{\text{eff},t}$

By (126), we have

$$\begin{aligned} \|\theta_0 - \theta_t^*\|_2 &\leq \lambda_{\min}^{-1/2}(\tilde{\mathbf{F}}_{t,\theta_t^*}) C_5 M_n t^\alpha \sqrt{p_*} \stackrel{(126)}{\leq} (C_3 n t)^{-1/2} C_5 M_n t^\alpha \sqrt{p_*} \\ &= (C_3^{-1/2} C_5) M_n t^{\alpha-1/2} p_*^{1/2} n^{-1/2} \stackrel{(127)}{\leq} 1/8. \end{aligned} \quad (131)$$

Also,

$$\begin{aligned}
 r_{\text{eff},t} &\stackrel{\text{(131)}}{\leq} \stackrel{\text{Lemma 18}}{M_n \lambda_{\min}^{-1/2}(\tilde{\mathbf{F}}_{t,\theta_t^*}) \lambda_{\max}^{1/2}(\mathbf{F}_{t,\theta_t^*}) p_*^{1/2}} \\
 &\stackrel{\text{(126)}}{\leq} \stackrel{\text{(131), (A2)}}{M_n (C_3 n t)^{-1/2} (K_{\max} n)^{1/2} \sqrt{p_*}} = (C_3^{-1/2} K_{\max}^{1/2}) M_n \sqrt{t^{-1} p_*}.
 \end{aligned} \tag{132}$$

For $\theta \in \Theta(\theta_t^*, \tilde{\mathbf{F}}_{t,\theta_t^*}, 4r_{\text{eff},t})$, we have

$$\begin{aligned}
 \|\theta - \theta_t^*\|_2 &\leq \lambda_{\min}^{-1/2}(\tilde{\mathbf{F}}_{t,\theta_t^*}) 4r_{\text{eff},t} \stackrel{\text{(126)}}{\leq} (C_3 n t)^{-1/2} 4r_{\text{eff},t} \\
 &\stackrel{\text{(132)}}{\leq} (4C_3^{-1} K_{\max}^{1/2}) M_n p_*^{1/2} n^{-1/2} t^{-1} \stackrel{\text{(127)}}{\leq} 1/8.
 \end{aligned} \tag{133}$$

By (131) and (133), we have

$$\Theta(\theta_t^*, \tilde{\mathbf{F}}_{t,\theta_t^*}, 4r_{\text{eff},t}) \subset \Theta(\theta_0, \mathbf{I}_p, 1/2).$$

Consequently, we have

$$\begin{aligned}
 \tau_{3,t}^* &\stackrel{\text{(A2)}}{\leq} \stackrel{\text{Lemma 42}}{(K_{\max} n) \lambda_{\min}^{-3/2}(\tilde{\mathbf{F}}_{t,\theta_t^*})} \stackrel{\text{(126)}}{\leq} (K_{\max} n) (C_3 n t)^{-3/2} \\
 &= (C_3^{-3/2} K_{\max}) t^{-3/2} n^{-1/2}.
 \end{aligned} \tag{134}$$

Step 2: $\hat{\tau}_{3,t}, \hat{\tau}_{4,t}$

By (132) and (134), we have

$$\begin{aligned}
 \tau_{3,t}^* r_{\text{eff},t} &\leq \left[\frac{K_{\max}}{C_3^{3/2}} t^{-3/2} n^{-1/2} \right] \left[\frac{K_{\max}^{1/2}}{C_3^{1/2}} M_n \sqrt{t^{-1} p_*} \right] = \frac{K_{\max}^{3/2}}{C_3^2} t^{-2} n^{-1/2} M_n p_*^{1/2} \\
 &\stackrel{\text{(127)}}{\leq} 1/32.
 \end{aligned} \tag{135}$$

Therefore,

$$\hat{\theta}_t \in \Theta(\theta_t^*, \tilde{\mathbf{F}}_{t,\theta_t^*}, 4r_{\text{eff},t}) \tag{136}$$

by Theorem 5. It follows that

$$\|\hat{\theta}_t - \theta_0\|_2 \leq \|\hat{\theta}_t - \theta_t^*\|_2 + \|\theta_t^* - \theta_0\|_2 \stackrel{\text{(131)}}{\stackrel{\text{(133)}}{\leq}} 1/8 + 1/8 = 1/4.$$

Also,

$$r_{\text{LA}} = 2\sqrt{p} + \sqrt{2 \log N} \leq 2\sqrt{p} + 2\sqrt{p_*} \leq 4\sqrt{p_*}. \tag{137}$$

For $\theta \in \Theta(\hat{\theta}_t, \tilde{\mathbf{F}}_{t,\hat{\theta}_t}, 4r_{\text{LA}})$, note that

$$\|\theta - \hat{\theta}_t\|_2 \stackrel{\text{(126)}}{\leq} (C_3 n t)^{-1/2} 4r_{\text{LA}} \leq (C_3 n t)^{-1/2} 16p_*^{1/2} = \frac{16}{\sqrt{C_3}} \sqrt{\frac{p_*}{n t}} \stackrel{\text{(127)}}{\leq} 1/4.$$

It follows that $\Theta(\widehat{\theta}_t, \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}, 4r_{\text{LA}}) \subset \Theta(\theta_0, \mathbf{I}_p, 1/2)$. By Lemma 42 and **(A2)**, therefore, we have

$$\begin{aligned}\widehat{\tau}_{3,t} &\leq (K_{\max} n) \lambda_{\min}^{-3/2}(\widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}) \leq (K_{\max} n) (C_3 n t)^{-3/2} = (C_3^{-3/2} K_{\max}) t^{-3/2} n^{-1/2}, \\ \widehat{\tau}_{4,t} &\leq (K_{\max} n) \lambda_{\min}^{-2}(\widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}) \leq (K_{\max} n) (C_3 n t)^{-2} = (C_3^{-2} K_{\max}) t^{-2} n^{-1}.\end{aligned}\tag{138}$$

Step 3: $\Delta_t, \epsilon_{n,t,\text{KL}}$

For $r > 4r_{\text{LA}}$ and $\theta \in \Theta(\widehat{\theta}_t, \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}, r)$, note that

$$\|\theta - \theta_0\|_2 \leq \|\theta - \widehat{\theta}_t\|_2 + \|\widehat{\theta}_t - \theta_0\|_2 \stackrel{(126)}{\leq} (C_3 n t)^{-1/2} r + \frac{1}{2} \stackrel{(127)}{\leq} r + \frac{1}{2}.$$

It follows that

$$\Theta(\widehat{\theta}_t, \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}, r) \subseteq \Theta(\theta_0, \mathbf{I}_p, 1/2 + r).$$

By (126), Lemma 42 and (36) in **(A2)**, the last display implies

$$\begin{aligned}\widehat{\tau}_{3,t,r} &\leq (C_3 n t)^{-3/2} K_{\max} N e^{8p} \exp\left(\left[\sqrt{p} + \sqrt{2 \log N} - 3\right] r\right) \\ &\stackrel{(127)}{\leq} N e^{8p} \exp\left(\left[\sqrt{p} + \sqrt{2 \log N} - 3\right] r\right).\end{aligned}\tag{139}$$

Also,

$$\begin{aligned}\widehat{\tau}_{3,t}^2 r_{\text{LA}}^2 &\stackrel{(137),(138)}{\leq} (C_3^{-3/2} K_{\max}) t^{-3/2} n^{-1/2} (16p_*) = (16C_3^{-3/2} K_{\max}) t^{-3/2} n^{-1/2} p_* \\ &\stackrel{(127)}{\leq} 1/8, \\ \widehat{\tau}_{4,t} p^2 &\stackrel{(138)}{\leq} (C_3^{-2} K_{\max}) t^{-2} n^{-1} p^2 \stackrel{(127)}{\leq} 1/8.\end{aligned}\tag{140}$$

By the last two displays, we can apply Theorem 3. Recall that

$$\epsilon_{n,t,\text{KL}}^2 = [\widehat{\tau}_{4,t} + \widehat{\tau}_{3,t}^2] p^2 + \widehat{\tau}_{3,t}^3 \log^3 N + e^{-7 \log N}.$$

By (138), $\epsilon_{n,t,\text{KL}}^2$ is bounded by

$$\begin{aligned}&\frac{K_{\max}}{C_3^2} t^{-2} n^{-1} p^2 + \frac{K_{\max}^2}{C_3^3} t^{-3} n^{-1} p^2 + \frac{K_{\max}^3}{C_3^{9/2}} t^{-9/2} n^{-3/2} \log^3 N + e^{-7 \log N} \\ &\leq c_1 \left[1 + t^{-1} + t^{-5/2} n^{-1/2} p_*^{-2} \log^3 N + e^{-5 \log N}\right] t^{-2} n^{-1} p_*^2 \\ &\lesssim c_1 \left[1 + t^{-1} + t^{-5/2} n^{-1/2} p_*^{-2} (\log^3 n + \log^3 T) + e^{-5 \log N}\right] t^{-2} n^{-1} p_*^2 \\ &\lesssim c_1 \left[1 + t^{-1} + t^{-5/2} n^{-1/2} (\log n + \log T) + e^{-5 \log N}\right] t^{-2} n^{-1} p_*^2 \\ &\stackrel{(127)}{\lesssim} c_1 t^{-2} n^{-1} p_*^2.\end{aligned}$$

for some constant $c_1 = c_1(K_{\min}, K_{\max}) > 0$. It follows that

$$\epsilon_{n,t,\text{KL}} \stackrel{(127)}{\leq} \delta$$

for a small enough constant $\delta > 0$, which further implies that

$$\Delta_t \stackrel{\text{Corollary 17}}{\leq} c'_1 \epsilon_{n,t,\text{KL}} \leq c_2 t^{-1} n^{-1/2} p_*, \quad (141)$$

where c'_1 is the universal constant in Corollary 17 and $c_2 = c_2(K_{\min}, K_{\max}) > 0$. By the results in **Step 1-3**, we completes the proof of (128).

Step 4: $\lambda_{\min}(\mathbf{\Omega}_t)$, $\lambda_{\max}(\mathbf{\Omega}_t)$

By (141), we have

$$\left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \mathbf{\Omega}_t \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_2 \stackrel{(100)}{\leq} \Delta_t \leq c_2 t^{-1} n^{-1/2} p_*.$$

It follows that

$$\begin{aligned} \lambda_{\min}(\mathbf{\Omega}_t) &\geq (1 - \Delta_t) \lambda_{\min}(\tilde{\mathbf{F}}_{t,\hat{\theta}_t}) \stackrel{(126)}{\geq} C_3 (1 - \Delta_t) nt, \\ \lambda_{\max}(\mathbf{\Omega}_t) &\leq (1 + \Delta_t) \lambda_{\max}(\tilde{\mathbf{F}}_{t,\hat{\theta}_t}) \stackrel{(126)}{\leq} C_4 (1 + \Delta_t) nt. \end{aligned} \quad (142)$$

Step 5: $\|\theta_{t+1}^* - \theta_0\|_2$, $\lambda_{\min}(\tilde{\mathbf{F}}_{t+1,\theta_{t+1}^*})$, $\lambda_{\max}(\tilde{\mathbf{F}}_{t+1,\theta_{t+1}^*})$, $\lambda_{\min}(\tilde{\mathbf{F}}_{t+1,\hat{\theta}_{t+1}})$, $\lambda_{\max}(\tilde{\mathbf{F}}_{t+1,\hat{\theta}_{t+1}})$

The results in **Step 1-4** and the assumptions can be summarized as follows:

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\theta_0 - \theta_t^*) \right\|_2 &\stackrel{(126)}{\leq} C_5 M_n t^\alpha \sqrt{p_*}, \\ \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\hat{\theta}_t - \theta_t^*) \right\|_2 &\stackrel{(136)}{\leq} 4r_{\text{eff},t} \stackrel{(132)}{\leq} (4C_3^{-1/2} K_{\max}^{1/2}) M_n \sqrt{t^{-1} p_*}, \\ \left\| \mathbf{\Omega}_t^{1/2} (\hat{\theta}_t - \mu_t) \right\|_2 \vee \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \mathbf{\Omega}_t \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_2 &\stackrel{(100)}{\leq} \Delta_t \stackrel{(141)}{\leq} c_2 t^{-1} n^{-1/2} p_* \stackrel{(127)}{\leq} 1/4, \\ \lambda_{\min}(\tilde{\mathbf{F}}_{t,\theta_t^*}) &\stackrel{(126)}{\geq} C_3 nt, \\ \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \tilde{\mathbf{F}}_{t,\hat{\theta}_t} \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} - \mathbf{I}_p \right\|_2 &\stackrel{(136)}{\leq} \stackrel{\text{Lemma 40}}{\leq} 4\tau_{3,t}^* r_{\text{eff},t} \stackrel{(135),(127)}{\leq} 1/8, \\ \|\theta_t^* - \theta_0\|_2 &\stackrel{(131)}{\leq} 1/8. \end{aligned} \quad (143)$$

By the last display, we can apply Lemma 48 with

$$D_1 = C_5, \quad D_2 = 4C_3^{-1/2} K_{\max}^{1/2}, \quad D_3 = c_2, \quad D_4 = C_3,$$

where the choice $D_3 = c_2$ is justified by $p_* \leq n$. By Lemma 48, we have

$$\left\| \mathbf{F}_{t+1,\theta_t^*}^{1/2} (\theta_{t+1}^* - \theta_t^*) \right\|_2 \leq C' M_n t^{\alpha-1/2} \sqrt{p_*}.$$

where $C' = C'(K_{\min}, K_{\max}, C_5) > 0$. Consequently, we have

$$\begin{aligned} \|\theta_{t+1}^* - \theta_0\|_2 &\leq \|\theta_{t+1}^* - \theta_t^*\|_2 + \|\theta_t^* - \theta_0\|_2 \\ &\stackrel{(143), (\mathbf{A2})}{\leq} (K_{\min}n)^{-1/2} C' M_n t^{\alpha-1/2} \sqrt{p_*} + 1/8 \\ &= (K_{\min}^{-1/2} C') (M_n^2 t^{2\alpha-1} p_* n^{-1})^{1/2} + 1/8 \stackrel{(127)}{\leq} 1/4. \end{aligned} \quad (144)$$

Since $\theta_{t+1}^* \in \Theta(\theta_0, \mathbf{I}_p, 1/2)$, combining the results in **Step 4** and **(A2)**, we have

$$\begin{aligned} \lambda_{\min}(\tilde{\mathbf{F}}_{t+1, \theta_{t+1}^*}) &\geq \lambda_{\min}(\mathbf{\Omega}_t) + \lambda_{\min}(\mathbf{F}_{t+1, \theta_{t+1}^*}) \stackrel{(142)}{\geq} C_3 (1 - \Delta_t) nt + K_{\min}n \\ &\geq c_{3,t}n(t+1), \end{aligned} \quad (145)$$

$$\begin{aligned} \lambda_{\max}(\tilde{\mathbf{F}}_{t+1, \theta_{t+1}^*}) &\leq \lambda_{\max}(\mathbf{\Omega}_t) + \lambda_{\max}(\mathbf{F}_{t+1, \theta_{t+1}^*}) \stackrel{(142)}{\leq} C_4 (1 + \Delta_t) nt + K_{\max}n \\ &\leq c_{4,t}n(t+1), \end{aligned}$$

where

$$\begin{aligned} c_{3,t} &= C_3 (1 - \Delta_t) \in [3K_{\min}/8, K_{\min}], \quad (\because \Delta_t \leq 1/4) \\ c_{4,t} &= C_4 (1 + \Delta_t) \in [K_{\max}, 5K_{\max}/2], \quad (\because \Delta_t \leq 1/4). \end{aligned} \quad (146)$$

The proofs of $\lambda_{\min}(\tilde{\mathbf{F}}_{t+1, \hat{\theta}_{t+1}})$ and $\lambda_{\max}(\tilde{\mathbf{F}}_{t+1, \hat{\theta}_{t+1}})$ are similar to those in **Step 1-2**. Hence, we give a sketch of the proof. Since $c_{3,t} \geq 3K_{\min}/8$, we have

$$\begin{aligned} r_{\text{eff}, t+1} &\stackrel{(144)}{\leq} \stackrel{\text{Lemma 18}}{M_n} \lambda_{\min}^{-1/2}(\tilde{\mathbf{F}}_{t+1, \theta_{t+1}^*}) \lambda_{\max}^{1/2}(\mathbf{F}_{t+1, \theta_{t+1}^*}) p_*^{1/2} \\ &\stackrel{(144), (\mathbf{A2})}{\leq} M_n (c_{3,t}n[t+1])^{-1/2} (K_{\max}n)^{1/2} p_*^{1/2} = \left(c_{3,t}^{-1/2} K_{\max}^{1/2}\right) M_n \sqrt{(t+1)^{-1} p_*}, \\ \sup_{\theta \in \Theta_{n, t+1}} \|\theta - \theta_0\|_2 &\leq \sup_{\theta \in \Theta_{n, t+1}} \|\theta - \theta_{t+1}^*\|_2 + \|\theta_{t+1}^* - \theta_0\|_2 \stackrel{(144)}{\leq} \lambda_{\min}^{-1/2}(\tilde{\mathbf{F}}_{t+1, \theta_{t+1}^*}) 4r_{\text{eff}, t+1} + 1/4 \\ &\stackrel{(145)}{\leq} (4c_{3,t}^{-1} K_{\max}^{1/2}) M_n p_*^{1/2} n^{-1/2} (t+1)^{-1} + 1/4 \stackrel{(127)}{\leq} 1/2, \\ \tau_{3, t+1}^* &\stackrel{(\mathbf{A2})}{\leq} \stackrel{\text{Lemma 42}}{(K_{\max}n)} \lambda_{\min}^{-3/2}(\tilde{\mathbf{F}}_{t+1, \theta_{t+1}^*}) \stackrel{(126)}{\leq} (K_{\max}n) (c_{3,t}n[t+1])^{-3/2} \\ &= (c_{3,t}^{-3/2} K_{\max}) (t+1)^{-3/2} n^{-1/2}, \\ \tau_{3, t+1}^* r_{\text{eff}, t+1} &\leq (c_{3,t}^{-2} K_{\max}^{3/2}) M_n p_*^{1/2} (t+1)^{-2} n^{-1/2} \stackrel{(127)}{\leq} 1/16, \\ \hat{\theta}_{t+1} &\stackrel{\text{Theorem 5}}{\in} \Theta_{n, t+1}, \quad (\because \tau_{3, t+1}^* r_{\text{eff}, t+1} \leq 1/16) \end{aligned}$$

where $\Theta_{n, t+1} = \Theta(\theta_{t+1}^*, \tilde{\mathbf{F}}_{t+1, \theta_{t+1}^*}, 4r_{\text{eff}, t+1})$. Since the last display gives

$$\|\hat{\theta}_{t+1} - \theta_0\|_2 \leq 1/2,$$

we have

$$\begin{aligned}
 \lambda_{\min}(\tilde{\mathbf{F}}_{t+1, \hat{\theta}_{t+1}}) &\geq \lambda_{\min}(\mathbf{\Omega}_t) + \lambda_{\min}(\mathbf{F}_{t+1, \hat{\theta}_{t+1}}) \stackrel{(142)}{\geq} C_3(1 - \Delta_t)nt + K_{\min}n \\
 &\geq c_{3,t}n(t+1), \\
 \lambda_{\max}(\tilde{\mathbf{F}}_{t+1, \hat{\theta}_{t+1}}) &\leq \lambda_{\max}(\mathbf{\Omega}_t) + \lambda_{\max}(\mathbf{F}_{t+1, \hat{\theta}_{t+1}}) \stackrel{(142)}{\leq} C_4(1 + \Delta_t)nt + K_{\max}n \\
 &\leq c_{4,t}n(t+1),
 \end{aligned}$$

which completes the proof of the first three assertions in (129).

Step 6: $\tau_{3,t+1, \text{bias}}, \tau_{4,t+1, \text{bias}}$

Note that

$$\begin{aligned}
 b_{n,t+1} &= \left\| \tilde{\mathbf{F}}_{t+1, \hat{\theta}_{t+1}}^{-1/2} \mathbf{\Omega}_t (\theta_0 - \mu_t) \right\|_2 \leq \left\| \mathbf{\Omega}_t^{-1/2} \mathbf{\Omega}_t (\theta_0 - \mu_t) \right\|_2 = \left\| \mathbf{\Omega}_t^{1/2} (\theta_0 - \mu_t) \right\|_2 \\
 &\leq \left\| \mathbf{\Omega}_t^{1/2} (\theta_0 - \theta_t^*) \right\|_2 + \left\| \mathbf{\Omega}_t^{1/2} (\theta_t^* - \hat{\theta}_t) \right\|_2 + \left\| \mathbf{\Omega}_t^{1/2} (\hat{\theta}_t - \mu_t) \right\|_2 \\
 &\leq \left\| \mathbf{\Omega}_t^{1/2} \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} \right\|_2 \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} \tilde{\mathbf{F}}_{t, \theta_t^*}^{-1/2} \right\|_2 \left[\left\| \tilde{\mathbf{F}}_{t, \theta_t^*}^{1/2} (\theta_0 - \theta_t^*) \right\|_2 + \left\| \tilde{\mathbf{F}}_{t, \theta_t^*}^{1/2} (\theta_t^* - \hat{\theta}_t) \right\|_2 \right] + \left\| \mathbf{\Omega}_t^{1/2} (\hat{\theta}_t - \mu_t) \right\|_2 \\
 &\leq \left(1 + \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} \mathbf{\Omega}_t \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_2 \right)^{1/2} \left(1 + \left\| \tilde{\mathbf{F}}_{t, \theta_t^*}^{-1/2} \tilde{\mathbf{F}}_{t, \hat{\theta}_t} \tilde{\mathbf{F}}_{t, \theta_t^*}^{-1/2} - \mathbf{I}_p \right\|_2 \right)^{1/2} \\
 &\quad \times \left[\left\| \tilde{\mathbf{F}}_{t, \theta_t^*}^{1/2} (\theta_0 - \theta_t^*) \right\|_2 + \left\| \tilde{\mathbf{F}}_{t, \theta_t^*}^{1/2} (\theta_t^* - \hat{\theta}_t) \right\|_2 \right] + \left\| \mathbf{\Omega}_t^{1/2} (\hat{\theta}_t - \mu_t) \right\|_2 \\
 &\stackrel{(143), (126)}{\leq} (1 + \Delta_t)^{1/2} (1 + 4\tau_{3,t}^* r_{\text{eff},t})^{1/2} (C_5 M_n t^\alpha \sqrt{p_*} + 4r_{\text{eff},t}) + \Delta_t \\
 &\stackrel{(132), (143)}{\leq} 2 \left(C_5 M_n t^\alpha \sqrt{p_*} + (4C_3^{-1/2} K_{\max}^{1/2}) M_n \sqrt{t^{-1} p_*} \right) + c_2 t^{-1} n^{-1/2} p_* \\
 &\stackrel{(127)}{\leq} c_5 M_n t^\alpha \sqrt{p_*},
 \end{aligned}$$

where $c_5 = c_5(K_{\min}, K_{\max}, C_5) > 0$. For $\theta \in \Theta(\theta_{t+1}^*, \tilde{\mathbf{F}}_{t+1, \theta_{t+1}^*}, 4b_{n,t+1})$, we have

$$\begin{aligned}
 \left\| \theta - \theta_{t+1}^* \right\|_2 &\stackrel{(145)}{\leq} (c_{3,t}n(t+1))^{-1/2} 4b_{n,t+1} \leq (c_{3,t}n(t+1))^{-1/2} 4c_5 M_n t^\alpha \sqrt{p_*} \\
 &\leq (4c_5(c_{3,t})^{-1/2}) \sqrt{M_n^2 t^{2\alpha-1} p_* n^{-1}} \stackrel{(127)}{\leq} 1/4.
 \end{aligned}$$

Combining with (144), we have

$$\Theta_n \left(\theta_{t+1}^*, \tilde{\mathbf{F}}_{t+1, \theta_{t+1}^*}, 4b_{n,t+1} \right) \subseteq \Theta(\theta_0, \mathbf{I}_p, 1/2).$$

By Lemma 42 and (A2), we have

$$\begin{aligned}
 \tau_{3,t+1, \text{bias}} &\leq (K_{\max}n) \lambda_{\min}^{-3/2}(\tilde{\mathbf{F}}_{t+1, \theta_{t+1}^*}) \leq (K_{\max}n) (c_{3,t}n(t+1))^{-3/2} \\
 &= K_{\max}(c_{3,t})^{-3/2} (t+1)^{-3/2} n^{-1/2}.
 \end{aligned} \tag{147}$$

Similarly,

$$\begin{aligned}\tau_{4,t+1,\text{bias}} &\leq (K_{\max}n) \lambda_{\min}^{-2}(\tilde{\mathbf{F}}_{t+1,\theta_{t+1}^*}) \leq (K_{\max}n) (c_{3,t}n(t+1))^{-2} \\ &= K_{\max}(c_{3,t})^{-2}(t+1)^{-2}n^{-1}.\end{aligned}\quad (148)$$

Step 7: $\|\tilde{\mathbf{F}}_{t+1,\theta_{t+1}^*}^{1/2}(\theta_0 - \theta_{t+1}^*)\|_2$

To complete the proof of the last inequality in (129), we will show that the assumptions in Lemma 20 are satisfied. By the results in **Step 1-6**, we have

$$\begin{aligned}\forall r > 4r_{\text{LA}}, \quad \hat{\tau}_{3,t,r} &\stackrel{(139)}{\leq} Ne^{8p} \exp\left(\left[\sqrt{p} + \sqrt{2\log N} - 3\right]r\right), \\ \hat{\tau}_{3,t}r_{\text{LA}}^2 &\stackrel{(140)}{\leq} c_6t^{-3/2}n^{-1/2}p_*, \quad \hat{\tau}_{4,t}p^2 \stackrel{(140)}{\leq} c_6t^{-2}n^{-1}p^2, \\ \tau_{3,t}^*r_{\text{eff},t} &\stackrel{(135)}{\leq} c_6t^{-2}n^{-1/2}M_n p_*^{1/2}, \quad \epsilon_{n,t,\text{KL}} \vee \Delta_t \stackrel{(141)}{\leq} c_6t^{-1}n^{-1/2}p_*, \\ \tau_{3,t+1,\text{bias}}(M_n t^{1/2+2\alpha}\sqrt{p_*}) &\stackrel{(147)}{\leq} c_6M_n t^{2\alpha-1}\sqrt{p_*}n^{-1/2}, \\ \tau_{4,t+1,\text{bias}}(M_n^2 t^{1/2+3\alpha}p_*) &\stackrel{(148)}{\leq} c_6M_n^2 t^{3\alpha-3/2}p_*n^{-1}, \\ \tau_{3,t+1,\text{bias}}^2(M_n^2 t^{1/2+3\alpha}p_*) &\stackrel{(147)}{\leq} c_6M_n^2 t^{3\alpha-5/2}p_*n^{-1}, \\ (\epsilon_{n,t,\text{KL}} \vee [\tau_{3,t}^*r_{\text{eff},t}]) t^{1/2+\alpha} &\stackrel{(135)}{\stackrel{(141)}{\leq}} c_6M_n t^{\alpha-1/2}p_*n^{-1/2},\end{aligned}$$

where $c_6 = c_6(K_{\min}, K_{\max})$. By the last display and (127), the assumptions (116), (117) and (118) in Lemma 20 are satisfied. Also, we have

$$r_{\text{eff},t} \stackrel{(132)}{\leq} (C_3^{-1/2}K_{\max}^{1/2})M_n\sqrt{t^{-1}p_*}, \quad \left\|\tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2}(\theta_0 - \theta_t^*)\right\|_2 \stackrel{(126)}{\leq} C_5M_n t^\alpha\sqrt{p_*}.$$

Hence, the assumption (115) in Lemma 20 are satisfied with $C_1 = C_3^{-1/2}K_{\max}^{1/2}$ and $C_2 = C_5$. Recall that

$$C_5 \geq 192C_3^{-1/2}K_{\max}^{1/2} \geq (2^{1/\alpha})48(C_3^{-1/2}K_{\max}^{1/2}), \quad C_5 \geq 1,$$

which is chosen to be sufficiently large so that we can replace K in (119) with C_5 ; see the remark following Lemma 20. By Lemma 20 with $K = C_5$, we have

$$\begin{aligned}\left\|\tilde{\mathbf{F}}_{t+1,\theta_{t+1}^*}^{1/2}(\theta_0 - \theta_{t+1}^*)\right\|_2 &\leq C_5M_n(t^{1/2-\alpha} + t)^\alpha\sqrt{p_*} \\ &\leq C_5M_n(t+1)^\alpha\sqrt{p_*}, \quad (\because \alpha \in [1/2, 1]),\end{aligned}$$

which completes the proof of (129).

All remaining proofs for (130) are similar to those in **Step 1-4**, but replace $c_{3,t}$, $c_{4,t}$ and t with $K_{\min}/2$, $2K_{\max}$ and $t+1$, respectively. \blacksquare

Proposition 23 *Suppose that (A0), (A1), (A2), (S) and (P) hold. Then, on $\mathcal{E}_{\text{est},1}$, the following inequalities hold uniformly for all $t \in [T]$:*

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2}(\theta_0 - \theta_t^*) \right\|_2 &\leq K_{\text{up}} M_n \sqrt{t p_*}, & \lambda_{\min}(\tilde{\mathbf{F}}_{t,\theta_t^*}) \wedge \lambda_{\min}(\tilde{\mathbf{F}}_{t,\hat{\theta}_t}) &\geq K_{\text{low}} n t, \\ \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2}(\hat{\theta}_t - \theta_t^*) \right\|_2 &\leq K_{\text{up}} M_n \sqrt{t^{-1} p_*}, & \lambda_{\max}(\tilde{\mathbf{F}}_{t,\theta_t^*}) \vee \lambda_{\max}(\tilde{\mathbf{F}}_{t,\hat{\theta}_t}) &\leq K_{\text{up}} n t, \end{aligned} \quad (149)$$

and

$$\begin{aligned} \hat{\tau}_{3,t} \vee \tau_{3,t}^* &\leq K_{\text{up}} t^{-3/2} n^{-1/2}, \\ \hat{\tau}_{4,t} &\leq K_{\text{up}} t^{-2} n^{-1}, \\ \Delta_t \vee \epsilon_{n,t,\text{KL}} &\leq K_{\text{up}} t^{-1} n^{-1/2} p_*, \end{aligned} \quad (150)$$

where $K_{\text{up}} = K_{\text{up}}(K_{\min}, K_{\max})$ and $K_{\text{low}} = K_{\text{low}}(K_{\min})$.

Proof In this proof, we will work on the event $\mathcal{E}_{\text{est},1}$ without explicitly mentioning it. Also, although the value of α in Lemma 22 can range over the interval $[1/2, 1]$, we fix $\alpha = 1/2$ in this proof. Later, we specify C_5 in Lemma 22 as a constant depending only on K_{\min} and K_{\max} . Consequently, (S) implies the condition (127) with $\alpha = 1/2$.

Step 1: Proof strategy

Suppose that the following inequalities hold for some $t \in [T]$:

$$\begin{aligned} \lambda_{\min}(\tilde{\mathbf{F}}_{t,\theta_t^*}) \wedge \lambda_{\min}(\tilde{\mathbf{F}}_{t,\hat{\theta}_t}) &\geq \frac{K_{\min}}{2} n t, \\ \lambda_{\max}(\tilde{\mathbf{F}}_{t,\theta_t^*}) \vee \lambda_{\max}(\tilde{\mathbf{F}}_{t,\hat{\theta}_t}) &\leq 2K_{\max} n t, \\ \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2}(\theta_0 - \theta_t^*) \right\|_2 &\leq (4 \vee [192\sqrt{2}K_{\min}^{-1/2}K_{\max}^{1/2}]) M_n \sqrt{t p_*}. \end{aligned} \quad (151)$$

By Lemma 22 with

$$C_3 = \frac{K_{\min}}{2}, \quad C_4 = 2K_{\max} n, \quad C_5 = 4 \vee (192\sqrt{2}K_{\min}^{-1/2}K_{\max}^{1/2}),$$

we have

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2}(\hat{\theta}_t - \theta_t^*) \right\|_2 &\leq K_1 M_n \sqrt{t^{-1} p_*}, \\ \tau_{3,t}^* \vee \hat{\tau}_{3,t} &\leq K_1 t^{-3/2} n^{-1/2}, \\ \hat{\tau}_{4,t} &\leq K_1 t^{-2} n^{-1}, \\ \Delta_t \vee \epsilon_{n,t,\text{KL}} &\leq K_1 t^{-1} n^{-1/2} p_*, \end{aligned}$$

where $K_1 = K_1(K_{\min}, K_{\max})$ denotes the constant K specified in Lemma 22. If (151) holds for all $t \in [T]$, then we complete the proof by taking

$$K_{\text{up}} = 2K_{\max} \vee K_1 \vee 4 \vee (192\sqrt{2}K_{\min}^{-1/2}K_{\max}^{1/2}), \quad K_{\text{low}} = K_{\min}/2.$$

Therefore, it suffices to show that (151) holds for all $t \in [T]$. The proof is divided into several steps. We first establish—through **Step 2-4**—that (151) holds for $t = 1$. Subsequently, by using induction, we will prove in **Step 5** that (151) holds for all $t \in [T]$.

Step 2: $\lambda_{\min}(\tilde{\mathbf{F}}_{1,\theta_1^*})$ and $\lambda_{\max}(\tilde{\mathbf{F}}_{1,\theta_1^*})$

By **(P)** with a small enough $\delta = \delta(K_{\min}, K_{\max})$, the assumption (211) in Lemma 45 is satisfied for $t = 0$. By Lemma 45, we have

$$\left\| \mathbf{F}_{1,\theta_0}^{1/2} (\theta_0 - \theta_1^*) \right\|_2 \leq \sqrt{2} \left\| \boldsymbol{\Omega}_0^{1/2} (\theta_0 - \mu_0) \right\|_2 \stackrel{\text{(P)}}{\leq} (\sqrt{2}\delta)n^{1/2},$$

which further implies that

$$\begin{aligned} \|\theta_0 - \theta_1^*\|_2 &\leq \lambda_{\min}^{-1/2}(\mathbf{F}_{1,\theta_0}) (\sqrt{2}\delta)n^{1/2} \stackrel{\text{(A2)}}{\leq} (K_{\min}n)^{-1/2} (\sqrt{2}\delta)n^{1/2} \stackrel{\text{(P)}}{\leq} 1/8, \\ \lambda_{\min}(\tilde{\mathbf{F}}_{1,\theta_1^*}) &\geq \lambda_{\min}(\mathbf{F}_{1,\theta_1^*}) + \lambda_{\min}(\boldsymbol{\Omega}_0) \stackrel{\text{(A2)}}{\geq} K_{\min}n, \\ \lambda_{\max}(\tilde{\mathbf{F}}_{1,\theta_1^*}) &\leq \lambda_{\max}(\boldsymbol{\Omega}_0) + \lambda_{\max}(\mathbf{F}_{1,\theta_1^*}) \stackrel{\text{(A2),(P)}}{\leq} K_{\max}p_* + K_{\max}n \stackrel{\text{(S)}}{\leq} \frac{4}{3}K_{\max}n. \end{aligned} \tag{152}$$

Step 3: $\|\tilde{\mathbf{F}}_{1,\theta_1^*}^{1/2} (\theta_0 - \theta_1^*)\|_2$

Note that

$$\begin{aligned} b_{n,1} &= \left\| \tilde{\mathbf{F}}_{1,\theta_1^*}^{-1/2} \boldsymbol{\Omega}_0 (\theta_0 - \mu_0) \right\|_2 \leq \lambda_{\min}^{-1/2}(\tilde{\mathbf{F}}_{1,\theta_1^*}) \|\boldsymbol{\Omega}_0\|_2^{1/2} \left\| \boldsymbol{\Omega}_0^{1/2} (\theta_0 - \mu_0) \right\|_2 \\ &\leq (K_{\min}n)^{-1/2} (K_{\max}p_*)^{1/2} (\delta n^{1/2}) = (K_{\min}^{-1/2} K_{\max}^{1/2} \delta) p_*^{1/2} \leq M_n p_*^{1/2}, \end{aligned} \tag{153}$$

where the last inequality holds by $M_n \geq 1$ and small enough δ . Also, for $\theta \in \Theta(\theta_1^*, \tilde{\mathbf{F}}_{1,\theta_1^*}, 4b_{n,1})$, we have

$$\|\theta - \theta_1^*\|_2 \leq \lambda_{\min}^{-1/2}(\tilde{\mathbf{F}}_{1,\theta_1^*}) 4b_{n,1} \stackrel{(152),(153)}{\leq} (K_{\min}n)^{-1/2} 4M_n \sqrt{p_*} \stackrel{\text{(S)}}{\leq} 1/4,$$

which, combining with $\|\theta_0 - \theta_1^*\|_2 \leq 1/8$, implies that $\Theta(\theta_1^*, \tilde{\mathbf{F}}_{1,\theta_1^*}, 4b_{n,1}) \subset \Theta(\theta_0, \mathbf{I}_p, 1/2)$. Hence, Lemma 42 implies that

$$\begin{aligned} \tau_{3,1,\text{bias}} &\stackrel{\text{(A2)}}{\leq} (K_{\max}n) (K_{\min}n)^{-3/2} = (K_{\max}K_{\min}^{-3/2})n^{-1/2}, \\ \tau_{4,1,\text{bias}} &\stackrel{\text{(A2)}}{\leq} (K_{\max}n) (K_{\min}n)^{-2} = (K_{\max}K_{\min}^{-2})n^{-1}. \end{aligned}$$

It follows that

$$\tau_{3,1,\text{bias}} b_{n,1} \leq \left((K_{\max}K_{\min}^{-3/2})n^{-1/2} \right) (M_n \sqrt{p_*}) \stackrel{\text{(S)}}{\leq} 1/16,$$

which allows us to utilize Lemma 41. By Lemma 41 with

$$\tau_3 = \tau_{3,1,\text{bias}}, \quad f(\theta) = \mathbb{E}_1 \tilde{L}_1(\theta), \quad \theta = \theta_1^*, \quad \tilde{\theta} = \theta_0, \quad \beta = \boldsymbol{\Omega}_0(\theta_0 - \mu_0), \quad r = b_{n,1},$$

we have

$$\left\| \tilde{\mathbf{F}}_{1,\theta_1^*}^{1/2} (\theta_0 - \theta_1^*) \right\|_2 \leq 4b_{n,1} \stackrel{(153)}{\leq} 4M_n \sqrt{p_*}. \tag{154}$$

Step 4: $\lambda_{\min}(\tilde{\mathbf{F}}_{1,\hat{\theta}_1})$ and $\lambda_{\max}(\tilde{\mathbf{F}}_{1,\hat{\theta}_1})$
 Combining with $\|\theta_0 - \theta_1^*\|_2 \leq 1/8$, we have

$$\begin{aligned} r_{\text{eff},1} &\stackrel{\text{Lemma 18}}{\leq} M_n \lambda_{\min}^{-1/2}(\tilde{\mathbf{F}}_{1,\theta_1^*}) \lambda_{\max}^{1/2}(\mathbf{F}_{1,\theta_1^*}) p_*^{1/2} \\ &\stackrel{(152),(\mathbf{A2})}{\leq} M_n (K_{\min} n)^{-1/2} (K_{\max} n)^{1/2} p_*^{1/2} = (K_{\min}^{-1} K_{\max})^{1/2} M_n p_*^{1/2}. \end{aligned} \quad (155)$$

Also, for $\theta \in \Theta(\theta_1^*, \tilde{\mathbf{F}}_{1,\theta_1^*}, 4r_{\text{eff},1})$,

$$\begin{aligned} \|\theta - \theta_1^*\|_2 &\leq \lambda_{\min}^{-1/2}(\tilde{\mathbf{F}}_{1,\theta_1^*}) 4r_{\text{eff},1} \stackrel{(152),(155)}{\leq} (K_{\min} n)^{-1/2} 4 (K_{\min}^{-1} K_{\max})^{1/2} M_n p_*^{1/2} \\ &= (4K_{\min}^{-1} K_{\max}^{1/2}) M_n p_*^{1/2} n^{-1/2} \stackrel{(\mathbf{S})}{\leq} 1/4. \end{aligned} \quad (156)$$

It follows that $\Theta(\theta_1^*, \tilde{\mathbf{F}}_{1,\theta_1^*}, 4r_{\text{eff},1}) \subset \Theta(\theta_0, \mathbf{I}_p, 1/2)$. Hence, we have

$$\begin{aligned} \tau_{3,1}^* &\stackrel{\text{Lemma 42}}{\stackrel{(\mathbf{A2}), (152)}{\leq}} (K_{\min} n)^{-3/2} (K_{\max} n) \leq (K_{\max} K_{\min}^{-3/2}) n^{-1/2}, \\ \tau_{3,1}^* r_{\text{eff},1} &\stackrel{(155)}{\leq} (K_{\max}^{3/2} K_{\min}^{-2}) M_n p_*^{1/2} n^{-1/2} \stackrel{(\mathbf{S})}{\leq} 1/16. \end{aligned}$$

By the last display, Theorem 5 implies that

$$\left\| \tilde{\mathbf{F}}_{1,\theta_1^*}^{1/2} (\hat{\theta}_1 - \theta_1^*) \right\|_2 \leq 4r_{\text{eff},1}, \quad \left\| \hat{\theta}_1 - \theta_1^* \right\|_2 \stackrel{(156)}{\leq} 1/4,$$

which further implies that $\hat{\theta}_1 \in \Theta(\theta_0, \mathbf{I}_p, 1/2)$. It follows that

$$\begin{aligned} \lambda_{\min}(\tilde{\mathbf{F}}_{1,\hat{\theta}_1}) &\geq \lambda_{\min}(\mathbf{F}_{1,\hat{\theta}_1}) + \lambda_{\min}(\mathbf{\Omega}_0) \stackrel{(\mathbf{A2})}{\geq} K_{\min} n, \\ \lambda_{\max}(\tilde{\mathbf{F}}_{1,\hat{\theta}_1}) &\leq \lambda_{\max}(\mathbf{\Omega}_0) + \lambda_{\max}(\mathbf{F}_{1,\hat{\theta}_1}) \stackrel{(\mathbf{A2}),(\mathbf{P})}{\leq} K_{\max} p_* + K_{\max} n \stackrel{(\mathbf{S})}{\leq} \frac{4}{3} K_{\max} n. \end{aligned} \quad (157)$$

Combining (152), (154) and (157), one can check that (151) holds for $t = 1$.

Step 5: Inductive argument

We complete this proof by employing an inductive argument. Let $t_0 \in \{2, 3, \dots, T\}$. Suppose that the following inequalities hold:

$$\begin{aligned} \lambda_{\min}(\tilde{\mathbf{F}}_{s,\theta_s^*}) \wedge \lambda_{\min}(\tilde{\mathbf{F}}_{s,\hat{\theta}_s}) &\geq c_1 n s, \\ \lambda_{\max}(\tilde{\mathbf{F}}_{s,\theta_s^*}) \vee \lambda_{\max}(\tilde{\mathbf{F}}_{s,\hat{\theta}_s}) &\leq c_2 n s, \\ \left\| \tilde{\mathbf{F}}_{s,\theta_s^*}^{1/2} (\theta_0 - \theta_s^*) \right\|_2 &\leq c_3 M_n s^{1/2} \sqrt{p_*} \end{aligned} \quad (158)$$

for all $s \in [t_0 - 1]$, where

$$c_1 \in [K_{\min}/2, K_{\min}], \quad c_2 \in [K_{\max}, 2K_{\max}], \quad c_3 = 4 \vee (192\sqrt{2}K_{\min}^{-1/2}K_{\max}^{1/2}).$$

By applying Lemma 22 with $\alpha = 1/2$, it follows from (129) that

$$\begin{aligned}
 \lambda_{\min}(\tilde{\mathbf{F}}_{t_0, \theta_{t_0}^*}) \wedge \lambda_{\min}(\tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}) &\geq c_{t_0} n t_0, \\
 \lambda_{\max}(\tilde{\mathbf{F}}_{t_0, \theta_{t_0}^*}) \vee \lambda_{\max}(\tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}) &\leq c'_{t_0} n t_0, \\
 \left\| \tilde{\mathbf{F}}_{t_0, \theta_{t_0}^*}^{1/2} (\theta_0 - \theta_{t_0}^*) \right\|_2 &\leq c_3 M_n t_0^{1/2} \sqrt{p_*}
 \end{aligned} \tag{159}$$

for some constants $c_{t_0}, c'_{t_0} > 0$. Next, we prove the following inequalities:

$$c_{t_0} \geq \frac{K_{\min}}{2}, \quad c'_{t_0} \leq 2K_{\max}, \quad \forall t_0 \in \{2, 3, \dots, T\}, \tag{160}$$

which, combining with the result in **Step 0**, completes the proof by induction.

By (128), (129) in Lemma 22, (158) implies that

$$\begin{aligned}
 \lambda_{\min}(\mathbf{F}_{t, \hat{\theta}_t}) \wedge \lambda_{\min}(\mathbf{F}_{t, \theta_t^*}) &\stackrel{\text{(A2)}}{\geq} K_{\min} n, & \forall t \in [t_0], \\
 \lambda_{\max}(\mathbf{F}_{t, \hat{\theta}_t}) \wedge \lambda_{\max}(\mathbf{F}_{t, \theta_t^*}) &\stackrel{\text{(A2)}}{\leq} K_{\max} n, & \forall t \in [t_0], \\
 \Delta_s &\stackrel{\text{(S)}}{\leq} K_1 s^{-1} n^{-1/2} p_* \leq 1/4, & \forall s \in [t_0 - 1],
 \end{aligned} \tag{161}$$

where the first two inequalities hold because $\theta_t^*, \hat{\theta}_t \in \Theta(\theta_0, \mathbf{I}_p, 1/2)$ for all $t \in [t_0]$. Also, for any $t \in [T]$, we have

$$\begin{aligned}
 \tilde{\mathbf{F}}_{t, \hat{\theta}_t} &= \mathbf{\Omega}_{t-1} + \mathbf{F}_{t, \hat{\theta}_t} \\
 &\succeq (1 - \Delta_{t-1}) \tilde{\mathbf{F}}_{t-1, \hat{\theta}_{t-1}} + \mathbf{F}_{t, \hat{\theta}_t} \\
 &\succeq (1 - \Delta_{t-2})(1 - \Delta_{t-1}) \tilde{\mathbf{F}}_{t-2, \hat{\theta}_{t-2}} + (1 - \Delta_{t-1}) \mathbf{F}_{t-1, \hat{\theta}_{t-1}} + \mathbf{F}_{t, \hat{\theta}_t} \\
 &\succeq \sum_{s=1}^{t-1} \left(\left[\prod_{r=1}^s (1 - \Delta_{t-r}) \right] \mathbf{F}_{t-s, \hat{\theta}_{t-s}} \right) + \mathbf{F}_{t, \hat{\theta}_t}, \\
 \tilde{\mathbf{F}}_{t, \theta_t^*} &= \mathbf{\Omega}_{t-1} + \mathbf{F}_{t, \theta_t^*} \\
 &\succeq \sum_{s=1}^{t-1} \left(\left[\prod_{r=1}^s (1 - \Delta_{t-r}) \right] \mathbf{F}_{t-s, \hat{\theta}_{t-s}} \right) + \mathbf{F}_{t, \theta_t^*}.
 \end{aligned} \tag{162}$$

It follows that

$$\begin{aligned}
 & \lambda_{\min}(\tilde{\mathbf{F}}_{t_0, \theta_{t_0}^*}) \wedge \lambda_{\min}(\tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}) \\
 & \stackrel{(161)}{\geq} \sum_{s=1}^{t_0-1} \left(\left[\prod_{r=1}^s (1 - \Delta_{t_0-r}) \right] K_{\min} n \right) + K_{\min} n \\
 & \stackrel{(162)}{\geq} \sum_{s=1}^{t_0-1} \left(\exp \left[-2 \sum_{r=1}^s \Delta_{t_0-r} \right] K_{\min} n \right) + K_{\min} n \\
 & \geq \sum_{s=1}^{t_0-1} \left(\left[1 - 2 \sum_{r=1}^s \Delta_{t_0-r} \right] K_{\min} n \right) + K_{\min} n \\
 & \geq \sum_{s=1}^{t_0-1} \left(\left[1 - 2 \sum_{r=1}^{t_0-1} \Delta_r \right] K_{\min} n \right) + K_{\min} n, \quad (\because \Delta_r \geq 0, \quad \forall r \in [t_0 - 1])
 \end{aligned}$$

where the second and third inequalities hold by $1-x \geq e^{-2x}$ for $x \in [0, 0.795]$ and $e^{-x} \geq 1-x$ for $x \in \mathbb{R}$, respectively. Suppose that the following inequality holds:

$$\sum_{s=1}^{t_0-1} \Delta_s \leq 1/4, \quad \forall t_0 \in \{2, 3, \dots, T\}. \quad (163)$$

Then, for any $t_0 \in \{2, 3, \dots, T\}$, we have

$$\sum_{s=1}^{t_0-1} \left(\left[1 - 2 \sum_{r=1}^{t_0-1} \Delta_{t_0-r} \right] K_{\min} n \right) + K_{\min} n \geq \frac{K_{\min}}{2} n t_0.$$

Hence, we need to show (163). Note that

$$\begin{aligned}
 \sum_{s=1}^{t_0-1} \Delta_s & \stackrel{(161)}{\leq} \sum_{s=1}^{t_0-1} K_1 p_* n^{-1/2} s^{-1} = K_1 p_* n^{-1/2} \sum_{s=1}^{t_0-1} s^{-1} \leq K_1 p_* n^{-1/2} \sum_{s=1}^T s^{-1} \\
 & \leq K_1 p_* n^{-1/2} (\log T + 1) \leq 2K_1 p_* n^{-1/2} \log(T \vee 3) \stackrel{(S)}{\leq} 1/4.
 \end{aligned}$$

As in (162), for any $t \in [T]$, we have

$$\begin{aligned}
 \tilde{\mathbf{F}}_{t, \hat{\theta}_t} & \preceq \left[\prod_{r=1}^{t-1} (1 + \Delta_r) \right] \mathbf{\Omega}_0 + \sum_{s=1}^{t-1} \left(\left[\prod_{r=1}^s (1 + \Delta_{t-r}) \right] \mathbf{F}_{t-s, \hat{\theta}_{t-s}} \right) + \mathbf{F}_{t, \hat{\theta}_t}, \\
 \tilde{\mathbf{F}}_{t, \theta_t^*} & \preceq \left[\prod_{r=1}^{t-1} (1 + \Delta_r) \right] \mathbf{\Omega}_0 + \sum_{s=1}^{t-1} \left(\left[\prod_{r=1}^s (1 + \Delta_{t-r}) \right] \mathbf{F}_{t-s, \hat{\theta}_{t-s}} \right) + \mathbf{F}_{t, \theta_t^*}.
 \end{aligned}$$

Also, note that

$$\|\mathbf{\Omega}_0\|_2 \stackrel{(P)}{\leq} K_{\max} p_* \stackrel{(S)}{\leq} \frac{K_{\max}}{3} n.$$

It follows that

$$\begin{aligned}
 & \lambda_{\max}(\tilde{\mathbf{F}}_{t_0, \theta_{t_0}^*}) \vee \lambda_{\max}(\tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}) \\
 \stackrel{(161)}{\leq} & \left[\prod_{r=1}^{t_0-1} (1 + \Delta_r) \right] \frac{K_{\max}}{3} n + \sum_{s=1}^{t_0-1} \left(\left[\prod_{r=1}^s (1 + \Delta_{t_0-r}) \right] K_{\max} n \right) + K_{\max} n \\
 \leq & \exp \left(\sum_{r=1}^{t_0-1} \Delta_r \right) \frac{K_{\max}}{3} n + \sum_{s=1}^{t_0-1} \left(\exp \left[\sum_{r=1}^s \Delta_{t_0-r} \right] K_{\max} n \right) + K_{\max} n \\
 \leq & \left(1 + 2 \sum_{r=1}^{t_0-1} \Delta_r \right) \frac{K_{\max}}{3} n + \sum_{s=1}^{t_0-1} \left(\left[1 + 2 \sum_{r=1}^s \Delta_{t_0-r} \right] K_{\max} n \right) + K_{\max} n \\
 \leq & \left(1 + 2 \sum_{r=1}^{t_0-1} \Delta_r \right) \frac{K_{\max}}{3} n + \sum_{s=1}^{t_0-1} \left(\left[1 + 2 \sum_{r=1}^{t_0-1} \Delta_r \right] K_{\max} n \right) + K_{\max} n \\
 \leq & \frac{3}{2} \cdot \frac{K_{\max}}{3} n + \frac{3}{2} K_{\max} n (t_0 - 1) + K_{\max} n = \frac{3}{2} K_{\max} n t_0 \\
 \leq & 2 K_{\max} n t_0,
 \end{aligned}$$

where the second and the third inequalities hold because $1 + x \leq e^x \leq 1 + 2x$ for $x \in [0, 1/4]$ and $\sum_{r=1}^{t_0-1} \Delta_r \leq 1/4$. This completes the proof of (160). \blacksquare

Appendix E. Proofs for Section 7

Lemma 24 *Suppose that (A0), (A1), (A2), (S) and (P) hold. Then, for all $t \in [T]$,*

$$\left\| \tilde{\mathbf{F}}_{1:t, \theta_{1:t}^*}^{1/2} (\theta_0 - \theta_{1:t}^*) \right\|_2 \leq 4 \left\| \tilde{\mathbf{F}}_{1:t, \theta_{1:t}^*}^{-1/2} \mathbf{\Omega}_0 (\theta_0 - \mu_0) \right\|_2, \quad (164)$$

and

$$\lambda_{\min}(\tilde{\mathbf{F}}_{1:t, \theta_{1:t}^*}) \geq K_{\min} n t, \quad \lambda_{\max}(\tilde{\mathbf{F}}_{1:t, \theta_{1:t}^*}) \leq \lambda_{\max}(\mathbf{\Omega}_0) + K_{\max} n t. \quad (165)$$

Proof Let $t \in [T]$. By the definition of $\theta_{1:t}^*$, we have

$$\begin{aligned}
 \mathbb{E} L_{1:t}(\theta_0) - \frac{1}{2} \left\| \mathbf{\Omega}_0^{1/2} (\theta_0 - \mu_0) \right\|_2^2 &= \mathbb{E} \tilde{L}_{1:t}(\theta_0) \leq \mathbb{E} \tilde{L}_{1:t}(\theta_{1:t}^*) \\
 &= \mathbb{E} L_{1:t}(\theta_{1:t}^*) - \frac{1}{2} \left\| \mathbf{\Omega}_0^{1/2} (\theta_{1:t}^* - \mu_0) \right\|_2^2 \leq \mathbb{E} L_{1:t}(\theta_{1:t}^*).
 \end{aligned}$$

It follows that

$$\mathbb{E} L_{1:t}(\theta_{1:t}^*) - \mathbb{E} L_{1:t}(\theta_0) \geq -\frac{1}{2} \left\| \mathbf{\Omega}_0^{1/2} (\theta_0 - \mu_0) \right\|_2^2. \quad (166)$$

In this proof, we denote $\Theta_{n,t} = \Theta(\theta_0, \mathbf{F}_{1:t, \theta_0}, \sqrt{2} \|\mathbf{\Omega}_0^{1/2} (\theta_0 - \mu_0)\|_2)$. For $\theta \in \Theta_{n,t}$, we have

$$\begin{aligned}
 \|\theta - \theta_0\|_2 &\leq \lambda_{\min}^{-1/2}(\mathbf{F}_{1:t, \theta_0}) \sqrt{2} \left\| \mathbf{\Omega}_0^{1/2} (\theta_0 - \mu_0) \right\|_2 \\
 &\stackrel{(A2)}{\leq} (K_{\min} n t)^{-1/2} \sqrt{2} \left\| \mathbf{\Omega}_0^{1/2} (\theta_0 - \mu_0) \right\|_2 \stackrel{(P)}{\leq} (K_{\min}^{-1/2} \sqrt{2} \delta) t^{-1/2} \leq \frac{1}{4},
 \end{aligned}$$

where the last inequality holds by a small enough δ . It follows that

$$\Theta_{n,t} \subset \{\theta \in \Theta : \|\theta - \theta_0\|_2 \leq 1/2\}. \quad (167)$$

By Lemma 42 and **(A2)**, $\mathbb{E}L_{1:t}(\theta)$ satisfies the third order smoothness at θ_0 with parameter

$$\left(K_{\max}K_{\min}^{-3/2}(nt)^{-1/2}, \mathbf{F}_{1:t,\theta_0}, \sqrt{2} \left\| \boldsymbol{\Omega}_0^{1/2}(\theta_0 - \mu_0) \right\|_2\right).$$

By contradiction, we will prove that

$$\left\| \mathbf{F}_{1:t,\theta_0}^{1/2}(\theta_{1:t}^* - \theta_0) \right\|_2 \leq \sqrt{2} \left\| \boldsymbol{\Omega}_0^{1/2}(\theta_0 - \mu_0) \right\|_2. \quad (168)$$

Suppose $\theta_{1:t}^* \notin \Theta_{n,t}$. Let

$$\partial\Theta_{n,t} = \left\{ \theta \in \Theta : \left\| \mathbf{F}_{1:t,\theta}^{1/2}(\theta - \theta_0) \right\|_2 = \sqrt{2} \left\| \boldsymbol{\Omega}_0^{1/2}(\theta_0 - \mu_0) \right\|_2 \right\}.$$

For $\theta^\circ \in \partial\Theta_{n,t}$, we have

$$\begin{aligned} \mathbb{E}L_{1:t}(\theta^\circ) - \mathbb{E}L_{1:t}(\theta_0) &\leq \nabla \mathbb{E}L_{1:t}(\theta_0) - \frac{1}{2} \inf_{\theta \in \Theta_{n,t}} \left\| \mathbf{F}_{1:t,\theta}^{1/2}(\theta^\circ - \theta_0) \right\|_2^2 \\ &= -\frac{1}{2} \inf_{\theta \in \Theta_{n,t}} \left\| \mathbf{F}_{1:t,\theta}^{1/2}(\theta^\circ - \theta_0) \right\|_2^2 \\ &\stackrel{\text{Lemma 40}}{\leq} -\frac{1}{2} \left(1 - K_{\max}K_{\min}^{-3/2}(nt)^{-1/2}\sqrt{2} \left\| \boldsymbol{\Omega}_0^{1/2}(\theta_0 - \mu_0) \right\|_2 \right) \left\| \mathbf{F}_{1:t,\theta_0}^{1/2}(\theta^\circ - \theta_0) \right\|_2^2 \\ &= -\frac{1}{2} \left(1 - (\sqrt{2}K_{\max}K_{\min}^{-3/2})(nt)^{-1/2} \left\| \boldsymbol{\Omega}_0^{1/2}(\theta_0 - \mu_0) \right\|_2 \right) \cdot 2 \left\| \boldsymbol{\Omega}_0^{1/2}(\theta_0 - \mu_0) \right\|_2^2 \\ &\stackrel{\text{(P)}}{\leq} - \left(1 - (\sqrt{2}K_{\max}K_{\min}^{-3/2}\delta)t^{-1/2} \right) \left\| \boldsymbol{\Omega}_0^{1/2}(\theta_0 - \mu_0) \right\|_2^2 \\ &< -\frac{1}{2} \left\| \boldsymbol{\Omega}_0^{1/2}(\theta_0 - \mu_0) \right\|_2^2. \end{aligned}$$

where the last inequality holds by a small enough δ . Consequently,

$$\mathbb{E}L_{1:t}(\theta_{1:t}^*) - \mathbb{E}L_{1:t}(\theta_0) < -\frac{1}{2} \left\| \boldsymbol{\Omega}_0^{1/2}(\theta_0 - \mu_0) \right\|_2^2$$

by the concavity of the map $\theta \mapsto \mathbb{E}L_{1:t}(\theta)$, which contradicts to (166). This completes the proof of (168).

By (167) and (168), we have

$$\begin{aligned} \lambda_{\min}(\tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}) &\geq \lambda_{\min}(\mathbf{F}_{1:t,\theta_{1:t}^*}) \geq K_{\min}nt, \\ \lambda_{\max}(\tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}) &\leq \lambda_{\max}(\boldsymbol{\Omega}_0) + \lambda_{\max}(\mathbf{F}_{1:t,\theta_{1:t}^*}) \leq \lambda_{\max}(\boldsymbol{\Omega}_0) + K_{\max}nt, \end{aligned}$$

which completes the proof of (165).

Next, we prove (164). Let

$$\rho_{n,t} = \left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{-1/2} \boldsymbol{\Omega}_0(\theta_0 - \mu_0) \right\|_2, \quad \Theta_{t,\text{bias}} = \Theta(\theta_{1:t}^*, \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}, 4\rho_{n,t}).$$

Note that

$$\begin{aligned}
 \rho_{n,t} &\leq \lambda_{\min}^{-1/2}(\tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}) \|\boldsymbol{\Omega}_0\|_2^{1/2} \left\| \boldsymbol{\Omega}_0^{1/2}(\theta_0 - \mu_0) \right\|_2 \\
 &\stackrel{\text{(P)}}{\leq} (K_{\min} nt)^{-1/2} (K_{\max} p_*)^{1/2} (\delta n^{1/2}) = (K_{\min}^{-1/2} K_{\max}^{1/2} \delta) t^{-1/2} p_*^{1/2} \\
 &\leq t^{-1/2} p_*^{1/2},
 \end{aligned} \tag{169}$$

where the last inequality holds by a small enough δ . For $\theta \in \Theta_{t,\text{bias}}$, note that

$$\|\theta - \theta_{1:t}^*\|_2 \leq (K_{\min} nt)^{-1/2} 4\rho_{n,t} \leq (4K_{\min}^{-1/2}) n^{-1/2} t^{-1} p_*^{1/2} \stackrel{\text{(S)}}{\leq} \frac{1}{4}.$$

Also, by $\theta_{1:t}^* \in \Theta_{n,t}$, we have $\|\theta_{1:t}^* - \theta_0\| \leq 1/4$. It follows that $\Theta_{t,\text{bias}} \subseteq \Theta(\theta_0, \mathbf{I}_p, 1/2)$, which, combining **(A2)** and Lemma 42, implies that $\mathbb{E}\tilde{L}_{1:t}(\cdot)$ satisfies the third order smoothness at $\theta_{1:t}^*$ with

$$\left(K_{\max} K_{\min}^{-3/2} (nt)^{-1/2}, \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}, 4\rho_{n,t} \right).$$

Let $\tau_{n,t} = K_{\max} K_{\min}^{-3/2} (nt)^{-1/2}$. Then, by **(S)**, one can easily check that $\tau_{n,t} \rho_{n,t} \leq 1/16$. By Lemma 41 with

$$\tau_3 = \tau_{n,t}, \quad f(\theta) = \mathbb{E}\tilde{L}_{1:t}(\theta), \quad \theta = \theta_{1:t}^*, \quad \tilde{\theta} = \theta_0, \quad \beta = \boldsymbol{\Omega}_0(\theta_0 - \mu_0), \quad r = \rho_{n,t},$$

$\tau_{n,t} \rho_{n,t} \leq 1/16$ implies that

$$\left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{1/2} (\theta_0 - \theta_{1:t}^*) \right\|_2 \leq 4\rho_{n,t},$$

which completes the proof of (164). \blacksquare

Lemma 25 *Suppose that **(A0)**, **(A1*)**, **(A2)**, **(S)** and **(P)** hold. Then, on $\mathcal{E}_{\text{est},2}$,*

$$\left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{1/2} (\hat{\theta}_{1:t} - \theta_{1:t}^*) \right\|_2 \leq 4M_n p_*^{1/2}, \quad \forall t \in [T].$$

Proof In this proof, we work on the event $\mathcal{E}_{\text{est},2}$ without explicitly mentioning it. The proof of this lemma is similar to that of Theorem 5. Hence, we provide a sketch of the proof.

Let $t \in [T]$. By Lemma 24, we have

$$\begin{aligned}
 \left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{1/2} (\theta_0 - \theta_{1:t}^*) \right\|_2 &\leq 4 \left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{-1/2} \boldsymbol{\Omega}_0(\theta_0 - \mu_0) \right\|_2 \stackrel{(169)}{\leq} 4t^{-1/2} p_*^{1/2}, \\
 \lambda_{\min}(\tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}) &\geq K_{\min}(nt),
 \end{aligned}$$

which implies that

$$\begin{aligned}
 \|\theta_0 - \theta_{1:t}^*\|_2 &\leq (K_{\min} nt)^{-1/2} 4t^{-1/2} p_*^{1/2} \stackrel{\text{(S)}}{\leq} 1/4, \\
 \tilde{\lambda}_{1:t} &= \left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{-1} \mathbf{V}_{1:t} \right\|_2 \leq \left\| \mathbf{F}_{1:t,\theta_{1:t}^*}^{-1} \mathbf{V}_{1:t} \right\|_2 \stackrel{\text{(A1*)}}{\leq} \frac{M_n^2}{9}, \\
 \tilde{r}_{\text{eff},1:t} &\leq \tilde{\lambda}_{1:t}^{1/2} \left[p^{1/2} + \sqrt{2(\log n + \log T)} \right] \leq 3\tilde{\lambda}_{1:t}^{1/2} p_*^{1/2} \leq M_n p_*^{1/2}.
 \end{aligned}$$

For all $\theta \in \Theta(\theta_{1:t}^*, \tilde{\mathbf{F}}_{1:t, \theta_{1:t}^*}, 4\tilde{r}_{\text{eff}, 1:t})$, it follows from the last display that

$$\|\theta - \theta_{1:t}^*\|_2 \leq (K_{\min} nt)^{-1/2} 4M_n p_*^{1/2} \stackrel{\text{(S)}}{\leq} 1/4.$$

Hence, we have

$$\Theta\left(\theta_{1:t}^*, \tilde{\mathbf{F}}_{1:t, \theta_{1:t}^*}, 4\tilde{r}_{\text{eff}, 1:t}\right) \subseteq \Theta(\theta_0, \mathbf{I}_p, 1/2),$$

which, combining with Lemma 42, implies that $\mathbb{E}\tilde{L}_{1:t}(\theta)$ satisfies the third order smoothness at $\theta_{1:t}^*$ with parameters

$$\left(K_{\max} K_{\min}^{-3/2} (nt)^{-1/2}, \tilde{\mathbf{F}}_{1:t, \theta_{1:t}^*}, 4\tilde{r}_{\text{eff}, 1:t}\right).$$

Also,

$$\left(K_{\max} K_{\min}^{-3/2} (nt)^{-1/2}\right) \tilde{r}_{\text{eff}, 1:t} \leq \left(K_{\max} K_{\min}^{-3/2} (nt)^{-1/2}\right) (M_n p_*^{1/2}) \stackrel{\text{(S)}}{\leq} 1/16,$$

which allows us to utilize Lemma 41. By Lemma 41 with

$$\tau_3 = K_{\max} K_{\min}^{-3/2} (nt)^{-1/2}, \quad f(\theta) = \mathbb{E}\tilde{L}_{1:t}(\theta), \quad \theta = \theta_{1:t}^*, \quad \tilde{\theta} = \hat{\theta}_{1:t}, \quad \beta = \nabla \zeta_{1:t}, \quad r = \tilde{r}_{\text{eff}, 1:t},$$

we have on $\mathcal{E}_{\text{est}, 2}$

$$\left\| \tilde{\mathbf{F}}_{1:t, \theta_{1:t}^*}^{1/2} \left(\hat{\theta}_{1:t} - \theta_{1:t}^* \right) \right\|_2 \leq 4\tilde{r}_{\text{eff}, 1:t} \leq 4M_n p_*^{1/2}, \quad \forall t \in [T],$$

which completes the proof. ■

Proof [Proof of Theorem 7] In this proof, we work on the event $\mathcal{E}_{\text{est}, 2}$ without explicitly mentioning it. The proof of this theorem is similar to the proofs of Theorem 2 and Proposition 23. Hence, we provide a sketch of the proof.

Let $t \in [T]$. Note that

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{1:t, \theta_{1:t}^*}^{1/2} (\theta_0 - \theta_{1:t}^*) \right\|_2 &\leq 4 \left\| \tilde{\mathbf{F}}_{1:t, \theta_{1:t}^*}^{-1/2} \boldsymbol{\Omega}_0 (\theta_0 - \mu_0) \right\|_2 \stackrel{(169)}{\leq} 4t^{-1/2} p_*^{1/2}, \\ \left\| \tilde{\mathbf{F}}_{1:t, \theta_{1:t}^*}^{1/2} (\hat{\theta}_{1:t} - \theta_{1:t}^*) \right\|_2 &\leq 4M_n p_*^{1/2}, \end{aligned} \tag{170}$$

which, combining with (S), implies that

$$\|\theta_0 - \theta_{1:t}^*\|_2 \leq 1/8, \quad \|\hat{\theta}_{1:t} - \theta_{1:t}^*\|_2 \leq 1/8. \tag{171}$$

From the last display, we have $\|\theta_0 - \hat{\theta}_{1:t}\|_2 \leq 1/4$, which implies that

$$\lambda_{\min}(\tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}) \geq \lambda_{\min}(\boldsymbol{\Omega}_0) + \lambda_{\min}(\mathbf{F}_{1:t, \hat{\theta}_{1:t}}) \geq \lambda_{\min}(\mathbf{F}_{1:t, \hat{\theta}_{1:t}}) \stackrel{\text{(A2)}}{\geq} K_{\min} nt. \tag{172}$$

Note that $r_{\text{LA}} = 2\sqrt{p} + \sqrt{2\log N} \leq 4\sqrt{p_*}$. Also, we have

$$\sup_{\theta \in \Theta(\widehat{\theta}_{1:t}, \widetilde{\mathbf{F}}_{1:t, \widehat{\theta}_{1:t}}, 4r_{\text{LA}})} \|\theta - \widehat{\theta}_{1:t}\|_2 \leq \lambda_{\min}^{-1/2}(\widetilde{\mathbf{F}}_{1:t, \widehat{\theta}_{1:t}}) 4r_{\text{LA}} \leq (K_{\min} nt)^{-1/2} 16\sqrt{p_*} \stackrel{\text{(S)}}{\leq} 1/4,$$

which implies that $\Theta(\widehat{\theta}_{1:t}, \widetilde{\mathbf{F}}_{1:t, \widehat{\theta}_{1:t}}, 4r_{\text{LA}}) \subseteq \Theta(\theta_0, \mathbf{I}_p, 1/2)$. By Lemma 42 and **(A2)**, $\widetilde{L}_{1:t}(\cdot)$ satisfies the third and fourth order smoothness at $\widehat{\theta}_{1:t}$ with parameter

$$\left(K_{\max} K_{\min}^{-3/2} (nt)^{-1/2}, \widetilde{\mathbf{F}}_{1:t, \widehat{\theta}_{1:t}}, 4r_{\text{LA}} \right) \quad \text{and} \quad \left(K_{\max} K_{\min}^{-2} (nt)^{-1}, \widetilde{\mathbf{F}}_{1:t, \widehat{\theta}_{1:t}}, 4r_{\text{LA}} \right),$$

respectively. In this proof, let

$$\tau_{3,t} = K_{\max} K_{\min}^{-3/2} (nt)^{-1/2}, \quad \tau_{4,t} = K_{\max} K_{\min}^{-2} (nt)^{-1}.$$

Then, we can apply the proof strategy in Theorem 2, which implies that

$$\begin{aligned} & d_V \left(\mathcal{N} \left(\widehat{\theta}_{1:t}, \widetilde{\mathbf{F}}_{1:t, \widehat{\theta}_{1:t}} \right), \Pi(\cdot \mid \mathbf{D}_{1:t}) \right) \\ & \leq c_1 \left([\tau_{4,t} + \tau_{3,t}^2] p^2 + \tau_{3,t} p + \tau_{3,t}^3 \log^3 N + e^{-8 \log N - 8p} \right) \\ & \leq c_2 \left((nt)^{-1} p^2 + (nt)^{-1/2} p + (nt)^{-3/2} \log^3 N + e^{-8 \log N - 8p} \right) \\ & \stackrel{\text{(S)}}{\leq} c_3 \sqrt{\frac{p_*^2}{nt}} \end{aligned}$$

for some constants $c_1, c_2, c_3 > 0$, depending only on (K_{\min}, K_{\max}) . ■

Proof [Proof of Theorem 8] In this proof, we work on the event $\mathcal{E}_{\text{est},2}$ without explicitly mentioning it. Let $t \in [T]$. To complete this proof, we utilize Lemma 43. Hence, we need to obtain upper bounds of the following quantities:

$$(i) = \left\| \widetilde{\mathbf{F}}_{1:t, \widehat{\theta}_{1:t}}^{1/2} \left(\widehat{\theta}_{1:t}^{\text{ML}} - \widehat{\theta}_{1:t} \right) \right\|_2, \quad (ii) = \left\| \mathbf{F}_{1:t, \theta_0}^{-1/2} \widetilde{\mathbf{F}}_{1:t, \widehat{\theta}_{1:t}} \mathbf{F}_{1:t, \theta_0}^{-1/2} - \mathbf{I}_p \right\|_{\text{F}}.$$

Step 1: (i)

Firstly, we will obtain an upper bound of (i). To complete the proof, we will show that

$$\left\| \mathbf{F}_{1:t, \theta_0}^{1/2} \left(\widehat{\theta}_{1:t}^{\text{ML}} - \theta_0 \right) \right\|_2 \leq 4M_n p_*^{1/2}$$

by utilizing Lemma 41. Note that required proof for the last display is similar to that of Theorem 5. Hence, in **Step 1**, we provide a sketch of the proof. Note that

$$\begin{aligned} \left\| \mathbf{F}_{1:t, \theta_0}^{-1} \mathbf{V}_{1:t} \right\|_2 & \stackrel{\text{(A1*)}}{\leq} \frac{M_n^2}{9}, \\ \lambda_{\min}(\mathbf{F}_{1:t, \theta_0}) & \stackrel{\text{(A2)}}{\geq} K_{\min}(nt) \\ r_{\text{eff},1:t} & \leq \lambda_{1:t}^{1/2} \left[p^{1/2} + \sqrt{2(\log n + \log T)} \right] \leq 3(\lambda_{1:t} p_*)^{1/2} \leq M_n p_*^{1/2}, \end{aligned}$$

which, combining with **(S)**, implies that

$$\Theta(\theta_0, \mathbf{F}_{1:t, \theta_0}, 4r_{\text{eff}, 1:t}) \subseteq \Theta(\theta_0, \mathbf{I}_p, 1/2).$$

By Lemma 42, the last display implies that $\mathbb{E}L_{1:t}(\theta)$ satisfies the third order smoothness at θ_0 with parameters

$$\left(K_{\max} K_{\min}^{-3/2} (nt)^{-1/2}, \mathbf{F}_{1:t, \theta_0}, 4r_{\text{eff}, 1:t} \right).$$

Also,

$$\begin{aligned} (K_{\max} K_{\min}^{-3/2} (nt)^{-1/2}) r_{\text{eff}, 1:t} &\leq (K_{\max} K_{\min}^{-3/2}) (M_n p_*^{1/2} (nt)^{-1/2}) \stackrel{\text{(S)}}{\leq} 1/16, \\ \left\| \mathbf{F}_{1:t, \theta_0}^{-1/2} \nabla \zeta_{1:t} \right\|_2 &\leq r_{\text{eff}, 1:t}, \end{aligned} \quad (173)$$

where the second inequality holds on $\mathcal{E}_{\text{eff}, 2}$ by the condition (41). Note that (173) allows us to utilize Lemma 41. By Lemma 41 with

$$\tau_3 = K_{\max} K_{\min}^{-3/2} (nt)^{-1/2}, \quad f(\theta) = \mathbb{E}L_{1:t}(\theta), \quad \theta = \theta_0, \quad \tilde{\theta} = \hat{\theta}_{1:t}^{\text{ML}}, \quad \beta = \nabla \zeta_{1:t}, \quad r = r_{\text{eff}, 1:t},$$

we have

$$\left\| \mathbf{F}_{1:t, \theta_0}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \theta_0) \right\|_2 \leq 4r_{\text{eff}, 1:t} \leq 4M_n p_*^{1/2}, \quad (174)$$

which implies $\|\hat{\theta}_{1:t}^{\text{ML}} - \theta_0\|_2 \leq 1/2$ by **(S)**.

By (165), (172) and **(A2)**, we have

$$\lambda_{\min}(\tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}) \wedge \lambda_{\min}(\tilde{\mathbf{F}}_{1:t, \theta_{1:t}^*}) \wedge \lambda_{\min}(\mathbf{F}_{1:t, \theta_0}) \wedge \lambda_{\min}(\mathbf{F}_{1:t, \hat{\theta}_{1:t}^{\text{ML}}}) \geq K_{\min}(nt). \quad (175)$$

It follows that

$$\begin{aligned} &\left\| \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1/2} \mathbf{\Omega}_0 (\hat{\theta}_{1:t}^{\text{ML}} - \mu_0) \right\|_2 \\ &\leq \left\| \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1/2} \mathbf{\Omega}_0 (\hat{\theta}_{1:t}^{\text{ML}} - \theta_0) \right\|_2 + \left\| \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1/2} \mathbf{\Omega}_0 (\theta_0 - \mu_0) \right\|_2 \\ &= \left\| \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1/2} \mathbf{\Omega}_0 \mathbf{F}_{1:t, \theta_0}^{-1/2} \mathbf{F}_{1:t, \theta_0}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \theta_0) \right\|_2 + \left\| \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1/2} \mathbf{\Omega}_0 (\theta_0 - \mu_0) \right\|_2 \\ &\leq (K_{\min} nt)^{-1} \|\mathbf{\Omega}_0\|_2 \left\| \mathbf{F}_{1:t, \theta_0}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \theta_0) \right\|_2 \\ &\quad + (K_{\min} nt)^{-1/2} \|\mathbf{\Omega}_0\|_2^{1/2} \left\| \mathbf{\Omega}_0^{1/2} (\theta_0 - \mu_0) \right\|_2 \\ &\leq (K_{\min} nt)^{-1} (K_{\max} p_*) \left(4M_n p_*^{1/2} \right) \\ &\quad + (K_{\min} nt)^{-1/2} (K_{\max} p_*)^{1/2} (K_{\max} M_n p_*^{1/2}) \\ &\stackrel{\text{(S)}}{\leq} \left(1 + K_{\min}^{-1/2} K_{\max}^{3/2} \right) M_n \left(\frac{p_*^2}{nt} \right)^{1/2} = c_1 M_n \left(\frac{p_*^2}{nt} \right)^{1/2}, \end{aligned} \quad (176)$$

where $c_1 = 1 + K_{\min}^{-1/2} K_{\max}^{3/2}$, and the third inequality holds by (174), (175) and (\mathbf{P}^*) . In this proof, let

$$b_t = \left\| \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1/2} \mathbf{\Omega}_0(\hat{\theta}_{1:t}^{\text{ML}} - \mu_0) \right\|_2.$$

Then, for $\theta \in \Theta(\hat{\theta}_{1:t}, \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}, 4b_t)$, note that

$$\begin{aligned} \|\theta - \theta_0\|_2 &\leq \|\theta - \hat{\theta}_{1:t}\|_2 + \|\hat{\theta}_{1:t} - \theta_0\|_2 \leq \lambda_{\min}^{-1/2}(\tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}) 4b_t + \|\hat{\theta}_{1:t} - \theta_0\|_2 \\ &\leq (K_{\min} nt)^{-1/2} \left[4c_1 M_n \left(\frac{p_*^2}{nt} \right)^{1/2} \right] + 1/4 \stackrel{(\mathbf{S})}{\leq} 1/2, \end{aligned}$$

where the third inequality holds by (171), (175) and (176). It follows that $\Theta(\hat{\theta}_{1:t}, \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}, 4b_t) \subseteq \Theta(\theta_0, \mathbf{I}_p, 1/2)$. By Lemma 42, $\tilde{L}_{1:t}(\theta)$ satisfies the third order smoothness at $\hat{\theta}_{1:t}$ with parameters

$$\left(K_{\max} K_{\min}^{-3/2} (nt)^{-1/2}, \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}, 4b_t \right).$$

Also,

$$(K_{\max} K_{\min}^{-3/2} (nt)^{-1/2}) b_t \leq K_{\max} K_{\min}^{-3/2} (nt)^{-1/2} \left[c_1 M_n \left(\frac{p_*^2}{nt} \right)^{1/2} \right] \stackrel{(\mathbf{S})}{\leq} 1/16,$$

which allows us to utilize Lemma 41. By Lemma 41 with

$$\begin{aligned} \tau_3 &= K_{\max} K_{\min}^{-3/2} (nt)^{-1/2}, \quad f(\theta) = \tilde{L}_{1:t}(\theta), \quad \theta = \hat{\theta}_{1:t}, \quad \tilde{\theta} = \hat{\theta}_{1:t}^{\text{ML}}, \\ \beta &= \mathbf{\Omega}_0(\hat{\theta}_{1:t}^{\text{ML}} - \mu_0), \quad r = b_t, \end{aligned}$$

we have

$$\left\| \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{1/2} \left(\hat{\theta}_{1:t}^{\text{ML}} - \hat{\theta}_{1:t} \right) \right\|_2 \leq 4b_t \leq 4c_1 M_n \left(\frac{p_*^2}{nt} \right)^{1/2}. \quad (177)$$

Step 2: (ii)

Next, we will obtain an upper bound of (ii). By Lemmas 24 and 25, we have

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{1:t, \theta_{1:t}^*}^{1/2} (\theta_0 - \theta_{1:t}^*) \right\|_2 &\leq 4 \left\| \tilde{\mathbf{F}}_{1:t, \theta_{1:t}^*}^{-1/2} \mathbf{\Omega}_0(\theta_0 - \mu_0) \right\|_2 \stackrel{(169)}{\leq} 4t^{-1/2} p_*^{1/2} \\ &\leq 4M_n p_*^{1/2}, \\ \left\| \tilde{\mathbf{F}}_{1:t, \theta_{1:t}^*}^{1/2} (\hat{\theta}_{1:t} - \theta_{1:t}^*) \right\|_2 &\leq 4M_n p_*^{1/2}. \end{aligned} \quad (178)$$

Combining (\mathbf{S}) , (175) and (178), one can easily check that

$$\begin{aligned} \Theta(\theta_{1:t}^*, \tilde{\mathbf{F}}_{1:t, \theta_{1:t}^*}, 4M_n p_*^{1/2}) &\subseteq \Theta(\theta_0, \mathbf{I}_p, 1/2), \\ \Theta(\theta_{1:t}^*, \mathbf{F}_{1:t, \theta_{1:t}^*}, 10M_n p_*^{1/2}) &\subseteq \Theta(\theta_0, \mathbf{I}_p, 1/2). \end{aligned}$$

Hence, by Lemma 42, $\tilde{L}_{1:t}(\cdot)$ satisfies the third order smoothness at $\theta_{1:t}^*$ with parameters

$$\left(K_{\max}K_{\min}^{-3/2}(nt)^{-1/2}, \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}, 4M_n p_*^{1/2}\right).$$

Since $\hat{\theta}_{1:t} \in \Theta(\theta_{1:t}^*, \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}, 4M_n p_*^{1/2})$ by (178), we have

$$\begin{aligned} & \left\| \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{1/2}(\theta_0 - \hat{\theta}_{1:t}) \right\|_2 \leq \left\| \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{1/2} \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{-1/2} \right\|_2 \left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{1/2}(\theta_0 - \hat{\theta}_{1:t}) \right\|_2 \\ & \stackrel{\text{Lemma 40}}{\leq} \left(1 + K_{\max}K_{\min}^{-3/2}(nt)^{-1/2}(4M_n p_*^{1/2})\right)^{1/2} \left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{1/2}(\theta_0 - \hat{\theta}_{1:t}) \right\|_2 \\ & \stackrel{(178)}{\leq} \left(1 + K_{\max}K_{\min}^{-3/2}(nt)^{-1/2}(4M_n p_*^{1/2})\right)^{1/2} 8M_n p_*^{1/2} \stackrel{\text{(S)}}{\leq} 10M_n p_*^{1/2}, \end{aligned}$$

which implies that

$$\theta_0 \in \Theta(\hat{\theta}_{1:t}, \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}, 10M_n p_*^{1/2}) \subset \Theta(\hat{\theta}_{1:t}, \mathbf{F}_{1:t,\hat{\theta}_{1:t}}, 10M_n p_*^{1/2}) \subseteq \Theta(\theta_0, \mathbf{I}_p, 1/2).$$

Consequently, by Lemma 42, $L_{1:t}(\cdot)$ satisfies the third order smoothness at $\hat{\theta}_{1:t}$ with parameters

$$\left(K_{\max}K_{\min}^{-3/2}(nt)^{-1/2}, \mathbf{F}_{1:t,\hat{\theta}_{1:t}}, 10M_n p_*^{1/2}\right).$$

Hence, combining with the last two displays, Lemma 40 and **(A2)** give that

$$\begin{aligned} \left\| \mathbf{F}_{1:t,\hat{\theta}_{1:t}}^{-1/2} \mathbf{F}_{1:t,\theta_0} \mathbf{F}_{1:t,\hat{\theta}_{1:t}}^{-1/2} - \mathbf{I}_p \right\|_2 & \leq K_{\max}K_{\min}^{-3/2}(nt)^{-1/2}(10M_n p_*^{1/2}), \\ \left\| \mathbf{F}_{1:t,\hat{\theta}_{1:t}} \right\|_2 & \leq K_{\max}(nt). \end{aligned} \tag{179}$$

Note that

$$\begin{aligned} & \left\| \mathbf{F}_{1:t,\theta_0} - \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \right\|_2 \leq \left\| \mathbf{\Omega}_0 \right\|_2 + \left\| \mathbf{F}_{1:t,\theta_0} - \mathbf{F}_{1:t,\hat{\theta}_{1:t}} \right\|_2 \\ & \leq \left\| \mathbf{\Omega}_0 \right\|_2 + \left\| \mathbf{F}_{1:t,\hat{\theta}_{1:t}} \right\|_2 \left\| \mathbf{F}_{1:t,\hat{\theta}_{1:t}}^{-1/2} \mathbf{F}_{1:t,\theta_0} \mathbf{F}_{1:t,\hat{\theta}_{1:t}}^{-1/2} - \mathbf{I}_p \right\|_2 \\ & \stackrel{(179),(\mathbf{P}^*)}{\leq} K_{\max} p_* + (K_{\max} nt) \left(K_{\max} K_{\min}^{-3/2} (nt)^{-1/2} (10M_n p_*^{1/2}) \right) \\ & \stackrel{\text{(S)}}{\leq} (1 + 10K_{\max}^2 K_{\min}^{-3/2}) M_n (nt)^{1/2} p_*^{1/2}. \end{aligned}$$

It follows that

$$\begin{aligned} \left\| \mathbf{F}_{1:t,\theta_0}^{-1/2} \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \mathbf{F}_{1:t,\theta_0}^{-1/2} - \mathbf{I}_p \right\|_2 & \leq \lambda_{\min}^{-1}(\mathbf{F}_{1:t,\theta_0}) \left\| \mathbf{F}_{1:t,\theta_0} - \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \right\|_2 \\ & \leq (K_{\min} nt)^{-1} \left[1 + 10K_{\max}^2 K_{\min}^{-3/2} \right] M_n (nt)^{1/2} p_*^{1/2} \\ & = (K_{\min}^{-1} + 10K_{\max}^2 K_{\min}^{-5/2}) M_n \left(\frac{p_*}{nt} \right)^{1/2}, \end{aligned}$$

which further implies that

$$\begin{aligned}
 \left\| \mathbf{F}_{1:t,\theta_0}^{-1/2} \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \mathbf{F}_{1:t,\theta_0}^{-1/2} - \mathbf{I}_p \right\|_{\mathbf{F}} &\leq \sqrt{p} \left\| \mathbf{F}_{1:t,\theta_0}^{-1/2} \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \mathbf{F}_{1:t,\theta_0}^{-1/2} - \mathbf{I}_p \right\|_2 \\
 &\leq (K_{\min}^{-1} + 10K_{\max}^2 K_{\min}^{-5/2}) M_n \left(\frac{p_*^2}{nt} \right)^{1/2} \\
 &= c_2 M_n \left(\frac{p_*^2}{nt} \right)^{1/2},
 \end{aligned} \tag{180}$$

where $c_2 = K_{\min}^{-1} + 10K_{\max}^2 K_{\min}^{-5/2}$.

Step 3: Applying Lemma 43

By (S), we have

$$\left\| \mathbf{F}_{1:t,\theta_0}^{-1/2} \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \mathbf{F}_{1:t,\theta_0}^{-1/2} - \mathbf{I}_p \right\|_2 \leq 0.684.$$

By Lemma 43, we have

$$\begin{aligned}
 &d_V \left(\mathcal{N} \left(\hat{\theta}_{1:t}^{\text{ML}}, \mathbf{F}_{1:t,\theta_0}^{-1} \right), \mathcal{N} \left(\hat{\theta}_{1:t}, \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{-1} \right) \right) \\
 &\leq \frac{1}{2} \left(\left\| \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{1/2} \left(\hat{\theta}_{1:t}^{\text{ML}} - \hat{\theta}_{1:t} \right) \right\|_2^2 + \left\| \mathbf{F}_{1:t,\theta_0}^{-1/2} \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \mathbf{F}_{1:t,\theta_0}^{-1/2} - \mathbf{I}_p \right\|_{\mathbf{F}}^2 \right)^{1/2} \\
 &\stackrel{(177)}{\leq} \frac{1}{2} (16c_1^2 + c_2^2)^{1/2} M_n \left(\frac{p_*^2}{nt} \right)^{1/2} \stackrel{(180)}{=} c_3 M_n \left(\frac{p_*^2}{nt} \right)^{1/2}
 \end{aligned}$$

for some constant $c_3 = c_3(K_{\min}, K_{\max}) > 0$. This completes the proof of the first assertion in (46).

Combining with Theorem 7, we have

$$\begin{aligned}
 &d_V \left(\mathcal{N} \left(\hat{\theta}_{1:t}^{\text{ML}}, \mathbf{F}_{1:t,\theta_0}^{-1} \right), \Pi(\cdot \mid \mathbf{D}_{1:t}) \right) \\
 &\leq d_V \left(\mathcal{N} \left(\hat{\theta}_{1:t}^{\text{ML}}, \mathbf{F}_{1:t,\theta_0}^{-1} \right), \mathcal{N} \left(\hat{\theta}_{1:t}, \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{-1} \right) \right) + d_V \left(\mathcal{N} \left(\hat{\theta}_{1:t}, \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{-1} \right), \Pi(\cdot \mid \mathbf{D}_{1:t}) \right) \\
 &\leq c_3 M_n \left(\frac{p_*^2}{nt} \right)^{1/2} + c_4 \left(\frac{p_*^2}{nt} \right)^{1/2} \leq c_5 M_n \left(\frac{p_*^2}{nt} \right)^{1/2}
 \end{aligned}$$

for some constants $c_4, c_5 > 0$, depending only on (K_{\min}, K_{\max}) . This completes the proof of the second assertion in (46). \blacksquare

Appendix F. Proofs for Section 8

Proof [Proof of Proposition 9] In this proof, we work on the event $\mathcal{E}_{\text{est},1} \cap \mathcal{E}_{\text{est},2}$ without explicitly referring to it. By Lemmas 24, 25 and Proposition 23, we have

$$\begin{aligned}
 \left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{1/2} (\theta_0 - \theta_{1:t}^*) \right\|_2 &\leq 4M_n p_*^{1/2}, & \lambda_{\min}(\tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}) &\geq K_{\min} n t \\
 \left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{1/2} (\hat{\theta}_{1:t} - \theta_{1:t}^*) \right\|_2 &\leq 4M_n p_*^{1/2}, & \lambda_{\max}(\tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}) &\leq \frac{4}{3} K_{\max} n t, \\
 \lambda_{\min}(\tilde{\mathbf{F}}_{t,\hat{\theta}_t}) \wedge \lambda_{\min}(\tilde{\mathbf{F}}_{t,\theta_t^*}) &\geq K_{\text{low}} n t, & \lambda_{\max}(\tilde{\mathbf{F}}_{t,\hat{\theta}_t}) \wedge \lambda_{\max}(\tilde{\mathbf{F}}_{t,\theta_t^*}) &\leq K_{\text{up}} n t, \\
 \hat{\theta}_t, \theta_t^*, \theta_{1:t}^*, \hat{\theta}_{1:t} &\in \Theta(\theta_0, \mathbf{I}_p, 1/4), & \Delta_t \vee \epsilon_{n,t,\text{KL}} &\leq K_{\text{up}} t^{-1} n^{-1/2} p_*
 \end{aligned} \tag{181}$$

for all $t \in [T]$.

By utilizing Lemma 41 and an inductive argument, we prove the following inequalities with some constants $D_1, D_2 > 0$:

$$\left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} (\hat{\theta}_{1:t} - \hat{\theta}_t) \right\|_2 \leq D_1 M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2}, \quad \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} (\theta_0 - \hat{\theta}_t) \right\|_2 \leq D_2 M_n \sqrt{p_*} \tag{182}$$

for all $t \in [T]$. Based on (182), we subsequently prove the following inequality:

$$\left\| \mathbf{\Omega}_t^{1/2} (\hat{\theta}_{1:t} - \mu_t) \right\|_2 \leq D_3 M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2} \quad \text{for all } t \in [T]$$

for some constant $D_3 = D_3(D_1)$.

Step 1: Inductive argument

We will first show that (182) holds for $t = 1$. Note that

$$\left\| \tilde{\mathbf{F}}_{1,\hat{\theta}_1}^{1/2} (\hat{\theta}_{1:1} - \hat{\theta}_1) \right\|_2 = 0$$

because $\hat{\theta}_{1:t} = \hat{\theta}_t$ at $t = 1$. Also, we have

$$\begin{aligned}
 \left\| \tilde{\mathbf{F}}_{1,\hat{\theta}_1}^{1/2} (\theta_0 - \hat{\theta}_1) \right\|_2 &= \left\| \tilde{\mathbf{F}}_{1,\hat{\theta}_1}^{1/2} \tilde{\mathbf{F}}_{1:1,\theta_{1:1}^*}^{-1/2} \tilde{\mathbf{F}}_{1:1,\theta_{1:1}^*}^{1/2} (\theta_0 - \hat{\theta}_{1:1}) \right\|_2 \\
 &\leq \left\| \tilde{\mathbf{F}}_{1,\hat{\theta}_1}^{1/2} \tilde{\mathbf{F}}_{1:1,\theta_{1:1}^*}^{-1/2} \right\|_2 \left(\left\| \tilde{\mathbf{F}}_{1:1,\theta_{1:1}^*}^{1/2} (\theta_0 - \theta_{1:1}^*) \right\|_2 + \left\| \tilde{\mathbf{F}}_{1:1,\theta_{1:1}^*}^{1/2} (\theta_{1:1}^* - \hat{\theta}_{1:1}) \right\|_2 \right) \\
 &\stackrel{(181)}{\leq} (K_{\min} n)^{-1/2} (K_{\text{up}} n)^{1/2} (4M_n \sqrt{p_*} + 4M_n \sqrt{p_*}) \\
 &= (8K_{\min}^{-1/2} K_{\text{up}}^{1/2}) M_n \sqrt{p_*}.
 \end{aligned}$$

Hence, for $t = 1$, the inequalities in (182) hold with $D_1 = 0$ and $D_2 = 8K_{\min}^{-1/2} K_{\text{up}}^{1/2}$.

Let $t_0 \in \{2, 3, \dots, T\}$. To prove (182) by induction, suppose that

$$\left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} (\theta_0 - \hat{\theta}_t) \right\|_2 \leq (8K_{\min}^{-1/2} K_{\text{up}}^{1/2} + 1) M_n \sqrt{p_*}, \quad \forall t \in [t_0 - 1]. \tag{183}$$

Based on (183), we will show that

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{1/2} (\hat{\theta}_{1:t_0} - \hat{\theta}_{t_0}) \right\|_2 &\leq D_1 M_n^2 \left(\frac{p^*}{n} \right)^{1/2}, \\ \left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{1/2} (\theta_0 - \hat{\theta}_{t_0}) \right\|_2 &\leq (8K_{\min}^{-1/2} K_{\text{up}}^{1/2} + 1) M_n \sqrt{p^*}, \end{aligned} \quad (184)$$

where $D_1 = D_1(K_{\min}, K_{\max}, K_{\text{low}}, K_{\text{up}})$. It then follows by induction that (182) holds for all $t \in [T]$.

The proof is divided into several steps. In **Step 2**, we introduce the theoretical framework needed to prove (184). Then, in **Step 3-6**, we will prove that (184) holds for every $t_0 \in \{2, 3, \dots, T\}$.

Step 2: Framework for applying Lemma 41

To prove (184), we will utilize Lemma 41. In **Step 2**, therefore, we introduce some theoretical preliminaries needed to apply Lemma 41.

Recall the definition of $\nabla \eta_t(\theta)$ given in (51). For $\theta \in \Theta$ and $t \in [T]$, note that

$$\begin{aligned} &\nabla \eta_t(\theta) \\ &= \nabla \tilde{L}_t(\theta) + \mathbf{\Omega}_t(\theta - \mu_t) = \nabla \tilde{L}_t(\theta) + \tilde{\mathbf{F}}_{t, \hat{\theta}_t}(\theta - \hat{\theta}_t) - \tilde{\mathbf{F}}_{t, \hat{\theta}_t}(\theta - \hat{\theta}_t) + \mathbf{\Omega}_t(\theta - \mu_t) \\ &= \nabla \tilde{L}_t(\theta) - \nabla \tilde{L}_t(\hat{\theta}_t) + \tilde{\mathbf{F}}_{t, \hat{\theta}_t}(\theta - \hat{\theta}_t) - \tilde{\mathbf{F}}_{t, \hat{\theta}_t}(\theta - \hat{\theta}_t) + \mathbf{\Omega}_t(\theta - \mu_t) \\ &= \dot{\mathcal{R}}_{t,3}(\hat{\theta}_t, \theta - \hat{\theta}_t) - \tilde{\mathbf{F}}_{t, \hat{\theta}_t}(\theta - \hat{\theta}_t) + \mathbf{\Omega}_t(\theta - \mu_t) \\ &= \dot{\mathcal{R}}_{t,3}(\hat{\theta}_t, \theta - \hat{\theta}_t) + \left(\mathbf{\Omega}_t - \tilde{\mathbf{F}}_{t, \hat{\theta}_t} \right) (\theta - \hat{\theta}_t) + \mathbf{\Omega}_t(\hat{\theta}_t - \mu_t), \end{aligned} \quad (185)$$

where $\dot{\mathcal{R}}_{t,3}(\cdot, \cdot)$ is defined as

$$\dot{\mathcal{R}}_{t,3}(\theta, u) = \nabla \tilde{L}_t(\theta + u) - \nabla \tilde{L}_t(\theta) - \nabla^2 \tilde{L}_t(\theta)u, \quad \forall \theta, u \in \Theta.$$

For $\theta \in \Theta$, define a linear perturbation version of $\tilde{L}_{t_0}(\theta)$ by

$$g_{n,t_0}(\theta) = \tilde{L}_{t_0}(\theta) + \left\langle \sum_{t=1}^{t_0-1} \nabla \eta_t(\hat{\theta}_{1:t_0}), \theta \right\rangle.$$

Note that

$$\nabla g_{n,t_0}(\hat{\theta}_{1:t_0}) = \nabla \tilde{L}_{t_0}(\hat{\theta}_{1:t_0}) + \sum_{t=1}^{t_0-1} \nabla \eta_t(\hat{\theta}_{1:t_0}) \stackrel{(52)}{=} \nabla \tilde{L}_{1:t_0}(\hat{\theta}_{1:t_0}) = 0.$$

If $\tilde{L}_{t_0}(\cdot)$ satisfies the third order smoothness at $\hat{\theta}_{t_0}$ with parameter

$$\left(\tau_{n,t_0}, \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}, 4 \left\| \sum_{t=1}^{t_0-1} \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \nabla \eta_t(\hat{\theta}_{1:t_0}) \right\|_2 \right)$$

for some $\tau_{n,t_0} \geq 0$ and

$$\tau_{n,t_0} \left\| \sum_{t=1}^{t_0-1} \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \nabla \eta_t(\hat{\theta}_{1:t_0}) \right\|_2 \leq \frac{1}{16},$$

one can apply Lemma 41 and then obtain the following inequality:

$$\left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{1/2} (\hat{\theta}_{1:t_0} - \hat{\theta}_{t_0}) \right\|_2 \leq 4 \left\| \sum_{t=1}^{t_0-1} \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \nabla \eta_t(\hat{\theta}_{1:t_0}) \right\|_2.$$

Later, we will show that τ_{n, t_0} can be chosen as $\tau_{n, t_0} = K_{\max} K_{\text{low}}^{-3/2} t_0^{-3/2} n^{-1/2}$. Therefore, to apply Lemma 41, we need to obtain an upper bound of

$$\left\| \sum_{t=1}^{t_0-1} \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \nabla \eta_t(\hat{\theta}_{1:t_0}) \right\|_2.$$

By (185), we have

$$\begin{aligned} & \left\| \sum_{t=1}^{t_0-1} \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \nabla \eta_t(\hat{\theta}_{1:t_0}) \right\|_2 \\ &= \left\| \sum_{t=1}^{t_0-1} \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \left[\dot{\mathcal{R}}_{t,3}(\hat{\theta}_t, \hat{\theta}_{1:t_0} - \hat{\theta}_t) + (\boldsymbol{\Omega}_t - \tilde{\mathbf{F}}_{t, \hat{\theta}_t}) (\hat{\theta}_{1:t_0} - \hat{\theta}_t) + \boldsymbol{\Omega}_t (\hat{\theta}_t - \mu_t) \right] \right\|_2 \\ &\leq \left\| \sum_{t=1}^{t_0-1} \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \dot{\mathcal{R}}_{t,3}(\hat{\theta}_t, \hat{\theta}_{1:t_0} - \hat{\theta}_t) \right\|_2 + \left\| \sum_{t=1}^{t_0-1} \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} (\boldsymbol{\Omega}_t - \tilde{\mathbf{F}}_{t, \hat{\theta}_t}) (\hat{\theta}_{1:t_0} - \hat{\theta}_t) \right\|_2 \\ &\quad + \left\| \sum_{t=1}^{t_0-1} \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \boldsymbol{\Omega}_t (\hat{\theta}_t - \mu_t) \right\|_2 \\ &\leq \sum_{t=1}^{t_0-1} \left[\left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \dot{\mathcal{R}}_{t,3}(\hat{\theta}_t, \hat{\theta}_{1:t_0} - \hat{\theta}_t) \right\|_2 + \left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} (\boldsymbol{\Omega}_t - \tilde{\mathbf{F}}_{t, \hat{\theta}_t}) (\hat{\theta}_{1:t_0} - \hat{\theta}_t) \right\|_2 \right. \\ &\quad \left. + \left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \boldsymbol{\Omega}_t (\hat{\theta}_t - \mu_t) \right\|_2 \right]. \end{aligned}$$

For $t \in [t_0 - 1]$, let

$$\begin{aligned} \text{(i)}_t &= \left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \dot{\mathcal{R}}_{t,3}(\hat{\theta}_t, \hat{\theta}_{1:t_0} - \hat{\theta}_t) \right\|_2, \\ \text{(ii)}_t &= \left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} (\boldsymbol{\Omega}_t - \tilde{\mathbf{F}}_{t, \hat{\theta}_t}) (\hat{\theta}_{1:t_0} - \hat{\theta}_t) \right\|_2, \\ \text{(iii)}_t &= \left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \boldsymbol{\Omega}_t (\hat{\theta}_t - \mu_t) \right\|_2. \end{aligned}$$

We will obtain upper bounds of these quantities throughout **Step 3-5**.

To bound $(i)_t$, $(ii)_t$ and $(iii)_t$, $\widehat{\theta}_{1:t_0}$ should be located in a sufficiently small neighborhood of $\widehat{\theta}_t$ for all $t \in [t_0 - 1]$. Let $t \in [t_0 - 1]$. Note that

$$\begin{aligned}
 & \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{1/2} (\widehat{\theta}_{1:t_0} - \widehat{\theta}_t) \right\|_2 \\
 & \leq \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{1/2} (\widehat{\theta}_{1:t_0} - \theta_0) \right\|_2 + \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{1/2} (\theta_0 - \widehat{\theta}_t) \right\|_2 \\
 & \leq \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{1/2} \widetilde{\mathbf{F}}_{1:t_0, \theta_{1:t_0}^*}^{-1/2} \right\|_2 \left\| \widetilde{\mathbf{F}}_{1:t_0, \theta_{1:t_0}^*}^{1/2} (\widehat{\theta}_{1:t_0} - \theta_0) \right\|_2 + \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{1/2} (\theta_0 - \widehat{\theta}_t) \right\|_2 \\
 & \leq \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{1/2} \widetilde{\mathbf{F}}_{1:t_0, \theta_{1:t_0}^*}^{-1/2} \right\|_2 \left[\left\| \widetilde{\mathbf{F}}_{1:t_0, \theta_{1:t_0}^*}^{1/2} (\widehat{\theta}_{1:t_0} - \theta_{1:t_0}^*) \right\|_2 + \left\| \widetilde{\mathbf{F}}_{1:t_0, \theta_{1:t_0}^*}^{1/2} (\theta_{1:t_0}^* - \theta_0) \right\|_2 \right] \\
 & \quad + \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{1/2} (\theta_0 - \widehat{\theta}_t) \right\|_2 \\
 & \leq (K_{\min} n t_0)^{-1/2} (K_{\text{up}} n t)^{1/2} (4M_n \sqrt{p_*} + 4M_n \sqrt{p_*}) + (8K_{\min}^{-1/2} K_{\text{up}}^{1/2} + 1) M_n \sqrt{p_*} \\
 & \leq (16K_{\min}^{-1/2} K_{\text{up}}^{1/2} + 1) M_n \sqrt{p_*}.
 \end{aligned} \tag{186}$$

where the fourth inequality holds by (181) and (183). It follows that

$$\widehat{\theta}_{1:t_0} \in \Theta_{n,t} := \Theta \left(\widehat{\theta}_t, \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}, [16K_{\min}^{-1/2} K_{\text{up}}^{1/2} + 1] M_n \sqrt{p_*} \right).$$

For $\theta \in \Theta_{n,t}$, we have

$$\left\| \theta - \widehat{\theta}_t \right\|_2 \stackrel{(181)}{\leq} (K_{\text{low}} n t)^{-1/2} (16K_{\min}^{-1/2} K_{\text{up}}^{1/2} + 1) M_n \sqrt{p_*} \stackrel{(S)}{\leq} 1/4.$$

It follows that $\Theta_{n,t} \subseteq \Theta(\theta_0, \mathbf{I}_p, 1/2)$ because $\|\theta_0 - \widehat{\theta}_t\|_2 \leq 1/4$ by (181). Combining **(A2)** and Lemma 42, $\widetilde{L}_t(\cdot)$ satisfies the third order smoothness at $\widehat{\theta}_t$ with parameters

$$\left(K_{\max} K_{\text{low}}^{-3/2} t^{-3/2} n^{-1/2}, \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}, (16K_{\min}^{-1/2} K_{\text{up}}^{1/2} + 1) M_n \sqrt{p_*} \right). \tag{187}$$

Now, we are ready to obtain upper bounds of $(i)_t$, $(ii)_t$ and $(iii)_t$.

Step 3: $(i)_t$

Note that

$$\begin{aligned}
 \left\| \widetilde{\mathbf{F}}_{t_0, \widehat{\theta}_{t_0}}^{-1/2} \dot{\mathcal{R}}_{t,3}(\widehat{\theta}_t, \widehat{\theta}_{1:t_0} - \widehat{\theta}_t) \right\|_2 & \leq \left\| \widetilde{\mathbf{F}}_{t_0, \widehat{\theta}_{t_0}}^{-1/2} \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{1/2} \right\|_2 \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{-1/2} \dot{\mathcal{R}}_{t,3}(\widehat{\theta}_t, \widehat{\theta}_{1:t_0} - \widehat{\theta}_t) \right\|_2 \\
 & \stackrel{(181)}{\leq} (K_{\text{low}}^{-1/2} K_{\text{up}}^{1/2}) \sqrt{\frac{t}{t_0}} \left\| \widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}^{-1/2} \dot{\mathcal{R}}_{t,3}(\widehat{\theta}_t, \widehat{\theta}_{1:t_0} - \widehat{\theta}_t) \right\|_2.
 \end{aligned} \tag{188}$$

Also, by $\widehat{\theta}_{1:t_0} \in \Theta_{n,t}$ and Taylor's theorem, we have

$$\begin{aligned}
 & \left\| \widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}^{-1/2} \dot{\mathcal{R}}_{t,3}(\widehat{\theta}_t, \widehat{\theta}_{1:t_0} - \widehat{\theta}_t) \right\|_2 \\
 & \leq \frac{1}{2} \sup_{\widetilde{u} \in \overline{\Theta}_{n,t}} \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \left| \left\langle \nabla^3 \widetilde{L}_t(\widehat{\theta}_t + \widetilde{u}), (\widehat{\theta}_{1:t_0} - \widehat{\theta}_t)^{\otimes 2} \otimes (\widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}^{-1/2} u) \right\rangle \right| \\
 & \stackrel{(187)}{\leq} \frac{1}{2} (K_{\max} K_{\text{low}}^{-3/2} t^{-3/2} n^{-1/2}) \left\| \widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}^{1/2} (\widehat{\theta}_{1:t_0} - \widehat{\theta}_t) \right\|_2^2 \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \left\| \widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}^{1/2} \widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}^{-1/2} u \right\|_2 \\
 & \stackrel{(186)}{\leq} \frac{1}{2} (K_{\max} K_{\text{low}}^{-3/2} t^{-3/2} n^{-1/2}) (16K_{\min}^{-1/2} K_{\text{up}}^{1/2} + 1)^2 M_n^2 p_* \\
 & = \left(\frac{K_{\max}}{2K_{\text{low}}^{3/2}} (16K_{\min}^{-1/2} K_{\text{up}}^{1/2} + 1)^2 \right) M_n^2 t^{-3/2} \left(\frac{p_*^2}{n} \right)^{1/2},
 \end{aligned}$$

where

$$\overline{\Theta}_{n,t} = \Theta \left(\widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}, [16K_{\min}^{-1/2} K_{\text{up}}^{1/2} + 1] M_n \sqrt{p_*} \right).$$

Combining with (188), therefore, we have

$$\left\| \widetilde{\mathbf{F}}_{t_0,\widehat{\theta}_{t_0}}^{-1/2} \dot{\mathcal{R}}_{t,3}(\widehat{\theta}_t, \widehat{\theta}_{1:t_0} - \widehat{\theta}_t) \right\|_2 \leq c_1 M_n^2 t_0^{-1/2} t^{-1} \left(\frac{p_*^2}{n} \right)^{1/2}$$

for some positive constant $c_1 = c_1(K_{\min}, K_{\max}, K_{\text{low}}, K_{\text{up}})$. Consequently, we have

$$\begin{aligned}
 & \sum_{t=1}^{t_0-1} \left\| \widetilde{\mathbf{F}}_{t_0,\widehat{\theta}_{t_0}}^{-1/2} \dot{\mathcal{R}}_{t,3}(\widehat{\theta}_t, \widehat{\theta}_{1:t_0} - \widehat{\theta}_t) \right\|_2 \leq \sum_{t=1}^{t_0-1} c_1 M_n^2 t_0^{-1/2} t^{-1} \left(\frac{p_*^2}{n} \right)^{1/2} \\
 & = c_1 M_n^2 t_0^{-1/2} \left(\frac{p_*^2}{n} \right)^{1/2} \sum_{t=1}^{t_0-1} t^{-1} \leq c_1 M_n^2 (\log t_0 + 1) t_0^{-1/2} \left(\frac{p_*^2}{n} \right)^{1/2} \\
 & \leq 2c_1 M_n^2 \left(\frac{p_*^2}{n} \right)^{1/2},
 \end{aligned}$$

where the last two inequalities hold by $\sum_{t=1}^{t_0-1} t^{-1} \leq \log t_0 + 1$ and $t_0^{-1/2} (\log t_0 + 1) \leq 2$.

Step 4: (ii)_t

Next, we will obtain an upper bound of (ii)_t. Note that

$$\begin{aligned}
 & \left\| \widetilde{\mathbf{F}}_{t_0,\widehat{\theta}_{t_0}}^{-1/2} \left(\boldsymbol{\Omega}_t - \widetilde{\mathbf{F}}_{t,\widehat{\theta}_t} \right) (\widehat{\theta}_{1:t_0} - \widehat{\theta}_t) \right\|_2 \leq \left\| \widetilde{\mathbf{F}}_{t_0,\widehat{\theta}_{t_0}}^{-1/2} \left(\boldsymbol{\Omega}_t - \widetilde{\mathbf{F}}_{t,\widehat{\theta}_t} \right) \widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}^{-1/2} \right\|_2 \left\| \widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}^{1/2} (\widehat{\theta}_{1:t_0} - \widehat{\theta}_t) \right\|_2 \\
 & = \left\| \widetilde{\mathbf{F}}_{t_0,\widehat{\theta}_{t_0}}^{-1/2} \widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}^{1/2} \left(\widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}^{-1/2} \boldsymbol{\Omega}_t \widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}^{-1/2} - \mathbf{I}_p \right) \right\|_2 \left\| \widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}^{1/2} (\widehat{\theta}_{1:t_0} - \widehat{\theta}_t) \right\|_2 \\
 & \leq \left\| \widetilde{\mathbf{F}}_{t_0,\widehat{\theta}_{t_0}}^{-1/2} \widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}^{1/2} \right\|_2 \left\| \widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}^{-1/2} \boldsymbol{\Omega}_t \widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_2 \left\| \widetilde{\mathbf{F}}_{t,\widehat{\theta}_t}^{1/2} (\widehat{\theta}_{1:t_0} - \widehat{\theta}_t) \right\|_2.
 \end{aligned}$$

Note that

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} \right\|_2 &\stackrel{(181)}{\leq} (K_{\text{low}}^{-1/2} K_{\text{up}}^{1/2}) \sqrt{\frac{t}{t_0}}, \\ \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} \boldsymbol{\Omega}_t \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_2 &\stackrel{(181)}{\leq} K_{\text{up}} \left(\frac{p_*^2}{nt^2} \right)^{1/2}. \end{aligned}$$

Also,

$$\left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} (\hat{\theta}_{1:t_0} - \hat{\theta}_t) \right\|_2 \stackrel{(186)}{\leq} (16K_{\text{min}}^{-1/2} K_{\text{up}}^{1/2} + 1) M_n \sqrt{p_*}.$$

Combining the above three inequalities, we have

$$\begin{aligned} &\left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \left(\boldsymbol{\Omega}_t - \tilde{\mathbf{F}}_{t, \hat{\theta}_t} \right) (\hat{\theta}_{1:t_0} - \hat{\theta}_t) \right\|_2 \\ &\leq \left(K_{\text{low}}^{-1/2} K_{\text{up}}^{3/2} (16K_{\text{min}}^{-1/2} K_{\text{up}}^{1/2} + 1) \right) M_n t^{-1/2} t_0^{-1/2} \left(\frac{p_*^3}{n} \right)^{1/2} \\ &= c_2 M_n t^{-1/2} t_0^{-1/2} \left(\frac{p_*^3}{n} \right)^{1/2}, \end{aligned}$$

where $c_2 = K_{\text{low}}^{-1/2} K_{\text{up}}^{3/2} (16K_{\text{min}}^{-1/2} K_{\text{up}}^{1/2} + 1)$. Consequently, we have

$$\begin{aligned} &\sum_{t=1}^{t_0-1} \left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \left(\boldsymbol{\Omega}_t - \tilde{\mathbf{F}}_{t, \hat{\theta}_t} \right) (\hat{\theta}_{1:t_0} - \hat{\theta}_t) \right\|_2 \leq \sum_{t=1}^{t_0-1} c_2 M_n t^{-1/2} t_0^{-1/2} \left(\frac{p_*^3}{n} \right)^{1/2} \\ &= c_2 M_n t_0^{-1/2} \left(\frac{p_*^3}{n} \right)^{1/2} \sum_{t=1}^{t_0-1} t^{-1/2} \leq c_2 M_n t_0^{-1/2} \left(\frac{p_*^3}{n} \right)^{1/2} (2t_0^{1/2} - 1) \\ &\leq 2c_2 M_n \left(\frac{p_*^3}{n} \right)^{1/2}. \end{aligned}$$

Step 5: (iii)_t

Next, we will obtain an upper bound of (iii)_t. Note that

$$\left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \boldsymbol{\Omega}_t (\hat{\theta}_t - \mu_t) \right\|_2 \leq \left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} \right\|_2 \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} \boldsymbol{\Omega}_t^{1/2} \right\|_2 \left\| \boldsymbol{\Omega}_t^{1/2} (\hat{\theta}_t - \mu_t) \right\|_2.$$

Note that

$$\begin{aligned} &\left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} \right\|_2 \stackrel{(181)}{\leq} (K_{\text{low}}^{-1/2} K_{\text{up}}^{1/2}) \sqrt{\frac{t}{t_0}}, \\ &\left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} \boldsymbol{\Omega}_t^{1/2} \right\|_2 \leq \left(\left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} \boldsymbol{\Omega}_t \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_2 + 1 \right)^{1/2} \stackrel{(181)}{\leq} \left(K_{\text{up}} \left(\frac{p_*^2}{nt^2} \right)^{1/2} + 1 \right)^{1/2} \stackrel{\text{(S)}}{\leq} 2, \\ &\left\| \boldsymbol{\Omega}_t^{1/2} (\hat{\theta}_t - \mu_t) \right\|_2 \stackrel{(181)}{\leq} K_{\text{up}} \left(\frac{p_*^2}{nt^2} \right)^{1/2}. \end{aligned}$$

By the last display, we have

$$\left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \boldsymbol{\Omega}_t(\hat{\theta}_t - \mu_t) \right\|_2 \leq (2K_{\text{low}}^{-1/2} K_{\text{up}}^{3/2}) t^{-1/2} t_0^{-1/2} \left(\frac{p_*^2}{n} \right)^{1/2} = c_3 t^{-1/2} t_0^{-1/2} \left(\frac{p_*^2}{n} \right)^{1/2},$$

where $c_3 = 2K_{\text{low}}^{-1/2} K_{\text{up}}^{3/2}$. Consequently, we have

$$\begin{aligned} \sum_{t=1}^{t_0-1} \left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \boldsymbol{\Omega}_t(\hat{\theta}_t - \mu_t) \right\|_2 &\leq \sum_{t=1}^{t_0-1} c_3 t^{-1/2} t_0^{-1/2} \left(\frac{p_*^2}{n} \right)^{1/2} \\ &= c_3 t_0^{-1/2} \left(\frac{p_*^2}{n} \right)^{1/2} \sum_{t=1}^{t_0-1} t^{-1/2} \leq c_3 t_0^{-1/2} \left(\frac{p_*^2}{n} \right)^{1/2} (2t_0^{1/2} - 1) \leq 2c_3 \left(\frac{p_*^2}{n} \right)^{1/2}. \end{aligned}$$

Step 6: Upper bound of $\left\| \sum_{t=1}^{t_0-1} \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \nabla \eta_t(\hat{\theta}_{1:t_0}) \right\|_2$

Let

$$\varrho_{n, t_0} = \left\| \sum_{t=1}^{t_0-1} \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{-1/2} \nabla \eta_t(\hat{\theta}_{1:t_0}) \right\|_2.$$

Combining the results in **Step 3-5**, note that

$$\begin{aligned} \varrho_{n, t_0} &\leq 2c_1 M_n^2 \left(\frac{p_*^2}{n} \right)^{1/2} + 2c_2 M_n \left(\frac{p_*^3}{n} \right)^{1/2} + 2c_3 \left(\frac{p_*^2}{n} \right)^{1/2} \\ &\leq 2(c_1 + c_2 + c_3) M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2} = c_4 M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2} \end{aligned} \tag{189}$$

where $c_4 = 2(c_1 + c_2 + c_3)$, and the second inequality holds by $M_n \wedge p_* \geq 1$. In this proof, we denote

$$\tilde{\Theta}_{n, t_0} = \Theta(\hat{\theta}_{t_0}, \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}, 4\varrho_{n, t_0}).$$

By (181), we have $\|\hat{\theta}_{t_0} - \theta_0\|_2 \leq 1/4$. Also, for $\theta \in \tilde{\Theta}_{n, t_0}$, we have

$$\begin{aligned} \|\theta - \hat{\theta}_{t_0}\|_2 &\stackrel{(181)}{\leq} (K_{\text{low}} n t_0)^{-1/2} 4\varrho_{n, t_0} \\ &\stackrel{(189)}{\leq} (K_{\text{low}} n t_0)^{-1/2} 4c_4 M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2} \\ &= (4K_{\text{low}}^{-1/2} c_4) M_n^2 p_*^{3/2} n^{-1} t_0^{-1/2} \stackrel{(S)}{\leq} 1/4. \end{aligned}$$

It follows that $\tilde{\Theta}_{n, t_0} \subseteq \Theta(\theta_0, \mathbf{I}_p, 1/2)$. Then, by Lemma 42 and (181), $\tilde{L}_{t_0}(\cdot)$ satisfies the third order smoothness at $\hat{\theta}_{t_0}$ with parameters

$$\left(K_{\text{max}} K_{\text{low}}^{-3/2} t_0^{-3/2} n^{-1/2}, \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}, 4\varrho_{n, t_0} \right). \tag{190}$$

Step 7: (184) and (54)

By **Step 1-6**, we are ready to prove (184). Note that

$$K_{\max} K_{\text{low}}^{-3/2} t_0^{-3/2} n^{-1/2} \varrho_{n,t_0} \leq (K_{\max} K_{\text{low}}^{-3/2} c_4) M_n^2 p_*^{3/2} n^{-1} t_0^{-3/2} \stackrel{\text{(S)}}{\leq} 1/16, \quad (191)$$

which allows us to apply Lemma 41. By Lemma 41 with

$$\begin{aligned} \tau_3 &= K_{\max} K_{\text{low}}^{-3/2} t_0^{-3/2} n^{-1/2}, \quad f(\theta) = \tilde{L}_{t_0}(\theta), \quad \theta = \hat{\theta}_{t_0}, \quad \tilde{\theta} = \hat{\theta}_{1:t_0}, \\ \beta &= \sum_{t=1}^{t_0-1} \nabla \eta_t(\hat{\theta}_{1:t_0}), \quad r = \varrho_{n,t_0}, \end{aligned}$$

we have

$$\left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{1/2} (\hat{\theta}_{1:t_0} - \hat{\theta}_{t_0}) \right\|_2 \leq 4\varrho_{n,t_0} \leq 4c_4 M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2}, \quad (192)$$

which completes the proof of the first assertion in (184). It follows that

$$\begin{aligned} & \left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{1/2} (\theta_0 - \hat{\theta}_{t_0}) \right\|_2 \leq \left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{1/2} (\theta_0 - \hat{\theta}_{1:t_0}) \right\|_2 + \left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{1/2} (\hat{\theta}_{1:t_0} - \hat{\theta}_{t_0}) \right\|_2 \\ & \leq \left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{1/2} \tilde{\mathbf{F}}_{1:t_0, \hat{\theta}_{1:t_0}}^{-1/2} \right\|_2 \left\| \tilde{\mathbf{F}}_{1:t_0, \hat{\theta}_{1:t_0}}^{1/2} (\theta_0 - \hat{\theta}_{1:t_0}) \right\|_2 + \left\| \tilde{\mathbf{F}}_{t_0, \hat{\theta}_{t_0}}^{1/2} (\hat{\theta}_{1:t_0} - \hat{\theta}_{t_0}) \right\|_2 \\ & \stackrel{(181)}{\leq} (K_{\min} n t_0)^{-1/2} (K_{\text{up}} n t_0)^{1/2} (4M_n \sqrt{p_*} + 4M_n \sqrt{p_*}) + 4c_4 M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2} \\ & = (8K_{\min}^{-1/2} K_{\text{up}}^{1/2} + 4c_4 M_n p_* n^{-1/2}) M_n \sqrt{p_*} \\ & \stackrel{\text{(S)}}{\leq} (8K_{\min}^{-1/2} K_{\text{up}}^{1/2} + 1) M_n \sqrt{p_*} = c_5 M_n \sqrt{p_*} \end{aligned}$$

for some constant $c_5 = 8K_{\min}^{-1/2} K_{\text{up}}^{1/2} + 1$. This completes the proof of (184).

It follows by induction that

$$\left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} (\hat{\theta}_{1:t} - \hat{\theta}_t) \right\|_2 \leq 4c_4 M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2}, \quad \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} (\theta_0 - \hat{\theta}_t) \right\|_2 \leq c_5 M_n \sqrt{p_*}, \quad \forall t \in [T].$$

Also, for any $t \in [T]$,

$$\begin{aligned} & \left\| \mathbf{\Omega}_t^{1/2} (\hat{\theta}_{1:t} - \mu_t) \right\|_2 \leq \left\| \mathbf{\Omega}_t^{1/2} (\hat{\theta}_{1:t} - \hat{\theta}_t) \right\|_2 + \left\| \mathbf{\Omega}_t^{1/2} (\hat{\theta}_t - \mu_t) \right\|_2 \\ & \stackrel{\text{Corollary 17}}{\leq} (1 + \Delta_t)^{1/2} \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} (\hat{\theta}_{1:t} - \hat{\theta}_t) \right\|_2 + \Delta_t \leq (1 + \Delta_t) \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} (\hat{\theta}_{1:t} - \hat{\theta}_t) \right\|_2 + \Delta_t \\ & \stackrel{(181)}{\leq} (1 + K_{\text{up}} t^{-1} n^{-1/2} p_*) 4c_4 M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2} + K_{\text{up}} t^{-1} n^{-1/2} p_* \\ & = \left[4c_4 (1 + K_{\text{up}} t^{-1} n^{-1/2} p_*) + K_{\text{up}} M_n^{-2} t^{-1} p_*^{-1/2} \right] M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2} \\ & \stackrel{\text{(S)}}{\leq} c_6 M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2} \end{aligned}$$

for some positive constant $c_6 = c_6(c_4, K_{\text{up}})$. This completes the proof of (54). \blacksquare

Proof [Proof of Proposition 10] In this proof, we work on the event $\mathcal{E}_{\text{est},1} \cap \mathcal{E}_{\text{est},2}$ without explicitly referring to it. Given the conditions in Proposition 9, the conditions required for Lemmas 24, 25 and Proposition 23 are satisfied. It follows that

$$\begin{aligned}
 \left\| \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{1/2} (\theta_0 - \theta_{1:t}^*) \right\|_2 &\leq 4M_n p_*^{1/2}, & \lambda_{\min}(\tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}) &\geq K_{\min} n t \\
 \left\| \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{1/2} (\hat{\theta}_{1:t} - \theta_{1:t}^*) \right\|_2 &\leq 4M_n p_*^{1/2}, & \lambda_{\max}(\tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}) &\leq \frac{4}{3} K_{\max} n t, \\
 \lambda_{\min}(\tilde{\mathbf{F}}_{t,\hat{\theta}_t}) \wedge \lambda_{\min}(\tilde{\mathbf{F}}_{t,\theta_t^*}) &\geq K_{\text{low}} n t, & \lambda_{\max}(\tilde{\mathbf{F}}_{t,\hat{\theta}_t}) \wedge \lambda_{\max}(\tilde{\mathbf{F}}_{t,\theta_t^*}) &\leq K_{\text{up}} n t, \\
 \hat{\theta}_t, \theta_t^*, \theta_{1:t}^*, \hat{\theta}_{1:t} &\in \Theta(\theta_0, \mathbf{I}_p, 1/4), & \Delta_t \vee \epsilon_{n,t,\text{KL}} &\leq K_{\text{up}} t^{-1} n^{-1/2} p_*
 \end{aligned} \tag{193}$$

for all $t \in [T]$.

Step 1: Proof framework

For $t \in [T]$, note that

$$\begin{aligned}
 &\left\| \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{-1/2} \boldsymbol{\Omega}_t \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{-1/2} - \mathbf{I}_p \right\|_{\text{F}} \leq \lambda_{\min}^{-1}(\tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}) \left\| \boldsymbol{\Omega}_t - \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \right\|_{\text{F}} \\
 &\leq \lambda_{\min}^{-1}(\tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}) \left(\left\| \boldsymbol{\Omega}_t - \tilde{\mathbf{F}}_{t,\hat{\theta}_t} \right\|_{\text{F}} + \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t} - \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \right\|_{\text{F}} \right) \\
 &\leq \lambda_{\min}^{-1}(\tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}) \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t} \right\|_2 \left[\left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \boldsymbol{\Omega}_t \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_{\text{F}} + \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_{\text{F}} \right] \\
 &\stackrel{(193)}{\leq} K_{\min}^{-1} K_{\text{up}} \left(\left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \boldsymbol{\Omega}_t \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_{\text{F}} + \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_{\text{F}} \right) \\
 &\stackrel{(193)}{\leq} K_{\min}^{-1} K_{\text{up}} \left[K_{\text{up}} t^{-1} \left(\frac{p_*^2}{n} \right)^{1/2} + \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_{\text{F}} \right].
 \end{aligned}$$

Hence, we only need to obtain an upper bound of

$$\left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_{\text{F}}.$$

Note that

$$\tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} = \tilde{\mathbf{F}}_{t,\hat{\theta}_t} + \sum_{s=1}^{t-1} \left[\left(\tilde{\mathbf{F}}_{s,\hat{\theta}_s} - \boldsymbol{\Omega}_s \right) + \left(\mathbf{F}_{s,\hat{\theta}_{1:t}} - \mathbf{F}_{s,\hat{\theta}_s} \right) \right] + \left(\mathbf{F}_{t,\hat{\theta}_{1:t}} - \mathbf{F}_{t,\hat{\theta}_t} \right)$$

It follows that

$$\begin{aligned}
 & \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} - \mathbf{I}_p \\
 &= \sum_{s=1}^{t-1} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \left[\left(\tilde{\mathbf{F}}_{s,\hat{\theta}_s} - \Omega_s \right) + \left(\mathbf{F}_{s,\hat{\theta}_{1:t}} - \mathbf{F}_{s,\hat{\theta}_s} \right) \right] \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} + \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \left(\mathbf{F}_{t,\hat{\theta}_{1:t}} - \mathbf{F}_{t,\hat{\theta}_t} \right) \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \\
 &= \left(- \sum_{s=1}^{t-1} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \tilde{\mathbf{F}}_{s,\hat{\theta}_s}^{1/2} \left[\tilde{\mathbf{F}}_{s,\hat{\theta}_s}^{-1/2} \Omega_s \tilde{\mathbf{F}}_{s,\hat{\theta}_s}^{-1/2} - \mathbf{I}_p \right] \tilde{\mathbf{F}}_{s,\hat{\theta}_s}^{1/2} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \right) \\
 & \quad + \left(\sum_{s=1}^t \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \mathbf{F}_{s,\hat{\theta}_s}^{1/2} \left[\mathbf{F}_{s,\hat{\theta}_s}^{-1/2} \mathbf{F}_{s,\hat{\theta}_{1:t}} \mathbf{F}_{s,\hat{\theta}_s}^{-1/2} - \mathbf{I}_p \right] \mathbf{F}_{s,\hat{\theta}_s}^{1/2} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \right).
 \end{aligned}$$

Let

$$\begin{aligned}
 \text{(i)} &= \left\| \sum_{s=1}^{t-1} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \tilde{\mathbf{F}}_{s,\hat{\theta}_s}^{1/2} \left[\tilde{\mathbf{F}}_{s,\hat{\theta}_s}^{-1/2} \Omega_s \tilde{\mathbf{F}}_{s,\hat{\theta}_s}^{-1/2} - \mathbf{I}_p \right] \tilde{\mathbf{F}}_{s,\hat{\theta}_s}^{1/2} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \right\|_{\mathbf{F}}, \\
 \text{(ii)} &= \left\| \sum_{s=1}^t \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \mathbf{F}_{s,\hat{\theta}_s}^{1/2} \left[\mathbf{F}_{s,\hat{\theta}_s}^{-1/2} \mathbf{F}_{s,\hat{\theta}_{1:t}} \mathbf{F}_{s,\hat{\theta}_s}^{-1/2} - \mathbf{I}_p \right] \mathbf{F}_{s,\hat{\theta}_s}^{1/2} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \right\|_{\mathbf{F}}.
 \end{aligned}$$

Note that

$$\left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_{\mathbf{F}} \leq \text{(i)} + \text{(ii)}.$$

Step 2: (i)

Let $s \in [t-1]$. Note that

$$\begin{aligned}
 & \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \tilde{\mathbf{F}}_{s,\hat{\theta}_s}^{1/2} \left[\tilde{\mathbf{F}}_{s,\hat{\theta}_s}^{-1/2} \Omega_s \tilde{\mathbf{F}}_{s,\hat{\theta}_s}^{-1/2} - \mathbf{I}_p \right] \tilde{\mathbf{F}}_{s,\hat{\theta}_s}^{1/2} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \right\|_{\mathbf{F}} \\
 & \leq \lambda_{\min}^{-1}(\tilde{\mathbf{F}}_{t,\hat{\theta}_t}) \lambda_{\max}(\tilde{\mathbf{F}}_{s,\hat{\theta}_s}) \left\| \tilde{\mathbf{F}}_{s,\hat{\theta}_s}^{-1/2} \Omega_s \tilde{\mathbf{F}}_{s,\hat{\theta}_s}^{-1/2} - \mathbf{I}_p \right\|_{\mathbf{F}} \\
 & \stackrel{\text{(193)}}{\leq} \underset{\text{Corollary 17}}{\left[\frac{K_{\text{up}} s}{K_{\text{low}} t} \right]} \left[K_{\text{up}} s^{-1} \left(\frac{p_*^2}{n} \right)^{1/2} \right] = (K_{\text{up}}^2 K_{\text{low}}^{-1}) t^{-1} \left(\frac{p_*^2}{n} \right)^{1/2}.
 \end{aligned}$$

By the last display, we have

$$\text{(i)} \leq \sum_{s=1}^{t-1} (K_{\text{up}}^2 K_{\text{low}}^{-1}) t^{-1} \left(\frac{p_*^2}{n} \right)^{1/2} = (K_{\text{up}}^2 K_{\text{low}}^{-1}) t^{-1} \left(\frac{p_*^2}{n} \right)^{1/2} \sum_{s=1}^{t-1} 1 \leq c_1 \left(\frac{p_*^2}{n} \right)^{1/2},$$

where $c_1 = K_{\text{up}}^2 K_{\text{low}}^{-1}$.

Step 3: (ii)

Let $s \in [t]$. Note that

$$\begin{aligned}
 & \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \mathbf{F}_{s,\hat{\theta}_s}^{1/2} \left[\mathbf{F}_{s,\hat{\theta}_s}^{-1/2} \mathbf{F}_{s,\hat{\theta}_{1:t}} \mathbf{F}_{s,\hat{\theta}_s}^{-1/2} - \mathbf{I}_p \right] \mathbf{F}_{s,\hat{\theta}_s}^{1/2} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \right\|_{\mathbf{F}} \\
 & \leq \lambda_{\min}^{-1}(\tilde{\mathbf{F}}_{t,\hat{\theta}_t}) \lambda_{\max}(\mathbf{F}_{s,\hat{\theta}_s}) \left\| \mathbf{F}_{s,\hat{\theta}_s}^{-1/2} \mathbf{F}_{s,\hat{\theta}_{1:t}} \mathbf{F}_{s,\hat{\theta}_s}^{-1/2} - \mathbf{I}_p \right\|_{\mathbf{F}} \\
 & \leq \left(\frac{K_{\text{max}}}{K_{\text{low}}} t^{-1} \right) \sqrt{p} \left\| \mathbf{F}_{s,\hat{\theta}_s}^{-1/2} \mathbf{F}_{s,\hat{\theta}_{1:t}} \mathbf{F}_{s,\hat{\theta}_s}^{-1/2} - \mathbf{I}_p \right\|_2.
 \end{aligned} \tag{194}$$

where the last inequality holds by (193) and **(A2)**. Also, by Proposition 9, we have

$$\left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} (\hat{\theta}_{1:t} - \hat{\theta}_t) \right\|_2 \leq KM_n^2 \left(\frac{p_*^3}{n} \right)^{1/2}, \quad \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} (\theta_0 - \hat{\theta}_t) \right\|_2 \leq KM_n p_*^{1/2}, \quad \forall t \in [T],$$

where $K = K(K_{\text{low}}, K_{\text{up}}, K_{\text{min}}, K_{\text{max}}) > 0$. It follows that

$$\begin{aligned} & \left\| \mathbf{F}_{s, \hat{\theta}_s}^{1/2} (\hat{\theta}_{1:t} - \hat{\theta}_s) \right\|_2 \\ & \leq \left\| \mathbf{F}_{s, \hat{\theta}_s}^{1/2} (\hat{\theta}_{1:t} - \hat{\theta}_t) \right\|_2 + \left\| \mathbf{F}_{s, \hat{\theta}_s}^{1/2} (\hat{\theta}_t - \theta_0) \right\|_2 + \left\| \mathbf{F}_{s, \hat{\theta}_s}^{1/2} (\theta_0 - \hat{\theta}_s) \right\|_2 \\ & \leq \left\| \mathbf{F}_{s, \hat{\theta}_s}^{1/2} \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} \right\|_2 \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} (\hat{\theta}_{1:t} - \hat{\theta}_t) \right\|_2 + \left\| \mathbf{F}_{s, \hat{\theta}_s}^{1/2} \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{-1/2} \right\|_2 \left\| \tilde{\mathbf{F}}_{t, \hat{\theta}_t}^{1/2} (\hat{\theta}_t - \theta_0) \right\|_2 + \left\| \tilde{\mathbf{F}}_{s, \hat{\theta}_s}^{1/2} (\theta_0 - \hat{\theta}_s) \right\|_2 \\ & \leq \left(\frac{K_{\text{max}}}{K_{\text{low}} t} \right)^{1/2} (KM_n^2 p_*^{3/2} n^{-1/2}) + \left(\frac{K_{\text{max}}}{K_{\text{low}} t} \right)^{1/2} (KM_n p_*^{1/2}) + KM_n p_*^{1/2} \\ & = \left(KK_{\text{max}}^{1/2} K_{\text{low}}^{-1/2} t^{-1/2} M_n p_* n^{-1/2} + K_{\text{max}}^{1/2} K_{\text{low}}^{-1/2} t^{-1/2} K + K \right) M_n p_*^{1/2} \\ & \stackrel{\text{(S)}}{\leq} K (K_{\text{max}}^{1/2} K_{\text{low}}^{-1/2} + 2) M_n p_*^{1/2} = c_2 M_n p_*^{1/2}, \end{aligned}$$

where the third inequality holds by (193) and $\mathbf{F}_{s, \hat{\theta}_s} \preceq \tilde{\mathbf{F}}_{s, \hat{\theta}_s}$, and $c_2 = K(K_{\text{max}}^{1/2} K_{\text{low}}^{-1/2} + 2)$. It follows that

$$\begin{aligned} \left\| \theta_0 - \hat{\theta}_s \right\|_2 & \stackrel{\text{(193)}}{\leq} 1/4, \\ \left\| \theta - \hat{\theta}_s \right\|_2 & \stackrel{\text{(A2)}}{\leq} (K_{\text{min}} n)^{-1/2} c_2 M_n p_*^{1/2} \stackrel{\text{(S)}}{\leq} 1/4, \quad \forall \theta \in \Theta(\hat{\theta}_s, \mathbf{F}_{s, \hat{\theta}_s}, c_2 M_n p_*^{1/2}) \\ \hat{\theta}_{1:t} & \in \Theta(\hat{\theta}_s, \mathbf{F}_{s, \hat{\theta}_s}, c_2 M_n p_*^{1/2}) \subseteq \Theta(\theta_0, \mathbf{I}_p, 1/2). \end{aligned}$$

which, combining with Lemma 42, implies that $L_s(\cdot)$ satisfies the third order smoothness at $\hat{\theta}_s$ with parameters

$$\left(K_{\text{max}} K_{\text{min}}^{-3/2} n^{-1/2}, \mathbf{F}_{s, \hat{\theta}_s}, c_2 M_n p_*^{1/2} \right).$$

By Lemma 40, it follows that

$$\left\| \mathbf{F}_{s, \hat{\theta}_s}^{-1/2} \mathbf{F}_{s, \hat{\theta}_{1:t}} \mathbf{F}_{s, \hat{\theta}_s}^{-1/2} - \mathbf{I}_p \right\|_2 \leq (K_{\text{max}} K_{\text{min}}^{-3/2} n^{-1/2}) (c_2 M_n p_*^{1/2}) = c_3 M_n \left(\frac{p_*}{n} \right)^{1/2},$$

where $c_3 = K_{\text{max}} K_{\text{min}}^{-3/2} c_2$. Hence, the right-hand side of (194) is bounded by

$$\left[\frac{K_{\text{max}}}{K_{\text{low}}} t^{-1} \right] \sqrt{p} \left(c_3 M_n \left(\frac{p_*}{n} \right)^{1/2} \right) \leq c_4 M_n t^{-1} \left(\frac{p_*^2}{n} \right)^{1/2}$$

where $c_4 = K_{\text{low}}^{-1} K_{\text{max}} c_3$. Consequently, we have

$$\text{(ii)} \leq \sum_{s=1}^t c_4 M_n t^{-1} \left(\frac{p_*^2}{n} \right)^{1/2} = c_4 M_n \left(\frac{p_*^2}{n} \right)^{1/2}.$$

Step 4

By **Step 2-3**, we have

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_{\mathbf{F}} &\leq \text{(i)} + \text{(ii)} \leq c_1 \left(\frac{p_*^2}{n} \right)^{1/2} + c_4 M_n \left(\frac{p_*^2}{n} \right)^{1/2} \\ &\leq (c_1 + c_4) M_n \left(\frac{p_*^2}{n} \right)^{1/2}. \end{aligned}$$

By **Step 1** and the last display, we have

$$\begin{aligned} &\left\| \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{-1/2} \boldsymbol{\Omega}_t \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{-1/2} - \mathbf{I}_p \right\|_{\mathbf{F}} \\ &\leq K_{\min}^{-1} K_{\text{up}} \left[K_{\text{up}} t^{-1} \left(\frac{p_*^2}{n} \right)^{1/2} + \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_{\mathbf{F}} \right] \\ &\leq K_{\min}^{-1} K_{\text{up}} \left[K_{\text{up}} t^{-1} \left(\frac{p_*^2}{n} \right)^{1/2} + (c_1 + c_4) M_n \left(\frac{p_*^2}{n} \right)^{1/2} \right] \\ &\leq c_5 M_n \left(\frac{p_*^2}{n} \right)^{1/2}, \end{aligned}$$

where $c_5 = c_5(K_{\min}, K_{\text{up}}, c_1, c_4) > 0$. This completes the proof. \blacksquare

Proof [Proof of Theorem 11] In this proof, we work on the event $\mathcal{E}_{\text{est},1} \cap \mathcal{E}_{\text{est},2}$ without explicitly referring to it. Note that $\mathbb{P}_0^{(N)}(\mathcal{E}_{\text{est},1} \cap \mathcal{E}_{\text{est},2}) \geq 1 - 3n^{-1}$. For all $t \in [T]$, we have

$$\begin{aligned} &d_V \left(\Pi_t(\cdot), \Pi(\cdot \mid \mathbf{D}_{1:t}) \right) \\ &\leq d_V \left(\Pi_t(\cdot), \mathcal{N} \left(\hat{\theta}_{1:t}, \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{-1} \right) \right) + d_V \left(\mathcal{N} \left(\hat{\theta}_{1:t}, \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{-1} \right), \Pi(\cdot \mid \mathbf{D}_{1:t}) \right) \\ &\stackrel{\text{Theorem 7}}{\leq} d_V \left(\Pi_t(\cdot), \mathcal{N} \left(\hat{\theta}_{1:t}, \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{-1} \right) \right) + K_1 \left(\frac{p_*^2}{nt} \right)^{1/2}, \end{aligned}$$

where $K_1 = K_1(K_{\min}, K_{\max}) > 0$ is the constant K specified in Theorem 7. Hence, we only need to obtain an upper bound of

$$d_V \left(\Pi_t(\cdot), \mathcal{N} \left(\hat{\theta}_{1:t}, \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{-1} \right) \right).$$

By Propositions 9 and 10, we have

$$\begin{aligned} \left\| \boldsymbol{\Omega}_t^{1/2} (\hat{\theta}_{1:t} - \mu_t) \right\|_2 &\leq K_2 M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2}, \\ \left\| \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{-1/2} \boldsymbol{\Omega}_t \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{-1/2} - \mathbf{I}_p \right\|_{\mathbf{F}} &\leq K_3 M_n \left(\frac{p_*^2}{n} \right)^{1/2}, \end{aligned}$$

where K_2 and K_3 , depending only on $(K_{\min}, K_{\max}, K_{\text{low}}, K_{\text{up}})$, are the constants specified in Propositions 9 and 10, respectively. Also,

$$\left\| \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1/2} \boldsymbol{\Omega}_t \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1/2} - \mathbf{I}_p \right\|_2 \leq K_3 M_n \left(\frac{p_*^2}{n} \right)^{1/2} \stackrel{\mathbf{(S)}}{\leq} 0.684,$$

which, combining with Lemma 43, implies that

$$\begin{aligned} & d_V \left(\Pi_t(\cdot), \mathcal{N} \left(\hat{\theta}_{1:t}, \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1} \right) \right) \\ & \leq \frac{1}{2} \left(\left\| \boldsymbol{\Omega}_t^{1/2} (\hat{\theta}_{1:t} - \mu_t) \right\|_2^2 + \left\| \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1/2} \boldsymbol{\Omega}_t \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1/2} - \mathbf{I}_p \right\|_F^2 \right)^{1/2} \\ & \leq \frac{1}{2} \left(K_2 M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2} + K_3 M_n \left(\frac{p_*^2}{n} \right)^{1/2} \right) \\ & \leq \frac{1}{2} (K_2 + K_3) M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2}, \end{aligned}$$

where the second inequality holds by $x^2 + y^2 \leq (x + y)^2$ for $x, y \geq 0$. Therefore, we have

$$\begin{aligned} d_V \left(\Pi_t(\cdot), \Pi(\cdot \mid \mathbf{D}_{1:t}) \right) & \leq \frac{K_2 + K_3}{2} M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2} + K_1 \left(\frac{p_*^2}{nt} \right)^{1/2} \\ & \leq \frac{K_2 + K_3 + 2K_1}{2} M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2} = K_4 M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2}, \end{aligned}$$

where $K_4 = (K_2 + K_3 + 2K_1)/2$.

If we further assume (\mathbf{P}^*) , we can employ Theorem 8. For all $t \in [T]$, it holds that

$$d_V \left(\mathcal{N} \left(\hat{\theta}_{1:t}^{\text{ML}}, \mathbf{F}_{1:t, \theta_0}^{-1} \right), \Pi(\cdot \mid \mathbf{D}_{1:t}) \right) \leq K_5 M_n \left(\frac{p_*^2}{nt} \right)^{1/2},$$

where $K_5 = K_5(K_{\min}, K_{\max})$ is the constant specified in Theorem 8. It follows that

$$\begin{aligned} d_V \left(\Pi_t, \mathcal{N} \left(\hat{\theta}_{1:t}^{\text{ML}}, \mathbf{F}_{1:t, \theta_0}^{-1} \right) \right) & \leq d_V \left(\Pi_t(\cdot), \Pi(\cdot \mid \mathbf{D}_{1:t}) \right) + d_V \left(\mathcal{N} \left(\hat{\theta}_{1:t}^{\text{ML}}, \mathbf{F}_{1:t, \theta_0}^{-1} \right), \Pi(\cdot \mid \mathbf{D}_{1:t}) \right) \\ & \leq K_4 M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2} + K_5 M_n \left(\frac{p_*^2}{nt} \right)^{1/2} \\ & \leq (K_4 + K_5) M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2}, \end{aligned}$$

which completes the proof. ■

Corollary 26 *Suppose that $(\mathbf{A0})$, $(\mathbf{A1}^*)$, $(\mathbf{A2})$, (\mathbf{S}) and (\mathbf{P}^*) hold. Then, on $\mathcal{E}_{\text{est}, 2}$, for all $t \in [T]$, $\mathbb{E}L_{1:t}(\theta)$ satisfies the third order smoothness at θ_0 with parameters*

$$\left(K_{\max} K_{\min}^{-3/2} (nt)^{-1/2}, \mathbf{F}_{1:t, \theta_0}, 4r_{\text{eff}, 1:t} \right).$$

Furthermore, the following inequalities hold uniformly for all $t \in [T]$:

$$\begin{aligned} \left\| \mathbf{F}_{1:t, \theta_0}^{1/2} (\widehat{\theta}_{1:t}^{\text{ML}} - \theta_0) \right\|_2 &\leq 4r_{\text{eff}, 1:t} \leq 4M_n p_*^{1/2}, \\ \left\| \widetilde{\mathbf{F}}_{1:t, \widehat{\theta}_{1:t}}^{1/2} (\widehat{\theta}_{1:t}^{\text{ML}} - \widehat{\theta}_{1:t}) \right\|_2 &\leq KM_n \left(\frac{p_*^2}{nt} \right)^{1/2}, \end{aligned}$$

where $K = K(K_{\min}, K_{\max})$.

Proof See **Step 1** in the proof of Theorem 8. ■

Corollary 27 Suppose that $(\mathbf{A0})$, $(\mathbf{A1}^*)$, $(\mathbf{A2})$, (\mathbf{S}) and (\mathbf{P}^*) hold. Then, on $\mathcal{E}_{\text{est}, 1} \cap \mathcal{E}_{\text{est}, 2}$, the following inequalities holds uniformly for all $t \in [T]$:

$$\begin{aligned} \left\| \mathbf{F}_{1:t, \theta_0}^{1/2} (\widehat{\theta}_{1:t}^{\text{ML}} - \theta_0) - \mathbf{F}_{1:t, \theta_0}^{-1/2} \nabla L_{1:t}(\theta_0) \right\|_2 &\leq KM_n^2 \left(\frac{p_*^2}{n} \right)^{1/2}, \\ \left\| \mathbf{F}_{1:t, \theta_0}^{1/2} (\widehat{\theta}_t - \theta_0) - \mathbf{F}_{1:t, \theta_0}^{-1/2} \nabla L_{1:t}(\theta_0) \right\|_2 &\leq KM_n^2 \left(\frac{p_*^3}{n} \right)^{1/2}, \\ \left\| \mathbf{F}_{1:t, \theta_0}^{1/2} (\mu_t - \theta_0) - \mathbf{F}_{1:t, \theta_0}^{-1/2} \nabla L_{1:t}(\theta_0) \right\|_2 &\leq KM_n^2 \left(\frac{p_*^3}{n} \right)^{1/2}, \end{aligned}$$

where $K = K(K_{\min}, K_{\max}, K_{\text{low}}, K_{\text{up}}) > 0$.

Proof In this proof, we work on the event $\mathcal{E}_{\text{est}, 1} \cap \mathcal{E}_{\text{est}, 2}$ without explicitly referring to it. By Lemmas 24, 25 and Proposition 23, for all $t \in [T]$, we have

$$\begin{aligned} \lambda_{\min}(\widetilde{\mathbf{F}}_{t, \widehat{\theta}_t}) &\geq K_{\text{low}} nt, \\ \widehat{\theta}_t, \widehat{\theta}_{1:t} &\in \Theta(\theta_0, \mathbf{I}_p, 1/4), \\ \Delta_t &\leq K_{\text{up}} t^{-1} n^{-1/2} p_*. \end{aligned} \tag{195}$$

Let $t \in [T]$. By Taylor's theorem, we have

$$\begin{aligned} -\nabla L_{1:t}(\theta_0) &= \nabla L_{1:t}(\widehat{\theta}_{1:t}^{\text{ML}}) - \nabla L_{1:t}(\theta_0) = -\overline{\mathbf{F}}_{1:t}(\widehat{\theta}_{1:t}^{\text{ML}} - \theta_0) \\ &= -\mathbf{F}_{1:t, \theta_0}(\widehat{\theta}_{1:t}^{\text{ML}} - \theta_0) - (\overline{\mathbf{F}}_{1:t} - \mathbf{F}_{1:t, \theta_0})(\widehat{\theta}_{1:t}^{\text{ML}} - \theta_0). \end{aligned}$$

where

$$\overline{\mathbf{F}}_{1:t} = \overline{\mathbf{F}}_{1:t}(\widehat{\theta}_{1:t}^{\text{ML}}, \theta_0) = - \int_0^1 \nabla^2 L_{1:t}(s\widehat{\theta}_{1:t}^{\text{ML}} + (1-s)\theta_0) ds.$$

It follows that

$$\begin{aligned} \mathbf{F}_{1:t, \theta_0}^{1/2} (\widehat{\theta}_{1:t}^{\text{ML}} - \theta_0) &= \mathbf{F}_{1:t, \theta_0}^{-1/2} \nabla L_{1:t}(\theta_0) + \mathbf{F}_{1:t, \theta_0}^{-1/2} (\overline{\mathbf{F}}_{1:t} - \mathbf{F}_{1:t, \theta_0}) \mathbf{F}_{1:t, \theta_0}^{-1/2} \mathbf{F}_{1:t, \theta_0}^{1/2} (\widehat{\theta}_{1:t}^{\text{ML}} - \theta_0) \\ &= \mathbf{F}_{1:t, \theta_0}^{-1/2} \nabla L_{1:t}(\theta_0) + (\mathbf{F}_{1:t, \theta_0}^{-1/2} \overline{\mathbf{F}}_{1:t} \mathbf{F}_{1:t, \theta_0}^{-1/2} - \mathbf{I}_p) \mathbf{F}_{1:t, \theta_0}^{1/2} (\widehat{\theta}_{1:t}^{\text{ML}} - \theta_0). \end{aligned}$$

By Corollary 26 and Lemma 40, we have

$$\left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \theta_0) \right\|_2 \leq 4M_n p_*^{1/2},$$

and

$$\begin{aligned} \left\| \mathbf{F}_{1:t,\theta_0}^{-1/2} \bar{\mathbf{F}}_{1:t} \mathbf{F}_{1:t,\theta_0}^{-1/2} - \mathbf{I}_p \right\|_2 &\leq (K_{\max} K_{\min}^{-3/2} (nt)^{-1/2}) (4M_n p_*^{1/2}) \\ &= c_1 M_n p_*^{1/2} (nt)^{-1/2}, \end{aligned}$$

where $c_1 = 4K_{\max} K_{\min}^{-3/2}$. Consequently, we have

$$\begin{aligned} &\left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \theta_0) - \mathbf{F}_{1:t,\theta_0}^{-1/2} \nabla L_{1:t}(\theta_0) \right\|_2 \\ &\leq \left\| \mathbf{F}_{1:t,\theta_0}^{-1/2} \bar{\mathbf{F}}_{1:t} \mathbf{F}_{1:t,\theta_0}^{-1/2} - \mathbf{I}_p \right\|_2 \left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \theta_0) \right\|_2 \leq 4c_1 M_n^2 p_* (nt)^{-1/2}. \end{aligned}$$

Note that

$$\begin{aligned} &\left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \hat{\theta}_t) \right\|_2 \leq \left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \hat{\theta}_{1:t}) \right\|_2 + \left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\hat{\theta}_{1:t} - \hat{\theta}_t) \right\|_2 \\ &\leq \left\| \mathbf{F}_{1:t,\theta_0}^{1/2} \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{-1/2} \right\|_2 \left\| \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \hat{\theta}_{1:t}) \right\|_2 + \left\| \mathbf{F}_{1:t,\theta_0}^{1/2} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \right\|_2 \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} (\hat{\theta}_{1:t} - \hat{\theta}_t) \right\|_2 \\ &\stackrel{\text{(A2)}}{\leq} \left\| \mathbf{F}_{1:t,\theta_0}^{1/2} \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{-1/2} \right\|_2 \left\| \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \hat{\theta}_{1:t}) \right\|_2 + K_{\max}^{1/2} K_{\text{low}}^{-1/2} \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} (\hat{\theta}_{1:t} - \hat{\theta}_t) \right\|_2 \\ &\stackrel{\text{Corollary 26}}{\leq} \left(M_n \sqrt{\frac{p_*^2}{nt}} + M_n^2 \sqrt{\frac{p_*^3}{n}} \right) \leq 2c_2 M_n^2 \sqrt{\frac{p_*^3}{n}} \end{aligned}$$

for some constant $c_2 = c_2(K_{\min}, K_{\max}, K_{\text{low}}, K_{\text{up}}) > 0$. Also, we have

$$\begin{aligned} &\left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \mu_t) \right\|_2 \leq \left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \hat{\theta}_t) \right\|_2 + \left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\hat{\theta}_t - \mu_t) \right\|_2 \\ &\leq \left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \hat{\theta}_t) \right\|_2 + \left\| \mathbf{F}_{1:t,\theta_0}^{1/2} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \right\|_2 \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} (\hat{\theta}_t - \mu_t) \right\|_2 \\ &\stackrel{(195)}{\leq} 2c_2 M_n^2 \sqrt{\frac{p_*^3}{n}} + K_{\max}^{1/2} K_{\text{low}}^{-1/2} K_{\text{up}} t^{-1} n^{-1/2} p_* \leq c_3 M_n^2 \sqrt{\frac{p_*^3}{n}} \end{aligned}$$

for some constant $c_3 = c_3(K_{\min}, K_{\max}, K_{\text{low}}, K_{\text{up}}) > 0$. Therefore, we have

$$\begin{aligned} &\left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\hat{\theta}_t - \theta_0) - \mathbf{F}_{1:t,\theta_0}^{-1/2} \nabla L_{1:t}(\theta_0) \right\|_2 \\ &\leq \left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \theta_0) - \mathbf{F}_{1:t,\theta_0}^{-1/2} \nabla L_{1:t}(\theta_0) \right\|_2 + \left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \hat{\theta}_t) \right\|_2 \\ &\leq 4c_1 M_n^2 p_* (nt)^{-1/2} + 2c_2 M_n^2 \sqrt{\frac{p_*^3}{n}} \leq (4c_1 + 2c_2) M_n^2 \sqrt{\frac{p_*^3}{n}}. \end{aligned}$$

Similarly,

$$\begin{aligned} &\left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\mu_t - \theta_0) - \mathbf{F}_{1:t,\theta_0}^{-1/2} \nabla L_{1:t}(\theta_0) \right\|_2 \\ &\leq \left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \theta_0) - \mathbf{F}_{1:t,\theta_0}^{-1/2} \nabla L_{1:t}(\theta_0) \right\|_2 + \left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \mu_t) \right\|_2 \\ &\leq 4c_1 M_n^2 p_* (nt)^{-1/2} + c_3 M_n^2 \sqrt{\frac{p_*^3}{n}} \leq (4c_1 + c_3) M_n^2 \sqrt{\frac{p_*^3}{n}}. \end{aligned}$$

This completes the proof. ■

Corollary 28 *Suppose that (A0), (A1*), (A2), (S) and (P*) hold. Then, on $\mathcal{E}_{\text{est},1} \cap \mathcal{E}_{\text{est},2}$, the following inequalities holds uniformly for all $t \in [T]$:*

$$\begin{aligned} \lambda_{\min}(\boldsymbol{\Omega}_t) \wedge \lambda_{\min}(\mathbf{F}_{1:t,\theta_0}) &\geq K_1 nt, \\ \lambda_{\max}(\tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}}) \wedge \lambda_{\max}(\mathbf{F}_{1:t,\theta_0}) &\leq K_2 nt, \\ \left\| \mathbf{F}_{1:t,\theta_0}^{-1/2} \tilde{\mathbf{F}}_{1:t,\hat{\theta}_{1:t}} \mathbf{F}_{1:t,\theta_0}^{-1/2} - \mathbf{I}_p \right\|_{\text{F}} &\leq K_2 M_n \left(\frac{p_*^2}{nt} \right)^{1/2}, \\ \left\| \mathbf{F}_{1:t,\theta_0}^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \mu_t) \right\|_2 &\leq K_2 M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2}, \end{aligned}$$

where K_1 and K_2 are positive constants depending only on (K_{\min}, K_{\max}) .

Proof The first two assertions directly follow from Proposition 23 and assumption (A2). For the proof of the third assertion, see **Step 2** in Theorem 8. The last assertion follows from the proof of Corollary 27. ■

Corollary 29 *Suppose that (A0), (A1*), (A2), (S) and (P*) hold. Then, on $\mathcal{E}_{\text{est},1} \cap \mathcal{E}_{\text{est},2}$, the following inequalities holds uniformly for all $t \in [T]$:*

$$\begin{aligned} nt \left\| \boldsymbol{\Omega}_t^{-1} - \mathbf{F}_{1:t,\theta_0}^{-1} \right\|_{\text{F}} &\leq K M_n \left(\frac{p_*^2}{n} \right)^{1/2}, \\ \left\| \mathbf{F}_{1:t,\theta_0}^{-1/2} \boldsymbol{\Omega}_t \mathbf{F}_{1:t,\theta_0}^{-1/2} - \mathbf{I}_p \right\|_{\text{F}} &\leq K M_n \left(\frac{p_*^2}{n} \right)^{1/2}, \\ \left\| \boldsymbol{\Omega}_t^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \mu_t) \right\|_2 &\leq K M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2} \end{aligned}$$

where $K = K(K_{\min}, K_{\max})$.

Proof In this proof, we work on the event $\mathcal{E}_{\text{est},1} \cap \mathcal{E}_{\text{est},2}$ without explicitly referring to it. Let $t \in [T]$ and $N_t = nt$ in this proof. Note that

$$\begin{aligned} \left\| \boldsymbol{\Omega}_t^{-1} - \mathbf{F}_{1:t,\theta_0}^{-1} \right\|_{\text{F}} &\leq \left\| \boldsymbol{\Omega}_t^{-1} \right\|_2 \left\| \mathbf{F}_{1:t,\theta_0}^{-1/2} \boldsymbol{\Omega}_t \mathbf{F}_{1:t,\theta_0}^{-1/2} - \mathbf{I}_p \right\|_{\text{F}} \\ &\leq \left\| \boldsymbol{\Omega}_t^{-1} \right\|_2 \left\| \mathbf{F}_{1:t,\theta_0}^{-1} \right\|_2 \left\| \boldsymbol{\Omega}_t - \mathbf{F}_{1:t,\theta_0} \right\|_{\text{F}} \\ &\stackrel{\text{Corollary 28}}{\leq} K_1^{-2} N_t^{-2} \left\| \boldsymbol{\Omega}_t - \mathbf{F}_{1:t,\theta_0} \right\|_{\text{F}}, \end{aligned}$$

where K_1 is the constant specified in Corollary 28. Hence, we only need to obtain an upper bound of $\left\| \boldsymbol{\Omega}_t - \mathbf{F}_{1:t,\theta_0} \right\|_{\text{F}}$.

Note that

$$\begin{aligned} \left\| \boldsymbol{\Omega}_t - \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}} \right\|_{\mathbb{F}} &\leq \left\| \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}} \right\|_2 \left\| \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1/2} \boldsymbol{\Omega}_t \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}}^{-1/2} - \mathbf{I}_p \right\|_{\mathbb{F}} \\ &\stackrel{\text{Corollary 28}}{\leq} \stackrel{\text{Proposition 10}}{\leq} (K_2 N_t) K_3 M_n \left(\frac{p_*^2}{n} \right)^{1/2} = (K_2 K_3) N_t M_n \left(\frac{p_*^2}{n} \right)^{1/2}, \end{aligned}$$

where K_2 is the constant specified in Corollary 28, and K_3 denotes the constant K in Proposition 10. Also,

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}} - \mathbf{F}_{1:t, \theta_0} \right\|_{\mathbb{F}} &\leq \left\| \mathbf{F}_{1:t, \theta_0} \right\|_2 \left\| \mathbf{F}_{1:t, \theta_0}^{-1/2} \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}} \mathbf{F}_{1:t, \theta_0}^{-1/2} - \mathbf{I}_p \right\|_{\mathbb{F}} \\ &\stackrel{\text{Corollary 28}}{\leq} (K_2 N_t) K_2 M_n \left(\frac{p_*^2}{N_t} \right)^{1/2} = K_2^2 M_n p_* N_t^{1/2}. \end{aligned}$$

Consequently, we have

$$\begin{aligned} \left\| \boldsymbol{\Omega}_t - \mathbf{F}_{1:t, \theta_0} \right\|_{\mathbb{F}} &\leq \left\| \boldsymbol{\Omega}_t - \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}} \right\|_{\mathbb{F}} + \left\| \tilde{\mathbf{F}}_{1:t, \hat{\theta}_{1:t}} - \mathbf{F}_{1:t, \theta_0} \right\|_{\mathbb{F}} \\ &\leq (K_2 K_3) N_t M_n \left(\frac{p_*^2}{n} \right)^{1/2} + K_2^2 M_n p_* N_t^{1/2} \\ &\leq (K_2 K_3 + K_2^2) N_t M_n \left(\frac{p_*^2}{n} \right)^{1/2}, \end{aligned}$$

which implies that

$$\begin{aligned} \left\| \boldsymbol{\Omega}_t^{-1} - \mathbf{F}_{1:t, \theta_0}^{-1} \right\|_{\mathbb{F}} &\leq K_1^{-2} N_t^{-2} (K_2 K_3 + K_2^2) N_t M_n \left(\frac{p_*^2}{n} \right)^{1/2} \\ &= K_1^{-2} (K_2 K_3 + K_2^2) N_t^{-1} M_n \left(\frac{p_*^2}{n} \right)^{1/2}. \end{aligned}$$

Also,

$$\begin{aligned} \left\| \mathbf{F}_{1:t, \theta_0}^{-1/2} \boldsymbol{\Omega}_t \mathbf{F}_{1:t, \theta_0}^{-1/2} - \mathbf{I}_p \right\|_{\mathbb{F}} &\leq \left\| \mathbf{F}_{1:t, \theta_0}^{-1} \right\|_2 \left\| \boldsymbol{\Omega}_t - \mathbf{F}_{1:t, \theta_0} \right\|_{\mathbb{F}} \\ &\stackrel{\text{Corollary 28}}{\leq} K_1^{-1} N_t^{-1} \left\| \boldsymbol{\Omega}_t - \mathbf{F}_{1:t, \theta_0} \right\|_{\mathbb{F}} \\ &\leq K_1^{-1} (K_2 K_3 + K_2^2) M_n \left(\frac{p_*^2}{n} \right)^{1/2}, \end{aligned}$$

which, combining with the last assertion in Corollary 28, implies that

$$\left\| \boldsymbol{\Omega}_t^{1/2} (\hat{\theta}_{1:t}^{\text{ML}} - \mu_t) \right\|_2 \leq K_4 M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2}$$

for some positive constant K_4 depending only on (K_{\min}, K_{\max}) . ■

Corollary 30 *Suppose that (A0), (A1*), (A2), (S) and (P*) hold. Then, on $\mathcal{E}_{\text{est},1} \cap \mathcal{E}_{\text{est},2}$, the following inequalities holds uniformly for all $t \in [T]$:*

$$\left| \|\boldsymbol{\Omega}_t^{1/2}(\theta_0 - \mu_t)\|_2 - \|\mathbf{F}_{1:t,\theta_0}^{1/2}(\theta_0 - \widehat{\theta}_{1:t}^{\text{ML}})\|_2 \right| \leq KM_n \left(\frac{p_*^3}{n} \right)^{1/2},$$

where $K = K(K_{\min}, K_{\max})$.

Proof In this proof, we work on the event $\mathcal{E}_{\text{est},1} \cap \mathcal{E}_{\text{est},2}$ without explicitly referring to it. Let $t \in [T]$. By Corollaries 26 and 29, we have

$$\begin{aligned} \|\mathbf{F}_{1:t,\theta_0}^{1/2}(\widehat{\theta}_{1:t}^{\text{ML}} - \theta_0)\|_2 &\leq 4M_n p_*^{1/2}, \\ \|\mathbf{F}_{1:t,\theta_0}^{-1/2} \boldsymbol{\Omega}_t \mathbf{F}_{1:t,\theta_0}^{-1/2} - \mathbf{I}_p\|_{\text{F}} &\leq K_1 M_n \left(\frac{p_*^2}{n} \right)^{1/2}, \\ \|\boldsymbol{\Omega}_t^{1/2}(\widehat{\theta}_{1:t}^{\text{ML}} - \mu_t)\|_2 &\leq K_1 M_n^2 \left(\frac{p_*^3}{n} \right)^{1/2}, \end{aligned}$$

where $K_1 = K_1(K_{\min}, K_{\max})$ denotes the constant K in Corollary 29. Let $\epsilon_{n,2} = K_1 M_n (p_*^2/n)^{1/2}$ and $\epsilon_{n,3} = K_1 M_n^2 (p_*^3/n)^{1/2}$ in this proof. Note that

$$\begin{aligned} \|\boldsymbol{\Omega}_t^{1/2}(\theta_0 - \mu_t)\|_2 &\leq \|\boldsymbol{\Omega}_t^{1/2}(\theta_0 - \widehat{\theta}_{1:t}^{\text{ML}})\|_2 + \|\boldsymbol{\Omega}_t^{1/2}(\widehat{\theta}_{1:t}^{\text{ML}} - \mu_t)\|_2 \\ &\leq (1 + \epsilon_{n,2})^{1/2} \|\mathbf{F}_{1:t,\theta_0}^{1/2}(\theta_0 - \widehat{\theta}_{1:t}^{\text{ML}})\|_2 + \|\boldsymbol{\Omega}_t^{1/2}(\widehat{\theta}_{1:t}^{\text{ML}} - \mu_t)\|_2 \\ &\leq (1 + \epsilon_{n,2}) \|\mathbf{F}_{1:t,\theta_0}^{1/2}(\theta_0 - \widehat{\theta}_{1:t}^{\text{ML}})\|_2 + \|\boldsymbol{\Omega}_t^{1/2}(\widehat{\theta}_{1:t}^{\text{ML}} - \mu_t)\|_2 \\ &\leq \|\mathbf{F}_{1:t,\theta_0}^{1/2}(\theta_0 - \widehat{\theta}_{1:t}^{\text{ML}})\|_2 + \epsilon_{n,2}(4M_n p_*^{1/2}) + \epsilon_{n,3}, \end{aligned}$$

which implies that

$$\|\boldsymbol{\Omega}_t^{1/2}(\theta_0 - \mu_t)\|_2 - \|\mathbf{F}_{1:t,\theta_0}^{1/2}(\theta_0 - \widehat{\theta}_{1:t}^{\text{ML}})\|_2 \leq 5\epsilon_{n,3}.$$

Also,

$$\begin{aligned} \|\boldsymbol{\Omega}_t^{1/2}(\theta_0 - \mu_t)\|_2 &\geq \|\boldsymbol{\Omega}_t^{1/2}(\theta_0 - \widehat{\theta}_{1:t}^{\text{ML}})\|_2 - \|\boldsymbol{\Omega}_t^{1/2}(\widehat{\theta}_{1:t}^{\text{ML}} - \mu_t)\|_2 \\ &\geq (1 - \epsilon_{n,2})^{1/2} \|\mathbf{F}_{1:t,\theta_0}^{1/2}(\theta_0 - \widehat{\theta}_{1:t}^{\text{ML}})\|_2 - \|\boldsymbol{\Omega}_t^{1/2}(\widehat{\theta}_{1:t}^{\text{ML}} - \mu_t)\|_2 \\ &\geq (1 - \epsilon_{n,2}) \|\mathbf{F}_{1:t,\theta_0}^{1/2}(\theta_0 - \widehat{\theta}_{1:t}^{\text{ML}})\|_2 - \|\boldsymbol{\Omega}_t^{1/2}(\widehat{\theta}_{1:t}^{\text{ML}} - \mu_t)\|_2 \\ &\geq \|\mathbf{F}_{1:t,\theta_0}^{1/2}(\theta_0 - \widehat{\theta}_{1:t}^{\text{ML}})\|_2 - \epsilon_{n,2}(4M_n p_*^{1/2}) - \epsilon_{n,3}, \end{aligned}$$

which implies that

$$\|\boldsymbol{\Omega}_t^{1/2}(\theta_0 - \mu_t)\|_2 - \|\mathbf{F}_{1:t,\theta_0}^{1/2}(\theta_0 - \widehat{\theta}_{1:t}^{\text{ML}})\|_2 \geq -5\epsilon_{n,3}.$$

This completes the proof. ■

Appendix G. Logistic regression with Gaussian design

In this section, we demonstrate that the main results in Section 8 hold under the logistic regression model. First, we introduce some notations needed for the theoretical verifications of this model. Let $\mathbf{Y} = (Y_i)_{i \in [N]} \in \mathbb{R}^N$ be the response vector and $\mathbf{X} = (X_{ij})_{i \in [N], j \in [p]} \in \mathbb{R}^{N \times p}$ be the design matrix. Also, for $t \in [T]$, let

$$\begin{aligned} I_t &= \{n(t-1) + 1, n(t-1) + 2, \dots, nt\}, & I_{1:t} &= \cup_{s=1}^t I_s, \\ \mathbf{Y}_t &= (Y_i)_{i \in I_t}, & \mathbf{X}_t &= (X_{ij})_{i \in I_t, j \in [p]}, & \mathbf{X}_{1:t} &= (X_{ij})_{i \in I_{1:t}, j \in [p]} \in \mathbb{R}^{N_t \times p}. \end{aligned}$$

With slight abuse of notation, we denote $\mathbf{D}_t = (\mathbf{X}_t, \mathbf{Y}_t)$ in this section.

For the logistic regression model, the likelihood function is given by

$$L_t(\theta) = \sum_{i \in I_t} \left[Y_i X_i^\top \theta - b(X_i^\top \theta) \right], \quad (196)$$

where $b(\cdot) = \log(1 + \exp(\cdot))$. Note that $b(\cdot)$ is four times differentiable with derivatives b', b'', b''' and b'''' , respectively.

In the sections in main text, we considered several regularity conditions, as assumed in **(A1)**, **(A1*)**, and **(A2)**. These conditions can be verified under the logistic regression model with a “well-posed” design \mathbf{X} . To see this, we consider a simple random matrix setup where each entry of the design matrix \mathbf{X} is an i.i.d. standard normal random variable, i.e., $X_{ij} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$. For simplicity, we take the covariance matrix to be the identity matrix \mathbf{I}_p ; this setting can be easily extended to a general covariance Σ satisfying

$$C^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C$$

for some constant $C > 0$. With slight abuse of notation, hereafter, let \mathbb{P} and \mathbb{E} denote the joint probability measure and expectation corresponding to (\mathbf{X}, \mathbf{Y}) , respectively.

Under the assumed random design setup, we can verify the conditions in **(A1)**, **(A1*)**, and **(A2)**. First, one can easily check that $L_t(\theta)$ in (196) is *stochastically linear* (with respect to the randomness in \mathbf{Y}) as follows:

$$\zeta_t(\theta) = L_t(\theta) - \mathbb{E}_t L_t(\theta) = \sum_{i \in I_t} \left[(Y_i - \mathbb{E}_t(Y_i | \mathbf{X})) X_i^\top \theta \right], \quad \forall t \in [T],$$

where $\mathbb{E}_t(Y_i | \mathbf{X}) = b'(X_i^\top \theta_0)$ for each $i \in I_t$. To verify the remaining assumptions, we impose **(EX)**. Under the assumption **(EX)**, we can prove the conditions in **(A1)**, **(A1*)**, and **(A2)** hold uniformly for all $t \in [T]$ with the following quantities:

$$\mathbf{V}_t = \mathbf{X}_t^\top \mathbf{X}_t / 4, \quad \mathbf{V}_{1:t} = \mathbf{X}_{1:t}^\top \mathbf{X}_{1:t} / 4, \quad M_n = C_1, \quad K_{\min} = C_2, \quad K_{\max} = C_3,$$

where C_1 and C_2 are positive constants depending only on K_1 , and C_3 is a universal constant. Technical statements and proofs are deferred to Appendix G.1; see Propositions 36, 38, 39 for precise statements.

We now state that the online BvM theorem holds for the logistic regression model.

Proposition 31 *Suppose that (\mathbf{EX}) holds. Then, with \mathbb{P} -probability at least $1 - 5n^{-1} - 10e^{-n/72} - 4(Np)^{-1}$, the following inequality holds uniformly for all $t \in [T]$:*

$$\begin{aligned} d_V\left(\Pi_t, \Pi(\cdot \mid \mathbf{D}_{1:t})\right) &\leq C \left(\frac{p_*^3}{n}\right)^{1/2}, \\ d_V\left(\Pi_t, \mathcal{N}\left(\hat{\theta}_{1:t}^{\text{ML}}, \mathbf{F}_{1:t, \theta_0}^{-1}\right)\right) &\leq C \left(\frac{p_*^3}{n}\right)^{1/2}, \end{aligned}$$

where $C = C(K_1)$.

G.1 Proof for Proposition 31

Throughout this subsection, we follow the notations given in Appendix G without explicitly referring to them.

Lemma 32 *Suppose that*

$$p \vee (4 \log T) \leq n. \quad (197)$$

Then,

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_{i \in I_t} X_i X_i^\top\right) \leq \frac{1}{9}n \quad \text{for some } t \in [T]\right\} \leq 2e^{-n/4} \quad (198)$$

and

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_{i \in I_t} X_i X_i^\top\right) \geq 9n \quad \text{for some } t \in [T]\right\} \leq 2e^{-n/4}. \quad (199)$$

Proof By the equation (60) in Wainwright (2009) and $p \leq n$, we have, for $t \in [T]$,

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_{i \in I_t} X_i X_i^\top\right) \leq \frac{1}{9}n\right\} \leq 2e^{-n/2}.$$

It follows that

$$\begin{aligned} &\mathbb{P}\left\{\lambda_{\min}\left(\sum_{i \in I_t} X_i X_i^\top\right) \leq \frac{1}{9}n \quad \text{for some } t \in [T]\right\} \\ &\leq T \cdot \max_{t \in [T]} \mathbb{P}\left\{\lambda_{\min}\left(\sum_{i \in I_t} X_i X_i^\top\right) \leq \frac{1}{9}n\right\} \leq 2e^{-n/2 + \log T} \stackrel{(197)}{\leq} 2e^{-n/4}, \end{aligned}$$

completing the proof of (198).

The proof of (199) is similar. By the equation (59) in Wainwright (2009) and $p \leq n$, we have, for $t \in [T]$,

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_{i \in I_t} X_i X_i^\top\right) \geq 9n\right\} \leq 2e^{-n/2}.$$

It follows that

$$\begin{aligned} & \mathbb{P} \left\{ \lambda_{\max} \left(\sum_{i \in I_t} X_i X_i^\top \right) \geq 9n \text{ for some } t \in [T] \right\} \\ & \leq T \cdot \max_{t \in [T]} \mathbb{P} \left\{ \lambda_{\max} \left(\sum_{i \in I_t} X_i X_i^\top \right) \geq 9n \right\} \leq 2e^{-n/2 + \log T} \stackrel{(197)}{\leq} 2e^{-n/4}, \end{aligned}$$

which completes the proof of (199). ■

Lemma 33 *We have*

$$\mathbb{P} \left\{ \max_{i \in [N], j \in [p]} |X_{i,j}| > 2\sqrt{\log(Np)} \right\} \leq 2(Np)^{-1} \quad (200)$$

and

$$\mathbb{P} \left\{ \max_{i \in [N]} \|X_i\|_2^2 > 4p \log(Np) \right\} \leq 2(Np)^{-1}, \quad (201)$$

where $X_i = (X_{ij})_{j \in [p]} \in \mathbb{R}^p$.

Proof Since $X_{ij} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$, we have, for all $\omega \geq 0$, $i \in [N]$ and $j \in [p]$,

$$\mathbb{P} \left(|X_{ij}| > \omega \right) \leq 2 \exp \left(-\frac{\omega^2}{2} \right).$$

It follows that

$$\mathbb{P} \left(\max_{i \in [N], j \in [p]} |X_{ij}| > \omega \right) \leq 2Np \exp \left(-\frac{\omega^2}{2} \right).$$

By taking $\omega = 2\sqrt{\log(Np)}$, we complete the proof of (200).

Also, on the same event where the following inequality holds:

$$\max_{i \in [N], j \in [p]} |X_{ij}| \leq 2\sqrt{\log(Np)},$$

we have

$$\max_{i \in [N]} \|X_i\|_2^2 \leq p \max_{i \in [N], j \in [p]} |X_{ij}| \leq p \left(2\sqrt{\log(Np)} \right)^2.$$

This completes the proof of (201). ■

Lemma 34 *Let $b(\cdot) = \log(1 + \exp(\cdot))$. Then,*

$$\frac{b''(\eta_1)}{b''(\eta_2)} \leq e^{3|\eta_1 - \eta_2|}, \quad \forall \eta_1, \eta_2 \in \mathbb{R}.$$

Proof Let $\eta_1, \eta_2 \in \mathbb{R}$. Since $b''(\eta) = e^\eta / (1 + e^\eta)^2$ for $\eta \in \mathbb{R}$, note that

$$\frac{b''(\eta_1)}{b''(\eta_2)} = e^{\eta_1 - \eta_2} \left(\frac{1 + e^{\eta_2}}{1 + e^{\eta_1}} \right)^2.$$

Also,

$$\frac{1 + e^{\eta_2}}{1 + e^{\eta_1}} = 1 + \frac{e^{\eta_2} - e^{\eta_1}}{1 + e^{\eta_1}} = 1 + \frac{e^{\eta_1} (e^{\eta_2 - \eta_1} - 1)}{1 + e^{\eta_1}} \leq 1 + e^{|\eta_2 - \eta_1|} - 1 = e^{|\eta_1 - \eta_2|}.$$

It follows that

$$\frac{b''(\eta_1)}{b''(\eta_2)} \leq e^{\eta_1 - \eta_2} \times e^{2|\eta_1 - \eta_2|} \leq e^{3|\eta_1 - \eta_2|},$$

which completes the proof. ■

Lemma 35 For $\tau > 0$, suppose that

$$n \geq C \left(\log T \vee [p \log(\tau^2 p \log N)] \right) \quad (202)$$

for a large enough universal constant $C > 0$. Then,

$$\frac{n}{1080e^{2(\tau+1)}} \leq \min_{t \in T} \inf_{\theta \in \Theta(\mathbf{I}_p, \tau)} \lambda_{\min}(\mathbf{F}_{t, \theta}) \leq \max_{t \in T} \sup_{\theta \in \Theta} \lambda_{\max}(\mathbf{F}_{t, \theta}) \leq \frac{9}{4}n \quad (203)$$

with \mathbb{P} -probability at least $1 - 6e^{-n/72} - 2(Np)^{-1}$.

Proof For $t \in [T]$ and $\theta \in \Theta$, note that

$$\mathbf{F}_{t, \theta} = \sum_{i \in I_t} \left[b''(X_i^\top \theta) X_i X_i^\top \right].$$

For $\tau > 0$ and $\epsilon \in (0, 1)$, let $\widehat{\Theta}_{\epsilon, \tau}$ be the ϵ -cover of $\Theta(\mathbf{I}_p, \tau)$. One can choose $\widehat{\Theta}_{\epsilon, \tau}$ so that $|\widehat{\Theta}_{\epsilon, \tau}| \leq (3\tau/\epsilon)^p$; see Proposition 1.3 of Section 15 in Lorentz et al. (1996). Let $\theta \in \Theta(\mathbf{I}_p, \tau)$. By the definition of $\widehat{\Theta}_{\epsilon, \tau}$, there exists $\widehat{\theta}(\theta) \in \widehat{\Theta}_{\epsilon, \tau}$ such that $\|\theta - \widehat{\theta}\|_2 \leq \epsilon$. For $\omega \geq 0$, let

$$\mathcal{I}_\omega(\widehat{\theta}, t) = \mathcal{I}_\omega(\widehat{\theta}, t, \tau) = \left\{ i \in I_t : |X_i^\top \widehat{\theta}| \leq \omega(\tau + 1) \right\}.$$

Note that

$$\begin{aligned} \lambda_{\min}(\mathbf{F}_{t, \theta}) &= \lambda_{\min} \left(\sum_{i \in I_t} b''(X_i^\top \theta) X_i X_i^\top \right) = \lambda_{\min} \left(\sum_{i \in I_t} \frac{b''(X_i^\top \theta)}{b''(X_i^\top \widehat{\theta})} b''(X_i^\top \widehat{\theta}) X_i X_i^\top \right) \\ &\geq \left[\min_{i \in [N]} \frac{b''(X_i^\top \theta)}{b''(X_i^\top \widehat{\theta})} \right] \lambda_{\min} \left(\sum_{i \in \mathcal{I}_\omega(\widehat{\theta}, t)} b''(X_i^\top \widehat{\theta}) X_i X_i^\top \right) \\ &\stackrel{\text{Lemma 34}}{\geq} \exp \left(-3 \|\theta - \widehat{\theta}\|_2 \max_{i \in [N]} \|X_i\|_2 \right) \lambda_{\min} \left(\sum_{i \in \mathcal{I}_\omega(\widehat{\theta}, t)} b''(X_i^\top \widehat{\theta}) X_i X_i^\top \right) \\ &\geq \exp \left(-3\epsilon \cdot \max_{i \in [N]} \|X_i\|_2 \right) b''(\omega(\tau + 1)) \lambda_{\min} \left(\sum_{i \in \mathcal{I}_\omega(\widehat{\theta}, t)} X_i X_i^\top \right) \end{aligned} \quad (204)$$

where the last inequality holds by the symmetry and monotonicity of $b''(\cdot)$ in the logistic regression model.

First, for $\hat{\theta} \in \hat{\Theta}_{\epsilon, \tau}$ and $t \in [T]$, we will prove that $|\mathcal{I}_2(\hat{\theta}, t)| \geq n/6$ with high probability. Since $X_i^\top \hat{\theta} \sim \mathcal{N}(0, \|\hat{\theta}\|_2^2)$ and

$$\|\hat{\theta}\|_2 \leq \|\theta\|_2 + \|\theta - \hat{\theta}\|_2 \leq \tau + \epsilon \leq \tau + 1,$$

we have, for $i \in I_t$,

$$\mathbb{P}\left(|X_i^\top \hat{\theta}| > \omega'(\tau + 1)\right) \leq \mathbb{P}\left(|X_i^\top \hat{\theta}| > \omega' \|\hat{\theta}\|_2\right) \leq 2e^{-(\omega')^2/2}, \quad \forall \omega' \geq 0.$$

By taking $\omega' = 2$, we have

$$\mathbb{P}\left(|X_i^\top \hat{\theta}| \leq 2(\tau + 1)\right) \geq 1 - 2e^{-2} \geq \frac{1}{3}.$$

We will utilize the Chernoff-type left tail inequality (see Section 2.3 in Vershynin (2018)). Let $S_n = \sum_{i=1}^n Z_i$, where $Z_i \stackrel{i.i.d.}{\sim} \text{Bernoulli}(\eta)$ for some $\eta \in (0, 1)$. Then, for any $\delta \in (0, 1)$,

$$\mathbb{P}\left\{S_n \leq (1 - \delta)\eta n\right\} \leq \exp\left(-\frac{\delta^2}{3}\eta n\right).$$

By taking $\delta = 1/2$ and $\eta = 1/3$ in the above display, we have, for $\hat{\theta} \in \hat{\Theta}_{\epsilon, \tau}$ and $t \in [T]$,

$$\mathbb{P}\left(|\mathcal{I}_2(\hat{\theta}, t)| \leq \frac{n}{6}\right) \leq e^{-n/36}.$$

By taking $\epsilon = (4\sqrt{p \log(Np)})^{-1}$, it follows that

$$\begin{aligned} \mathbb{P}\left(\min_{\hat{\theta} \in \hat{\Theta}_{\epsilon, \tau}} \min_{t \in [T]} |\mathcal{I}_2(\hat{\theta}, t)| \leq \frac{n}{6}\right) &\leq (3\tau/\epsilon)^p \cdot T \cdot e^{-n/36} \\ &= \exp\left(\frac{p}{2} \log(144\tau^2 p \log(Np)) + \log T - \frac{n}{36}\right) \stackrel{(202)}{\leq} e^{-n/72}. \end{aligned} \tag{205}$$

Let

$$\begin{aligned} \Omega_{n,1} &= \left\{|\mathcal{I}_2(\hat{\theta}, t)| \geq \frac{1}{6}n \text{ for all } t \in [T] \text{ and } \hat{\theta} \in \hat{\Theta}_{\epsilon, \tau}\right\}, \\ \Omega_{n,2} &= \left\{\lambda_{\min}\left(\sum_{i \in \mathcal{I}_2(\hat{\theta}, t)} X_i X_i^\top\right) \geq \frac{1}{9}|\mathcal{I}_2(\hat{\theta}, t)| \text{ for all } t \in [T] \text{ and } \hat{\theta} \in \hat{\Theta}_{\epsilon, \tau}\right\}, \\ \Omega_{n,3} &= \left\{\max_{i \in [N]} \|X_i\|_2 \leq 2\sqrt{p \log(Np)}\right\}. \end{aligned}$$

Since $p \vee 4 \log T \leq n/6 \leq |\mathcal{I}_2(\hat{\theta}, t)|$ for all $t \in [T]$ on $\Omega_{n,1}$, we can apply the results of Lemma 32 on $\Omega_{n,1}$. By the equation (205), Lemmas 32 and 33,

$$\begin{aligned} \mathbb{P}\{\Omega_{n,1}^c\} &\leq e^{-n/72}, \\ \mathbb{P}\{\Omega_{n,3}^c\} &\leq 2(Np)^{-1} \\ \mathbb{P}\{\Omega_{n,2}^c \mid \Omega_{n,1}\} &\leq (3\tau/\epsilon)^p \times 2e^{-(n/6)/4} = 2 \exp\left(-n/24 + \frac{p}{2} \log(144\tau^2 p \log(Np))\right) \\ &\stackrel{(202)}{\leq} 2e^{-n/48}. \end{aligned}$$

By $1 - x \geq e^{-2x}$ and $e^{-y} \geq 1 - y$ for $x \in [0, 0.795]$ and $y \in \mathbb{R}$, note that

$$\begin{aligned} \mathbb{P}\{\Omega_n\} &\geq 1 - \mathbb{P}\{\Omega_{n,1}^c\} - \mathbb{P}\{\Omega_{n,2}^c\} - \mathbb{P}\{\Omega_{n,3}^c\} \\ &\geq 1 - 2\mathbb{P}\{\Omega_{n,1}^c\} - \mathbb{P}\{\Omega_{n,2}^c \mid \Omega_{n,1}\} - \mathbb{P}\{\Omega_{n,3}^c\} \\ &\geq 1 - 2e^{-n/72} - 2e^{-n/48} - 2(Np)^{-1} \\ &\geq 1 - 4e^{-n/72} - 2(Np)^{-1}, \end{aligned}$$

where $\Omega_n = \Omega_{n,1} \cap \Omega_{n,2} \cap \Omega_{n,3}$. On Ω_n , therefore, we have

$$\begin{aligned} \min_{\theta \in \Theta_{n,\tau}} \min_{t \in [T]} \lambda_{\min}(\mathbf{F}_{t,\theta}) &\stackrel{(204)}{\geq} \exp\left(-3\epsilon \cdot \max_{i \in [N]} \|X_i\|_2\right) b''(2(\tau+1)) \left(\frac{1}{9} \times \frac{n}{6}\right) \\ &\geq e^{-3/2} \times \frac{\exp(2(\tau+1))}{[1 + \exp(2(\tau+1))]^2} \times \frac{n}{54} \\ &\geq \frac{n}{1080e^{2(\tau+1)}}, \end{aligned}$$

where the third inequality holds by $e^{-3/2} \geq 1/5$ and $e^x/(1+e^x)^2 \geq 1/(4e^x)$ for $x \geq 0$.

The proof of the third inequality in (203) is simple. Since $b''(\cdot) \leq b''(0) = 1/4$, with \mathbb{P} -probability at least $1 - 2e^{-n/4}$,

$$\lambda_{\max}(\mathbf{F}_{t,\theta}) = \lambda_{\max}\left(\sum_{i=1}^n \left[b''\left(X_i^\top \theta\right) X_i X_i^\top\right]\right) \leq \frac{1}{4} \lambda_{\max}\left(\sum_{i=1}^n X_i X_i^\top\right) \leq \frac{9}{4}n,$$

where the second inequality holds by Lemma 32. This completes the proof of (203). \blacksquare

The following proposition verifies (27), (34) and (42).

Proposition 36 *Suppose that (EX) holds. Then, the following inequalities hold with \mathbb{P} -probability at least $1 - 4e^{-n/72} - 2(Np)^{-1}$:*

$$\tilde{K}_{\min} n \leq \min_{t \in [T]} \inf_{\theta \in \Theta(\theta_0, \mathbf{I}_p, 1/2)} \lambda_{\min}(\mathbf{F}_{t,\theta}) \leq \max_{t \in [T]} \sup_{\theta \in \Theta} \lambda_{\max}(\mathbf{F}_{t,\theta}) \leq \tilde{K}_{\max} n, \quad (206)$$

$$\max_{t \in [T]} \sup_{\theta \in \Theta(\theta_0, \mathbf{I}_p, 1/2)} \left(\left\| \mathbf{F}_{t,\theta}^{-1} \left(\frac{\mathbf{X}_t^\top \mathbf{X}_t}{4} \right) \right\|_2 \vee \left\| \mathbf{F}_{1:t,\theta}^{-1} \left(\frac{\mathbf{X}_{1:t}^\top \mathbf{X}_{1:t}}{4} \right) \right\|_2 \right) \leq \frac{M^2}{9}, \quad (207)$$

where \tilde{K}_{\min} and M are constants depending only on K_1 , and \tilde{K}_{\max} is a universal constant.

Proof Since $\|\theta_0\|_2 \leq K_1$, we have

$$\Theta(\theta_0, \mathbf{I}_p, 1/2) \subseteq \Theta(\mathbf{I}_p, K_1 + 1/2).$$

By Lemma 35 with $\tau = K_1 + 1/2$, (206) holds with the constants

$$\tilde{K}_{\min} = \frac{1}{1080e^{2K_1+3}}, \quad \tilde{K}_{\max} = \frac{9}{4},$$

with \mathbb{P} -probability at least $1 - 8e^{-n/72} - 4(Np)^{-1}$.

By the last display and Lemma 32, on the same event on which (206) holds, we have

$$\begin{aligned} & \max_{t \in [T]} \sup_{\theta \in \Theta(\theta_0, \mathbf{I}_p, 1/2)} \left\| \mathbf{F}_{t,\theta}^{-1} \left(\mathbf{X}_t^\top \mathbf{X}_t / 4 \right) \right\|_2 \\ & \leq \left[\min_{t \in [T]} \inf_{\theta \in \Theta(\theta_0, \mathbf{I}_p, 1/2)} \lambda_{\min}(\mathbf{F}_{t,\theta}) \right]^{-1} \left[\max_{t \in [T]} \left(\mathbf{X}_t^\top \mathbf{X}_t / 4 \right) \right] \\ & \leq (1080e^{2K_1+3}) \left(\frac{9}{4} \right) = 2430e^{2K_1+3}. \end{aligned}$$

Also,

$$\begin{aligned} & \max_{t \in [T]} \sup_{\theta \in \Theta(\theta_0, \mathbf{I}_p, 1/2)} \left\| \mathbf{F}_{1:t,\theta}^{-1} \left(\mathbf{X}_{1:t}^\top \mathbf{X}_{1:t} / 4 \right) \right\|_2 \\ & \leq \left[\min_{t \in [T]} \inf_{\theta \in \Theta(\theta_0, \mathbf{I}_p, 1/2)} \lambda_{\min}(\mathbf{F}_{1:t,\theta}) \right]^{-1} \left[\max_{t \in [T]} \left(\mathbf{X}_{1:t}^\top \mathbf{X}_{1:t} / 4 \right) \right] \\ & \leq (1080e^{2K_1+3}) \left(\frac{9}{4} \right) = 2430e^{2K_1+3} = \frac{(27\sqrt{30}e^{K_1+3/2})^2}{9}. \end{aligned}$$

The last two displays complete the proof of (207) by taking $M = 27\sqrt{30}e^{K_1+3/2}$. ■

Lemma 37 *For any $\omega \geq 0$ and $t \in [T]$, we have*

$$\mathbb{P}_{0,t} \left(\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \zeta_t \right\|_2 \geq \sqrt{\text{tr} \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \frac{\mathbf{X}_t^\top \mathbf{X}_t}{4} \right)} + \sqrt{2 \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \frac{\mathbf{X}_t^\top \mathbf{X}_t}{4} \right\|_2 \omega} \mid \mathbf{X} \right) \leq e^{-\omega},$$

Proof Let $t \in [T]$ and $\mathcal{E}_t = (\epsilon_i)_{i \in I_t} \in \mathbb{R}^n$, where $\epsilon_i = Y_i - \mathbb{E}_t(Y_i \mid \mathbf{X})$. Note that

$$\nabla \zeta_t = \sum_{i \in I_t} \epsilon_i x_i = \mathbf{X}_t^\top \mathcal{E}_t, \quad \sup_{\eta \in \mathbb{R}} b''(\eta) \leq b''(0) = 1/4.$$

By Lemma 6.1 in Rigollet (2012), we have, for any $u = (u_i)_{i \in [n]} \in \mathbb{R}^n$ with $\|u\|_2 = 1$,

$$u^\top \mathcal{E}_t = \sum_{i=1}^n u_i \epsilon_{n(t-1)+i} \sim \text{subG}(1/4), \quad \forall t \in [T],$$

which is equivalent to $\mathcal{E}_t \sim \text{subG}(1/4)$ for all $t \in [T]$. By the last display and $\mathbb{E}_t(\mathcal{E}_t \mid \mathbf{X}) = 0$ for all $t \in [T]$, we can apply Hanson-Wright inequality. By Lemma 50, we have, for any $\omega \geq 0$,

$$\mathbb{P}_{0,t} \left(\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \mathbf{X}_t^\top \mathcal{E}_t \right\|_2 \geq \frac{1}{4} \left[\text{tr}(\mathbf{B}_t) + 2\sqrt{\text{tr}(\mathbf{B}_t^2)\omega} + 2\|\mathbf{B}_t\|_2\omega \right] \mid \mathbf{X} \right) \leq e^{-\omega},$$

where $\mathbf{B}_t = \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} (\mathbf{X}_t^\top \mathbf{X}_t) \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2}$. Since

$$\begin{aligned} \text{tr}(\mathbf{B}_t) + 2\sqrt{\text{tr}(\mathbf{B}_t^2)\omega} + 2\|\mathbf{B}_t\|_2\omega &= \text{tr}(\mathbf{B}_t) + 2\|\mathbf{B}_t\|_F\omega^{1/2} + 2\|\mathbf{B}_t\|_2\omega \\ &\leq \left[\sqrt{\text{tr}(\mathbf{B}_t)} + \sqrt{2\omega\|\mathbf{B}_t\|_2} \right]^2 \end{aligned}$$

for any $\omega \geq 0$, we have

$$\mathbb{P}_{0,t} \left(\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \mathbf{X}_t^\top \mathcal{E}_t \right\|_2 \geq \frac{1}{2} \left[\sqrt{\text{tr}(\mathbf{B}_t)} + \sqrt{2\omega\|\mathbf{B}_t\|_2} \right] \mid \mathbf{X} \right) \leq e^{-\omega}.$$

The above display is equivalent to

$$\mathbb{P}_{0,t} \left(\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \mathbf{X}_t^\top \mathcal{E}_t \right\|_2 \geq \sqrt{\text{tr}(\tilde{\mathbf{B}}_t)} + \sqrt{2\omega\|\tilde{\mathbf{B}}_t\|_2} \mid \mathbf{X} \right) \leq e^{-\omega},$$

where $\tilde{\mathbf{B}}_t = \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} (\mathbf{X}_t^\top \mathbf{X}_t / 4)$. This completes the proof. \blacksquare

In the following Proposition 38, we demonstrate that (26) and (41) are satisfied with the specified matrices $\mathbf{V}_t = \mathbf{X}_t^\top \mathbf{X}_t / 4$ and $\mathbf{V}_{1:t} = \mathbf{X}_{1:t}^\top \mathbf{X}_{1:t} / 4$, respectively.

Proposition 38 *Suppose that $(\mathbf{E}\mathbf{X})$ holds. Then, with \mathbb{P} -probability at least $1 - 3n^{-1} - 6e^{-n/72} - 2(Np)^{-1}$, the following inequalities hold uniformly for all $t \in [T]$:*

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \zeta_t \right\|_2 &\leq r(\tilde{\mathbf{F}}_{t,\theta_t^*}, \mathbf{X}_t^\top \mathbf{X}_t / 4, \log n + \log T), \\ \left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{-1/2} \nabla \zeta_{1:t} \right\|_2 &\leq r(\tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}, \mathbf{X}_{1:t}^\top \mathbf{X}_{1:t} / 4, \log n + \log T), \\ \left\| \mathbf{F}_{1:t,\theta_0}^{-1/2} \nabla \zeta_{1:t} \right\|_2 &\leq r(\mathbf{F}_{1:t,\theta_0}, \mathbf{X}_{1:t}^\top \mathbf{X}_{1:t} / 4, \log n + \log T), \end{aligned}$$

where

$$r(\mathbf{F}, \mathbf{V}, \omega) = \sqrt{\text{tr}(\mathbf{F}^{-1}\mathbf{V})} + \sqrt{2\omega\|\mathbf{F}^{-1}\mathbf{V}\|_2}, \quad \mathbf{F}, \mathbf{V} \in \mathcal{S}_{++}^p, \quad \omega \in \mathbb{R}_+.$$

Proof The proof of the first assertion directly follows from Lemma 37. Note that

$$\begin{aligned} &\mathbb{P} \left(\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \zeta_t \right\|_2 \geq \sqrt{\text{tr} \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \frac{\mathbf{X}_t^\top \mathbf{X}_t}{4} \right)} + \sqrt{2 \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \frac{\mathbf{X}_t^\top \mathbf{X}_t}{4} \right\|_2} (\log n + \log T) \text{ for some } t \in [T] \mid \mathbf{X} \right) \\ &\leq T \cdot \max_{t \in [T]} \mathbb{P}_t \left(\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \zeta_t \right\|_2 \geq \sqrt{\text{tr} \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \frac{\mathbf{X}_t^\top \mathbf{X}_t}{4} \right)} + \sqrt{2 \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \frac{\mathbf{X}_t^\top \mathbf{X}_t}{4} \right\|_2} (\log n + \log T) \mid \mathbf{X} \right) \\ &\leq T \cdot e^{-\log n - \log T} = n^{-1}. \end{aligned}$$

By integrating over the values of \mathbf{X} , we have

$$\begin{aligned} & \mathbb{P} \left(\left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \nabla \zeta_t \right\|_2 \geq \sqrt{\text{tr} \left(\tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \frac{\mathbf{X}_t^\top \mathbf{X}_t}{4} \right)} + \sqrt{2 \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1} \frac{\mathbf{X}_t^\top \mathbf{X}_t}{4} \right\|_2} (\log n + \log T) \text{ for some } t \in [T] \right) \\ & \leq n^{-1}, \end{aligned}$$

which completes the proof of the first assertion.

The proof of the second assertion is similar to that of Lemma 37. Hence, we will provide the sketch of the proof. Let $t \in [T]$ and $\mathcal{E}_{1:t} = (\epsilon_i)_{i \in I_{1:t}} \in \mathbb{R}^{nt}$. Note that $\nabla \zeta_{1:t} = \mathbf{X}_{1:t}^\top \mathcal{E}_{1:t}$. By Lemma 50, we have, for any $\omega \geq 0$,

$$\mathbb{P} \left(\left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{-1/2} \mathbf{X}_{1:t}^\top \mathcal{E}_{1:t} \right\|_2 \geq \frac{1}{4} \left[\text{tr}(\tilde{\mathbf{B}}_{1:t}) + 2\sqrt{\text{tr}(\tilde{\mathbf{B}}_{1:t}^2)\omega} + \|\tilde{\mathbf{B}}_{1:t}\|_2 \omega \right] \mid \mathbf{X} \right) \leq e^{-\omega},$$

where $\tilde{\mathbf{B}}_{1:t} = \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{-1/2} (\mathbf{X}_{1:t}^\top \mathbf{X}_{1:t}) \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{-1/2}$. Since

$$\text{tr}(\tilde{\mathbf{B}}_{1:t}) + 2\sqrt{\text{tr}(\tilde{\mathbf{B}}_{1:t}^2)\omega} + \|\tilde{\mathbf{B}}_{1:t}\|_2 \omega \leq \left[\sqrt{\text{tr}(\tilde{\mathbf{B}}_{1:t})} + \sqrt{2\omega \|\tilde{\mathbf{B}}_{1:t}\|_2} \right]^2,$$

for any $\omega \geq 0$, we have

$$\mathbb{P} \left(\left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{-1/2} \mathbf{X}_{1:t}^\top \mathcal{E}_{1:t} \right\|_2 \geq \frac{1}{2} \left[\sqrt{\text{tr}(\tilde{\mathbf{B}}_{1:t})} + \sqrt{2\omega \|\tilde{\mathbf{B}}_{1:t}\|_2} \right] \mid \mathbf{X} \right) \leq e^{-\omega}.$$

It follows that

$$\begin{aligned} & \mathbb{P} \left(\left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{-1/2} \mathbf{X}_{1:t}^\top \mathcal{E}_{1:t} \right\|_2 \geq \frac{1}{2} \left[\sqrt{\text{tr}(\tilde{\mathbf{B}}_{1:t})} + \sqrt{2\|\tilde{\mathbf{B}}_{1:t}\|_2} (\log n + \log T) \right] \text{ for some } t \in [T] \mid \mathbf{X} \right) \\ & \leq T \cdot e^{-\log n - \log T} = n^{-1}. \end{aligned}$$

By integrating over the values of \mathbf{X} , we have

$$\begin{aligned} & \mathbb{P} \left(\left\| \tilde{\mathbf{F}}_{1:t,\theta_{1:t}^*}^{-1/2} \mathbf{X}_{1:t}^\top \mathcal{E}_{1:t} \right\|_2 \geq \frac{1}{2} \left[\sqrt{\text{tr}(\tilde{\mathbf{B}}_{1:t})} + \sqrt{2\|\tilde{\mathbf{B}}_{1:t}\|_2} (\log n + \log T) \right] \text{ for some } t \in [T] \right) \\ & \leq n^{-1}, \end{aligned}$$

which completes the proof of the second assertion.

Next, we will prove the third assertion, which is similar to the second assertion. Let $\tilde{\mathcal{E}}$ be an event where $\mathbf{F}_{1:t,\theta_0}$ is nonsingular. By Lemma 37, there exists an event \mathcal{E} such that on \mathcal{E} , $\mathbf{F}_{1:t,\theta_0}$ is nonsingular and

$$\mathbb{P}(\mathcal{E}) \geq 1 - 6e^{-n/72} - 2(Np)^{-1}.$$

Hence, we have

$$\mathbb{P}(\tilde{\mathcal{E}}) \geq \mathbb{P}(\mathcal{E}) \geq 1 - 6e^{-n/72} - 2(Np)^{-1}.$$

In this proof, we denote

$$R_{n,T}(\mathbf{B}_{1:t}) = \frac{1}{2} \left(\sqrt{\text{tr}(\mathbf{B}_{1:t})} + \sqrt{2 \|\mathbf{B}_{1:t}\|_2 (\log n + \log T)} \right),$$

where $\mathbf{B}_{1:t} = \mathbf{F}_{1:t,\theta_0}^{-1/2} (\mathbf{X}_{1:t}^\top \mathbf{X}_{1:t}) \mathbf{F}_{1:t,\theta_0}^{-1/2}$. It follows from Lemma 50 that

$$\mathbb{P} \left(\left\| \mathbf{F}_{1:t,\theta_0}^{-1/2} \mathbf{X}_{1:t}^\top \mathcal{E}_{1:t} \right\|_2 \geq R_{n,T}(\mathbf{B}_{1:t}) \text{ for some } t \in [T] \mid \mathbf{X}, \tilde{\mathcal{E}} \right) \leq n^{-1},$$

Consequently, we have

$$\begin{aligned} & \mathbb{P} \left(\left\| \mathbf{F}_{1:t,\theta_0}^{-1/2} \mathbf{X}_{1:t}^\top \mathcal{E}_{1:t} \right\|_2 \geq R_{n,T}(\mathbf{B}_{1:t}) \text{ for some } t \in [T] \right) \\ & \leq \mathbb{E}_{\mathbf{X}} \left[\mathbb{P} \left(\left\| \mathbf{F}_{1:t,\theta_0}^{-1/2} \mathbf{X}_{1:t}^\top \mathcal{E}_{1:t} \right\|_2 \geq R_{n,T}(\mathbf{B}_{1:t}) \text{ for some } t \in [T] \mid \mathbf{X}, \tilde{\mathcal{E}} \right) \mathbf{1}_{\tilde{\mathcal{E}}} \right] + \mathbb{P}(\tilde{\mathcal{E}}^c) \\ & \leq n^{-1} + \mathbb{P}(\mathcal{E}^c) \leq n^{-1} + 6e^{-n/72} + 2(Np)^{-1}. \end{aligned}$$

This completes the proof of the third assertion. Therefore, with \mathbb{P} -probability at least $1 - 3n^{-1} - 6e^{-n/72} - 2(Np)^{-1}$, the three assertions hold uniformly for all $t \in [T]$. \blacksquare

The following proposition verifies (35) and (36).

Proposition 39 *Suppose that $(\mathbf{E}\mathbf{X})$ holds. Then,*

$$\mathbb{P} \left(\left[\max_{t \in [T]} \sup_{\theta \in \Theta} \|\nabla^3 L_t(\theta)\|_{\text{op}} \right] \vee \left[\max_{t \in [T]} \sup_{\theta \in \Theta} \|\nabla^4 L_t(\theta)\|_{\text{op}} \right] \leq \tilde{K}'_{\max} n \right) \geq 1 - 2n^{-1}, \quad (208)$$

where $\tilde{K}'_{\max} > 0$ is a universal constant.

Proof For $b(\cdot) = \log(1 + \exp(\cdot))$, note that

$$\sup_{\eta \in \mathbb{R}} \left[b'''(\eta) \vee b''''(\eta) \right] \leq 1.$$

For $t \in [T]$, we have

$$\begin{aligned} \sup_{\theta \in \Theta} \|\nabla^3 L_t(\theta)\|_{\text{op}} &= \sup_{\theta \in \Theta} \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \sum_{i \in I_t} b'''(X_i^\top \theta) (X_i^\top u)^3 \leq \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \sum_{i \in I_t} |X_i^\top u|^3, \\ \sup_{\theta \in \Theta} \|\nabla^4 L_t(\theta)\|_{\text{op}} &= \sup_{\theta \in \Theta} \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \sum_{i \in I_t} b''''(X_i^\top \theta) (X_i^\top u)^4 \leq \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \sum_{i \in I_t} |X_i^\top u|^4. \end{aligned}$$

By Lemma 52, for $t \in [T]$, $k \geq 2$ and $\omega, \tau \geq 1$, there exists some constants $D_1 = D_1(k) > 1$, $D_2 > 0$ and $D_3 = D_3(k) > 0$ such that

$$\begin{aligned} & \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \left| \frac{1}{n} \sum_{i \in I_t} \left(|\langle X_i, u \rangle|^k - \mathbb{E} |\langle X_i, u \rangle|^k \right) \right| \\ & \leq D_1 \tau \omega^{k-1} \log^{k-1} \left(\frac{2n}{p} \right) \sqrt{\frac{p}{n}} + D_1 \frac{\omega^k p^{k/2}}{n} + D_1 \left(\frac{p}{2n} \right)^\omega \end{aligned}$$

with a probability at least

$$1 - \exp(-D_2\omega\sqrt{p}) - \exp\left(-D_3\left\{\left[\tau^2\omega^{2k-2}p\log^{(2k-2)}\left(\frac{2n}{p}\right)\right] \wedge \left[\tau\omega^{-1}\sqrt{np}\log^{-1}\left(\frac{2n}{p}\right)\right]\right\}\right).$$

By **(EX)**, we have

$$(\log T + \log n)^{k-1} \log^{(k-1)}\left(\frac{2n}{p}\right) \sqrt{\frac{p}{n}} \leq \delta, \quad (\log T + \log n)^k p^{k/2} n^{-1} \leq \delta, \quad \frac{p}{2n} \leq \delta$$

for a small enough constant $\delta > 0$ depending only D_1 . Also,

$$\begin{aligned} (\log T + \log n)^4 \log^4(2n/p) &\geq C'(\log n + \log T), \\ \sqrt{np}\log^{-1}(2n/p)(\log T + \log n)^{-1} &\geq C'(\log n + \log T) \end{aligned}$$

for a large enough constant $C' > 0$ depending only D_3 . By taking

$$\omega = \log T + \log n, \quad \tau = 1,$$

after some algebra, for any $k \in \{3, 4\}$, there exist some positive constants $c_1 = c_1(D_1)$ and $c_2 = c_2(D_3)$ such that

$$\begin{aligned} &\max_{t \in [T]} \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \left| \frac{1}{n} \sum_{i \in I_t} \left(|\langle X_i, u \rangle|^k - \mathbb{E} |\langle X_i, u \rangle|^k \right) \right| \\ &\leq c_1 \left[(\log T + \log n)^{k-1} \log^{(k-1)}\left(\frac{2n}{p}\right) \sqrt{\frac{p}{n}} + (\log T + \log n)^k p^{k/2} n^{-1} + \frac{p}{2n} \right] \leq 3 \end{aligned}$$

with \mathbb{P} -probability at least

$$\begin{aligned} &1 - e^{-\sqrt{p}(\log T + \log n) + \log T} \\ &\quad - \exp\left(-c_2 \left[\{(\log T + \log n)^4 \log^4(2n/p)\} \wedge \{\sqrt{np}\log^{-1}(2n/p)(\log T + \log n)^{-1}\} \right] + \log T \right) \\ &\geq 1 - e^{-\log n - \log T + \log T} - e^{-\log n - \log T + \log T} = 1 - 2n^{-1}. \end{aligned}$$

Also, for any $i \in I_t$ and $u \in \mathbb{R}^p$ with $\|u\|_2 = 1$,

$$\mathbb{E} |\langle X_i, u \rangle|^3 = 2\sqrt{\frac{2}{\pi}}, \quad \mathbb{E} |\langle X_i, u \rangle|^4 = 15.$$

Therefore, we have

$$\begin{aligned} \max_{t \in [T]} \sup_{\theta \in \Theta} \|\nabla^3 L_t(\theta)\|_{\text{op}} &\leq \max_{t \in [T]} \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \sum_{i \in I_t} |X_i^\top u|^3 \leq \left(3 + 2\sqrt{\frac{2}{\pi}}\right) n, \\ \max_{t \in [T]} \sup_{\theta \in \Theta} \|\nabla^4 L_t(\theta)\|_{\text{op}} &\leq \max_{t \in [T]} \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \sum_{i \in I_t} |X_i^\top u|^4 \leq (3 + 15) n = 18n \end{aligned}$$

with \mathbb{P} -probability at least $1 - 2n^{-1}$. By taking $\tilde{K}'_{\max} = (3 + 2\sqrt{2/\pi}) \vee 18 = 18$, we complete the proof. \blacksquare

Proof [Proof of Proposition 31] Let

$$\begin{aligned}\Omega_1 &= \{ \text{The three assertions in Proposition 38 hold uniformly for all } t \in [T] \}, \\ \Omega_2 &= \{ (206) \text{ and } (207) \text{ hold with the constants } \tilde{K}_{\min}, \tilde{K}_{\max} \text{ and } M \}, \\ \Omega_3 &= \{ (208) \text{ holds with the constant } \tilde{K}'_{\max} \}.\end{aligned}$$

By Propositions 36, 38 and 39, we have

$$\begin{aligned}\mathbb{P}(\Omega_1) &\geq 1 - 3n^{-1} - 6e^{-n/72} - 2(Np)^{-1}, \\ \mathbb{P}(\Omega_2) &\geq 1 - 4e^{-n/72} - 2(Np)^{-1}, \\ \mathbb{P}(\Omega_3) &\geq 1 - 2n^{-1}.\end{aligned}$$

It follows that

$$\mathbb{P}(\Omega_1 \cap \Omega_2 \cap \Omega_3) \geq 1 - 5n^{-1} - 10e^{-n/72} - 4(Np)^{-1}.$$

Let $\Omega = \Omega_1 \cap \Omega_2 \cap \Omega_3$. On Ω , the assumptions in **(A1)**, **(A1*)** and **(A2)** are satisfied when K_{\min}, K_{\max} and M_n are replaced by $\tilde{K}_{\min}, (\tilde{K}_{\max} \vee \tilde{K}'_{\max} \vee K_2 \vee K_3)$ and M , respectively. Recall that

$$\tilde{K}_{\min} = (1080e^{2K_1+3})^{-1}, \quad \tilde{K}_{\max} \vee \tilde{K}'_{\max} = 18, \quad M = 27\sqrt{30}e^{K_1+3/2}.$$

By **(EX)**, all conditions specified in Theorem 11, Propositions 36, 38 and 39 hold. Therefore, the result of Proposition 31 follows from Theorem 11. \blacksquare

Appendix H. Technical lemmas

H.1 General technical lemmas

In this subsection, assume we are given a function $f : \Theta \rightarrow \mathbb{R}$ that is four times continuously differentiable. For $\theta \in \Theta$, let $\mathbf{F}_\theta = -\nabla^2 f(\theta)$, and assume \mathbf{F}_θ is nonsingular in this subsection.

Lemma 40 *For a given $\theta \in \Theta$, suppose that f satisfies the third order smoothness at θ with parameters $(\tau_3, \mathbf{F}_\theta, r)$. Then,*

$$\sup_{\theta' \in \Theta(\theta, \mathbf{F}_\theta, r)} \left\| \mathbf{F}_\theta^{-1/2} \mathbf{F}_{\theta'} \mathbf{F}_\theta^{-1/2} - \mathbf{I}_p \right\|_2 \leq \tau_3 r.$$

Consequently, for every $u \in \mathbb{R}^p$ and $\theta' \in \Theta(\theta, \mathbf{F}_\theta, r)$, we have

$$(1 - \tau_3 r) \left\| \mathbf{F}_\theta^{1/2} u \right\|_2^2 \leq \left\| \mathbf{F}_{\theta'}^{1/2} u \right\|_2^2 \leq (1 + \tau_3 r) \left\| \mathbf{F}_\theta^{1/2} u \right\|_2^2.$$

Proof For $\theta' \in \Theta(\theta, \mathbf{F}_\theta, r)$, we have

$$\begin{aligned}
 & \left\| \mathbf{F}_\theta^{-1/2} \mathbf{F}_{\theta'} \mathbf{F}_\theta^{-1/2} - \mathbf{I}_p \right\|_2 = \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \left| \langle \mathbf{F}_\theta^{-1/2} \mathbf{F}_{\theta'} \mathbf{F}_\theta^{-1/2} - \mathbf{I}_p, u^{\otimes 2} \rangle \right| \\
 & \leq \sup_{h \in \Theta(\mathbf{F}_\theta, r)} \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \left| \langle \mathbf{F}_\theta^{-1/2} \mathbf{F}_{\theta+h} \mathbf{F}_\theta^{-1/2} - \mathbf{I}_p, u^{\otimes 2} \rangle \right| \\
 & = \sup_{h \in \Theta(\mathbf{F}_\theta, r)} \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \left| \left\langle \mathbf{F}_{\theta+h} - \mathbf{F}_\theta, \left(\mathbf{F}_\theta^{-1/2} u \right)^{\otimes 2} \right\rangle \right| \\
 & \leq \sup_{h, h' \in \Theta(\mathbf{F}_\theta, r)} \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \left| \langle \nabla^3 f(\theta + h'), \left(\mathbf{F}_\theta^{-1/2} u \right)^{\otimes 2} \otimes h \rangle \right| \\
 & \stackrel{(12)}{\leq} \sup_{h \in \Theta(\mathbf{F}_\theta, r)} \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \tau_3 \left\| \mathbf{F}_\theta^{1/2} \mathbf{F}_\theta^{-1/2} u \right\|_2^2 \left\| \mathbf{F}_\theta^{1/2} h \right\|_2 \leq \tau_3 r.
 \end{aligned}$$

This completes the proof. ■

Lemma 41 *Let $\theta, \tilde{\theta}, \beta \in \Theta$ and $r = \|\mathbf{F}_\theta^{-1/2} \beta\|_2$. Suppose that f is concave and satisfies the third order smoothness at θ with parameters $(\tau_3, \mathbf{F}_\theta, 4r)$. Assume further that*

$$\tau_3 r \leq 1/16, \quad \nabla f(\theta) = 0, \quad \nabla f(\tilde{\theta}) + \beta = 0.$$

Then,

$$\left\| \mathbf{F}_\theta^{1/2} (\tilde{\theta} - \theta) \right\|_2 \leq 4r.$$

Proof Let $\Theta_{\theta, r} = \Theta(\theta, \mathbf{F}_\theta, 4r)$, $\bar{\Theta}_{\theta, r} = \Theta(\mathbf{F}_\theta, 4r)$ and

$$\partial\Theta_{\theta, r} = \left\{ \theta' \in \Theta : \left\| \mathbf{F}_\theta^{1/2} (\theta' - \theta) \right\|_2 = 4r \right\}.$$

Then, it suffices to prove that $\tilde{\theta} \in \Theta_{\theta, r}$.

For $\theta' \in \Theta$, let $g(\theta') = f(\theta') + \langle \beta, \theta' \rangle$. Then, $\nabla g(\tilde{\theta}) = 0$ and the map $\theta' \mapsto g(\theta')$ is concave. By the concavity of $g(\cdot)$, for any $\theta' \in \Theta_{\theta, r}^c$, we have

$$g(\bar{\theta}) \geq \omega g(\theta') + (1 - \omega)g(\theta), \tag{209}$$

where $\bar{\theta} = \omega \theta' + (1 - \omega)\theta$ and $\omega = 4r \|\mathbf{F}_\theta^{1/2} (\theta' - \theta)\|_2^{-1} \in (0, 1)$. One can easily check that $\bar{\theta} \in \partial\Theta_{\theta, r}$. At the end of this proof, we will show that

$$\sup_{\theta^\circ \in \partial\Theta_{\theta, r}} g(\theta^\circ) - g(\theta) \leq -2r^2 < 0. \tag{210}$$

It follows that, for any $\theta' \in \Theta_{\theta, r}^c$,

$$0 > -2r^2 \geq \sup_{\theta^\circ \in \partial\Theta_{\theta, r}} g(\theta^\circ) - g(x) \geq g(\bar{\theta}) - g(x) \stackrel{(209)}{\geq} \omega [g(\theta') - g(\theta)] \geq g(\theta') - g(\theta),$$

which implies that $\tilde{\theta} \in \Theta_{\theta,r}$.

To complete the proof, we only need to prove (210). Let $\theta^\circ \in \partial\Theta_{\theta,r}$ and $u = \theta^\circ - \theta$. By Taylor's theorem, there exists some $\tilde{u} \in \bar{\Theta}_{\theta,r}$ such that

$$\begin{aligned}
 g(\theta^\circ) - g(\theta) &= \nabla g(\theta)^\top u + \frac{1}{2} \langle \nabla^2 g(\theta + \tilde{u}), u^{\otimes 2} \rangle \\
 &= [\nabla f(\theta) + \beta]^\top u + \frac{1}{2} \langle \nabla^2 f(\theta + \tilde{u}), u^{\otimes 2} \rangle = \beta^\top u + \frac{1}{2} \langle \nabla^2 f(\theta + \tilde{u}), u^{\otimes 2} \rangle \\
 &= [\mathbf{F}_\theta^{-1/2} \beta]^\top \mathbf{F}_\theta^{1/2} u - \frac{1}{2} \langle \mathbf{F}_{\theta+\tilde{u}}, u^{\otimes 2} \rangle \\
 &\stackrel{\text{Lemma 40}}{\leq} \left(\left\| \mathbf{F}_\theta^{-1/2} \beta \right\|_2 - \frac{1}{2} (1 - 4\tau_3 r) \left\| \mathbf{F}_\theta^{1/2} u \right\|_2 \right) \left\| \mathbf{F}_\theta^{1/2} u \right\|_2 \\
 &\leq \left[r - 2(1 - 4\tau_3 r) r \right] \times 4r, \quad (\because \left\| \mathbf{F}_\theta^{1/2} u \right\|_2 = 4r) \\
 &\leq -2r^2,
 \end{aligned}$$

where the last inequality holds by $\tau_3 r \leq 1/16$. This completes the proof. \blacksquare

Lemma 42 *Let $\theta_c \in \Theta$, $\mathbf{F} \in \mathcal{S}_{++}^p$ and $r \geq 0$ be given. Then, f satisfies the third and fourth order smoothness at θ_c with parameters (τ_3, \mathbf{F}, r) and (τ_4, \mathbf{F}, r) , respectively, where*

$$\begin{aligned}
 \tau_3 &= \lambda_{\min}^{-3/2}(\mathbf{F}) \sup_{\theta \in \Theta(\theta_c, \mathbf{F}, r)} \left\| \nabla^3 f(\theta) \right\|_{\text{op}}, \\
 \tau_4 &= \lambda_{\min}^{-2}(\mathbf{F}) \sup_{\theta \in \Theta(\theta_c, \mathbf{F}, r)} \left\| \nabla^4 f(\theta) \right\|_{\text{op}}.
 \end{aligned}$$

Proof Note that

$$\begin{aligned}
 &\sup_{u \in \Theta(\mathbf{F}, r)} \sup_{z \in \mathbb{R}^p} \frac{|\langle \nabla^3 f(\theta_c + u), z^{\otimes 3} \rangle|}{\left\| \mathbf{F}^{1/2} z \right\|_2^3} = \sup_{u \in \Theta(\mathbf{F}, r)} \sup_{z \in \mathbb{R}^p} \left| \left\langle \nabla^3 f(\theta_c + u), \frac{z^{\otimes 3}}{\left\| \mathbf{F}^{1/2} z \right\|_2^3} \right\rangle \right| \\
 &\leq \lambda_{\min}^{-3/2}(\mathbf{F}) \sup_{u \in \Theta(\mathbf{F}, r)} \sup_{z \in \mathbb{R}^p: \|z\|_2=1} |\langle \nabla^3 f(\theta_c + u), z^{\otimes 3} \rangle| \\
 &\leq \lambda_{\min}^{-3/2}(\mathbf{F}) \sup_{\theta \in \Theta(\theta_c, \mathbf{F}, r)} \left\| \nabla^3 f(\theta) \right\|_{\text{op}}.
 \end{aligned}$$

Also,

$$\begin{aligned}
 &\sup_{u \in \Theta(\mathbf{F}, r)} \sup_{z \in \mathbb{R}^p} \frac{|\langle \nabla^4 f(\theta_c + u), z^{\otimes 4} \rangle|}{\left\| \mathbf{F}^{1/2} z \right\|_2^4} = \sup_{u \in \Theta(\mathbf{F}, r)} \sup_{z \in \mathbb{R}^p} \left| \left\langle \nabla^4 f(\theta_c + u), \frac{z^{\otimes 4}}{\left\| \mathbf{F}^{1/2} z \right\|_2^4} \right\rangle \right| \\
 &\leq \lambda_{\min}^{-2}(\mathbf{F}) \sup_{u \in \Theta(\mathbf{F}, r)} \sup_{z \in \mathbb{R}^p: \|z\|_2=1} |\langle \nabla^4 f(\theta_c + u), z^{\otimes 4} \rangle| \\
 &\leq \lambda_{\min}^{-2}(\mathbf{F}) \sup_{\theta \in \Theta(\theta_c, \mathbf{F}, r)} \left\| \nabla^4 f(\theta) \right\|_{\text{op}},
 \end{aligned}$$

which completes the proof. \blacksquare

H.2 Technical lemmas for TV distance

Lemma 43 For $\mu_1, \mu_2 \in \mathbb{R}^p$ and $\mathbf{\Omega}_1, \mathbf{\Omega}_2 \in \mathcal{S}_{++}^p$, let $Q_1 = \mathcal{N}(\mu_1, \mathbf{\Omega}_1^{-1})$, $Q_2 = \mathcal{N}(\mu_2, \mathbf{\Omega}_2^{-1})$. Suppose that

$$\left\| \mathbf{\Omega}_2^{-1/2} \mathbf{\Omega}_1 \mathbf{\Omega}_2^{-1/2} - \mathbf{I}_p \right\|_2 \leq 0.684.$$

Then,

$$d_V(Q_1, Q_2) \leq \frac{1}{2} \left(\left\| \mathbf{\Omega}_1^{1/2} (\mu_1 - \mu_2) \right\|_2^2 + \left\| \mathbf{\Omega}_2^{-1/2} \mathbf{\Omega}_1 \mathbf{\Omega}_2^{-1/2} - \mathbf{I}_p \right\|_F^2 \right)^{1/2}.$$

Proof By Pinsker's inequality, we have

$$d_V(Q_1, Q_2) \leq \left(\frac{1}{2} K(Q_1, Q_2) \right)^{1/2}.$$

By the definition of KL divergence, $K(Q_1, Q_2)$ is equal to

$$\frac{1}{2} \left[\left\| \mathbf{\Omega}_1^{1/2} (\mu_1 - \mu_2) \right\|_2^2 + \text{tr} \left(\mathbf{\Omega}_2^{-1/2} \mathbf{\Omega}_1 \mathbf{\Omega}_2^{-1/2} - \mathbf{I}_p \right) - \log \det \left(\mathbf{\Omega}_2^{-1/2} \mathbf{\Omega}_1 \mathbf{\Omega}_2^{-1/2} \right) \right].$$

Let $(\lambda_j)_{j \in [p]}$ be eigenvalues of $\mathbf{B} = \mathbf{\Omega}_2^{-1/2} \mathbf{\Omega}_1 \mathbf{\Omega}_2^{-1/2} - \mathbf{I}_p$. Then, the last display is represented by

$$\begin{aligned} & \frac{1}{2} \left[\left\| \mathbf{\Omega}_1^{1/2} (\mu_1 - \mu_2) \right\|_2^2 + \sum_{j=1}^p \lambda_j - \sum_{j=1}^p \log(1 + \lambda_j) \right] \\ & \leq \frac{1}{2} \left[\left\| \mathbf{\Omega}_1^{1/2} (\mu_1 - \mu_2) \right\|_2^2 + \sum_{j=1}^p \lambda_j - \sum_{j=1}^p (\lambda_j - \lambda_j^2) \right] = \frac{1}{2} \left[\left\| \mathbf{\Omega}_1^{1/2} (\mu_1 - \mu_2) \right\|_2^2 + \sum_{j=1}^p \lambda_j^2 \right] \\ & = \frac{1}{2} \left[\left\| \mathbf{\Omega}_1^{1/2} (\mu_1 - \mu_2) \right\|_2^2 + \text{tr}(\mathbf{B}^2) \right] = \frac{1}{2} \left[\left\| \mathbf{\Omega}_1^{1/2} (\mu_1 - \mu_2) \right\|_2^2 + \|\mathbf{B}\|_F^2 \right]. \end{aligned}$$

where the first inequality holds by $\max_{j \in [p]} |\lambda_j| \leq 0.684$. It follows that

$$d_V(Q_1, Q_2) \leq \left(\frac{1}{2} K(Q_1, Q_2) \right)^{1/2} \leq \frac{1}{2} \left[\left\| \mathbf{\Omega}_1^{1/2} (\mu_1 - \mu_2) \right\|_2^2 + \|\mathbf{B}\|_F^2 \right]^{1/2},$$

which completes the proof. ■

Lemma 44 For $\mu_1, \mu_2 \in \mathbb{R}^p$ and $\mathbf{\Omega}_1, \mathbf{\Omega}_2 \in \mathcal{S}_{++}^p$, let $Q_1 = \mathcal{N}(\mu_1, \mathbf{\Omega}_1^{-1})$, $Q_2 = \mathcal{N}(\mu_2, \mathbf{\Omega}_2^{-1})$. Suppose that

$$d_V(Q_1, Q_2) \leq \frac{1}{600}.$$

Let

$$\Delta = \left\| \mathbf{\Omega}_1^{1/2} (\mu_1 - \mu_2) \right\|_2 \vee \left\| \mathbf{\Omega}_2^{-1/2} \mathbf{\Omega}_1 \mathbf{\Omega}_2^{-1/2} - \mathbf{I}_p \right\|_F.$$

Then,

$$\frac{\Delta}{200} \leq d_V(Q_1, Q_2) \leq \frac{\Delta}{\sqrt{2}}.$$

Proof See Theorem 1.8 in Arbas et al. (2023). ■

H.3 Technical lemmas for eigenvalue analysis

Lemma 45 *Suppose that (A0)-(A2) hold. Also, assume that*

$$n \geq C \left\| \mathbf{\Omega}_t^{1/2} (\theta_0 - \mu_t) \right\|_2^2 \quad (211)$$

on an event \mathcal{E} , where $C = C(K_{\min}, K_{\max})$ is a large enough constant. Then, on \mathcal{E} ,

$$\left\| \mathbf{F}_{t+1, \theta_0}^{1/2} (\theta_0 - \theta_{t+1}^*) \right\|_2 \leq \sqrt{2} \left\| \mathbf{\Omega}_t^{1/2} (\theta_0 - \mu_t) \right\|_2. \quad (212)$$

Proof In this proof, we will work on the event \mathcal{E} without explicitly mentioning it, and assume $C = C(K_{\min}, K_{\max})$ in (211) is large enough. By the definition of θ_{t+1}^* , we have

$$\begin{aligned} \mathbb{E}_{t+1} L_{t+1}(\theta_0) - \frac{1}{2} \left\| \mathbf{\Omega}_t^{1/2} (\theta_0 - \mu_t) \right\|_2^2 &= \mathbb{E}_{t+1} \tilde{L}_{t+1}(\theta_0) \leq \mathbb{E}_{t+1} \tilde{L}_{t+1}(\theta_{t+1}^*) \\ &= \mathbb{E}_{t+1} L_{t+1}(\theta_{t+1}^*) - \frac{1}{2} \left\| \mathbf{\Omega}_t^{1/2} (\theta_{t+1}^* - \mu_t) \right\|_2^2 \\ &\leq \mathbb{E}_{t+1} L_{t+1}(\theta_{t+1}^*). \end{aligned}$$

It follows that

$$\mathbb{E}_{t+1} L_{t+1}(\theta_{t+1}^*) - \mathbb{E}_{t+1} L_{t+1}(\theta_0) \geq -\frac{1}{2} \left\| \mathbf{\Omega}_t^{1/2} (\theta_0 - \mu_t) \right\|_2^2. \quad (213)$$

In this proof, we denote $\Theta_{n, t+1} = \Theta(\theta_0, \mathbf{F}_{t+1, \theta_0}, \sqrt{2} \left\| \mathbf{\Omega}_t^{1/2} (\theta_0 - \mu_t) \right\|_2)$. For $\theta \in \Theta_{n, t+1}$, we have

$$\begin{aligned} \|\theta - \theta_0\|_2 &\leq \lambda_{\min}^{-1/2}(\mathbf{F}_{t+1, \theta_0}) \sqrt{2} \left\| \mathbf{\Omega}_t^{1/2} (\theta_0 - \mu_t) \right\|_2 \\ &\stackrel{\text{(A2)}}{\leq} (K_{\min} n)^{-1/2} \sqrt{2} \left\| \mathbf{\Omega}_t^{1/2} (\theta_0 - \mu_t) \right\|_2 \stackrel{(211)}{\leq} \frac{1}{2}. \end{aligned}$$

It follows that

$$\Theta_{n, t+1} \subseteq \{\theta \in \Theta : \|\theta - \theta_0\|_2 \leq 1/2\}.$$

By Lemma 42 and **(A2)**, $\mathbb{E}_{t+1}L_{t+1}(\theta)$ satisfies the third order smoothness at θ_0 with parameters

$$\left(K_{\max}K_{\min}^{-3/2}n^{-1/2}, \mathbf{F}_{t+1,\theta_0}, \sqrt{2}\left\|\boldsymbol{\Omega}_t^{1/2}(\theta_0 - \mu_t)\right\|_2\right).$$

We will prove (212) by contradiction. Suppose $\theta_{t+1}^* \notin \Theta_{n,t+1}$. Let

$$\partial\Theta_{n,t+1} = \left\{\theta \in \Theta : \left\|\mathbf{F}_{t+1,\theta}^{1/2}(\theta - \theta_0)\right\|_2 = \sqrt{2}\left\|\boldsymbol{\Omega}_t^{1/2}(\theta_0 - \mu_t)\right\|_2\right\}.$$

For $\theta^\circ \in \partial\Theta_{n,t+1}$, we have

$$\begin{aligned} & \mathbb{E}_{t+1}L_{t+1}(\theta^\circ) - \mathbb{E}_{t+1}L_{t+1}(\theta_0) \\ & \leq \left(\nabla\mathbb{E}_{t+1}L_{t+1}(\theta_0)\right)^\top (\theta^\circ - \theta_0) - \frac{1}{2} \inf_{\theta \in \Theta_{n,t+1}} \left\|\mathbf{F}_{t+1,\theta}^{1/2}(\theta^\circ - \theta_0)\right\|_2^2 \\ & = -\frac{1}{2} \inf_{\theta \in \Theta_{n,t+1}} \left\|\mathbf{F}_{t+1,\theta}^{1/2}(\theta^\circ - \theta_0)\right\|_2^2 \\ & \stackrel{\text{Lemma 40}}{\leq} -\frac{1}{2} \left(1 - K_{\max}K_{\min}^{-3/2}n^{-1/2}\sqrt{2}\left\|\boldsymbol{\Omega}_t^{1/2}(\theta_0 - \mu_t)\right\|_2\right) \left\|\mathbf{F}_{t+1,\theta_0}^{1/2}(\theta^\circ - \theta_0)\right\|_2^2 \\ & = -\frac{1}{2} \left(1 - K_{\max}K_{\min}^{-3/2}n^{-1/2}\sqrt{2}\left\|\boldsymbol{\Omega}_t^{1/2}(\theta_0 - \mu_t)\right\|_2\right) 2\left\|\boldsymbol{\Omega}_t^{1/2}(\theta_0 - \mu_t)\right\|_2^2 \\ & \stackrel{(211)}{<} -\frac{1}{2} \left\|\boldsymbol{\Omega}_t^{1/2}(\theta_0 - \mu_t)\right\|_2^2. \end{aligned}$$

where the second equality holds by the definition of $\partial\Theta_{n,t+1}$. Consequently, we have

$$\mathbb{E}_{t+1}L_{t+1}(\theta_{t+1}^*) - \mathbb{E}_{t+1}L_{t+1}(\theta_0) < -\frac{1}{2} \left\|\boldsymbol{\Omega}_t^{1/2}(\theta_0 - \mu_t)\right\|_2^2,$$

by the concavity of the map $\theta \mapsto \mathbb{E}_{t+1}L_{t+1}(\theta)$, which contradicts to (213). This completes the proof of (212). \blacksquare

Remark 46 *The constant C in (211) can be chosen as*

$$C = 8K_{\min}^{-1} \vee (8K_{\max}^2K_{\min}^{-3} + 1).$$

Lemma 47 *Suppose that **(A0)**-**(A2)** hold. Also, on an event \mathcal{E} , assume that $\|\theta_t^* - \theta_0\|_2 \leq 1/4$, and*

$$n \geq C \left(\left\|\mathbf{F}_{t+1,\theta_t^*}^{-1/2} \nabla\mathbb{E}_{t+1}L_{t+1}(\theta_t^*)\right\|_2^2 \vee \left\|\boldsymbol{\Omega}_t^{1/2}(\theta_t^* - \mu_t)\right\|_2^2 \right) \quad (214)$$

for a large enough constant $C = C(K_{\min}, K_{\max})$. Then, on \mathcal{E} ,

$$\begin{aligned} & \left\|\mathbf{F}_{t+1,\theta_t^*}^{1/2}(\theta_{t+1}^* - \theta_t^*)\right\|_2 \\ & \leq \left(4 \left\|\mathbf{F}_{t+1,\theta_t^*}^{-1/2} \nabla\mathbb{E}_{t+1}L_{t+1}(\theta_t^*)\right\|_2\right) \vee \left(2 \left\|\boldsymbol{\Omega}_t^{1/2}(\theta_t^* - \mu_t)\right\|_2\right). \end{aligned} \quad (215)$$

Proof In this proof, we work on the event \mathcal{E} without explicitly referring it, and assume $C = C(K_{\min}, K_{\max})$ in (214) is large enough. By the definition of θ_{t+1}^* , we have, for $t \in \{0, 1, \dots, T-1\}$,

$$\begin{aligned} & \mathbb{E}_{t+1} L_{t+1}(\theta_t^*) - \frac{1}{2} \left\| \boldsymbol{\Omega}_t^{1/2} (\theta_t^* - \mu_t) \right\|_2^2 = \mathbb{E}_{t+1} \tilde{L}_{t+1}(\theta_t^*) \leq \mathbb{E}_{t+1} \tilde{L}_{t+1}(\theta_{t+1}^*) \\ & = \mathbb{E}_{t+1} L_{t+1}(\theta_{t+1}^*) - \frac{1}{2} \left\| \boldsymbol{\Omega}_t^{1/2} (\theta_{t+1}^* - \mu_t) \right\|_2^2 \leq \mathbb{E}_{t+1} L_{t+1}(\theta_{t+1}^*). \end{aligned}$$

It follows that

$$\mathbb{E}_{t+1} L_{t+1}(\theta_{t+1}^*) - \mathbb{E}_{t+1} L_{t+1}(\theta_t^*) \geq -\frac{1}{2} \left\| \boldsymbol{\Omega}_t^{1/2} (\theta_t^* - \mu_t) \right\|_2^2. \quad (216)$$

Let

$$r_t = \left(4 \left\| \mathbf{F}_{t+1, \theta_t^*}^{-1/2} \nabla \mathbb{E}_{t+1} L_{t+1}(\theta_t^*) \right\|_2 \right) \vee \left(2 \left\| \boldsymbol{\Omega}_t^{1/2} (\theta_t^* - \mu_t) \right\|_2 \right).$$

In this proof, we denote $\Theta_{n, t+1} = \Theta(\theta_t^*, \mathbf{F}_{t+1, \theta_t^*}, r_t)$. For $\theta \in \Theta_{n, t+1}$, we have

$$\|\theta - \theta_t^*\|_2 \leq \lambda_{\min}^{-1/2}(\mathbf{F}_{t+1, \theta_t^*}) r_t \leq (K_{\min} n)^{-1/2} r_t \stackrel{(214)}{\leq} \frac{1}{4},$$

where the second inequality holds by $\|\theta_t^* - \theta_0\|_2 \leq 1/4$ and **(A2)**. It follows that

$$\Theta_{n, t+1} \subseteq \{\theta \in \Theta : \|\theta - \theta_0\|_2 \leq 1/2\}.$$

By Lemma 42 and **(A2)**, $\mathbb{E}_{t+1} L_{t+1}(\theta)$ satisfies the third order smoothness at θ_t^* with parameters

$$\left(K_{\max} K_{\min}^{-3/2} n^{-1/2}, \mathbf{F}_{t+1, \theta_t^*}, r_t \right).$$

Next, we will prove (215) by contradiction. Suppose $\theta_{t+1}^* \notin \Theta_{n, t+1}$. Let

$$\partial\Theta_{n, t+1} = \left\{ \theta \in \Theta : \left\| \mathbf{F}_{t+1, \theta_t^*}^{1/2} (\theta - \theta_t^*) \right\|_2 = r_t \right\}. \quad (217)$$

For $\theta^\circ \in \partial\Theta_{n, t+1}$, Taylor's theorem gives

$$\begin{aligned} & \mathbb{E}_{t+1} L_{t+1}(\theta^\circ) - \mathbb{E}_{t+1} L_{t+1}(\theta_t^*) \\ & \leq \left(\nabla \mathbb{E}_{t+1} L_{t+1}(\theta_t^*) \right)^\top (\theta^\circ - \theta_t^*) - \frac{1}{2} \inf_{\theta \in \Theta_{n, t+1}} \left\| \mathbf{F}_{t+1, \theta}^{1/2} (\theta^\circ - \theta_t^*) \right\|_2^2 \\ & = \left(\mathbf{F}_{t+1, \theta_t^*}^{-1/2} \nabla \mathbb{E}_{t+1} L_{t+1}(\theta_t^*) \right)^\top \mathbf{F}_{t+1, \theta_t^*}^{1/2} (\theta^\circ - \theta_t^*) - \frac{1}{2} \inf_{\theta \in \Theta_{n, t+1}} \left\| \mathbf{F}_{t+1, \theta}^{1/2} (\theta^\circ - \theta_t^*) \right\|_2^2 \\ & \stackrel{\text{Lemma 40}}{\leq} \left\| \mathbf{F}_{t+1, \theta_t^*}^{-1/2} \nabla \mathbb{E}_{t+1} L_{t+1}(\theta_t^*) \right\|_2 \left\| \mathbf{F}_{t+1, \theta_t^*}^{1/2} (\theta^\circ - \theta_t^*) \right\|_2 \\ & \quad - \frac{1}{2} \left(1 - K_{\max} K_{\min}^{-3/2} n^{-1/2} r_t \right) \left\| \mathbf{F}_{t+1, \theta_t^*}^{1/2} (\theta^\circ - \theta_t^*) \right\|_2^2 \\ & \stackrel{(217)}{=} \left[\left\| \mathbf{F}_{t+1, \theta_t^*}^{-1/2} \nabla \mathbb{E}_{t+1} L_{t+1}(\theta_t^*) \right\|_2 - \frac{1}{2} \left(1 - K_{\max} K_{\min}^{-3/2} n^{-1/2} r_t \right) r_t \right] r_t \\ & \stackrel{(214)}{<} \left[\frac{1}{4} r_t - \frac{1}{2} \left(1 - \frac{1}{4} \right) r_t \right] r_t = -\frac{1}{8} r_t^2. \end{aligned}$$

Consequently,

$$\mathbb{E}_{t+1}L_{t+1}(\theta_{t+1}^*) - \mathbb{E}_{t+1}L_{t+1}(\theta_t^*) < -\frac{1}{8}r_t^2 \leq -\frac{1}{2} \left\| \boldsymbol{\Omega}_t^{1/2} (\theta_t^* - \mu_t) \right\|_2^2,$$

by the concavity of the map $\theta \mapsto \mathbb{E}_{t+1}L_{t+1}(\theta)$, which contradicts to (216). This completes the proof of (215). \blacksquare

Lemma 48 *Suppose that (A0)-(A2) hold. Let $\alpha \in [1/2, 1]$. Also, assume that there exist some constants $D_1, D_2, D_3, D_4 > 0$ such that*

$$\begin{aligned} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\theta_t^* - \theta_0) \right\|_2 &\leq D_1 M_n t^\alpha \sqrt{p_*}, & \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\hat{\theta}_t - \theta_t^*) \right\|_2 &\leq D_2 M_n \sqrt{t^{-1} p_*}, \\ \left\| \boldsymbol{\Omega}_t^{1/2} (\hat{\theta}_t - \mu_t) \right\|_2 &\leq D_3 M_n \sqrt{t^{-1} p_*}, & \lambda_{\min} \left(\tilde{\mathbf{F}}_{t,\theta_t^*} \right) &\geq D_4 n t \end{aligned} \quad (218)$$

on an event \mathcal{E} . Assume further that, on \mathcal{E} ,

$$\left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \boldsymbol{\Omega}_t \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_2 \leq \frac{1}{4}, \quad \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \tilde{\mathbf{F}}_{t,\hat{\theta}_t} \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} - \mathbf{I}_p \right\|_2 \leq \frac{1}{8}, \quad \|\theta_t^* - \theta_0\|_2 \leq \frac{1}{4}, \quad (219)$$

and

$$n \geq C M_n^2 t^{2\alpha-1} p_*, \quad (220)$$

for a large enough constant $C = C(K_{\min}, K_{\max}, D_1, D_2, D_3, D_4)$. Then, on \mathcal{E} ,

$$\left\| \mathbf{F}_{t+1,\theta_t^*}^{1/2} (\theta_{t+1}^* - \theta_t^*) \right\|_2 \leq K M_n t^{\alpha-1/2} \sqrt{p_*},$$

where $K = K(D_1, D_2, D_3, D_4, K_{\min}, K_{\max})$.

Proof In this proof, we work on the event \mathcal{E} without explicitly referring it, and assume that $C = C(K_{\min}, K_{\max}, D_1, D_2, D_3, D_4)$ in (220) is sufficiently large. Note that

$$\begin{aligned} &\left\| \boldsymbol{\Omega}_t^{1/2} (\theta_t^* - \mu_t) \right\|_2 \leq \left\| \boldsymbol{\Omega}_t^{1/2} (\hat{\theta}_t - \theta_t^*) \right\|_2 + \left\| \boldsymbol{\Omega}_t^{1/2} (\hat{\theta}_t - \mu_t) \right\|_2 \\ &\leq \left\| \boldsymbol{\Omega}_t^{1/2} \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \right\|_2 \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{1/2} \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \right\|_2 \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\hat{\theta}_t - \theta_t^*) \right\|_2 + \left\| \boldsymbol{\Omega}_t^{1/2} (\hat{\theta}_t - \mu_t) \right\|_2 \\ &\leq \left(1 + \left\| \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} \boldsymbol{\Omega}_t \tilde{\mathbf{F}}_{t,\hat{\theta}_t}^{-1/2} - \mathbf{I}_p \right\|_2 \right)^{1/2} \left(1 + \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \tilde{\mathbf{F}}_{t,\hat{\theta}_t} \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} - \mathbf{I}_p \right\|_2 \right)^{1/2} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\hat{\theta}_t - \theta_t^*) \right\|_2 \\ &\quad + \left\| \boldsymbol{\Omega}_t^{1/2} (\hat{\theta}_t - \mu_t) \right\|_2 \\ &\stackrel{(219)}{\leq} \frac{3}{2} \left\| \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\hat{\theta}_t - \theta_t^*) \right\|_2 + \left\| \boldsymbol{\Omega}_t^{1/2} (\hat{\theta}_t - \mu_t) \right\|_2 \\ &\stackrel{(218)}{\leq} \frac{3}{2} \left[D_2 M_n \sqrt{t^{-1} p_*} + \frac{2}{3} D_3 M_n \sqrt{t^{-1} p_*} \right] \leq 3(D_2 \vee D_3) M_n \sqrt{t^{-1} p_*}. \end{aligned}$$

Also, by Taylor's theorem and $\nabla \mathbb{E}_{t+1}L_{t+1}(\theta_0) = 0$, we have

$$\begin{aligned} \nabla \mathbb{E}_{t+1}L_{t+1}(\theta_t^*) &= \nabla \mathbb{E}_{t+1}L_{t+1}(\theta_t^*) - \nabla \mathbb{E}_{t+1}L_{t+1}(\theta_0) = -\bar{\mathbf{F}}_{t+1}(\theta_t^*, \theta_0) (\theta_t^* - \theta_0) \\ &= -\bar{\mathbf{F}}_{t+1}(\theta_t^*, \theta_0) \tilde{\mathbf{F}}_{t,\theta_t^*}^{-1/2} \tilde{\mathbf{F}}_{t,\theta_t^*}^{1/2} (\theta_t^* - \theta_0) \end{aligned}$$

where

$$\bar{\mathbf{F}}_{t+1}(\theta_t^*, \theta_0) = - \int_0^1 \nabla^2 \mathbb{E}_{t+1} L_{t+1}(s\theta_t^* + (1-s)\theta_0) ds = - \int_0^1 \nabla^2 L_{t+1}(s\theta_t^* + (1-s)\theta_0) ds.$$

Note that $s\theta_t^* + (1-s)\theta_0 \in \Theta(\theta_0, \mathbf{I}_p, 1/2)$ for all $s \in [0, 1]$ because $\|\theta_t^* - \theta_0\|_2 \leq 1/4$. It follows that

$$\begin{aligned} & \left\| \mathbf{F}_{t+1, \theta_t^*}^{-1/2} \nabla \mathbb{E}_{t+1} L_{t+1}(\theta_t^*) \right\|_2 = \left\| \mathbf{F}_{t+1, \theta_t^*}^{-1/2} \bar{\mathbf{F}}_{t+1}(\theta_t^*, \theta_0) \tilde{\mathbf{F}}_{t, \theta_t^*}^{-1/2} \tilde{\mathbf{F}}_{t, \theta_t^*}^{1/2} (\theta_t^* - \theta_0) \right\|_2 \\ & \leq \left\| \mathbf{F}_{t+1, \theta_t^*}^{-1/2} \bar{\mathbf{F}}_{t+1}(\theta_t^*, \theta_0)^{1/2} \right\|_2 \left\| \bar{\mathbf{F}}_{t+1}(\theta_t^*, \theta_0)^{1/2} \tilde{\mathbf{F}}_{t, \theta_t^*}^{-1/2} \right\|_2 \left\| \tilde{\mathbf{F}}_{t, \theta_t^*}^{1/2} (\theta_t^* - \theta_0) \right\|_2 \\ & \leq \lambda_{\min}^{-1/2}(\mathbf{F}_{t+1, \theta_t^*}) \lambda_{\max}(\bar{\mathbf{F}}_{t+1}(\theta_t^*, \theta_0)) \lambda_{\min}^{-1/2}(\tilde{\mathbf{F}}_{t, \theta_t^*}) \left\| \tilde{\mathbf{F}}_{t, \theta_t^*}^{1/2} (\theta_t^* - \theta_0) \right\|_2 \\ & \stackrel{(218), (\mathbf{A}2)}{\leq} (K_{\min} n)^{-1/2} (K_{\max} n) (D_4 n t)^{-1/2} (D_1 M_n t^\alpha \sqrt{p_*}) \\ & = (K_{\min}^{-1/2} D_4^{-1/2} K_{\max} D_1) (M_n t^{\alpha-1/2} \sqrt{p_*}). \end{aligned}$$

Since

$$\left\| \boldsymbol{\Omega}_t^{1/2} (\theta_t^* - \mu_t) \right\|_2 \leq c_1 M_n \sqrt{t^{-1} p_*}, \quad \left\| \mathbf{F}_{t+1, \theta_t^*}^{-1/2} \nabla \mathbb{E}_{t+1} L_{t+1}(\theta_t^*) \right\|_2 \leq c_2 M_n t^{\alpha-1/2} \sqrt{p_*}$$

for some positive constants $c_1 = c_1(D_2, D_3)$ and $c_2 = c_2(K_{\min}, K_{\max}, D_1, D_4)$, combining (220) and the assumption $\|\theta_t^* - \theta_0\|_2 \leq 1/4$, we can utilize Lemma 47. Therefore, Lemma 47 gives

$$\left\| \mathbf{F}_{t+1, \theta_t^*}^{1/2} (\theta_{t+1}^* - \theta_t^*) \right\|_2 \leq \left(4 \left\| \mathbf{F}_{t+1, \theta_t^*}^{-1/2} \nabla \mathbb{E}_{t+1} L_{t+1}(\theta_t^*) \right\|_2 \right) \vee \left(2 \left\| \boldsymbol{\Omega}_t^{1/2} (\theta_t^* - \mu_t) \right\|_2 \right),$$

which further upper bounded by

$$(4c_2 M_n t^{\alpha-1/2} \sqrt{p_*}) \vee (2c_1 M_n \sqrt{t^{-1} p_*}) \leq (2c_1 + 4c_2) M_n t^{\alpha-1/2} \sqrt{p_*},$$

which completes the proof. \blacksquare

H.4 Deviation bounds for (sub-) Gaussian random vectors

Lemma 49 For $\mathbf{A} \in \mathcal{S}_{++}^p$ and $\mathbf{B} \in \mathcal{S}_{++}^p$, let $Z \sim \mathcal{N}(0, \mathbf{A}^{-1})$ and $\boldsymbol{\Omega} = \mathbf{A}^{-1/2} \mathbf{B} \mathbf{A}^{-1/2}$. Then, for every $\omega \geq 0$,

$$\mathbb{P} \left(\left| \langle \mathbf{B} Z, Z \rangle - \text{tr}(\boldsymbol{\Omega}) \right| > 2 \|\boldsymbol{\Omega}\|_{\text{F}} \sqrt{\omega} + 2 \|\boldsymbol{\Omega}\|_2 \omega \right) \leq 2e^{-\omega}.$$

Furthermore,

$$\mathbb{P} \left(\left\| \mathbf{B}^{1/2} Z \right\|_2 > \sqrt{\text{tr}(\boldsymbol{\Omega})} + \sqrt{2 \|\boldsymbol{\Omega}\|_2 \omega} \right) \leq e^{-\omega}.$$

Proof See Theorem B.4 in Spokoiny (2024). \blacksquare

Lemma 50 *Let $\mathbf{A} \in \mathbb{R}^{p \times p} \succeq 0$, and $Z = (Z_i)_{i=1}^n \in \mathbb{R}^n \sim \text{SubG}(\sigma^2)$ be a random vector whose components are independent with $\mathbb{E}Z = 0$. Then, for every $\omega \geq 0$,*

$$\mathbb{P}\left(\langle \mathbf{A}Z, Z \rangle > \sigma^2 \left[\text{tr}(\mathbf{A}) + 2\sqrt{\text{tr}(\mathbf{A}^2)\omega} + 2\|\mathbf{A}\|_2 \omega \right]\right) \leq e^{-\omega}.$$

Proof See Theorem 1 in Hsu et al. (2012). ■

Lemma 51 *Let $Z \sim \mathcal{N}(0, \mathbf{I}_p)$ and $\mathbf{A} \in \mathcal{S}_{++}^p$. Then,*

$$\mathbb{E}\langle \mathbf{A}Z, Z \rangle^4 \leq (\text{tr}(\mathbf{A}) + 3\|\mathbf{A}\|_2)^4.$$

Proof See Theorem B.1 in Spokoiny (2024). ■

Lemma 52 *Let $(Z_i)_{i \in [n]}$ be i.i.d. copies of an isotropic and log-concave probability measure on \mathbb{R}^p . Suppose that $n \geq p$. Then, for any $k \geq 2$, there exists some constants $K_1 = K_1(k) > 1$, $K_2 > 0$ and $K_3 = K_3(k) > 0$ such that*

$$\begin{aligned} & \sup_{u \in \mathbb{R}^p: \|u\|_2=1} \left| \frac{1}{n} \sum_{i=1}^n \left(|\langle Z_i, u \rangle|^k - \mathbb{E}|\langle Z_i, u \rangle|^k \right) \right| \\ & \leq K_1 t s^{k-1} \log^{k-1} \left(\frac{2n}{p} \right) \sqrt{\frac{p}{n}} + K_1 \frac{s^k p^{k/2}}{n} + K_1 \left(\frac{p}{2n} \right)^s, \quad \forall s, t \geq 1 \end{aligned}$$

with probability at least

$$1 - \exp(-K_2 s \sqrt{p}) - \exp\left(-K_3 \left\{ \left[t^2 s^{2k-2} p \log^{(2k-2)} \left(\frac{2n}{p} \right) \right] \wedge \left[t s^{-1} \sqrt{np} \log^{-1} \left(\frac{2n}{p} \right) \right] \right\}\right).$$

Proof See Proposition 4.4 in Adamczak et al. (2010). ■

H.5 3-order Gaussian tensor analysis

For $\theta, u \in \Theta$ and a three times differentiable function $f : \Theta \rightarrow \mathbb{R}$, let

$$\begin{aligned} \mathcal{R}_{3,f}(\theta, u) &= f(\theta + u) - f(\theta) - \langle \nabla f(\theta), u \rangle - \frac{1}{2} \langle \nabla^2 f(\theta), u^{\otimes 2} \rangle, \\ \mathcal{R}_{4,f}(\theta, u) &= f(\theta + u) - f(\theta) - \langle \nabla f(\theta), u \rangle - \frac{1}{2} \langle \nabla^2 f(\theta), u^{\otimes 2} \rangle - \frac{1}{6} \langle \nabla^3 f(\theta), u^{\otimes 3} \rangle. \end{aligned} \tag{221}$$

For a 3-order symmetric tensor $\mathbf{T} = (T_{ijk})_{i,j,k \in [p]} \in \mathbb{R}^{p \times p \times p}$, let

$$\mathbf{T}(u) = \langle \mathbf{T}, u^{\otimes 3} \rangle, \quad \mathbf{T}_i = (T_{ijk})_{j,k \in [p]} \in \mathbb{R}^{p \times p}, \quad \|\mathbf{T}\|_F^2 = \sum_{i,j,k \in [p]} T_{ijk}^2.$$

The following lemmas are from the Section B.7.1 in Spokoiny (2024). We reproduce them here for the sake of readability and completeness of proof.

Lemma 53 For a symmetric 3-order tensor $\mathbf{T} \in \mathbb{R}^{p \times p \times p}$, let $Z = (Z_i)_{i \in [p]} \sim \mathcal{N}(0, \mathbf{I}_p)$ and $M = (M_i)_{i \in [p]}$, where $M_i = \text{tr}(\mathbf{T}_i)$. Then,

$$\mathbb{E} \left(\mathbf{T}(Z) - 3 \langle M, Z \rangle \right)^2 = 6 \|\mathbf{T}\|_{\mathbf{F}}^2, \quad \mathbb{E} \mathbf{T}^2(Z) = 6 \|\mathbf{T}\|_{\mathbf{F}}^2 + 9 \|M\|_2^2.$$

Proof See Lemma B.32 in Spokoiny (2024). ■

Lemma 54 For a symmetric 3-order tensor $\mathbf{T} \in \mathbb{R}^{p \times p \times p}$, suppose that there exist some $\mathbf{F} \in \mathcal{S}_{++}^p$ and $\tau_3 \geq 0$ such that

$$\mathbf{T}(u) = \langle \mathbf{T}, u^{\otimes 3} \rangle \leq \tau_3 \|\mathbf{F}u\|_2^3, \quad \forall u \in \mathbb{R}^p.$$

Let $\tilde{Z} \sim \mathcal{N}(0, \mathbf{D}^{-1})$ for some $\mathbf{D} \in \mathcal{S}_{++}^p$ and $\mathbf{V} = \mathbf{D}^{-1/2} \mathbf{F} \mathbf{D}^{-1/2}$. Then,

$$\mathbb{E}[\{\mathbf{T}(\tilde{Z})\}^2] \leq 15\tau_3^2 \|\mathbf{V}\|_2 \text{tr}^2(\mathbf{V})$$

Proof See Lemma B.36 in Spokoiny (2024). ■

Lemma 55 For a three times differentiable function $f : \Theta \rightarrow \mathbb{R}$ and $\theta \in \Theta$, suppose that f satisfies the third order smoothness at θ with parameters (τ_3, \mathbf{F}, r) , where $\mathbf{F} \in \mathcal{S}_{++}^p$ and $\tau_3, r \geq 0$. Let $\tilde{Z} \sim \mathcal{N}(0, \mathbf{D}^{-1})$ for some $\mathbf{D} \in \mathcal{S}_{++}^p$ and $\mathbf{V} = \mathbf{D}^{-1/2} \mathbf{F} \mathbf{D}^{-1/2}$. Then, for a random variable G with $|G| \leq 1$,

$$\left| \mathbb{E} \left(e^X - 1 - X - \frac{X^2}{2} \right) G \right| \leq \frac{5}{3} \epsilon^3 \exp(\epsilon^2),$$

where

$$X = \mathcal{R}_{3,f}(\theta, \tilde{Z}) - \mathbb{E} \left[\mathcal{R}_{3,f}(\theta, \tilde{Z}) \mathbf{1}_{\Theta(\mathbf{F}, r)}(\tilde{Z}) \right], \quad \epsilon = \tau_3 \|\mathbf{V}\|_2 r^2 / 2$$

Proof Combining with Lemma B.42 in Spokoiny (2024), this lemma is a special case of Lemma B.39 in Spokoiny (2024). The proof can be found therein. ■