Understanding Expert Structures on Minimax Parameter Estimation in Contaminated Mixture of Experts

Fanqi Yan[⋆,⋄]   **Huy Nguyen**[⋆,†]   **Dung Le**[⋆,†]   **Pedram Akbarian**[‡]   **Nhat Ho**[†]

[⋄]Department of Computer Science

[†]Department of Statistics and Data Sciences

[‡]Department of Electrical and Computer Engineering

The University of Texas at Austin

Abstract

We conduct the convergence analysis of parameter estimation in the contaminated mixture of experts. This model is motivated from the prompt learning problem where ones utilize prompts, which can be formulated as experts, to fine-tune a large-scale pre-trained model for learning downstream tasks. There are two fundamental challenges emerging from the analysis: (i) the proportion in the mixture of the pre-trained model and the prompt may converge to zero during the training, leading to the prompt vanishing issue; (ii) the algebraic interaction among parameters of the pre-trained model and the prompt can occur via some partial differential equations and decelerate the prompt learning. In response, we introduce a distinguishability condition to control the previous parameter interaction. Additionally, we also investigate various types of expert structure to understand their effects on the convergence behavior of parameter estimation. In each scenario, we provide comprehensive convergence rates of parameter estimation along with the corresponding minimax lower bounds. Finally, we run several numerical experiments to empirically justify our theoretical findings.

1 INTRODUCTION

Originally introduced by Jacobs et al. (1991), mixture of experts (MoE) has been widely used as a statistical machine learning framework to integrate the power of several sub-models based on the principle of divide and conquer. In particular, it consists of two main components, namely multiple experts formulated as either regression functions (Ho et al., 2022) or classifiers (Nguyen et al., 2024b), each of which is responsible for handling a few specific tasks; and a gating function softly partitioning the input space into regions where corresponding specialized experts are assigned larger weights than the others. Thanks to its flexibility, there is a surge of interests in utilizing MoE model for various applications, including natural language processing (Jiang et al., 2024; Zhou et al., 2023; Fedus et al., 2022; Puigcerver et al., 2024; Du et al., 2022; Pham et al., 2024), computer vision (Riquelme et al., 2021; Liang et al., 2022; Ruiz et al., 2021), autonomous driving (Pini et al., 2023), multimodal fusion (Han et al., 2024), reinforcement learning (Chow et al., 2023; Obando Ceron et al., 2024) and others (Li et al., 2023; Hazimeh et al., 2021; Gupta et al., 2022; Ng and Carley, 2023). Recently, practitioners have also incorporated MoE models into parameter-efficient fine-tuning methods for large-scale pre-trained models (Zadouri et al., 2024; Shen et al., 2024; Le et al., 2024). More specifically, they attach small trainable experts called prompts to a frozen pre-trained model, which we refer to as a *contaminated MoE* in the sequel. By doing so, they are able to learn downstream tasks efficiently without having to train the whole model expensively. Nevertheless, the potentials for further enhancing this fine-tuning technique have been restricted owing to the absence of the theoretical foundation for the contaminated MoE model. Thus, our main objective in this work is to provide a comprehensive convergence analy-

sis of parameter estimation under the following probabilistic setting of the contaminated MoE to potentially shed some light on both the prompt convergence behavior and the prompt design.

Problem setting. To begin with, we assume that $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$ are i.i.d. samples of size n drawn from a contaminated MoE model whose conditional density function is given by

$$p_{\lambda^*, G_*}(Y|X) := (1 - \lambda^*) f_0(Y|\varphi(a_0^\top X + b_0), \nu_0) \\ + \lambda^* f(Y|\sigma((a^*)^\top X + b^*), \nu^*). \quad (1)$$

Above, $f_0(\cdot|\mu_0, \nu_0)$ is a known probability density function with mean μ_0 and variance ν_0, which represents for the pre-trained model. Meanwhile, $f(\cdot|\mu, \nu)$ is the univariate Gaussian density function with mean μ and variance ν, standing for the prompt. Next, $\lambda^* \in [0, 1]$ denotes a mixing proportion, $G_0 := (a_0, b_0, \nu_0)$ is a set of known parameters of the pre-trained model belonging to the parameter space $\Theta \subset \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^+$, whereas $G_* := (a^*, b^*, \nu^*) \in \Theta$ is a true yet unknown parameter set of the prompt. For the sake of presentation, we suppress the dependence of G_* on the sample size n. In addition, $X_1, X_2, \ldots, X_n$ are i.i.d. samples from some continuous probability distribution with the density function denoted by $\bar{f}(X)$. Lastly, note that the expectation of the response variable Y conditioning on the covariate (input) X is a mixture of experts $\varphi(a_0^\top X + b_0)$ and $\sigma((a^*)^\top X + b^*)$, that is,

$$\mathbb{E}[Y|X] = (1 - \lambda^*)\varphi(a_0^\top X + b_0) + \lambda^* \sigma((a^*)^\top X + b^*),$$

where $\varphi, \sigma : \mathbb{R} \to \mathbb{R}$ are two given real-valued activation functions. To ensure the dependence of expert values on the input X as in practice, we assume that $a^*, a_0 \neq 0_d$ and the first derivative of σ is different from zero in a non-zero measure set in $\mathbb{R}$ to eliminate the trivial case when the expert function σ is a constant.

Maximum likelihood estimation (MLE). Subsequently, to estimate the unknown parameters λ^* and $G_* = (a^*, b^*, \nu^*)$ of model (1), we leverage the maximum likelihood method (van de Geer, 2000) as follows:

$$(\widehat{\lambda}_n, \widehat{G}_n) \in \arg\max_{(\lambda, G) \in \Xi} \sum_{i=1}^n \log(p_{\lambda, G}(Y_i|X_i)), \quad (2)$$

where $\widehat{G}_n := (\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n)$ and $\Xi := [0, 1] \times \Theta$. In order to guarantee the convergence of parameter estimation, we assume that the parameter space Θ is compact, whereas the input space $\mathcal{X}$ is bounded. Note that these assumptions are mild and commonly employed in previous works (Ho et al., 2022; Do et al., 2022; Nguyen et al., 2024d).

Related works. There are early attempts to comprehend the properties of MoE models from the perspective of the parameter estimation problem. Firstly, Ho et al. (2022) investigated the convergence behavior of MLE under the covariate-free gating Gaussian MoE model. They discovered that the convergence of parameter estimation might be negatively affected by parameter interactions (via some partial differential equations), depending on the algebraic independence between expert functions. In particular, they showed that parameter estimation rates were determined by the solvability of a system of polynomial equations, which became significantly slow when the number of fitted parameters increased. Then, Do et al. (2022) generalized these results to the settings where the conditional distribution of the response variable given the covariate was a mixture of distributions other than Gaussians. Lastly, Nguyen et al. (2024d) explored a deviated Gaussian MoE where they attached a mixture of Gaussian distribution to the pre-trained model f_0 rather than a single Gaussian as in equation (1). However, they took only linear experts into account, i.e. the activation σ was an identity function. Furthermore, they also let the ground-truth parameters be independent of the sample size n, which simplified the problem considerably and induced only pointwise parameter estimation rates. Therefore, we aim to examine the impacts of various expert structures on the uniform convergence rates of parameter estimation as well as establish the corresponding minimax lower bounds for those rates in this work.

Challenges. There are two fundamental challenges arising from the dependence of ground-truth parameters λ^* and G_* on the sample size n:

(i) Prompt vanishing. The mixing proportion λ^* may converge to zero as n goes to infinity, making the prompt disappear from the contaminated MoE model (1). Consequently, the prompt's effects on fine-tuning the pre-trained model would become minor. Thus, it would be difficult to characterize the convergence behavior of estimation for the prompt parameters G_*.

(ii) Prompt merging. When f_0 belongs to the family of Gaussian densities and the two expert functions φ and σ are identical, that is, $\varphi = \sigma$, if the prompt parameters G_* converge to those of the pre-trained model G_0 as n goes to infinity, then the prompt in equation (1) will merge into the pre-trained model. This phenomenon not only hinders the ability of the contaminated MoE to learn downstream tasks but also induces several obstacles in capturing the convergence behavior of estimation for the mixing proportion λ^*.

Contributions. To address these challenges, we con-

Fanqi Yan[*,◇], Huy Nguyen[*,†], Dung Le[*,†], Pedram Akabrian[‡], Nhat Ho[†]

Table 1: Summary of the convergence rates of parameter estimation in the contaminated MoE model. Below, we denote $\Delta G_* := (a^* - a_0, b^* - b_0, \nu^* - \nu_0)$ and $\vartheta_{\Delta G_*}^{\alpha,\beta,\gamma} := \|a^* - a_0\|^\alpha + |b^* - b_0|^\beta + |\nu^* - \nu_0|^\gamma$. Additionally, $f_0 = f$ means that f_0 is a Gaussian density, while $f_0 \neq f$ indicates the opposite.

σ	f_0	φ	Thm.	$|\widehat{\lambda}_n - \lambda^*|$	$\|\widehat{a}_n - a^*\|, |\widehat{\nu}_n - \nu^*|$	$|\widehat{b}_n - b^*|$
linear	$f_0 \neq f$	–	2	$\widetilde{\mathcal{O}}(n^{-\frac{1}{2}})$	$\widetilde{\mathcal{O}}(n^{-\frac{1}{2}}(\lambda^*)^{-1})$	
	$f_0 = f$	linear	4	$\widetilde{\mathcal{O}}((n \cdot \vartheta_{\Delta G_*}^{4,8,4})^{-\frac{1}{2}})$	$\widetilde{\mathcal{O}}((n \cdot \vartheta_{\Delta G_*}^{2,4,2})^{-\frac{1}{2}}(\lambda^*)^{-1})$	$\widetilde{\mathcal{O}}((n \cdot \vartheta_{\Delta G_*}^{2,4,2})^{-\frac{1}{4}}(\lambda^*)^{-\frac{1}{2}})$
		non-linear	6	$\widetilde{\mathcal{O}}(n^{-\frac{1}{2}})$	$\widetilde{\mathcal{O}}(n^{-\frac{1}{2}}(\lambda^*)^{-1})$	
non-linear	$f_0 \neq f$	–	7	$\widetilde{\mathcal{O}}(n^{-\frac{1}{2}})$	$\widetilde{\mathcal{O}}(n^{-\frac{1}{2}}(\lambda^*)^{-1})$	
	$f_0 = f$	linear	8	$\widetilde{\mathcal{O}}(n^{-\frac{1}{2}})$	$\widetilde{\mathcal{O}}(n^{-\frac{1}{2}}(\lambda^*)^{-1})$	
		non-linear	9	$\widetilde{\mathcal{O}}(n^{-\frac{1}{2}} \cdot \|\Delta G_*\|^{-2})$	$\widetilde{\mathcal{O}}(n^{-\frac{1}{2}} \cdot (\|\Delta G_*\|\lambda^*)^{-1})$	

sider various structures of the expert functions φ and σ (whether they are linear or non-linear) and different families of the pre-trained model f_0 (whether it belongs to the Gaussian density family). Our contributions are three-fold and can be summarized as follows:

1. Uniform convergence rates: In each of the aforementioned scenarios, we establish the corresponding convergence rates of the mixing proportion ($\widehat{\lambda}_n$ to λ^*) and the prompt parameters (($\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n$) to ($a^*, b^*, \nu^*$)). Notably, we find that whether the proportion rate depends on the prompt parameter rates or not varies with the expert structures and the pretrained model family. By contrast, the latter rates always hinge upon the former rate under every scenario (see also Table 1).

2. Minimax lower bounds: To demonstrate that the above uniform convergence rates of parameter estimation are optimal, we also derive the associated minimax lower bound with each of those rates.

3. Theoretical insights into fine-tuning method: Finally, from the results in our theoretical analysis, we provide several useful insights into designing prompts to improve the efficiency of fine-tuning techniques.

Organization. The rest of the paper proceeds as follows. In Section 2, we introduce some necessary concepts for our analysis and capture the density estimation rates. Then, we present the convergence rates of parameter estimation along with their corresponding minimax lower bounds when the expert σ is linear and non-linear in Section 3 and Section 4, respectively. Next, we conduct numerical experiments to justify our theoretical results in Section 5. Lastly, we discuss and conclude the paper in Section 6. Additional experiments and full proofs are defer to the Appendices.

Notations. Firstly, we let $[n] := \{1, 2, \ldots, n\}$ for any $n \in \mathbb{N}$. Next, for any vector of real numbers u we denote $\|u\|$ as its Euclidean norm value. Given any two positive sequences $(a_n)_{n \geq 1}$ and $(b_n)_{n \geq 1}$, we write

$a_n = \mathcal{O}(b_n)$ or $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for all $n \in \mathbb{N}$, where $C > 0$ is some universal constant. Besides, the notation $a_n = \widetilde{\mathcal{O}}(b_n)$ indicates the previous inequality may depend on some logarithmic term. Lastly, for any two densities p and q (dominated by the Lebesgue measure m), their squared Hellinger distance is computed as $h^2(p, q) := \frac{1}{2} \int [\sqrt{p(x)} - \sqrt{q(x)}]^2 dm(x)$.

2 PRELIMINARIES

We start this section with the following *distinguishability condition* to control the merging of the prompt into the pre-trained model mentioned in Section 1.

Definition 1 (Distinguishability). *We say that f is distinguishable from f_0 if the following holds: Given any two distinct components $(a_i, b_i, \nu_i) \in \Theta$, for $i = 1, 2$, if we have real coefficients η_i for $0 \leq i \leq 2$ such that $\eta_1 \eta_2 \leq 0$ and*

$$\eta_0 f_0(Y|\varphi(a_0^\top X + b_0), \nu_0)$$
$$+ \sum_{i=1}^2 \eta_i f(Y|\sigma((a_i)^\top X + b_i), \nu_i) = 0,$$

for almost surely $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, then $\eta_i = 0, \forall i$.

Example. If f_0 belongs to the family of Student's t-distribution, we can verify that f is distinguishable from f_0. By contrast, if f_0 is a Gaussian density and $\varphi = \sigma$, then f is not distinguishable from f_0.

Note that if f is distinguishable from f_0, the merging $f_0(Y|\varphi(a_0^\top X + b_0), \nu_0) = f(Y|\sigma((a_i)^\top X + b_i), \nu_i)$, for almost surely $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, cannot occur as it violates the distinguishability condition. Furthermore, according to Proposition 1, whose proof is in Appendix D.2, this condition also ensures that the contaminated MoE model (1) is identifiable.

Proposition 1 (Identifiability). *Let $(\lambda, G), (\lambda', G')$ be two components in Ξ. Suppose that f is distinguishable from f_0, then if the identifiability equa-*

tion $p_{\lambda,G}(Y|X) = p_{\lambda',G'}(Y|X)$ holds for almost surely $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, we obtain that $(\lambda, G) = (\lambda', G')$.

Subsequently, we proceed to characterize the convergence behavior of density estimation in Theorem 1.

Theorem 1. *Assume that the density f_0 is bounded with tail $\mathbb{E}_X \left[-\log f_0(Y|\varphi(a_0^\top X + b_0), \nu_0) \right] \gtrsim Y^q$, for almost surely $Y \in \mathcal{Y}$ for some $q > 0$. Then, there exists a constant $C > 0$ depending on f_0, λ^*, G_* and Ξ such that the following holds true:*

$$\sup_{(\lambda^*, G_*) \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} h(p_{\widehat{\lambda}_n, \widehat{G}_n}, p_{\lambda^*, G_*}) \leq C\sqrt{\log n / n}.$$

Proof of Theorem 1 is in Appendix C. The above bound shows that the density estimator $p_{\widehat{\lambda}_n, \widehat{G}_n}$ converges to the true density p_{λ^*, G_*} at the parametric rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$. Moreover, if we can construct some parameter loss function $D((\lambda, G), (\lambda^*, G_*))$ that satisfies $D((\widehat{\lambda}_n, \widehat{G}_n), (\lambda^*, G_*)) \lesssim h(p_{\widehat{\lambda}_n, \widehat{G}_n}, p_{\lambda^*, G_*})$, we achieve our desired parameter estimation rates (see Appendix A for further details). Finally, to utilize Theorem 1, we assume that the density f_0 is bounded with the tail $\mathbb{E}_X \left[-\log f_0(Y|\varphi(a_0^\top X + b_0), \nu_0) \right] \gtrsim Y^q$, for almost surely $Y \in \mathcal{Y}$ for some $q > 0$, throughout this paper unless stating otherwise.

3 LINEAR EXPERT σ

In this section, we determine the convergence rates of parameter estimation when the expert σ is linear, that is, it takes the form $\sigma(z) = a'z + b'$ for any $z \in \mathbb{R}$, where a', b' are known constants. Since the rates vary with the family of the density f_0, we further consider two scenarios when f_0 is and is not a Gaussian density.

3.1 When f_0 is not a Gaussian density

We begin with the simpler scenario when f_0 is not a Gaussian density as we can verify that f is distinguishable from f_0 under this scenario in Proposition 2.

Proposition 2. *If f_0 does not belong to the family of Gaussian densities, then f is distinguishable from f_0.*

Proof of Proposition 2 is in Appendix D.3. Given this result, we are now ready to capture the convergence of parameter estimation in the following theorem:

Theorem 2 (MLE rates). *When the expert function σ is of linear form and f_0 is not a Gaussian density, we obtain the MLE convergence rates as follows:*

$$\sup_{(\lambda^*, G_*) \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} \left[|\widehat{\lambda}_n - \lambda^*|^2 \right] \lesssim \frac{\log n}{n},$$

$$\sup_{(\lambda^*, G_*) \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} \left[(\lambda^*)^2 \|\widehat{G}_n - G_*\|^2 \right] \lesssim \frac{\log n}{n}.$$

The proof of Theorem 2 is in Appendix A.1. The first bound indicates that the mixing proportion estimator $\widehat{\lambda}_n$ converges to the ground-truth value λ^* at the rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$, which is parametric on the sample size n. On the other hand, it follows from the second bound that the convergence rate of the prompt parameter estimation $\widehat{G}_n = (\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n)$ to its true counterpart $G_* = (a^*, b^*, \nu^*)$ is slower than $\widetilde{\mathcal{O}}(n^{-1/2})$ as it hinges upon the vanishing rate of λ^*. This rate dependence reflects the prompt vanishing issue mentioned in the "Challenges" paragraph in Section 1.

Next, we establish the minimax lower bounds for estimating the true parameters (λ^*, G_*) in Theorem 3.

Theorem 3 (Minimax lower bounds). *Under the setting of Theorem 2, we have for any $0 < r < 1$ that*

$$\inf_{(\overline{\lambda}_n, \overline{G}_n) \in \Xi} \sup_{(\lambda, G) \in \Xi} \mathbb{E}_{p_{\lambda, G}} \left(|\overline{\lambda}_n - \lambda|^2 \right) \gtrsim n^{-1/r},$$

$$\inf_{(\overline{\lambda}_n, \overline{G}_n) \in \Xi} \sup_{(\lambda, G) \in \Xi} \mathbb{E}_{p_{\lambda, G}} \left(\lambda^2 \|\overline{G}_n - G\|^2 \right) \gtrsim n^{-1/r}.$$

Proof of Theorem 3 is in Appendix B.1. The results of this theorem imply that the convergence rates of the MLE $(\widehat{\lambda}_n, \widehat{G}_n)$ provided in Theorem 2 are optimal.

3.2 When f_0 is a Gaussian density

Next, we consider the scenario when f_0 belongs to the family of Gaussian densities. Under this scenario, it can be checked that f might not be distinguishable from f_0, depending on the structure of the expert φ in the pre-trained model. To this end, we continue to divide the analysis into two smaller scenarios where the function φ is linear and non-linear, respectively.

3.2.1 Linear expert φ

Recall that in this case, both the experts φ and σ are of linear forms, and the density $f_0 = f$ is a Gaussian density. Notably, this setting induces two potential obstacles in our analysis.

Firstly, the prompt might merge into the pre-trained model. In particular, let us assume that $\varphi(a_0^\top X + b_0) = \alpha_0(a_0^\top X + b_0) + \beta_0$ and $\sigma((a^*)^\top X + b^*) = \alpha^*((a^*)^\top X + b^*) + \beta^*$, where $\alpha_0, \alpha^* \neq 0$ and β_0, β^* are some known constants. Then, if $a^* \to \frac{\alpha_0}{\alpha^*} a_0$ and $b^* \to \frac{\alpha_0 b_0 + \beta_0 - \beta^*}{\alpha^*}$ and $\nu^* \to \nu_0$ as $n \to \infty$, it follows that the expert $\sigma((a^*)^\top X + b^*)$ will converge to its counterpart $\varphi(a_0^\top X + b_0)$. Consequently, we can justify that f is not distinguishable from f_0. Furthermore, since the prompt $f(Y|\sigma((a^*)^\top X + b^*), \nu^*)$ will also converge to the pre-trained model $f_0(Y|\varphi(a_0^\top X + b_0), \nu_0)$, we have to cope with the prompt merging issue mentioned in the "Challenges" paragraph in Section 1 as well. This

issue will be demonstrated to make the convergence behavior of the mixing proportion estimator become complicated in Theorem 4. For simplicity, we will focus only on the case when φ and σ are identity functions, that is, when $\alpha_0 = \alpha^* = 1$ and $\beta_0 = \beta^* = 0$. Other cases can be naturally generalized based on the aforementioned limits of a^* and b^*.

Secondly, there is an interaction between parameters b and ν via the following heat equation:

$$\frac{\partial^2 f}{\partial b^2}(Y|a^\top X + b, \nu) = 2\frac{\partial f}{\partial \nu}(Y|a^\top X + b, \nu). \quad (3)$$

The derivation of this equation can be seen in Appendix D. Note that such interaction has been captured in prior works on Gaussian MoE (Ho et al., 2022; Nguyen et al., 2024e,a) where it is shown to decelerate the convergence rates of involved parameters significantly. However, to the best of our knowledge, the effects of that interaction has never been explored when the ground-truth parameter values λ^*, G_* depend on the sample size n. Thus, we will demistify this problem in Theorem 4.

For simplicity, let us denote $\Delta G := (\Delta a, \Delta b, \Delta \nu) := (a - a_0, b - b_0, \nu - \nu_0)$ for any component $(a, b, \nu) \in \Theta$.

Theorem 4 (MLE rates). *Suppose that f_0 is a Gaussian density and both the experts φ and σ are identity functions. For any sequence $(l_n)_{n \geq 1}$, we denote*

$$\Xi_1(l_n) := \left\{(\lambda, G) = (\lambda, a, b, \nu) \in \Xi : \right.$$
$$\left. \frac{l_n}{\min_{1 \leq i \leq d}\{|(\Delta a)_i|^2, |\Delta b|^4, |\Delta \nu|^2\}\sqrt{n}} \leq \lambda\right\}.$$

Then, the followings hold for any sequence $(l_n)_{n \geq 1}$ such that $l_n/\log n \to \infty$ as $n \to \infty$:

$$\sup_{(\lambda^*, G_*) \in \Xi_1(l_n)} \mathbb{E}_{p_{\lambda^*, G_*}}\left[(\|\Delta a^*\|^4 + |\Delta b^*|^8 + |\Delta \nu^*|^4)\right.$$
$$\left. \times |\widehat{\lambda}_n - \lambda^*|^2\right] \lesssim \frac{\log n}{n},$$

$$\sup_{(\lambda^*, G_*) \in \Xi_1(l_n)} \mathbb{E}_{p_{\lambda^*, G_*}}\left[(\lambda^*)^2(\|\Delta a^*\|^2 + |\Delta b^*|^4 + |\Delta \nu^*|^2)\right.$$
$$\left. \times (\|\widehat{a}_n - a^*\|^2 + |\widehat{b}_n - b^*|^4 + |\widehat{\nu}_n - \nu^*|^2)\right] \lesssim \frac{\log n}{n}.$$

The proof of Theorem 4 is in Appendix A.6. Above, we take the set $\Xi_1(l_n)$ into account to guarantee the consistency of the MLE estimators under this scenario. Next, there are two main implications from the results of Theorem 4:

(i) In contrast to Theorem 2, the first bound in Theorem 4 indicates that the convergence rate of the mixing proportion estimator $\widehat{\lambda}_n$ is now slower than $\widetilde{\mathcal{O}}(n^{-1/2})$ and is determined by that of the quantity

$(\|\Delta a^*\|^4 + |\Delta b^*|^8 + |\Delta \nu^*|^4)$. This rate deceleration occurs as a result of the prompt merging issue brought up at the beginning of Section 3.2.1.

(ii) Again, different from Theorem 2 where we observe that the prompt parameter estimators $\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n$ share the same convergence rate, the second bound in Theorem 4 points out a mismatch among their rates under this scenario. In particular, the convergence rate of $\widehat{b}_n$ is the squared root of those of $\widehat{a}_n$ and $\widehat{\nu}_n$. Moreover, all these rates depend on the vanishing rate of not only λ^* but also that of the term $(\|\Delta a^*\|^2 + |\Delta b^*|^4 + |\Delta \nu^*|^2)$. This intriguing phenomenon primarily arises from the parameter interaction through the heat equation (3).

Similar to Section 3.1, we also derive the corresponding minimax lower bounds for estimating the ground-truth parameters (λ^*, G_*) in Theorem 4.

Theorem 5 (Minimax lower bounds). *Under the setting of Theorem 4, we have for any $0 < r < 1$ and sequence $(l_n)_{n \geq 1}$ that*

$$\inf_{(\overline{\lambda}_n, \overline{G}_n) \in \Xi} \sup_{(\lambda, G) \in \Xi_1(l_n)} \mathbb{E}_{p_{\lambda, G}}\left[(\|\Delta a\|^4 + \|\Delta b\|^8 + \|\Delta \nu\|^4)\right.$$
$$\left. \times |\overline{\lambda}_n - \lambda|^2\right] \gtrsim n^{-1/r},$$

$$\inf_{(\overline{\lambda}_n, \overline{G}_n) \in \Xi} \sup_{(\lambda, G) \in \Xi_1(l_n)} \mathbb{E}_{p_{\lambda, G}}\left[\lambda^2(\|\Delta a\|^2 + \|\Delta b\|^4 + \|\Delta \nu\|^2)\right.$$
$$\left. \times (\|\overline{a}_n - a\|^2 + \|\overline{b}_n - b\|^4 + \|\overline{\nu}_n - \nu\|^2)\right] \gtrsim n^{-1/r}.$$

Proof of Theorem 5 is in Appendix B.2. The above minimax lower bounds indicate that the convergence rates of the MLE presented in Theorem 4 are optimal.

3.2.2 Non-linear expert φ

In this section, we draw our attention to the scenario where f_0 is a Gaussian density, the expert σ is of linear form, while φ is a non-linear expert. For instance, φ can be one among the activation functions sigmoid, ReLU and tanh, which are commonly used in practice.

Under this scenario, since the expert $\sigma((a^*)^\top X + b^*)$ cannot converge to its counterpart $\varphi(a_0^\top X + b_0)$ due to their structure distinction, the prompt $f(Y|\sigma((a^*)^\top X + b^*), \nu^*)$ does not converge to the pretrained model $f_0(Y|\varphi(a_0^\top X + b_0), \nu_0)$. Therefore, the prompt merging phenomenon does not take place as in Section 3.2.1 when φ is a linear expert. Additionally, we can also validate that f is distinguishable from f_0. As a consequence, we illustrate in the following theorem that the convergence behavior of the MLE in this case is identical to that in Section 3.1.

Theorem 6 (MLE rates). *Suppose that f_0 is a Gaussian density, the expert function σ is linear, while its counterpart φ is a non-linear expert. Then, we obtain*

the following MLE convergence rates:

$$\sup_{(\lambda^*, G_*) \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} \left[|\widehat{\lambda}_n - \lambda^*|^2 \right] \lesssim \frac{\log n}{n},$$

$$\sup_{(\lambda^*, G_*) \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} \left[(\lambda^*)^2 \| \widehat{G}_n - G_* \|^2 \right] \lesssim \frac{\log n}{n}.$$

The proof of Theorem 6 is in Appendix A.2. As mentioned above, since the two experts σ and φ have different structures (linear vs. non-linear), the issue that the prompt merges into the pre-trained model does not exist. Therefore, $\widehat{\lambda}_n$ converges to λ^* at the rate of order $\widetilde{\mathcal{O}}(n^{-1/2})$, which does not depend on any other factors. On the other hand, note that the prompt vanishing problem still holds when λ^* approaches zero. Thus, the convergence rate of $\widehat{G}_n = (\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n)$ is slower than $\widetilde{\mathcal{O}}(n^{-1/2})$ as it is governed by the vanishing rate of λ^*. Lastly, we can also arrive at the minimax lower bounds in Theorem 3 by arguing similarly. Consequently, the above MLE convergence rates are optimal.

4 NON-LINEAR EXPERT σ

Moving to this section, we will look into the convergence behavior of the MLE when the expert σ in the prompt takes a non-linear form such as sigmoid, ReLU and tanh functions, etc. Similar to Section 3, we also separate the analysis into two scenarios when f_0 is excluded from and included in the family of Gaussian densities, respectively.

4.1 When f_0 is not a Gaussian density

Firstly, when f_0 does not belong to the family of Gaussian densities, we can verify that f is distinguishable from f_0 by employing similar arguments for Proposition 2. Thus, the MLE admits the same convergence behavior as that in Section 3.1, which is exhibited in Theorem 7 whose proof is in Appendix A.3.

Theorem 7 (MLE rates). *When the expert function σ is of non-linear form and f_0 is not a Gaussian density, we obtain the MLE convergence rates as follows:*

$$\sup_{(\lambda^*, G_*) \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} \left[|\widehat{\lambda}_n - \lambda^*|^2 \right] \lesssim \frac{\log n}{n},$$

$$\sup_{(\lambda^*, G_*) \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} \left[(\lambda^*)^2 \| \widehat{G}_n - G_* \|^2 \right] \lesssim \frac{\log n}{n}.$$

Notably, it can be validated that the minimax lower bounds in Theorem 3 also hold under the setting of this section by employing the same arguments. Thus, we can claim that the above MLE rates are optimal.

4.2 When f_0 is a Gaussian density

Subsequently, we examine the convergence behavior of parameter estimation when f_0 belongs to the family of Gaussian densities. Since the distinguishability of the prompt f from the pre-trained model f_0 hinges upon the structure of the expert function φ, we proceed to divide the analysis into two smaller cases: when φ is a linear and non-linear expert function, respectively.

4.2.1 Linear expert φ

In this section, we concentrate on the scenario where the expert σ takes a non-linear form, while φ is a linear expert function.

Although this scenario is opposite to that in Section 3.2.2 where the expert σ is linear and its counterpart φ is non-linear, these two scenarios have one thing in common, which is the expert structure distinction. Therefore, it can be validated that they share several convergence properties. In particular, since the structures of the two experts σ and φ are different from each other, it is obvious that $\sigma((a^*)^\top X + b^*)$ cannot converge to $\varphi(a_0^\top X + b_0)$ as $n \to \infty$ for almost surely $X \in \mathcal{X}$. As a result, the prompt $f(Y|\sigma((a^*)^\top X + b^*), \nu^*)$ does not converge to the pre-trained model $f_0(Y|\varphi(a_0^\top X + b_0), \nu_0)$, either. This means that the expert structure difference helps tackle the prompt merging issue. Furthermore, it also guarantees that f is distinguishable from f_0. Consequently, we demonstrate in the following theorem that the convergence behavior of the MLE in this scenario resembles that in Section 3.2.2.

Theorem 8 (MLE rates). *Suppose that f_0 is a Gaussian density, the expert function σ is non-linear, while its counterpart φ is a linear expert. Then, we obtain the following MLE convergence rates:*

$$\sup_{(\lambda^*, G_*) \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} \left[|\widehat{\lambda}_n - \lambda^*|^2 \right] \lesssim \frac{\log n}{n},$$

$$\sup_{(\lambda^*, G_*) \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} \left[(\lambda^*)^2 \| \widehat{G}_n - G_* \|^2 \right] \lesssim \frac{\log n}{n}.$$

The proof of Theorem 8 is in Appendix A.4. Note that we can also derive the minimax lower bounds in Theorem 3 under the setting of Theorem 8 by arguing in a similar fashion and conclude that the above MLE convergence rates are optimal.

4.2.2 Non-linear expert φ

In this section, we take into account the last scenario when both the expert functions φ and σ are non-linear.

Since both φ and σ are non-linear expert functions, we observe that if σ is not equal to φ almost everywhere,

Fanqi Yan[⋆,◇]**, Huy Nguyen**[⋆,†]**, Dung Le**[⋆,†]**, Pedram Akabrian**[‡]**, Nhat Ho**[†]

then the fact $\sigma((a^*)^\top X + b^*)$ converges to $\varphi(a_0^\top X + b_0)$ as $n \to \infty$ for almost surely $X \in \mathcal{X}$ does not hold. Consequently, the prompt f become distinguishable from the the pre-trained model f_0, which allows us to demonstrate that the MLE convergence behavior in this case is totally similar to that in Theorem 8 by using the same arguments. Therefore, we will focus only on the challenging case when $\varphi = \sigma$ almost everywhere as the prompt might not be distinguishable from the pre-trained model in this situation, inducing the prompt merging issue if $(a^*, b^*, \nu^*) \to (a_0, b_0, \nu_0)$ when $n \to \infty$ as in Section 3.2.1.

On the other hand, thanks to the non-linear structures of the experts φ and σ, we exhibit in Proposition 3 a property of the prompt f that wipes out the interaction between parameters b and ν as in equation (3).

Proposition 3. *Suppose that the expert function* σ *is twice differentiable for almost everywhere. For any* $(a, b, \nu) \in \Theta$, *if we have a set of scalars* α_η, *for* $\eta = (\eta_1, \eta_2, \eta_3) \in \mathbb{N}^3$, *such that*

$$\sum_{\ell=0}^{2} \sum_{|\eta|=\ell} \alpha_\eta \cdot \frac{\partial^{|\eta|} f}{\partial a^{\eta_1} \partial b^{\eta_2} \partial \nu^{\eta_3}} (Y | \sigma(a^\top X + b), \nu) = 0$$

for almost surely $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, *then we get that* $\alpha_\eta = 0$ *for all* $\eta \in \mathbb{N}^3$ *such that* $|\eta| := \eta_1 + \eta_2 + \eta_3 \leq 2$.

The proof of Proposition 3 is in Appendix D.4. Given this result, we then illustrate in Theorem 9 that due to the parameter interaction disappearance, the convergence rate of parameter estimation is improved compared to that in Theorem 4 where the expert functions φ and σ also share the same structure but of linear form rather than non-linear form. Before presenting the theorem statement, let us recall that we denote $\Delta G := (\Delta a, \Delta b, \Delta \nu) := (a - a_0, b - b_0, \nu - \nu_0)$.

Theorem 9 (MLE rates). *Suppose that* f_0 *is a Gaussian density, the non-linear expert* σ *is twice differentiable and* $\sigma = \varphi$ *almost everywhere. For any sequence* $(l_n)_{n \geq 1}$, *we denote*

$$\Xi_2(l_n) := \left\{ (\lambda, G) = (\lambda, a, b, \nu) \in \Xi : \right.$$
$$\left. \frac{l_n}{\min_{1 \leq i \leq d} \{ |(\Delta a)_i|^2, |\Delta b|^2, |\Delta \nu|^2 \} \sqrt{n}} \leq \lambda \right\}.$$

Then, the followings hold for any sequence $(l_n)_{n \geq 1}$ *such that* $l_n / \log n \to \infty$ *as* $n \to \infty$:

$$\sup_{(\lambda^*, G_*) \in \Xi_2(l_n)} \mathbb{E}_{p_{\lambda^*, G_*}} \left[\|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|^4 |\widehat{\lambda}_n - \lambda^*|^2 \right]$$
$$\lesssim \frac{\log n}{n},$$

$$\sup_{(\lambda^*, G_*) \in \Xi_2(l_n)} \mathbb{E}_{p_{\lambda^*, G_*}} \left[(\lambda^*)^2 \|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|^2 \right.$$
$$\left. \times \|(\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n) - (a^*, b^*, \nu^*)\|^2 \right] \lesssim \frac{\log n}{n}.$$

The proof of Theorem 9 is in Appendix A.5. Above, we involve the set $\Xi_2(l_n)$ to ensure that the MLE $(\widehat{\lambda}_n, \widehat{G}_n) = (\widehat{\lambda}_n, \widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n)$ is a consistent estimator of its ground-truth value $(\lambda^*, G_*) = (\lambda^*, a^*, b^*, \nu^*)$. A few comments regarding Theorem 9 are in order.

(i) The first bound in Theorem 9 reveals that the convergence rate of $\widehat{\lambda}_n$ becomes slower than $\widetilde{\mathcal{O}}(n^{-1/2})$ and is subject to the vanishing rate of the quantity $\|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|^4$ as a result of the prompt merging issue. Note that $\|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|^4$ cannot approach zero faster than the term $(\|\Delta a^*\|^4 + |\Delta b^*|^8 + |\Delta \nu^*|^4)$ in Theorem 4. Thus, the convergence rate of $\widehat{\lambda}_n$ in this theorem is technically faster than that in Theorem 4.

(ii) The second bound in Theorem 9 suggests that the prompt parameter estimators $\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n$ share the same convergence rate which depend on the vanishing rates of both λ^* and the term $\|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|^2$. Such convergence behavior is different from that in Theorem 4 where the rate of $\widehat{b}_n$ is only the squared root of those of $\widehat{a}_n$ and $\widehat{\nu}_n$. Moreover, since $\|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|^2$ cannot vanish faster than its counterpart $(\|\Delta a^*\|^2 + |\Delta b^*|^4 + |\Delta \nu^*|^2)$ in Theorem 4, we deduce that the convergence rates of $\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n$ are faster than those in Theorem 4. These rate improvements are owing to the disappearance of the parameter interaction in equation (3).

Finally, we establish the minimax lower bounds for estimating true parameters (λ^*, G_*) under this section setting to demonstrate that the above MLE convergence rates are optimal.

Theorem 10 (Minimax lower bounds). *Under the setting of Theorem 9, we have for any* $0 < r < 1$ *and sequence* $(l_n)_{n \geq 1}$ *that*

$$\inf_{(\overline{\lambda}_n, \overline{G}_n) \in \Xi} \sup_{(\lambda, G) \in \Xi_2(l_n)} \mathbb{E}_{p_{\lambda, G}} \left[\|(\Delta a, \Delta b, \Delta \nu)\|^4 \right.$$
$$\left. \times |\overline{\lambda}_n - \lambda|^2 \right] \gtrsim n^{-1/r},$$

$$\inf_{(\overline{\lambda}_n, \overline{G}_n) \in \Xi} \sup_{(\lambda, G) \in \Xi_2(l_n)} \mathbb{E}_{p_{\lambda, G}} \left[\lambda^2 \|(\Delta a, \Delta b, \Delta \nu)\|^2 \right.$$
$$\left. \times \|(\overline{a}_n, \overline{b}_n, \overline{\nu}_n) - (a, b, \nu)\|^2 \right] \gtrsim n^{-1/r}.$$

Proof of Theorem 10 is in Appendix B.3.

5 NUMERICAL EXPERIMENTS

In this section, we conduct numerical experiments to verify the theoretical results presented in Theorem 2. Meanwhile, we defer additional experiments for other theorems to Appendix E.

Experimental setup. In Theorem 2, since the pre-trained model f_0 does not belong to the Gaussian den-
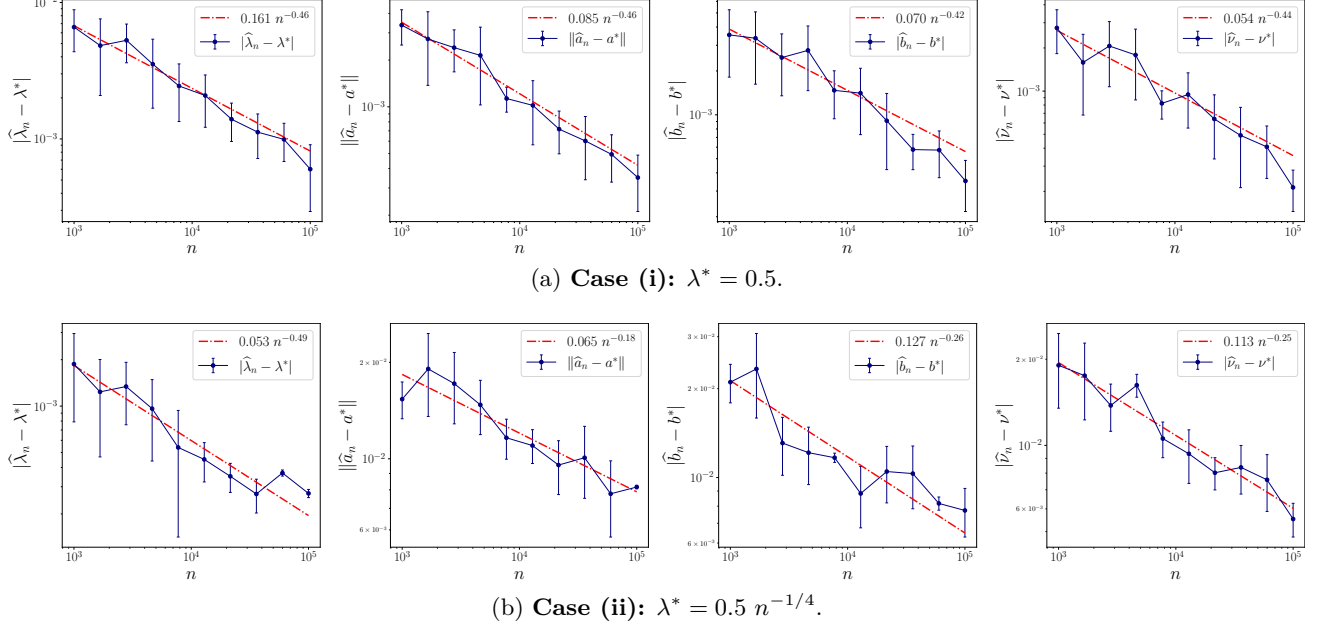
(a) **Case (i):** $\lambda^* = 0.5$.

(b) **Case (ii):** $\lambda^* = 0.5 \; n^{-1/4}$.

Figure 1: Log-log graphs depicting the empirical convergence rates of the MLE $(\widehat{\lambda}_n, \widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n)$ to the ground-truth values $(\lambda^*, a^*, b^*, \nu^*)$. The blue lines display the parameter estimation errors, while the red dashed dotted lines are the fitted lines, highlighting the empirical MLE convergence rates. Figure 1a and Figure 1b illustrates results for the cases when $\lambda^* = 0.5$ and $\lambda^* = 0.5 \; n^{-1/4}$, respectively.

sity family, we let f_0 be the density of a Student's t-distribution, characterized by mean $\varphi(a_0^\top X + b_0) = a_0^\top X + b_0$ and degrees of freedom $\nu_0 = 4$. Here, a_0 is a d-dimension vector $(1, 0, \ldots, 0)$ and $b_0 = 0$. Meanwhile, the prompt f is formulated as a Gaussian density with mean $\sigma((a^*)^\top X + b^*) = (a^*)^\top X + b^*$ and variance ν^*.

Synthetic data generation. We create synthetic datasets following the model outlined in equation (1). Specifically, we generate data pairs $\{(X_i, Y_i)\}_{i=1}^n \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$ by initially drawing X_i from a uniform distribution on the interval $[-1, 1]^d$ for $i = 1, \ldots, n$, consistently setting $d = 8$ across all trials. Subsequently, Y_i is produced according to the true density $p_{\lambda^*, G_*}(Y|X)$. Here, we consider two following cases of the true parameters $(\lambda^*, G_*) = (\lambda^*, a^*, b^*, \nu^*)$ to observe the difference in the MLE convergence behavior when λ^* is fixed versus when it varies with n:

(i) $\lambda^* = 0.5, a^* = \mathbf{1}_d, b^* = 1, \nu^* = 0.01$;

(ii) $\lambda^* = 0.5 \; n^{-1/4}, a^* = \mathbf{1}_d, b^* = 1, \nu^* = 0.01$.

Training procedure. We conducted 20 experiments for each sample size n, ranging from 10^3 to 10^5. Prior to each experiment, parameter initialization was carefully set close to the true parameter values to mitigate potential optimization instabilities. Each experiment employed an EM algorithm (Jordan and Jacobs, 1994) to derive the MLE, utilizing an SGD optimizer in Py-

Torch for the maximization step due to the absence of a universal closed-form solution.

Results. Note that the convergence rates of $\widehat{\lambda}_n$ under both cases are close to $\mathcal{O}(n^{-1/2})$. However, there are differences among those of $(\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n)$ in the two cases. In particular, in Figure 1a when λ^* is fixed, they share similar rates of orders roughly $\mathcal{O}(n^{-1/2})$. By contrast, when λ^* is set to vanish at the rate $\mathcal{O}(n^{-1/4})$, their rates become substantially slower as shown in Figure 1b, which are of orders around $\mathcal{O}(n^{-1/4})$. Those empirical results totally align with Theorem 2.

6 DISCUSSION

In this paper, we carry out the convergence analysis of maximum likelihood parameter estimation under the contaminated MoE model which is the mixture of a frozen pre-trained model and a trainable prompt for learning downstream tasks. Due to the dependence of ground-truth parameters on the sample size, there are several challenges for the theoretical analysis, namely, the prompt vanishing when the mixing proportion tends to zero; the prompt merging when it converges to the pre-trained model; and the parameter interaction via the heat equation. For better understanding, we divide our analysis into multiple scenarios based upon the linear and non-linear structures of the expert functions as well as the distribution

Fanqi Yan[★,◇], Huy Nguyen[★,†], Dung Le[★,†], Pedram Akabrian[‡], Nhat Ho[†]

family of the pre-trained model. In each scenario, we provide corresponding MLE convergence rates along with their minimax lower bounds to show that those rates are optimal. Additionally, we also conduct experiments to verify our theoretical results.

Practical implications. From our convergence analysis whose result is summarized in Table 1, we observe that the MLE convergence behaves better when the expert functions associated with the pre-trained model and the prompt do not share the same structure. Since several large-scale pre-trained models often adopt non-linear experts in practice, our theory implies that we can achieve a good performance on downstream tasks even when attaching the prompts with simple linear experts to those models. However, if ones still would like to employ non-linear experts in the prompt, then our analysis suggests that it is better to use different expert functions from those in the pre-trained model.

Limitations and Future directions. A few natural directions arise from our work. Firstly, the current contaminated MoE model considers only one prompt. It is of practical importance to extend the current analysis to the setting where several prompts are incorporated. The convergence rates for parameter estimation under that setting will provide us new insights into reduce the prompt length in fine-tuning large-scale AI models based on merging the close experts. Secondly, the current model assumes that the density f belongs to the Gaussian family. We plan to generalize the current theoretical analysis for general family of density functions f. Finally, the gating function (mixing proportion) considered in this paper is input-independent. Thus, we can examine popular input-dependent gating functions, namely softmax gating (Shazeer et al., 2017; Nguyen et al., 2023) and sigmoid gating (Csordás et al., 2023; Nguyen et al., 2024c), in future works.

Acknowledgements

References

Chow, Y., Tulepbergenov, A., Nachum, O., Gupta, D., Ryu, M., Ghavamzadeh, M., and Boutilier, C. (2023). A Mixture-of-Expert Approach to RL-based Dialogue Management. In *The Eleventh International Conference on Learning Representations.* (Cited on page 1.)

Csordás, R., Irie, K., and Schmidhuber, J. (2023). Approximating two-layer feedforward networks for efficient transformers. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 674–692, Singapore. Association for Computational Linguistics. (Cited on page 9.)

Do, D., Do, L., and Nguyen, X. (2022). Strong identifiability and parameter learning in regression with heterogeneous response. *arXiv preprint arXiv:2212.04091.* (Cited on page 2.)

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A., Firat, O., Zoph, B., Fedus, L., Bosma, M., Zhou, Z., Wang, T., Wang, E., Webster, K., Pellat, M., Robinson, K., Meier-Hellstern, K., Duke, T., Dixon, L., Zhang, K., Le, Q., Wu, Y., Chen, Z., and Cui, C. (2022). Glam: Efficient scaling of language models with mixture-of-experts. In *ICML.* (Cited on page 1.)

Fedus, W., Zoph, B., and Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23:1–39. (Cited on page 1.)

Gadat, S., Kahn, J., Marteau, C., and Maugis-Rabusseau, C. (2020). Parameter recovery in two-component contamination mixtures: The l^2 strategy. (Cited on pages 40 and 42.)

Gupta, S., Mukherjee, S., Subudhi, K., Gonzalez, E., Jose, D., Awadallah, A., and Gao, J. (2022). Sparsely activated mixture-of-experts are robust multi-task learners. *arXiv preprint arxiv 2204.0768.* (Cited on page 1.)

Han, X., Nguyen, H., Harris, C., Ho, N., and Saria, S. (2024). FuseMoE: Mixture-of-experts transformers for fleximodal fusion. In *Advances on Neural Information Processing Systems (NeurIPS).* (Cited on page 1.)

Hazimeh, H., Zhao, Z., Chowdhery, A., Sathiamoorthy, M., Chen, Y., Mazumder, R., Hong, L., and Chi, E. (2021). DSelect-k: Differentiable Selection in the Mixture of Experts with Applications to Multi-Task Learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29335–29347. Curran Associates, Inc. (Cited on page 1.)

Ho, N. and Nguyen, X. (2016a). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44:2726–2755. (Cited on page 30.)

Ho, N. and Nguyen, X. (2016b). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10:271–307. (Cited on page 13.)

Ho, N., Yang, C.-Y., and Jordan, M. I. (2022). Convergence rates for Gaussian mixtures of experts.

Journal of Machine Learning Research, 23(323):1–81. (Cited on pages 1, 2, 5, and 46.)

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3. (Cited on page 1.)

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2024). Mixtral of experts. *arxiv preprint arxiv 2401.04088*. (Cited on page 1.)

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214. (Cited on page 8.)

Le, M., Nguyen, A., Nguyen, H., Nguyen, T., Pham, T., Van Ngo, L., and Ho, N. (2024). Mixture of experts meets prompt-based continual learning. In *Advances in Neural Information Processing Systems*. (Cited on page 1.)

Li, B., Shen, Y., Yang, J., Wang, Y., Ren, J., Che, T., Zhang, J., and Liu, Z. (2023). Sparse mixture-of-experts are domain generalizable learners. In *The Eleventh International Conference on Learning Representations*. (Cited on page 1.)

Liang, H., Fan, Z., Sarkar, R., Jiang, Z., Chen, T., Zou, K., Cheng, Y., Hao, C., and Wang, Z. (2022). M^{3}ViT: Mixture-of-Experts Vision Transformer for Efficient Multi-task Learning with Model-Accelerator Co-design. In *NeurIPS*. (Cited on page 1.)

Ng, L. H. X. and Carley, K. M. (2023). Botbuster: Multi-platform bot detection using a mixture of experts. In *Proceedings of the international AAAI conference on web and social media*, volume 17, pages 686–697. (Cited on page 1.)

Nguyen, H., Akbarian, P., and Ho, N. (2024a). Is temperature sample efficient for softmax Gaussian mixture of experts? In *Proceedings of the ICML*. (Cited on page 5.)

Nguyen, H., Akbarian, P., Nguyen, T., and Ho, N. (2024b). A general theory for softmax gating multinomial logistic mixture of experts. In *Proceedings of the ICML*. (Cited on page 1.)

Nguyen, H., Ho, N., and Rinaldo, A. (2024c). Sigmoid gating is more sample efficient than softmax gating in mixture of experts. In *Advances in Neural Information Processing Systems*. (Cited on page 9.)

Nguyen, H., Nguyen, K., and Ho, N. (2024d). On parameter estimation in deviated Gaussian mixture

of experts. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. (Cited on page 2.)

Nguyen, H., Nguyen, T., and Ho, N. (2023). Demystifying softmax gating function in Gaussian mixture of experts. In *Advances in Neural Information Processing Systems*. (Cited on page 9.)

Nguyen, H., Nguyen, T., Nguyen, K., and Ho, N. (2024e). Towards convergence rates for parameter estimation in Gaussian-gated mixture of experts. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. (Cited on page 5.)

Obando Ceron, J. S., Sokar, G., Willi, T., Lyle, C., Farebrother, J., Foerster, J. N., Dziugaite, G. K., Precup, D., and Castro, P. S. (2024). Mixtures of experts unlock parameter scaling for deep RL. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 38520–38540. PMLR. (Cited on page 1.)

Pham, Q., Do, G., Nguyen, H., Nguyen, T., Liu, C., Sartipi, M., Nguyen, B. T., Ramasamy, S., Li, X., Hoi, S., and Ho, N. (2024). Competesmoe – effective training of sparse mixture of experts via competition. *arXiv preprint arXiv:2402.02526*. (Cited on page 1.)

Pini, S., Perone, C. S., Ahuja, A., Ferreira, A. S. R., Niendorf, M., and Zagoruyko, S. (2023). Safe real-world autonomous driving by learning to predict and plan with a mixture of experts. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10069–10075. IEEE. (Cited on page 1.)

Puigcerver, J., Riquelme, C., Mustafa, B., and Houlsby, N. (2024). From sparse to soft mixtures of experts. In *The Twelfth International Conference on Learning Representations*. (Cited on page 1.)

Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Pint, A. S., Keysers, D., and Houlsby, N. (2021). Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 34, pages 8583–8595. Curran Associates, Inc. (Cited on page 1.)

Ruiz, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Pinto, A., Keysers, D., and Houlsby, N. (2021). Scaling vision with sparse mixture of experts. In *NeurIPS*. (Cited on page 1.)

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *In International Con-*

Fanqi Yan[*,◇]**, Huy Nguyen**[*,†]**, Dung Le**[*,†]**, Pedram Akabrian**[‡]**, Nhat Ho**[†]

ference on Learning Representations. (Cited on page 9.)

Shen, S., Hou, L., Zhou, Y., Du, N., Longpre, S., Wei, J., Chung, H. W., Zoph, B., Fedus, W., Chen, X., Vu, T., Wu, Y., Chen, W., Webson, A., Li, Y., Zhao, V. Y., Yu, H., Keutzer, K., Darrell, T., and Zhou, D. (2024). Mixture-of-experts meets instruction tuning: A winning combination for large language models. In *The International Conference on Learning Representations.* (Cited on page 1.)

van de Geer, S. (2000). *Empirical Processes in M-estimation.* Cambridge University Press. (Cited on pages 2, 43, and 44.)

Zadouri, T., Üstün, A., Ahmadian, A., Ermis, B., Locatelli, A., and Hooker, S. (2024). Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. In *The Twelfth International Conference on Learning Representations.* (Cited on page 1.)

Zhou, Y., Du, N., Huang, Y., Peng, D., Lan, C., Huang, D., Shakeri, S., So, D., Dai, A., Lu, Y., Chen, Z., Le, Q., Cui, C., Laudon, J., and Dean, J. (2023). Brainformers: Trading simplicity for efficiency. In *International Conference on Machine Learning*, pages 42531–42542. PMLR. (Cited on page 1.)

Supplement to "Understanding Expert Structures on Minimax Parameter Estimation in Contaminated Mixture of Experts"

In this supplementary material, we provide additional results and proofs. At first, Appendix A will provide the proofs for convergence rates of parameter estimation. Appendix B focuses on the proofs for the minimax lower bounds for parameter estimation. Appendix C contains the general theory for convergence rates of densities (Theorem 1) and their corresponding proofs. Appendix D includes auxiliary results and their proof. Finally, Appendix E illustrates the numerical experiments for justifying the theoretical findings.

A PROOFS FOR PARAMETER ESTIMATION RATES

In this section, we will provide the proofs for the MLE rate theorems discussed in Sections 3 and 4, focusing on various expert structures.

A.1 Proof of Theorem 2

Now we consider the first situation we meet, the expert function σ is of linear form and f_0 is not a Gaussian density. For the MLE $(\widehat{\lambda}_n, \widehat{G}_n) = (\widehat{\lambda}_n, \widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n)$ as defined in equation (2), we obtain the MLE convergence rates as follows:

$$\sup_{(\lambda^*, G_*) \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} \left[|\widehat{\lambda}_n - \lambda^*|^2 \right] \lesssim \frac{\log n}{n},$$

$$\sup_{(\lambda^*, G_*) \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} \left[(\lambda^*)^2 \|(\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n) - (a^*, b^*, \nu^*)\|^2 \right] \lesssim \frac{\log n}{n}.$$

Proof of Theorem 2. In order to prove the above results, for any $(\lambda, G), (\lambda^*, G_*)$, we define a loss

$$D_1((\lambda, G), (\lambda^*, G_*)) = |\lambda - \lambda^*| + (\lambda + \lambda^*)\|(a, b, \nu) - (a^*, b^*, \nu^*)\|. \tag{4}$$

Then, we claim that there exists a positive constant C_1 that depends on Θ, f_0 and satisfies the following for all (λ, G) and (λ^*, G_*):

$$V(p_{\lambda, G}, p_{\lambda^*, G_*}) \geq C_1 D_1((\lambda, G), (\lambda^*, G_*)). \tag{5}$$

Then it follows from Theorem 1 that:

$$\mathbb{E}_{p_{\lambda^*, G_*}} \left(|\widehat{\lambda}_n - \lambda^*| + \lambda^* \|(\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n) - (a^*, b^*, \nu^*)\| \right) \lesssim \mathbb{E}_{p_{\lambda^*, G_*}} V(p_{\widehat{\lambda}_n, \widehat{G}_n}, p_{\lambda^*, G_*}) \leq \mathbb{E}_{p_{\lambda^*, G_*}} h(p_{\widehat{\lambda}_n, \widehat{G}_n}, p_{\lambda^*, G_*})$$

$$\lesssim^{Thm1} \sqrt{\frac{\log n}{n}}.$$

Since all inequalities are uniform with respect to (λ^*, G_*), we can conclude the theorem. $\qquad \square$

We will now prove the claim stated in equation (5). To establish the inequality in equation (5), we require a distinguishability condition between f_0 and f, which involves the higher-order derivatives of the density.

Lemma 1. *Under the setting of Theorem 2, i.e. the expert function σ is of linear form and f_0 is not a Gaussian density, for any component $(a, b, \nu) \in \Theta$, if we have real coefficients η, τ_α for all $\alpha = (\alpha_1, \alpha_2, \alpha_3) \in \mathbb{N}^d \times \mathbb{N} \times \mathbb{N}$, $|\alpha| = |\alpha_1| + |\alpha_2| + |\alpha_3| \leq 1$ such that*

$$\eta f_0(Y|\varphi(a_0^\top X + b_0), \nu_0) + \sum_{|\alpha| \leq 1} \tau_\alpha \frac{\partial^{|\alpha|} f}{\partial a^{\alpha_1} \partial b^{\alpha_2} \partial \nu^{\alpha_3}}(Y|\sigma(a^\top X + b), \nu) = 0 \tag{6}$$

for almost surely $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, then we will have $\eta = \tau_\alpha = 0$ for all $|\alpha| \leq 1$.

Fanqi Yan$^{\star,\diamond}$, Huy Nguyen*,†, Dung Le*,†, Pedram Akabrian‡, Nhat Ho†

Proof of Lemma 1. Under the setting of Theorem 2, equation (6) could be expressed as

$$\eta f_0 + \tau_{\alpha_0} f + \sum_{i=1}^{d} \tau_{\alpha_{1,i}} \frac{\partial f}{\partial a_i} + \tau_{\alpha_2} \frac{\partial f}{\partial b} + \tau_{\alpha_3} \frac{\partial f}{\partial \nu} = \eta f_0 + \tau_{\alpha_0} f + \left(\sum_{i=1}^{d} \tau_{\alpha_{1,i}} \frac{\partial \sigma}{\partial a_i} + \tau_{\alpha_2} \frac{\partial \sigma}{\partial b} \right) \frac{\partial f}{\partial \sigma} + \tau_{\alpha_3} \frac{1}{2} \frac{\partial^2 f}{\partial \sigma^2} = 0.$$

Since $f_0 \neq f$ and $\frac{\partial^k f}{\partial \sigma^k}, k = 0, 1, 2$ are linear independent (see Lemma 9). The coefficients of f_0, f, $\frac{\partial f}{\partial \sigma}$ and $\frac{\partial^2 f}{\partial \sigma^2}$ must be zero. Now consider the coefficient for $\frac{\partial f}{\partial \sigma}$:

$$\sum_{i=1}^{d} \tau_{\alpha_{1,i}} \frac{\partial \sigma}{\partial a_i} + \tau_{\alpha_2} \frac{\partial \sigma}{\partial b} = 0 \iff \left(\tau_{\alpha_1}^{\top} X + \tau_{\alpha_2} \right) \sigma'(a^{\top} X + b) = 0. \tag{7}$$

Here we consider σ is a linear function and $\sigma' \neq 0$ in a non-zero measure set in $\mathbb{R}$, so only when $\tau_{\alpha_1} = 0$ and $\tau_{\alpha_2} = 0$, equation (7) will occur, which means that $\eta = \tau_\alpha = 0$ for all $|\alpha| \leq 1$. $\square$

Proof of the claim in equation (5). Now we will prove the inequality in equation (5). Let $\overline{G} = (\bar{a}, \bar{b}, \bar{\nu})$, we need to demonstrate that

$$\lim_{\varepsilon \to 0} \inf_{(\lambda, G), (\lambda^*, G_*)} \left\{ \frac{V(p_{\lambda, G}, p_{\lambda^*, G_*})}{D_1((\lambda, G), (\lambda^*, G_*))} : D_1((\lambda, G), (\overline{\lambda}, \overline{G})) \vee D_1((\lambda^*, G_*), (\overline{\lambda}, \overline{G})) \leq \varepsilon \right\} > 0.$$

Using the argument with Fatou's lemma as in Theorem 3.1, (Ho and Nguyen, 2016b) , it is sufficient to show that

$$\lim_{\varepsilon \to 0} \inf_{(\lambda, G), (\lambda^*, G_*)} \left\{ \frac{\|p_{\lambda, G} - p_{\lambda^*, G_*}\|_\infty}{D_1((\lambda, G), (\lambda^*, G_*))} : D_1((\lambda, G), (\overline{\lambda}, \overline{G})) \vee D_1((\lambda^*, G_*), (\overline{\lambda}, \overline{G})) \leq \varepsilon \right\} > 0.$$

Assume by contrary that the above claim is not true. Then, there exist two sequences $G_n = (a_n, b_n, \nu_n)$ and $G_{*,n} = (a_n^*, b_n^*, \nu_n^*)$, and two sequences of mixing proportions λ_n and $\lambda_n^* \in [0, 1]$ such that when n tends to infinity, we get

$$\begin{cases} D_1((\lambda_n, G_n), (\overline{\lambda}, \overline{G})) \to 0, \\ D_1((\lambda_n^*, G_{*,n}), (\overline{\lambda}, \overline{G})) \to 0, \\ \|p_{\lambda_n, G_n} - p_{\lambda_n^*, G_{*,n}}\|_\infty / D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0. \end{cases}$$

In this proof, we will take into account only the most challenging setting of $G_n = (a_n, b_n, \nu_n)$ and $G_n^* = (a_n^*, b_n^*, \nu_n^*)$ when they converge to the same limit point (a', b', ν'), where (a', b', ν') is not necessarily equal to $(\bar{a}, \bar{b}, \bar{\nu})$ as $\lambda_n, \lambda_n^* \in [0, 1]$ can go to 0 or 1 in the limit. Next, we consider

$$p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y) = (\lambda_n^* - \lambda_n)[f_0(Y|\varphi((a_0)^\top X + b_0), \nu_0) - f(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)]\bar{f}(X)$$
$$+ \lambda_n[f(Y|\sigma((a_n)^\top X + b_n), \nu_n) - f(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)]\bar{f}(X).$$

By applying the Taylor expansion up to the first order, we get

$$f(Y|\sigma((a_n)^\top X + b_n), \nu_n) - f(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)$$
$$= \left[\sum_{u=1}^{d} (a_n - a_n^*)^{(u)} \frac{\partial f}{\partial a^{(u)}} + (b_n - b_n^*) \frac{\partial f}{\partial b} + (\nu_n - \nu_n^*) \frac{\partial f}{\partial \nu} \right] (Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*) + R(X, Y),$$

where $R(X, Y) = \mathcal{O}(\|(a_n, b_n, \nu_n) - (a_n^*, b_n^*, \nu_n^*)\|^{1+\gamma})$ for some $\gamma > 0$ is the remainder from the Taylor expansion. Since f is the p.d.f of the univariate location-scale Gaussian distribution, we have the following partial differential equation (PDE):

$$\frac{1}{2} \frac{\partial^2 f}{\partial \sigma^2} (Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*) = \frac{\partial f}{\partial \nu} (Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*).$$

Thus, the difference $p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y)$ is reduced to

$$p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y) = (\lambda_n^* - \lambda_n)[f_0(Y|\varphi((a_0)^\top X + b_0), \nu_0) - f(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)]\bar{f}(X)$$

$$+ \lambda_n \Big[\sum_{u=1}^d (a_n - a_n^*)^{(u)} \frac{\partial \sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*) \frac{\partial f}{\partial \sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*) \tag{8}$$

$$+ (b_n - b_n^*) \frac{\partial \sigma}{\partial b}(X, a_n^*, b_n^*) \frac{\partial f}{\partial \sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*) \tag{9}$$

$$+ \frac{1}{2}(\nu_n - \nu_n^*) \frac{\partial^2 f}{\partial \sigma^2}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*) + R(X, Y) \Big] \bar{f}(X). \tag{10}$$

From the formulation of $D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))$, it is obvious that

$$R(X, Y)\bar{f}(X)/D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) = \mathcal{O}(\|(a_n, b_n, \nu_n) - (a_n^*, b_n^*, \nu_n^*)\|^\gamma) \to 0$$

as $n \to \infty$. Therefore, the quantity $[p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y)]/D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))$ can be seen as a linear combination of elements in the following set

$$\mathcal{S} = \Big\{ f_0(Y|\varphi((a_0)^\top X + b_0), \nu_0)\bar{f}(X), f(Y|\sigma((a_n^*)^\top X + b_n^*, \nu_n^*))\bar{f}(X),$$

$$\frac{\partial \sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*) \frac{\partial f}{\partial \sigma}(Y|\sigma((a_n^*)^\top X + b_n^*, \nu_n^*))\bar{f}(X),$$

$$\frac{\partial \sigma}{\partial b}(X, a_n^*, b_n^*) \frac{\partial f}{\partial \sigma}(Y|\sigma((a_n^*)^\top X + b_n^*, \nu_n^*))\bar{f}(X),$$

$$\frac{\partial^2 f}{\partial \sigma^2}(Y|\sigma((a_n^*)^\top X + b_n^*, \nu_n^*))\bar{f}(X) : u \in [d] \Big\}.$$

Assume that the coefficients associated with those terms go to 0 as $n \to \infty$. Thus, we obtain that

$$\frac{\lambda_n^* - \lambda_n}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0, \qquad \frac{\lambda_n(a_n - a_n^*)^{(u)}}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0,$$

$$\frac{\lambda_n(b_n - b_n^*)}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0, \qquad \frac{\lambda_n(\nu_n - \nu_n^*)}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0,$$

as $n \to \infty$ for all $u \in [d]$. The last three limits imply that

$$(\lambda_n + \lambda_n^*)\|(a_n, b_n, \nu_n) - (a_n^*, b_n^*, \nu_n^*)\|/D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0.$$

Consequently, we achieve that

$$1 = [|\lambda_n^* - \lambda_n| + (\lambda_n^* + \lambda_n)\|(a_n, b_n, \nu_n) - (a_n^*, b_n^*, \nu_n^*)\|]/D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0,$$

which is a contradiction. Thus, not all the coefficients of elements in the set $\mathcal{S}$ tend to 0 as $n \to \infty$. Let us denote by m_n the maximum of the absolute values of those coefficients. It follows from the previous result that $1/m_n \not\to \infty$. Therefore, we get

$$\frac{1}{m_n} \frac{p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y)}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \tag{11}$$

$$\to \eta f_0(X|Y|\varphi((a_0)^\top X + b_0), \nu_0) + \sum_{\ell=0}^2 \alpha_\ell(X) \frac{\partial^\ell f}{\partial \sigma^\ell}(Y|\sigma((a_n')^\top X + b_n'), \nu_n') = 0, \tag{12}$$

where

$$\eta = \lim_{n \to \infty} \frac{1}{m_n} \frac{(\lambda_n^* - \lambda_n)}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))}, \quad \alpha_0(X) = \lim_{n \to \infty} \frac{1}{m_n} \frac{-(\lambda_n^* - \lambda_n)}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))},$$

$$\alpha_1(X) = \lim_{n \to \infty} \left[\sum_{u=1}^d \frac{1}{m_n} \frac{\lambda_n(a_n - a_n^*)^{(u)}}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \frac{\partial \sigma}{\partial a^{(u)}}(X, a', b') + \frac{1}{m_n} \frac{\lambda_n(b_n - b_n^*)}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \frac{\partial \sigma}{\partial b}(X, a', b') \right],$$

$$\alpha_2(X) = \lim_{n \to \infty} \frac{1}{2m_n} \frac{\lambda_n(\nu_n - \nu_n^*)}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))}.$$

Recall that f is distinguishable from f_0, we obtain that $\eta = \alpha_0(X) = \alpha_1(X) = \alpha_2(X) = 0$ for almost surely X. Since $\sigma(a^\top X + b)$ and ν are algebraically independent, this result implies that

$$\lim_{n \to \infty} \frac{\lambda_n (a_n - a_n^*)^{(u)}}{D_1((\lambda_n, G_n),(\lambda_n^*, G_{*,n}))} = \lim_{n \to \infty} \frac{\lambda_n(b_n - b_n^*)}{D_1((\lambda_n, G_n),(\lambda_n^*, G_{*,n}))} = \lim_{n \to \infty} \frac{\lambda_n(\nu_n - \nu_n^*)}{D_1((\lambda_n, G_n),(\lambda_n^*, G_{*,n}))} = 0$$

for all $u \in [d]$. This contradicts the fact that not all the coefficients of elements in the set $\mathcal{S}$ vanish as $n \to \infty$. Hence, we reach the desired conclusion. □

A.2 Proof of Theorem 6

In this section, we draw our attention to the scenario where f_0 is a Gaussian density, the expert σ is of linear form, while φ is a non-linear expert.

Proof of Theorem 6. As discussed in Section 3.2.2, the expert model $\sigma((a^*)^\top X + b^*)$ cannot converge to its counterpart $\varphi(a_0^\top X + b_0)$ due to the inherent differences in their structures. Consequently, the prompt $f(Y|\sigma((a^*)^\top X + b^*), \nu^*)$ won't converge to the pre-trained model $f_0(Y|\varphi(a_0^\top X + b_0), \nu_0)$. Following the same reasoning as in Lemma 1, we can validate that for any component $(a, b, \nu) \in \Theta$, if we have real coefficients η, τ_α for all $\alpha = (\alpha_1, \alpha_2, \alpha_3) \in \mathbb{N}^d \times \mathbb{N} \times \mathbb{N}$, $|\alpha| = |\alpha_1| + |\alpha_2| + |\alpha_3| \leq 1$ such that

$$\eta f_0(Y|\varphi(a_0^\top X + b_0), \nu_0) + \sum_{|\alpha| \leq 1} \tau_\alpha \frac{\partial^{|\alpha|} f}{\partial a^{\alpha_1} \partial b^{\alpha_2} \partial \nu^{\alpha_3}}(Y|\sigma(a^\top X + b), \nu) = 0$$

for almost surely $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, then we will have $\eta = \tau_\alpha = 0$ for all $|\alpha| \leq 1$. Building on this, and similar to the discussion in the proof of the claim in equation (5), we can conclude that there exists a positive constant C_1 that depends on Θ, f_0 and satisfies the following for all (λ, G) and (λ^*, G_*):

$$C_1 \left(|\lambda - \lambda^*| + (\lambda + \lambda^*) \|(a, b, \nu) - (a^*, b^*, \nu^*)\| \right) \leq V(p_{\lambda, G}, p_{\lambda^*, G_*}).$$

Recall Theorem 1, we will have that:

$$\mathbb{E}_{p_{\lambda^*, G_*}} \left(|\widehat{\lambda}_n - \lambda^*| + \lambda^* \|(\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n) - (a^*, b^*, \nu^*)\| \right) \lesssim \sqrt{\frac{\log n}{n}}.$$

Since all inequalities hold uniformly with respect to (λ^*, G_*), we can thus conclude the theorem. □

A.3 Proof of Theorem 7

In this section, we focus on the scenario where f_0 does not belong to the family of Gaussian densities and the expert σ takes a non-linear form.

Proof of Theorem 7. Firstly, when f_0 does not belong to the family of Gaussian densities, $f_0 \neq f$ and $\partial^k f / \partial \sigma^k, k = 0, 1, 2$ are linear independent. When equation (6) holds, we will have $\eta = \tau_{\alpha_0} = \tau_{\alpha_3} = 0$ and $\left(\tau_{\alpha_1}^\top X + \tau_{\alpha_2} \right) \sigma'(a^\top X + b) = 0$. Since $\sigma'(a^\top X + b)$ is a non-linear function, it is natural to assume that the first derivative of σ is non-zero on a set of non-zero measure in $\mathbb{R}$, which serves as our global assumption. Then we can verify that f is distinguishable from f_0 as in Lemma 1. Based on this distinguishability condition, following the same argument in Appendix A.1 we will have $\forall (\lambda, G), (\lambda^*, G_*) \in \Xi$,

$$|\lambda - \lambda^*| + (\lambda + \lambda^*) \|(a, b, \nu) - (a^*, b^*, \nu^*)\| \lesssim V(p_{\lambda, G}, p_{\lambda^*, G_*}).$$

Recall Theorem 1, we obtain the following:

$$\mathbb{E}_{p_{\lambda^*, G_*}} \left(|\widehat{\lambda}_n - \lambda^*| + \lambda^* \|(\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n) - (a^*, b^*, \nu^*)\| \right) \lesssim \sqrt{\frac{\log n}{n}}.$$

Since all inequalities hold uniformly with respect to (λ^*, G_*), we can thus conclude the theorem. □

A.4 Proof of Theorem 8

In this section, we concentrate on the scenario where f_0 and f are both Gaussian distribution and the expert σ takes a non-linear form, while φ is a linear expert function.

Proof of Theorem 8. In fact, we can view this scenario as the opposite of Appendix A.2, where the expert σ is linear and its counterpart φ is non-linear. Both scenarios share a common feature: the distinction in the structure of the experts. As a result, the prompt $f(Y|\sigma((a^*)^\top X + b^*), \nu^*)$ does not converge to the pre-trained model $f_0(Y|\varphi(a_0^\top X + b_0), \nu_0)$. Therefore, following the reasoning in Appendix A.2 and assuming that the first derivative of the non-linear function $\sigma'(a^\top X + b)$ is non-zero on a set of non-zero measure in $\mathbb{R}$, as our global assumption, we can conclude the theorem.

A more detailed explanation is that this situation guarantees that f is distinguishable from f_0 as in Lemma 1. Then the claim in equation (5) will hold, i.e. $\forall (\lambda, G), (\lambda^*, G_*) \in \Xi$,

$$|\lambda - \lambda^*| + (\lambda + \lambda^*)\|(a, b, \nu) - (a^*, b^*, \nu^*)\| \lesssim V(p_{\lambda,G}, p_{\lambda^*, G_*}).$$

Thus, we can conclude the theorem, as all inequalities hold uniformly with respect to (λ^*, G_*), in conjunction with Theorem 1:

$$\sup_{(\lambda^*, G_*) \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} \left(|\widehat{\lambda}_n - \lambda^*| \right) \lesssim \sqrt{\frac{\log n}{n}}, \quad \text{and} \quad \sup_{(\lambda^*, G_*) \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} \left(\lambda^* \|(\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n) - (a^*, b^*, \nu^*)\| \right) \lesssim \sqrt{\frac{\log n}{n}}.$$

$\square$

A.5 Proof of Theorem 9

In this section, we consider the scenario where both f_0 and f belong to the family of Gaussian densities, and both expert functions, φ and σ, are non-linear. Building on the explanation in Section 4.2.2, we will focus on the challenging case where $\varphi = \sigma$ and they are twice differentiable almost everywhere, as in this situation, the prompt might not be distinguishable from the pre-trained model. Recall for the simplicity of the presentation in the paper, we define $\Delta G = (\Delta a, \Delta b, \Delta \nu) = (a - a_0, b - b_0, \nu - \nu_0)$ and $\Delta G_* = (\Delta a^*, \Delta b^*, \Delta \nu^*) = (a^* - a_0, b^* - b_0, \nu^* - \nu_0)$ for any element $(a, b, \nu), (a^*, b^*, \nu^*) \in \Theta$. For any sequence $(l_n)_{n \geq 1}$, we denote

$$\Xi_2(l_n) := \left\{ (\lambda, G) = (\lambda, a, b, \nu) \in \Xi : \frac{l_n}{\min_{1 \leq i \leq d} \{|(\Delta a)_i|^2, |\Delta b|^2, |\Delta \nu|^2\} \sqrt{n}} \leq \lambda \right\}.$$

Then, from Theorem 9 the followings will hold for any sequence $(l_n)_{n \geq 1}$ such that $l_n / \log n \to \infty$ as $n \to \infty$:

$$\sup_{(\lambda^*, G_*) \in \Xi_2(l_n)} \mathbb{E}_{p_{\lambda^*, G_*}} \left[\|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|^4 |\widehat{\lambda}_n - \lambda^*|^2 \right] \lesssim \frac{\log n}{n},$$

$$\sup_{(\lambda^*, G_*) \in \Xi_2(l_n)} \mathbb{E}_{p_{\lambda^*, G_*}} \left[(\lambda^*)^2 \|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|^2 \times \|(\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n) - (a^*, b^*, \nu^*)\|^2 \right] \lesssim \frac{\log n}{n}.$$

To prove the above results, drawing from Proposition 3, we can establish the following result concerning the lower bound of $V(p_{\lambda,G}, p_{\lambda^*, G_*})$.

Lemma 2. *Assume that f_0 and f are both Gaussians, and the mean experts φ in f_0 and σ in f are both non-linear functions. Futhermore, if $f_0(Y|\varphi(a_0^\top X + b_0), \nu_0) = f(Y|\sigma(a_0^\top X + b_0), \nu_0)$ for some $(a_0, b_0, \nu_0) \in \Theta$. For any $(\lambda, G), (\lambda^*, G_*)$, we define*

$$D_2((\lambda, G), (\lambda^*, G_*))$$
$$= \lambda \|(\Delta a, \Delta b, \Delta \nu)\|^2 + \lambda^* \|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|^2 - \min\{\lambda, \lambda^*\} \left(\|(\Delta a, \Delta b, \Delta \nu)\|^2 + \|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|^2 \right)$$
$$+ (\lambda \|(\Delta a, \Delta b, \Delta \nu)\| + \lambda^* \|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|) \times \|(\Delta a - \Delta a^*, \Delta b - \Delta b^*, \Delta \nu - \Delta \nu^*)\|.$$

Then, there exists a positive constant C_2 that depends on Θ, (a_0, b_0, ν_0) and satisfies

$$V(p_{\lambda,G}, p_{\lambda^*, G_*}) \geq C_2 D_2((\lambda, G), (\lambda^*, G_*)).$$

for all (λ, G) and (λ^*, G_*).

Fanqi Yan[⋆,◇]**, Huy Nguyen**[⋆,†]**, Dung Le**[⋆,†]**, Pedram Akabrian**[‡]**, Nhat Ho**[†]

Proof of Theorem 9. First we denote $\overline{D_2}((\lambda, G), (\lambda^*, G_*))$ by

$$\overline{D_2}((\lambda, G), (\lambda^*, G_*)) := |\lambda^* - \lambda| \|(\Delta a, \Delta b, \Delta \nu)\| \|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|$$
$$+ \|(\Delta a, \Delta b, \Delta \nu) - (\Delta a^*, \Delta b^*, \Delta \nu^*)\| \big(\lambda \|(\Delta a, \Delta b, \Delta \nu)\| + \lambda^* \|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|\big)$$

It's easily to verify that $D_2((\lambda, G), (\lambda^*, G_*)) \asymp \overline{D_2}((\lambda, G), (\lambda^*, G_*))$. By leveraging this result together with Theorem 1 and Lemma 2 , we immediately establish the following convergence rates

$$\sup_{\lambda^*, G_* \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} \left(\|(\Delta \widehat{a}_n, \Delta \widehat{b}_n, \Delta \widehat{\nu}_n)\|^2 \|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|^2 |\widehat{\lambda}_n - \lambda^*|^2 \right) \lesssim \frac{\log n}{n}, \qquad (13)$$

$$\sup_{\lambda^*, G_* \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} \left((\lambda^*)^2 \|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|^2 \|(\Delta \widehat{a}_n, \Delta \widehat{b}_n, \Delta \widehat{\nu}_n) - (\Delta a^*, \Delta b^*, \Delta \nu^*)\|^2 \right) \lesssim \frac{\log n}{n}. \qquad (14)$$

Although the result in equation (13) does not align with the first formulae of Theorem 9, we can bypass this issue using the fact that $(\lambda^*, a^*, b^*, \nu^*) \in \Xi_2(l_n)$. In fact, observe that $(\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n) - (a^*, b^*, \nu^*) = (\Delta \widehat{a}_n, \Delta \widehat{b}_n, \Delta \widehat{\nu}_n) - (\Delta a^*, \Delta b^*, \Delta \nu^*)$, applying the Cauchy-Schwartz inequality, we have

$$\mathbb{E}_{p_{\lambda^*, G_*}} \left(\|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|^4 |\widehat{\lambda}_n - \lambda^*|^2 \right) \lesssim \mathbb{E}_{p_{\lambda^*, G_*}} \left(\|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|^2 \|(\Delta \widehat{a}_n, \Delta \widehat{b}_n, \Delta \widehat{\nu}_n)\|^2 |\widehat{\lambda}_n - \lambda^*|^2 \right)$$
$$+ \mathbb{E}_{p_{\lambda^*, G_*}} \left(\|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|^2 \|(\Delta \widehat{a}_n, \Delta \widehat{b}_n, \Delta \widehat{\nu}_n) - (\Delta a^*, \Delta b^*, \Delta \nu^*)\|^2 |\widehat{\lambda}_n - \lambda^*|^2 \right)$$

The first term in the right-hand-side has been estimated in equation (13). For the second term, if $\lambda^* \neq 0$ it is in fact an implication of equation (14), noting that $|\widehat{\lambda}_n - \lambda^*|/\lambda^*$ is uniformly bounded by $2/\lambda^*$. Otherwise, if $\lambda^* = 0$, we have

$$D_2((\widehat{\lambda}_n, \widehat{G}_n), (\lambda^*, G_*)) = \widehat{\lambda}_n \|(\Delta \widehat{a}_n, \Delta \widehat{b}_n, \Delta \widehat{\nu}_n)\|^2 + \widehat{\lambda}_n \|(\Delta \widehat{a}_n, \Delta \widehat{b}_n, \Delta \widehat{\nu}_n)\| \|(\Delta \widehat{a}_n - \Delta a^*, \Delta \widehat{b}_n - \Delta b^*, \Delta \widehat{\nu}_n - \Delta \nu^*)\|$$
$$= \widehat{\lambda}_n \left(\|(\Delta \widehat{a}_n, \Delta \widehat{b}_n, \Delta \widehat{\nu}_n)\|^2 + \|(\Delta \widehat{a}_n, \Delta \widehat{b}_n, \Delta \widehat{\nu}_n)\| \|(\Delta \widehat{a}_n - \Delta a^*, \Delta \widehat{b}_n - \Delta b^*, \Delta \widehat{\nu}_n - \Delta \nu^*)\| \right)$$
$$\geq \widehat{\lambda}_n \|(\Delta \widehat{a}_n, \Delta \widehat{b}_n, \Delta \widehat{\nu}_n)\| \|(\Delta \widehat{a}_n, \Delta \widehat{b}_n, \Delta \widehat{\nu}_n) - (\Delta a^*, \Delta b^*, \Delta \nu^*)\|.$$

The conclusion follows from direct application of Lemma 2 and Theorem 1. □

Now we will prove Lemma 2.

Proof of Lemma 2. Let $\overline{G} = (\bar{a}, \bar{b}, \bar{\nu})$ such that $\bar{\lambda} \in [0, 1]$ and $(\bar{a}, \bar{b}, \bar{\nu})$ can be identical to (a_0, b_0, ν_0). Then, we will show that

(i) When $(a_0, b_0, \nu_0) \neq (\bar{a}, \bar{b}, \bar{\nu})$ and $\bar{\lambda} > 0$,

$$\lim_{\varepsilon \to 0} \inf_{(\lambda, G), (\lambda^*, G_*)} \left\{ \frac{\|p_{\lambda, G} - p_{\lambda^*, G_*}\|_\infty}{D_1((\lambda, G), (\lambda^*, G_*))} : D_1((\lambda, G), (\bar{\lambda}, \overline{G})) \vee D_1((\lambda^*, G_*), (\bar{\lambda}, \overline{G})) \leq \varepsilon \right\} > 0.$$

(ii) When $(a_0, b_0, \nu_0) = (\bar{a}, \bar{b}, \bar{\nu})$ or $(a_0, b_0, \nu_0) \neq (\bar{a}, \bar{b}, \bar{\nu})$ and $\bar{\lambda} = 0$,

$$\lim_{\varepsilon \to 0} \inf_{(\lambda, G), (\lambda^*, G_*)} \left\{ \frac{\|p_{\lambda, G} - p_{\lambda^*, G_*}\|_\infty}{D_2((\lambda, G), (\lambda^*, G_*))} : D_2((\lambda, G), (\bar{\lambda}, \overline{G})) \vee D_2((\lambda^*, G_*), (\bar{\lambda}, \overline{G})) \leq \varepsilon \right\} > 0. \qquad (15)$$

Part (i) can be proved by using the same arguments as in the proof of Theorem 2. Thus, we will consider only part (ii) in this section, specifically the most challenging setting that $(a_0, b_0, \nu_0) = (\bar{a}, \bar{b}, \bar{\nu})$. Under this assumption, we know that φ and σ are the same nonlinear function, s.t. $f_0(Y|\varphi(a_0^\top X + b_0), \nu_0) = f(Y|\sigma(a_0^\top X + b_0), \nu_0)$ for all most surly $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. Assume that the above claim in equation (15) does not hold, then there exist two sequences $G_n = (a_n, b_n, \nu_n)$ and $G_{*,n} = (a_n^*, b_n^*, \nu_n^*)$, and two sequences of mixing proportions λ_n and $\lambda_n^* \in [0, 1]$ such that

$$\begin{cases} D_2((\lambda_n, G_n), (\bar{\lambda}, \overline{G})) = \lambda_n \|(\Delta a_n, \Delta b_n, \Delta \nu_n)\|^2 \to 0, \\ D_2((\lambda_n^*, G_{*,n}), (\bar{\lambda}, \overline{G})) = \lambda_n^* \|(\Delta a_n^*, \Delta b_n^*, \Delta \nu_n^*)\|^2 \to 0, \\ \|p_{\lambda_n, G_n} - p_{\lambda_n^*, G_{*,n}}\|_\infty / D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0. \end{cases}$$

To facilitate the presentation, let us denote

$$
\begin{aligned}
A_n &= \|(\Delta a_n, \Delta b_n, \Delta \nu_n)\|, \\
B_n &= \|(\Delta a_n^*, \Delta b_n^*, \Delta \nu_n^*)\|, \\
C_n &= \|(\Delta a_n - \Delta a_n^*, \Delta b_n - \Delta b_n^*, \Delta \nu_n - \Delta \nu_n^*)\|.
\end{aligned}
$$

Now, we have three primary cases regarding the convergence behaviors between $G_n = (a_n, b_n, \nu_n)$ and $G_{*,n} = (a_n^*, b_n^*, \nu_n^*)$.

Case 1:

Both $A_n \to 0$ and $B_n \to 0$ as $n \to \infty$, or equivalently, (a_n, b_n, ν_n) and (a_n^*, b_n^*, ν_n^*) share the same limit of (a_0, b_0, ν_0). Without loss of generality, we assume that $\lambda_n^* \geq \lambda_n$ for all $n \in \mathbb{N}$. Then, the metric $D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))$ is reduced to

$$
D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) = (\lambda_n^* - \lambda_n)B_n^2 + (\lambda_n A_n + \lambda_n^* B_n)C_n.
$$

Subsequently, we consider

$$
\begin{aligned}
p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y) &= (\lambda_n^* - \lambda_n)\left[f(Y|\sigma(a_0^\top X + b_0), \nu_0) - f(Y|\sigma\left((a_n^*)^\top X + b_n^*\right), \nu_n^*)\right]\bar{f}(X) \\
&\quad + \lambda_n\left[f(Y|\sigma\left(a_n^\top X + b_n\right), \nu) - f(Y|\sigma\left((a_n^*)^\top X + b_n^*\right), \nu_n^*)\right]\bar{f}(X) \\
&= S_n + T_n,
\end{aligned}
$$

where S_n and T_n denote the first term and second term in the equation above, respectively.

By using the Taylor expansion up to the second order, we get

$$
\begin{aligned}
S_n &= (\lambda_n^* - \lambda_n)\Bigg[\sum_{1 \leq |\alpha| + |\beta| \leq 2} \frac{1}{\alpha!\beta!} \prod_{u=1}^{d} \{-(\Delta a_n^*)^{(u)}\}^{\alpha_{1u}} \{-(\Delta b_n^*)\}^{\alpha_2} \{-(\Delta \nu_n^*)\}^{\beta} \\
&\qquad\qquad \times \frac{\partial^{|\alpha| + |\beta|} f}{\partial a^{\alpha_1} \partial b^{\alpha_2} \partial \nu^\beta}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X) + R_1(X, Y)\Bigg], \\
T_n &= \lambda_n\Bigg[\sum_{1 \leq |\alpha| + |\beta| \leq 2} \frac{1}{\alpha!\beta!} \prod_{u=1}^{d} \{(\Delta a_n - \Delta a_n^*)^{(u)}\}^{\alpha_{1u}} \{(\Delta b_n - \Delta b_n^*)\}^{\alpha_2} \{(\Delta \nu_n - \Delta \nu_n^*)\}^{\beta} \\
&\qquad\qquad \times \frac{\partial^{|\alpha| + |\beta|} f}{\partial a^{\alpha_1} \partial b^{\alpha_2} \partial \nu^\beta}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X) + R_2(X, Y)\Bigg],
\end{aligned}
$$

where $R_1(X, Y) = \mathcal{O}(B_n^{2+\gamma})$ and $R_2(X, Y) = \mathcal{O}(C_n^{2+\gamma})$ are Taylor remainders for some $\gamma > 0$. According to the triangle inequality, we get $A_n + B_n \geq C_n$. It follows from this result and the assumption both $A_n \to 0$ and $B_n \to 0$ that

$$
\begin{aligned}
(\lambda_n^* - \lambda_n)|R_1(X, Y)|/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) &\leq |R_1(X, Y)|/B_n^2 = \mathcal{O}(B_n^\gamma) \to 0, \\
\lambda_n|R_2(X, Y)|/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) &\leq |R_2(X, Y)|/[(A_n + B_n)C_n] = \mathcal{O}(C_n^\gamma) \to 0,
\end{aligned}
$$

for all $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. Then we will have

$$
\frac{(\lambda_n - \lambda_n^*)|R_1(X, Y)| + \lambda_n|R_2(X, Y)|}{D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0.
$$

Due to the following PDE:

$$
\frac{\partial^2 f}{\partial \sigma^2}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*) = 2\frac{\partial f}{\partial \nu}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*),
$$

for all $(a, b, \nu) \in \Theta$, S_n can be rewritten as follows:

$$
S_n = \sum_{\tau=1}^{4} S_{n,\tau}(X)\frac{\partial^\tau f}{\partial \sigma^\tau}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X) + R_1(X, Y)
$$

Fanqi Yan$^{\star,\diamond}$**, Huy Nguyen***,†**, Dung Le***,†**, Pedram Akabrian**‡**, Nhat Ho**†

where

$$S_{n,1}(X) = (\lambda_n^* - \lambda_n)\Big[\sum_{u=1}^{d}\{-(\Delta a_n^*)^{(u)}\}\frac{\partial\sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*) + \{-(\Delta b_n^*)\}\frac{\partial\sigma}{\partial b}(X, a_n^*, b_n^*)$$

$$+ \sum_{1\leq u,v\leq d}\frac{\{-(\Delta a_n^*)^{(u)}\}\{-(\Delta a_n^*)^{(v)}\}}{1 + \mathbf{1}_{u=v}}\frac{\partial^2\sigma}{\partial a^{(u)}\partial a^{(v)}}(X, a_n^*, b_n^*)$$

$$+ \sum_{1\leq u\leq d}\{-(\Delta a_n^*)^{(u)}\}\{-(\Delta b_n^*)\}\frac{\partial^2\sigma}{\partial a^{(u)}\partial b}(X, a_n^*, b_n^*) + \frac{1}{2}\{-(\Delta b_n^*)\}^2\frac{\partial^2\sigma}{\partial b^2}(X, a_n^*, b_n^*)\Big],$$

$$S_{n,2}(X) = (\lambda_n^* - \lambda_n)\Big[\frac{1}{2}\{-(\Delta\nu_n^*)\} + \sum_{1\leq u,v\leq d}\frac{\{-(\Delta a_n^*)^{(u)}\}\{-(\Delta a_n^*)^{(v)}\}}{1 + \mathbf{1}_{u=v}}\frac{\partial\sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*)\frac{\partial\sigma}{\partial a^{(v)}}(X, a_n^*, b_n^*)$$

$$+ \sum_{1\leq u\leq d}\{-(\Delta a_n^*)^{(u)}\}\{-(\Delta b_n^*)\}\frac{\partial\sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*)\frac{\partial\sigma}{\partial b}(X, a_n^*, b_n^*) + \frac{1}{2}\{-(\Delta b_n^*)\}^2(\frac{\partial\sigma}{\partial b}(X, a_n^*, b_n^*))^2\Big],$$

$$S_{n,3}(X) = \frac{\lambda_n^* - \lambda_n}{2}\Big[\sum_{u=1}^{d}\{-(\Delta a_n^*)^{(u)}\}\{-(\Delta\nu_n^*)\}\frac{\partial\sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*) + \{-(\Delta b_n^*)\}\{-(\Delta\nu_n^*)\}\frac{\partial\sigma}{\partial b}(X, a_n^*, b_n^*),\Big]$$

$$S_{n,4}(X) = \frac{\lambda_n^* - \lambda_n}{8}\{-(\Delta\nu_n^*)\}^2.$$

Similarly, we can rewrite T_n in the same fashion as follows:

$$T_n = \sum_{\tau=1}^{4}T_{n,\tau}(X)\frac{\partial^\tau f}{\partial\sigma^\tau}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X) + R_2(X, Y),$$

where

$$T_{n,1}(X) = \lambda_n\Big[\sum_{u=1}^{d}\{(\Delta a_n - \Delta a_n^*)^{(u)}\}\frac{\partial\sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*) + \{(\Delta b_n - \Delta b_n^*)\}\frac{\partial\sigma}{\partial b}(X, a_n^*, b_n^*)$$

$$+ \sum_{1\leq u,v\leq d}\frac{\{(\Delta a_n - \Delta a_n^*)^{(u)}\}\{(\Delta a_n - \Delta a_n^*)^{(v)}\}}{1 + \mathbf{1}_{u=v}}\frac{\partial^2\sigma}{\partial a^{(u)}\partial a^{(v)}}(X, a_n^*, b_n^*)$$

$$+ \sum_{1\leq u\leq d}\{(\Delta a_n - \Delta a_n^*)^{(u)}\}\{(\Delta b_n - \Delta b_n^*)\}\frac{\partial^2\sigma}{\partial a^{(u)}\partial b}(X, a_n^*, b_n^*) + \frac{1}{2}\{(\Delta b_n - \Delta b_n)\}^2\frac{\partial^2\sigma}{\partial b^2}(X, a_n^*, b_n^*)\Big],$$

$$T_{n,2}(X) = \lambda_n\Big[\frac{1}{2}\{(\Delta\nu_n - \Delta\nu_n^*)\} + \sum_{1\leq u,v\leq d}\frac{\{(\Delta a_n - \Delta a_n^*)^{(u)}\}\{(\Delta a_n - \Delta a_n^*)^{(v)}\}}{1 + \mathbf{1}_{u=v}}\frac{\partial\sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*)\frac{\partial\sigma}{\partial a^{(v)}}(X, a_n^*, b_n^*)$$

$$+ \sum_{1\leq u\leq d}\{(\Delta a_n - \Delta a_n^*)^{(u)}\}\{(\Delta b_n - \Delta b_n^*)\}\frac{\partial\sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*)\frac{\partial\sigma}{\partial b}(X, a_n^*, b_n^*)$$

$$+ \frac{1}{2}\{(\Delta b_n - \Delta b_n^*)\}^2(\frac{\partial\sigma}{\partial b}(X, a_n^*, b_n^*))^2\Big],$$

$$T_{n,3}(X) = \frac{\lambda_n}{2}\Big[\sum_{u=1}^{d}\{(\Delta a_n - \Delta a_n^*)^{(u)}\}\{(\Delta\nu_n - \Delta\nu_n^*)\}\frac{\partial\sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*)$$

$$+ \{(\Delta b_n - \Delta b_n^*)\}\{(\Delta\nu_n - \Delta\nu_n^*)\}\frac{\partial\sigma}{\partial b}(X, a_n^*, b_n^*),\Big]$$

$$T_{n,4}(X) = \frac{\lambda_n}{8}\{(\Delta\nu_n - \Delta\nu_n^*)\}^2.$$

Therefore, we can view the quantity $[p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y)]/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})))$ as a linear com-

bination of elements of the set $\mathcal{L} = \cup_{\tau=1}^{4}\mathcal{L}_\tau$, where

$$\mathcal{L}_1 = \left\{\frac{\partial\sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*)\frac{\partial f}{\partial\sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X) : u \in [d]\right\}$$

$$\cup \left\{\frac{\partial\sigma}{\partial b}(X, a_n^*, b_n^*)\frac{\partial f}{\partial\sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X)\right\}$$

$$\cup \left\{\frac{\partial^2\sigma}{\partial a^{(u)}\partial a^{(v)}}(X, a_n^*, b_n^*)\frac{\partial f}{\partial\sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X) : u, v \in [d]\right\},$$

$$\cup \left\{\frac{\partial^2\sigma}{\partial a^{(u)}\partial b}(X, a_n^*, b_n^*)\frac{\partial f}{\partial\sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X) : u \in [d]\right\},$$

$$\cup \left\{\frac{\partial^2\sigma}{\partial b^2}(X, a_n^*, b_n^*)\frac{\partial f}{\partial\sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X)\right\},$$

$$\mathcal{L}_2 = \left\{\frac{\partial^2 f}{\partial\sigma^2}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X)\right\}$$

$$\cup \left\{\frac{\partial\sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*)\frac{\partial\sigma}{\partial a^{(v)}}(X, a_n^*, b_n^*)\frac{\partial^2 f}{\partial\sigma^2}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X) : u, v \in [d]\right\}$$

$$\cup \left\{\frac{\partial\sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*)\frac{\partial\sigma}{\partial b}(X, a_n^*, b_n^*)\frac{\partial^2 f}{\partial\sigma^2}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X) : u \in [d]\right\}$$

$$\cup \left\{(\frac{\partial\sigma}{\partial b}(X, a_n^*, b_n^*))^2\frac{\partial^2 f}{\partial\sigma^2}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X)\right\},$$

$$\mathcal{L}_3 = \left\{\frac{\partial\sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*)\frac{\partial^3 f}{\partial\sigma^3}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X) : u \in [d]\right\}$$

$$\cup \left\{\frac{\partial\sigma}{\partial b}(X, a_n^*, b_n^*)\frac{\partial^3 f}{\partial\sigma^3}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X)\right\},$$

$$\mathcal{L}_4 = \left\{\frac{\partial^4 f}{\partial\sigma^4}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X)\right\}.$$

Assume by contrary that all the coefficients of these elements vanish when $n \to \infty$. Looking at the coefficients of $\frac{\partial^2\sigma}{\partial a^{(u)}\partial a^{(v)}}(X, a_n^*, b_n^*)\frac{\partial f}{\partial\sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X)$ and $\frac{\partial^2\sigma}{\partial a^{(u)}\partial b}(X, a_n^*, b_n^*)\frac{\partial f}{\partial\sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X)$ and $\frac{\partial^2\sigma}{\partial b^2}(X, a_n^*, b_n^*)\frac{\partial f}{\partial\sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X)$, we get for all $u, v \in [d]$,

$$[(\lambda_n^* - \lambda_n)(\Delta a_n^*)^{(u)}(\Delta a_n^*)^{(v)} + \lambda_n(\Delta a_n - \Delta a_n^*)^{(u)}(\Delta a_n - \Delta a_n^*)^{(v)}]/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0,$$

$$[(\lambda_n^* - \lambda_n)(\Delta a_n^*)^{(u)}(\Delta b_n^*) + \lambda_n(\Delta a_n - \Delta a_n^*)^{(u)}(\Delta b_n - \Delta b_n^*)]/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0,$$

$$[(\lambda_n^* - \lambda_n)(\Delta b_n^*)^2 + \lambda_n(\Delta b_n - \Delta b_n^*)^2]/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0. \tag{16}$$

Let $u = v$ in the first equation in equation (16), we achieve that for all $u \in [d]$,

$$[(\lambda_n^* - \lambda_n)\{(\Delta a_n^*)^{(u)}\}^2 + \lambda_n\{(\Delta a_n - \Delta a_n^*)^{(u)}\}^2]/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0, \tag{17}$$

which implies that

$$[(\lambda_n^* - \lambda_n)\|\Delta a_n^*\|^2 + \lambda_n\|\Delta a_n - \Delta a_n^*\|^2]/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0,$$

$$[(\lambda_n^* - \lambda_n)\|\Delta b_n^*\|^2 + \lambda_n\|\Delta b_n - \Delta b_n^*\|^2]/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0. \tag{18}$$

We also have each term inside equation (18) is non-negative, thus

$$(\lambda_n^* - \lambda_n)\|\Delta a_n^*\|^2/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0, \; \lambda_n\|\Delta a_n - \Delta a_n^*\|^2/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0,$$

$$(\lambda_n^* - \lambda_n)\|\Delta b_n^*\|^2/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0, \; \lambda_n\|\Delta b_n - \Delta b_n^*\|^2/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0. \tag{19}$$

Fanqi Yan$^{\star,\diamond}$, Huy Nguyen*,†, Dung Le*,†, Pedram Akabrian‡, Nhat Ho†

Applying the AM-GM inequality, we have for all $u, v \in [d]$,

$$\frac{(\lambda_n^* - \lambda_n)(\Delta a_n^*)^{(u)}(\Delta a_n^*)^{(v)}}{D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0, \quad \frac{\lambda_n(\Delta a_n - \Delta a_n^*)^{(u)}(\Delta a_n - \Delta a_n^*)^{(v)}}{D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0, \tag{20}$$

$$\frac{(\lambda_n^* - \lambda_n)(\Delta a_n^*)^{(u)}(\Delta b_n^*)}{D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0, \quad \frac{\lambda_n(\Delta a_n - \Delta a_n^*)^{(u)}(\Delta b_n - \Delta b_n^*)}{D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0. \tag{21}$$

Next, by considering the coefficients of $\frac{\partial \sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*)\frac{\partial f}{\partial \sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X)$, $\frac{\partial \sigma}{\partial b}(X, a_n^*, b_n^*)\frac{\partial f}{\partial \sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X)$ and $\frac{\partial^2 f}{\partial \sigma^2}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X)$, we have

$$[\lambda_n(\Delta a_n)^{(u)} - \lambda_n^*(\Delta a_n^*)^{(u)}]/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0, \quad u \in [d], \tag{22}$$

$$[\lambda_n(\Delta b_n) - \lambda_n^*(\Delta b_n^*)]/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0, \tag{23}$$

$$[\lambda_n(\Delta \nu_n) - \lambda_n^*(\Delta \nu_n^*)]/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0. \tag{24}$$

Noting that for $u, v \in [d]$,

$$\lambda_n^*(\Delta a_n^*)^{(u)}(\Delta a_n - \Delta a_n^*)^{(v)} = (\lambda_n(\Delta a_n)^{(v)} - \lambda_n^*(\Delta a_n^*)^{(v)})(\Delta a_n^*)^{(u)} + (\lambda_n^* - \lambda_n)(\Delta a_n)^{(v)}(\Delta a_n^*)^{(u)},$$

$$\lambda_n(\Delta a_n)^{(u)}(\Delta a_n - \Delta a_n^*)^{(v)} = \lambda_n^*(\Delta a_n^*)^{(u)}(\Delta a_n - \Delta a_n^*)^{(v)} - (\lambda_n(\Delta a_n)^{(u)} - \lambda_n^*(\Delta a_n^*)^{(u)})(\Delta a_n - \Delta a_n^*)^{(v)}.$$

Thus, from equation (20) and equation (22), we achieve that for $u, v \in [d]$,

$$\lambda_n^*(\Delta a_n^*)^{(u)}(\Delta a_n - \Delta a_n^*)^{(v)}/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0,$$

$$\lambda_n(\Delta a_n)^{(u)}(\Delta a_n - \Delta a_n^*)^{(v)}/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0.$$

Noting that for all $u \in [d]$,

$$\lambda_n^*(\Delta a_n^*)^{(u)}(\Delta b_n - \Delta b_n^*) = (\lambda_n \Delta b_n - \lambda_n^* \Delta b_n^*)(\Delta a_n^*)^{(u)} + (\lambda_n^* - \lambda_n)(\Delta b_n)(\Delta a_n^*)^{(u)},$$

$$\lambda_n(\Delta a_n)^{(u)}(\Delta b_n - \Delta b_n^*) = \lambda_n^*(\Delta a_n^*)^{(u)}(\Delta b_n - \Delta b_n^*) - \left(\lambda_n(\Delta a_n)^{(u)} - \lambda_n^*(\Delta a_n^*)^{(u)}\right)(\Delta b_n - \Delta b_n^*).$$

Thus, from equation (22) and equation (23), we have for all $u \in [d]$,

$$\lambda_n^*(\Delta a_n^*)^{(u)}(\Delta b_n - \Delta b_n^*)/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0$$

$$\lambda_n(\Delta a_n)^{(u)}(\Delta b_n - \Delta b_n^*)/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0.$$

By using the same arguments we will derive

$$\lambda_n \|\Delta a_n\|.\|\Delta b_n - \Delta b_n^*\|/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0, \tag{25}$$

$$\lambda_n^* \|\Delta a_n^*\|.\|\Delta b_n - \Delta b_n^*\|/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0, \tag{26}$$

$$\lambda_n \|\Delta b_n\|.\|\Delta b_n - \Delta b_n^*\|/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0, \tag{27}$$

$$\lambda_n^* \|\Delta b_n^*\|.\|\Delta b_n - \Delta b_n^*\|/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0. \tag{28}$$

By using the same arguments to derive equation (17), equation (19) and equation (20), we can point out that

$$[(\lambda_n^* - \lambda_n)\|\Delta \nu_n^*\|^2 + \lambda_n\|\Delta \nu_n - \Delta \nu_n^*\|^2]/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0,$$

$$\lambda_n\|\Delta \nu_n\|.\|\Delta \nu_n - \Delta \nu_n^*\|/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0,$$

$$\lambda_n^*\|\Delta \nu_n^*\|.\|\Delta \nu_n - \Delta \nu_n^*\|/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0, \tag{29}$$

$$\lambda_n\|\Delta a_n\|.\|\Delta \nu_n - \Delta \nu_n^*\|/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0,$$

$$\lambda_n^*\|\Delta a_n^*\|.\|\Delta \nu_n - \Delta \nu_n^*\|/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0$$

$$\lambda_n\|\Delta b_n\|.\|\Delta \nu_n - \Delta \nu_n^*\|/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0,$$

$$\lambda_n^*\|\Delta b_n^*\|.\|\Delta \nu_n - \Delta \nu_n^*\|/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0.$$

Collecting results in equation (17), and equations (19) to (29), we obtain that

$$1 = D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0,$$

which is a contradiction. Therefore, not all the coefficients in the representation of $p_{\lambda_n,G_n}(X,Y) - p_{\lambda_n^*,G_{*,n}}(X,Y)/D_2((\lambda_n,G_n),(\lambda_n^*,G_{*,n}))$ tend to 0 as $n \to \infty$. Let us denote by m_n the maximum of the absolute values of those coefficients. Based on the previous result, $1/m_n \not\to \infty$. Additionally, we define

$$[(\lambda_n^* - \lambda_n)\{-(\Delta a_n^*)^{(u)}\} + \lambda_n(\Delta a_n - \Delta a_n^*)^{(u)}]/m_n \to \alpha_{1u},$$
$$[(\lambda_n^* - \lambda_n)\{-(\Delta b_n^*)\} + \lambda_n(\Delta b_n - \Delta b_n^*)]/m_n \to \alpha_{20},$$
$$[(\lambda_n^* - \lambda_n)\{-(\Delta \nu_n^*)\} + \lambda_n(\Delta \nu_n - \Delta \nu_n^*)]/m_n \to \alpha_{30},$$
$$[(\lambda_n^* - \lambda_n)(\Delta a_n^*)^{(u)}(\Delta a_n^*)^{(v)} + \lambda_n(\Delta a_n - \Delta a_n^*)^{(u)}(\Delta a_n - \Delta a_n^*)^{(v)}]/m_n \to \beta_{1uv},$$
$$[(\lambda_n^* - \lambda_n)(\Delta a_n^*)^{(u)}(\Delta b_n^*) + \lambda_n(\Delta a_n - \Delta a_n^*)^{(u)}(\Delta b_n - \Delta b_n^*)]/m_n \to \beta_{2u0},$$
$$[(\lambda_n^* - \lambda_n)(\Delta b_n^*)^2 + \lambda_n(\Delta b_n - \Delta b_n^*)^2]/m_n \to \beta_{300},$$
$$[(\lambda_n^* - \lambda_n)(\Delta a_n^*)^{(u)}(\Delta \nu_n^*) + \lambda_n(\Delta a_n - \Delta a_n^*)^{(u)}(\Delta \nu_n - \Delta \nu_n^*)]/m_n \to \gamma_{1u},$$
$$[(\lambda_n^* - \lambda_n)(\Delta b_n^*)(\Delta \nu_n^*) + \lambda_n(\Delta b_n - \Delta b_n^*)(\Delta \nu_n - \Delta \nu_n^*)]/m_n \to \gamma_{20},$$
$$[(\lambda_n^* - \lambda_n)(\Delta \nu_n^*)^2 + \lambda_n(\Delta \nu_n - \Delta \nu_n^*)^2]/m_n \to \gamma_{30},$$

when $n \to \infty$ for all $u, v \in [d]$. Note that at least one among $\alpha_{\tau u}, \beta_{\tau uv}$ and $\gamma_{\tau u}$ where $\tau \in \{1,2,3\}$ must be different from zero. By applying the Fatou's lemma, we get

$$0 = \lim_{n\to\infty} \frac{1}{m_n} \frac{2V(p_{\lambda_n,G_n}, p_{\lambda_n^*,G_{*,n}})}{D_2((\lambda_n,G_n),(\lambda_n^*,G_{*,n}))} \geq \int \liminf_{n\to\infty} \frac{1}{m_n} \frac{|p_{\lambda_n,G_n}(X,Y) - p_{\lambda_n^*,G_{*,n}}(X,Y)|}{D_2((\lambda_n,G_n),(\lambda_n^*,G_{*,n}))} d(X,Y).$$

On the other hand,

$$\frac{1}{m_n} \frac{p_{\lambda_n,G_n}(X,Y) - p_{\lambda_n^*,G_{*,n}}(X,Y)}{D_2((\lambda_n,G_n),(\lambda_n^*,G_{*,n}))} \to \sum_{\tau=1}^{4} E_\tau(X) \frac{\partial^\tau f}{\partial \sigma^\tau}(Y|\sigma((a_0)^\top X + b_0), \nu_0) \bar{f}(X),$$

where

$$E_1(X) = \sum_{u=1}^{d} \alpha_{1u} \frac{\partial \sigma}{\partial a^{(u)}}(X, a_0, b_0) + \alpha_{20} \frac{\partial \sigma}{\partial b}(X, a_0, b_0) + \sum_{1 \leq u,v \leq d} \frac{\beta_{1uv}}{1 + \mathbf{1}_{\{u=v\}}} \frac{\partial^2 \sigma}{\partial a^{(u)} \partial a^{(v)}}(X, a_0, b_0)$$

$$+ \sum_{1 \leq u \leq d} \beta_{2u0} \frac{\partial^2 \sigma}{\partial a^{(u)} \partial b}(X, a_0, b_0) + \beta_{300} \frac{\partial^2 \sigma}{\partial b^2}(X, a_0, b_0),$$

$$E_2(X) = \frac{1}{2}\alpha_{30} + \sum_{1 \leq u,v \leq d} \frac{\beta_{1uv}}{1 + \mathbf{1}_{\{u=v\}}} \frac{\partial \sigma}{\partial a^{(v)}}(X, a_0, b_0) \frac{\partial \sigma}{\partial a^{(u)}}(X, a_0, b_0)$$

$$+ \sum_{1 \leq u \leq d} \beta_{2u0} \frac{\partial \sigma}{\partial a^{(u)}}(X, a_0, b_0) \frac{\partial \sigma}{\partial b}(X, a_0, b_0) + \frac{1}{2}\beta_{300}(\frac{\partial \sigma}{\partial b}(X, a_0, b_0))^2$$

$$E_3(X) = \frac{1}{2} \sum_{u=1}^{d} \gamma_{1u} \frac{\partial \sigma}{\partial a^{(u)}}(X, a_0, b_0) + \frac{1}{2}\gamma_{20} \frac{\partial \sigma}{\partial b}(X, a_0, b_0),$$

$$E_4(X) = \frac{1}{8}\gamma_{30}.$$

Then, we have

$$\sum_{\tau=1}^{4} E_\tau(X) \frac{\partial^\tau f}{\partial \sigma^\tau}(Y|\sigma((a_0)^\top X + b_0), \nu_0) = 0.$$

It is worth noting that for almost surely (X, Y), the set

$$\left\{ \frac{\partial^\tau f}{\partial \sigma^\tau}(Y|\sigma((a_0)^\top X + b_0), \nu_0) : \tau \in [4] \right\}$$

Fanqi Yan$^{\star,\diamond}$, Huy Nguyen*,†, Dung Le*,†, Pedram Akabrian‡, Nhat Ho†

is linearly independent, which leads to the fact that $E_\tau(X) = 0$ for almost surely X for any $\tau \in [4]$.

Since $\sigma(X, a, b)$ and ν are algebraically independent and $E_3(X) = E_4(X) = 0$ for almost surely X, we get $\gamma_{\tau u} = 0$ for all u. It follows from the result $E_2(X) = 0$ and $\gamma_{\tau u} = 0$ that $\alpha_{30} = \beta_{\tau uv} = 0$ for all u, v. Next, $E_1(X) = 0$ implies that $\alpha_{\tau u} = 0$ for all u. This contradicts the fact that not all $\alpha_{\tau u}$, $\beta_{\tau uv}$ and γ_{uv} vanish. Thus, we obtain the conclusion for this case.

Case 2:

Either $A_n \to 0$ or $B_n \to 0$ as $n \to \infty$. WLOG, we assume that $A_n \not\to 0$ and $B_n \to 0$, which means that $(a_n, b_n, \nu_n) \not\to (a_0, b_0, \nu_0)$ and $(a_n^*, b_n^*, \nu_n^*) \to (a_0, b_0, \nu_0)$. Let us denote

$$D_2'((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) := |\lambda_n^* - \lambda_n|B_n + \lambda_n A_n + \lambda_n^* B_n.$$

As $D_2'((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \gtrsim D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))$ and $[p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y)]/D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0$, we get $[p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y)]/D_2'((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0$. By using the Taylor expansion up to the first order, we have

$$p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y) = (\lambda_n^* - \lambda_n)[f(Y|\sigma((a_0)^\top X + b_0), \nu_0) - f(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)]\bar{f}(X)$$
$$+ \lambda_n[f(Y|\sigma((a_n)^\top X + b_n), \nu_n) - f(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)]\bar{f}(X)$$
$$= (\lambda_n^* - \lambda_n)\Big[\sum_{u=1}^{d}(-\Delta a_n^*)^{(u)}\frac{\partial \sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*)\frac{\partial f}{\partial \sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)$$
$$+ (-\Delta b_n^*)\frac{\partial \sigma}{\partial b}(X, a_n^*, b_n^*)\frac{\partial f}{\partial \sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*) + \frac{1}{2}(-\Delta \nu_n^*)\frac{\partial^2 f}{\partial \sigma^2}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*) + R_3(X, Y)\Big]\bar{f}(X)$$
$$+ \lambda_n[f(Y|\sigma((a_n)^\top X + b_n), \nu_n) - f(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)]\bar{f}(X),$$

where $R_3(X, Y)$ is a Taylor remainder that satisfies

$$(\lambda_n^* - \lambda_n)|R_3(X, Y)|/D_2'((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) = \mathcal{O}(B_n^{\gamma'}) \to 0,$$

for some $\gamma' > 0$. Thus, we can treat $[p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y)]/D_2'((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))$ as a linear combination of elements of the set

$$\mathcal{K} = \Big\{f(Y|\sigma((a_n)^\top X + b_n), \nu_n)\bar{f}(X), f(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X),$$
$$\frac{\partial \sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*)\frac{\partial f}{\partial \sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X), \frac{\partial \sigma}{\partial b}(X, a_n^*, b_n^*)\frac{\partial f}{\partial \sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X),$$
$$\frac{\partial^2 f}{\partial \sigma^2}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)\bar{f}(X) : u \in [d]\Big\}. \tag{30}$$

Assume by contrary that all the coefficients of those elements go to 0 as $n \to \infty$, then we get for all $u \in [d]$,

$$\frac{\lambda_n}{D_2'((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0, \quad (\lambda_n^* - \lambda_n)\frac{(-\Delta a_n^*)^{(u)}}{D_2'((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0,$$
$$(\lambda_n^* - \lambda_n)\frac{(-\Delta b_n^*)}{D_2'((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0, \quad (\lambda_n^* - \lambda_n)\frac{(-\Delta \nu_n^*)}{D_2'((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0,$$

which leads to

$$\frac{(\lambda_n^* - \lambda_n)B_n}{D_2'((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0, \quad \frac{\lambda_n A_n}{D_2'((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0, \quad \frac{\lambda_n^* B_n}{D_2'((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0.$$

As a result,

$$1 = \frac{|\lambda_n^* - \lambda_n|B_n + \lambda_n A_n + \lambda_n^* B_n}{D_2'((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0,$$

which is a contradiction. Consequently, not all the aforementioned coefficients tend to 0 as $n \to \infty$.

Subsequently, we denote by m'_n the maximum of the absolute values of those coefficients, then $1/m'_n \nrightarrow \infty$. Recall that $D_2((\lambda_n, G_n), (\overline{\lambda}, \overline{G})) = \lambda_n \|(\Delta a_n, \Delta b_n, \Delta \nu_n)\|^2 \to 0$. Since $(a_n, b_n, \nu_n) \nrightarrow (a_0, b_0, \nu_0)$, we get $\lambda_n \to 0$. Thus,

$$\frac{1}{m'_n} \frac{p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y)}{D'_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to \sum_{\ell=1}^{2} \alpha'_\ell(X) \frac{\partial^\ell f}{\partial \sigma^\ell}(Y|\sigma((a_0)^\top X + b_0), \nu_0) = 0.$$

Recall that for almost surely $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, the set

$$\left\{ \frac{\partial^\ell f}{\partial \sigma^\ell}(Y|\sigma((a_0)^\top X + b_0), \nu_0) : 0 \leq \ell \leq 2 \right\} \tag{31}$$

is linearly independent. Then, equation (31) indicates that

$$\alpha'_1(X) = \lim_{n \to \infty} \left[\sum_{u=1}^{d} \frac{1}{m'_n} \frac{(\lambda_n^* - \lambda_n)(-\Delta a_n^*)^{(u)}}{D'_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \frac{\partial \sigma}{\partial a^{(u)}}(X, a_0, b_0) + \frac{1}{m'_n} \frac{(\lambda_n^* - \lambda_n)(-\Delta b_n^*)}{D'_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \frac{\partial \sigma}{\partial b}(X, a_0, b_0) \right] = 0,$$

$$\alpha'_2(X) = \lim_{n \to \infty} \frac{1}{m'_n} \frac{(\lambda_n^* - \lambda_n)(-\Delta \nu_n^*)}{D'_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} = 0.$$

Due to the algebraic independence of $\sigma(X, a, b)$ and ν, we obtain that for all $u \in [d]$,

$$\lim_{n \to \infty} \frac{1}{m'_n} \frac{(\lambda_n^* - \lambda_n)(-\Delta a_n^*)^{(u)}}{D'_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} = 0,$$

$$\lim_{n \to \infty} \frac{1}{m'_n} \frac{(\lambda_n^* - \lambda_n)(-\Delta b_n^*)}{D'_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} = 0,$$

$$\lim_{n \to \infty} \frac{1}{m'_n} \frac{(\lambda_n^* - \lambda_n)(-\Delta \nu_n^*)}{D'_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} = 0.$$

Since $1/m'_n$ is bounded, the above limits imply that the coefficients of the following items

$$\frac{\partial \sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*) \frac{\partial f}{\partial \sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*),$$

$$\frac{\partial \sigma}{\partial b}(X, a_n^*, b_n^*) \frac{\partial f}{\partial \sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*),$$

$$\frac{\partial^2 f}{\partial \sigma^2}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)$$

vanish when $n \to \infty$ for all $u \in [d]$. Moreover, those of $f(Y|\sigma((a_n)^\top X + b_n), \nu_n)$ and $f(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)$, which are $\lambda_n/D'_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))$ and $-\lambda_n/D'_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))$, also go to 0. This contradicts the assumption that at least one among the coefficients of elements in the set $\mathcal{K}$ not vanishing. Thus, we reach the conclusion for this case.

Case 3:

Both A_n and B_n do not vanish. This means that $(a_n, b_n, \nu_n) \nrightarrow (a_0, b_0, \nu_0)$ and $(a_n^*, b_n^*, \nu_n^*) \nrightarrow (a_0, b_0, \nu_0)$. In this case, we will consider only the most difficult case when (a_n, b_n, ν_n) and (a_n^*, b_n^*, ν_n^*) converge to the same point $(a', b', \nu') \neq (a_0, b_0, \nu_0)$.

From the formulation of the metric $D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))$ in the proof of Theorem 2, it is clear that $D_2((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \leq D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))$. Therefore, we get $[p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y)]/D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0$ as $n \to \infty$. By following the same steps to derive equation (8) and

Fanqi Yan[⋆,⋄], Huy Nguyen[⋆,†], Dung Le[⋆,†], Pedram Akabrian[‡], Nhat Ho[†]

equation (11) and abuse notations defined for those equations, we obtain that

$$p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y) = (\lambda_n^* - \lambda_n)[f(Y|\sigma((a_0)^\top X + b_0), \nu_0) - f(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)]\bar{f}(X)$$

$$+ \lambda_n \Big[\sum_{u=1}^d (\Delta a_n - \Delta a_n^*)^{(u)} \frac{\partial \sigma}{\partial a^{(u)}}(X, a_n^*, b_n^*) \frac{\partial f}{\partial \sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)$$

$$+ (\Delta b_n - \Delta b_n^*) \frac{\partial \sigma}{\partial b}(X, a_n^*, b_n^*) \frac{\partial f}{\partial \sigma}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*)$$

$$+ \frac{1}{2}(\Delta \nu_n - \Delta \nu_n^*) \frac{\partial^2 f}{\partial \sigma^2}(Y|\sigma((a_n^*)^\top X + b_n^*), \nu_n^*) + R(X, Y) \Big] \bar{f}(X).$$

with a note that not all the coefficients of elements of the set $\mathcal{K}$ (defined in equation (30)) go to 0 as $n \to \infty$. Furthermore,

$$\frac{1}{m_n} \frac{p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y)}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to \eta f(Y|\sigma((a_0)^\top X + b_0), \nu_0)\bar{f}(X)$$

$$+ \sum_{\ell=0}^2 \alpha_\ell(X) \frac{\partial^\ell f}{\partial \sigma^\ell}(Y|\sigma((a')^\top X + b'), \nu')\bar{f}(X) = 0,$$

where

$$\eta = \lim_{n\to\infty} \frac{1}{m_n} \frac{(\lambda_n^* - \lambda_n)}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))}, \quad \alpha_0(X) = \lim_{n\to\infty} \frac{1}{m_n} \frac{-(\lambda_n^* - \lambda_n)}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))},$$

$$\alpha_1(X) = \lim_{n\to\infty} \Big[\sum_{u=1}^d \frac{1}{m_n} \frac{\lambda_n(\Delta a_n - \Delta a_n^*)^{(u)}}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \frac{\partial \sigma}{\partial a^{(u)}}(X, a', b') + \frac{1}{m_n} \frac{\lambda_n(\Delta b_n - \Delta b_n^*)}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \frac{\partial \sigma}{\partial b}(X, a', b') \Big],$$

$$\alpha_2(X) = \lim_{n\to\infty} \frac{1}{2m_n} \frac{\lambda_n(\Delta \nu_n - \Delta \nu_n^*)}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))}.$$

Recall that $D_2((\lambda_n, G_n), (\overline{\lambda}, \overline{G})) = \lambda_n \|(\Delta a_n, \Delta b_n, \Delta \nu_n)\|^2 \to 0$ and $D_2((\lambda_n^*, G_{*,n}), (\overline{\lambda}, \overline{G})) = \lambda_n^* \|(\Delta a_n^*, \Delta b_n^*, \Delta \nu_n^*)\|^2 \to 0$. Since $(a_n, b_n, \nu_n) \not\to (a_0, b_0, \nu_0)$ and $(a_n^*, b_n^*, \nu_n^*) \not\to (a_0, b_0, \nu_0)$, we get that $\lambda_n \to 0$ and $\lambda_n^* \to 0$. As a result, we achieve that $\eta = \alpha_0(X) = 0$, which leads to

$$\sum_{\ell=0}^2 \alpha_\ell(X) \frac{\partial^\ell f}{\partial \sigma^\ell}(Y|\sigma((a')^\top X + b'), \nu') = 0.$$

Additionally, as the set

$$\left\{ \frac{\partial^\tau f}{\partial \sigma^\tau}(Y|\sigma((a')^\top X + b'), \nu') : 0 \le \tau \le 2 \right\}$$

is linearly independent, then $\alpha_1(X) = \alpha_2(X) = 0$ for almost surely X. Moreover, $\sigma(X, a, b)$ and ν are algebraically independent and $1/m_n$ is bounded, we get that for all $u \in [d]$,

$$\lim_{n\to\infty} \frac{\lambda_n(\Delta a_n - \Delta a_n^*)^{(u)}}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} = 0,$$

$$\lim_{n\to\infty} \frac{\lambda_n(\Delta b_n - \Delta b_n^*)}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} = 0,$$

$$\lim_{n\to\infty} \frac{\lambda_n(\Delta \nu_n - \Delta \nu_n^*)}{D_1((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} = 0.$$

This contradicts the fact that not all the coefficients of elements in the set $\mathcal{K}$ vanish as $n \to \infty$. Hence, we reach the conclusion for this case. $\qquad \square$

A.6 Proof of Theorem 4

In this section, we consider both the experts φ and σ are of linear forms, and the density $f_0 = f$ is a Gaussian density. Building on the explanation in Section 3.2.1, we will focus only on the case when φ and σ are the same linear functions. The most challenging part is that there is an interaction between parameters b and ν via the following heat equation:

$$\frac{\partial^2 f}{\partial b^2}(Y|a^\top X + b, \nu) = 2\frac{\partial f}{\partial \nu}(Y|a^\top X + b, \nu).$$

Recall for the simplicity of the presentation in the paper, we define $\Delta G = (\Delta a, \Delta b, \Delta \nu) = (a - a_0, b - b_0, \nu - \nu_0)$ and $\Delta G_* = (\Delta a^*, \Delta b^*, \Delta \nu^*) = (a^* - a_0, b^* - b_0, \nu^* - \nu_0)$ for any element $(a, b, \nu), (a^*, b^*, \nu^*) \in \Theta$. For the convenience of readers, we will first restate the result of Theorem 4: denote for any sequence $(l_n)_{n \geq 1}$,

$$\Xi_1(l_n) := \left\{ (\lambda, G) = (\lambda, a, b, \nu) \in \Xi : \frac{l_n}{\min_{1 \leq i \leq d}\{|(\Delta a)_i|^2, |\Delta b|^4, |\Delta \nu|^2\}\sqrt{n}} \leq \lambda \right\}.$$

Then, the followings hold for any sequence $(l_n)_{n \geq 1}$ such that $l_n/\log n \to \infty$ as $n \to \infty$:

$$\sup_{(\lambda^*, G_*) \in \Xi_1(l_n)} \mathbb{E}_{p_{\lambda^*, G_*}}\left[(\|\Delta a^*\|^4 + |\Delta b^*|^8 + |\Delta \nu^*|^4) \times |\widehat{\lambda}_n - \lambda^*|^2 \right] \lesssim \frac{\log n}{n},$$

$$\sup_{(\lambda^*, G_*) \in \Xi_1(l_n)} \mathbb{E}_{p_{\lambda^*, G_*}}\left[(\lambda^*)^2 (\|\Delta a^*\|^2 + |\Delta b^*|^4 + |\Delta \nu^*|^2) \right.$$

$$\left. \times (\|\Delta \widehat{a}_n - \Delta a^*\|^2 + |\Delta \widehat{b}_n - \Delta b^*|^4 + |\Delta \widehat{\nu}_n - \Delta \nu^*|^2) \right] \lesssim \frac{\log n}{n}.$$

To prove the above results, we can establish the following result concerning the lower bound of $V(p_{\lambda, G}, p_{\lambda^*, G_*})$.

Lemma 3. *Assume that f_0 and f are both Gaussians, and the mean experts φ in f_0 and σ in f are both linear functions. Futhermore, if $f_0(Y|\varphi(a_0^\top X + b_0), \nu_0) = f(Y|\sigma(a_0^\top X + b_0), \nu_0)$ for some $(a_0, b_0, \nu_0) \in \Theta$. For any $(\lambda, G), (\lambda^*, G_*)$, let us define*

$$D_4((\lambda, G), (\lambda^*, G_*)) = \lambda \left(\|\Delta a\|^2 + |\Delta b|^4 + |\Delta \nu|^2 \right) + \lambda^* \left(\|\Delta a^*\|^2 + |\Delta b^*|^4 + |\Delta \nu^*|^2 \right)$$

$$- \min\{\lambda, \lambda^*\} \left[\|\Delta a\|^2 + |\Delta b|^4 + |\Delta \nu|^2 + \|\Delta a^*\|^2 + |\Delta b^*|^4 + |\Delta \nu^*|^2 \right]$$

$$+ \left[\lambda \left(\|\Delta a\| + |\Delta b|^2 + |\Delta \nu| \right) + \lambda^* \left(\|\Delta a^*\| + |\Delta b^*|^2 + |\Delta \nu^*| \right) \right] \times \left(\|\Delta a - \Delta a^*\| + |\Delta b - \Delta b^*|^2 + |\Delta \nu - \Delta \nu^*| \right).$$

Then, there exists a positive constant C_4 that depends on Θ, (a_0, b_0, ν_0) and satisfies the following:

$$V(p_{\lambda, G}, p_{\lambda^*, G_*}) \geq C_4 D_4((\lambda, G), (\lambda^*, G_*)).$$

for all (λ, G) and (λ^*, G_*)

Proof of Theorem 4. Similar to the proof in the previous section, we define:

$$\overline{D_4}((\lambda, G), (\lambda^*, G_*)) = (\lambda^* - \lambda)\left(\|\Delta a\| + |\Delta b|^2 + |\Delta \nu| \right)\left(\|\Delta a^*\| + |\Delta b^*|^2 + |\Delta \nu^*| \right)$$

$$+ \left(\lambda(\|\Delta a\| + |\Delta b|^2 + |\Delta \nu|) + \lambda^*(\|\Delta a^*\| + |\Delta b^*|^2 + |\Delta \nu^*|) \right) \cdot \left(\|\Delta a - \Delta a^*\| + |\Delta b - \Delta b^*|^2 + |\Delta \nu - \Delta \nu^*| \right).$$

It will be shown that $D_4((\lambda, G), (\lambda^*, G_*)) \asymp \overline{D_4}((\lambda, G), (\lambda^*, G_*))$. Considering the above formulation of $\overline{D_4}((\lambda, G), (\lambda^*, G_*))$ and assuming, without loss of generality, that $\lambda^* \geq \lambda$, we can infer that

$$\overline{D_4}((\lambda, G), (\lambda^*, G_*))$$

$$\lesssim (\lambda^* - \lambda)\left(\|\Delta a^*\| + \|\Delta a - \Delta a^*\| + |\Delta b^*|^2 + |\Delta b - \Delta b^*|^2 + |\Delta \nu^*| + |\Delta \nu - \Delta \nu^*| \right) \cdot \left(\|\Delta a^*\| + |\Delta b^*|^2 + |\Delta \nu^*| \right)$$

$$+ \left(\lambda(\|\Delta a\| + |\Delta b|^2 + |\Delta \nu|) + \lambda^*(\|\Delta a^*\| + |\Delta b^*|^2 + |\Delta \nu^*|) \right) \cdot \left(\|\Delta a - \Delta a^*\| + |\Delta b - \Delta b^*|^2 + |\Delta \nu - \Delta \nu^*| \right)$$

$$\lesssim (\lambda^* - \lambda)\left(\|\Delta a^*\| + |\Delta b^*|^2 + |\Delta \nu^*| \right)\left(\|\Delta a^*\| + |\Delta b^*|^2 + |\Delta \nu^*| \right)$$

$$+ \lambda^* \left(\|\Delta a^*\| + |\Delta b^*|^2 + |\Delta \nu^*| \right)\left(\|\Delta a - \Delta a^*\| + |\Delta b - \Delta b^*|^2 + |\Delta \nu - \Delta \nu^*| \right)$$

$$+ \left(\lambda(\|\Delta a\| + |\Delta b|^2 + |\Delta \nu|) + \lambda^*(\|\Delta a^*\| + |\Delta b^*|^2 + |\Delta \nu^*|) \right)\left(\|\Delta a - \Delta a^*\| + |\Delta b - \Delta b^*|^2 + |\Delta \nu - \Delta \nu^*| \right)$$

$$\lesssim D_4((\lambda, G), (\lambda^*, G_*)).$$

Fanqi Yan$^{\star,\diamond}$, Huy Nguyen*,†, Dung Le*,†, Pedram Akabrian‡, Nhat Ho†

On the other hand,

$$\overline{D_4}((\lambda, G), (\lambda^*, G_*)) \gtrsim (\lambda^* - \lambda)\Big(\|\Delta a\| + \|\Delta a - \Delta a^*\| + |\Delta b|^2 + |\Delta b - \Delta b^*|^2 + |\Delta \nu| + |\Delta \nu - \Delta \nu^*|\Big) \cdot$$
$$\Big(\|\Delta a^*\| + |\Delta b^*|^2 + |\Delta \nu^*|\Big) + \Big(\lambda(\|\Delta a\| + |\Delta b|^2 + |\Delta \nu|) + \lambda^*(\|\Delta a^*\| + |\Delta b^*|^2 + |\Delta \nu^*|)\Big) \cdot \Big(\|\Delta a - \Delta a^*\| +$$
$$|\Delta b - \Delta b^*|^2 + |\Delta \nu - \Delta \nu^*|\Big) \gtrsim D_4((\lambda, G), (\lambda^*, G_*))$$

where the last inequality come from the triangle inequality and Cauchy–Schwarz inequality. Now we could conclude that $D_4((\lambda, G), (\lambda^*, G_*)) \asymp \overline{D_4}((\lambda, G), (\lambda^*, G_*))$, i.e.

$$D_4((\lambda, G), (\lambda^*, G_*)) \asymp |\lambda - \lambda^*|(\|\Delta a\| + |\Delta b|^2 + |\Delta \nu|)(\|\Delta a^*\| + |\Delta b^*|^2 + |\Delta \nu^*|)$$
$$+ (\lambda^*(\|\Delta a^*\| + |\Delta b^*|^2 + |\Delta \nu^*|) + \lambda(\|\Delta a\| + |\Delta b|^2 + |\Delta \nu|))(\|a - a^*\| + |b - b^*|^2 + |\nu - \nu^*|)$$

Combining this result with Theorem 1 and Lemma 3, we have

$$\sup_{(\lambda^*, G_*) \in \Xi_1(l_n)} \mathbb{E}_{p_{\lambda^*, G_*}} \Big((\|\Delta \widehat{a}_n\|^2 + |\Delta \widehat{b}_n|^4 + |\Delta \widehat{\nu}_n|^2)(\|\Delta a^*\|^2 + |\Delta b^*|^4 + |\Delta \nu^*|^2)|\widehat{\lambda}_n - \lambda^*|^2 \Big) \lesssim \frac{\log n}{n}, \quad (32)$$

$$\sup_{(\lambda^*, G_*) \in \Xi_1(l_n)} \mathbb{E}_{p_{\lambda^*, G_*}} \Big((\lambda^*)^2 (\|\Delta a^*\|^2 + |\Delta b^*|^4 + |\Delta \nu^*|^2)(\|\widehat{a}_n - a^*\|^2 + |\widehat{b}_n - b^*|^4 + |\widehat{\nu}_n - \nu^*|^2) \Big) \lesssim \frac{\log n}{n}. \quad (33)$$

We follow the same argument as in the proof of Theorem 9. Applying the Cauchy-Schwartz inequality, we can bound $\mathbb{E}_{p_{\lambda^*, G_*}} \Big(\|\Delta a^*\|^4 + |\Delta b^*|^8 + |\Delta \nu^*|^4 |\widehat{\lambda}_n - \lambda^*|^2 \Big)$ by

$$\mathbb{E}_{p_{\lambda^*, G_*}} \Big((\|\Delta a^*\|^2 + |\Delta b^*|^4 + |\Delta \nu^*|^2)(\|\Delta \widehat{a}_n\|^2 + |\Delta \widehat{b}_n|^4 + |\Delta \widehat{\nu}_n|^2)|\widehat{\lambda}_n - \lambda^*|^2 \Big)$$
$$+ \mathbb{E}_{p_{\lambda^*, G_*}} \Big((\|\Delta a^*\|^2 + |\Delta b^*|^4 + |\Delta \nu^*|^2)(\|\Delta \widehat{a}_n - \Delta a^*\|^2 + |\Delta \widehat{b}_n - \Delta b^*|^4 + |\Delta \widehat{\nu}_n - \Delta \nu^*|^2)|\widehat{\lambda}_n - \lambda^*|^2 \Big)$$

The first term has been estimated in equation (32). For the second term, if $\lambda^* \neq 0$ it is in fact an implication of equation (33), noting that $\limsup \frac{|\widehat{\lambda}_n - \lambda^*|}{\lambda^*} \leq \frac{2}{\lambda^*}$. Otherwise, if $\lambda^* = 0$, we have

$$D_4((\widehat{\lambda}_n, \widehat{G}_n), (\lambda^*, G_*)) = \widehat{\lambda}_n(\|\Delta \widehat{a}_n\|^2 + |\Delta \widehat{b}_n|^4 + |\Delta \widehat{\nu}_n|^2)$$
$$+ \widehat{\lambda}_n(\|\Delta \widehat{a}_n\| + |\Delta \widehat{b}_n|^2 + |\Delta \widehat{\nu}_n|)(\|\Delta \widehat{a}_n - \Delta a^*\| + |\Delta \widehat{b}_n - \Delta b^*|^2 + |\Delta \widehat{\nu}_n - \Delta \nu^*|)$$
$$\geq \widehat{\lambda}_n(\|\Delta \widehat{a}_n\| + |\Delta \widehat{b}_n|^2 + |\Delta \widehat{\nu}_n|)(\|\Delta \widehat{a}_n - \Delta a^*\| + |\Delta \widehat{b}_n - \Delta b^*|^2 + |\Delta \widehat{\nu}_n - \Delta \nu^*|).$$

The conclusion follows from direct application of Lemma 3 and Theorem 1. $\qquad \square$

We will now prove Lemma 3.

Proof of Lemma 3. Given $\overline{G} = (\bar{a}, \bar{b}, \bar{\nu})$, where $\bar{\lambda} \in [0, 1]$ and $(\bar{a}, \bar{b}, \bar{\nu})$ can be identical to (a_0, b_0, ν_0), we will prove that

(i) If $(a_0, b_0, \nu_0) \neq (\bar{a}, \bar{b}, \bar{\nu})$ and $\bar{\lambda} > 0$, then

$$\lim_{\varepsilon \to 0} \inf_{(\lambda, G), (\lambda^*, G_*)} \left\{ \frac{\|p_{\lambda, G} - p_{\lambda^*, G_*}\|_\infty}{D_1((\lambda, G), (\lambda^*, G_*))} : D_1((\lambda, G), (\bar{\lambda}, \overline{G})) \vee D_1((\lambda^*, G_*), (\bar{\lambda}, \overline{G})) \leq \varepsilon \right\} > 0.$$

(ii) If $(a_0, b_0, \nu_0) = (\bar{a}, \bar{b}, \bar{\nu})$ or $(a_0, b_0, \nu_0) \neq (\bar{a}, \bar{b}, \bar{\nu})$ and $\bar{\lambda} = 0$, then

$$\lim_{\varepsilon \to 0} \inf_{(\lambda, G), (\lambda^*, G_*)} \left\{ \frac{\|p_{\lambda, G} - p_{\lambda^*, G_*}\|_\infty}{D_4((\lambda, G), (\lambda^*, G_*))} : D_4((\lambda, G), (\bar{\lambda}, \overline{G})) \vee D_4((\lambda^*, G_*), (\bar{\lambda}, \overline{G})) \leq \varepsilon \right\} > 0.$$

Note that we can apply the arguments in the proof of Theorem 2 for part (i). Therefore, we will provide proof for only part (ii) in this section, particularly the most demanding setting when $(a_0, b_0, \nu_0) = (\bar{a}, \bar{b}, \bar{\nu})$.

Assume that the conclusion in part (ii) does not hold, we can find two sequences $G_n = (a_n, b_n, \nu_n)$ and $G_{*,n} = (a_n^*, b_n^*, \nu_n^*)$ and two sequences of mixing proportions λ_n and $\lambda_n^* \in [0, 1]$ that satisfy the following properties

$$\begin{cases} D_4((\lambda_n, G_n), (\bar{\lambda}, \overline{G})) \to 0, \\ D_4((\lambda_n^*, G_{*,n}), (\bar{\lambda}, \overline{G})) \to 0, \\ \|p_{\lambda_n, G_n} - p_{\lambda_n^*, G_{*,n}}\|_\infty / D_4((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0. \end{cases}$$

when $n \to \infty$. WLOG, we assume that $\lambda_n^* \geq \lambda_n$. Now, we take into account only the case when $\|(\Delta a_n, \Delta b_n, \Delta \nu_n)\| \to 0$ and $\|(\Delta a_n^*, \Delta b_n^*, \Delta \nu_n^*)\| \to 0$, while the other two cases can be argued in the same way as in the proof of Theorem 9 , Case 2 and Case 3. Here $\Delta a_n = a_n - a_0$, $\Delta a_n^* = a_n^* - a_0$.

We consider

$$\begin{aligned} p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y) &= (\lambda_n^* - \lambda_n)[f(Y|\varphi(a_0^\top X + b_0), \nu_0) - f(Y|\sigma\left((a_n^*)^\top X + b_n^*\right), \nu_n^*)]\bar{f}(X) \\ &\quad + \lambda_n[f(Y|\sigma\left(a_n^\top X + b_n\right), \nu) - f(Y|\sigma\left((a_n^*)^\top X + b_n^*\right), \nu_n^*)]\bar{f}(X) \\ &= S_n + T_n. \end{aligned}$$

WLOG, since $\varphi(\cdot)$ and $\sigma(\cdot)$ are the same linear function, for simplicity, we could just consider that they are both identity. Also for the simplicity of the proof, we will consider the parameter vector a in the mean expert is just a one-dimension vector. By means of Taylor expansion up to the fourth order, we get that

$$\begin{aligned} S_n &= (\lambda_n^* - \lambda_n)\Big[\sum_{|\alpha|=1}^{4} \frac{1}{\alpha!}(-\Delta a_n^*)^{\alpha_1}(-\Delta b_n^*)^{\alpha_2}(-\Delta \nu_n^*)^{\alpha_3} \\ &\quad \times \frac{\partial^{|\alpha|} f}{\partial a^{\alpha_1} \partial b^{\alpha_2} \partial \nu^{\alpha_3}}(Y|\sigma\left((a_n^*)^\top X + b_n^*\right), \nu_n^*) + R_1(X, Y)\Big]\bar{f}(X) \\ &= (\lambda_n^* - \lambda_n)\Big[\sum_{|\alpha|=1}^{4} \frac{1}{\alpha!}\{-\Delta a_n^*\}^{\alpha_1}\{-\Delta b_n^*\}^{\alpha_2}\{-\Delta \nu_n^*\}^{\alpha_3} \\ &\quad \times \frac{X^{\alpha_1}}{2^{\alpha_3}} \frac{\partial^{\alpha_1+\alpha_2+2\alpha_3} f}{\partial \sigma^{\alpha_1+\alpha_2+2\alpha_3}}(Y|\sigma\left((a_n^*)^\top X + b_n^*\right), \nu_n^*) + R_1(X, Y)\Big]\bar{f}(X) \\ &= (\lambda_n^* - \lambda_n)\Big[\sum_{\alpha_1=0}^{4} \sum_{\ell=0}^{2(4-\alpha_1)} \sum_{\alpha_2, \alpha_3} \frac{1}{2^{\alpha_3}\alpha!}\{-\Delta a_n^*\}^{\alpha_1}\{-\Delta b_n^*\}^{\alpha_2}\{-\Delta \nu_n^*\}^{\alpha_3} \\ &\quad \times X^{\alpha_1} \frac{\partial^{\ell+\alpha_1} f}{\partial \sigma^{\ell+\alpha_1}}(Y|\sigma\left((a_n^*)^\top X + b_n^*\right), \nu_n^*) + R_1(X, Y)\Big]\bar{f}(X), \end{aligned} \tag{34}$$

where α_2 and α_3 in the above sum satisfy $\alpha_2 + 2\alpha_3 = \ell$ and $0 \vee 1 - \alpha_1 \leq \alpha_2 + \alpha_3 \leq 4 - \alpha_1$. In addition, $R_1(X, Y)$ is the Taylor remainder such that $(\lambda_n^* - \lambda_n)|R_1(X, Y)|/D_4((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0$ as $n \to \infty$. Analogously, we have

$$\begin{aligned} T_n &= \lambda_n\Big[\sum_{\alpha_1=0}^{4} \sum_{\ell=0}^{2(4-\alpha_1)} \sum_{\alpha_2, \alpha_3} \frac{1}{2^{\alpha_3}\alpha!}\{\Delta a_n - \Delta a_n^*\}^{\alpha_1}\{\Delta b_n - \Delta b_n^*\}^{\alpha_2}\{\Delta \nu_n - \Delta \nu_n^*\}^{\alpha_3} \\ &\quad \times X^{\alpha_1} \frac{\partial^{\ell+\alpha_1} f}{\partial \sigma^{\ell+\alpha_1}}(Y|\sigma\left((a_n^*)^\top X + b_n^*\right), \nu_n^*) + R_2(X, Y)\Big]\bar{f}(X), \end{aligned} \tag{35}$$

where $R_2(X, Y)$ is the Taylor remainder such that $\lambda_n|R_2(X, Y)|/D_4((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0$.

Subsequently, we define

$$\mathcal{F}_n = \left\{ X^{\alpha_1} \frac{\partial^{\ell+\alpha_1} f}{\partial \sigma^{\ell+\alpha_1}}(Y|\sigma\left((a_n^*)^\top X + b_n^*\right), \nu_n^*)\bar{f}(X) : 0 \leq \alpha_1 \leq 4, 0 \leq \ell \leq 2(4-\alpha_1) \right\}.$$

Fanqi Yan[★,◇], **Huy Nguyen**[★,†], **Dung Le**[★,†], **Pedram Akabrian**[‡], **Nhat Ho**[†]

From equations (34) and (35), the quantity $p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y)$ can be viewed as a linear combination of elements in the set $\mathcal{F}_n$. Let us denote by $E_{\alpha_1, \ell}$ the coefficients of those elements in the representation of $[p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y)]$, i.e.

$$E_{\alpha_1, \ell} = \sum_{\alpha_2 + 2\alpha_3 = \ell} \frac{1}{2^{\alpha_3} \alpha!} \Big[(\lambda_n^* - \lambda_n)(-\Delta a_n^*)^{\alpha_1} (-\Delta b_n^*)^{\alpha_2} (-\Delta \nu_n^*)^{\alpha_3}$$
$$+ \lambda_n \{\Delta a_n - \Delta a_n^*\}^{\alpha_1} \{\Delta b_n - \Delta b_n^*\}^{\alpha_2} \{\Delta \nu_n - \Delta \nu_n^*\}^{\alpha_3} \Big].$$

Assume by contrary that all the coefficients of elements in the set $\mathcal{F}_n$ in the representation of $[p_{\lambda_n, G_n}(X, Y) - p_{\lambda_n^*, G_{*,n}}(X, Y)]/D_4((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))$ vanish when n tends to infinity, for all $0 \leq \alpha_1 \leq 4$ and $0 \vee 1 - \alpha_1 \leq \ell \leq 2(4 - \alpha_1)$, i.e.

$$E_{\alpha_1, \ell}/D_4((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0. \tag{36}$$

Now, we will divide the proof into two main scenarios:

Case 1:

The following ratio does not tend to infinity:

$$R_n = \frac{\lambda_n(|\Delta a_n| + |\Delta b_n|^2 + |\Delta \nu_n|) + \lambda_n^*(|\Delta a_n^*| + |\Delta b_n^*|^2 + |\Delta \nu_n^*|)}{\lambda_n(|\Delta a_n - \Delta a_n^*| + |\Delta b_n - \Delta b_n^*|^2 + |\Delta \nu_n - \Delta \nu_n^*|)} \not\to \infty.$$

It follows that as n is large enough

$$D_4((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \lesssim (\lambda_n^* - \lambda_n)(|\Delta a_n^*|^2 + |\Delta b_n^*|^4 + |\Delta \nu_n^*|^2)$$
$$+ \lambda_n(|\Delta a_n - \Delta a_n^*|^2 + |\Delta b_n - \Delta b_n^*|^4 + |\Delta \nu_n - \Delta \nu_n^*|^2)$$
$$=: H_n.$$

Putting this result together with equation (36), we have $E_{\alpha_1, \ell}/H_n \to 0$ as $n \to \infty$ for any feasible α_1, ℓ. Then, for $\alpha_1 = 2$ and $\ell = 0$, we obtain

$$\frac{E_{2,0}}{H_n} = \frac{(\lambda_n^* - \lambda_n)|\Delta a_n^*|^2 + \lambda_n|\Delta a_n - \Delta a_n^*|^2}{H_n} \to 0,$$

which leads to

$$\frac{(\lambda_n^* - \lambda_n)(|\Delta b_n^*|^4 + |\Delta \nu_n^*|^2) + \lambda_n(|\Delta b_n - \Delta b_n^*|^4 + |\Delta \nu_n - \Delta \nu_n^*|^2)}{H_n} \to 1.$$

As a consequence, we get

$$T_{0,\ell} := \frac{\sum_{\substack{\alpha_2 + 2\alpha_3 = \ell \\ \alpha_2 + \alpha_3 \leq 4}} \frac{1}{2^{\alpha_3} \alpha_2! \alpha_3!} [(\lambda_n^* - \lambda_n)\{-\Delta b_n^*\}^{\alpha_2}\{-\Delta \nu_n^*\}^{\alpha_3} + \lambda_n\{\Delta b_n - \Delta b_n^*\}^{\alpha_2}\{\Delta \nu_n - \Delta \nu_n^*\}^{\alpha_3}]}{(\lambda_n^* - \lambda_n)(|\Delta b_n^*|^4 + |\Delta \nu_n^*|^2) + \lambda_n(|\Delta b_n - \Delta b_n^*|^4 + |\Delta \nu_n - \Delta \nu_n^*|^2)}$$
$$= \frac{E_{0,\ell}}{(\lambda_n^* - \lambda_n)(|\Delta b_n^*|^4 + |\Delta \nu_n^*|^2) + \lambda_n(|\Delta b_n - \Delta b_n^*|^4 + |\Delta \nu_n - \Delta \nu_n^*|^2)}$$
$$= \frac{E_{0,\ell}/H_n}{\Big((\lambda_n^* - \lambda_n)(|\Delta b_n^*|^4 + |\Delta \nu_n^*|^2) + \lambda_n(|\Delta b_n - \Delta b_n^*|^4 + |\Delta \nu_n - \Delta \nu_n^*|^2)\Big)/H_n}$$
$$\to \frac{0}{1} = 0.$$

Since $(\lambda_n^* - \lambda_n)$ and λ_n play the same role, we can assume that $\lambda_n^* - \lambda_n \leq \lambda_n$ for all $n \to \infty$.

Case 1.1: $(\lambda_n^* - \lambda_n)/\lambda_n \not\to 0$ as $n \to \infty$.

Under this setting, we define $p_n = \max\{\lambda_n^* - \lambda_n, \lambda_n\}$ and

$$M_n = \max\{|\Delta b_n^*|, |\Delta b_n - \Delta b_n^*|, |\Delta \nu_n^*|^{1/2}, |\Delta \nu_n - \Delta \nu_n^*|^{1/2}\}.$$

In addition, we denote

$$
\begin{aligned}
(\lambda_n^* - \lambda_n)/p_n &\to c_1^2, & \lambda_n/p_n &\to c_2^2, \\
\Delta b_n^*/M_n &\to -a_1, & (\Delta b_n - \Delta b_n^*)/M_n &\to a_2, \\
\tfrac{1}{2}\Delta \nu_n^*/M_n^2 &\to -b_1, & \tfrac{1}{2}(\Delta \nu_n - \Delta \nu_n^*)/M_n^2 &\to b_2.
\end{aligned}
$$

Note that at least one among a_1, b_1, a_2, b_2 and both c_1, c_2 are different from zero. Next, for each $\ell \in [4]$, we divide both the numerators and the denominators of $T_{0,\ell}$ by $p_n M_n^\ell$, then we achieve the following system of polynomial equations

$$\sum_{\alpha_2 + 2\alpha_3 = \ell} \frac{1}{\alpha_2! \alpha_3!}(c_1^2 a_1^{\alpha_2} b_1^{\alpha_3} + c_2^2 a_2^{\alpha_2} b_2^{\alpha_3}), \quad \forall \ell \in [4].$$

According to Propsition 2.1 in Ho and Nguyen (2016a), this system admits only the trivial solution $a_1 = b_1 = a_2 = b_2 = 0$, which is a contradiction. Thus, the setting that $(\lambda_n^* - \lambda_n)/\lambda_n \nrightarrow 0$ as $n \to \infty$ cannot occur.

Case 1.2: $(\lambda_n^* - \lambda_n)/\lambda_n \to 0$, or equivalently $\lambda_n^*/\lambda_n \to 1$ as $n \to \infty$.

Now, we consider the following quantity:

$$T_{0,1} = \frac{(\lambda_n^* - \lambda_n)(-\Delta b_n^*) + \lambda_n(\Delta b_n - \Delta b_n^*)}{(\lambda_n^* - \lambda_n)(|\Delta b_n^*|^4 + |\Delta \nu_n^*|^2) + \lambda_n(|\Delta b_n - \Delta b_n^*|^4 + |\Delta \nu_n - \Delta \nu_n^*|^2)}.$$

By dividing both the numerator and the denominator of $T_{0,1}$ by $\lambda_n M_n$ (with a convention that if the new denominator goes to 0, then the new numerator also goes to 0), we get

$$\frac{\lambda_n^* - \lambda_n}{\lambda_n} \cdot \frac{(-\Delta b_n^*)}{M_n} + \frac{\Delta b_n - \Delta b_n^*}{M_n} \to 0.$$

As $(\lambda_n^* - \lambda_n)/\lambda_n \to 0$ and $|\Delta b_n^*| \le M_n$, then the first term in the above sum tends to 0. This implies that

$$(\Delta b_n - \Delta b_n^*)/M_n \to 0. \tag{37}$$

Similarly, we continue to divide both the numerator and the denominator of $T_{0,2}$ by $\lambda_n M_n^2$, and obtain that

$$\frac{(\lambda_n^* - \lambda_n)(-\Delta b_n^*)^2 + \lambda_n(\Delta b_n - \Delta b_n^*)^2}{\lambda_n M_n^2} + \frac{(\lambda_n^* - \lambda_n)(-\Delta \nu_n^*) + \lambda_n(\Delta \nu_n - \Delta \nu_n^*)}{\lambda_n M_n^2} \to 0.$$

As demonstrated above, the first term in the above sum goes to 0 when $n \to \infty$, therefore, the second term also tends to 0, i.e.

$$\frac{\lambda_n^* - \lambda_n}{\lambda_n} \cdot \frac{\Delta \nu_n^*}{M_n^2} + \frac{\Delta \nu_n - \Delta \nu_n^*}{M_n^2} \to 0.$$

Recall that $(\lambda_n^* - \lambda_n)/\lambda_n \to 0$ and $|\Delta \nu_n^*| \le M_n^2$, then $\dfrac{\lambda_n^* - \lambda_n}{\lambda_n} \cdot \dfrac{\Delta \nu_n^*}{M_n^2} \to 0$, leading to

$$\frac{\Delta \nu_n - \Delta \nu_n^*}{M_n^2} \to 0 \tag{38}$$

Assume by contrary that $M_n \in \{|\Delta b_n - \Delta b_n^*|, |\Delta \nu_n - \Delta \nu_n^*|^{1/2}\}$, then it follows from equations (37) and (38) that

$$1 = \frac{\max\{|\Delta b_n - \Delta b_n^*|^2, |\Delta \nu_n - \Delta \nu_n^*|\}}{M_n^2} \to 0,$$

which is a contradiction. As a result, $M_n \in \{|\Delta b_n^*|, |\Delta \nu_n^*|^{1/2}\}$.

Fanqi Yan[⋆,◇]**, Huy Nguyen**[⋆,†]**, Dung Le**[⋆,†]**, Pedram Akabrian**[‡]**, Nhat Ho**[†]

For the sake of simplicity, we will consider only the setting when $M_n = |\Delta b_n^*|$, whereas the other case can be argued similarly. Subsequently, we continue to consider two small cases of this setting.

Case 1.2.1: $(\lambda_n^* - \lambda_n)|\Delta b_n^*|^4 \leq \lambda_n|\Delta b_n - \Delta b_n^*|^4$ for all n.

Assume by contrary that $|\Delta b_n - \Delta b_n^*|/|\Delta \nu_n - \Delta \nu_n^*|^{1/2} \not\to 0$. By dividing both the numerator and the denominator of $T_{0,1}$ by $\lambda_n|\Delta b_n - \Delta b_n^*|$, we get

$$\frac{\lambda_n^* - \lambda_n}{\lambda_n} \cdot \frac{\Delta b_n^*}{\Delta b_n - \Delta b_n^*} \to 1.$$

From equation (37) , we have $|\Delta b_n - \Delta b_n^*|/|\Delta b_n^*| \to 0$, which together with the above equation deduces that

$$\frac{\lambda_n^* - \lambda_n}{\lambda_n} \cdot \frac{|\Delta b_n^*|^4}{|\Delta b_n - \Delta b_n^*|^4} \to \infty.$$

This contradicts the assumption of Case 1.2.1. Consequently, we must have that

$$\frac{|\Delta b_n - \Delta b_n^*|}{|\Delta \nu_n - \Delta \nu_n^*|^{1/2}} \to 0.$$

WLOG, assume that $\Delta \nu_n - \Delta \nu_n^* > 0$ for all n. Additionally, let us denote

$$-\Delta b_n^* = r_1^n (\Delta \nu_n - \Delta \nu_n^*)^{1/2},$$
$$\Delta \nu_n^* = r_2^n (\Delta \nu_n - \Delta \nu_n^*).$$

By dividing the numerators and the denominators of $T_{0,\ell}$ by $\lambda_n(\Delta \nu_n - \Delta \nu_n^*)^{\ell/2}$ for any $\ell \in [3]$, then the new denominators will go to zero. Thus, all the new numerators also tend to 0, i.e.

$$\begin{cases} \dfrac{\lambda_n^* - \lambda_n}{\lambda_n} \cdot r_1^n \to 0, \\[2mm] \dfrac{\lambda_n^* - \lambda_n}{\lambda_n} \cdot [(r_1^n)^2 + r_2^n] + 1 \to 0, \\[2mm] \dfrac{\lambda_n^* - \lambda_n}{\lambda_n} \cdot \left[\dfrac{(r_1^n)^3}{6} + \dfrac{r_1^n r_2^n}{2} \right] \to 0. \end{cases}$$

Note that equation (38) indicates that $1/|r_1^n| \to 0$ as $n \to \infty$. Therefore, the third limit in the above system is reduced to

$$\frac{\lambda_n^* - \lambda_n}{\lambda_n} \cdot \left[\frac{(r_1^n)^2}{3} + r_2^n \right] \to 0.$$

Putting this result together with the second limit, we obtain that

$$\frac{\lambda_n^* - \lambda_n}{\lambda_n} \cdot \frac{2}{3}(r_1^n)^2 + 1 \to 0.$$

However, the left hand side of the above equation cannot converge to 0 as it is not less than one. Thus, case 1.2.1 cannot occur.

Case 1.2.2: $(\lambda_n^* - \lambda_n)|\Delta b_n^*|^4 > \lambda_n|\Delta b_n - \Delta b_n^*|^4$ for all n.

It is worth noting that we can achieve that $(\lambda_n^* - \lambda_n)|\Delta b_n^*|^4 > \lambda_n|\Delta \nu_n - \Delta \nu_n^*|^2$ for all n by arguing similarly to case 1.2.1. Next, let us denote

$$\Delta b_n - \Delta b_n^* = s_1^n (-\Delta b_n^*),$$
$$-\Delta \nu_n^* = s_2^n (\Delta b_n^*)^2,$$
$$\Delta \nu_n - \Delta \nu_n^* = s_3^n (\Delta b_n^*)^2.$$

For $\ell \in [4]$, we divide both the numerators and the denominators of $T_{0,\ell}$ by $(\lambda_n^* - \lambda_n)(-\Delta b_n^*)^\ell$. Then, we see that the new numerators goes to 0, i.e.

$$1 + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot s_1^n \to 0,$$

$$\left[1 + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot (s_1^n)^2\right] + s_2^n + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot s_3^n \to 0,$$

$$\frac{1}{6}\left[1 + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot (s_1^n)^3\right] + \frac{1}{2}\left[s_2^n + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot s_1^n s_3^n\right] \to 0,$$

$$\frac{1}{24}\left[1 + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot (s_1^n)^4\right] + \frac{1}{4}\left[s_2^n + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot (s_3^n)^2\right] + \frac{1}{8}\left[(s_2^n)^2 + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot (s_3^n)^2\right] \to 0.$$

Recall that $M_n = |\Delta b_n^*|$, then $|s_i^n| \le 1$ for all $i \in [3]$. Suppose that $s_i^n \to s_i$ for any $i \in [3]$. From equations (37) and (38), we know that $s_1 = s_3 = 0$. Looking at the above system, the first and third limits yield that $s_2 = -\frac{1}{3}$, while the second and the fourth indicate that

$$\frac{1}{6} + s_2 + \frac{1}{2}s_2^2 = 0,$$

which contradicts the fact that $s_2 = -\frac{1}{3}$. Thus, case 1.2.2 does not hold.

Case 2:

The following ratio tends to infinity:

$$R_n = \frac{\lambda_n(|\Delta a_n| + |\Delta b_n|^2 + |\Delta \nu_n|) + \lambda_n^*(|\Delta a_n^*| + |\Delta b_n^*|^2 + |\Delta \nu_n^*|)}{\lambda_n(|\Delta a_n - \Delta a_n^*| + |\Delta b_n - \Delta b_n^*|^2 + |\Delta \nu_n - \Delta \nu_n^*|)} \to \infty. \tag{39}$$

Recall we have defined

$$\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) = (\lambda_n^* - \lambda_n)\Big(|\Delta a_n| + |\Delta b_n|^2 + |\Delta \nu_n|\Big)\Big(|\Delta a_n^*| + |\Delta b_n^*|^2 + |\Delta \nu_n^*|\Big)$$

$$+ \Big(\lambda_n(|\Delta a_n| + |\Delta b_n|^2 + |\Delta \nu_n|) + \lambda_n^*(|\Delta a_n^*| + |\Delta b_n^*|^2 + |\Delta \nu_n^*|)\Big) \cdot \Big(|\Delta a_n - \Delta a_n^*| + |\Delta b_n - \Delta b_n^*|^2 + |\Delta \nu_n - \Delta \nu_n^*|\Big).$$

In the proof of Theorem 4, we already have $D_4((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \asymp \overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))$. Recall equation (36), we have $E_{\alpha_1,\ell}/D_4((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0$ as $n \to \infty$ for any feasible α_1, ℓ. Then, for $\alpha_1 = [4]$ and $0 \le \ell \le 2(4 - \alpha_1)$, we would have that

$$F_{\alpha_1,\ell} = E_{\alpha_1,\ell}/\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \to 0.$$

For convenience, we denote and recall the following terms that are frequently employed in our proof

$$A_n := -F_{1,0} = \frac{(\lambda_n^* - \lambda_n)\Delta a_n^* + \lambda_n(\Delta a_n^* - \Delta a_n)}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} = \frac{\lambda_n^* \Delta a_n^* - \lambda_n \Delta a_n}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0 \tag{40}$$

$$B_n := -F_{0,1} = \frac{(\lambda_n^* - \lambda_n)\Delta b_n^* + \lambda_n(\Delta b_n^* - \Delta b_n)}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} = \frac{\lambda_n^* \Delta b_n^* - \lambda_n \Delta b_n}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0 \tag{41}$$

$$C_n := F_{0,2} = \frac{(\lambda_n^* - \lambda_n)(\Delta b_n^*)^2 + \lambda_n(\Delta b_n^* - \Delta b_n)^2 + (\lambda_n^* - \lambda_n)(-\Delta \nu_n^*) + \lambda_n(\Delta \nu_n - \Delta \nu_n^*)}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0. \tag{42}$$

$$F_{0,3} := \frac{1}{6}\frac{(\lambda_n^* - \lambda_n)(-\Delta b_n^*)^3 + \lambda_n(\Delta b_n - \Delta b_n^*)^3}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \tag{43}$$

$$+ \frac{1}{2}\frac{(\lambda_n^* - \lambda_n)(-\Delta \nu_n^*)(-\Delta b_n^*) + \lambda_n(\Delta b_n - \Delta b_n^*)(\Delta \nu_n - \Delta \nu_n^*)}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0$$

$$F_{0,4} := \frac{1}{24}\frac{(\lambda_n^* - \lambda_n)(-\Delta b_n^*)^4 + \lambda_n(\Delta b_n - \Delta b_n^*)^4}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} + \frac{1}{8}\frac{(\lambda_n^* - \lambda_n)(-\Delta \nu_n^*)^2 + \lambda_n(\Delta \nu_n - \Delta \nu_n^*)^2}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \tag{44}$$

$$+ \frac{1}{4}\frac{(\lambda_n^* - \lambda_n)(-\Delta \nu_n^*)(-\Delta b_n^*)^2 + \lambda_n(\Delta b_n - \Delta b_n^*)(\Delta \nu_n - \Delta \nu_n^*)}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0.$$

Fanqi Yan$^{\star,\diamond}$, **Huy Nguyen***,†, **Dung Le***,†, **Pedram Akabrian**‡, **Nhat Ho**†

Also, let $M_n' = \max\{|\Delta a_n|, |\Delta b_n|^2, |\Delta \nu_n|, |\Delta a_n^*|, |\Delta b_n^*|^2, |\Delta \nu_n^*|\}$. We consider sub-cases based on the asymptotic behavior of λ_n^*/λ_n.

Case 2.1: $\lambda_n^*/\lambda_n \not\to 1$ and $\lambda_n^*/\lambda_n \not\to \infty$.

From the assumption about R_n at the beginning of Case 2, we would have

$$\frac{|\Delta a_n - \Delta a_n^*|}{M_n'} \to 0 \ , \ \frac{|\Delta b_n - \Delta b_n^*|^2}{M_n'} \to 0, \ \frac{|\Delta \nu_n - \Delta \nu_n^*|}{M_n'} \to 0. \tag{45}$$

Given the symmetry between $(|\Delta a_n|, |\Delta b_n|^2, |\Delta \nu_n|)$ and $(|\Delta a_n^*|, |\Delta b_n^*|^2, |\Delta \nu_n^*|)$, we can assume, without loss of generality, that $M_n' \in \max\{|\Delta a_n|, |\Delta b_n|^2, |\Delta \nu_n|\}$. Under this assumption, we encounter three distinct cases.

Case 2.1.1: $M_n' = |\Delta b_n|^2$. In this case, given that $\frac{|\Delta b_n - \Delta b_n^*|^2}{M_n'} \to 0$, which is equivalent to $\frac{|\Delta b_n - \Delta b_n^*|^2}{|\Delta b_n|^2} \to 0$, we achieve that $\frac{\Delta b_n}{\Delta b_n^*} \to 1$. Noting that $\lambda_n^*/\lambda_n \not\to 1$, we would have

$$(\lambda_n^* - \lambda_n)\Big(|\Delta a_n| + |\Delta b_n|^2 + |\Delta \nu_n|\Big)\Big(|\Delta a_n^*| + |\Delta b_n^*|^2 + |\Delta \nu_n^*|\Big) \asymp \lambda_n^*|\Delta b_n|^4$$

and

$$\Big(\lambda_n(|\Delta a_n| + |\Delta b_n|^2 + |\Delta \nu_n|) + \lambda_n^*(|\Delta a_n^*| + |\Delta b_n^*|^2 + |\Delta \nu_n^*|)\Big)\Big(|\Delta a_n - \Delta a_n^*| + |\Delta b_n - \Delta b_n^*|^2 + |\Delta \nu_n - \Delta \nu_n^*|\Big)$$
$$\lesssim \lambda_n^*|\Delta b_n|^2\Big(|\Delta a_n - \Delta a_n^*| + |\Delta b_n - \Delta b_n^*|^2 + |\Delta \nu_n - \Delta \nu_n^*|\Big) \lesssim \lambda_n^*|\Delta b_n|^4.$$

Thus, we have an asymptotic estimation $\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \asymp \lambda_n^*|\Delta b_n|^4$. By substituting this estimation to the equation (41), we would achieve

$$(\Delta b_n^*)^3 B_n \asymp \frac{(\lambda_n^* - \lambda_n)(\Delta b_n^*)^4 + \lambda_n(\Delta b_n^* - \Delta b_n)(\Delta b_n^*)^3}{\lambda_n^*|\Delta b_n|^4} \to 0$$

Meanwhile, following the discussion at the beginning of Case 2.1.1, we have the estimation for the terms in previous equation

$$\frac{(\lambda_n^* - \lambda_n)|\Delta b_n^*|^4}{\lambda_n^*|\Delta b_n|^4} \asymp 1, \quad \frac{\lambda_n(|\Delta b_n^* - \Delta b_n||\Delta b_n^*|^3}{\lambda_n^*|\Delta b_n|^4} \lesssim \frac{\lambda_n(\Delta b_n^* - \Delta b_n)}{\lambda_n|\Delta b_n|} \to 0$$

This leads to a contradiction, so this case is invalid.

Case 2.1.2: $M_n' = |\Delta \nu_n|$. In this case, given that $\frac{|\Delta \nu_n - \Delta \nu_n^*|}{M_n'} \to 0$, which is equivalent to $\frac{|\Delta \nu_n - \Delta \nu_n^*|}{|\Delta \nu_n|} \to 0$, we achieve that $\frac{\Delta \nu_n}{\Delta \nu_n^*} \to 1$. Using the same argument as in the Case 2.1.1, we would achieve the asymptotic behavior $\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \asymp \lambda_n^*|\Delta \nu_n|^2$. By substituting this estimation to the equation (42), we would have

$$|\Delta \nu_n|C_n \asymp \frac{(\lambda_n^* - \lambda_n)|\Delta b_n^*|^2|\Delta \nu_n| + \lambda_n(\Delta b_n^* - \Delta b_n)^2|\Delta \nu_n| + (\lambda_n^* - \lambda_n)(-\Delta \nu_n^*)|\Delta \nu_n| + \lambda_n(\Delta \nu_n - \Delta \nu_n^*)|\Delta \nu_n|}{\lambda_n^*|\Delta \nu_n|^2} \to 0.$$

Following the discussion at the beginning of Case 2.1 and Case 2.1.2, noting that $\lambda_n^*/\lambda_n \not\to \infty$, we would have the following estimation for the terms in previous equation

$$\frac{\lambda_n(\Delta b_n^* - \Delta b_n)^2|\Delta \nu_n|}{\lambda_n^*|\Delta \nu_n|^2} \asymp \frac{(\Delta b_n^* - \Delta b_n)^2}{|\Delta \nu_n|} \to 0, \quad \frac{\lambda_n(\Delta \nu_n - \Delta \nu_n^*)|\Delta \nu_n|}{\lambda_n^*|\Delta \nu_n|^2} \asymp \frac{\Delta \nu_n - \Delta \nu_n^*}{|\Delta \nu_n|} \to 0.$$

This would lead to

$$\frac{(\Delta b_n^*)^2 - \Delta \nu_n^*}{\nu_n^*} \asymp \frac{(\lambda_n^* - \lambda_n)(\Delta b_n^*)^2|\Delta \nu_n| + (\lambda_n^* - \lambda_n)(-\nu_n^*)|\Delta \nu_n|}{\lambda_n^*|\Delta \nu_n|^2} \to 0.$$

In other words, $(\Delta b_n^*)^2/\Delta \nu_n^* \to 1$. By substituting this estimation to B_n as in Case 2.1.1, we receive a contradiction.

Case 2.1.3: $M_n' = |\Delta a_n|$. Using the same argument as in Case 2.1.1 and 2.1.2, we have $\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \asymp \lambda_n^* |\Delta a_n^*|^2$. By substituting this estimation to equation (40), we would have

$$\Delta a_n^* A_n \asymp \frac{(\lambda_n^* - \lambda_n)(\Delta a_n^*)^2 + \lambda_n^*(\Delta a_n^* - \Delta a_n)\Delta a_n^*}{\lambda_n |\Delta a_n^*|^2} \to 1.$$

Meanwhile, following the estimation in equation (45), noting that $\lambda_n^*/\lambda_n \not\to 1$, we have

$$\frac{(\lambda_n^* - \lambda_n)(\Delta a_n^*)^2}{\lambda_n^* |\Delta a_n^*|^2} \asymp 1, \quad \frac{\lambda_n(\Delta a_n^* - \Delta a_n)\Delta a_n^*}{\lambda_n^* |\Delta a_n^*|^2} \asymp \frac{\Delta a_n^* - \Delta a_n}{\Delta a_n^*} \to 0.$$

This leads to a contradiction, so this case is also invalid.

Case 2.2: $\lambda_n^*/\lambda_n \to 1$. Then the equation (45) still holds. Given the symmetry between $(|\Delta a_n|, |\Delta b_n|^2, |\Delta \nu_n|)$ and $(|\Delta a_n^*|, |\Delta b_n^*|^2, |\Delta \nu_n^*|)$, we can assume, without loss of generality, that $M_n' \in \max\{|\Delta a_n|, |\Delta b_n|^2, |\Delta \nu_n|\}$. Under this assumption, we consider three distinct cases.

Case 2.2.1: $M_n' = |\Delta b_n|^2$. Using the same argument as at the beginiing of Case 2.1.1, noting that the equation (45) holds, we achieve that $\Delta b_n^*/\Delta b_n \to 1$. Assume that

$$\max\left\{\frac{\lambda_n |\Delta a_n - \Delta a_n^*|}{(\lambda_n^* - \lambda_n)|\Delta b_n|^2}, \frac{\lambda_n |\Delta b_n - \Delta b_n^*|^2}{(\lambda_n^* - \lambda_n)|\Delta b_n|^2}, \frac{\lambda_n |\Delta \nu_n - \Delta \nu_n^*|}{(\lambda_n^* - \lambda_n)|\Delta b_n|^2}\right\} \to \infty,$$

then we could conclude that

$$B_{2,n} := \frac{(\lambda_n^* - \lambda_n)(\Delta b_n \Delta b_n^*)^2}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \le \min\left\{\frac{(\lambda_n^* - \lambda_n)|\Delta b_n|^2}{\lambda_n |\Delta a_n - \Delta a_n^*|}, \frac{(\lambda_n^* - \lambda_n)|\Delta b_n|^2}{\lambda_n |\Delta b_n - \Delta b_n^*|^2}, \frac{(\lambda_n^* - \lambda_n)|\Delta b_n|^2}{\lambda_n |\Delta \nu_n - \Delta \nu_n^*|}\right\} \to 0.$$

From this, noting that $\Delta b_n \asymp \Delta b_n^*$, we could have

$$\frac{(\lambda_n^* - \lambda_n)(|\Delta a_n^*| + |\Delta b_n^*|^2 + |\Delta \nu_n^*|)(|\Delta a_n| + |\Delta b_n|^2 + |\Delta \nu_n|)}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \asymp \frac{(\lambda_n^* - \lambda_n)(\Delta b_n \Delta b_n^*)^2}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0. \quad (46)$$

Meanwhile, we also could have following estimation. Firstly,

$$\frac{\lambda_n (\Delta b_n)^2 (\Delta a_n - \Delta a_n^*)}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} = \frac{(\lambda_n \Delta a_n - \lambda_n^* \Delta a_n^*)(\Delta b_n)^2}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} + \frac{(\lambda_n^* - \lambda_n)(\Delta b_n)^2 \Delta a_n^*}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))}$$

$$= -(\Delta b_n)^2 A_n + \frac{\Delta a_n^*}{(\Delta b_n^*)^2} B_{2,n}.$$

Noting that $A_n, B_{2,n} \to 0$, and $|\Delta a_n^*| \lesssim |\Delta b_n^*|^2$, we have

$$\frac{\lambda_n (\Delta b_n)^2 (\Delta a_n - \Delta a_n^*)}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0.$$

Secondly,

$$\frac{\lambda_n (\Delta b_n)^2 |\Delta b_n - \Delta b_n^*|^2}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} = \frac{(\lambda_n \Delta b_n - \lambda_n^* \Delta b_n^*)(\Delta b_n)^2 (\Delta b_n - \Delta b_n^*)}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} + \frac{(\lambda_n^* - \lambda_n)(\Delta b_n)^2 (\Delta b_n - \Delta b_n^*)\Delta b_n^*}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))}$$

$$= (\Delta b_n)^2 (\Delta b_n - \Delta b_n^*) B_n + \frac{\Delta b_n - \Delta b_n^*}{\Delta b_n^*} B_{2,n}$$

Given that $B_n, B_{2,n} \to 0$ and $\frac{\Delta b_n - \Delta b_n^*}{\Delta b_n^*} \to 0$ (an implication from equation (45)), we achieve

$$\frac{\lambda_n (\Delta b_n)^2 |\Delta b_n - \Delta b_n^*|^2}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0.$$

Fanqi Yan[⋆,⋄], **Huy Nguyen**[⋆,†], **Dung Le**[⋆,†], **Pedram Akabrian**[‡], **Nhat Ho**[†]

Thirdly, by multiplying C_n with $(\Delta b_n)^2$, we achieve

$$
\begin{aligned}
(\Delta b_n)^2 C_n &= \frac{(\lambda_n^* - \lambda_n)(\Delta b_n^*)^4 + \lambda_n(\Delta b_n^* - \Delta b_n)^2(\Delta b_n)^2 + (\lambda_n \Delta \nu_n - \lambda_n^* \Delta \nu_n^*)(\Delta b_n)^2}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \\
&= \frac{(\lambda_n^* - \lambda_n)(\Delta b_n^*)(\Delta b_n)^3 + (\lambda_n^* \Delta b_n^* - \lambda_n \Delta b_n)(\Delta b_n^* - \Delta b_n)(\Delta b_n)^2 + (\lambda_n \Delta \nu_n - \lambda_n^* \Delta \nu_n^*)(\Delta b_n)^2}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \\
&= \frac{\Delta b_n}{\Delta b_n^*} B_{2,n} + (\Delta b_n^* - \Delta b_n)(\Delta b_n)^2 B_n + \frac{(\lambda_n \Delta \nu_n - \lambda_n^* \Delta \nu_n^*)(\Delta b_n)^2}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))}.
\end{aligned}
$$

As $C_n, B_{2,n}, B_n \to 0$, and $\Delta b_n^*/\Delta b_n \to 1$ as discussed at the beginning of Case 2.2.1, we obtain

$$
NB_n := \frac{(\lambda_n \Delta \nu_n - \lambda_n^* \Delta \nu_n^*)(\Delta b_n)^2}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0.
$$

Thus, we have the following estimation:

$$
\begin{aligned}
\frac{\lambda_n(\Delta b_n)^2 |\Delta \nu_n^* - \Delta \nu_n|^2}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} &= \frac{(\lambda_n \Delta \nu_n - \lambda_n^* \Delta \nu_n^*)(\Delta b_n)^2}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} + \frac{(\lambda_n^* - \lambda_n)(\Delta b_n)^2 \Delta \nu_n^*}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \\
&= NB_n + \frac{\Delta \nu_n^*}{(\Delta b_n^*)^2} B_{2,n} \to 0
\end{aligned}
$$

thanks to $NB_n, B_{2,n} \to 0$, and $\dfrac{|\Delta \nu_n^*|}{(\Delta b_n^*)^2} \lesssim \dfrac{|\Delta \nu_n^*|}{(\Delta b_n)^2} \lesssim \dfrac{|\Delta \nu_n^*|}{(\Delta b_n)^2} = \dfrac{|\Delta \nu_n^*|}{M_n'} \lesssim 1$.

From these estimation, we achieve

$$
\begin{aligned}
&\frac{(\lambda_n(|\Delta a_n| + |\Delta b_n|^2 + |\Delta \nu_n|) + \lambda_n^*(|\Delta a_n^*| + |\Delta b_n^*|^2 + |\Delta \nu_n^*|))(|\Delta a_n - \Delta a_n^*| + |\Delta b_n - \Delta b_n^*|^2 + |\Delta \nu_n - \Delta \nu_n^*|)}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \\
&\lesssim \frac{\lambda_n |\Delta b_n|^2(|\Delta a_n - \Delta a_n^*| + |\Delta b_n - \Delta b_n^*|^2 + |\Delta \nu_n - \Delta \nu_n^*|)}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0.
\end{aligned}
\tag{47}
$$

By adding equation (46) and equation (47), we would receive that

$$
\frac{\overline{\overline{D_4}}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0,
$$

which is a contradiction. Thus, we conclude that

$$
\max\left\{ \frac{\lambda_n |\Delta a_n - \Delta a_n^*|}{(\lambda_n^* - \lambda_n)|\Delta b_n|^2}, \; \frac{\lambda_n |\Delta b_n - \Delta b_n^*|^2}{(\lambda_n^* - \lambda_n)|\Delta b_n|^2}, \; \frac{\lambda_n |\Delta \nu_n - \Delta \nu_n^*|}{(\lambda_n^* - \lambda_n)|\Delta b_n|^2} \right\} \not\to \infty.
\tag{48}
$$

Thus, from the formulation of $\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))$, noting that $\Delta b_n^*/\Delta b_n \to 1$, we could conclude that

$$
\begin{aligned}
\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) &\asymp (\lambda_n^* - \lambda_n)(\Delta b_n^*)^2(\Delta b_n)^2 + \lambda_n^* |\Delta b_n|^2(|\Delta a_n - \Delta a_n^*| + |\Delta b_n - \Delta b_n^*|^2 + |\Delta \nu_n - \Delta \nu_n^*|) \\
&\asymp (\lambda_n^* - \lambda_n)(\Delta b_n^*)^4 + (\lambda_n^* - \lambda_n)(\Delta b_n^*)^4 \quad \text{(by equation (48))} \\
&\asymp (\lambda_n^* - \lambda_n)(\Delta b_n^*)^4.
\end{aligned}
$$

Additionally, let us denote

$$
\Delta b_n - \Delta b_n^* = r_1^n \Delta b_n^*, \quad \Delta \nu_n - \Delta \nu_n^* = r_2^n (\Delta b_n^*)^2, \quad -\Delta \nu_n^* = r_3^n (\Delta b_n^*)^2.
$$

Then, based on the discussion at the beginning of Case 2.2 that the equation (45) holds, we have $r_1^n, r_2^n \to 0$, and $|r_3^n| \not\to \infty$. Moreover, the equation (48) indicates that

$$
\frac{\lambda_n |r_1^n|}{\lambda_n^* - \lambda_n} \not\to \infty, \; \frac{\lambda_n |r_2^n|}{\lambda_n^* - \lambda_n} \not\to \infty.
$$

We divide the numerators and the denominators of $F_{0,3}$ in equation (43) by $(\lambda_n^* - \lambda_n)(\Delta b_n^*)^3$ and $F_{0,4}$ in equation (44) by $(\lambda_n^* - \lambda_n)(\Delta b_n^*)^4$, respectively. Then, we see that the new numerators go to 0 as the new denominators do not go to infinity, i.e.

$$\frac{1}{6}\left[-1 + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot (r_1^n)^3\right] + \frac{1}{2}\left[r_3^n + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot r_1^n r_2^n\right] \to 0,$$

$$\frac{1}{24}\left[1 + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot (r_2^n)^4\right] + \frac{1}{4}\left[r_3^n + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot r_1^n r_2^n\right] + \frac{1}{8}\left[(r_3^n)^2 + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot (r_2^n)^2\right] \to 0.$$

From the discussion above, we would have

$$\frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot (r_1^n)^3 \to 0, \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot (r_2^n)^4 \to 0, \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot r_1^n r_2^n \to 0, \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot (r_2^n)^2 \to 0.$$

Suppose that $r_3^n \to r_3$, the above system will become

$$\begin{cases} -\frac{1}{6} + \frac{1}{2}r_3 = 0, \\ \frac{1}{24} - \frac{1}{4}r_3 + \frac{1}{8}(r_3)^2 = 0, \end{cases}$$

without a solution, which is a contradiction. Thus, case 2.2.1 does not hold.

Case 2.2.2: $M_n' = |\Delta \nu_n|$. Using the same argument as in Case 2.2.1, we would have that

$$\max\left\{\frac{\lambda_n|\Delta a_n - \Delta a_n^*|}{(\lambda_n^* - \lambda_n)|\Delta \nu_n|}, \frac{\lambda_n|\Delta b_n - \Delta b_n^*|^2}{(\lambda_n^* - \lambda_n)|\Delta \nu_n|}, \frac{\lambda_n|\Delta \nu_n - \Delta \nu_n^*|}{(\lambda_n^* - \lambda_n)|\Delta \nu_n|}\right\} \not\to \infty, \tag{49}$$

which is an important ingredient for the estimation that

$$\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \asymp (\lambda_n^* - \lambda_n)(\Delta \nu_n^*)^2.$$

Next, let us denote

$$\Delta \nu_n - \Delta \nu_n^* = s_1^n |\Delta \nu_n^*|, \quad \Delta b_n - \Delta b_n^* = s_2^n |\Delta \nu_n^*|^{1/2}, \quad \Delta b_n^* = s_3^n |\Delta \nu_n^*|^{1/2}.$$

From equation (45), we could conclude that $s_1^n \to 0, s_2^n \to 0$, and $|s_3^n| \not\to \infty$. From the equation (49), we would have

$$\frac{\lambda_n|s_1^n|}{\lambda_n^* - \lambda_n} \not\to \infty, \frac{\lambda_n|s_2^n|^2}{\lambda_n^* - \lambda_n} \not\to \infty,$$

Now we divide both the numerators and the denominators of $F_{0,3}$ and $F_{0,4}$ by $(\lambda_n^* - \lambda_n)|\Delta \nu_n^*|^{3/2}$ and $(\lambda_n^* - \lambda_n)|\Delta \nu_n^*|^2$, respectively. Then, we see that the new numerators go to 0 while the new denominators do not go to infinity, i.e.

We divide the numerators and the denominators of $F_{0,3}$ in equation (43) by $(\lambda_n^* - \lambda_n)|\Delta \nu_n^*|^{3/2}$ and in equation (44) by $(\lambda_n^* - \lambda_n)|\Delta \nu_n^*|^2$, respectively. Then, we see that the new numerators go to 0 as the new denominators do not go to infinity, i.e.

$$\frac{1}{6}\left[-(s_3^n)^3 + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot (s_2^n)^3\right] + \frac{1}{2}\left[s_3^n + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot s_1^n s_2^n\right] \to 0,$$

$$\frac{1}{24}\left[(s_3^n)^4 + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot (s_2^n)^4\right] + \frac{1}{4}\left[-(s_3^n)^2 + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot s_1^n (s_2^n)^2\right] + \frac{1}{8}\left[1 + \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot (s_1^n)^2\right] \to 0.$$

Based on the assumptions about $s_i^n, i \in [3]$, we could get that

$$\frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot s_1^n s_2^n \to 0, \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot (s_2^n)^4 \to 0, \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot s_1^n (s_2^n)^2 \to 0, \frac{\lambda_n}{\lambda_n^* - \lambda_n} \cdot (s_1^n)^2 \to 0.$$

Fanqi Yan[*,◇], **Huy Nguyen**[*,†], **Dung Le**[*,†], **Pedram Akabrian**[‡], **Nhat Ho**[†]

Suppose that $s_3^n \to s_3$, the above system will become

$$\begin{cases} -\frac{1}{6}s_3^3 + \frac{1}{2}s_3 = 0, \\ \frac{1}{24}s_3^4 - \frac{1}{4}s_3^2 + \frac{1}{8} = 0, \end{cases}$$

without a solution, which is a contradiction. Thus, case 2.2.2 does not hold.

Case 2.2.3: $M_n' = |\Delta a_n|$. Using the same argument as in Case 2.2.1, we would have that

$$\max \left\{ \frac{\lambda_n |\Delta a_n - \Delta a_n^*|}{(\lambda_n^* - \lambda_n)|\Delta a_n|}, \ \frac{\lambda_n |\Delta b_n - \Delta b_n^*|^2}{(\lambda_n^* - \lambda_n)|\Delta a_n|}, \ \frac{\lambda_n |\Delta \nu_n - \Delta \nu_n^*|}{(\lambda_n^* - \lambda_n)|\Delta a_n|} \right\} \not\to \infty, \tag{50}$$

which is an important ingredient for the estimation that

$$\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \asymp (\lambda_n^* - \lambda_n)(\Delta a_n^*)^2.$$

Meanwhile, the term $F_{2,0}$ gives us that

$$\begin{aligned} F_{2,0} &= \frac{(\lambda_n^* - \lambda_n)(\Delta a_n^*)^2 + \lambda_n(\Delta a_n^* - \Delta a_n)^2}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \asymp \frac{(\lambda_n^* - \lambda_n)(\Delta a_n^*)^2 + \lambda_n(\Delta a_n^* - \Delta a_n)^2}{(\lambda_n^* - \lambda_n)(\Delta a_n^*)^2} \\ &= 1 + \frac{\lambda_n(\Delta a_n^* - \Delta a_n)^2}{(\lambda_n^* - \lambda_n)(\Delta a_n^*)^2} \to 0 \end{aligned}$$

which is a contradiction, as $0 \lesssim \frac{\lambda_n(\Delta a_n^* - \Delta a_n)^2}{(\lambda_n^* - \lambda_n)(\Delta a_n^*)^2}$. In overall, the case 2.2 cannot hold.

Case 2.3: $\lambda_n^*/\lambda_n \to \infty$. Recall at the beginning of case 2 that

$$M_n' = \max\{|\Delta a_n|, |\Delta b_n|^2, |\Delta \nu_n|, |\Delta a_n^*|, |\Delta b_n^*|^2, |\Delta \nu_n^*|\}.$$

Obviously we have $\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \lesssim \lambda_n^*(M_n')^2$. Dividing both the numerator and the denominator of A_n (defined in equation (40)) and B_n (defined in equation (41)) by $\lambda_n^*(M_n')^{1/2}$ and $\lambda_n^* M_n'$, respectively, and considering that the new denominators tend to 0, we can conclude that the new numerators also tend to 0. In other words, we obtain the following results:

$$\frac{\lambda_n \Delta b_n - \lambda_n^* \Delta b_n^*}{\lambda_n^*(M_n')^{1/2}} = \frac{\lambda_n \Delta b_n}{\lambda_n^*(M_n')^{1/2}} - \frac{\Delta b_n^*}{(M_n')^{1/2}} \to 0$$

$$\frac{\lambda_n \Delta a_n - \lambda_n^* \Delta a_n^*}{\lambda_n^* M_n'} = \frac{\lambda_n \Delta a_n}{\lambda_n^* M_n'} - \frac{\Delta a_n^*}{M_n'} \to 0$$

Now we have $\lambda_n/\lambda_n^* \to 0$, so we can conclude that $\Delta b_n^*/(M_n')^{1/2} \to 0$ and $\Delta a_n^*/M_n' \to 0$.

Dividing both the numerator and the denominator of C_n by $\lambda_n^* M_n'$, similarly we obtain the following results:

$$\frac{(\lambda_n^* - \lambda_n)\Delta b_n \Delta b_n^* + \lambda_n \Delta \nu_n - \lambda_n^* \Delta \nu_n^*}{\lambda_n^* M_n'} = \frac{(\lambda_n^* - \lambda_n)}{\lambda_n^*} \frac{\Delta b_n}{(M_n')^{1/2}} \frac{\Delta b_n^*}{(M_n')^{1/2}} + \frac{\lambda_n}{\lambda_n^*} \frac{\Delta \nu_n}{M_n'} - \frac{\Delta \nu_n^*}{M_n'} \to 0.$$

Reminding that $\frac{\lambda_n^*}{\lambda_n} \to \infty$ as a hypothesis of this case, $|\Delta b_n| \leq (M_n')^{1/2}$, $|\Delta \nu_n| \leq M_n'$, and more importantly, $|\Delta b_n^*|/(M_n')^{1/2} \to 0$, we achieve $\Delta \nu_n^*/M_n' \to 0$. To sum up the estimation above:

$$\Delta a_n^*/M_n' \to 0, \ \Delta b_n^*/(M_n')^{1/2} \to 0, \ \Delta \nu_n^*/M_n' \to 0. \tag{51}$$

Thus, WLOG, we can suppose that $M_n' \in \{|\Delta a_n|, |\Delta b_n|^2, |\Delta \nu_n|\}$. Moreover, by considering three situations $M_n' = |\Delta a_n|$, $M_n' = |\Delta b_n|^2$, $M_n' = |\Delta \nu_n|$ separately, combining with the estimation in equation (51), we can easily achieve that

$$|\Delta a_n - \Delta a_n^*| + |\Delta b_n - \Delta b_n^*|^2 + |\Delta \nu_n - \Delta \nu_n^*| \asymp M_n' \tag{52}$$

Combining this estimation, the obvious fact that $\lambda_n(|\Delta a_n| + |\Delta b_n|^2 + |\Delta \nu_n|) \asymp \lambda_n M_n'$, and the assumption about R_n in equation (39), we have

$$\lambda_n^* \max\{|\Delta a_n^*|, |\Delta b_n^*|^2, |\Delta \nu_n^*|\}/(\lambda_n M_n') \to \infty. \tag{53}$$

Let $N_n' = \max\{|\Delta a_n^*|, |\Delta b_n^*|^2, |\Delta \nu_n^*|\}$. It is evident that $M_n' \to 0, N_n' \to 0$. Based on this assumption, by estimating each term in $\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))$ separately, combining with equation (52) and equation (53), we have

$$\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \asymp (\lambda_n^* - \lambda_n)M_n'N_n' + \lambda_n^* M_n'N_n'$$
$$\asymp \lambda_n^* M_n'N_n'.$$

Let discuss three distinct cases based on N_n'.

Case 2.3.1: $N_n' = |\Delta b_n^*|^2$ for all n, then $\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n})) \asymp \lambda_n^* M_n' |\Delta b_n^*|^2$. By dividing both the numerator and denominator of B_n (in equation (41)) by $\lambda_n^* \Delta b_n^*$, since the new denominator of B_n goes to 0, its new numerator must go to 0, i.e.,

$$\frac{\lambda_n^* \Delta b_n^* - \lambda_n \Delta b_n}{\lambda_n^* \Delta b_n^*} \to 0 \Leftrightarrow (\lambda_n^* \Delta b_n^*)/(\lambda_n \Delta b_n) \to 1 \Leftrightarrow (\lambda_n^* |\Delta b_n^*|^2)/(\lambda_n \Delta b_n \Delta b_n^*) \to 1.$$

Moreover, $\lambda_n |\Delta b_n \Delta b_n^*|/\lambda_n M_n' \lesssim 1$, thus

$$\lambda_n^* |\Delta b_n^*|^2/\lambda_n M_n' \lesssim 1,$$

which is a contradiction to equation (53). Noting that from equation (51), $(\lambda_n^* |\Delta b_n^*|^2)/(\lambda_n |\Delta b_n|^2) \to 0$. This is a contradiction with (53).

Case 2.3.2: $N_n' = |\Delta a_n^*|$ for all n, by dividing both the numerator and denominator of A_n by $\lambda_n^* \Delta a_n^*$, since the new denominator of A_1 goes to 0, its new numerator must go to 0, i.e.,

$$\frac{\lambda_n^* \Delta a_n^* - \lambda_n \Delta a_n}{\lambda_n^* \Delta a_n^*} \to 0 \Leftrightarrow (\lambda_n^* \Delta a_n^*)/(\lambda_n \Delta a_n) \to 1.$$

Thus, $(\lambda_n^* \Delta a_n^*)/(\lambda_n M_n') \lesssim 1$, which is also a contradiction with equation (53).

Case 2.3.3: $N_n' = |\Delta \nu_n^*|$. Firstly, we mention a small remark which would be helpful during our proof:

Claim: Given three sequences of real numbers $(\mathfrak{a}_n)_n, (\mathfrak{b}_n)_n, (\mathfrak{c}_n)_n$ satisfying $(\mathfrak{a}_n - \mathfrak{b}_n)/\mathfrak{c}_n \to 0$. If $\mathfrak{a}_n/\mathfrak{b}_n \not\to 1$ (i.e. there does not exist a subsequence $\mathfrak{a}_{n_i}, \mathfrak{b}_{n_i}$ such that $\mathfrak{a}_{n_i}/\mathfrak{b}_{n_i} \to 1$), then

$$\frac{\mathfrak{a}_n}{\mathfrak{c}_n} \to 0, \frac{\mathfrak{b}_n}{\mathfrak{c}_n} \to 0.$$

In fact, by writing $(\mathfrak{a}_n - \mathfrak{b}_n)/\mathfrak{c}_n$ as

$$(\mathfrak{a}_n - \mathfrak{b}_n)/\mathfrak{c}_n = \frac{\mathfrak{b}_n}{\mathfrak{c}_n}\left(\frac{\mathfrak{a}_n}{\mathfrak{b}_n} - 1\right),$$

then $\mathfrak{a}_n/\mathfrak{b}_n - 1 \not\to 0$, thus $\mathfrak{b}_n/\mathfrak{c}_n \to 0$, thus $\mathfrak{a}_n/\mathfrak{c}_n = \mathfrak{b}_n/\mathfrak{c}_n + (\mathfrak{a}_n - \mathfrak{b}_n)/\mathfrak{c}_n \to 0$

Return to our problem. Denote the term that are useful in the next step of our proof:

$$TB_n = \frac{(\lambda_n^* - \lambda_n)\Delta b_n^*}{\lambda_n(\Delta b_n - \Delta b_n^*)},$$
$$TBN_n = \frac{(\lambda_n^* - \lambda_n)(\Delta b_n^*)^2 + \lambda_n(\Delta b_n^* - \Delta b_n)^2}{(\lambda_n^* - \lambda_n)(\Delta \nu_n^*) + \lambda_n(\Delta \nu_n^* - \Delta \nu_n)},$$
$$G_n = \frac{\lambda_n^* \Delta \nu_n^* - \lambda_n \Delta \nu_n}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))}.$$

We have

$$\frac{(\lambda_n^* - \lambda_n)\Delta b_n^*}{\lambda_n(\Delta b_n^* - \Delta b_n)} \to 1 \Leftrightarrow \frac{\lambda_n^* \Delta b_n^*}{\lambda_n(\Delta b_n - \Delta b_n^*)} \to 1 \Leftrightarrow \frac{\lambda_n(\Delta b_n - \Delta b_n^*)}{\lambda_n^* \Delta b_n^*} \to 1 \tag{54}$$
$$\Leftrightarrow \frac{\lambda_n \Delta b_n}{\lambda_n^* \Delta b_n^*} - \frac{\lambda_n}{\lambda_n^*} \to 1 \Leftrightarrow \frac{\lambda_n \Delta b_n}{\lambda_n^* \Delta b_n^*} \to 1.$$

Case 2.3.3.1: $TB_n \nrightarrow 1$. Applying the claim to $\mathfrak{a}_n = (\lambda_n^* - \lambda_n)\Delta b_n^*$, $\mathfrak{b}_n = \lambda_n(\Delta b_n - \Delta b_n^*)$, $\mathfrak{c}_n = \overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))$ in the formulae of B_n (in equation (41)), then we achieve that

$$\frac{(\lambda_n^* - \lambda_n)\Delta b_n^*}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0, \frac{\lambda_n(\Delta b_n - \Delta b_n^*)}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0.$$

Thus, from the hypothesis that $\Delta b_n^* \to 0$, $\Delta b_n \to 0$

$$\frac{(\lambda_n^* - \lambda_n)(\Delta b_n^*)^2}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0, \frac{\lambda_n(\Delta b_n - \Delta b_n^*)^2}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0.$$

By substituting these estimation to the formulae of C_n (in equation (42)), we achieve that

$$G_n = \frac{\lambda_n^* \Delta \nu_n^* - \lambda_n \Delta \nu_n}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0.$$

Using the same argument as in Case 2.3.2, we get the contradiction.

Case 2.3.3.2: $TBN_n \nrightarrow 1$. Then, applying the claim to $\mathfrak{a}_n = (\lambda_n^* - \lambda_n)(\Delta b_n^*)^2 + \lambda_n(\Delta b_n^* - \Delta b_n)^2$, $\mathfrak{b}_n = (\lambda_n^* - \lambda_n)(\Delta \nu_n^*) + \lambda_n(\Delta \nu_n^* - \Delta \nu_n)$, $\mathfrak{c}_n = \overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))$ in the formulae of C_n (in equation (42)), then we achieve that

$$G_n = \frac{\lambda_n^* \Delta \nu_n^* - \lambda_n \Delta \nu_n}{\overline{D_4}((\lambda_n, G_n), (\lambda_n^*, G_{*,n}))} \to 0.$$

Using the same argument as in Case 2.3.2, we get the contradiction.

Case 2.3.3.3: Both $TB_n, TBN_n \to 1$. Then, as discussed in equation (54), $TB_n \to 1$ implies that

$$(\lambda_n^* \Delta b_n^*)/(\lambda_n \Delta b_n) \to 1.$$

As $\lambda_n^*/\lambda_n \to \infty$, the previous equation implies that

$$\Delta b_n/\Delta b_n^* \to \infty \text{ or } \Delta b_n^*/\Delta b_n \to 0. \tag{55}$$

Next, we would like to find the equivalent form of the nominator and denominator of TBN_n. For the nominator, we write

$$\frac{(\lambda_n^* - \lambda_n)(\Delta b_n^*)^2 + \lambda_n(\Delta b_n^* - \Delta b_n)^2}{\lambda_n(\Delta b_n)^2} = \frac{\lambda_n^* - \lambda_n}{\lambda_n^*}\left(\frac{\Delta b_n^*}{\Delta b_n}\right)^2 + \left(1 - \frac{\Delta b_n^*}{\Delta b_n}\right)^2.$$

From the hypothesis that $\lambda_n^*/\lambda_n \to \infty$ as well as equation (55), we have

$$\frac{(\lambda_n^* - \lambda_n)(\Delta b_n^*)^2 + \lambda_n(\Delta b_n^* - \Delta b_n)^2}{\lambda_n(\Delta b_n)^2} \to 1. \tag{56}$$

For the denominator of TBN_n, according to equation (53), noting that $|\Delta \nu_n^*| = \max\{|\Delta a_n^*|, |\Delta b_n^*|^2, |\Delta \nu_n^*|\}$ (hypothesis of Case 2.3.3), $|\Delta \nu_n| \le M_n'$, then $(\lambda_n \Delta \nu_n)/(\lambda_n^* \Delta \nu_n^*) \to 0$. Thus,

$$\frac{\lambda_n^* \Delta \nu_n^* - \lambda_n \Delta \nu_n}{\lambda_n^* \Delta \nu_n^*} = 1 - \frac{\lambda_n \Delta \nu_n}{\lambda_n^* \Delta \nu_n^*} \to 1. \tag{57}$$

By combining equation (56) and equation (57), and the formulae of TBN_n, we achieve that

$$\frac{\lambda_n^* \Delta \nu_n^*}{\lambda_n(\Delta b_n)^2} \to 1 \text{ or } \frac{\lambda_n^* |\Delta \nu_n^*|}{\lambda_n |\Delta b_n|^2} \to 1.$$

As $|\Delta \nu_n^*| = \max\{|\Delta a_n^*|, |\Delta b_n^*|^2, |\Delta \nu_n^*|, |\Delta b_n|^2 \le M_n'$, this equation implies that $\lambda_n^* \max\{|\Delta a_n^*|, |\Delta b_n^*|^2, |\Delta \nu_n^*|\}/(\lambda_n M_n') \lesssim 1$, which is a contradiction to equation (53).

We conclude the proof of our theorem. $\qquad\square$

B   PROOFS FOR THE MINIMAX LOWER BOUNDS

In this appendix, we present the proofs for the minimax lower bounds introduced in Section 3 and Section 4.

B.1   Proof of Theorem 3

In order to prove the result, we will at first define two distances:

$$d_1((\lambda_1, G_1), (\lambda_2, G_2)) = \lambda_1 \|(a_1, b_1, \nu_1) - (a_2, b_2, \nu_2)\|,$$
$$d_2((\lambda_1, G_1), (\lambda_2, G_2)) = |\lambda_1 - \lambda_2|^2,$$

for any $(\lambda_1, G_1) = (\lambda_1, a_1, b_1, \nu_1) \in \Xi$ and $(\lambda_2, G_2) = (\lambda_2, a_2, b_2, \nu_2) \in \Xi$. Obviously $d_2((\lambda_1, G_1), (\lambda_2, G_2))$ is a proper distance. The structure for $d_1((\lambda_1, G_1), (\lambda_2, G_2))$ tells us that it is not symmetric. Only when $\lambda_1 = \lambda_2 = \lambda$, $d_1((\lambda, G_1), (\lambda, G_2))$ is symmetric. Also $d_1((\lambda_1, G_1), (\lambda_2, G_2))$ still satisfies a weak triangle inequality:

$$d_1((\lambda_1, G_1), (\lambda_2, G_2)) + d_1((\lambda_2, G_2), (\lambda_3, G_3)) \geq \min\{d_1((\lambda_1, G_1), (\lambda_2, G_2)), d_1((\lambda_2, G_2), (\lambda_3, G_3))\}.$$

Therefore, we will apply the modified Le Cam method for nonsymmetric loss, as outlined in Lemma C.1 of Gadat et al. (2020), to handle this distance. For f satisfies all assumptions in Theorem 3, based on the Taylor expansion, we have the following results:

Lemma 4. *Given f in Theorem 3, we achieve for any $r < 1$ that*

(i) $\lim\limits_{\epsilon \to 0} \inf\limits_{(\lambda, G_1) = (\lambda, a_1, b_1, \nu_1), (\lambda, G_2) = (\lambda, a_2, b_2, \nu_2)} \left\{ \dfrac{h(p_{\lambda, G_1}, p_{\lambda, G_2})}{d_1^r((\lambda, G_1), (\lambda, G_2))} : d_1((\lambda, G_1), (\lambda, G_2)) \leq \epsilon \right\} = 0,$

(ii) $\lim\limits_{\epsilon \to 0} \inf\limits_{(\lambda_1, G) = (\lambda_1, a, b, \nu), (\lambda_2, G) = (\lambda_2, a, b, \nu)} \left\{ \dfrac{h(p_{\lambda_1, G}, p_{\lambda_2, G})}{d_2^r((\lambda_1, G), (\lambda_2, G))} : d_2((\lambda_1, G), (\lambda_2, G)) \leq \epsilon \right\} = 0.$

We will prove this lemma later.

Proof of Theorem 3. Denote $(\lambda^*, G_*) = (\lambda^*, a^*, b^*, \nu^*)$ and assume $r < 1$. Given Lemma 4 part (i) , for any sufficiently small $\epsilon > 0$, there exists $(\lambda^*, G'_*) = (\lambda^*, a_1^*, b_1^*, \nu_1^*)$ such that $d_1((\lambda^*, G_*), (\lambda^*, G'_*)) = d_1((\lambda^*, G'_*), (\lambda^*, G_*)) = \epsilon$, there exists a constant C_0, s.t.

$$h(p_{\lambda^*, G_*}, p_{\lambda^*, G'_*}) \leq C_0 \epsilon^r. \tag{58}$$

Now we will denote p_{λ^*, G_*}^n as the density of the n-i.i.d. sample $(X_1, Y_1), \cdots, (X_n, Y_n)$. Lemma C.1 in Gadat et al. (2020) tells us that

$$\inf_{(\overline{\lambda}_n, \overline{G}_n) \in \Xi} \sup_{(\lambda, G) \in \Xi} \mathbb{E}_{p_{\lambda, G}} \left(\lambda^2 \|(\overline{a}_n, \overline{b}_n, \overline{\nu}_n) - (a, b, \nu)\|^2 \right) \geq \frac{\epsilon^2}{2} \left(1 - V(p_{\lambda^*, G_*}^n, p_{\lambda^*, G'_*}^n) \right) \geq \frac{\epsilon^2}{2} \sqrt{1 - (1 - C_0^2 \epsilon^{2r})^n}.$$

Last inequality is from the definition of the Total Variation distance and Hellinger distance and equation (58). Let $\epsilon^{2r} = \dfrac{1}{C_0^2 n}$, then for any $r < 1$ we have

$$\inf_{(\overline{\lambda}_n, \overline{G}_n) \in \Xi} \sup_{(\lambda, G) \in \Xi} \mathbb{E}_{p_{\lambda, G}} \left(\lambda^2 \|(\overline{a}_n, \overline{b}_n, \overline{\nu}_n) - (a, b, \nu)\|^2 \right) \geq c_1 n^{-1/r},$$

where c_1 is some positive constant. Following a similar reasoning and using Lemma 4 part (ii) , we will obtain

$$\inf_{(\overline{\lambda}_n, \overline{G}_n) \in \Xi} \sup_{(\lambda, G) \in \Xi} \mathbb{E}_{p_{\lambda, G}} \left(|\overline{\lambda}_n - \lambda|^2 \right) \geq c_2 n^{-1/r},$$

for some positive constant c_2. Consequently, we establish all of the results for Theorem 3. $\qquad \square$

Fanqi Yan$^{\star,\diamond}$, **Huy Nguyen***,†, **Dung Le***,†, **Pedram Akabrian**‡, **Nhat Ho**†

Proof of Lemma 4. (i) Now consider two sequences

$$\begin{cases} (\lambda_n, G_{1,n}) = (\lambda_n, a_{1,n}, b_{1,n}, \nu_{1,n}) \\ (\lambda_n, G_{2,n}) = (\lambda_n, a_{2,n}, b_{2,n}, \nu_{2,n}) \end{cases}$$

with the same proportion λ_n. Just refer to the contaminated MoE model definition (1) we will have

$$h^2\left(p_{\lambda_n, G_{1,n}}, p_{\lambda_n, G_{2,n}}\right) \le \lambda_n \int \frac{\left[f\left(Y|\sigma((a_{1,n})^\top X + b_{1,n}), \nu_{1,n}\right) - f\left(Y|\sigma((a_{2,n})^\top X + b_{2,n}), \nu_{2,n}\right)\right]^2}{f\left(Y|\sigma((a_{2,n})^\top X + b_{2,n}), \nu_{2,n}\right)} d(X,Y)$$

due to $\sqrt{p_{\lambda_n, G_{1,n}}(Y|X)} + \sqrt{p_{\lambda_n, G_{2,n}}(Y|X)} > \sqrt{\lambda_n f\left(Y|\sigma((a_{2,n})^\top X + b_{2,n}), \nu_{2,n}\right)}$ and $p_{\lambda_n, G_{1,n}}(Y|X) - p_{\lambda_n, G_{2,n}}(Y|X) = \lambda_n \left(f\left(Y|\sigma((a_{1,n})^\top X + b_{1,n}), \nu_{1,n}\right) - f\left(Y|\sigma((a_{2,n})^\top X + b_{2,n}), \nu_{2,n}\right)\right)$. Consider the Taylor expansion up to the first order with remainder in integral form, we have

$$f\left(Y|\sigma((a_{1,n})^\top X + b_{1,n}), \nu_{1,n}\right) - f\left(Y|\sigma((a_{2,n})^\top X + b_{2,n}), \nu_{2,n}\right)$$
$$= \sum_{|\alpha|=1} \frac{(a_{1,n} - a_{2,n})^{\alpha_1}(b_{1,n} - b_{2,n})^{\alpha_2}(\nu_{1,n} - \nu_{2,n})^{\alpha_3}}{\alpha_1! \alpha_2! \alpha_3!} \frac{\partial f^{|\alpha|}}{\partial a^{\alpha_1} \partial b^{\alpha_2} \partial \nu^{\alpha_3}}(Y|\sigma((a_{2,n})^\top X + b_{2,n}), \nu_{2,n})$$
$$+ \sum_{|\alpha|=1} \frac{(a_{1,n} - a_{2,n})^{\alpha_1}(b_{1,n} - b_{2,n})^{\alpha_2}(\nu_{1,n} - \nu_{2,n})^{\alpha_3}}{\alpha_1! \alpha_2! \alpha_3!} \times$$
$$\int_0^1 \frac{\partial f^{|\alpha|}}{\partial a^{\alpha_1} \partial b^{\alpha_2} \partial \nu^{\alpha_3}}\left(Y|\sigma\left((a_{2,n} + t(a_{1,n} - a_{2,n}))^\top X + b_{2,n} + t(b_{1,n} - b_{2,n})\right), \nu_{2,n} + t(\nu_{1,n} - \nu_{2,n})\right) dt.$$

So it follows that

$$\frac{h^2(p_{\lambda_n, G_{1,n}}, p_{\lambda_n, G_{2,n}})}{d_1^{2r}((\lambda_n, G_{1,n}), (\lambda_n, G_{2,n}))} \to 0.$$

when we have $d_1^{2-2r}((\lambda_n, G_{1,n}), (\lambda_n, G_{2,n}))/\lambda_n = \lambda_n^{1-2r}\|(a_{1,n}, b_{1,n}, \nu_{1,n}) - (a_{2,n}, b_{2,n}, \nu_{2,n})\|^{2-2r} \to 0$ and $d_1((\lambda_n, G_{1,n}), (\lambda_n, G_{2,n}))/\lambda_n = \|(a_{1,n}, b_{1,n}, \nu_{1,n}) - (a_{2,n}, b_{2,n}, \nu_{2,n})\| \to 0$. Naturally we got part (i).

(ii) The proof for part (ii) is basically the same as in part (i). Similarly, consider two sequences

$$\begin{cases} (\lambda_{1,n}, G_n) = (\lambda_{1,n}, a_n, b_n, \nu_n), \\ (\lambda_{2,n}, G_n) = (\lambda_{2,n}, a_n, b_n, \nu_n), \end{cases}$$

with different proportion and the same $G_n = (a_n, b_n \nu_n) \in \Theta$. we will have

$$\frac{h^2(p_{\lambda_{1,n}, G_n}, p_{\lambda_{2,n}, G_n})}{d_2^{2r}((\lambda_{1,n}, G_n), (\lambda_{2,n}, G_n))} \le \frac{(\lambda_{1,n} - \lambda_{2,n})^{2-2r}}{(1 - \lambda_{1,n}) \wedge \lambda_{1,n}} \int \frac{(f_0(Y|\sigma((a_0)^\top X + b_0), \nu_0) - f(Y|\sigma((a_n)^\top X + b_n), \nu_n))^2}{f_0(Y|\sigma((a_0)^\top X + b_0), \nu_0) + f(Y|\sigma((a_n)^\top X + b_n), \nu_n)} d(X,Y)$$
$$\le \frac{2(\lambda_{1,n} - \lambda_{2,n})^{2-2r}}{(1 - \lambda_{1,n}) \wedge \lambda_{1,n}}$$

Now consider when $\frac{(\lambda_{1,n} - \lambda_{2,n})^{2-2r}}{(1 - \lambda_{1,n}) \wedge \lambda_{1,n}} \to 0$, we will reach the conclusion of part (ii). □

B.2 Proof of Theorem 5

Proof of Theorem 5. The proof follows similar steps to the argument in the last section. To be precise, we define

$$\begin{cases} d_3((\lambda_1, G_1), (\lambda_2, G_2)) & = \lambda_1 \|((a_1, b_1, \nu_1) - (a_2, b_2, \nu_2))\|^4, \\ d_4((\lambda_1, G_1), (\lambda_2, G_2)) & = |\lambda_1 - \lambda_2| \|((\Delta a_1, \Delta b_1, \Delta \nu_1) - (\Delta a_2, \Delta b_2, \Delta \nu_2))\|^4. \end{cases}$$

It is still straighforward that d_4 still satisfies the weak triangle inequality, meanwhile d_3 no longer satisfies this inequality. However, a close look at d_3 shows that

$$d_3(\lambda_1, G_1), (\lambda_3, G_3)) + d_3(\lambda_2, G_2), (\lambda_3, G_3)) \ge \frac{\min\{d_3((\lambda_1, G_1), (\lambda_2, G_2)), d_3((\lambda_2, G_2), (\lambda_1, G_1))\}}{8}.$$

Under this remark, we can modify the Le Cam method in Lemma C.1, (Gadat et al., 2020) to achieve the following estimation

$$\inf_{(\overline{\lambda}_n, \overline{G}_n) \in \Xi} \sup_{(\lambda, G) \in \Xi_1(l_n)} \mathbb{E}_{p_{\lambda, G}}(d_3^2((\lambda, G), (\overline{\lambda}_n, \overline{G}_n))) \geq \frac{\epsilon^2}{128} \left\{ 1 - V(p_{\lambda_1, G_1}^n, p_{\lambda_2, G_2}^n) \right\}$$

for $(\lambda_1, G_1), (\lambda_2, G_2) \in \Xi_1(l_n)$ satisfying $d_3((\lambda_1, G_1), (\lambda_2, G_2)) \wedge d_3((\lambda_2, G_2), (\lambda_1, G_1)) \geq \epsilon/4$. Following the same schema as in Lemma 4, we can demonstrate two subsequent results for any $r > 1$:

(i) Two sequences can be found

$$\begin{cases} (\lambda_n, G_{1,n}) = (\lambda_n, a_{1,n}, b_{1,n}, \nu_{1,n}) \in \Xi_1(l_n), \\ (\lambda_n, G_{2,n}) = (\lambda_n, a_{2,n}, b_{2,n}, \nu_{2,n}) \in \Xi_1(l_n), \end{cases}$$

such that $d_3((\lambda_n, G_{1,n}), (\lambda_n, G_{2,n})) \to 0$ and $h(p_{\lambda_n, G_{1,n}}, p_{\lambda_n, G_{2,n}})/d_3^r((\lambda_n, G_{1,n}), (\lambda_n, G_{2,n})) \to 0$ as $n \to \infty$.

(ii) Two sequences can also be found

$$\begin{cases} (\lambda_{1,n}, G_n) = (\lambda_{1,n}, a_n, b_n, \nu_n) \in \Xi_1(l_n), \\ (\lambda_{2,n}, G_n) = (\lambda_{2,n}, a_n, b_n, \nu_n) \in \Xi_1(l_n), \end{cases}$$

such that $d_4(p_{\lambda_{1,n}, G_n}, p_{\lambda_{2,n}, G_n}) \to 0$ and $h(p_{\lambda_{1,n}, G_n}, p_{\lambda_{2,n}, G_n})/d_4^r((\lambda_{1,n}, G_n), (\lambda_{2,n}, G_n)) \to 0$ as $n \to \infty$.

The justification for the above results can be omitted, as it follows a similar approach to that of Lemma 4. This brings us to the conclusion of the theorem. □

B.3 Proof of Theorem 10

Proof of Theorem 10. The proof follows similar steps to the arguments in the previous two sections. Concretely, define for $(\lambda_1, G_1) = (\lambda_1, a_1, b_1, \nu_1)$, $(\lambda_2, G_2) = (\lambda_2, a_2, b_2, \nu_2)$:

$$\begin{cases} d_5((\lambda_1, G_1), (\lambda_2, G_2)) = \lambda_1 \|\Delta a_1, \Delta b_1, \Delta \nu_1\| \|(a_1, b_1, \nu_1) - (a_2, b_2, \nu_2)\|, \\ d_6((\lambda_1, G_1), (\lambda_2, G_2)) = |\lambda_1 - \lambda_2| \|\Delta a_1, \Delta b_1, \Delta \nu_1\|^2. \end{cases}$$

It is straightforward that both d_5 and d_6 satisfy the weak triangle inequality. Following the same schema as in Lemma 4, we can demonstrate two subsequent results for any $r > 1$:

(i) Two sequences can be found

$$\begin{cases} (\lambda_n, G_{1,n}) = (\lambda_n, a_{1,n}, b_{1,n}, \nu_{1,n}) \in \Xi_2(l_n), \\ (\lambda_n, G_{2,n}) = (\lambda_n, a_{2,n}, b_{2,n}, \nu_{2,n}) \in \Xi_2(l_n), \end{cases}$$

such that $d_5((\lambda_n, G_{1,n}), (\lambda_n, G_{2,n})) \to 0$ and $h(p_{\lambda_n, G_{1,n}}, p_{\lambda_n, G_{2,n}})/d_5^r((\lambda_n, G_{1,n}), (\lambda_n, G_{2,n}) \to 0$ as $n \to \infty$.

(ii) Two sequences can also be found

$$\begin{cases} (\lambda_{1,n}, G_n) = (\lambda_{1,n}, a_n, b_n, \nu_n) \in \Xi_2(l_n), \\ (\lambda_{2,n}, G_n) = (\lambda_{2,n}, a_n, b_n, \nu_n) \in \Xi_2(l_n), \end{cases}$$

such that $d_6(p_{\lambda_{1,n}, G_n}, p_{\lambda_{2,n}, G_{2,n}}) \to 0$ and $h(p_{\lambda_{1,n}, G_n}, p_{\lambda_{2,n}, G_{2,n}})/d_6^r((\lambda_{1,n}, G_n), (\lambda_{2,n}, G_{2,n})) \to 0$ as $n \to \infty$.

We can omit the justification for the above results as it can follow a similar approach as in Lemma 4. This leads to the conclusion of the theorem. □

Fanqi Yan[*,⋄], Huy Nguyen[*,†], Dung Le[*,†], Pedram Akabrian[‡], Nhat Ho[†]

C PROOF FOR DENSITY ESTIMATION RATE

C.1 General theory for the Proof of Theorem 1

At first we restate Theorem 1:

Theorem 11. *Assume that the function f_0 is bounded with tail $\mathbb{E}_X\left(-\log f_0(Y|\varphi(a_0^\top X + b_0), \nu_0)\right) \gtrsim Y^q$ for almost surely $Y \in \mathcal{Y}$ for some $q > 0$. and f is the density function of an univariate Gaussian distribution. Then, there exists a constant $C > 0$ depending only on Ξ such that for all $n \geq 1$,*

$$\sup_{(\lambda^*, G_*) \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} h(p_{\widehat{\lambda}_n, \widehat{G}_n}, p_{\lambda^*, G_*}) \leq C\sqrt{\log n / n}.$$

First we will introduce some necessary notations used throughout this appendix. $(\mathcal{P}, d)$ is a metric space, where d is the metric in $\mathcal{P}$. An ϵ-net for $(\mathcal{P}, d)$ is a collection of balls with radius ϵ, and the union of the balls contains $\mathcal{P}$. $N(\epsilon, \mathcal{P}, d)$ is the covering number for the ϵ-net (minimal cardinality). $H(\epsilon, \mathcal{P}, d) = \log N(\epsilon, \mathcal{P}, d)$ is the entropy number. $N_B(\epsilon, \mathcal{P}, d)$ is the bracketing number which means the minimal number n such that there exists n pairs $(\underline{f_i}, \overline{f_i})_{i=1}^n$ such that $\underline{f_i} < \overline{f_i}, d(\underline{f_i}, \overline{f_i}) < \epsilon$, and $\mathcal{P}$ was covered by their unions. $H_B(\epsilon, \mathcal{P}, d) = \log N_B(\epsilon, \mathcal{P}, d)$ is called the bracketing entropy number. If $\mathcal{P}$ is a density family, then d is the distance associated with $L^2(m)$, where m m is the Lebesgue measure.

In particular, we denote $\mathcal{P}(\Xi) = \{p_{\lambda, G} : (\lambda, G) \in \Xi\}$, and $\bar{p}_{\lambda, G} = \dfrac{p_{\lambda^*, G_*} + p_{\lambda, G}}{2}$ then we define

$$\overline{\mathcal{P}}(\Xi) = \{\bar{p}_{\lambda, G} : (\lambda, G), \in \Xi\},$$
$$\overline{\mathcal{P}}^{1/2}(\Xi) = \{\bar{p}_{\lambda, G}^{1/2} : \bar{p}_{\lambda, G} \in \overline{\mathcal{P}}(\Xi)\}.$$

The convergence rate can be inferred from the complexity of the set:

$$\overline{\mathcal{P}}^{1/2}(\Xi, \epsilon) := \{\bar{p}_{\lambda, G}^{1/2} \in \overline{\mathcal{P}}^{1/2}(\Xi) : h(\bar{p}_{\lambda, G}, p_{\lambda^*, G_*}) \leq \epsilon\}.$$

We assess the complexity of this class using the bracketing entropy integral in van de Geer (2000) :

$$\mathcal{J}_B(\epsilon, \overline{\mathcal{P}}^{1/2}(\Xi, \epsilon), m) = \int_{\epsilon^2/2^{13}}^{\epsilon} H_B^{1/2}(u, \overline{\mathcal{P}}^{1/2}(\Xi, \epsilon), m) du \vee \epsilon$$

where $u \vee \epsilon = \max\{u, \epsilon\}$ and $H_B(\epsilon, \mathcal{P}, m)$ represents the ϵ-bracketing entropy of a set $\mathcal{P}$ under the Lebesgue measure m, we may omit m for the simplicity.

Lemma 5. *Assume the following assumption hold: Given a universal constant $J > 0$, there exists $N > 0$, possibly depending on Ξ, such that for all $n \geq N$ and all $\epsilon > (\log(n)/n)^{1/2}$, we have*

$$\mathcal{J}_B(\epsilon, \overline{P}^{1/2}(\Xi, \epsilon)) \leq J\sqrt{n}\epsilon^2. \tag{59}$$

Then, there exists a constant $C > 0$ depending only on Ξ such that for all $n \geq 1$,

$$\sup_{(\lambda^*, G_*) \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} h(p_{\widehat{\lambda}_n, \widehat{G}_n}, p_{\lambda^*, G_*}) \leq C\sqrt{\log n / n}.$$

This lemma told us we just need to prove the model distribution satisfy the above assumption in equation (59) to get the conclusion we want. But usually the assumption in equation (59) is too complex to prove, instead we could prove the following lemma:

Lemma 6. *If the distribution satisfies*

$$H_B(\epsilon, \mathcal{P}(\Xi), h) \lesssim \log(1/\epsilon), \tag{60}$$

it will meet the assumption in equation (59).

Though we have simplified the assumption in equation (59) to equation (60). It is still hard to prove. But recall the contaminated model

$$p_{\lambda,G}(Y|X) = (1-\lambda)f_0(Y|\varphi(a_0^\top X + b_0), \nu_0) + \lambda f(Y|\sigma\big((a)^\top X + b\big), \nu),$$

and we have the assumption that f_0 is bounded and light tail, and f is univariate Gaussian density. We have the following lemma that could help us get the conclusion:

Lemma 7. *Let Θ be a bounded subsets of $\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^+$, f is a univariate Gaussian density and f_0 is bounded with tail $\mathbb{E}_X\big(-\log f_0(Y|\varphi(a_0^\top X + b_0), \nu_0)\big) \gtrsim Y^q$ for almost surely $Y \in \mathcal{Y}$ for some $q > 0$. Then, for any $0 < \varepsilon < \frac{1}{2}$, the following results hold:*

(i) $\log N(\epsilon, \mathcal{P}(\Xi), \|\cdot\|_\infty) \lesssim \log(1/\epsilon)$,

(ii) $H_B(\epsilon, \mathcal{P}(\Xi), h) \lesssim \log(1/\epsilon)$.

Now we could get the conclusion for Theorem 1.

C.2   Proof of Lemmas for Theorem 1

Now we will prove Lemma 5, Lemma 6 and Lemma 7 in order. At first we need to introduced another Lemma 8 before we prove Lemma 5. Lemma 8 is Theorem 5.11 in van de Geer (2000) and its proof can also be found in van de Geer (2000).

Lemma 8. *Let $R > 0$, $k \geq 1$ and $\mathcal{G}$ is a subset in Ξ where $G_* \in \mathcal{G} \subset \Xi$. Given $C_1 < \infty$, for all C sufficiently large, and for $n \in \mathbb{N}$ and $t > 0$ is in the following range*

$$t \leq (8\sqrt{n}R) \wedge (C_1\sqrt{n}R^2/K), \tag{61}$$

$$t \geq C^2(C_1+1)\left(R \vee \int_{t/(2^6\sqrt{n})}^{R} H_B^{1/2}\big(\frac{u}{\sqrt{2}}, \overline{\mathcal{P}}^{1/2}(\Xi, R), \mu\big)du\right), \tag{62}$$

then we will have

$$\mathbb{P}_{\lambda^*,G_*}\left(\sup_{G \in \mathcal{G}, h(\bar{p}_{\lambda,G}, p_{\lambda^*,G_*}) \leq R} |\mu_n(G)| \geq t\right) \leq C \exp\left(-\frac{t^2}{C^2(C_1+1)R^2}\right). \tag{63}$$

Proof of Lemma 5. Firstly, by Lemma 4.1 and 4.2 in van de Geer (2000), we have

$$\frac{1}{16}h^2(p_{\widehat{\lambda}_n, \widehat{G}_n}, p_{\lambda^*,G_*}) \leq h^2(\bar{p}_{\widehat{\lambda}_n, \widehat{G}_n}, p_{\lambda^*,G_*}) \leq \frac{1}{\sqrt{n}}\mu_n(\widehat{\lambda}_n, \widehat{G}_n),$$

here $\mu_n(\widehat{\lambda}_n, \widehat{G}_n)$ is an empirical process defined as

$$\mu_n(\widehat{\lambda}_n, \widehat{G}_n) := \sqrt{n}\int_{p_{\lambda^*,G_*}>0} \frac{1}{2}\log\big(\frac{\bar{p}_{\widehat{\lambda}_n, \widehat{G}_n}}{p_{\lambda^*,G_*}}\big)(\bar{p}_{\widehat{\lambda}_n, \widehat{G}_n} - p_{\lambda^*,G_*})d(X,Y).$$

Thus, for any $\delta > \delta_n := \sqrt{\log n/n}$, we have

$$\mathbb{P}_{\lambda^*,G_*}(h(p_{\widehat{\lambda}_n, \widehat{G}_n}, p_{\lambda^*,G_*}) \geq \delta)$$

$$\leq \mathbb{P}_{\lambda^*,G_*}\left(\mu_n(\widehat{\lambda}_n, \widehat{G}_n) - \sqrt{n}h^2(p_{\widehat{\lambda}_n, \widehat{G}_n}, p_{\lambda^*,G_*}) \geq 0, h(p_{\widehat{\lambda}_n, \widehat{G}_n}, p_{\lambda^*,G_*}) \geq \frac{\delta}{4}\right)$$

$$\leq \mathbb{P}_{\lambda^*,G_*}\left(\sup_{\lambda,G:h(\bar{p}_{\lambda,G}, p_{\lambda^*,G_*}) \geq \delta/4} \big[\mu_n(\lambda, G) - \sqrt{n}h^2(\bar{p}_{\lambda,G}, p_{\lambda^*,G_*})\big] \geq 0\right)$$

$$\leq \sum_{s=0}^{S} \mathbb{P}_{\lambda^*,G_*}\left(\sup_{\lambda,G:2^s\delta/4 \leq h(\bar{p}_{\lambda,G}, p_{\lambda^*,G_*}) \leq 2^{s+1}\delta/4} |\mu_n(\lambda, G)| \geq \sqrt{n}2^{2s}(\frac{\delta}{4})^2\right)$$

$$\leq \sum_{s=0}^{S} \mathbb{P}_{\lambda^*,G_*}\left(\sup_{\lambda,G:h(\bar{p}_{\lambda,G}, p_{\lambda^*,G_*}) \leq 2^{s+1}\delta/4} |\mu_n(\lambda, G)| \geq \sqrt{n}2^{2s}(\frac{\delta}{4})^2\right)$$

Fanqi Yan$^{\star,\diamond}$, Huy Nguyen*,†, Dung Le*,†, Pedram Akabrian‡, Nhat Ho†

where S is a smallest number such that $2^S\delta/4 > 1$.

Now we will use Lemma 8: choose $R = 2^{s+1}\delta, C_1 = 15$ and $t = \sqrt{n}2^{2s}(\delta/4)^2$. We can confirm that condition (i) in Lemma 3 is met since $2^{s-1}\delta/4 \leq 1$ for all $s \leq S$. For the condition (ii), it is still satisfied since

$$\int_{t/2^6\sqrt{n}}^{R} H_B^{1/2}\left(\frac{u}{\sqrt{2}}, \mathcal{P}^{1/2}(\Xi, R), \mu\right) du \vee 2^{s+1}\delta$$
$$= \sqrt{2}\int_{R^2/2^{13}}^{R/\sqrt{2}} H_B^{1/2}\left(u, \mathcal{P}^{1/2}(\Xi, R), \mu\right) du \vee 2^{s+1}\delta$$
$$\leq 2\mathcal{J}_B\left(R, \mathcal{P}^{1/2}(\Xi, R), \mu\right)$$
$$\leq 2J\sqrt{n}2^{2s+1}\delta^2$$
$$= 2^6 Jt.$$

Now since the two conditions in Lemma 8 are all satisfied, we could conclude that

$$\mathbb{P}_{\lambda^*,G_*}\left(h(p_{\widehat{\lambda}_n,\widehat{G}_n}, p_{\lambda^*,G_*}) > \delta\right) \leq C\sum_{s=0}^{\infty}\exp\left(\frac{2^{2s}n\delta^2}{J^2 2^{14}}\right) \leq c\exp\left(-\frac{n\delta^2}{c}\right), \tag{64}$$

here constant c is a large constant that does not depend on G_*. Now we could derive the bound on supremum of expectation:

$$\mathbb{E}_{p_{\lambda^*,G_*}} h(p_{\widehat{\lambda}_n,\widehat{G}_n}, p_{\lambda^*,G_*}) = \int_0^{\infty} \mathbb{P}\left(h(p_{\widehat{\lambda}_n,\widehat{G}_n}, p_{\lambda^*,G_*}) > \delta\right) d\delta$$
$$\leq \delta_n + c\int_{\delta_n}^{\infty}\exp\left(-\frac{n\delta^2}{c^2}\right) d\delta$$
$$\leq \tilde{c}\delta_n,$$

here $\tilde{c}$ is independent from λ^*, G_* and $\delta_n := \sqrt{\log n/n}$. So we can conclude that

$$\sup_{(\lambda^*,G_*)\in\Xi} \mathbb{E}_{p_{\lambda^*,G_*}} h(p_{\widehat{\lambda}_n,\widehat{G}_n}, p_{\lambda^*,G_*}) \leq C\sqrt{\log n/n}.$$

$\square$

Proof of Lemma 6. Because $\overline{\mathcal{P}}^{1/2}(\Xi, \delta) \subset \overline{\mathcal{P}}^{1/2}(\Xi)$ and from the definition of Hellinger distance, we have

$$H_B(\delta, \overline{\mathcal{P}}^{1/2}(\Xi, \delta), \mu) \leq H_B(\delta, \overline{\mathcal{P}}^{1/2}(\Xi), \mu) = H_B\left(\frac{\delta}{\sqrt{2}}, \overline{\mathcal{P}}(\Xi), h\right).$$

Now, using the fact that for densities f^*, f_1, f_2, we have $h^2\left(\frac{f_1+f^*}{2}, \frac{f_2+f^*}{2}\right) \leq \frac{h^2(f_1,f_2)}{2}$, it is easy to verify that $H_B(\delta/\sqrt{2}, \overline{\mathcal{P}}(\Xi), h) \leq H_B(\delta, \mathcal{P}(\Xi), h)$. Hence, if equation (60) holds true, then

$$H_B(\delta, \overline{\mathcal{P}}^{1/2}(\Xi, \delta), \mu) \leq H_B(\delta, \mathcal{P}(\Xi), h) \lesssim \log\left(\frac{1}{\delta}\right).$$

This implies that

$$\mathcal{J}_B\left(\epsilon, \overline{\mathcal{P}}^{1/2}(\Xi, \delta), \mu\right) \lesssim \epsilon\left(\log(\frac{2^{13}}{\epsilon^2})\right)^{\frac{1}{2}} < n\epsilon^2, \quad \text{for all } d\epsilon > \sqrt{\frac{\log n}{n}}.$$

$\square$

Proof of Lemma 7. **Proof for (i):** For any set S, we denote $\mathcal{E}_\epsilon(S)$ an $\epsilon-$net of S if each element of S is within ϵ distance from some elements of $\mathcal{E}_\epsilon(S)$. So $|\mathcal{E}_\epsilon(S)| = N(\epsilon, S, \|\cdot\|_\infty)$ by the definition of the covering number.

Denote $\mathcal{P}(\Theta) = \{p_G : G \in \Theta\}$ and $p_G(X,Y) := f(Y|\sigma(a^\top X + b), \nu)\bar{f}(X)$. From Lemma 6 in Ho et al. (2022), we have

$$\log |\mathcal{E}_\epsilon(\mathcal{P}(\Theta))| = N(\epsilon, \mathcal{P}(\Theta), \|\cdot\|_\infty) \lesssim \log(1/\epsilon).$$

Indeed, for any $\lambda \in [0,1], G \in \Theta$, there exists $\widetilde{\lambda} \in \mathcal{E}_\epsilon([0,1]), \widetilde{G} \in \mathcal{E}_\epsilon(\mathcal{P}(\Theta))$ s.t. $|\lambda - \widetilde{\lambda}| \leq \epsilon$ and $\|p_G - p_{\widetilde{G}}\|_\infty \leq \epsilon$. By triangle inequalities we will have

$$\|p_{\lambda,G} - p_{\widetilde{\lambda},\widetilde{G}}\|_\infty \leq |\lambda - \widetilde{\lambda}|(\|f_0\|_\infty + \|f\|_\infty) + \widetilde{\lambda}\|p_G - p_{\widetilde{G}}\|_\infty \lesssim \epsilon.$$

Hence we have the conclusion that

$$\log N(\epsilon, \mathcal{P}(\Xi), \|\cdot\|_\infty) \lesssim \log(|\mathcal{E}_\epsilon([0,1])|) + \log(|\mathcal{E}_\epsilon(\mathcal{P}(\Theta))|) \lesssim \log(1/\epsilon).$$

Proof for (ii): First, let $\eta \leq \varepsilon$ be a positive number, which will be chosen later. We consider f is the density function of an univariate Gaussian distribution, so f is light tail: for any $|Y| \geq 2a$ and $X \in \mathcal{X}$,

$$f(Y|\sigma(a^\top X + b), \nu) \leq \frac{1}{\sqrt{2\pi\ell}} \exp\left(-\frac{Y^2}{8u^2}\right).$$

Also f_0 is bounded with tail $\log f_0(Y|\varphi(a_0^\top X + b_0), \nu_0) \lesssim -Y^q$ and $f_0(Y|\varphi(a_0^\top X + b_0), \nu_0) \leq M$ for almost surely $Y \in \mathcal{Y}$ for some $M, q > 0$. Now let $q = \min\{p, 2\}$ and $C_2 = \max\{M, 1/\sqrt{2\pi\ell}\}$, we will have

$$H(X,Y) = \begin{cases} C_1 \exp(-Y^q)\bar{f}(X), & |Y| \geq 2a \\ C_2\bar{f}(X), & |Y| < 2a \end{cases} \tag{65}$$

here C_1 is a positive constant depending on ℓ and f_0. Moreover $H(X,Y)$ is an envelope of $\mathcal{P}(\Xi)$. Next, let $g_1, \ldots, g_N$ represent an η-net over $\mathcal{P}_k(\Xi)$. Then, we construct the brackets $[p_i^L(X,Y), p_i^U(X,Y)]$ as follows:

$$\begin{cases} p_i^L(X,Y) := \max\{g_i(X,Y) - \eta, 0\} \\ p_i^U(X,Y) := \min\{g_i(X,Y) + \eta, H(X,Y)\} \end{cases}$$

for $i = 1, \cdots, N$. As a result, $\mathcal{P}_k(\Xi) \subset \bigcup_{i=1}^N [p_i^L(X,Y), p_i^U(X,Y)]$ and $p_i^U(X,Y) - p_i^L(X,Y) \leq \min\{2\eta, H(X,Y)\}$. Consequently,

$$\int \left(p_i^U(X,Y) - p_i^L(X,Y)\right) d(X,Y)$$
$$\leq \int_{|Y|<2a} \left(p_i^U(X,Y) - p_i^L(X,Y)\right) d(X,Y) + \int_{|Y|\geq 2a} \left(p_i^U(X,Y) - p_i^L(X,Y)\right) d(X,Y)$$
$$\leq \int_{|Y|<2a} 2\eta d(X,Y) + \int_{|Y|\geq 2a} H(X,Y)d(X,Y) \lesssim \eta.$$

This shows that

$$H_B(c\eta, \mathcal{P}(\Xi), \|\cdot\|_1) \leq N \lesssim \log(1/\eta).$$

Setting $\eta = \epsilon/c$, we find

$$H_B(\epsilon, \mathcal{P}(\Xi), \|\cdot\|_1) \lesssim \log(1/\epsilon).$$

Since $h^2 \leq \|\cdot\|_1$ holds between the Hellinger distance and the total variation distance, we conclude the bracketing entropy bound. $\square$

D   PROOFS FOR AUXILIARY RESULTS

D.1   The derivation of the heat equation

If f represents a family of univariate Gaussian distributions:

$$f(Y|\sigma(a^\top X + b), \nu) = \frac{1}{\sqrt{2\pi\nu}} \exp\left(-\frac{(Y - \sigma(a^\top X + b))^2}{2\nu}\right).$$

Fanqi Yan[⋆,◇]**, Huy Nguyen**[⋆,†]**, Dung Le**[⋆,†]**, Pedram Akabrian**[‡]**, Nhat Ho**[†]

We will have

$$
\frac{\partial^2 f}{\partial \sigma^2} = f(Y|\sigma(a^\top X + b), \nu) \cdot \left[\frac{(Y - \sigma(a^\top X + b))^2}{\nu^2} - \frac{1}{\nu} \right],
$$

$$
\frac{\partial f}{\partial \nu} = f(Y|\sigma(a^\top X + b), \nu) \cdot \left[\frac{(Y - \sigma(a^\top X + b))^2}{2\nu^2} - \frac{1}{2\nu} \right].
$$

So $f(Y|\sigma(a^\top X + b), \nu)$ satisfies the heat partial differential equation (PDE) with respect to the location and covariance parameters:

$$
\frac{\partial^2 f}{\partial \sigma^2}(Y|\sigma(a^\top X + b), \nu) = 2\frac{\partial f}{\partial \nu}(Y|\sigma(a^\top X + b), \nu), \tag{66}
$$

for almost surly $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ and $(a, b, \nu) \in \Theta$. In Section 3.2.1, we consider σ as an identity function, i.e. $\sigma = a^\top X + b$. Thus, we obtain:

$$
\frac{\partial \sigma}{\partial b} = 1.
$$

Then, the second derivative for f respect to b is:

$$
\frac{\partial^2 f}{\partial b^2} = \frac{\partial}{\partial b}\left(\frac{\partial f}{\partial b}\right) = \frac{\partial}{\partial b}\left(\frac{\partial f}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial b}\right) = \frac{\partial}{\partial b}\left(\frac{\partial f}{\partial \sigma} \cdot 1\right) = \frac{\partial^2 f}{\partial \sigma^2}.
$$

So we will have

$$
\frac{\partial^2 f}{\partial b^2} = 2\frac{\partial f}{\partial \nu},
$$

where the equation is what we stated in equation (3).

D.2   Proof of Proposition 1

Proof. Suppose that $p_{\lambda,G}(Y|X) = p_{\lambda',G'}(Y|X)$ holds for almost sure every $(X, Y) \in \mathcal{X} \times \mathcal{Y}$:

$$
\lambda f_0(Y|\varphi(a_0^\top X + b_0), \nu_0) + (1-\lambda)f(Y|\sigma(a^\top X + b), \nu) = \lambda' f_0(Y|\varphi(a_0^\top X + b_0), \nu_0) + (1-\lambda')f(Y|\sigma((a')^\top X + b'), \nu').
$$

This means that

$$
(\lambda' - \lambda)f_0(Y|\varphi(a_0^\top X + b_0), \nu_0) + (1-\lambda')f(Y|\sigma((a')^\top X + b'), \nu') - (1-\lambda)f(Y|\sigma(a^\top X + b), \nu) = 0. \tag{67}
$$

If $G = G'$, let $\tilde{G}$ be other parameter in the parameter set. Then, equation (67) becomes

$$
(\lambda' - \lambda)f_0(Y|\varphi(a_0^\top X + b_0), \nu_0) + (\lambda - \lambda')f(Y|\sigma(a^\top X + b), \nu) + 0f(Y|\sigma(\tilde{a}^\top X + \tilde{b}), \tilde{\nu}) = 0.
$$

Applying the equation in Definition 1 with $\eta_0 = \lambda' - \lambda$, $\eta_1 = \lambda - \lambda'$, and $\eta_2 = 0$, we get that $\lambda = \lambda'$, which means that $(\lambda, G) = (\lambda', G')$.

Otherwise, if $G \neq G'$, then by pplying the equation in Definition 1 with $\eta_0 = \lambda' - \lambda$, $\eta_1 = 1 - \lambda$, and $\eta_2 = \lambda' - 1$ to equation (67), we get $\lambda = \lambda'$, and $1 - \lambda = 0$. This mean that $\lambda = \lambda' = 1$, which means that the parts G and G' have no contribution. Thus, it is natural to suppose that $(\lambda, G) = (\lambda', G')$. □

D.3   Proof of Proposition 2

Proof. Recall the Definition 1 for the distinguishibility . At first if $\eta_0 \neq 0$, then

$$
f_0(Y|\varphi(a_0^\top X + b_0), \nu_0) = -\sum_{i=1}^{2} \frac{\eta_i}{\eta_0} f(Y|a_i^\top X + b_i, \nu_i)
$$

contradicts that $f_0 \neq f$, so $\eta_0 = 0$. Then since $(a_i, b_i, \nu_i), i = 1, 2$ are two distinct components, means that the mean and variance are different, then from the property of Gaussian distribution, we will have $\eta_1 = \eta_2 = 0$. So that now f is distinguishable from f_0. □

D.4   Proof of Proposition 3

Proof. In this proof, we assume technically that the prior density distribution $\bar{f}(X)$ is non-vanishing almost everywhere. This assumption is actually realistic, because adding a quite small independent Gaussian noise can make the density function to be on-vanishing almost everywhere. This assumption is crucial for guarantee that X can attain any value with a sufficient probability. Moreover, we also assume that the prompt function σ should not have the form $\pm \log\left(x + c_1 + \sqrt{(x + c_1)^2 + c_2}\right) + C$. To our knowledge, this expert function has not appeared in common activation function in Deep Learning, which makes our assumption is reasonable.

Suppose that there exists α_η such that

$$\sum_{\ell=0}^{2} \sum_{|\eta|=\ell} \alpha_\eta \cdot \frac{\partial^{|\eta|} f}{\partial a^{\eta_1} \partial b^{\eta_2} \partial \nu^{\eta_3}} (Y|\sigma(a^\top X + b), \nu) = 0. \tag{68}$$

In order to do this exercise, it is necessary to calculate the derivative of order at most of f. The first-order derivative are:

$$\frac{\partial f}{\partial a_i} = \frac{\partial f}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial a_i}, \quad \frac{\partial f}{\partial b} = \frac{\partial f}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial b},$$

The second order derivatives are:

$$\frac{\partial^2 f}{\partial a_i \partial a_j} = \frac{\partial^2 f}{\partial \sigma^2} \cdot \frac{\partial \sigma}{\partial a_i} \frac{\partial \sigma}{\partial a_j} + \frac{\partial f}{\partial \sigma} \cdot \frac{\partial^2 \sigma}{\partial a_i \partial a_j}, \quad \frac{\partial^2 f}{\partial b^2} = \frac{\partial^2 f}{\partial \sigma^2} \cdot \left(\frac{\partial \sigma}{\partial b}\right)^2 + \frac{\partial f}{\partial \sigma} \cdot \frac{\partial^2 \sigma}{\partial b^2},$$

$$\frac{\partial^2 f}{\partial a_i \partial b} = \frac{\partial^2 f}{\partial \sigma^2} \cdot \frac{\partial \sigma}{\partial a_i} \frac{\partial \sigma}{\partial b} + \frac{\partial f}{\partial \sigma} \cdot \frac{\partial^2 \sigma}{\partial a_i \partial b}, \quad \frac{\partial f}{\partial \nu} = \frac{1}{2} \frac{\partial^2 f}{\partial \sigma^2}.$$

The third order derivatives are:

$$\frac{\partial^2 f}{\partial a_i \partial \nu} = \frac{1}{2} \frac{\partial^3 f}{\partial \sigma^3} \cdot \frac{\partial \sigma}{\partial a_i}, \quad \frac{\partial^2 f}{\partial b \partial \nu} = \frac{1}{2} \frac{\partial^3 f}{\partial \sigma^3} \cdot \frac{\partial \sigma}{\partial b}.$$

The forth order derivatives are:

$$\frac{\partial^2 f}{\partial \nu^2} = \frac{1}{4} \frac{\partial^4 f}{\partial \sigma^4}.$$

Then, the equation (68) can be considered as a linear combination of $\partial^k f / \partial \sigma^k$. According to Lemma 9, these coefficients in the linear combination are 0's. We consider these coefficients in decreasing order of the derivative.

Order 4:

There is only one term that contribute to the $\partial^4 f / \partial \sigma^4$, that is $\partial^2 f / \partial \nu^2$. As the coefficient of $\partial^4 f / \partial \sigma^4$ is 0, the coefficient of $\partial^2 f / \partial \nu^2$ is also 0.

Order 3:

The terms contributing to the $\partial^3 f / \partial \sigma^3$ are $\dfrac{\partial^2 f}{\partial a_i \partial \nu}$ and $\dfrac{\partial^2 f}{\partial b \partial \nu}$. Let ρ_i be the coefficient of $\dfrac{\partial^2 f}{\partial a_i \partial \nu}$ and ω be the coefficient of $\dfrac{\partial^2 f}{\partial b \partial \nu}$. Then, by summing up all the coefficient with respect to $\partial^3 f / \partial \sigma^3$, we achieve that

$$\omega \frac{\partial \sigma}{\partial b} + \sum_{i=1}^{n} \rho_i \frac{\partial \sigma}{\partial a_i} = 0 \Leftrightarrow \left(\omega + \rho^\top X\right) \sigma'(a^\top X + b) = 0, \tag{69}$$

where $\rho = (\rho_1 \ldots \rho_n)^\top$ for brevity. As our random vector X has density function that does not vanish almost everywhere in $\mathbb{R}^n$, equation (69) is equivalent to

$$\left(\omega + \rho^\top x\right) \sigma'(a^\top x + b) = 0$$

almost everywhere in $\mathbb{R}^n$. As σ is not linear, its derivative does not equal to 0 in a set $B \subset \mathbb{R}$ of positive measure. As $a \neq 0$, we can find subset $\tilde{B} \subset \mathbb{R}^n$ of positive measure such that $a^\top x + b \in B$ for all $x \in \tilde{B}$, in other words, $\sigma'(a^\top x + b) \neq 0$. Thus, $\omega + \rho^\top x = 0$ for $x \in \tilde{B}$. This equality occurs if and only if $\omega = \rho = 0$. Thus, all the coefficients of $\dfrac{\partial^2 f}{\partial a_i \partial \nu}$ and $\dfrac{\partial^2 f}{\partial b \partial \nu}$ are 0.

Fanqi Yan[⋆,⋄], **Huy Nguyen**[⋆,†], **Dung Le**[⋆,†], **Pedram Akabrian**[‡], **Nhat Ho**[†]

Order 2:

The terms contributing to the $\partial^2 f / \partial \sigma^2$ are $\dfrac{\partial^2 f}{\partial a_i \partial a_j}, \dfrac{\partial^2 f}{\partial a_i \partial b}, \dfrac{\partial^2 f}{\partial b^2}, \dfrac{\partial f}{\partial \nu}$. Let $\alpha_{i,j}$ be the coefficients of $\dfrac{\partial^2 f}{\partial a_i \partial a_j}$, β_i be the coefficients of $\dfrac{\partial^2 f}{\partial a_i \partial b}$, γ be the coefficient of $\dfrac{\partial^2 f}{\partial b^2}$, and ξ be the coefficient of $\dfrac{\partial f}{\partial \nu}$. Then, by summing up all the coefficient with respect to $\partial^2 f / \partial \sigma^2$, we achieve that

$$\sum_{i,j=1}^{n} \alpha_{i,j} \frac{\partial \sigma}{\partial a_i} \frac{\partial \sigma}{\partial a_j} + \sum_{i=1}^{n} \beta_i \frac{\partial \sigma}{\partial a_i} \frac{\partial \sigma}{\partial b} + \gamma \left(\frac{\partial \sigma}{\partial b} \right)^2 + \frac{1}{2} \xi = 0 \Leftrightarrow \left(\sum_{i,j=1}^{n} \alpha_{i,j} X_i X_j + \sum_{i=1}^{n} \beta_i X_i + \gamma \right) (\sigma'(a^\top X + b))^2 + \frac{1}{2} \xi = 0. \tag{70}$$

With the hypothesis that $a \neq 0$, by a change of base, we can suppose that $a = (1, 0, \ldots, 0)^\top$. Then, the equation above can be considred as

$$\left(\sum_{i,j=1}^{n} \alpha_{i,j} X_i X_j + \sum_{i=1}^{n} \beta_i X_i + \gamma \right) (\sigma'(X_1 + b))^2 + \frac{1}{2} \xi = 0.$$

As our random vector X has density function that does not vanish almost everywhere in $\mathbb{R}^n$, equation (70) is equivalent to

$$\left(\sum_{i,j=1}^{n} \alpha_{i,j} x_i x_j + \sum_{i=1}^{n} \beta_i x_i + \gamma \right) (\sigma'(x_1 + b))^2 + \frac{1}{2} \xi = 0. \tag{71}$$

WLOG, we can suppose that $\alpha_{i,j} = \alpha_{j,i}$. By letting $x_2, \ldots, x_n = 0$, we achieve that

$$(\alpha_{1,1} x_1^2 + \beta_1 x_1 + \gamma)(\sigma'(x_1 + b))^2 + \frac{1}{2} \xi = 0. \tag{72}$$

By substituting equation (72) in equation (71), we receive that

$$\left(\sum_{i,j \neq 1}^{n} \alpha_{i,j} x_i x_j + 2 \sum_{i=1, j \neq 1}^{n} \alpha_{i,j} x_i x_j + \sum_{i=2}^{n} \beta_i \right) (\sigma'(x_1 + b))^2 = 0.$$

As σ is not a linear function, there exists a set $B \subset \mathbb{R}$ of positive measure such that $\sigma' \neq 0$. Thus, it is possible to find a set $B' \subset \mathbb{R}$ such that $\sigma'(x_1 + b) \neq 0$. In this set,

$$\sum_{i,j \neq 1}^{n} \alpha_{i,j} x_i x_j + 2 \sum_{i=1, j \neq 1}^{n} \alpha_{i,j} x_i x_j + \sum_{i=2}^{n} \beta_i = 0$$

This equation implies that $\alpha_{i,j} = 0$ for $(i,j) \neq (1,1)$, and $\beta_i = 0$ for $i = 2, \cdots, n$.

Next, we consider the equation (72). Suppose that not all the coefficient $\alpha_{1,1}$, β_1, and γ is equal to 0. We examine it in three different setting:

Case 1: If $\alpha_{1,1} = 0$. Then, if $\beta_1 \neq 0$, then $(\beta_1 x_1 + \gamma)(\sigma'(x_1 + b))^2 + \frac{1}{2} \xi = 0$. Thus, $(\sigma'(x_1 + b))^2 = -\xi/2(\beta_1 x_1 + \gamma) \geq 0$ almost sure, this occurs if and only if $\xi = 0$. Then, $\sigma'(x_1 + b) = 0$, which implies that σ be a constant, a contradiction to the hypothesis that σ is not linear.

Otherwise, if $\beta_1 = 0$, then $\gamma(\sigma'(x_1+b))^2 + \xi/2 = 0$. If $\gamma \neq 0$, then this equation occurs if and only if $\sigma'(x_1+b) = 0$ for all $x_1 \in B'$, which is a contradiction to our hypothesis that $\sigma'(x_1 + b) \neq 0$ in B'. If $\gamma = 0$, then $\xi = 0$, which is also a contradiction to the hypothesis that not all the coefficient $\alpha_{1,1}$, β_1, and γ is equal to 0.

Case 2: If $\alpha_{1,1} \neq 0$. Then, by letting $y = x_1 + b$, $c_1 = \beta_1/(2\alpha_{1,1}) - b$, $c_2 = \gamma/\alpha_{1,1} - (\beta_1/(2\alpha))^2$, $c = -\xi/(2\alpha_{1,1})$, equation (72) becomes

$$\sigma'(y)^2 = \frac{c}{(y + c_1)^2 + c_2}.$$

If $c_2 < 0$, then for $y \in (-c_1 - \sqrt{c_2}, -c_1 + \sqrt{c_2})$ we have $\sigma'(y)^2 < 0$, which is not possible.

If $c_2 = 0$, then $|\sigma'(y)| \to \infty$ as $y \to -c_1$, which is not possible as σ' be bounded in a compact set due to its continuous property.

If $c_2 > 0$, then we have the famous formulae $\sigma(y) = \pm \log\left(y + c_1 + \sqrt{(y + c_1)^2 + c_2}\right) + C$, which is also a contradiction with the hypothesis of σ

Order 1:

The terms contributing to the $\partial f / \partial \sigma$ are $\dfrac{\partial f}{\partial a_i}, \dfrac{\partial f}{\partial b}, \dfrac{\partial^2 f}{\partial a_i \partial a_j}, \dfrac{\partial^2 f}{\partial a_i \partial b}, \dfrac{\partial^2 f}{\partial b^2}$. Recall we have $\alpha_{i,j}$ as the coefficients

of $\dfrac{\partial^2 f}{\partial a_i \partial a_j}$, β_i as the coefficients of $\dfrac{\partial^2 f}{\partial a_i \partial b}$, γ as the coefficient of $\dfrac{\partial^2 f}{\partial b^2}$, let δ_i be the coefficient of $\dfrac{\partial f}{\partial a_i}$, η be the

coefficient of $\dfrac{\partial f}{\partial b}$. Then, by summing up all the coefficient with respect to $\partial f / \partial \sigma$, we achieve that

$$\sum_{i,j=1}^{n} \alpha_{i,j} \frac{\partial^2 \sigma}{\partial a_i \partial a_j} + \sum_{i=1} \beta_i \frac{\partial^2 \sigma}{\partial a_i \partial b} + \gamma \frac{\partial^2 \sigma}{\partial b^2} + \sum_{i=1} \delta_i \frac{\partial \sigma}{\partial a_i} + \eta \frac{\partial \sigma}{\partial b} = 0$$

$$\Leftrightarrow \left(\sum_{i,j=1}^{n} \alpha_{i,j} X_i X_j + \sum_{i=1}^{n} \beta_i X_i + \gamma \right) \sigma''(a^\top X + b) + \left(\delta^\top X + \eta \right) \sigma'(a^\top X + b) = 0.$$

By the proof of above for order 2, all the coefficient of $\dfrac{\partial^2 f}{\partial a_i \partial a_j}, \dfrac{\partial^2 f}{\partial a_i \partial b}, \dfrac{\partial^2 f}{\partial b^2}$ are equal to 0. Thus, we only take

into account the terms $\dfrac{\partial f}{\partial a_i}, \dfrac{\partial f}{\partial b}$. Denote δ_i be the coefficients of $\dfrac{\partial f}{\partial a_i}$, and η be the coefficient of $\dfrac{\partial f}{\partial b}$. Then,

$$\eta \frac{\partial \sigma}{\partial b} + \sum_{i=1}^{n} \delta_i \frac{\partial \sigma}{\partial a_i} = 0 \Leftrightarrow \left(\eta + \delta^\top X \right) \sigma'(a^\top X + b) = 0,$$

where $\delta = (\delta_1 \ldots \delta_n)^\top$ for brevity. Using the same argument as for the case of order 3, we receive that all the coefficient of $\dfrac{\partial f}{\partial a_i}, \dfrac{\partial f}{\partial b}$ are equal to 0.

Order 0:

The rest term that was not considered is f without derivative. It is obvious that the coefficient of f is equal to 0. $\qquad\square$

D.5 Linear Independence of Gaussian Derivatives

Lemma 9. *Let $f(x, \sigma) = \frac{1}{\sqrt{2\pi\nu}} \exp\left(-\frac{(x-\sigma)^2}{2\nu}\right)$ be the Gaussian density/heat kernel. Consider the scalar (with respect to x) $a_0, \ldots, a_n$ such that*

$$\sum_{i=0}^{n} a_i \frac{\partial^n f}{\partial \sigma^n}(x, \sigma) = 0, \tag{73}$$

for almost every x, σ. Then, $a_i = 0$ for $i = 0, \ldots, n$.

Proof. By taking the Fourier transformation with respect to σ in both side of equation (73), we have

$$\left(\sum_{k=0}^{n} a_k (i\hat\sigma)^k \right) \hat{f}(x, \hat\sigma) = 0$$

for almost every $x, \hat\sigma$, where $\hat{f}$ be the Fourier transformation of f, which is also a Gaussian density/heat kernel. Thus, we have

$$\sum_{k=0}^{n} a_k (i\hat\sigma)^k = 0$$

for almost every $\hat\sigma$, which means that $a_k = 0$, $0 \le k \le n$. $\qquad\square$

Fanqi Yan[*,◇], **Huy Nguyen**[*,†], **Dung Le**[*,†], **Pedram Akabrian**[‡], **Nhat Ho**[†]

E   ADDITIONAL EXPERIMENTS

Synthetic data generation.

In this section, we perform numerical experiments to validate the theoretical results of Theorem 6, Theorem 7, and Theorem 8 in Appendix E.1 as they share the same convergence behavior of parameter estimation. Then, we continue to empirically justify the results of Theorem 4 and Theorem 9 in Appendix E.2 and Appendix E.3, respectively.

E.1   Theorems 6, 7 and 8

Let us begin with the experimental details for Theorems 6, 7 and 8.

Experimental setup.

- In Theorem 6, f_0 belongs to the Gaussian density family, φ is a non-linear function, we let φ be the sigmoid function, characterized by $\varphi(a_0^\top X + b_0) = 1/(1 + \exp\{-(a_0^\top X + b_0)\})$. Meanwhile, the prompt f is formulated as a Gaussian density with mean $\sigma((a^*)^\top X + b^*) = (a^*)^\top X + b^*$ and variance ν^*.

- In Theorem 7, since the pre-trained model f_0 does not belong to the Gaussian density family, we let f_0 be the density of a Student's t-distribution, characterized by mean $\varphi(a_0^\top X + b_0) = a_0^\top X + b_0$ and degrees of freedom $\nu_0 = 4$. Meanwhile, the prompt f is modeled as a Gaussian distribution with a non-linear sigmoid mean, given by $\sigma((a^*)^\top X + b^*) = 1/(1 + \exp\{-((a^*)^\top X + b^*)\})$, and a variance of ν^*.

- In Theorem 8, f_0 belongs to the Gaussian density family, φ is a linear function but σ is a non-linear function. We let φ be the identity function and σ be the sigmoid function. Specifically, the prompt f is formulated as a Gaussian density with mean $\sigma((a^*)^\top X + b^*) = 1/(1 + \exp\{-((a^*)^\top X + b^*)\})$ and variance ν^*.

True parameters. We consider two cases for the true parameters $(\lambda^*, G_*) = (\lambda^*, a^*, b^*, \nu^*)$ to examine the difference in the MLE convergence behavior when λ^* is fixed versus when it varies with n:

(i) $\lambda^* = 0.5, a^* = \mathbf{1}_d, b^* = 1, \nu^* = 0.01$;

(ii) $\lambda^* = 0.5 \ n^{-1/4}, a^* = \mathbf{1}_d, b^* = 1, \nu^* = 0.01$.

Results. We display the empirical illustration for Theorems 6, 7 and 8 in Figures 2, 3 and 4, respectively. Although these three scenarios differ significantly, they share the same property that the pre-trained model f is distinguishable from the prompt f_0. As a result, the convergence behavior of parameter estimation under the settings of those theorems are quite similar. First, we observe that the convergence rates of $\widehat{\lambda}_n$ in both cases across all three scenarios are approximately $\mathcal{O}(n^{-1/2})$. However, there are mismatches in the convergence rates of $(\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n)$ between the two cases. In case (i), where λ^* is fixed, $(\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n)$ exhibit similar rates, approximately $\mathcal{O}(n^{-1/2})$, as seen in Figures 2a, 3a and 4a. In case (ii), where λ^* vanishes at the rate of $\mathcal{O}(n^{-1/4})$, their rates slow down significantly to around $\mathcal{O}(n^{-1/4})$, as shown in Figures 2b, 3b and 4b. These empirical results are fully consistent with the theoretical findings presented in the three theorems, which say that

$$\sup_{(\lambda^*, G_*) \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} \left[|\widehat{\lambda}_n - \lambda^*|^2 \right] = \widetilde{\mathcal{O}}(n^{-1}),$$

$$\sup_{(\lambda^*, G_*) \in \Xi} \mathbb{E}_{p_{\lambda^*, G_*}} \left[(\lambda^*)^2 \|(\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n) - (a^*, b^*, \nu^*)\|^2 \right] = \widetilde{\mathcal{O}}(n^{-1}).$$

E.2   Theorem 4

Subsequently, we perform numerical experiments to validate theoretical results presented in Theorem 4 where both the pre-trained model f and the prompt f_0 belong to the family of Gaussian densities, while the experts φ and σ are the same linear functions. Here, we let φ and σ be identity functions.

True parameters. We consider two following cases of the true parameters $(\lambda^*, G_*) = (\lambda^*, a^*, b^*, \nu^*)$ to capture the difference in the MLE convergence behavior when $a^* \to a_0$ and $\nu^* \to \nu_0$ versus when $b^* \to b_0$ as $n \to \infty$:

(i) $\lambda^* = 0.5, a^* = (1 + n^{-1/8}) \cdot e_1, b^* = 0, \nu^* = 0.01 + n^{-1/8}$;

(ii) $\lambda^* = 0.5, a^* = e_1, b^* = n^{-1/8}, \nu^* = 0.01$.

Results. We exhibit the empirical results of Theorem 4 under the above two cases in Figure 5. Prior to analyzing them, let us recall the theoretical rates in Theorem 4:

$$\sup_{(\lambda^*, G_*) \in \Xi_1(l_n)} \mathbb{E}_{p_{\lambda^*, G_*}} \left[(\|\Delta a^*\|^4 + |\Delta b^*|^8 + |\Delta \nu^*|^4) |\widehat{\lambda}_n - \lambda^*|^2 \right] = \widetilde{\mathcal{O}}(n^{-1}),$$

$$\sup_{(\lambda^*, G_*) \in \Xi_1(l_n)} \mathbb{E}_{p_{\lambda^*, G_*}} \left[(\lambda^*)^2 (\|\Delta a^*\|^2 + |\Delta b^*|^4 + |\Delta \nu^*|^2)(\|\widehat{a}_n - a^*\|^2 + |\widehat{b}_n - b^*|^4 + |\widehat{\nu}_n - \nu^*|^2) \right] = \widetilde{\mathcal{O}}(n^{-1}).$$

(i) In the first case when λ^*, b^* are fixed while $a^* \to a_0$ and $\nu^* \to \nu_0$ at the rate of $\mathcal{O}(n^{-1/8})$, the first bound indicates that $\widehat{\lambda}_n$ converges to λ^* at the theoretical rate of $\widetilde{\mathcal{O}}(n^{-1/4})$. Meanwhile, the convergence rates of $\widehat{a}_n, \widehat{b}_n$ and $\widehat{\nu}_n$ should be $\widetilde{\mathcal{O}}(n^{-3/8})$, $\widetilde{\mathcal{O}}(n^{-3/16})$ and $\widetilde{\mathcal{O}}(n^{-3/8})$, respectively. Now, looking at Figure 5a, we observe that the empirical convergence rates $\mathcal{O}(n^{-0.22})$, $\mathcal{O}(n^{-0.35})$, $\mathcal{O}(n^{-0.15})$ and $\mathcal{O}(n^{-0.38})$ of $\widehat{\lambda}_n, \widehat{a}_n, \widehat{b}_n$ and $\widehat{\nu}_n$ totally align with the theory.

(ii) In the second case when λ^*, a^*, ν^* are fixed while $b^* \to b_0$ at the rate of $\mathcal{O}(n^{-1/8})$, the first bound implies that $\widehat{\lambda}_n$ might not converge to λ^*. However, in case the convergence occurs, the rate would be significantly slow. This property is empirically illustrated in Figure 5b. Additionally, we also see from that figure that the empirical convergence rates of $\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n$ totally match their theoretical counterparts, specifically $\mathcal{O}(n^{-0.24})$, $\mathcal{O}(n^{-0.12})$, $\mathcal{O}(n^{-0.25})$ compared to $\widetilde{\mathcal{O}}(n^{-1/4})$, $\widetilde{\mathcal{O}}(n^{-1/8})$, $\widetilde{\mathcal{O}}(n^{-1/4})$.

(iii) It can be seen from the above observations that the convergence of b^* to b_0 causes more severe effects on the convergence rates of parameter estimation than those induced by the convergences of a^*, ν^* to a_0, ν_0. In particular, the former makes the mixing proportion $\widehat{\lambda}_n$ possibly not converge to the true value, while the convergence rates of $\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n$ are slower than those generated from the latter.

E.3 Theorem 9

Lastly, we present the empirical results for Theorem 9 where both f and f_0 belong to the Gaussian density family, φ and σ are the same nonlinear expert functions. Here, we let φ and σ be the sigmoid function, that is,

$$\varphi(z) = \sigma(z) = \frac{1}{1 + \exp(-z)},$$

for any $z \in \mathbb{R}$.

True parameters. We take two following cases of the true parameters $(\lambda^*, G_*) = (\lambda^*, a^*, b^*, \nu^*)$ into account to understand the effects of the convergence of a^*, b^*, ν^* to a_0, b_0, ν_0 on the parameter estimation rates:

(i) $\lambda^* = 0.5 \cdot n^{-1/4}, a^* = e_1, b^* = 0, \nu^* = 0.01$;

(ii) $\lambda^* = 0.5, a^* = (1 + n^{-1/4}), b^* = n^{-1/4}, \nu^* = 0.01 + n^{-1/4}$.

Results. Figure 6 empirically illustrates the parameter estimation rates presented in Theorem 9, which are given by

$$\sup_{(\lambda^*, G_*) \in \Xi_2(l_n)} \mathbb{E}_{p_{\lambda^*, G_*}} \left[\|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|^4 |\widehat{\lambda}_n - \lambda^*|^2 \right] = \widetilde{\mathcal{O}}(n^{-1}),$$
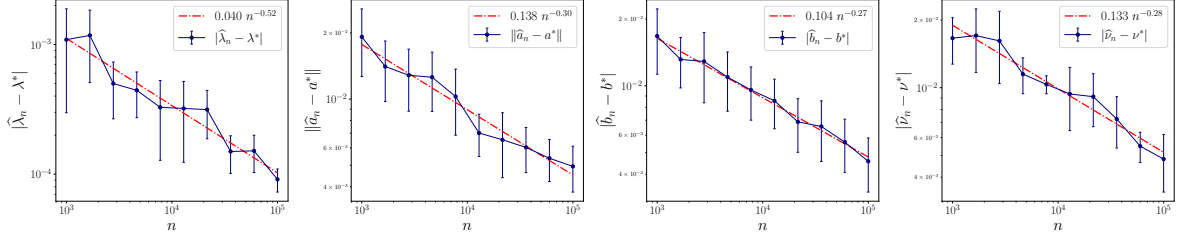
$$\sup_{(\lambda^*, G_*) \in \Xi_2(l_n)} \mathbb{E}_{p_{\lambda^*, G_*}} \left[(\lambda^*)^2 \|(\Delta a^*, \Delta b^*, \Delta \nu^*)\|^2 \|(\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n) - (a^*, b^*, \nu^*)\|^2 \right] = \widetilde{\mathcal{O}}(n^{-1}).$$

In case (i) when a^*, b^*, ν^* are fixed, Figure 6a shows that the convergence rate of $\widehat{\lambda}_n$ to λ^* is of order $\mathcal{O}(n^{-0.52})$, while those for $\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n$ are slower, standing at around $\mathcal{O}(n^{-0.25})$, as they hinges upon the vanishing rate $\mathcal{O}(n^{-1/4})$ of λ^*. These empirical rates are in line with the above theoretical rates.

In case (ii) when a^*, b^*, ν^* converge to a_0, b_0, ν_0 at the rate of $\mathcal{O}(n^{-1/4})$, respectively, Figure 6b reveals that $\widehat{\lambda}_n$ might not converge to λ^* as the displayed rate is significantly slow of order $\mathcal{O}(n^{-0.05})$. This observation totally aligns with the first bound. On the other hand, the MLE $\widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n$ still empirically converges to a^*, b^*, ν^* at the rates of $\mathcal{O}(n^{-0.22})$, $\mathcal{O}(n^{-0.26})$, $\mathcal{O}(n^{-0.22})$, which approximates the theoretical rates $\widetilde{\mathcal{O}}(n^{-1/4})$ very well.

Fanqi Yan[⋆,⋄], Huy Nguyen[⋆,†], Dung Le[⋆,†], Pedram Akabrian[‡], Nhat Ho[†]
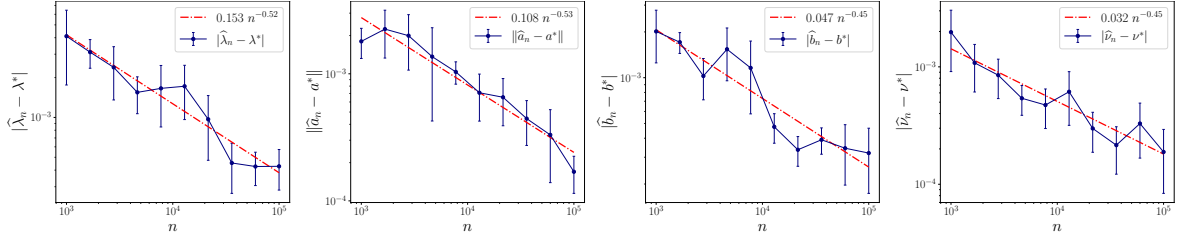

(a) **Case (i):** $\lambda^* = 0.5$.

(b) **Case (ii):** $\lambda^* = 0.5 \ n^{-1/4}$.

Figure 2: (**Theorem 6:** f_0 is Gaussian, $\sigma(z) = z$, $\varphi(z) = 1/(1 + e^{-z})$.) Log-log graphs depicting the empirical convergence rates of the MLE $(\widehat{\lambda}_n, \widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n)$ to the ground-truth values $(\lambda^*, a^*, b^*, \nu^*)$. The blue lines display the parameter estimation errors, while the red dashed dotted lines are the fitted lines, highlighting the empirical MLE convergence rates. Figure 2a and Figure 2b illustrates results for the cases when $\lambda^* = 0.5$ and $\lambda^* = 0.5 \ n^{-1/4}$, respectively.
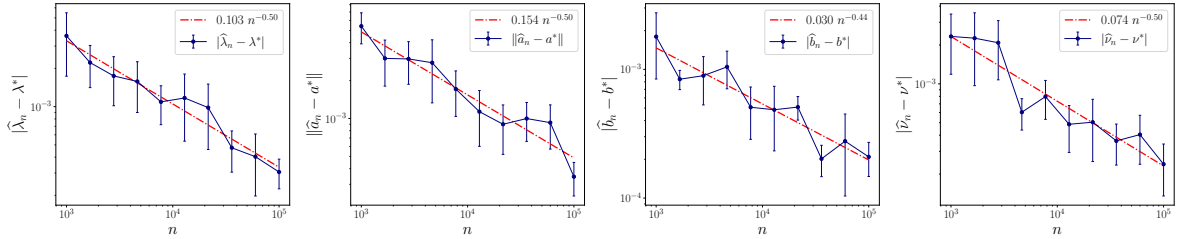
(a) **Case (i):** $\lambda^* = 0.5$.
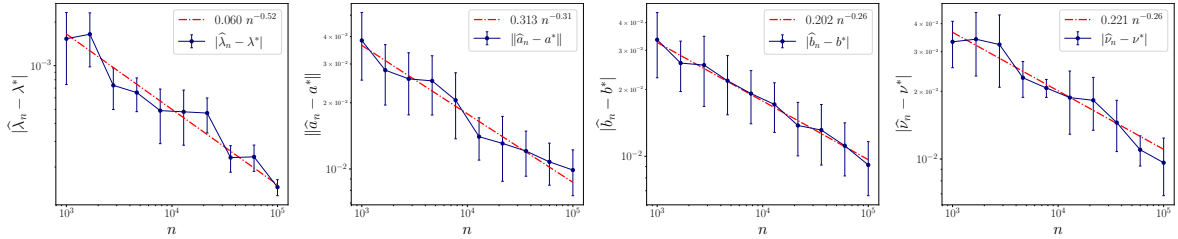
(b) **Case (ii):** $\lambda^* = 0.5 \ n^{-1/4}$.

Figure 3: (**Theorem 7:** f_0 is Student's t-distribution, $\sigma(z) = 1/(1 + e^{-z})$, $\varphi(z) = z$.) Log-log graphs depicting the empirical convergence rates of the MLE $(\widehat{\lambda}_n, \widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n)$ to the ground-truth values $(\lambda^*, a^*, b^*, \nu^*)$. The blue lines display the parameter estimation errors, while the red dashed dotted lines are the fitted lines, highlighting the empirical MLE convergence rates. Figure 3a and Figure 3b illustrates results for the cases when $\lambda^* = 0.5$ and $\lambda^* = 0.5 \ n^{-1/4}$, respectively.
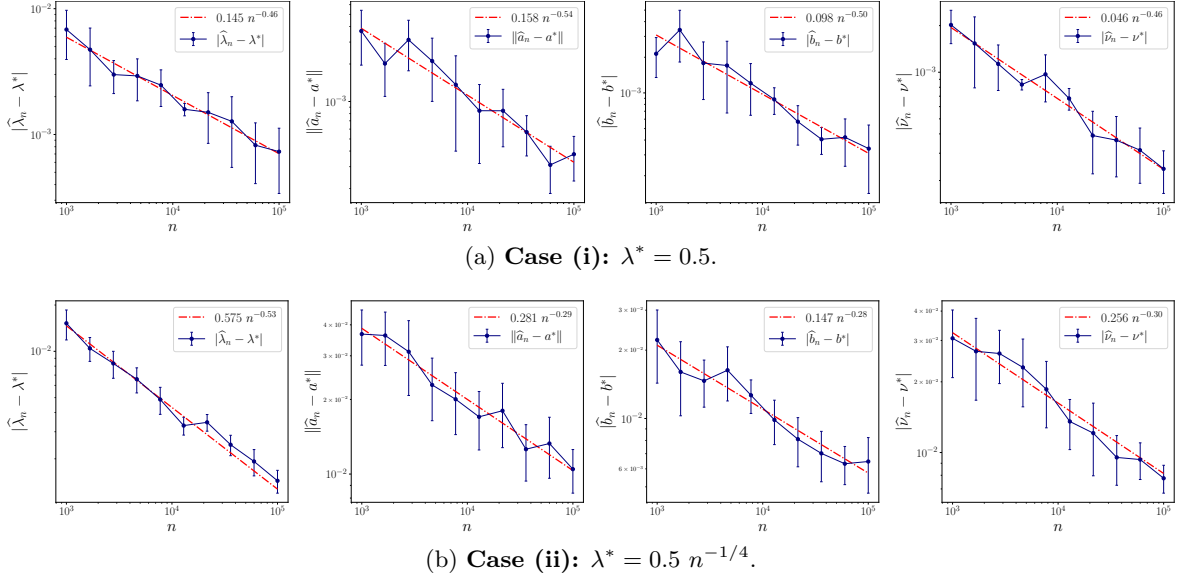
(a) **Case (i):** $\lambda^* = 0.5$.

(b) **Case (ii):** $\lambda^* = 0.5 \; n^{-1/4}$.

Figure 4: (**Theorem 8:** f_0 is Gaussian, $\sigma(z) = 1/(1 + e^{-z})$, $\varphi(z) = z$.) Log-log graphs depicting the empirical convergence rates of the MLE $(\widehat{\lambda}_n, \widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n)$ to the ground-truth values $(\lambda^*, a^*, b^*, \nu^*)$. The blue lines display the parameter estimation errors, while the red dashed dotted lines are the fitted lines, highlighting the empirical MLE convergence rates. Figure 4a and Figure 4b illustrates results for the cases when $\lambda^* = 0.5$ and $\lambda^* = 0.5 \; n^{-1/4}$, respectively.

(a) **Case (i):** $\lambda^* = 0.5, a^* = (1 + n^{-1/8})e_1, b^* = 0, \nu^* = 0.01 + n^{-1/8}$.

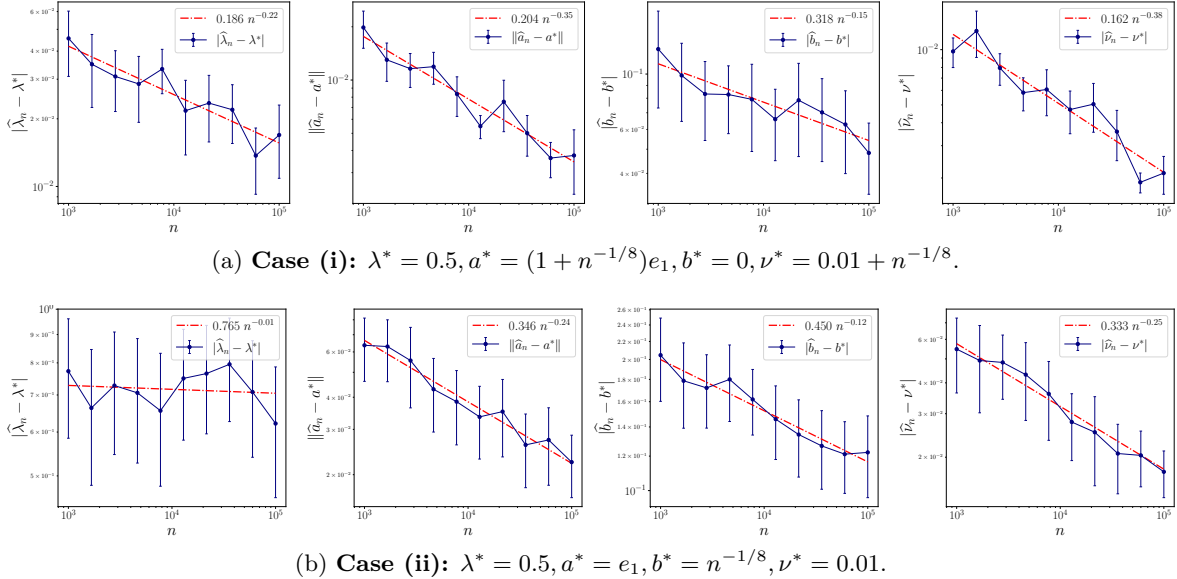(b) **Case (ii):** $\lambda^* = 0.5, a^* = e_1, b^* = n^{-1/8}, \nu^* = 0.01$.

Figure 5: (**Theorem 4:** f_0 is Gaussian, $\sigma(z) = \varphi(z) = z$.) Log-log graphs depicting the empirical convergence rates of the MLE $(\widehat{\lambda}_n, \widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n)$ to the ground-truth values $(\lambda^*, a^*, b^*, \nu^*)$. The blue lines display the parameter estimation errors, while the red dashed dotted lines are the fitted lines, highlighting the empirical MLE convergence rates. Figure 5a and Figure 5b illustrates results for Case (i) and Case (ii), respectively.
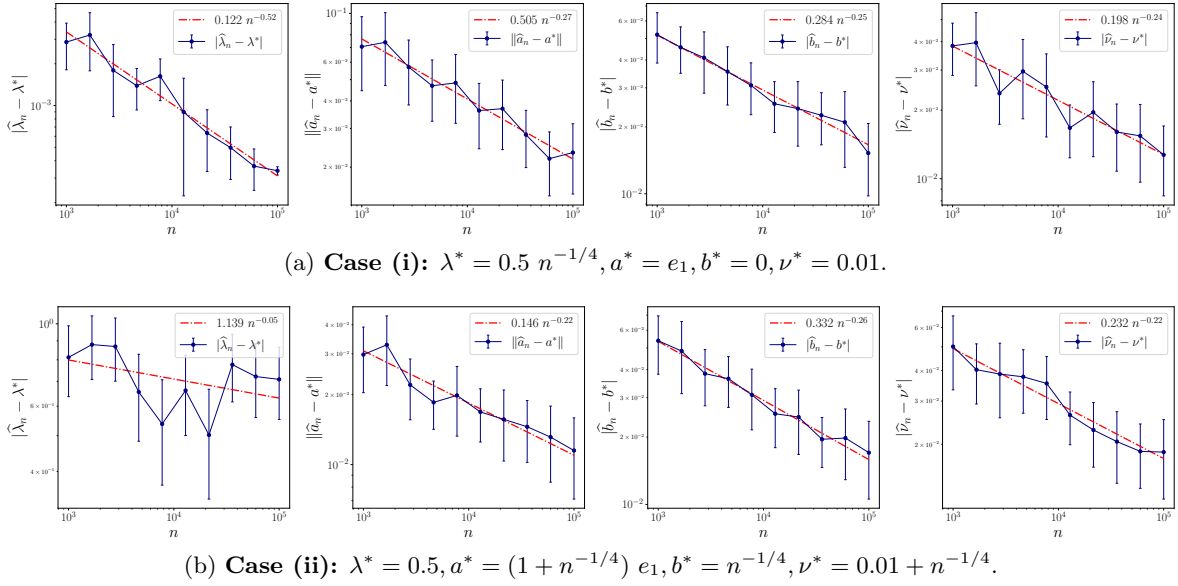
Fanqi Yan[*,◇], Huy Nguyen[*,†], Dung Le[*,†], Pedram Akabrian[‡], Nhat Ho[†]

(a) **Case (i):** $\lambda^* = 0.5 \ n^{-1/4}, a^* = e_1, b^* = 0, \nu^* = 0.01$.

(b) **Case (ii):** $\lambda^* = 0.5, a^* = (1 + n^{-1/4}) \ e_1, b^* = n^{-1/4}, \nu^* = 0.01 + n^{-1/4}$.

Figure 6: (**Theorem 9:** f_0 is Gaussian, $\sigma(z) = \varphi(z) = 1/(1 + e^{-z})$.) Log-log graphs depicting the empirical convergence rates of the MLE $(\widehat{\lambda}_n, \widehat{a}_n, \widehat{b}_n, \widehat{\nu}_n)$ to the ground-truth values $(\lambda^*, a^*, b^*, \nu^*)$. The blue lines display the parameter estimation errors, while the red dashed dotted lines are the fitted lines, highlighting the empirical MLE convergence rates. Figure 6a and Figure 6b illustrates results for Case (i) and Case (ii), respectively.