On the Optimal Second-Order Convergence Rate of Minimax Estimation Under Weighted MSE

Tianren Peng, Qi Li, Shao-Lun Huang
Data Science and Information Technology Research Center
Tsinghua Shenzhen International Graduate School
Shenzhen, China
Email: ptr22@mails.tsinghua.edu.cn, lqinfdim@163.com, shaolun.huang@sz.tsinghua.edu.cn

***Abstract*—Estimating unknown parameters with side information that the parameters are restricted to a given subspace is often considered in information theory, statistics, and machine learning. In such scenarios, minimax parameter estimation (MPE) is a widely adopted formulation, which aims at designing the estimators that optimize the worst-case estimation performance. In this paper, we study the minimax parameter estimation in the asymptotic regime, where a sequence of i.i.d. data can be observed from the parametric model, and the goal is to investigate the optimal second-order convergence rate of the minimax estimators under the Weighted MSE loss. In particular, we show that the optimal second-order convergence rate can be characterized by the local smoothness of the weighted Fisher information matrix of the parametric model along the row vectors of the Fisher information matrix, which reveals the fundamental connection between the second-order convergence rate and Fisher information matrix. Finally, our results also provide a construction of the second-order optimal minimax estimators, which can be beneficial in statistics and machine learning scenarios.**

I. Introduction

A. Motivation

The *minimax parameter estimation* (MPE) framework, introduced by Wald [1], provides a robust method to manage uncertainty by minimizing the worst-case estimation error for broad applications in communication systems, statistics [2], game theory [3], and control systems [4]. This framework has gained renewed interest in the design of algorithms for machine learning and related tasks [5]–[7].

At its essence, MPE focuses on minimizing the maximum possible estimation error for an unknown parameter within a *restricted space*. For instance, consider a basic example, given a parametric model $P_X(x;\theta)$, where θ is a scalar unknown parameter to be estimated from observed i.i.d. data $x = \{x_1, x_2, \dots, x_n\}$, denoted as $\hat{\theta}(x)$. If θ is constrained to an interval $[\alpha, \beta]$, the estimation problem can be formulated as a minimax optimization task:

$$r_n^*(\alpha,\beta) = \inf_{\hat{\theta}} \sup_{\theta\in[\alpha,\beta]} \mathbb{E}\left[\left(\hat{\theta}(x)-\theta\right)^2\right]. \tag{1}$$

This formulation captures several classical problems such as [8]–[10]. For instance, the bounded normal mean estimation (BNME) problem [11], where θ is the mean of a Gaussian distribution constrained to $[-D, D]$, emerges in signal detection with power constraints in additive white Gaussian noise (AWGN) channels. Similarly, for a Bernoulli distribution with $\theta \in [0, 1]$, the minimax formulation relates to worst-case distribution estimation which has been discussed in [12], [13].

Beyond these classical settings, MPE offers a robust tool for designing estimators in *transfer learning*. Consider a transfer learning scenario where the source and target models are $P_X(x;\boldsymbol{\theta}_S)$ and $P_X(x;\boldsymbol{\theta}_T)$, with parameters $\boldsymbol{\theta}_S, \boldsymbol{\theta}_T \in \mathbb{R}^l$. Training source datasets $\mathcal{S}_S = \{x'_1, \dots, x'_m\}$ and target dataset $\mathcal{S}_T = \{x_1, \dots, x_n\}$ are sampled independently from these models. The objective is to estimate $\boldsymbol{\theta}_T$, denoted as $\hat{\boldsymbol{\theta}}_T$, using all available samples. We assume that the source task $\mathcal{T}_S$ is well understood, with $\boldsymbol{\theta}_S$ can be accurately estimated. Further, the source and target tasks $\mathcal{T}_S$ and $\mathcal{T}_T$ are assumed to be sufficiently similar, expressed as $\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_S\| \leq D$, where $\|\cdot\|$ is the Euclidean norm, and D quantifies the similarity. Under this framework, an MPE formulation for transfer learning becomes:

$$\hat{\boldsymbol{\theta}}_T = \mathop{\arg\min}_{\hat{\boldsymbol{\theta}}:\mathcal{X}^n\mapsto\mathbb{R}^l} \max_{\|\boldsymbol{\theta}_T-\boldsymbol{\theta}_S\|\leq D} \mathbb{E}\left[\left\|\hat{\boldsymbol{\theta}}(\mathbf{X})-\boldsymbol{\theta}_T\right\|^2\right]. \tag{2}$$

Solving problem (2) leads to robust transfer learning algorithms that adapt to differences between source and target datasets. Despite its theoretical significance, much of the literature has focused on *first-order characterizations* of MPE risks. These studies provide insights into the leading-order behavior of minimax estimators. However, the second-order characterizations of general distribution $P_X(x;\boldsymbol{\theta})$ remain largely *under-explored*, while existing works focus on special distribution cases like Gaussian. This motivates the following research question: *What's the second-order characterization of MPE for general distribution?*

B. Main Contributions

This paper aims to bridge this gap by exploring the *second-order convergence rate* of minimax risk in the asymptotic regime. Specifically, we consider a family of distributions $P_X(x;\boldsymbol{\theta})$ parameterized by the unknown deterministic vector $\boldsymbol{\theta}$, which takes values in a restricted set denoted by $\mathcal{D} \subset \mathbb{R}^l$, where $l \in \mathbb{N}$. This study investigates the estimator based on a sequence of observed i.i.d. data $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ from the unknown distribution $P_X(x;\boldsymbol{\theta})$, which minimizes the maximum MSE within the predefined $\mathcal{D}$. All these analyses are under the weighted MSE metric with arbitrary weight. The risk of MPE problem with weighted MSE metric can be formulated as follows:

$$r^*_{\boldsymbol{W},n}(\mathcal{D}) \\ = \inf_{\hat{\boldsymbol{\theta}}:\mathcal{X}^n\mapsto\mathbb{R}^l} \max_{\boldsymbol{\theta}\in\mathcal{D}} \mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}(\mathbf{X})-\boldsymbol{\theta}\right)^{\mathrm{T}} \boldsymbol{W}(\boldsymbol{\theta})\left(\hat{\boldsymbol{\theta}}(\mathbf{X})-\boldsymbol{\theta}\right)\right], \tag{3}$$

where the weight matrix $\boldsymbol{W}(\boldsymbol{\theta})$ is the positive-definite matrix parameterized by parameter $\boldsymbol{\theta}$.

Specifically, we analyze the asymptotic behavior of the minimax MPE. We begin by investigating the MPE problem under the weighted MSE via residual estimator in the section III-A. Next, in section III-B, we then establish the optimal second-order convergence rate by deriving the lower bound and the upper bound of the risk for weighted MSE.

In summary, our contributions include:

- *Linking Minimax Risk to Minimax differential equations:* We establish that the minimax risk (3) is closely tied to a minimax partial differential equation (PDE) of the bias function of the estimator $\hat{\boldsymbol{\theta}}(\mathbf{X})$.
- *Optimal Second-Order Convergence Rate:* For a parametric model $P_X(x;\boldsymbol{\theta})$ satisfying specific regularity and singularity conditions, we demonstrate that the optimal second-order convergence rate of (3) is $\Theta\left(n^{-\frac{2d+2}{d+2}}\right)$, where d is the degree of the Fisher information $\boldsymbol{J}(\boldsymbol{\theta})$.
- *Second-Order Optimal Minimax Estimator:* By adopting the residual estimator, we construct the second-order optimal minimax estimator (ref. (21)), which can help to gain insight into the design of the estimators.

C. Related Works

The study of minimax estimation is deeply rooted in the domains of statistics and information theory. A significant portion of this research involves determining optimal estimators by identifying the least favorable prior distributions [14]. However, finding these priors is often fraught with theoretical and computational challenges, which restricts the application of this approach to specific cases [15].

A prominent subset of this field concerns the bounded normal mean problem, where the probabilistic model is Gaussian with its mean parameter constrained to an interval [11]. Research in this area has extensively explored minimizing estimation risk while adhering to the constraint. For example, earlier studies have focused on single-dimensional scenarios, such as estimating the mean of a normal distribution constrained by a small interval [16], [17]. In these cases, the least favorable distribution typically concentrates on the interval's endpoints. Furthermore, extensions to multi-dimensional Gaussian distributions have also been rigorously investigated [11].

In addition to Gaussian models, recent works have expanded the analysis to general distributions. For instance, studies have examined binomial distributions within restricted parameter spaces [18] and extended these findings to multinomial distributions [19]. Both basic cases, such as single-variable mean estimation, and more complex cases, including multivariate distributions [20], [21], general mean estimation [22], and estimation over small intervals [23], have been effectively addressed in these works.

Another critical research direction has involved connecting minimax optimization problems to theoretical bounds. Notable advancements have been made in deriving upper and lower bounds for minimax risk [24], [25], often leveraging the Cramer-Rao lower bound [26], [27]. By fitting bias functions to meet these bounds, researchers have obtained suboptimal estimators [28]. In particular, the Cramer-Rao lower bound has proven to be an accurate benchmark for minimax risks under specific conditions, such as Gaussian distributions [29] and binomial distributions with natural parameter constraints [30]. Additionally, investigations into higher-order moments of minimax estimators have further refined these results [31].

Substantial progress has been achieved in some works under single-dimensional settings, e.g. in [32]. It investigates the second-order behavior of the minimax problem for MPE in the interval, and the second-order convergence rate, the differential equation characterization, and the coefficients of some second-order terms are presented. However, significant gaps remain in understanding the second-order characterization of minimax risk for *general distributions*. Moreover, multi-dimensional extensions of these problems continue to present theoretical and practical challenges. In this work, we aim to address these gaps by focusing on the optimal second-order convergence rate of the multi-dimensional MPE under weighted MSE for general distributions in section III.

D. Notations

In this paper, we adopt the uppercase, such as $X, X_i, \mathbf{X}$, to denote the random variables, and the lowercase, such as $x, x_i, \mathbf{x}$, to denote the corresponding realizations. Moreover, we employ the conventional big-O, o, ω and Θ-notations in the asymptotic regime. For example, $r_n = O\left(f(n)\right)$ implies that there exists a constant M and some n_0, such that for all $n > n_0$, $|r_n| \le M f(n)$.

II. Preliminaries and Definitions

In this section, we first give the definition of fundamental concepts for characterizing the distribution, for which we are going to conduct an estimation. Then the useful estimator constructions and lemmas are introduced for discussing convergence rate.

A. The Regularity and Degree

Consider a parametric model $P_X(x;\boldsymbol{\theta})$, where $\boldsymbol{\theta}\in\mathbb{R}^l$. The model is called *regular* if

$$\mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log P_X(X;\boldsymbol{\theta})\right] = \mathbf{0}, \quad \forall\boldsymbol{\theta}. \tag{4}$$

The *Fisher information matrix* $\boldsymbol{J}(\boldsymbol{\theta})$ is defined as:

$$J(\boldsymbol{\theta}) = \mathbb{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log P_X(X;\boldsymbol{\theta})\right)^2\right] \tag{5}$$

For an estimator $\hat{\boldsymbol{\theta}} : \mathcal{X}\mapsto\mathbb{R}^l$, the *bias* of the estimator, denoted by $\boldsymbol{b}(\boldsymbol{\theta})$, is given by:

$$\boldsymbol{b}(\boldsymbol{\theta}) = \mathbb{E}\left[\hat{\boldsymbol{\theta}}(X)\right] - \boldsymbol{\theta}. \tag{6}$$

An estimator is called *unbiased* if $\boldsymbol{b}(\boldsymbol{\theta}) = \mathbf{0}, \forall \boldsymbol{\theta}$. For regular models, the Cramér-Rao lower bound (CRLB) [26] provides the following inequality for any estimator $\hat{\boldsymbol{\theta}}(x)$:

$$\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}(X)-\boldsymbol{\theta}\right)^2\right] \geq \|\boldsymbol{b}(\boldsymbol{\theta})\|^2 + \mathrm{tr}\left(\boldsymbol{M}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta})\boldsymbol{M}^{\mathrm{T}}(\boldsymbol{\theta})\right),$$

where the $\boldsymbol{M}(\boldsymbol{\theta}) \triangleq \boldsymbol{I}+\boldsymbol{B}(\boldsymbol{\theta})$ and $\boldsymbol{B}(\boldsymbol{\theta}) \triangleq \frac{\partial \boldsymbol{b}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{l\times l}$. The $b(\boldsymbol{\theta})$ is differentiable due to the regularity of the model.

We define the *reduced Fisher information* in a given domain $\mathcal{D}$ as:

$$J^* = \left(\max_{\boldsymbol{\theta}\in\mathcal{D}} \mathrm{tr}\left(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta})\right)\right)^{-1}. \tag{7}$$

The set of all $\boldsymbol{\theta}$ achieving this quantity is denoted as:

$$\mathcal{S} = \left\{\boldsymbol{\theta}\in\mathcal{D} : \mathrm{tr}(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta})) = J^{*-1}\right\}. \tag{8}$$

If $\mathcal{S}$ is either a singleton or a simply connected open subset of $\mathcal{D}$, the model is called *singular*.

For singular models, if $\mathcal{S}$ is a singleton, the global minimum of $\boldsymbol{J}(\boldsymbol{\theta})$ is given by:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}\in\mathcal{D}} \mathrm{tr}\left(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta})\right). \tag{9}$$

We also define $\mathcal{V}$ as follows:

$$\begin{aligned}\mathcal{V} \triangleq \{&\boldsymbol{v}\in\mathcal{R}(\boldsymbol{W}^{\frac{1}{2}}(\boldsymbol{\theta}^*)\boldsymbol{J}^{-1}(\boldsymbol{\theta}^*)) | \exists\epsilon>0,\ \boldsymbol{\theta}^*+t\boldsymbol{v}\in\mathcal{D}\\ &\text{for either all } t\in(0,\epsilon) \text{ or all } t\in(-\epsilon,0)\},\end{aligned} \tag{10}$$

where $\mathcal{R}(\boldsymbol{W}^{\frac{1}{2}}(\boldsymbol{\theta}^*)\boldsymbol{J}^{-1}(\boldsymbol{\theta}^*))$ denotes the set consisting of the row vectors of the matrix $\boldsymbol{W}^{\frac{1}{2}}(\boldsymbol{\theta}^*)\boldsymbol{J}^{-1}(\boldsymbol{\theta}^*)$.

Definition 1 (Degree of Fisher Information)**.** *The directional degree d_v of $J(\boldsymbol{\theta})$ along direction $\boldsymbol{v}\in\mathbb{R}^l$ is defined as:*

$$d_{\boldsymbol{v}} \triangleq \arg\min_k \left\{ k \,\middle|\, \frac{\mathrm{d}^k\ \mathrm{tr}(\boldsymbol{W}(\boldsymbol{\theta}^*+t\boldsymbol{v})\boldsymbol{J}^{-1}(\boldsymbol{\theta}^*+t\boldsymbol{v}))}{\mathrm{d}\ t^k} \neq 0 \right\},$$

when $\mathcal{S}=\{\boldsymbol{\theta}^*\}$. The degree $d \triangleq \min_{\boldsymbol{v}\in\mathcal{V}} d_{\boldsymbol{v}}$.

Example. *When estimating the mean parameter $\boldsymbol{\theta}$ of the Gaussian distribution $P_X(x;\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta},\boldsymbol{\Sigma})$ with given variance matrix $\boldsymbol{\Sigma}$, the Fisher information matrix is $\boldsymbol{\Sigma}^{-1}$ and the degree $d=\infty$ if the interior of the domain $\mathcal{D}$ is not empty.*

B. Bias Corrected Estimator and Comparison Theorem

For the regular distribution $P_X(x;\boldsymbol{\theta})$, we define the *bias-corrected estimator* as:

$$\hat{\boldsymbol{\theta}}_{BC}(\mathbf{x}) \triangleq \hat{\boldsymbol{\theta}}_{ML}(\mathbf{x}) - \hat{\boldsymbol{b}}(\hat{\boldsymbol{\theta}}_{ML}(\mathbf{x})), \tag{11}$$

where

$$\hat{\boldsymbol{b}}(\boldsymbol{\theta}) \triangleq \mathbb{E}_{\mathbf{X}\sim P_X(x;\boldsymbol{\theta})}\left[\boldsymbol{R}_1(\mathbf{X})\right], \tag{12}$$

$$\begin{aligned}\boldsymbol{R}_1(\mathbf{x}) \triangleq \frac{1}{n^2}\boldsymbol{J}^{-1}(\boldsymbol{\theta})\Bigg[&\boldsymbol{L}^{(2)}(\mathbf{X})\boldsymbol{J}(\boldsymbol{\theta})\boldsymbol{L}^{(1)}(\mathbf{X})\\ &+\frac{1}{2}\sum_{i,j=1}^{l}\mathbb{E}\left[\boldsymbol{L}_{:,i,j}(X)\right]\\ &\qquad\cdot\mathbb{E}\left[\boldsymbol{L}_{i,:}(X)\right]\boldsymbol{L}^{(1)}(\mathbf{X})\\ &\qquad\cdot\mathbb{E}\left[\boldsymbol{L}_{j,:}(X)\right]\boldsymbol{L}^{(1)}(\mathbf{X})\Bigg],\end{aligned} \tag{13}$$

and where

$$\boldsymbol{L}_{:,i,j}(x) \triangleq \frac{\partial^3}{\partial\boldsymbol{\theta}\partial\theta_i\partial\theta_j}\log P_X(x;\boldsymbol{\theta}), \tag{14}$$

$$\boldsymbol{L}_{i,:}(x) \triangleq \frac{\partial^2}{\partial\theta_i\partial\boldsymbol{\theta}}\log P_X(x;\boldsymbol{\theta}), \tag{15}$$

$$\boldsymbol{L}^{(1)}(\mathbf{x}) \triangleq \sum_{k=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\log P_X(x_k;\boldsymbol{\theta}), \tag{16}$$

$$\boldsymbol{L}^{(2)}(\mathbf{x}) \triangleq \sum_{k=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\log P_X(x_k;\boldsymbol{\theta}). \tag{17}$$

Moreover, the following comparison theorem [33] will be useful when bounding the asymptotic convergence rates of the bias functions in the next section.

Lemma 2. *Suppose that $u(\theta)$ and $v(\theta)$ are differentiable on the interval $[a,b]\subset\mathbb{R}$, and*

$$u^2(\theta)+\frac{2u'(\theta)}{n} < v^2(\theta)+\frac{2v'(\theta)}{n},\quad \forall\theta\in[a,b]. \tag{18}$$

Then,

$$u(a)\leq v(a) \iff u(\theta)\leq v(\theta),\quad \forall\theta\in[a,b]; \tag{19}$$

$$u(b)\geq v(b) \iff u(\theta)\geq v(\theta),\quad \forall\theta\in[a,b]. \tag{20}$$

III. Characterization of MPE Risk under Weighted MSE Metric

In this section, we present the second-order characterization of (3) by connecting the MPE problem with the minimax partial differential equation. Specifically, we establish the *optimal second-order convergence rate* of $r^*_{\boldsymbol{W},n}(\mathcal{D})$. Before diving into the details, we first introduce some key concepts.

A. Residual Estimator of MPE for Weighted MSE Case

In this part, we provide an estimator for MPE under the weighted MSE. Prior to this, we first introduce the concept of the *residual estimator*. Consider the observations $\mathbf{x} = \{x_1,x_2,\ldots,x_n\}$, which are i.i.d. samples from the distribution $P_X(x;\boldsymbol{\theta})$. Then, the estimator can be expressed as:

$$\hat{\boldsymbol{\theta}}(\mathbf{x}) = \hat{\boldsymbol{\theta}}_{BC}(\mathbf{x}) + \boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_{BC}(\mathbf{x})), \tag{21}$$

where $\boldsymbol{\psi}$ is referred to as the *residual estimator*. Moreover, the bias of the estimator in (21) can be calculated as:

$$b(\boldsymbol{\theta}) \triangleq \mathbb{E}\left[\hat{\boldsymbol{\theta}}(\mathbf{X})\right] - \boldsymbol{\theta} = \mathbb{E}\left[\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_{BC}(\mathbf{X}))\right]. \tag{22}$$

Building upon these definitions, the weighted MSE of MPE is formally established in the following lemma. For convenience, we denote $\boldsymbol{G}(\mathbf{x}) \triangleq \frac{1}{n}\boldsymbol{J}^{-1}(\boldsymbol{\theta})\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\log P_X(x_i;\boldsymbol{\theta})$ and $\boldsymbol{B}(\boldsymbol{\theta}) \triangleq \frac{\partial \boldsymbol{b}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \in \mathbb{R}^{l\times l}$.

Lemma 3. *The weighted MSE of estimating $\boldsymbol{\theta}$ based on (21) is given by*

$$\begin{aligned}&\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}(\mathbf{X})-\boldsymbol{\theta}\right)^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{\theta})\left(\hat{\boldsymbol{\theta}}(\mathbf{X})-\boldsymbol{\theta}\right)\right]\\ &=\frac{1}{n}\mathrm{tr}\left(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta})\right)+\boldsymbol{b}^{\mathrm{T}}(\boldsymbol{\theta})\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{b}(\boldsymbol{\theta})+O(n^{-3})\\ &\quad+\frac{2}{n}\mathrm{tr}\left(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta})\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{\theta})\right)+\mathrm{var}_{\boldsymbol{W}}\left(\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_{BC}(\mathbf{x}))\right),\end{aligned} \tag{23}$$

where the variance term $\mathrm{var}_{\boldsymbol{W}}\left(\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_{BC}(\mathbf{x}))\right)$ *is*

$$\mathbb{E}\left[(\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_{BC}(\mathbf{x}))-\boldsymbol{b}(\boldsymbol{\theta}))^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{\theta})(\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_{BC}(\mathbf{x}))-\boldsymbol{b}(\boldsymbol{\theta}))\right].$$

Proof. We have the maximal likelihood (ML) estimator as:

$$\hat{\boldsymbol{\theta}}_{ML}(\mathbf{x}) = \boldsymbol{\theta} + \boldsymbol{G}(\mathbf{x}) + \boldsymbol{R}_1(\mathbf{x}) + \boldsymbol{R}_2(\mathbf{x}),$$

where the random variable $\boldsymbol{R}_2(\mathbf{x})$ satisfies $\|\mathbb{E}\left[\boldsymbol{R}_2(\mathbf{X})\right]\| = O(n^{-2})$ and $\mathbb{E}\left[\boldsymbol{R}_2^{\mathrm{T}}(\mathbf{X})\boldsymbol{R}_2(\mathbf{X})\right] = O(n^{-3})$.

Note that from the regular condition (4), we have

$$\begin{aligned}\mathbb{E}&\left[\left(\hat{\boldsymbol{\theta}}_{BC}(\mathbf{X})-\boldsymbol{\theta}\right)^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{\theta})\left(\hat{\boldsymbol{\theta}}_{BC}(\mathbf{X})-\boldsymbol{\theta}\right)\right]\\ &=\frac{1}{n}\mathrm{tr}(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta}))+O(n^{-2}),\end{aligned}\tag{24}$$

where (24) follows from

$$\mathbb{E}\left[\boldsymbol{G}^{\mathrm{T}}(\mathbf{X})\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{G}(\mathbf{X})\right]=\frac{\mathrm{tr}(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta}))}{n},\tag{25}$$

and

$$\mathbb{E}\left[\left\|R_1(\mathbf{X})-\hat{\boldsymbol{b}}(\boldsymbol{\theta})+R_2(\mathbf{X})\right\|^2\right]=O(n^{-3}).\tag{26}$$

We also have

$$\begin{aligned}&\mathbb{E}\left[\boldsymbol{\psi}^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_{BC}(\mathbf{X}))\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{G}(\mathbf{X})\right]\\ &\quad=\frac{1}{n}\mathrm{tr}\left(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta})\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{\theta})\right).\end{aligned}\tag{27}$$

Then, it follows from (22), (24), (26) and (27) that

$$\begin{aligned}\mathbb{E}&\left[\left(\hat{\boldsymbol{\theta}}(\mathbf{X})-\boldsymbol{\theta}^*\right)^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{\theta})\left(\hat{\boldsymbol{\theta}}(\mathbf{X})-\boldsymbol{\theta}^*\right)\right]\\ &=\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}_{BC}(\mathbf{X})-\boldsymbol{\theta}\right)^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{\theta})\left(\hat{\boldsymbol{\theta}}_{BC}(\mathbf{X})-\boldsymbol{\theta}\right)\right]\\ &\quad+\mathbb{E}\left[\boldsymbol{\psi}^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_{BC}(\mathbf{X}))\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_{BC}(\mathbf{X}))\right]\\ &\quad+2\mathbb{E}\left[\boldsymbol{\psi}^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_{BC}(\mathbf{X}))\boldsymbol{W}(\boldsymbol{\theta})\left(\hat{\boldsymbol{\theta}}_{BC}(\mathbf{X})-\boldsymbol{\theta}\right)\right] \qquad (28)\\ &=\frac{1}{n}\mathrm{tr}\left(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta})\right)+\boldsymbol{b}^{\mathrm{T}}(\boldsymbol{\theta})\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{b}(\boldsymbol{\theta})+O(n^{-3})\\ &\quad+\frac{2}{n}\mathrm{tr}\left(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta})\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{\theta})\right)+\mathrm{var}_{\boldsymbol{W}}\left(\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_{BC}(\mathbf{X}))\right).\end{aligned}$$

□

From Lemma 3, by the non-negativeness of the variance term, the following lower bound holds for $r^*_{\boldsymbol{W},n}(\mathcal{D})$:

$$\begin{aligned}r^*_{\boldsymbol{W},n}(\mathcal{D})\geq&\frac{1}{n}\mathrm{tr}\left(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta})\right)+\boldsymbol{b}^{\mathrm{T}}(\boldsymbol{\theta})\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{b}(\boldsymbol{\theta})\\ &+\frac{2}{n}\mathrm{tr}\left(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta})\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{\theta})\right)\\ \triangleq&\ \tilde{r}^*_{\boldsymbol{W},n}(\mathcal{D}).\end{aligned}\tag{29}$$

B. Main Results

In this part, we first characterize the second-order behavior of the minimax PDE risk $\tilde{r}^*_{\boldsymbol{W},n}(\mathcal{D})$ under weighted MSE. We show that, $\tilde{r}^*_{\boldsymbol{W},n}(\mathcal{D})$ is indeed equal to $r^*_{\boldsymbol{W},n}(\mathcal{D})$ up to the second-order term, which provides the second-order characterization of MPE risk with weighted MSE at (3). Our results also reveal that the convergence rate of $r^*_{\boldsymbol{W},n}(\mathcal{D})$ is optimal in asymptotic regime.

Theorem 4. *Suppose that the model $P_X(x;\boldsymbol{\theta})$ is singular, then*

$$r^*_{\boldsymbol{W},n}(\mathcal{D})=\frac{1}{nJ^*}-\Theta\left(n^{-\frac{2d+2}{d+2}}\right),\tag{30}$$

where J^* and d are as defined in (7) and Definition 1, respectively.

Proof. Please refer to Appendix A for full proof. □

C. Example

Take the MPE problem of estimating the probability parameters of the multinomial distribution as an example.

Example. *Consider the multinomial distribution $P_X(i;\boldsymbol{\theta}) = p_i$, $i = 1,2,3$, where the paramter $\boldsymbol{\theta} = (p_1,p_2)$ is to be estimated, and where $p_3 = 1-p_1-p_2$. We denote the estimator as $\hat{\boldsymbol{\theta}} \triangleq (\hat{p}_1,\hat{p}_2)$. There are two ways to evaluate the estimation loss, firstly the MSE, i.e., $\mathbb{E}\left[(\hat{p}_1-p_1)^2+(\hat{p}_2-p_2)^2\right]$, and secondly,*

$$\mathbb{E}\left[(\hat{p}_1-p_1)^2+(\hat{p}_2-p_2)^2+(\hat{p}_3-p_3)^2\right].\tag{31}$$

Note that (31) can also be written as the weighted MSE

$$\mathbb{E}\left[(\hat{p}_1-p_1,\hat{p}_2-p_2)\boldsymbol{W}(\hat{p}_1-p_1,\hat{p}_2-p_2)^{\mathrm{T}}\right],\tag{32}$$

where the weight matrix $\boldsymbol{W} = \begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}$. The inverse of the Fisher information is $\boldsymbol{J}^{-1}(\boldsymbol{\theta}) = \begin{pmatrix}p_1(1-p_1) & -p_1p_2\\ -p_1p_2 & p_2(1-p_2)\end{pmatrix}$.

We discuss the two cases as follows, where the domain is defined as $\mathcal{D} = \{\boldsymbol{\theta}|p_1+2p_2\leq 1\}$.

(i) For the first type of metric, the global maximum of $\mathrm{tr}\left(\boldsymbol{J}^{-1}(\boldsymbol{\theta})\right)$ in $\mathcal{D}$ is $\boldsymbol{\theta}_1^* = (2/5,3/10)^{\mathrm{T}}$.

The degree of $\boldsymbol{\theta}_1^*$ is $d=1$, and the minimax risk (3) satisfies

$$r^*_{\boldsymbol{I},n}(\mathcal{D})=\frac{9}{20n}-\Theta\left(n^{-\frac{4}{3}}\right).\tag{33}$$

(ii) While for the second type of metric, the global minimum of $\mathrm{tr}\left(\boldsymbol{W}\boldsymbol{J}^{-1}(\boldsymbol{\theta})\right)$ in $\mathcal{D}$ is $\boldsymbol{\theta}_2^* = (1/3,1/3)^{\mathrm{T}}$.

The degree of $\boldsymbol{\theta}_2^*$ is $d=2$, and the minimax risk (3) satisfies

$$r^*_{\boldsymbol{W},n}(\mathcal{D})=\frac{2}{3n}-\Theta\left(n^{-\frac{3}{2}}\right),\tag{34}$$

which coincides to the results in [12] and [13].

Remark 5. *In this example, we demonstrate that when estimating parameters, the second-order convergence rates of the risk $r^*(\mathcal{D})$ can vary depending on the choice of the loss function even for the same model.*

IV. Conclusion

This study examines the second-order characterization of MPE in constrained parameter spaces, focusing on the interplay between estimator bias and minimax risk through partial differential equation analysis. We derived the optimal second-order convergence rates and demonstrated their dependence on Fisher information and the degree of local regularity. Additionally, our results also provide a construction of the second-order optimal minimax estimator, which can benefit researchers in statistics and machine learning.

References

[1] A. Wald, “Statistical decision functions which minimize the maximum risk,” *Annals of Mathematics*, vol. 46, no. 2, pp. 265–280, 1945.

[2] M. Mintz, “On a minimax parameter estimation problem with a compact parameter space,” in *IEEE Conference on Decision and Control*, 1972.

[3] D. Benko, D. I. Coroian, P. D. Dragnev, and R. Orive, “Probability, minimax approximation, and nash-equilibrium. estimating the parameter of a biased coin,” *ETNA - Electronic Transactions on Numerical Analysis*, 2016.

[4] J. S. Gibson and G. Lee, “Minimax parameter estimation for linear systems,” *Proceedings of 1994 American Control Conference - ACC ’94*, vol. 1, pp. 654–655 vol.1, 1994.

[5] A. Shafahi, P. Saadatpanah, C. Zhu, A. Ghiasi, C. Studer, D. W. Jacobs, and T. Goldstein, “Adversarially robust transfer learning,” *ArXiv*, vol. abs/1905.08232, 2019.

[6] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, “Maximum classifier discrepancy for unsupervised domain adaptation,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3723–3732, 2017.

[7] X. Tong, X. Xu, S.-L. Huang, and L. Zheng, “Robust transfer learning based on minimax principle,” *2023 59th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1–7, 2023.

[8] S. Verdu and H. Poor, “On minimax robustness: A general approach and applications,” *IEEE transactions on Information Theory*, vol. 30, no. 2, pp. 328–340, 1984.

[9] J. Jiao, K. Venkat, Y. Han, and T. Weissman, “Minimax estimation of functionals of discrete distributions,” *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2835–2885, 2015.

[10] S. Hayakawa and T. Suzuki, “On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces,” *Neural Networks*, vol. 123, pp. 343–361, 2020.

[11] J. C. Berry, “Minimax estimation of a bounded normal mean vector,” *Journal of Multivariate Analysis*, vol. 35, pp. 130–139, 1990.

[12] S. Kamath, A. Orlitsky, D. Pichapati, and A. T. Suresh, “On learning distributions from their samples,” in *Annual Conference Computational Learning Theory*, 2015.

[13] J. L. Hodges and E. L. Lehmann, “Some problems in minimax point estimation,” *Annals of Mathematical Statistics*, vol. 21, pp. 15–30, 1950.

[14] A. Dytso, H. V. Poor, R. Bustin, and S. Shamai, “On the structure of the least favorable prior distributions,” *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 1081–1085, 2018.

[15] M. N. Ghosh, “Uniform approximation of minimax pint estimates,” *Annals of Mathematical Statistics*, vol. 35, pp. 1031–1047, 1964.

[16] G. Casella and W. E. Strawderman, “Estimating a bounded normal mean,” *Annals of Statistics*, vol. 9, pp. 870–878, 1981.

[17] P. J. Kempthorne, “Dominating inadmissible procedures using compromise decision theory,” *Statistical Decision Theory and Related Topics IV*, vol. 1, pp. 381–396, 1988.

[18] É. Marchand, F. Perron, and R. Gueye, “Minimax estimation of a constrained binomial proportion p when $|p - 1/2|$ is small,” *Sankhyā: The Indian Journal of Statistics*, pp. 526–537, 2005.

[19] I. Olkin and M. Sobel, “Admissible and minimax estimation for the multinomial distribution and for k independent binomial distributions,” *Annals of Statistics*, vol. 7, pp. 284–290, 1979.

[20] S. C. Trybuła, “Some problems of simultaneous minimax estimation,” *Annals of Mathematical Statistics*, vol. 29, pp. 245–253, 1958.

[21] H. S. M. Erol, E. Sula, and L. Zheng, “On semi-supervised estimation of distributions,” *2023 IEEE International Symposium on Information Theory (ISIT)*, pp. 1866–1871, 2023.

[22] A. A. Melkman and Y. Ritov, “Minimax estimation of the mean of a general distribution when the parameter space is restricted,” *Annals of Statistics*, vol. 15, pp. 432–442, 1987.

[23] A. Dasgupta, “Bayes minimax estimation in multiparameter families when the parameter space is restricted to a bounded convex set,” *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, vol. 47, no. 3, pp. 326–332, 1985.

[24] J. Hodges and E. Lehmann, “Some applications of the Cramer-Rao inequality,” in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, vol. 2. University of California Press, 1951, pp. 13–23.

[25] H. M. Hudson, “A natural identity for exponential families with applications in multiparameter estimation,” *Annals of Statistics*, vol. 6, pp. 473–484, 1978.

[26] H. Cramér, *Mathematical methods of statistics*. Princeton university press, 1999, vol. 26.

[27] Y. V. Linnik, “A note on Rao-Cramer and Bhattacharya inequalities,” *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, vol. 32, no. 4, pp. 449–452, 1970.

[28] L. D. Brown and R. H. Farrell, “A lower bound for the risk in estimating the value of a probability density,” *Journal of the American Statistical Association*, vol. 85, no. 412, pp. 1147–1153, 1990.

[29] P. J. Bickel, “Minimax estimation of the mean of a normal distribution when the parameter space is restricted,” *The Annals of Statistics*, vol. 9, no. 6, pp. 1301–1309, 1981.

[30] D. Braess and H. Dette, “The asymptotic minimax risk for the estimation of constrained binomial and multinomial probabilities,” *Technical reports*, 2004.

[31] S. Batalama and D. Kazakos, “On the generalized Cramer-Rao bound for the estimation of the location,” *IEEE Transactions on Signal Processing*, vol. 45, no. 2, pp. 487–492, 1997.

[32] T. Peng, X. Tong, and S.-L. Huang, “Second-order characterization of minimax parameter estimation in restricted parameter space,” in *2024 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2024, pp. 3420–3425.

[33] A. McNabb, “Comparison theorems for differential equations,” *Journal of Mathematical Analysis and Applications*, vol. 119, no. 1, pp. 417–428, 1986.

[34] H.-K. Tang and C.-T. See, “Variance inequalities using first derivatives,” *Statistics & Probability Letters*, vol. 79, no. 9, pp. 1277–1281, 2009.

[35] R. M. Dudley, *Real analysis and probability*. CRC Press, 2018.

[36] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[37] V. A. Zorich, *Differential Calculus*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 171–329.

Appendix A
Full proof of Main Results

To prove theorem 4, we first introduce some necessary lemma. We first prove the low bound of equation (30) as:

$$r^*_{\boldsymbol{W},n}(\mathcal{D}) \geq \frac{1}{nJ^*} - c_1 n^{-\frac{2d+2}{d+2}}, \tag{35}$$

where the c_1 is an constant. We have the following theorem:

Theorem 6. *Suppose that the model $P_X(x;\boldsymbol{\theta})$ is singular, then there exists an constant c_1 such that*

$$r^*_{\boldsymbol{W},n}(\mathcal{D}) \geq \frac{1}{nJ^*} - c_1 n^{-\frac{2d+2}{d+2}}. \tag{36}$$

where J^* and d are as defined in (7) and Definition 1, respectively.

Proof. Suppose that for sufficiently large n, $\tilde{r}^*_{\boldsymbol{W},n}(\mathcal{D})$ as defined in (29) satisfies

$$\tilde{r}^*_{\boldsymbol{W},n}(\mathcal{D}) = \frac{1}{nJ^*} - C(n), \text{ where } C(n) = \omega\left(n^{-\frac{2d+2}{d+2}}\right) > 0.$$

Then, there exists $\boldsymbol{b}(\boldsymbol{\theta})$ such that $\forall \boldsymbol{\theta} \in \mathcal{D}$,

$$\begin{aligned}&\boldsymbol{b}^{\mathrm{T}}(\boldsymbol{\theta})\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{b}(\boldsymbol{\theta}) + \frac{1}{n}\mathrm{tr}\left(\boldsymbol{W}(\theta)\boldsymbol{J}^{-1}(\boldsymbol{\theta})\right)\\ &+ \frac{2}{n}\mathrm{tr}\left(\boldsymbol{W}(\theta)\boldsymbol{J}^{-1}(\boldsymbol{\theta})\boldsymbol{B}^{\mathrm{T}}(\theta)\right) \leq \frac{1}{nJ^*} - C(n).\end{aligned} \tag{37}$$

By the definition of d, there exists $i \in \{1, \ldots, l\}$, such that for $\boldsymbol{\theta}$ on the segment $\boldsymbol{\Lambda} \triangleq \{\boldsymbol{\theta}^* + t\boldsymbol{v}_i \mid t \in [0, n^{-\frac{1}{d+2}}]\}$, where $\boldsymbol{v}_i$ is the i-th row vector of $\boldsymbol{W}^{\frac{1}{2}}(\boldsymbol{\theta}^*)\boldsymbol{J}^{-1}(\boldsymbol{\theta}^*)$, we have that the i-th entry of $\boldsymbol{b}(\boldsymbol{\theta})$ denoted by $b_i(\boldsymbol{\theta})$ is bounded and

$$\mathrm{tr}\left(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\theta)\right) = \frac{1}{J^*} + O\left(n^{-\frac{d}{d+2}}\right). \tag{38}$$

From (37) and (38), we let

$$\begin{aligned}\tilde{C}(n) &\triangleq \min_{\theta\in\Lambda}\frac{1}{n}\left(\mathrm{tr}\left(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta})\right) - \frac{1}{J^*}\right) + \frac{1}{2}C(n)\\ &= \omega\left(n^{-\frac{2d+2}{d+2}}\right) > 0,\end{aligned} \tag{39}$$

and then

$$b_i^2(\boldsymbol{\theta}) + \frac{2}{n}\boldsymbol{v}_i^{\mathrm{T}}\frac{\partial b_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} < -\tilde{C}(n). \tag{40}$$

Let us define

$$f(t) \triangleq -\sqrt{\tilde{C}(n)}\tan\left(\frac{n\sqrt{\tilde{C}(n)}}{2}t - U(n)\right), \tag{41}$$

where $U(n) \triangleq \tan^{-1}\frac{b_i(\boldsymbol{\theta}^*)}{\sqrt{\tilde{C}(n)}}$. We have $f(0) = b_i(\boldsymbol{\theta}^*)$ and

$$\begin{aligned}&f^2(t) + \frac{2}{n}f'(t) \geq -\tilde{C}(n)\\ &> b_i^2(\boldsymbol{\theta}^* + \boldsymbol{v}_i t) + \frac{2}{n}\frac{\mathrm{d}\, b_i(\boldsymbol{\theta}^* + \boldsymbol{v}_i t)}{\mathrm{d}\, t}.\end{aligned} \tag{42}$$

Then, it follows from Lemma 2 that

$$b_i(\boldsymbol{\theta}^* + \boldsymbol{v}_i t) \leq f(t),\ \forall t \in \left[0, n^{-\frac{1}{d+2}}\right]. \tag{43}$$

Note that $\lim_{t\to t_1} f(t) = -\infty$, where

$$t_1 \triangleq \frac{2U(n)}{nJ^*\sqrt{\tilde{C}(n)}} = o\left(n^{-\frac{1}{d+2}}\right) < n^{-\frac{1}{d+2}},$$

which breaks the existence of $\boldsymbol{b}(\boldsymbol{\theta})$. Thus, from (29), we have

$$r^*_{\boldsymbol{W},n}(\mathcal{D}) \geq \tilde{r}^*_{\boldsymbol{W},n}(\mathcal{D}) \geq \frac{1}{nJ^*} - \Theta\left(n^{-\frac{2d+2}{d+2}}\right). \tag{44}$$

□

Since we have proved the lower bound of the equation (35), we now turn to investigate the upper bound, which means we would like to show

$$r^*_{\boldsymbol{W},n}(\mathcal{D}) \leq \frac{1}{nJ^*} - c_2 n^{-\frac{2d+2}{d+2}}, \tag{45}$$

where J^* and d are as defined in (7) and Definition 1, respectively. c_2 is an constant. To achieve this, we first define the following estimator:

Definition 7. *The estimator $\tilde{\boldsymbol{b}}(\boldsymbol{\theta})$ with i-th entry is defined as*

$$\tilde{b}_i(\boldsymbol{\theta}) \triangleq \begin{cases}\sqrt{mn^{-1}\left|\boldsymbol{v}_i^{\mathrm{T}}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^*\right)\right|^{d_{\boldsymbol{v}_i}} - \tilde{K}_i n^{-\frac{2d_{\boldsymbol{v}_i}+2}{d_{\boldsymbol{v}_i}+2}}},\\ \qquad\qquad \text{if } \boldsymbol{v}_i^{\mathrm{T}}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^*\right) < \underline{t}\|\boldsymbol{v}_i\|^2,\\ b_{0,i}(\boldsymbol{\theta}),\\ \qquad\qquad \text{if } \boldsymbol{v}_i^{\mathrm{T}}\left(\theta - \theta^*\right) \in \left[\underline{t}\|\boldsymbol{v}_i\|^2, \bar{t}\|\boldsymbol{v}_i\|^2\right],\\ -\sqrt{mn^{-1}\left|\boldsymbol{v}_i^{\mathrm{T}}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^*\right)\right|^{d_{\boldsymbol{v}_i}} - \tilde{K}_i n^{-\frac{2d_{\boldsymbol{v}_i}+2}{d_{\boldsymbol{v}_i}+2}}},\\ \qquad\qquad \text{if } \boldsymbol{v}_i^{\mathrm{T}}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^*\right) > \bar{t}\|\boldsymbol{v}_i\|^2,\end{cases} \tag{46}$$

where positive numbers K_i and $\tilde{K}_i$ are the solutions of

$$\begin{aligned}\frac{2K_i}{J^*}\tan K_i\gamma_i &= \sqrt{m\gamma^{d_{\boldsymbol{v}_i}} - \tilde{K}_i},\\ \frac{2K_i^2}{J^*}\cos^{-2}K_i\gamma_i &= \frac{md\gamma^{d_{\boldsymbol{v}_i}-1}}{2\sqrt{m\gamma^{d_{\boldsymbol{v}_i}} - \tilde{K}_i}}.\end{aligned} \tag{47}$$

and $b_{0,i}(\boldsymbol{\theta})$ is

$$b_{0,i}(\boldsymbol{\theta}) \triangleq -\frac{2K_i}{J^*}n^{-\frac{d_{\boldsymbol{v}_i}+1}{d_{\boldsymbol{v}_i}+2}}\tan\left(K_i n^{\frac{1}{d_{\boldsymbol{v}_i}+2}}\frac{\boldsymbol{v}_i^{\mathrm{T}}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^*\right)}{\|\boldsymbol{v}_i\|^2}\right).$$

For properly chosen $\gamma_i > 0$, we let $\underline{t}_i \triangleq -\gamma n^{-\frac{1}{d_{\boldsymbol{v}_i}+2}}$ and $\bar{t}_i \triangleq \gamma n^{-\frac{1}{d_{\boldsymbol{v}_i}+2}}$. Fix $i \in \{1, 2, \ldots, l\}$. Note that when $d_{\boldsymbol{v}_i} < \infty$, there exists $m > 0$ and $M > m$, such that $\forall t \in \{t \mid \boldsymbol{\theta}^* + \boldsymbol{v}_i t \in \mathcal{D}\}$, we have

$$-M\left|t\right|^{d_{\boldsymbol{v}_i}} \leq \mathrm{tr}(\boldsymbol{W}(\boldsymbol{\theta}^* + \boldsymbol{v}_i t)\boldsymbol{J}^{-1}(\boldsymbol{\theta}^* + \boldsymbol{v}_i t)) - \frac{1}{J^*} \leq -m\left|t\right|^{d_{\boldsymbol{v}_i}},$$

which can be easily proved by bounding

$$\left(\mathrm{tr}(\boldsymbol{W}(\boldsymbol{\theta} + \boldsymbol{v}_i t)\boldsymbol{J}^{-1}(\boldsymbol{\theta} + \boldsymbol{v}_i t)) - \frac{1}{J^*}\right) \Big/ \left|t\right|^{d_{\boldsymbol{v}_i}}.$$

Especially, we may choose

$$\begin{aligned}\gamma_i &= \left(\frac{\alpha_i^3\sin\frac{\alpha_i}{2}}{md_{\boldsymbol{v}_i}J^{*3}\cos^3\frac{\alpha_i}{2}}\right)^{\frac{1}{d+2}},\ K_i = \frac{\alpha_i}{2\gamma_i},\\ \tilde{K}_i &= \frac{\alpha_i^3\sin\frac{\alpha_i}{2}}{d_{\boldsymbol{v}_i}J^{*3}\cos^3\frac{\alpha_i}{2}} - \frac{\alpha_i^2}{J^{*2}}\tan^2\frac{\alpha_i}{2},\end{aligned} \tag{48}$$

where $\alpha_i \in (0, \pi)$ satisfies $\frac{\sin\alpha_i}{\alpha_i} < \frac{2}{d_{\boldsymbol{v}_i}J^*}$.

The estimator defined in Definition 7 has the following property.

Theorem 8. *Note that (47) establishes the first-order smoothness of $\tilde{b}$, and we can easily derive from (46) that $\forall \theta \in \mathcal{D}$,*

$$\tilde{b}_i(\boldsymbol{\theta}) = O\left(n^{-\frac{1}{2}}\right), \tag{49}$$

$$\left\|\nabla \tilde{b}_i(\boldsymbol{\theta})\right\| = O\left(n^{\max\{-\frac{1}{2}, -\frac{d}{d+2}\}}\right), \tag{50}$$

$$\left\|\frac{\partial^2 \tilde{b}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right\| = O\left(n^{\max\{-\frac{1}{2}, -\frac{d-1}{d+2}\}}\right), \tag{51}$$

$$\left\|\frac{\partial^3 \tilde{b}_i(\boldsymbol{\theta}^* + \boldsymbol{v}_i t)}{\partial t}\right\| = O\left(n^{\max\{-\frac{1}{2}, -\frac{d-2}{d+2}\}}\right). \tag{52}$$

Proof. We first have (49) by considering

$$mn^{-1}\left|\boldsymbol{v}_i^{\mathrm{T}}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^*\right)\right|^{d_{\boldsymbol{v}_i}} = O(n^{-1}) \tag{53}$$

holds for all $\boldsymbol{\theta}$, and

$$\tan\left(K_i n^{\frac{1}{d_{\boldsymbol{v}_i}+2}} \frac{\boldsymbol{v}_i^{\mathrm{T}}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^*\right)}{\|\boldsymbol{v}_i\|^2}\right) = O(1) \tag{54}$$

for $\boldsymbol{\theta}$ satisfying $\boldsymbol{v}_i^{\mathrm{T}}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^*\right) \in \left[\underline{t}\|\boldsymbol{v}_i\|^2, \bar{t}\|\boldsymbol{v}_i\|^2\right]$.

Then we have (50) by considering

$$\nabla \tilde{b}_i(\boldsymbol{\theta}) \triangleq \begin{cases} \dfrac{mn^{-1}\left|\boldsymbol{v}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right|^{d_{\boldsymbol{v}_i}-1}\boldsymbol{v}_i}{\sqrt{mn^{-1}\left|\boldsymbol{v}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right|^{d_{\boldsymbol{v}_i}} - \tilde{K}_i n^{-\frac{2d_{\boldsymbol{v}_i}+2}{d_{\boldsymbol{v}_i}+2}}}}, \\ \text{if } \boldsymbol{v}_i^{\mathrm{T}}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^*\right) \in \left(-\infty, \underline{t}\|\boldsymbol{v}_i\|^2\right) \cup \left(\bar{t}\|\boldsymbol{v}_i\|^2, \infty\right), \\ -\dfrac{2K_i^2 \boldsymbol{v}_i}{J^*\|\boldsymbol{v}_i\|^2} n^{-\frac{d_{\boldsymbol{v}_i}}{d_{\boldsymbol{v}_i}+2}} \cos^{-2}\left(K_i n^{\frac{1}{d_{\boldsymbol{v}_i}+2}} \dfrac{\boldsymbol{v}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)}{\|\boldsymbol{v}_i\|^2}\right), \\ \qquad \text{if } \boldsymbol{v}_i^{\mathrm{T}}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^*\right) \in \left[\underline{t}\|\boldsymbol{v}_i\|^2, \bar{t}\|\boldsymbol{v}_i\|^2\right], \end{cases} \tag{55}$$

where $mn^{-1}\left|\boldsymbol{v}_i^{\mathrm{T}}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^*\right)\right|^{d_{\boldsymbol{v}_i}} = O(n^{-1})$ holds for all $\boldsymbol{\theta}$, and $\cos^{-2}\left(K_i n^{\frac{1}{d_{\boldsymbol{v}_i}+2}} \frac{\boldsymbol{v}_i^{\mathrm{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)}{\|\boldsymbol{v}_i\|^2}\right) = O(1)$ for $\boldsymbol{\theta}$ satisfying $\boldsymbol{v}_i^{\mathrm{T}}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^*\right) \in \left[\underline{t}\|\boldsymbol{v}_i\|^2, \bar{t}\|\boldsymbol{v}_i\|^2\right]$. □

Moreover, we have:

Corollary 9. *$\tilde{\boldsymbol{b}}(\boldsymbol{\theta})$ satisfies $\forall \boldsymbol{\theta}$ such that $\underline{t}_i\|\boldsymbol{v}_i\|^2 < \boldsymbol{v}_i^{\mathrm{T}}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) < \bar{t}_i\|\boldsymbol{v}_i\|^2$, $i = 1, \dots, l$,*

$$\tilde{\boldsymbol{b}}^{\mathrm{T}}(\boldsymbol{\theta})\boldsymbol{W}(\boldsymbol{\theta}^*)\tilde{\boldsymbol{b}}(\boldsymbol{\theta}) + \frac{1}{n}\operatorname{tr}(\boldsymbol{W}(\boldsymbol{\theta}^*)\boldsymbol{J}^{-1}(\boldsymbol{\theta}^*))$$
$$+\frac{2}{n}\operatorname{tr}(\boldsymbol{W}(\boldsymbol{\theta}^*)\boldsymbol{J}^{-1}(\boldsymbol{\theta}^*)\tilde{\boldsymbol{B}}(\boldsymbol{\theta})) = \frac{1}{nJ^*} - \frac{4K^2}{J^{*2}}n^{-\frac{2d+2}{d+2}}, \tag{56}$$

where $K \triangleq \min_{i\in\{1,\dots,l\}} K_i$. We also define $\tilde{K} \triangleq \min_{i\in\{1,\dots,l\}} \tilde{K}_i$ It follows from (48), (50), and (56) that

$$\max_{\boldsymbol{\theta}\in\mathcal{D}} \tilde{\boldsymbol{b}}^{\mathrm{T}}(\boldsymbol{\theta})\boldsymbol{W}(\boldsymbol{\theta})\tilde{\boldsymbol{b}}(\boldsymbol{\theta}) + \frac{1}{n}\operatorname{tr}(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta}))$$
$$+\frac{2}{n}\operatorname{tr}\left(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta})\tilde{\boldsymbol{B}}(\boldsymbol{\theta})\right)$$
$$\leq \frac{1}{nJ^*} - \frac{\min\left\{\frac{4K^2}{J^{*2}}, \tilde{K}\right\}}{n^{\frac{2d+2}{d+2}}}. \tag{57}$$

□

We have the following definition of the residual estimator:

Definition 10. *we define $\psi(\cdot)$ (cf. (21)) with i-th entry $\psi_i(\cdot)$ as*

$$\psi_i(\boldsymbol{\theta}) \triangleq \begin{cases} \tilde{b}_i(\underline{\boldsymbol{\theta}}) + \nabla\tilde{b}_i(\underline{\boldsymbol{\theta}})^{\mathrm{T}}(\boldsymbol{\theta} - \underline{\boldsymbol{\theta}}), & \boldsymbol{v}_i^{\mathrm{T}}(\boldsymbol{\theta} - \underline{\boldsymbol{\theta}}) < 0, \\ \tilde{b}_i(\boldsymbol{\theta}), & \boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{\theta} \in [\boldsymbol{v}_i^{\mathrm{T}}\underline{\boldsymbol{\theta}}, \boldsymbol{v}_i^{\mathrm{T}}\bar{\boldsymbol{\theta}}], \\ \tilde{b}_i(\bar{\boldsymbol{\theta}}) + \nabla\tilde{b}_i(\bar{\boldsymbol{\theta}})^{\mathrm{T}}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}), & \boldsymbol{v}_i^{\mathrm{T}}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) > 0, \end{cases} \tag{58}$$

where $\underline{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\theta}}$ is chosen such that $\boldsymbol{v}_i^{\mathrm{T}}\underline{\boldsymbol{\theta}} < \boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{\theta} < \boldsymbol{v}_i^{\mathrm{T}}\bar{\boldsymbol{\theta}}$, $\forall \boldsymbol{\theta} \in \mathcal{D}$, and where $\tilde{\boldsymbol{b}}_i(\boldsymbol{\theta})$ is defined in (46).

Corollary 11. *Note that from Lemma 3, we have $\forall \boldsymbol{\theta} \in \mathcal{D}$,*

$$\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\theta}\right)^{\mathrm{T}} \boldsymbol{W}(\boldsymbol{\theta}) \left(\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\theta}\right)\right]$$
$$= \boldsymbol{b}^{\mathrm{T}}(\boldsymbol{\theta})\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{b}(\boldsymbol{\theta}) + \frac{1}{n}\operatorname{tr}(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta}))$$
$$+ \frac{2}{n}\operatorname{tr}\left(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta})\frac{\partial \boldsymbol{b}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) + \operatorname{var}_{\boldsymbol{W}}\left(\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x}))\right), \tag{59}$$

where $\boldsymbol{b}(\boldsymbol{\theta}) \triangleq \mathbb{E}\left[\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x}))\right]$. From (50), we have

$$\operatorname{var}_{\boldsymbol{W}}(\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x}))) \leq \left(\max_{\boldsymbol{\theta}\in\mathcal{D}}\left\|\frac{\partial \tilde{\boldsymbol{b}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\|\right)^2 \operatorname{var}_{\boldsymbol{W}}(\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x}))$$
$$= o\left(n^{-\frac{2d+2}{d+2}}\right), \tag{60}$$

where the inequality is proved in [34], and we have used the fact $\operatorname{var}(\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x})) = \frac{1}{n}\operatorname{tr}(\boldsymbol{J}^{-1}(\theta)) + o(n^{-\frac{2d+2}{d+2}})$.

Theorem 12. *In addition, the next step is to prove $\forall \boldsymbol{\theta} \in \mathcal{D}$,*

$$|\boldsymbol{b}^{\mathrm{T}}(\boldsymbol{\theta})\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{b}(\theta) - \tilde{\boldsymbol{b}}^{\mathrm{T}}(\theta)\boldsymbol{W}(\theta)\tilde{\boldsymbol{b}}(\theta)| = o\left(n^{-\frac{2d+2}{d+2}}\right), \tag{61}$$

$$\left\|\boldsymbol{B}(\boldsymbol{\theta}) - \tilde{\boldsymbol{B}}(\boldsymbol{\theta})\right\| = o\left(n^{-\frac{d}{d+2}}\right). \tag{62}$$

Proof. To prove (61), we have used $\forall \theta \in \mathcal{D}$,

$$|\boldsymbol{b}^{\mathrm{T}}(\boldsymbol{\theta})\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{b}(\boldsymbol{\theta}) - \tilde{\boldsymbol{b}}^{\mathrm{T}}(\theta)\boldsymbol{W}(\boldsymbol{\theta})\tilde{\boldsymbol{b}}(\theta)|$$
$$\leq \|\boldsymbol{b}(\boldsymbol{\theta}\| \cdot \|\boldsymbol{W}(\boldsymbol{\theta})\| \cdot \|\boldsymbol{b}(\boldsymbol{\theta}) - \tilde{\boldsymbol{b}}(\boldsymbol{\theta})\|$$
$$+ \|\boldsymbol{b}(\boldsymbol{\theta}) - \tilde{\boldsymbol{b}}(\boldsymbol{\theta})\| \cdot \|\boldsymbol{W}(\boldsymbol{\theta})\| \cdot \|\tilde{\boldsymbol{b}}(\boldsymbol{\theta})\| \tag{63}$$

where

$$\|\boldsymbol{b}(\boldsymbol{\theta}) - \tilde{\boldsymbol{b}}(\boldsymbol{\theta}\|$$
$$= \left\|\mathbb{E}\left[\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x})) - \tilde{\boldsymbol{b}}(\boldsymbol{\theta})\right]\right\|$$
$$\leq \left\|\mathbb{E}\left[\left(\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x})) - \tilde{\boldsymbol{b}}(\boldsymbol{\theta})\right) \cdot \mathbb{1}\left\{\mathbf{X} \in S\right\}\right]\right\| \tag{64}$$
$$+ \left\|\mathbb{E}\left[\left(\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x})) - \tilde{\boldsymbol{b}}(\boldsymbol{\theta})\right) \cdot \mathbb{1}\left\{\mathbf{X} \notin S\right\}\right]\right\|, \tag{65}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function [35, Chapter 1], and where $S \triangleq \{\mathbf{x} \colon \|\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x}) - \boldsymbol{\theta}\| \leq \epsilon\}$ for sufficiently small $\epsilon > 0$. Note that from Sanov's theorem [36], the probability $\mathbb{E}\left[\mathbb{1}\left\{\mathbf{X} \notin S\right\}\right]$ is exponentially small. Thus, we have

$$\mathbb{E}\left[\left(\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x})) - \tilde{\boldsymbol{b}}(\boldsymbol{\theta})\right) \mathbb{1}\left\{\mathbf{X} \notin S\right\}\right] = o\left(n^{-\frac{3d+2}{2d+4}}\right). \tag{66}$$

Moreover, we obtain from Lagrange mean value theorem [37]

$$\|\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x})) - \tilde{\boldsymbol{b}}(\boldsymbol{\theta}) - (\nabla\boldsymbol{\psi}(\boldsymbol{\theta}))^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x}) - \boldsymbol{\theta})\|$$
$$\leq \frac{1}{2}\left(\max_{\boldsymbol{\theta}\in\mathcal{D}}\left\|\frac{\partial^2 \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right\|\right)\|\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x}) - \boldsymbol{\theta}\|^2. \tag{67}$$

We have also used the fact that for any given function $f(\cdot)$ and $g(\cdot)$

$$|\mathbb{E}\left[f(\mathbf{X})g(\mathbf{X})\right]| \leq \max_{\mathbf{x}} |f(\mathbf{x})| \cdot \mathbb{E}\left[|g(\mathbf{X})|\right]. \tag{68}$$

Then, from (51), $\frac{\partial^2 \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = \frac{\partial^2 \tilde{\boldsymbol{b}}(\theta)}{\partial \boldsymbol{\theta}^2}$ in $\mathcal{D}$, and the fact that $\mathbb{E}[\|\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x}) - \boldsymbol{\theta}\|^2] = \frac{1}{n}\operatorname{tr}(\boldsymbol{J}^{-1}(\boldsymbol{\theta})) + o(n^{-\frac{2d+2}{d+2}})$, we have

$$\mathbb{E}\left[\left(\max_{\boldsymbol{\theta}\in\mathcal{D}} \left\|\frac{\partial^2 \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right\|\right) \|\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x}) - \theta\|^2 \cdot \mathbb{1}\left\{\mathbf{X} \in S\right\}\right] = o\left(n^{-\frac{3d+2}{2d+4}}\right). \tag{69}$$

From $\mathbb{E}\left[(\nabla \boldsymbol{\psi}(\boldsymbol{\theta}))^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x}) - \boldsymbol{\theta})\right] = o(n^{-\frac{2d+2}{d+2}})$, we have

$$\begin{aligned}&\mathbb{E}\left[(\nabla \boldsymbol{\psi}(\boldsymbol{\theta}))^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x}) - \boldsymbol{\theta}) \cdot \mathbb{1}\left\{\mathbf{X} \in S\right\}\right] \\ &= -\mathbb{E}\left[(\nabla \boldsymbol{\psi}(\boldsymbol{\theta}))^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_{BC}(\boldsymbol{x}) - \boldsymbol{\theta}) \cdot \mathbb{1}\left\{\mathbf{X} \notin S\right\}\right] = o\left(n^{-\frac{3d+2}{2d+4}}\right).\end{aligned} \tag{70}$$

From (66), (67), (69), and (70), we can conclude that

$$\left\|\boldsymbol{b}(\theta) - \tilde{\boldsymbol{b}}(\boldsymbol{\theta})\right\| = o\left(n^{-\frac{3d+2}{2d+4}}\right), \tag{71}$$

which leads to (61) by combining (63) and $\|\tilde{\boldsymbol{b}}(\boldsymbol{\theta})\| = O\left(n^{-\frac{1}{2}}\right)$ (ref. (49)). Similar to the proof of (61), we can also prove (62) from the order of $\frac{\partial^3 \tilde{\boldsymbol{b}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^3}$ in (52). □

Theorem 13. *Suppose that the model $P_X(x;\boldsymbol{\theta})$ is singular, then there exists an constant c_2, such that*

$$r^*_{\boldsymbol{W},n}(\mathcal{D}) \leq \frac{1}{nJ^*} - c_2 n^{-\frac{2d+2}{d+2}}, \tag{72}$$

where J^* and d are as defined in (7) and Definition 1, respectively.

Proof. Once (61) and (62) are proved, we have $\forall \boldsymbol{\theta} \in \mathcal{D}$,

$$\begin{aligned}&\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\theta}\right)^{\mathrm{T}} \boldsymbol{W}(\boldsymbol{\theta}) \left(\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\theta}\right)\right] \\ &= \tilde{\boldsymbol{b}}^{\mathrm{T}}(\boldsymbol{\theta})\boldsymbol{W}(\boldsymbol{\theta})\tilde{\boldsymbol{b}}(\boldsymbol{\theta}) + \frac{1}{n}\operatorname{tr}(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta})) \\ &\quad + \frac{2}{n}\operatorname{tr}\left(\boldsymbol{W}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta})\tilde{\boldsymbol{B}}(\boldsymbol{\theta})\right) + o\left(n^{-\frac{2d+2}{d+2}}\right) \qquad (73) \\ &\leq \frac{1}{nJ^*} - \frac{\min\left\{\frac{4K^2}{J^{*2}}, \tilde{K}\right\}}{n^{\frac{2d+2}{d+2}}} + o\left(n^{-\frac{2d+2}{d+2}}\right), \qquad (74)\end{aligned}$$

where to obtain (73) we have used (59) and (60), and where to obtain (74) we have used (57). □

Now, we have proved the upper and lower bound in theorem 13 and theorem 6 respectively, we now can conclude:

Theorem 4. Suppose that the model $P_X(x;\boldsymbol{\theta})$ is singular, then

$$r^*_{\boldsymbol{W},n}(\mathcal{D}) = \frac{1}{nJ^*} - \Theta\left(n^{-\frac{2d+2}{d+2}}\right), \tag{75}$$

where J^* and d are as defined in (7) and Definition 1, respectively.

Proof. from the theorem 13 and theorem 6, we have

$$r^*_{\boldsymbol{W},n}(\mathcal{D}) = \frac{1}{nJ^*} - \Theta\left(n^{-\frac{2d+2}{d+2}}\right), \tag{76}$$

□