Dissecting the Impact of Model Misspecification in Data-Driven Optimization

Adam N. Elmachtoub
Columbia University
adam@ieor.columbia.edu

Henry Lam
Columbia University
henry.lam@columbia.edu

Haixiang Lan
Columbia University
haixiang.lan@columbia.edu

Haofeng Zhang *
Columbia University
Morgan Stanley
hz2553@columbia.edu

Abstract

Data-driven optimization aims to translate a machine learning model into decision-making by optimizing decisions on estimated costs. Such a pipeline can be conducted by fitting a distributional model which is then plugged into the target optimization problem. While this fitting can utilize traditional methods such as maximum likelihood, a more recent approach uses estimation-optimization integration that minimizes decision error instead of estimation error. Although intuitive, the statistical benefit of the latter approach is not well understood yet is important to guide the prescriptive usage of machine learning. In this paper, we dissect the performance comparisons between these approaches in terms of the amount of model misspecification. In particular, we show how the integrated approach offers a "universal double benefit" on the top two dominating terms of regret when the underlying model is misspecified, while the traditional approach can be advantageous when the model is nearly well-specified. Our comparison is powered by finite-sample tail regret bounds that are derived via new higher-order expansions of regrets and the leveraging of a recent Berry-Esseen theorem.

1 INTRODUCTION

In data-driven decision-making, optimal decisions are determined by minimizing a cost function that can only be estimated from observed data. Unlike standard machine learning prediction, this calls for a *prescriptive* use of data, where the goal is to obtain statistically outperforming decisions rather than predictions. To this end, a natural approach is "estimate-then-optimize (ETO)". Namely, we estimate unknown parameters from data via established statistical methods such as maximum likelihood estimation (MLE), and then plug into the downstream optimization task. Recently, an integrated estimation-optimization (IEO) approach has gained popularity. This approach estimates parameters by minimizing the empirical costs directly, thus accounting for the downstream optimization in the estimation procedure. Such an integrated perspective is intuitive and has propelled an array of studies in data- or machine-learning-driven optimization (Kao et al., 2009; Donti et al., 2017; Elmachtoub and Grigas, 2022). Further properties and methods are studied in the integrated framework, including surrogate loss functions (Loke et al., 2022; Chung et al., 2022), calibration properties (Liu and Grigas, 2021; Ho-Nguyen and Kılınç-Karzan, 2022), online decision-making (Liu and Grigas, 2022), active learning (Liu et al., 2023), differentiable optimizers (Berthet et al., 2020; Blondel et al., 2020), combinatorial optimization (Mandi et al., 2020; Muñoz et al., 2022; Jeong et al., 2022) and tree-based approaches (Elmachtoub et al., 2020; Kallus and Mao, 2022).

Despite the growing popularity, the theoretical understanding of IEO has been mostly confined to algorithmic analysis (Amos and Kolter, 2017; Huang and Gupta, 2024; Bennouna et al., 2024) or generalization bounds based on individual analysis as opposed to comparisons with ETO (El Balghiti et al., 2023). While they provide some insights on the performances of IEO, they are unable to distinguish the benefits relative to ETO, especially at an instance-specific level. An exception is the recent work Elmachtoub et al. (2023); Hu et al. (2022), which compares the regrets

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

*Authors are listed alphabetically.

of IEO and ETO as data size grows. More specifically, Elmachtoub et al. (2023) shows that IEO has a preferable regret when the model is misspecified, and vice versa when the model is well-specified, where the preference is induced via stochastic dominance. However, because of the asymptotic nature of their analysis, the comparison characterization in Elmachtoub et al. (2023) is notably "discontinuous" when transitioning from well-specified to misspecified models, in that the preference ordering of IEO and ETO changes abruptly when the model is misspecified even infinitesimally. Since in reality "all models are wrong", the result in Elmachtoub et al. (2023) bears a conceptual gap with practical usage. On the other hand, Hu et al. (2022) considers a setting when the cost function is linear, whereas we consider nonlinear cost functions.

The main goal of this paper is to provide finite-sample performance comparisons between ETO and IEO via a sharper analysis that notably captures the *amount of model misspecification* in a smooth manner. Specifically, we provide a detailed dissection on the difference between the tail probabilities of ETO's and IEO's regrets, in terms of a model misspecification measure δ that, roughly speaking, signifies the suboptimality due to misspecification, as well as the tail threshold. On a high level, our results imply that IEO exhibits a lighter regret tail than ETO when the model is sufficiently misspecified, while ETO exhibits a lighter regret tail when the model is nearly, but not necessarily completely, well-specified. Importantly, both of these cases do arise in practice while the previous literature focus only on the unrealistic well-specified setting.

Besides the above implications, our another key contribution lies in the creation of a technical roadmap that allows us to achieve the aforementioned sharp regret comparison. More precisely, we quantify the differences of regret probabilities at an accuracy that can conclude *two-sided* bounds. In contrast, as we will discuss, conventional generalization analyses in machine learning based on uniformity arguments would give worst-case bounds too loose to provide comparative insights. To this end, our regret comparison roadmap consists of two novel developments. One is the derivation of higher-order expansions of IEO's and ETO's regrets. These expansions reveal a "universal double benefit" of IEO in that the coefficients in the two most dominant terms of the IEO regret are always at most those of ETO, and these play an important role in justifying the benefit of IEO in misspecified settings. Second is the leveraging of recent Berry-Esseen bounds for M-estimation that allows the conversion of estimation errors into finite-sample regret bounds that sufficiently reflect the dominant error terms in the expansions.

2 SETUP AND PRELIMINARIES

Consider a stochastic optimization problem

$$\boldsymbol{w}^* \in \operatorname*{argmin}_{\boldsymbol{w}\in\Omega} \{v_0(\boldsymbol{w}) := \mathbb{E}_P[c(\boldsymbol{w}, \boldsymbol{z})]\} \tag{1}$$

where $\boldsymbol{w} \in \Omega \subset \mathbb{R}^p$ is the decision and Ω is an open set in $\mathbb{R}^p$, $\boldsymbol{z} \in \mathcal{Z} \subset \mathbb{R}^d$ is a random vector distributed according to an unknown distribution $P \in \mathcal{P}$ and $\mathcal{P}$ is the collection of all potential distributions, $c(\cdot, \cdot)$ is a known cost function, and $v_0(\cdot)$ is the expected cost under P. We aim to find an optimal decision $\boldsymbol{w}^*$. In data-driven stochastic optimization, the true distribution P is unknown, but we have i.i.d. data $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ generated from P.

We use a parametric approach to infer P, i.e., we will estimate P by a distribution in the family $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\} \subset \mathcal{P}$, parameterized by $\boldsymbol{\theta}$. We introduce the oracle

$$\boldsymbol{w}_{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{w}\in\Omega}\{v(\boldsymbol{w}, \boldsymbol{\theta}) := \mathbb{E}_{P_{\boldsymbol{\theta}}}[c(\boldsymbol{w}, \boldsymbol{z})]\}, \tag{2}$$

where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$ is the parameter of the underlying distribution $P_{\boldsymbol{\theta}}$ and Θ is an open set in $\mathbb{R}^q$, $\boldsymbol{z}$ is a random vector distributed according to $P_{\boldsymbol{\theta}}$, and $v(\cdot, \boldsymbol{\theta})$ is the expected cost under distribution $P_{\boldsymbol{\theta}}$. Note that $\boldsymbol{w}_{\boldsymbol{\theta}}$ is a minimizer of problem (2) when $P_{\boldsymbol{\theta}}$ is the distribution. The true distribution P may or may not be in the distribution family $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. More precisely:

Definition 1 (Well-Specified Model Family)**.** The distribution family $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is *well-specified* if there exists a $\boldsymbol{\theta}_0 \in \Theta$ such that $P = P_{\boldsymbol{\theta}_0}$.

Definition 2 (Misspecified Model Family)**.** The distribution family $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is *misspecified* if for all $\boldsymbol{\theta} \in \Theta$, $P \neq P_{\boldsymbol{\theta}}$.

We use the regret, i.e., the optimality gap or excess risk, to evaluate the quality of a decision $\boldsymbol{w}$.

Definition 3 (Regret)**.** For any $\boldsymbol{w} \in \Omega$, the *regret* of $\boldsymbol{w}$ is given by $R(\boldsymbol{w}) := v_0(\boldsymbol{w}) - v_0(\boldsymbol{w}^*)$, where $\boldsymbol{w}^*$ is an optimal solution to (1).

2.1 Data-Driven Optimization Approaches

We consider two approaches to obtain a data-driven solution of (1). Both approaches rely on using the data $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ to estimate a $\hat{\boldsymbol{\theta}}$, which is then plugged into (2) to generate a solution $\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}$.

Estimate-Then-Optimize (ETO): We use maximum likelihood estimation (MLE) to infer $\boldsymbol{\theta}$, i.e., $\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} := \arg\sup_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^n \log p_{\boldsymbol{\theta}}(\boldsymbol{z}_i)$ where $P_{\boldsymbol{\theta}}$ has probability density function $p_{\boldsymbol{\theta}}$. By plugging $\hat{\boldsymbol{\theta}}^{\mathrm{ETO}}$ into the objective, we obtain the decision $\hat{\boldsymbol{w}}^{\mathrm{ETO}} := \boldsymbol{w}_{\hat{\boldsymbol{\theta}}^{\mathrm{ETO}}} = \operatorname{argmin}_{\boldsymbol{w}\in\Omega} v(\boldsymbol{w}, \hat{\boldsymbol{\theta}}^{\mathrm{ETO}})$. ETO uses MLE to

infer $\boldsymbol{\theta}$ and then plugs $\hat{\boldsymbol{\theta}}^{\mathrm{ETO}}$ into the optimization problem (2) to obtain $\hat{\boldsymbol{w}}^{\mathrm{ETO}}$. We denote $\boldsymbol{\theta}^{\mathrm{KL}} := \operatorname{argmin}_{\boldsymbol{\theta}\in\Theta} \mathrm{KL}(P, P_{\boldsymbol{\theta}}) = \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_P[\log p_{\boldsymbol{\theta}}(\boldsymbol{z})]$.

Integrated-Estimation-Optimization (IEO): We infer $\boldsymbol{\theta}$ by solving

$$\inf_{\boldsymbol{\theta}\in\Theta} \hat{v}_0(\boldsymbol{w}_{\boldsymbol{\theta}}) := \frac{1}{n}\sum_{i=1}^{n} c(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z}_i) \tag{3}$$

where $\boldsymbol{w}_{\boldsymbol{\theta}}$ is the oracle solution defined in (2). The function $c(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z})$ is called the *IEO loss* function with respect to $\boldsymbol{\theta}$. IEO integrates optimization with estimation in that the loss function used to "train" $\boldsymbol{\theta}$ is the decision-making optimization problem evaluated on $\boldsymbol{w}_{\boldsymbol{\theta}}$. In other words, when we make decisions from a model parameterized by $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}^{\mathrm{IEO}}$ is the choice that leads to the lowest empirical risk. By plugging $\hat{\boldsymbol{\theta}}^{\mathrm{IEO}}$ into the objective, we obtain the decision $\hat{\boldsymbol{w}}^{\mathrm{IEO}} := \boldsymbol{w}_{\hat{\boldsymbol{\theta}}^{\mathrm{IEO}}} = \operatorname{argmin}_{\boldsymbol{w}\in\Omega} v(\boldsymbol{w}, \hat{\boldsymbol{\theta}}^{\mathrm{IEO}})$. We denote $\boldsymbol{\theta}^* := \operatorname{argmin}_{\boldsymbol{\theta}\in\Theta} v_0(\boldsymbol{w}_{\boldsymbol{\theta}})$.

Since the data is random, $R(\hat{\boldsymbol{w}})$ is a random variable. In this work, we shall directly compare ETO and IEO by comparing the tail probabilities of the regret distributions $R(\hat{\boldsymbol{w}}^{\mathrm{ETO}})$ and $R(\hat{\boldsymbol{w}}^{\mathrm{IEO}})$. We utilize the notion of first-order stochastic dominance (Quirk and Saposnik, 1962) and second-order stochastic dominance (Rothschild and Stiglitz, 1978) to rank two random variables, as defined below.

Definition 4 (First Order Stochastic Dominance)**.** For any two random variables X, Y, we say that X is first-order stochastically dominated by Y, written as $X \preceq_{\mathrm{st}} Y$, or $Y \succeq_{\mathrm{st}} X$, if

$$\mathbb{P}[X > x] \le \mathbb{P}[Y > x] \quad \text{for all } x \in \mathbb{R}. \tag{4}$$

Definition 5 (Second Order Stochastic Dominance)**.** For any two random variables X and Y, we say that X is second-order stochastically dominated by Y, written as $X \preceq_{\text{s-st}} Y$, or $Y \succeq_{\text{s-st}} Y$, if $\mathbb{E}[u(X)] \le \mathbb{E}[u(Y)]$ for all non-decreasing convex function $u : \mathbb{R} \to \mathbb{R}$.

We provide two stylized examples in data-driven decision making in our framework: the newsvendor problem and portfolio optimization.

Example 1 (Multi-product Newsvendor)**.** The multi-product newsvendor problem aims to find the optimal order quantities of p products. The cost function can be represented as $c(\boldsymbol{w}, \boldsymbol{z}) := \boldsymbol{h}^\top(\boldsymbol{w}-\boldsymbol{z})^+ + \boldsymbol{b}^\top(\boldsymbol{z}-\boldsymbol{w})^+$, where $\boldsymbol{z} = (z_1, \ldots, z_p)$ is the random demand each product and $\boldsymbol{w} = (w_1, \ldots, w_p)$ is the order quantity for each product. The holding and backlogging cost are $\boldsymbol{h} = (h_1, \ldots, h_p)$ and $\boldsymbol{b} = (b_1, \ldots, b_p)$, respectively.

Example 2 (Portfolio Optimization (Iyengar et al., 2023))**.** The risk-averse portfolio optimization problem with exponential utility aims to find the optimal allocation of p assets. The problem has the cost function $c(\boldsymbol{w}, \boldsymbol{z}) := \exp(-\boldsymbol{z}^\top \boldsymbol{w}) + \gamma\|\boldsymbol{w}\|^2$, where $\boldsymbol{z}$ denotes the random return of each asset and $\boldsymbol{w}$ denote the allocation weight of each asset.

Notation. For nonnegative sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n \lesssim b_n$ or $a_n = O(b_n)$ or $b_n = \Omega(a_n)$ if there exists a positive constant C, independent of n, such that $a_n \le C b_n$. In addition, we write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For a general distribution $\tilde{P}$, we write $\mathbb{E}_{\tilde{P}}[\cdot]$ and $\mathrm{var}_{\tilde{P}}(\cdot)$ as the expectation and (co)variance with respect to the distribution $\tilde{P}$. Unless otherwise specified, we use $\|\cdot\|$ to denote $\|\cdot\|_2$ of a vector or a matrix. For a random vector $\boldsymbol{X}$ and for any $p \ge 1$, let $\|\boldsymbol{X}\|_p := (\mathbb{E}[\|\boldsymbol{X}\|^p])^{1/p}$ be the L_p-norm of $\boldsymbol{X}$. When the differentiable map $y(\boldsymbol{u}) : \mathbb{R}^{d_1} \to \mathbb{R}$ is real-valued, the gradient $\nabla y(\boldsymbol{u})$ is a row vector $1 \times d_1$. For any symmetric matrix Q, we write $Q \ge 0$ if Q is positive semi-definite and $Q > 0$ if Q is positive definite. For two symmetric matrices Q_1 and Q_2, we write $Q_1 \ge Q_2$ if $Q_1 - Q_2 \ge 0$. Similarly, we write $Q_1 > Q_2$ if $Q_1 - Q_2 > 0$. We denote $\boldsymbol{Y}_0$ as the standard multivariate Gaussian vector $N(\boldsymbol{0}, \boldsymbol{I})$ with a proper dimension.

3 GENERALIZATION BOUNDS

We first derive finite-sample guarantees of IEO by utilizing conventional generalization error analysis. While this approach gives rise to regret bounds, we will also see its limitation in generating comparative insights that motivates our approach in Section 4.

In this section, we further assume the feasible region, the parameter space, and the cost function satisfy the following standard properties in learning theory.

Assumption 1. *The feasible decision region Ω is convex and the parameter space Θ is bounded with $E_\Theta := \sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{\theta}\|$.*

Assumption 2.A (Convexity)**.** *For any fixed $\boldsymbol{z}$, $c(\cdot, \boldsymbol{z})$ is a convex function of $\boldsymbol{w}$. For any $Q \in \mathcal{P}$, $v(\boldsymbol{w}, Q) := \mathbb{E}_Q c(\boldsymbol{w}, \boldsymbol{z})$ is ρ_c-strongly convex with respect to $\boldsymbol{w}$.*

Assumption 2.B (Lipschitzness)**.** *For any fixed $\boldsymbol{z}$, the cost function $c(\cdot, \boldsymbol{z})$ is L_c-Lipschitz: $\forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in \Omega, \boldsymbol{z} \in \mathcal{Z}$, $|c(\boldsymbol{w}_1, \boldsymbol{z}) - c(\boldsymbol{w}_2, \boldsymbol{z})| \le L_c\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|$.*

Assumption 2.C (Boundedness)**.** *The cost function $c(\cdot, \cdot)$ is bounded: $B_c := \sup_{\boldsymbol{w}\in\Omega, \boldsymbol{z}\in\mathcal{Z}} c(\boldsymbol{w}, \boldsymbol{z}) - \inf_{\boldsymbol{w}\in\Omega, \boldsymbol{z}\in\mathcal{Z}} c(\boldsymbol{w}, \boldsymbol{z}) < \infty$.*

We further assume the distribution family is a *good parametrization*, adapted from Doss et al. (2023), so that the distance between two parametric distributions are controlled by the distance between their corresponding parameters. We will give many examples that satisfy Assumption 3 in Appendix A.1.

Table 1: Summary of the comparisons between ETO and IEO. The results are expressed in terms of the tail probability difference $\mathcal{D} = \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \geq t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \geq t)$. $\epsilon_{n,t}$ is a small constant vanishing as the sample size n increases.

Model Misspecification	$\kappa_0^{\mathbf{IEO}} < t < \kappa_0^{\mathbf{ETO}}$	$t > \kappa_0^{\mathbf{ETO}}$
$\delta \gg 0$	$\mathcal{D} \geq 1 - \epsilon_{n,t}$	$\mathcal{D} \geq -\epsilon_{n,t}$
$\delta \approx 0$ and $B_0 \approx 0$	$\mathcal{D} \geq 1 - \epsilon_{n,t}$	$\mathcal{D} \leq \epsilon_{n,t}$ (for only large t)
$\delta = 0$ and $B_0 \approx 0$	N/A (since $\kappa_0^{\mathrm{IEO}} = \kappa_0^{\mathrm{ETO}}$)	$\mathcal{D} \leq \epsilon_{n,t}$
$\delta = 0$ and $B_0 = 0$	N/A (since $\kappa_0^{\mathrm{IEO}} = \kappa_0^{\mathrm{ETO}}$)	$\mathcal{D} \leq -C + \epsilon_{n,t}$

Assumption 3 (Good Parametrization)**.** *The family of distributions $\mathcal{P}_\Theta := \{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ is a good parametrization with respect to the total variation distance d_{TV} if there exists $D_\Theta > 0$ such that for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, $d_{\mathrm{TV}}(P_{\boldsymbol{\theta}_1}, P_{\boldsymbol{\theta}_2}) \leq D_\Theta \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$.*

Under the assumptions, we have the following generalization bound:

Lemma 1. *Under Assumptions 1, 2 and 3, there exists an absolute constant C_{abs} such that for any $1 > \tilde{\delta} > 0$, with probability at least $1 - \tilde{\delta}$ the following holds for all $\boldsymbol{\theta} \in \Theta$:*

$$v_0(\boldsymbol{w}_{\boldsymbol{\theta}}) \leq \hat{v}_0(\boldsymbol{w}_{\boldsymbol{\theta}}) + \frac{4\sqrt{2}L_c^2 C_{\mathrm{abs}} D_\Theta E_\Theta}{\rho_c}\sqrt{\frac{q}{n}} + B_c\sqrt{\frac{\log(1/\tilde{\delta})}{2n}}.$$

Furthermore, since $\hat{\boldsymbol{w}}^{\mathrm{IEO}}$ is an empirical minimizer of $\hat{v}_0(\cdot)$, we can bound its regret as follows.

Theorem 1. *Under Assumptions 1, 2 and 3, there exists an absolute constant C_{abs} such that for any $\tilde{\delta} > 0$, with probability at least $1 - \tilde{\delta}$, the decision $\hat{\boldsymbol{w}}^{\mathrm{IEO}}$ returned by IEO satisfies:*

$$R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \leq R(\boldsymbol{w}_{\boldsymbol{\theta}^*}) + \frac{4\sqrt{2}L_c^2 C_{\mathrm{abs}} D_\Theta E_\Theta}{\rho_c}\sqrt{\frac{q}{n}} + 2B_c\sqrt{\frac{\log(2/\tilde{\delta})}{2n}}.$$

Note that the result in Lemma 1, which subsequently gives rise to the regret bound of IEO in Theorem 1, holds uniformly for all $\boldsymbol{\theta} \in \Theta$ including $\hat{\boldsymbol{\theta}}^{\mathrm{ETO}}$ and $\hat{\boldsymbol{\theta}}^{\mathrm{IEO}}$. In other words, it is too loose to distinguish the differences between IEO and ETO. From another perspective, the attained bound has a $O(\frac{1}{\sqrt{n}})$ rate, but since we are considering parametric approaches, the convergence of the empirical minimizer $\hat{\boldsymbol{w}}^{\mathrm{IEO}}$ is usually at a faster $O(\frac{1}{n})$ rate. Note that established fast rate analysis, even if doable for our setting, provides at best upper bounds on regrets and is still insufficient to compare ETO and IEO at an instance-specific level.

We conclude this section by noting that El Balghiti et al. (2023) and Qi et al. (2021) also study generalization bounds for IEO of the type described above, but El Balghiti et al. (2023) only considers linear cost functions, and Qi et al. (2021) nonlinear cost but with finite discrete support on the uncertain parameter. To this end, we appear the first to analyze generalization bounds with nonlinear objectives and general parametric distributions. Nonetheless, the limitations in such bounds motivate us to consider a new analysis roadmap that we present next.

4 DISSECTED COMPARISON

In this section, we present the analysis of regret comparisons between IEO and ETO, which bypasses the limitations of conventional generalization bounds. Table 1 overviews our main findings. We use two key measurements of model misspecification that control the performance comparisons between ETO and IEO, δ and B_0. Herein, δ is defined as $\delta := v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})$, which describes loosely the distinction between the population-level ETO and IEO solutions. B_0 is used to characterize the distance between the true distribution and the parametric family, and its precise definition is deferred to Section 4.1.

Table 1 summarizes the tail probability difference of ETO and IEO, $\mathcal{D} := \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \geq t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \geq t)$, across different scenarios of δ and B_0. If $\mathcal{D} \leq 0$, i.e., $\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \geq t) \leq \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \geq t)$, then ETO has a smaller regret tail than IEO and thus the occurrence of large regrets in ETO is less often than IEO, making ETO advantageous and vice versa.

To interpret Table 1, we look at two cases for example. Case 1: when $\delta \gg 0$ and $\kappa_0^{\mathrm{IEO}} < t < \kappa_0^{\mathrm{ETO}}$, we have $\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \geq t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \geq t) \geq 1 - \epsilon_{n,t}$ where ϵ_n is a statistical error vanishing as $n \to \infty$. This suggests that IEO has a much lighter tail probability than ETO, and thus IEO tends to have a lower regret in this case. Case 2: When $\delta = 0, B_0 \approx 0$ and $t > \kappa_0^{\mathrm{ETO}}$, we have $\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \geq t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \geq t) \leq \epsilon_{n,t}$. In this case, ETO shows an advantage of alleviating the occurrence of extremely large regrets compared to IEO

in a slightly misspecified model.

In the following, we lay out our technical developments in attaining the insights in Table 1, which comprise two ingredients. First is high-order Taylor expansions of IEO's and ETO's regrets (Section 4.1). We discuss the zeroth-order ($O(1)$), first-order ($O(1/\sqrt{n})$) and second-order ($O(1/n)$) term from the expansion. These expansions reveal a universal double benefit of IEO in that its zeroth- and first-order terms are both at most those of ETO, regardless of model specification. In the second-order term, however, a universal ordering does not exist, but the ordering can be established when the model is well-specified or slightly misspecified. Our second key ingredient is the conversion of estimation errors to finite-sample regret bounds (Section 4.2) via the leveraging of recent Berry-Esseen bounds (Shao and Zhang (2022)) at an accuracy that allows dissected comparisons between IEO and ETO (Section 4.3).

4.1 Higher-Order Regret Expansions

To facilitate discussion, we define $\kappa_0^{\mathrm{ETO}} := v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) - v_0(\boldsymbol{w}^*)$ and $\kappa_0^{\mathrm{IEO}} := v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) - v_0(\boldsymbol{w}^*)$, the regret error of the population-level ETO and IEO solutions caused by misspecification. In addition, we denote two matrices $\boldsymbol{M}_1^{\mathrm{ETO}}$, $\boldsymbol{M}_1^{\mathrm{IEO}}$ as $\boldsymbol{M}_1^{\mathrm{ETO}} := \big((\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbb{E}_P[\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})])^{-1} \operatorname{var}_P(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})) (\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbb{E}_P[\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})])^{-1}\big)^{\frac{1}{2}}$ and $\boldsymbol{M}_1^{\mathrm{IEO}} := \Big(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})^{-1} \operatorname{var}_P(\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}^*}, \boldsymbol{z})) \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})^{-1}\Big)^{\frac{1}{2}}$. We define two Gaussian distributions $\mathbb{N}_1^{\mathrm{ETO}} \overset{d}{=} \boldsymbol{M}_1^{\mathrm{ETO}}\boldsymbol{Y}_0$ and $\mathbb{N}_1^{\mathrm{IEO}} \overset{d}{=} \boldsymbol{M}_1^{\mathrm{IEO}}\boldsymbol{Y}_0$ where $\boldsymbol{Y}_0$ denotes the standard normal distribution with the corresponding dimension. The rationale for defining these notations originates from the standard asymptotic normality of M-estimation. $\boldsymbol{M}_1^{\mathrm{ETO}}$ (resp. $\boldsymbol{M}_1^{\mathrm{IEO}}$) is the asymptotic normality variance of ETO (resp. IEO); See Proposition 3 in Appendix B. Some well-established standard assumptions (Assumptions 6, 7, and 8) and supporting auxiliary asymptotic results are also listed in Appendix B.

The following theorem shows the expansions of the regrets of IEO and ETO.

Theorem 2 (Double Benefit of IEO)**.** *Under Assumptions 6 and 7, we have:*

1. (Zeroth order)

$$R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \xrightarrow{P} \kappa_0^{\mathrm{ETO}},$$
$$R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \xrightarrow{P} \kappa_0^{\mathrm{IEO}}.$$

Moreover, $\kappa_0^{\mathrm{ETO}} \geq \kappa_0^{\mathrm{IEO}} \geq 0$.

2. (First order).

$$\sqrt{n}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) - \kappa_0^{\mathrm{ETO}}) \xrightarrow{d} \nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\mathbb{N}_1^{\mathrm{ETO}},$$
$$\sqrt{n}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - \kappa_0^{\mathrm{IEO}}) \xrightarrow{P} 0.$$

Moreover, $0 \preceq_{\text{s-st}} \nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\mathbb{N}_1^{\mathrm{ETO}}$. *Here* $\preceq_{\text{s-st}}$ *represents second-order stochastic dominance.*

3. (Second order).

$$n(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) - \kappa_0^{\mathrm{ETO}} - \nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}}))$$
$$\xrightarrow{d} \mathbb{G}^{\mathrm{ETO}} := \frac{1}{2}{\mathbb{N}_1^{\mathrm{ETO}}}^{\top} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\mathbb{N}_1^{\mathrm{ETO}},$$
$$n(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - \kappa_0^{\mathrm{IEO}})$$
$$\xrightarrow{d} \mathbb{G}^{\mathrm{IEO}} := \frac{1}{2}{\mathbb{N}_1^{\mathrm{IEO}}}^{\top} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})\mathbb{N}_1^{\mathrm{IEO}}.$$

Theorem 2 develops a higher-order expansion of the regret as a random variable. In particular, Theorems 2.1 and 2.2 establish a *universal* ordering between ETO and IEO in terms of the zeroth- and first-order terms, in that IEO is always at least as good as ETO regardless of the used model. Therefore, IEO not only *has a lower asymptotic regret but also a faster convergence rate compared to ETO*. In fact, the first-order regret of IEO is 0, which implies that the regret of IEO minus the model bias converges at the rate of $\frac{1}{n}$. However, the regret of ETO minus the model bias has a convergence rate $\frac{1}{\sqrt{n}}$.

Unlike Theorems 2.1 and 2.2, Theorem 2.3 does not immediately imply a universal ordering among IEO and ETO at the second-order expansion ($\frac{1}{n}$-term). It is tempting to believe that the comparison of $\mathbb{G}^{\mathrm{ETO}}$, $\mathbb{G}^{\mathrm{IEO}}$ in the well-specified case established in previous literature (Elmachtoub et al., 2023) would also hold in the misspecified case. However, this is not always the case. The comparison in terms of the second-order expansion, $\mathbb{G}^{\mathrm{IEO}}$ vs $\mathbb{G}^{\mathrm{ETO}}$, is delicate and depends on the model specification. We will zoom into the analysis of this term for the rest of this section.

We introduce some notations first. Let $P^{\mathrm{KL}} = P_{\boldsymbol{\theta}^{\mathrm{KL}}}$ and $P^* = P_{\boldsymbol{\theta}^*}$ where we omit the $\boldsymbol{\theta}$ notation in the distributions to avoid the confusion when we apply the gradient. For instance, $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbb{E}_{P^{\mathrm{KL}}}[\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})] = (\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbb{E}_{P^{\mathrm{KL}}}[\log p_{\boldsymbol{\theta}}(\boldsymbol{z})])\mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\mathrm{KL}}}$ which clearly states that the Hessian $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}$ is with respect to the $\boldsymbol{\theta}$ in $\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})$, free of P^{KL}. We denote two matrices $\tilde{\boldsymbol{M}}_1^{\mathrm{ETO}}$, $\tilde{\boldsymbol{M}}_1^{\mathrm{IEO}}$ as $\tilde{\boldsymbol{M}}_1^{\mathrm{ETO}} := \big((\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbb{E}_{P^{\mathrm{KL}}}[\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})])^{-1} \operatorname{var}_{P^{\mathrm{KL}}}(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})) (\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbb{E}_{P^{\mathrm{KL}}}[\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})])^{-1}\big)^{\frac{1}{2}}$ and $\tilde{\boldsymbol{M}}_1^{\mathrm{IEO}} := \Big(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}, P^{\mathrm{KL}})^{-1} \operatorname{var}_{P^{\mathrm{KL}}}(\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}, \boldsymbol{z})) \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}, P^{\mathrm{KL}})^{-1}\Big)^{\frac{1}{2}}$.

Note that the difference between $\tilde{\boldsymbol{M}}_1^{\mathrm{IEO}}$ (resp. $\tilde{\boldsymbol{M}}_1^{\mathrm{ETO}}$) and $\boldsymbol{M}_1^{\mathrm{IEO}}$ (resp. $\boldsymbol{M}_1^{\mathrm{ETO}}$) is that the mean and variance are taken with respect to P^{KL} instead of the ground-truth P. So $\tilde{\boldsymbol{M}}_1^{\mathrm{IEO}}$ (resp. $\tilde{\boldsymbol{M}}_1^{\mathrm{ETO}}$) is the asymptotic normality variance of IEO (resp. ETO) when P^{KL} is viewed as the ground-truth data distribution.

Next we introduce how to measure the degree of model misspecification. The most natural way is via κ_0^{ETO} and κ_0^{IEO}, the regret contributions of the population-level ETO and IEO solutions caused by misspecification. However, these measurements are not tight enough to trade off with data uncertainty. This points to the need of other measurements to describe a *slightly misspecified* model besides $\kappa_0^{\mathrm{ETO}} \approx \kappa_0^{\mathrm{IEO}}$. The intuition is that when the degree of misspecification is small, we would expect $P \approx P^* \approx P^{\mathrm{KL}}$ and $\boldsymbol{\theta}^{\mathrm{KL}} \approx \boldsymbol{\theta}^*$. To provide a rigorous definition,

Assumption 4 (Degree of misspecification)**.** *Suppose that the distributions P^{KL}, P^* and the true distribution P satisfy that there exists $B_0 \geq 0$ such that*

$$\begin{aligned}
\|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbb{E}_P[\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})]) - \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbb{E}_{P^{\mathrm{KL}}}[\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})])\| &\leq B_0,\\
\|\mathrm{var}_P(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})) - \mathrm{var}_{P^{\mathrm{KL}}}(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z}))\| &\leq B_0,\\
\left\|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) - \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}, P^{\mathrm{KL}})\right\| &\leq B_0,\\
\|\mathrm{var}_P(\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}^*}, \boldsymbol{z})) - \mathrm{var}_{P^{\mathrm{KL}}}(\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}, \boldsymbol{z}))\| &\leq B_0,\\
\|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) - \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\| &\leq B_0.
\end{aligned}$$

The assumption on the distributional similarity $P \approx P^* \approx P^{\mathrm{KL}}$ is similar to the integral probability metrics (Müller, 1997), which are a well-known type of distance measurement between probability distributions. If the model is well-specified, then $P = P^* = P^{\mathrm{KL}}$ and all parameters coincide with each other. In this case, Assumption 4 is satisfied with $B_0 = 0$. If the model is "almost" well-specified, from the continuity perspective, we expect the difference between the terms in the formulas in Assumption 4 to be still close to 0, say some $B_0 > 0$. Under Assumption 4, the next result demonstrates that the second-order regret of ETO is dominated by that of IEO up to the degree of misspecification.

Theorem 3. *Under Assumptions 4, 6, 7, and 8, there exist two random variables Z^{ETO}, Z^{IEO}, a matrix Δ and the standard Gaussian random vector $\boldsymbol{Y}_0 \sim N(\boldsymbol{0}, \boldsymbol{I})$ such that $Z^{\mathrm{ETO}} \overset{d}{=} \mathbb{G}^{\mathrm{ETO}}$, $Z^{\mathrm{IEO}} \overset{d}{=} \mathbb{G}^{\mathrm{IEO}}$, and*

$$\begin{aligned}
Z^{\mathrm{IEO}} &= Z^{\mathrm{ETO}} + \frac{1}{2}\boldsymbol{Y}_0^\top \Big(\boldsymbol{M}_1^{\mathrm{IEO}} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \boldsymbol{M}_1^{\mathrm{IEO}}\\
&\quad - \boldsymbol{M}_1^{\mathrm{ETO}} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) \boldsymbol{M}_1^{\mathrm{ETO}}\Big)\boldsymbol{Y}_0\\
&= Z^{\mathrm{ETO}} + \frac{1}{2}\boldsymbol{Y}_0^\top \Big(\tilde{\boldsymbol{M}}_1^{\mathrm{IEO}} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{IEO}}\\
&\quad - \tilde{\boldsymbol{M}}_1^{\mathrm{ETO}} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{ETO}} + \Delta\Big)\boldsymbol{Y}_0.
\end{aligned}$$

Herein, the matrix Δ satisfies $\|\Delta\| \leq C_{\mathrm{mis}} B_0$, where C_{mis} is a problem-dependent constant, and

$$\tilde{\boldsymbol{M}}_1^{\mathrm{IEO}} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{IEO}} - \tilde{\boldsymbol{M}}_1^{\mathrm{ETO}} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{ETO}}$$

is positive semi-definite.

Theorem 3 provides a precise distance measurement between $\mathbb{G}^{\mathrm{ETO}}$ and $\mathbb{G}^{\mathrm{IEO}}$ that depends on the model misspecification. Clearly, to measure the difference between $\mathbb{G}^{\mathrm{IEO}}$ and $\mathbb{G}^{\mathrm{ETO}}$, we need to study the matrix $\boldsymbol{M}_1^{\mathrm{IEO}} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \boldsymbol{M}_1^{\mathrm{IEO}} - \boldsymbol{M}_1^{\mathrm{ETO}} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) \boldsymbol{M}_1^{\mathrm{ETO}}$. Let τ_1 denote its smallest eigenvalue and τ_2 denote its largest eigenvalue.

Theorem 3 encompasses both well-specified and misspecified cases:

1. When the model is well-specified, then we can set $B_0 = 0$, and thus $\tau_1 \geq 0$ naturally holds. This immediately implies $\mathbb{G}^{\mathrm{ETO}}$ is first-order stochastically dominated by $\mathbb{G}^{\mathrm{IEO}}$.

2. When the model misspecification is small (Assumption 4), $\mathbb{G}^{\mathrm{ETO}}$ is first-order stochastically dominated by $\mathbb{G}^{\mathrm{IEO}}$ with an error related to the degree of the model misspecification. This means that when the model transits from well-specified to misspecified, the relation between $\mathbb{G}^{\mathrm{IEO}}$ and $\mathbb{G}^{\mathrm{ETO}}$ has "continuity". We prove this rigorously in Corollary 1.

Corollary 1. *Under the conditions in Theorem 3, let*

$$0 \leq \tau_3 := \text{ the smallest eigenvalue of } \tilde{\boldsymbol{M}}_1^{\mathrm{IEO}} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{IEO}} - \tilde{\boldsymbol{M}}_1^{\mathrm{ETO}} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{ETO}}.$$

If B_0 satisfies $B_0 \leq \tau_3 / C_{\mathrm{mis}}$, then $\mathbb{G}^{\mathrm{ETO}}$ is first-order stochastically dominated by $\mathbb{G}^{\mathrm{IEO}}$.

Corollary 1 indicates that for the second-order term of the regret distribution, $\mathbb{G}^{\mathrm{ETO}}$ is first-order stochastically dominated by $\mathbb{G}^{\mathrm{IEO}}$ when the model misspecification is small (B_0 is small). However, this fact generally does *not* imply that the total regret of ETO is less than the total regret of IEO. We shall also take into account the double benefit of IEO in the first two dominating terms of regret (Theorem 2). It is delicate to balance these performance differences in each of the higher-order terms of the regret distribution. This will be our task in Section 4.3.

4.2 Finite-Sample Bounds

In this section, we derive finite-sample regret error bounds, achieved by utilizing Berry-Esseen-type bounds on the estimation error and then converting the estimation error to the finite-sample regrets.

First, via an existing result on the finite-sample guarantee of M-estimation (Lemma 8 in Appendix B), we immediately obtain the finite-sample performance of the $\boldsymbol{\theta}$ solution in ETO and IEO (Proposition 4 in Appendix B). In the following, we integrate it with the regret analysis in Section 4.1 to establish finite-sample performance guarantees. We introduce the following Lipschitz assumption on the Hessian of the expected cost.

Assumption 5. *We assume that Θ is bounded and the function $v_0(\boldsymbol{w}_{\boldsymbol{\theta}})$ satisfies for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$,*

$$\|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}_1}) - \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}_2})\| \le L_1 \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

From Assumption 5, it immediately follows that the Hessian function $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{\theta})$ is a bounded function on Θ with $L_2 := \sup_{\boldsymbol{\theta}\in\Theta} \|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}})\| < \infty$.

Proposition 1.A (Finite-sample regret for ETO, part I)**.** *Under Assumptions 5, 6.A, and 7.A, for all $t \in \mathbb{R}$:*

$$\begin{aligned}&|\mathbb{P}(\sqrt{n}(v_0(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})) \ge t)\\ &- \mathbb{P}(\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) \mathbb{N}_1^{\mathrm{ETO}} \ge t)| \le G_{n,q}^{\mathrm{ETO}}\\ &:\lesssim \frac{\left\|\boldsymbol{M}_1^{\mathrm{ETO}}\right\|^2 L_2}{\left\|\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) \boldsymbol{M}_1^{\mathrm{ETO}}\right\|} q n^{-\frac{1}{2}} \log n + C_{n,q}^{\mathrm{ETO}}.\end{aligned}$$

Herein, "$\lesssim$" hides the constant independent of n and q. Moreover, $G_{n,q}^{\mathrm{ETO}}$ is independent of t. $C_{n,q}^{\mathrm{ETO}} \lesssim q^{9/4} n^{-1/2}$ is given in Proposition 4.

Proposition 1.B (Finite-sample regret for IEO, part I)**.** *Under Assumptions 5, 6.B, and 7.B, for all $t > 0$:*

$$\begin{aligned}&\mathbb{P}(\sqrt{n}(v_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})) \ge t)\\ &\le G_{n,q,t}^{\mathrm{IEO}} :\lesssim q \exp\left(-\frac{\sqrt{n} t}{q L_2 \left\|\boldsymbol{M}_1^{\mathrm{IEO}}\right\|^2}\right) + C_{n,q}^{\mathrm{IEO}}.\end{aligned}$$

Herein, "$\lesssim$" hides the constant independent of n and q. $C_{n,q}^{\mathrm{IEO}} \lesssim q^{9/4} n^{-1/2}$ is given in Proposition 4.

Proposition 1 indicates the finite-sample regret bounds for ETO and IEO with respect to the first-order term $O(\frac{1}{\sqrt{n}})$. Importantly, it shows that the finite-sample regret error vanishes at the rate (essentially) $O(\frac{1}{\sqrt{n}})$, which is the same as the finite-sample estimation error (Proposition 4). There is an additional $(\log n)$ in the bound of ETO since the first-order regret of ETO is non-zero (Theorem 2), which introduces an additional small error.

The next proposition indicates the finite-sample regret bounds with respect to the second-order term $O(\frac{1}{n})$.

Proposition 2.A (Finite-sample regret for ETO, part II)**.** *Under Assumptions 5, 6.A, and 7.A, and assume that $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) \ge 0$, we have for all $t > 0$,*

$$\begin{aligned}&|\mathbb{P}(n(v_0(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) - \nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}})) \\ &\ge t) - \mathbb{P}(\mathbb{G}^{\mathrm{ETO}} \ge t)| \lesssim D_{n,q}^{\mathrm{ETO}}.\end{aligned}$$

When $q = 1$,

$$D_{n,1}^{\mathrm{ETO}} \lesssim \frac{1}{\sqrt{\lambda_1}} L_1^{\frac{1}{2}} \left\|\boldsymbol{M}_1^{\mathrm{ETO}}\right\|^{\frac{3}{2}} (\log n)^{\frac{3}{4}} n^{-\frac{1}{4}} + C_{n,1}^{\mathrm{ETO}}$$

and when $q \ge 2$,

$$\begin{aligned}&D_{n,q}^{\mathrm{ETO}} \lesssim C_{n,q}^{\mathrm{ETO}} +\\ &\left(\left(\sum_{i=1}^q \lambda_i^2\right)\left(\sum_{i=2}^q \lambda_i^2\right)\right)^{-\frac{1}{4}} L_1 \left\|\boldsymbol{M}_1^{\mathrm{ETO}}\right\|^3 q^{3/2} (\log n)^{\frac{3}{2}} n^{-\frac{1}{2}}\end{aligned}$$

Herein, "$\lesssim$" hides constants independent of n and q. $D_{n,q}^{\mathrm{ETO}}$ is independent of t. $C_{n,q}^{\mathrm{ETO}} \lesssim q^{9/4} n^{-1/2}$ is given in Proposition 4. $\lambda_1, \ldots, \lambda_q \ge 0$ are the eigenvalues of $\frac{1}{2}\boldsymbol{M}_1^{\mathrm{ETO}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) \boldsymbol{M}_1^{\mathrm{ETO}}$.

Proposition 2.B (Finite-sample regret for IEO, part II)**.** *Under Assumptions 5, 6.B, and 7.B, for all $t > 0$,*

$$\left|\mathbb{P}(n v_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - n v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \ge t) - \mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \ge t)\right| \le D_{n,q}^{\mathrm{IEO}}.$$

When $q = 1$,

$$D_{n,1}^{\mathrm{IEO}} \lesssim \frac{1}{\sqrt{\lambda_1}} L_1^{\frac{1}{2}} \left\|\boldsymbol{M}_1^{\mathrm{IEO}}\right\|^{\frac{3}{2}} (\log n)^{\frac{3}{4}} n^{-\frac{1}{4}} + C_{n,1}^{\mathrm{IEO}}$$

and when $q \ge 2$,

$$\begin{aligned}&D_{n,q}^{\mathrm{IEO}} \lesssim C_{n,q}^{\mathrm{IEO}} +\\ &\left(\left(\sum_{i=1}^q \lambda_i^2\right)\left(\sum_{i=2}^q \lambda_i^2\right)\right)^{-\frac{1}{4}} L_1 \left\|\boldsymbol{M}_1^{\mathrm{IEO}}\right\|^3 q^{3/2} (\log n)^{\frac{3}{2}} n^{-\frac{1}{2}}\end{aligned}$$

Herein, "$\lesssim$" hides constants independent of n and q. $D_{n,q}^{\mathrm{IEO}}$ is independent of t. $C_{n,q}^{\mathrm{IEO}} \lesssim q^{9/4} n^{-1/2}$ is given in Proposition 4. $\lambda_1, \ldots, \lambda_q \ge 0$ are the eigenvalues of $\frac{1}{2}\boldsymbol{M}_1^{\mathrm{IEO}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \boldsymbol{M}_1^{\mathrm{IEO}}$.

Propositions 1 and 2 formally establish the individual finite-sample regret bounds for ETO and IEO by converting Berry–Esseen bounds for the estimation error to finite-sample regrets, using the higher-order expansion result we develop in Section 4.1. Unlike generalization bound in Theorem 1 (which are upper bounds), our bounds consists of both upper and lower bounds and thus provide a tighter characterization of the finite-sample regret distribution. These results will be used to derive the ultimate regret comparisons in Section 4.3.

An immediate implication of the above theorems is a high probability bound of the IEO regret and its fast rate compared to previous generalization bounds.

Corollary 2. *Under Assumptions 5, 6.B, and 7.B, there exists a problem dependent C_{prob}, such that for any $\varepsilon > 0$, when n satisfies $C_{\mathrm{prob}}(\log n)^{\frac{3}{2}} n^{-\frac{1}{2}} \le \varepsilon/2$ (for $q \ge 2$) or $C_{\mathrm{prob}}(\log n)^{\frac{3}{4}} n^{-\frac{1}{4}} \le \varepsilon/2$ (for $q = 1$), with probability at least $1 - \varepsilon$, $R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \le \kappa_0^{\mathrm{IEO}} + \frac{F_{\mathbb{G}^{\mathrm{IEO}}}^{-1}(1-\varepsilon/2)}{n}$.*

The above corollary implies when the sample size n is larger than an explicit threshold, the convergence rate of the empirical minimizer $\hat{\boldsymbol{w}}^{\mathrm{IEO}}$ is $O(1/n)$, matching the results in the asymptotic analysis.

4.3 Comparisons on Finite-Sample Regret

With the above developments, we now derive regret comparisons that draw our insights in Table 1. Our main results consist of two theorems that formally summarize the comparison in terms of the tail probability of the regret distribution, one in a generally misspecified model and one in a "slightly misspecified" model (including the well-specified model). The assumptions in the two theorems are not mutually exclusive.

Recall $\delta = \kappa_0^{\mathrm{ETO}} - \kappa_0^{\mathrm{IEO}}$. In Section 4.1, we have established the "double benefit" property of IEO. Intuitively, when δ is relatively large or the sample size n is relatively large, the major component of the regret distribution would be the first two order terms of the regret where IEO is strictly better. Combining this observation with the finite-sample regret bounds in Section 4.2, we can build an explicit finite-sample regret comparison.

Theorem 4 (Lower Bound on $\mathcal{D}$)**.** *Suppose Assumptions 4, 5, 6, 7, and 8 hold. We have:*

Case 1: $t \le \kappa_0^{\mathrm{IEO}}$:

$$\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge t) = 0$$

Case 2: $t > \kappa_0^{\mathrm{ETO}}$:

$$\begin{aligned}&\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge t)\\ &\quad\ge C - G^{\mathrm{IEO}}_{n,q,\sqrt{n}(t-\kappa_0^{\mathrm{IEO}})} - G^{\mathrm{ETO}}_{n,q}\end{aligned}$$

Herein, $C = 1 - \exp\left(-\frac{n(\kappa_0^{\mathrm{ETO}}-t)^2}{2\|\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}}\|^2}\right)$ *if* $\kappa_0^{\mathrm{IEO}} < t < \kappa_0^{\mathrm{ETO}}$, $C = 0$ *if* $t > \kappa_0^{\mathrm{ETO}}$, *and* $C = \frac{1}{2}$ *if* $t = \kappa_0^{\mathrm{ETO}}$. $G^{\mathrm{ETO}}_{n,q}$ *and* $G^{\mathrm{IEO}}_{n,q,\sqrt{n}(t-\kappa_0^{\mathrm{IEO}})}$ *are given in Proposition 1. All the error terms* $G^{\mathrm{IEO}}_{n,q,\sqrt{n}(t-\kappa_0^{\mathrm{IEO}})}$, $G^{\mathrm{ETO}}_{n,q}$ *go to* 0 *as* $n \to \infty$.

We discuss Theorem 4:

1. When $t < \kappa_0^{\mathrm{ETO}}$, the benefit of IEO is outstanding, since 1 is generally much larger than the statistical error terms $G^{\mathrm{IEO}}_{n,q,\sqrt{n}(t-\kappa_0^{\mathrm{IEO}})} + G^{\mathrm{ETO}}_{n,q}$ when n is not too small. To provide intuition, $\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge t) \to 1$ as n increases, since $R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \xrightarrow{P} \kappa_0^{\mathrm{ETO}} > t$. $\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge t) \to 0$ as n increases, since $R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \xrightarrow{P} \kappa_0^{\mathrm{IEO}} < t$. This clearly shows that $\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge t) \le \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge t)$ when n is large. Therefore, even in a situation with very slight misspecification where $\mathbb{G}^{\mathrm{ETO}}$ is first-order stochastically dominated by $\mathbb{G}^{\mathrm{IEO}}$ (Corollary 1), we cannot claim the total regret of ETO is first-order stochastically dominated by that of IEO.

2. When $t > \kappa_0^{\mathrm{ETO}}$, Theorem 4 does not provide a strict distinction between IEO and ETO. However, we can still assert that the tail probability of the IEO regret is dominated by that of ETO, up to a finite-sample statistical error ($G^{\mathrm{IEO}}_{n,q,\sqrt{n}(t-\kappa_0^{\mathrm{IEO}})} + G^{\mathrm{ETO}}_{n,q}$). In other words, even in cases where IEO, if at all, performs worse than ETO, any potential performance degradation would be limited.

3. When $t > \kappa_0^{\mathrm{ETO}}$, the bound in Theorem 4 and subsequent Theorem 5 becomes trivial if we only consider $n \to \infty$. This is because when $t > \kappa_0^{\mathrm{ETO}}$, both $\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge t) \to 0$ and $\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge t) \to 0$. Hence, while a standard asymptotic analysis cannot capture the subtle finite-sample tail difference in IEO and ETO, our Theorem 4 and subsequent Theorem 5 can as they hold for any $n \ge 1$.

Our next theorem shows that in a well-specified or slightly misspecified model, we can establish an upper bound for $\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge t)$. Recall that another measurement of model specification besides δ is based on the eigenvalues of $\boldsymbol{M}_1^{\mathrm{IEO}}\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})\boldsymbol{M}_1^{\mathrm{IEO}} - \boldsymbol{M}_1^{\mathrm{ETO}}\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}}$, which quantifies the difference in the second-order terms $\mathbb{G}^{\mathrm{IEO}}$ and $\mathbb{G}^{\mathrm{ETO}}$, as shown in Theorem 3. When $\tau_1 \ge 0$, from the previous discussion, we have $\mathbb{G}^{\mathrm{ETO}} \preceq_{\mathrm{st}} \mathbb{G}^{\mathrm{IEO}}$, which inherits the properties from the well-specified cases. In this case, ETO has some benefits in the $O(1/n)$ term, and such benefits can be formally stated in the following theorem.

Theorem 5 (Upper Bound on $\mathcal{D}$)**.** *Suppose Assumptions 4, 5, 6, 7, and 8 hold. Suppose that* $\tau_1 \ge 0$. *For instance, this holds when* $B_0 \le \tau_3/C_{\mathrm{mis}}$ *by Corollary 1. In addition, let* τ_6 *denote the largest eigenvalue of the matrix* $\boldsymbol{M}_1^{\mathrm{ETO}}\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}}$. *Then we have the following results for any sample size* $n \ge 1$.

Case 1: $t \le \kappa_0^{\mathrm{IEO}}$:

$$\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge t) = 0$$

Case 2: $t > \kappa_0^{\mathrm{IEO}} + \frac{\tau_6+\tau_1}{\tau_1}\delta$: *for any* $0 < \varepsilon < \frac{\tau_1}{\tau_1+\tau_6}(t -$

$\kappa_0^{\mathrm{IEO}}) - \delta$,

$$\begin{aligned}&\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \geq t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \geq t)\\ \leq& -\mathbb{P}(nt - n\kappa_0^{\mathrm{IEO}} \leq \mathbb{G}^{\mathrm{IEO}} \leq \\ &(1+\frac{\tau_1}{\tau_6})(nt - n\kappa_0^{\mathrm{IEO}} - n\delta - n\varepsilon))\\ &+ D_{n,q}^{\mathrm{ETO}} + D_{n,q}^{\mathrm{IEO}} + C_{n,q}^{\mathrm{ETO}} + E_n^{\delta,\varepsilon}.\end{aligned}$$

Particularly when $\delta = 0$, for any $t > \kappa_0^{\mathrm{IEO}}$,

$$\begin{aligned}&\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \geq t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \geq t)\\ \leq& -\mathbb{P}(nt - n\kappa_0^{\mathrm{IEO}} \leq \mathbb{G}^{\mathrm{IEO}} \leq (1+\frac{\tau_1}{\tau_6})(nt - n\kappa_0^{\mathrm{IEO}}))\\ &+ D_{n,q}^{\mathrm{ETO}} + D_{n,q}^{\mathrm{IEO}}.\end{aligned}$$

Herein,

$$E_n^{\delta,\varepsilon} = \exp\left(-\frac{n\varepsilon^2}{2\|\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}}\|^2}\right)$$

depends on the model misspecification. $D_{n,q}^{\mathrm{ETO}}$, $D_{n,q}^{\mathrm{IEO}}$, $C_{n,q}^{\mathrm{ETO}}$ are given by Propositions 2 and 4. All the error terms $D_{n,q}^{\mathrm{ETO}}$, $D_{n,q}^{\mathrm{IEO}}$, $C_{n,q}^{\mathrm{ETO}}$, $E_n^{\delta,\varepsilon}$ go to 0 as $n \to \infty$.

Note that Theorem 4 provides a lower bound on $\mathcal{D} = \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \geq t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \geq t)$ for all tail probabilities. In contrast, Theorem 5 provides an upper bound on $\mathcal{D}$, but only for part of the regret tail probabilities (not covering $\kappa_0^{\mathrm{IEO}} < t \leq \kappa_0^{\mathrm{IEO}} + \frac{\tau_6+\tau_1}{\tau_1}\delta$). We generally do not have a universal ordering in the intermediate region $\kappa_0^{\mathrm{IEO}} < t < \kappa_0^{\mathrm{IEO}} + \frac{\tau_6+\tau_1}{\tau_1}\delta$ when $\delta > 0$. However, when δ is sufficiently small, the intermediate region between Case 1 and Case 2 will be negligible. We discuss Theorem 5 further:

1. When the model is well-specified, then $\delta = 0$, $B_0 = 0$ and $\tau_1 \geq 0$. Then Theorem 5 immediately applies to the well-specified case. In this case, Case 1 and Case 2 together cover all the tail probabilities, showing that any tail probability of ETO is less than that of IEO with a finite-sample statistical error ($D_{n,q}^{\mathrm{ETO}} + D_{n,q}^{\mathrm{IEO}}$). This finite-sample estimation error will vanish as $n \to \infty$. To obtain an asymptotic result, letting $t = \frac{\tilde{t}}{n}$, we have

$$\begin{aligned}&\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \geq \frac{\tilde{t}}{n}) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \geq \frac{\tilde{t}}{n})\\ \leq& -\mathbb{P}(\tilde{t} \leq \mathbb{G}^{\mathrm{IEO}} \leq (1+\frac{\tau_1}{\tau_6})\tilde{t}) + D_{n,q}^{\mathrm{ETO}} + D_{n,q}^{\mathrm{IEO}}.\end{aligned}$$

   Note that the first term $\mathbb{P}(\tilde{t} \leq \mathbb{G}^{\mathrm{IEO}} \leq (1+\frac{\tau_1}{\tau_6})\tilde{t})$ is a constant independent of n, so taking $n \to \infty$,

$$\begin{aligned}&\lim_{n\to\infty}\left(\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \geq \frac{\tilde{t}}{n}) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \geq \frac{\tilde{t}}{n})\right)\\ \leq& -\mathbb{P}(\tilde{t} \leq \mathbb{G}^{\mathrm{IEO}} \leq (1+\frac{\tau_1}{\tau_6})\tilde{t}).\end{aligned}$$

   This shows that asymptotically, $nR(\hat{\boldsymbol{w}}^{\mathrm{ETO}})$ is first-order stochastically dominated by $nR(\hat{\boldsymbol{w}}^{\mathrm{IEO}})$, which leads to the asymptotic result in Elmachtoub et al. (2023).

2. When the model is misspecified, the double-benefit effect of IEO, as shown in Theorem 2, renders any potential advantage of ETO relatively minor. Moreover, such an advantage is only observed under a relatively restrictive setting. Theorem 5 demonstrates that ETO exhibits a small advantage when model misspecification is zero or small, corresponding to a small or zero δ and $\tau_1 \geq 0$ stipulated by Corollary 1. In this case, within a large tail region ($t > \kappa_0^{\mathrm{IEO}} + \frac{\tau_6+\tau_1}{\tau_1}\delta$), the regret tail probability of ETO is always dominated by that of IEO, up to a finite-sample statistical error ($D_{n,q}^{\mathrm{ETO}} + D_{n,q}^{\mathrm{IEO}} + C_{n,q}^{\mathrm{ETO}} + E_n^{\delta,\varepsilon}$). This suggests that ETO helps mitigate the occurrence of extremely large regrets compared to IEO.

5 Conclusions and Discussions

In this paper, we present a comprehensive theoretical analysis comparing the performance of IEO and ETO, focusing on how varying degrees of model misspecification influence their regrets across diverse scenarios. Our investigation employs several advanced techniques, including higher-order expansions of regret, the derivation of finite-sample regret bounds using recent Berry-Esseen results, and an in-depth characterization of regret tail behaviors in the finite-sample regime.

Our results demonstrate that IEO consistently enjoys a "universal double benefit" in the two leading dominant terms of regret under general model misspecification. This fundamental advantage enables IEO to outperform ETO across a broad range of practical settings, providing statistical evidence for its superior empirical performance. On the other hand, when the underlying model is nearly well-specified, ETO could exhibit advantages thanks to its smaller estimation variability, and these advantages show up in the second-order term of regret.

Acknowledgements

We gratefully acknowledge support from the National Science Foundation under grant CMMI-1763000, InnoHK initiative, the Government of the HKSAR, Laboratory for AI-Powered Financial Technologies, and Columbia SEAS Innovation Hub Award. The authors thank the anonymous reviewers for their constructive comments, which have greatly improved the quality of our paper.

References

B. Amos and J. Z. Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International conference on machine learning*, pages 136–145. PMLR, 2017.

A. Beck. *First-order methods in optimization*. SIAM, 2017.

O. Bennouna, J. Zhang, S. Amin, and A. Ozdaglar. Addressing misspecification in contextual optimization. *arXiv preprint arXiv:2409.10479*, 2024.

Q. Berthet, M. Blondel, O. Teboul, M. Cuturi, J.-P. Vert, and F. Bach. Learning with differentiable pertubed optimizers. *Advances in neural information processing systems*, 33:9508–9519, 2020.

M. Blondel, A. F. Martins, and V. Niculae. Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

S. G. Bobkov, A. A. Naumov, and V. V. Ulyanov. Two-sided inequalities for the density function's maximum of weighted sum of chi-square variables. *arXiv preprint arXiv:2012.10747*, 2020.

R. Busa-Fekete, D. Fotakis, B. Szörényi, and M. Zampetakis. Optimal learning of mallows block model. In *Conference on learning theory*, pages 529–532. PMLR, 2019.

T.-H. Chung, V. Rostami, H. Bastani, and O. Bastani. Decision-aware learning for optimizing health supply chains. *arXiv preprint arXiv:2211.08507*, 2022.

L. Devroye, A. Mehrabian, and T. Reddad. The total variation distance between high-dimensional gaussians with the same mean. *arXiv preprint arXiv:1810.08693*, 2018.

P. Donti, B. Amos, and J. Z. Kolter. Task-based end-to-end model learning in stochastic optimization. *Advances in neural information processing systems*, 30, 2017.

N. Doss, Y. Wu, P. Yang, and H. H. Zhou. Optimal estimation of high-dimensional gaussian location mixtures. *The Annals of Statistics*, 51(1):62–95, 2023.

O. El Balghiti, A. N. Elmachtoub, P. Grigas, and A. Tewari. Generalization bounds in the predict-then-optimize framework. *Mathematics of Operations Research*, 48(4):2043–2065, 2023.

A. N. Elmachtoub and P. Grigas. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022.

A. N. Elmachtoub, J. C. N. Liang, and R. McNellis. Decision trees for decision-making under the predict-then-optimize framework. In *International Conference on Machine Learning*, pages 2858–2867. PMLR, 2020.

A. N. Elmachtoub, H. Lam, H. Zhang, and Y. Zhao. Estimate-then-optimize versus integrated-estimation-optimization: A stochastic dominance perspective. *arXiv preprint arXiv:2304.06833*, 2023.

N. Ho-Nguyen and F. Kılınç-Karzan. Risk guarantees for end-to-end prediction and optimization processes. *Management Science*, 68(12):8680–8698, 2022.

Y. Hu, N. Kallus, and X. Mao. Fast rates for contextual linear optimization. *Management Science*, 2022.

M. Huang and V. Gupta. Decision-focused learning with directional gradients. *arXiv preprint arXiv:2402.03256*, 2024.

G. Iyengar, H. Lam, and T. Wang. Optimizer's information criterion: Dissecting and correcting bias in data-driven optimization. *arXiv preprint arXiv:2306.10081*, 2023.

J. Jeong, P. Jaggi, A. Butler, and S. Sanner. An exact symbolic reduction of linear smart predict+ optimize to mixed integer linear programming. In *International Conference on Machine Learning*, pages 10053–10067. PMLR, 2022.

N. Kallus and X. Mao. Stochastic optimization forests. *Management Science*, 2022.

Y.-h. Kao, B. Roy, and X. Yan. Directed regression. *Advances in Neural Information Processing Systems*, 22, 2009.

H. Liu and P. Grigas. Risk bounds and calibration for a smart predict-then-optimize method. *Advances in Neural Information Processing Systems*, 34:22083–22094, 2021.

H. Liu and P. Grigas. Online contextual decision-making with a smart predict-then-optimize method. *arXiv preprint arXiv:2206.07316*, 2022.

M. Liu, P. Grigas, H. Liu, and Z.-J. M. Shen. Active learning in the predict-then-optimize framework: A margin-based approach. *arXiv preprint arXiv:2305.06584*, 2023.

G. G. Loke, Q. Tang, and Y. Xiao. Decision-driven regularization: A blended model for predict-then-optimize. *Available at SSRN 3623006*, 2022.

J. Mandi, E. Demirović, P. J. Stuckey, and T. Guns. Smart predict–and–optimize for hard combinatorial optimization problems. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 1603–1610. AAAI Press, 2020.

A. Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pages 3–17. Springer, 2016.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.

R. J. Muirhead. *Quadratic Forms in Statistics*. Wiley, 2nd edition, 1982. ISBN 978-0824786914.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997.

M. A. Muñoz, S. Pineda, and J. M. Morales. A bilevel framework for decision-making under uncertainty with contextual information. *Omega*, 108:102575, 2022.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

M. Qi, P. Grigas, and Z.-J. M. Shen. Integrated conditional estimation-optimization. *arXiv preprint arXiv:2110.12351*, 2021.

J. P. Quirk and R. Saposnik. Admissibility and measurable utility functions. *The Review of Economic Studies*, 29(2):140–146, 1962.

M. Rothschild and J. E. Stiglitz. Increasing risk: I. a definition. In *Uncertainty in economics*, pages 99–121. Elsevier, 1978.

B. Sen. A gentle introduction to empirical process theory and applications. *Lecture Notes, Columbia University*, 11:28–29, 2018.

Q.-M. Shao and Z.-S. Zhang. Berry–esseen bounds for multivariate nonlinear statistics with applications to m-estimators and stochastic gradient descent algorithms. *Bernoulli*, 28(3):1548–1576, 2022.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

S. Wang, W. Zhou, H. Lu, A. Maleki, and V. Mirrokni. Approximate leave-one-out for fast parameter tuning in high dimensions. In *International Conference on Machine Learning*, pages 5228–5237. PMLR, 2018.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. NA
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). NA
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). NA
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). NA
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).NA
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. NA
   (b) The license information of the assets, if applicable. NA
   (c) New assets either in the supplemental material or as a URL, if applicable. NA
   (d) Information about consent from data providers/curators. NA
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. NA
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. NA
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. NA
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. NA

Dissecting the Impact of Model Misspecification in Data-Driven Optimization: Supplementary Materials

A Further Details and Proofs for Section 3

A.1 Examples of Parametric Distributions that Satisfy Assumption 3

We show the generality of Assumption 3 by giving concrete examples that satisfy the assumption.

Example 3 (Discrete Random Variables with Finite Support (Qi et al., 2021))**.** Suppose $\boldsymbol{z}$ has finite discrete support, i.e., $\boldsymbol{z} \in \mathcal{Z} := \{\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_k\}$. The distribution of $\boldsymbol{z}$ is characterized by a vector $\boldsymbol{\theta} \in \Delta^{q-1} := \{\boldsymbol{\theta} \in \mathbb{R}^q : \boldsymbol{\theta} \geq 0, \sum_{k=1}^q \theta_k = 1\}$. The total

$$d_{\mathrm{TV}}(\mathbb{P}_{\boldsymbol{\theta}_1}, \mathbb{P}_{\boldsymbol{\theta}_2}) = \frac{1}{2}\sum_{k=1}^{q} |\theta_{1k} - \theta_{2k}| = \frac{1}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_1 \leq D_\Theta \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

Example 4 (Multivariate Gaussian (Devroye et al., 2018))**.** Consider the distribution family $\mathcal{P}_\Theta := \{N(\boldsymbol{\theta}, \boldsymbol{\Sigma}) : \boldsymbol{\theta} \in \Theta\}$ where $\boldsymbol{\Sigma}$ is fixed. We have a closed form for the total variation distance:

$$\begin{aligned} d_{\mathrm{TV}}(N(\boldsymbol{\theta}_1, \boldsymbol{\Sigma}), N(\boldsymbol{\theta}_2, \boldsymbol{\Sigma})) &= \mathbb{P}\left(N(0,1) \in \left[-\frac{\sqrt{(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)}}{2}, \frac{\sqrt{(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)}}{2}\right]\right) \\ &= \Phi\left(\frac{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_{\boldsymbol{\Sigma}}}{2}\right) - \Phi\left(-\frac{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_{\boldsymbol{\Sigma}}}{2}\right) \lesssim \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_{\boldsymbol{\Sigma}} \leq D_\Theta \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|. \end{aligned}$$

Example 5 (Mixture Gaussian (Doss et al., 2023))**.** Let $\mathcal{G}_{k,d}$ denote the collection of k-atomic distributions supported on a ball of radius R is d dimensions, i.e.,

$$\begin{aligned} \mathcal{G}_{k,d} &:= \left\{\Gamma = \sum_{j=1}^{k} w_j \delta_{\mu_j} : \mu_j \in \mathbb{R}^d, \|\mu_j\|_2 \leq R, w_j \geq 0, \sum_{j=1}^{k} w_j = 1\right\}. \\ \mathcal{P}_{k,d} &:= \{P_\Gamma : \Gamma_{k,d}, P_\Gamma = \Gamma * N(0, \boldsymbol{I}_d)\}. \end{aligned}$$

A parametrization that satisfies the identifiability assumption is provided by the moment tensors. The degree-l moment tensor of the mixing distribution Γ is the symmetric tensor

$$\mathcal{M}_l(\Gamma) := \sum_{j=1}^{k} w_j \mu_j^{\otimes l}.$$

It can be shown that any k-atomic distribution is uniquely determined by its first $2k-1$ moment tensors $\boldsymbol{\mathcal{M}}_{2k-1}(\Gamma) = [\mathcal{M}_1(\Gamma), ..., \mathcal{M}_{2k-1}(\Gamma)]$. Consequently, moment tensors provides a valid parametrization of the k-Gaussian-Mixture in the sense that $\boldsymbol{\mathcal{M}}_{2k-1}(\Gamma) = \boldsymbol{\mathcal{M}}_{2k-1}(\Gamma') \Leftrightarrow P_\Gamma = P_{\Gamma'}$. We have the following property:

$$d_{\mathrm{TV}}(P_\Gamma, P'_\Gamma) \leq D_\Theta \max_{l \leq 2k-1} \|\mathcal{M}_l(\Gamma) - \mathcal{M}_l(\Gamma')\|_{\mathrm{F}}.$$

Example 6 (Exponential Family (Busa-Fekete et al., 2019)). Let μ be a measure on $\mathbb{R}^d$ and $h : \mathbb{R}^d \to \mathbb{R}_+, \boldsymbol{T} : \mathbb{R}^d \to \mathbb{R}^k$ be measurable functions. We define the logarithmic partition function $\alpha_{\boldsymbol{T},h} : \mathbb{R}^k \to \mathbb{R}_+$ as $\alpha(\boldsymbol{\eta}) = \alpha_{\boldsymbol{T},h}(\boldsymbol{\eta}) = \ln(\int \exp(\boldsymbol{\eta}^\top \boldsymbol{T}(\boldsymbol{x}))h(\boldsymbol{x})d\mu(\boldsymbol{x}))$. We also define the range of natural parameters $\mathcal{H}_{\boldsymbol{T},h} = \{\boldsymbol{\eta} \in \mathbb{R}^k | \alpha_{\boldsymbol{T},h}(\boldsymbol{\eta}) < \infty\}$. The exponential family $\mathcal{E}(\boldsymbol{T}, h)$ with sufficient statistics $\boldsymbol{T}$, carrier measure h and natural parameters $\boldsymbol{\eta}$ is the family of distributions $\mathcal{E}(\boldsymbol{T}, h) = \{P_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \mathcal{H}_{\boldsymbol{T},h}\}$ where the probability distribution $P_{\boldsymbol{\eta}}$ has density

$$p_{\boldsymbol{\eta}}(\boldsymbol{x}) = \exp(\boldsymbol{\eta}^\top \boldsymbol{T}(\boldsymbol{x}) - \alpha(\boldsymbol{\eta}))h(\boldsymbol{x}).$$

It is known that the total variation of two distributions in the same exponential family has the following closed form. Suppose there are $\boldsymbol{\eta}, \boldsymbol{\eta}'$, there exists $\boldsymbol{\xi}$ in the line segment of $[\boldsymbol{\eta}, \boldsymbol{\eta}']$ such that

$$d_{\mathrm{TV}}(P_{\boldsymbol{\eta}}, P_{\boldsymbol{\eta}'}) = \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim P_{\boldsymbol{\xi}}} \left[\mathrm{sign}(P_{\boldsymbol{\eta}}(\boldsymbol{x}) - P_{\boldsymbol{\eta}'}(\boldsymbol{x}))(\boldsymbol{\eta} - \boldsymbol{\eta}')^\top \left(\boldsymbol{T}(\boldsymbol{x}) - \mathop{\mathbb{E}}_{\boldsymbol{y}\sim P_{\boldsymbol{\xi}}} [\boldsymbol{T}(\boldsymbol{y})] \right) \right].$$

If we further assume the parameter space $\Theta \subset \mathcal{H}_{\boldsymbol{T},h}$ is bounded and that $D_\Theta = \sup_{\boldsymbol{\eta}\in\Theta} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim P_{\boldsymbol{\eta}}} (\|\boldsymbol{T}(\boldsymbol{x})\|) < \infty$, then

$$\begin{aligned}
d_{\mathrm{TV}}(P_{\boldsymbol{\eta}}, P_{\boldsymbol{\eta}'}) &= \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim P_{\boldsymbol{\xi}}} \left[\mathrm{sign}(P_{\boldsymbol{\eta}}(\boldsymbol{x}) - P_{\boldsymbol{\eta}'}(\boldsymbol{x}))(\boldsymbol{\eta} - \boldsymbol{\eta}')^\top \left(\boldsymbol{T}(\boldsymbol{x}) - \mathop{\mathbb{E}}_{\boldsymbol{y}\sim P_{\boldsymbol{\xi}}} [\boldsymbol{T}(\boldsymbol{y})] \right) \right] . \\
&\le \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim P_{\boldsymbol{\xi}}} \left[\|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \left\| \boldsymbol{T}(\boldsymbol{x}) - \mathop{\mathbb{E}}_{\boldsymbol{y}\sim P_{\boldsymbol{\xi}}} [\boldsymbol{T}(\boldsymbol{y})] \right\| \right] \\
&\le \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim P_{\boldsymbol{\xi}}} \left[\|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \left(\|\boldsymbol{T}(\boldsymbol{x})\| + \left\| \mathop{\mathbb{E}}_{\boldsymbol{y}\sim P_{\boldsymbol{\xi}}} [\boldsymbol{T}(\boldsymbol{y})] \right\| \right) \right] \\
&\le D_\Theta \|\boldsymbol{\eta} - \boldsymbol{\eta}'\|.
\end{aligned}$$

Concrete examples in the exponential family include: Univariate Gaussian, Multivariate Gaussian, Poisson, Centered Laplacian, Bernoulli, Binomial (with fixed number of trials), Multinomial (with fixed number of trials), negative binomial (with fixed number of failures) Rayleigh, Gamma, Beta, chi-squared, exponential, Dirichlet, geometric, etc.

More generally, if the density $P_{\boldsymbol{\theta}}$ is uniformly smooth in the distribution family, which usually holds if Θ is bounded, then the distribution family satisfies Assumption 3, as presented in the following:

Example 7 (Smooth Bounded Distribution Family). Suppose $\mathcal{Z}$ is bounded and $E_1 := \sup_{\boldsymbol{z}\in\mathcal{Z}} \|\boldsymbol{z}\|$, and for all $\boldsymbol{z} \in \mathcal{Z}$, for all $P_{\boldsymbol{\theta}}$, the density $p_{\boldsymbol{\theta}}(\boldsymbol{z})$, as a function of $\boldsymbol{\theta}$ is continuously differentiable with respect to $\boldsymbol{\theta}$. Furthermore,$\sup_{\boldsymbol{z}\in\mathcal{Z}} \sup_{\boldsymbol{\theta}\in\Theta} \|\nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{z})\| =: E_2 < \infty$, then for any $\boldsymbol{z} \in \mathcal{Z}$, $|p_{\boldsymbol{\theta}_1}(\boldsymbol{z}) - p_{\boldsymbol{\theta}_2}(\boldsymbol{z})| \le D_\Theta \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$. This is because, by mean value theorem, there exists $\boldsymbol{\theta}_3$ such that

$$p_{\boldsymbol{\theta}_1}(\boldsymbol{z}) - p_{\boldsymbol{\theta}_2}(\boldsymbol{z}) = \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{z})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_3}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \le \|\nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{w})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_3}\| \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \le E_2 \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

Since $\mathcal{Z}$ is bounded, then we can get an upper bound of the total variation

$$2d_{\mathrm{TV}}(P_{\boldsymbol{\theta}_1}, P_{\boldsymbol{\theta}_2}) = \int_{\boldsymbol{z}\in Z} |p_{\boldsymbol{\theta}_1}(\boldsymbol{z}) - p_{\boldsymbol{\theta}_2}(\boldsymbol{z})| d\boldsymbol{z} \le \int_{\boldsymbol{z}\in Z} E_2 \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| d\boldsymbol{z} \le E_2 E_1^d \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

In this case, it satisfies Assumption 3 with $D_\Theta = \frac{1}{2} E_2 E_1^d$.

Finally, any distribution family that satisfies Assumption 3 can still preserve such property after truncation and normalization.

Example 8 (Truncated Distribution Family). Suppose a distribution family on $\mathcal{Z}$ satisfies Assumption 3. The truncated distribution family $\tilde{p}_{\boldsymbol{\theta}}$ defined on the restricted region $\tilde{\mathcal{Z}}$ is given by $\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{z}) = 1(\boldsymbol{z} \in \tilde{\mathcal{Z}}) p_{\boldsymbol{\theta}}(\boldsymbol{z}) / \int_{\boldsymbol{z}\in\tilde{\mathcal{Z}}} p_{\boldsymbol{\theta}}(z)$. If we assume $\lambda := \sup_{\boldsymbol{\theta}\in\Theta} 1/\int_{\boldsymbol{z}\in\tilde{\mathcal{Z}}} p_{\boldsymbol{\theta}}(z)$, then for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, $d_{\mathrm{TV}}(\tilde{p}_{\boldsymbol{\theta}_1}, \tilde{p}_{\boldsymbol{\theta}_2}) \le 2\lambda d_{\mathrm{TV}}(p_{\boldsymbol{\theta}_1}, p_{\boldsymbol{\theta}_2}) \le 2D\lambda \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$. To see this, define $\lambda_1 = 1/\int_{\boldsymbol{z}\in\tilde{\mathcal{Z}}} p_{\boldsymbol{\theta}_1}(z)$ and $\lambda_2 = 1/\int_{\boldsymbol{z}\in\tilde{\mathcal{Z}}} p_{\boldsymbol{\theta}_2}(z)$ and $\lambda_1 \le \lambda_2 \le \lambda$ without loss of generality.We

define $\mathcal{Z}_1 := \{\boldsymbol{z} \in \tilde{\mathcal{Z}} : \lambda_1 p(\boldsymbol{z}) > \lambda_2 q(\boldsymbol{z})\}$ and $\mathcal{Y}_2 := \{\boldsymbol{z} \in \tilde{\mathcal{Z}} : \lambda_1 p(\boldsymbol{z}) \leq \lambda_2 q(\boldsymbol{z})\}$. The total variation satisfies

$$\begin{aligned}
2d_{\mathrm{TV}}(\tilde{p}_{\boldsymbol{\theta}_1}, \tilde{p}_{\boldsymbol{\theta}_2}) &= \int_{\boldsymbol{z}\in\tilde{\mathcal{Z}}} |\lambda_1 p_{\boldsymbol{\theta}_1}(\boldsymbol{z}) - \lambda_2 p_{\boldsymbol{\theta}_2}(\boldsymbol{z})| d\boldsymbol{z} \\
&= \int_{\boldsymbol{z}\in\mathcal{Z}_1} (\lambda_1 p_{\boldsymbol{\theta}_1}(\boldsymbol{z}) - \lambda_2 p_{\boldsymbol{\theta}_2}(\boldsymbol{z})) d\boldsymbol{z} + \int_{\boldsymbol{z}\in\mathcal{Z}_2} (\lambda_2 p_{\boldsymbol{\theta}_2}(\boldsymbol{z}) - \lambda_1 p_{\boldsymbol{\theta}_1}(\boldsymbol{z})) d\boldsymbol{z} \\
&= 2\int_{\boldsymbol{z}\in\mathcal{Z}_1} (\lambda_1 p_{\boldsymbol{\theta}_1}(\boldsymbol{z}) - \lambda_2 p_{\boldsymbol{\theta}_2}(\boldsymbol{z})) d\boldsymbol{z} \leq 2\int_{\boldsymbol{z}\in\mathcal{Z}_1} (\lambda_2 p_{\boldsymbol{\theta}_1}(\boldsymbol{z}) - \lambda_2 p_{\boldsymbol{\theta}_2}(\boldsymbol{z})) d\boldsymbol{z} \\
&= 2\int_{\boldsymbol{z}\in\mathcal{Z}_1} |\lambda_2 p_{\boldsymbol{\theta}_1}(\boldsymbol{z}) - \lambda_2 p_{\boldsymbol{\theta}_2}(\boldsymbol{z})| d\boldsymbol{z} \leq 2\int_{\boldsymbol{z}\in\mathcal{Z}} |\lambda_2 p_{\boldsymbol{\theta}_1}(\boldsymbol{z}) - \lambda_2 p_{\boldsymbol{\theta}_2}(\boldsymbol{z})| d\boldsymbol{z} \\
&= 4\lambda_2 d_{\mathrm{TV}}(p_{\boldsymbol{\theta}_1}, p_{\boldsymbol{\theta}_2}) \leq 4\lambda d_{\mathrm{TV}}(p_{\boldsymbol{\theta}_1}, p_{\boldsymbol{\theta}_2}).
\end{aligned}$$

In most cases, the distribution family may be truncated to the concentrated region so that λ will be slightly larger than 1.

A.2 Proof of Lemma 1 and Theorem 1

Since the expected cost function $v(\boldsymbol{w}, Q)$ (resp. $v(\boldsymbol{w}, \boldsymbol{\theta})$) is strongly convex for any $Q \in \mathcal{P}$ (resp. $\boldsymbol{\theta} \in \Theta$), we have the following first order property, as presented in Lemma 2.

Lemma 2 (Theorem 5.24 in Beck (2017))**.** *Under Assumption 2.A, for any $Q \in \mathcal{P}$, for any $\boldsymbol{w} \in \Omega$ and for any $\boldsymbol{\theta} \in \Theta$*

$$\begin{aligned}
v(\boldsymbol{w}, Q) - v(\boldsymbol{w}_Q, Q) &\geq \nabla_{\boldsymbol{w}} v(\boldsymbol{w}_Q, Q)(\boldsymbol{w} - \boldsymbol{w}_Q) + \frac{\rho_c}{2}\|\boldsymbol{w} - \boldsymbol{w}_Q\|^2 \geq \frac{\rho_c}{2}\|\boldsymbol{w} - \boldsymbol{w}_Q\|^2. \\
v(\boldsymbol{w}, \boldsymbol{\theta}) - v(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{\theta}) &\geq \nabla_{\boldsymbol{w}} v(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{\theta})(\boldsymbol{w} - \boldsymbol{w}_{\boldsymbol{\theta}}) + \frac{\rho_c}{2}\|\boldsymbol{w} - \boldsymbol{w}_{\boldsymbol{\theta}}\|^2 \geq \frac{\rho_c}{2}\|\boldsymbol{w} - \boldsymbol{w}_{\boldsymbol{\theta}}\|^2.
\end{aligned}$$

To establish the generalization bound for IEO, we rely on both single-variate and multi-variate Rademacher complexity. Given a distribution family $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, we can apply generalization bounds that directly use the Rademacher complexity of the function class $c \circ \Theta$. Given a sample $\{\boldsymbol{z}_i\}_{i=1}^n$, the *empirical Rademacher complexity* $\hat{\mathfrak{R}}_n^{\mathrm{IEO}}(\Theta)$ of the function class $c \circ \Theta$ is defined by

$$\hat{\mathfrak{R}}_n^{\mathrm{IEO}}(\Theta) := \mathbb{E}_\sigma \left[\sup_{\boldsymbol{\theta}\in\Theta} \frac{1}{n} \sum_{i=1}^n \sigma_i c(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z}_i) \right],$$

where σ_i are independent Rademacher random variables, i.e. $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}$ for all $i \in [n]$. The *expected Rademacher complexity* $\mathfrak{R}_n^{\mathrm{IEO}}(\Theta)$ is then defined as the expectation of $\hat{\mathfrak{R}}_n^{\mathrm{IEO}}(\Theta)$ with respect to the i.i.d. sample $\{\boldsymbol{z}_i\}_{i=1}^n$ drawn from the distribution P:

$$\mathfrak{R}_n^{\mathrm{IEO}}(\Theta) := \mathbb{E}_{\boldsymbol{z}\sim P} \left[\hat{\mathfrak{R}}_n^{\mathrm{IEO}}(\Theta) \right].$$

By directly using $\mathfrak{R}_n^{\mathrm{IEO}}(\Theta)$, we have the following theorem is an adaptation of the classical generalization bounds based on Rademacher complexity due to Mohri et al. (2018) to our setting.

Lemma 3 (Theorem 3.3 in Mohri et al. (2018))**.** *Under Assumption 2.C, for any $\tilde{\delta} > 0$, with probability at least $1 - \tilde{\delta}$ over an i.i.d. sample of $\{\boldsymbol{z}_i\}_{i=1}^n$ drawn from P, the following holds for all $\boldsymbol{\theta} \in \Theta$:*

$$v_0(\boldsymbol{w}_{\boldsymbol{\theta}}) \leq \hat{v}_0(\boldsymbol{w}_{\boldsymbol{\theta}}) + 2\mathfrak{R}_n^{\mathrm{IEO}}(\Theta) + B_c\sqrt{\frac{\log(1/\tilde{\delta})}{2n}}.$$

Equipped with Lemma 3, which holds for all $\boldsymbol{\theta} \in \Theta$, we have the following guarantee on the excess risk of the empirical minimizer of IEO $\hat{\boldsymbol{w}}^{\mathrm{IEO}}$. Corollary 3 below is the combination of Lemma 3 and the Hoeffding inequality.

Corollary 3. *Under Assumption 2.C, for any $\tilde{\delta} > 0$, with probability at least $1 - \tilde{\delta}$ over an i.i.d. sample of $\{\boldsymbol{z}_i\}_{i=1}^n$ drawn from P, the decision $\hat{\boldsymbol{w}}^{\mathrm{IEO}}$ returned by the IEO satisfies:*

$$R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \leq R(\boldsymbol{w}_{\boldsymbol{\theta}^*}) + 2\mathfrak{R}_n^{\mathrm{IEO}}(\Theta) + 2B_c\sqrt{\frac{\log(2/\tilde{\delta})}{2n}}.$$

Proof of Corollary 3. By the definition of the regret of $\hat{\boldsymbol{w}}^{\mathrm{IEO}}$,

$$\begin{aligned} R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) &= v_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \\ &= v_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - \hat{v}_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) + \hat{v}_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - \hat{v_0}(\boldsymbol{w}_{\boldsymbol{\theta}^*}) + \hat{v_0}(\boldsymbol{w}_{\boldsymbol{\theta}^*}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \\ &\leq v_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - \hat{v}_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) + \hat{v_0}(\boldsymbol{w}_{\boldsymbol{\theta}^*}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}). \end{aligned}$$

The inequality follows from the fact that $\hat{\boldsymbol{\theta}}^{\mathrm{IEO}}$ is the empirical minimizer of $\hat{v}_0(\boldsymbol{w}_{\boldsymbol{\theta}})$. From Lemma 3 we know that

$$v_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \leq \hat{v}_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) + 2\mathfrak{R}_n^{\mathrm{IEO}}(\Theta) + B_c\sqrt{\frac{\log(2/\tilde{\delta})}{2n}}$$

with probability at least $1 - \tilde{\delta}/2$. From Hoeffding's inequality, we have that

$$\hat{v}_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \leq B_c\sqrt{\frac{\log(2/\tilde{\delta})}{2n}}$$

with probability at least $1 - \tilde{\delta}/2$. Thus, with probability at least $1 - \tilde{\delta}$ we can combine the three previous inequalities and obtain the desired result. □

Next, we introduce the multivariate Rademacher complexity as an extension of the regular Rademacher complexity to a class of vector-valued functions. Given a fixed sample $\{(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z}_i)\}_{i=1}^n$, we define the *empirical multivariate Rademacher complexity* of Θ as

$$\hat{\mathfrak{R}}_{\boldsymbol{w}}^n(\Theta) := \mathbb{E}_\sigma\left[\sup_{\boldsymbol{\theta}\in\Theta}\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^p \sigma_{ij}^\top \boldsymbol{w}_{\boldsymbol{\theta}j}\right] = \mathbb{E}_\sigma\left[\sup_{\boldsymbol{\theta}\in\Theta}\frac{1}{n}\sum_{i=1}^n \boldsymbol{\sigma}_i^\top \boldsymbol{w}_{\boldsymbol{\theta}}\right].$$

In this case, $\hat{\mathfrak{R}}_{\boldsymbol{w}}^n(\Theta)$ is deterministic and it always holds that the *expected multivariate Rademacher complexity* $\mathfrak{R}_{\boldsymbol{w}}^n(\Theta) := \mathop{\mathbb{E}}_{\boldsymbol{z}\sim P}\hat{\mathfrak{R}}_{\boldsymbol{w}}^n(\Theta) \equiv \hat{\mathfrak{R}}_{\boldsymbol{w}}^n(\Theta)$. A famous result of the multivariate Rademacher complexity is the *vector contraction inequalities*. Let $\Phi_i : \mathbb{R}^p \to \mathbb{R}$ for $i \in [n]$ be a collection of L-Lipschitz functions with respect to the give norm $\|\cdot\|$ defined on $\mathbb{R}^p$:

$$|\Phi_i(\boldsymbol{u}) - \Phi_i(\boldsymbol{v})| \leq L\|\boldsymbol{u} - \boldsymbol{v}\| \quad \text{for all } \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d.$$

A vector contraction inequality (Maurer, 2016) takes the form:

$$\mathbb{E}_\sigma\left[\sup_{\boldsymbol{\theta}\in\Theta}\frac{1}{n}\sum_{i=1}^n \sigma_i\Phi_i(\boldsymbol{w}_{\boldsymbol{\theta}})\right] \leq CL\cdot\mathbb{E}_\sigma\left[\sup_{\boldsymbol{\theta}\in\Theta}\frac{1}{n}\sum_{i=1}^n \boldsymbol{\sigma}_i^\top \boldsymbol{w}_{\boldsymbol{\theta}}\right] = CL\cdot\hat{\mathfrak{R}}_{\boldsymbol{w}}^n(\Theta),$$

where C is a constant. In our setting, $\Phi_i(\cdot) = c(\cdot, \boldsymbol{z}_i)$ so $\Phi_i(\cdot)$ is L-Lipschitz for all $i \in [n]$ by Assumption 2.B. If $\|\cdot\|$ is ℓ_2 norm, we can apply an elegant vector contraction inequality due to Maurer (2016), which exactly takes the form $C = \sqrt{2}$. We have the following relations between $\mathfrak{R}_n^{\mathrm{IEO}}(\Theta)$ and $\mathfrak{R}_{\boldsymbol{w}}^n(\Theta)$.

Lemma 4 (Corollary 4 in Maurer (2016)). *Under Assumption 2.B,* $\mathfrak{R}_n^{\mathrm{IEO}}(\Theta) \leq \sqrt{2}L_c\mathfrak{R}_{\boldsymbol{w}}^n(\Theta)$.

We are now going to prove a Lipschitz property of the optimization oracle $\boldsymbol{w}_{\boldsymbol{\theta}}$ with respect to the total variation distances of two distributions. We will prove a more general result. Let $\boldsymbol{w}_Q$ denote the optimization oracle, i.e., $\boldsymbol{w}_Q := \mathrm{argmin}_{\boldsymbol{w}\in\Omega}\mathop{\mathbb{E}}_{\boldsymbol{z}\sim Q}[c(\boldsymbol{w}, \boldsymbol{z})]$. We show that $\boldsymbol{w}_{\boldsymbol{\theta}}$ is "Lipschitz" with respect to $\boldsymbol{\theta}$ in some sense.

Lemma 5. *Under Assumption 2.A and 2.B, for any two distributions $Q_1, Q_2 \in \mathcal{P}$,*

1. *If they have density f_{Q_1} and f_{Q_2}, then $\|\boldsymbol{w}_{\boldsymbol{Q}_1} - \boldsymbol{w}_{Q_2}\| \leq \frac{2L_c}{\rho_c}d_{\mathrm{TV}}(Q_1, Q_2)$.*
2. *If they have the same finite support $\{\boldsymbol{z}_k\}_{k=1}^K$, with probability mass function $\boldsymbol{q}^1, \boldsymbol{q}^2 \in \Delta^{K-1}$, then $\|\boldsymbol{w}_{\boldsymbol{Q}_1} - \boldsymbol{w}_{Q_2}\| \leq \frac{2L_c}{\rho_c}d_{\mathrm{TV}}(Q_1, Q_2)$.*

3. If the parametric family $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ further satisfies Assumption 3, then $\|\boldsymbol{w}_{\boldsymbol{\theta}_1} - \boldsymbol{w}_{\boldsymbol{\theta}_2}\| \leq \frac{2L_c D_\Theta}{\rho_c}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$.

Proof of Lemma 5. By the strong convexity of Assumption 2.A, for $Q_1, Q_2 \in \mathcal{P}$,

$$v(\boldsymbol{w}_{Q_2}, Q_1) - v(\boldsymbol{w}_{Q_1}, Q_1) \geq \frac{\rho_c}{2}\|\boldsymbol{w} - \boldsymbol{w}_{Q_1}\|^2$$
$$v(\boldsymbol{w}_{Q_1}, Q_2) - v(\boldsymbol{w}_{Q_2}, Q_2) \geq \frac{\rho_c}{2}\|\boldsymbol{w} - \boldsymbol{w}_{Q_2}\|^2.$$

In the first scenario,

$$\begin{aligned}
\rho_c\|\boldsymbol{w}_{Q_2} - \boldsymbol{w}_{Q_1}\|^2 &\leq v(\boldsymbol{w}_{Q_2}, Q_1) - v(\boldsymbol{w}_{Q_1}, Q_1) + v(\boldsymbol{w}_{Q_1}, Q_2) - v(\boldsymbol{w}_{Q_2}, Q_2) \\
&= \int c(\boldsymbol{w}_{Q_2}, \boldsymbol{z})dQ_1 - c(\boldsymbol{w}_{Q_1}, \boldsymbol{z})dQ_1 + c(\boldsymbol{w}_{Q_1}, \boldsymbol{z})dQ_2 - c(\boldsymbol{w}_{Q_2}, \boldsymbol{z})dQ_2 \\
&= \int \left(c(\boldsymbol{w}_{Q_2}, \boldsymbol{z}) - c(\boldsymbol{w}_{Q_1}, \boldsymbol{z})\right)(dQ_1 - dQ_2) \\
&= \int_{\boldsymbol{z}\in\mathcal{Z}} (c(\boldsymbol{w}_{Q_2}, \boldsymbol{z}) - c(\boldsymbol{w}_{\boldsymbol{\theta}_1}, \boldsymbol{z}))(f_{Q_1}(\boldsymbol{z}) - f_{Q_2}(\boldsymbol{z}))d\boldsymbol{z} \\
&\leq \int_{\boldsymbol{z}\in\mathcal{Z}} |c(\boldsymbol{w}_{Q_2}, \boldsymbol{z}) - c(\boldsymbol{w}_{Q_1}, \boldsymbol{z})||f_{Q_1}(\boldsymbol{z}) - f_{Q_2}(\boldsymbol{z})|d\boldsymbol{z} \\
&\leq L\|\boldsymbol{w}_{Q_2} - \boldsymbol{w}_{Q_1}\| \int_{\boldsymbol{z}\in\mathcal{Z}} |f_{Q_1}(\boldsymbol{z}) - f_{Q_2}(\boldsymbol{z})|d\boldsymbol{z} \\
&= 2L\|\boldsymbol{w}_{Q_2} - \boldsymbol{w}_{Q_1}\| d_{\mathrm{TV}}(Q_1, Q_2).
\end{aligned}$$

In the second scenario,

$$\begin{aligned}
\rho_c\|\boldsymbol{w}_{Q_2} - \boldsymbol{w}_{Q_1}\|^2 &\leq v(\boldsymbol{w}_{Q_2}, Q_1) - v(\boldsymbol{w}_{Q_1}, Q_1) + v(\boldsymbol{w}_{Q_1}, Q_2) - v(\boldsymbol{w}_{Q_2}, Q_2) \\
&= \sum_{k=1}^K q_k^1 c(\boldsymbol{w}_{Q_2}, \boldsymbol{z}_k) - \sum_{k=1}^K q_k^1 c(\boldsymbol{w}_{Q_1}, \boldsymbol{z}_k) + \sum_{k=1}^K q_k^2 c(\boldsymbol{w}_{Q_1}, \boldsymbol{z}_k) - \sum_{k=1}^K q_k^2 c(\boldsymbol{w}_{Q_2}, \boldsymbol{z}_k) \\
&= \sum_{k=1}^K (q_k^1 - q_k^2)\left(c(\boldsymbol{w}_{Q_2}, \boldsymbol{z}_k) - c(\boldsymbol{w}_{Q_1}, \boldsymbol{z}_k)\right) \\
&\leq \sum_{k=1}^K |q_k^1 - q_k^2|\left|\left(c(\boldsymbol{w}_{Q_2}, \boldsymbol{z}_k) - c(\boldsymbol{w}_{Q_1}, \boldsymbol{z}_k)\right|\right. \\
&\leq \sum_{k=1}^K |q_k^1 - q_k^2| L_c\|\boldsymbol{w}_{Q_2} - \boldsymbol{w}_{Q_1}\| \\
&= 2L_c\|\boldsymbol{w}_{Q_2} - \boldsymbol{w}_{Q_1}\| d_{\mathrm{TV}}(Q_1, Q_2).
\end{aligned}$$

In particular, the result holds for a distribution family $\mathcal{P}_\Theta$ that satisfies either scenario in Lemma 5: for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta, \|\boldsymbol{w}_{\boldsymbol{\theta}_1} - \boldsymbol{w}_{\boldsymbol{\theta}_2}\| \leq \frac{2L_c}{\rho_c} d_{\mathrm{TV}}(P_{\boldsymbol{\theta}_1}, P_{\boldsymbol{\theta}_2})$. If the parametric family further satisfies Assumption 3, then for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta, \|\boldsymbol{w}_{\boldsymbol{\theta}_1} - \boldsymbol{w}_{\boldsymbol{\theta}_2}\| \leq \frac{2L_c}{\rho_c} d_{\mathrm{TV}}(P_{\boldsymbol{\theta}_1}, P_{\boldsymbol{\theta}_2}) \leq \frac{2L_c D_\Theta}{\rho_c}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$. □

In the remainder of this section, we will leverage the *covering number* argument to bound $\mathfrak{R}^n_{\boldsymbol{w}}(\Theta)$. We first review some basic ingredients of covering number.

Definition 6 (Covering)**.** Consider the metric space $(\mathbb{R}^q, \|\cdot\|)$ and $\Theta \subset \mathbb{R}^q$.

1. We say $\{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N\} \subset \mathbb{R}^q$ is an *ε-covering* of Θ if for all $\boldsymbol{\theta} \in \Theta$, there exists $i \in [N]$ such that $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}\| < \varepsilon$.

2. The *ε-covering number* of Θ is $\mathbb{N}(\varepsilon, \Theta, \|\cdot\|) := \inf\{N \in \mathbb{Z}^+ : \exists \text{ an } \varepsilon\text{-covering } \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N \text{ of } \Theta\}$.

Lemma 6. *When Θ is bounded, we denote $E_\Theta := \sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{\theta}\|$. For $\varepsilon > 0$, the covering number of Θ satisfies*

$$\mathcal{N}(\varepsilon, \Theta, \|\cdot\|) \leq E_\Theta^q \left(1 + \frac{2}{\varepsilon}\right)^q.$$

Proof of Lemma 6. By Lemma 6.27 in Mohri et al. (2018), for any $\varepsilon > 0$

$$\left(\frac{1}{\varepsilon}\right)^q \leq \mathcal{N}(\varepsilon, B(0,1), \|\cdot\|) \leq \left(1+\frac{2}{\varepsilon}\right)^q.$$

For general bounded set Θ, it satisfies

$$\left(\frac{1}{\varepsilon}\right)^q \frac{\text{vol}(\Theta)}{\text{vol}B(0,1)} \leq \mathcal{N}(\varepsilon, \Theta, \|\cdot\|) \leq \left(1+\frac{2}{\varepsilon}\right)^q \frac{\text{vol}(\Theta)}{\text{vol}B(0,1)} \leq \max_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{\theta}\|^q \left(1+\frac{2}{\varepsilon}\right)^q = E_\Theta^q \left(1+\frac{2}{\varepsilon}\right)^q.$$

□

We are now ready to provide an explicit upper bound of $\mathfrak{R}_{\boldsymbol{w}}^n(\Theta)$:

Lemma 7. *There exists a universal constant C_{abs} such that the multivariate Rademacher complexity $\mathfrak{R}_{\boldsymbol{w}}^n(\Theta)$ satisfies*

$$\mathfrak{R}_{\boldsymbol{w}}^n(\Theta) \leq \frac{2L_c C_{\text{abs}} D_\Theta E_\Theta}{\rho_c}\sqrt{\frac{q}{n}}.$$

Proof of Lemma 7. For simplicity, we denote $L' := 2L_c D_\Theta/\rho_c$, $X_{\boldsymbol{\theta}} := \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\sigma}_i^\top \boldsymbol{w}_{\boldsymbol{\theta}}$. It is not hard to see that since for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$,

$$X_{\boldsymbol{\theta}_1} - X_{\boldsymbol{\theta}_2} = \sum_{i=1}^n \frac{1}{\sqrt{n}}\boldsymbol{\sigma}_i^\top(\boldsymbol{w}_{\boldsymbol{\theta}_1} - \boldsymbol{w}_{\boldsymbol{\theta}_2}) = \sum_{i=1}^n\sum_{j=1}^p \sigma_{ij}\frac{1}{\sqrt{n}}(\boldsymbol{w}_{\boldsymbol{\theta}_1 j} - \boldsymbol{w}_{\boldsymbol{\theta}_2 j}).$$

By Lemma 3.12 in Sen (2018), $X_{\boldsymbol{\theta}_1} - X_{\boldsymbol{\theta}_2}$ is $\|\boldsymbol{w}_{\boldsymbol{\theta}_1} - \boldsymbol{w}_{\boldsymbol{\theta}_2}\|^2$ sub-Gaussian, and thus $L'^2\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2$ sub-Gaussian. In conclusion, $\{X_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\Theta}$ is a sub-Gaussian process. By Theorem 4.3 in Sen (2018), for any fixed $\boldsymbol{\theta}_0 \in \Theta$,

$$\begin{aligned}
\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} X_{\boldsymbol{\theta}} &= \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}(X_{\boldsymbol{\theta}} - X_{\boldsymbol{\theta}_0}) \leq \mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta}|X_{\boldsymbol{\theta}} - X_{\boldsymbol{\theta}_0}| \lesssim \int_0^\infty \sqrt{\log\mathcal{N}(\varepsilon, \Theta, L'\|\cdot\|)}d\varepsilon \\
&= \int_0^{L'E_\Theta} \sqrt{\log\mathcal{N}(\varepsilon, \Theta, L'\|\cdot\|)}d\varepsilon = \int_0^{L'E_\Theta} \sqrt{\log\mathcal{N}(\frac{\varepsilon}{L'}, \Theta, \|\cdot\|)}d\varepsilon \\
&= L'\int_0^{E_\Theta} \sqrt{\log\mathcal{N}(\varepsilon, \Theta, \|\cdot\|)}d\varepsilon \leq L'\sqrt{q}\int_0^{E_\Theta} \sqrt{\log(1+\frac{2}{\varepsilon})}d\varepsilon \\
&= L'\sqrt{q}\left(\int_0^{E\wedge 1} \sqrt{\log(1+\frac{2}{\varepsilon})}d\varepsilon + \int_{E_\Theta\wedge 1}^{E_\Theta} \sqrt{\log(1+\frac{2}{\varepsilon})}d\varepsilon\right) \\
&\leq L'\sqrt{q}\left(\underbrace{\int_0^1 \sqrt{\log(1+\frac{2}{\varepsilon})}d\varepsilon}_{\text{a universal constant}} + \underbrace{\int_{E_\Theta\wedge 1}^{E_\Theta} \sqrt{\log(1+\frac{2}{\varepsilon})}d\varepsilon}_{\leq\sqrt{\log 3}E_\Theta}\right) \\
&\leq C_{\text{abs}}L'\sqrt{q}E_\Theta.
\end{aligned}$$

The multivariate Rademacher complexity thus satisfies

$$\mathfrak{R}_{\boldsymbol{w}}^n(\Theta) = \frac{1}{\sqrt{n}}\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} X_{\boldsymbol{\theta}} \leq C_{\text{abs}}L'E_\Theta\sqrt{\frac{q}{n}} = \frac{2L_c C_{\text{abs}} D_\Theta E_\Theta}{\rho_c}\sqrt{\frac{q}{n}}.$$

□

Finally, we can prove Lemma 1 and Theorem 1:

Proof of Lemma 1. By previous discussion, we have

$$\begin{aligned} v_0(\boldsymbol{w}_{\boldsymbol{\theta}}) &\le \hat{v}_0(\boldsymbol{w}_{\boldsymbol{\theta}}) + 2\mathfrak{R}_n^{\mathrm{IEO}}(\Theta) + B_c\sqrt{\frac{\log(1/\tilde{\delta})}{2n}} \\ &\le \hat{v}_0(\boldsymbol{w}_{\boldsymbol{\theta}}) + \frac{4\sqrt{2}L_c^2 C D_\Theta E_\Theta}{\rho_c}\sqrt{\frac{q}{n}} + B_c\sqrt{\frac{\log(1/\tilde{\delta})}{2n}}. \end{aligned}$$

□

Proof of Theorem 1. By Corollary 3, we have

$$\begin{aligned} R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) &\le R(\boldsymbol{w}_{\boldsymbol{\theta}^*}) + 2\mathfrak{R}_n^{\mathrm{IEO}}(\Theta) + 2B_c\sqrt{\frac{\log(2/\tilde{\delta})}{2n}} \\ &\le R(\boldsymbol{w}_{\boldsymbol{\theta}^*}) + 2\sqrt{2}L_c\mathfrak{R}_{\boldsymbol{w}}^n(\Theta) + 2B_c\sqrt{\frac{\log(2/\tilde{\delta})}{2n}} \\ &\le R(\boldsymbol{w}_{\boldsymbol{\theta}^*}) + \frac{4\sqrt{2}L_c^2 C_{\mathrm{abs}} D_\Theta E_\Theta}{\rho_c}\sqrt{\frac{q}{n}} + 2B_c\sqrt{\frac{\log(2/\tilde{\delta})}{2n}}. \end{aligned}$$

□

B Further Details and Proofs in Section 4

B.1 Assumptions and Supporting Theorems

We first list out standard assumptions that lead to the consistency of the ETO and IEO, which are direct consequences of asymptotic statistical theory (e.g., Theorem 5.7 in Van der Vaart (2000).)

Assumption 6.A (Consistency conditions for ETO)**.** *Suppose that:*

1. $\sup_{\boldsymbol{\theta}\in\Theta} |\frac{1}{n}\sum_{i=1}^n \log p_{\boldsymbol{\theta}}(\boldsymbol{z}_i) - \mathbb{E}_P[\log p_{\boldsymbol{\theta}}(\boldsymbol{z})]| \xrightarrow{P} 0.$
2. $\forall \varepsilon > 0$, $\sup_{\boldsymbol{\theta}\in\Theta:\|\boldsymbol{\theta}-\boldsymbol{\theta}^{\mathrm{KL}}\|\ge\varepsilon} \mathbb{E}_P[\log p_{\boldsymbol{\theta}}(\boldsymbol{z})] < \mathbb{E}_P[\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})]$.

Assumption 6.B (Consistency conditions for IEO)**.** *Suppose that:*

1. $\sup_{\boldsymbol{\theta}\in\Theta} |\hat{v}_0(\boldsymbol{w}_{\boldsymbol{\theta}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}})| \xrightarrow{P} 0.$
2. $\forall \varepsilon > 0$, $\inf_{\boldsymbol{\theta}\in\Theta:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\ge\varepsilon} v_0(\boldsymbol{w}_{\boldsymbol{\theta}}) > v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})$.

In each of Assumptions 6.A and 6.B, the first part is a uniform law of large numbers that are satisfied via Glivenko-Cantelli conditions. The second part ensure the uniqueness of $\boldsymbol{\theta}^{\mathrm{KL}}$ or $\boldsymbol{\theta}^*$.

We also list out standard conditions and asymptotic normality guarantees for IEO and ETO which follow directly from established results in asymptotic statistical theory (e.g. Shao and Zhang (2022).)

Assumption 7.A (Regularity conditions for ETO)**.** *The function* $\log p_{\boldsymbol{\theta}}(\cdot)$ *is twice differentiable with respect to* $\boldsymbol{\theta}$ *and there exist constant* $\mu^{\mathrm{ETO}} > 0, c_1^{\mathrm{ETO}} > 0, c_2^{\mathrm{ETO}} > 0$ *and two nonnegative functions* $K_1^{\mathrm{ETO}}, K_2^{\mathrm{ETO}} : \mathcal{Z} \to \mathbb{R}$ *with* $\|K_1^{\mathrm{ETO}}(\boldsymbol{z})\|_9 \le c_1^{\mathrm{ETO}}$ *and* $\|K_2^{\mathrm{ETO}}(\boldsymbol{z})\|_4 \le c_2^{\mathrm{ETO}}$ *such that for any* $\boldsymbol{\theta} \in \Theta$*,*

$$\begin{aligned} \mathbb{E}_P(\log p_{\boldsymbol{\theta}}(\boldsymbol{z})) - \mathbb{E}_P(\log P^{\mathrm{KL}}(\boldsymbol{z})) &\ge \mu^{\mathrm{ETO}}\|\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{KL}}\|^2, \\ |\log p_{\boldsymbol{\theta}}(\boldsymbol{z}) - \log P^{\mathrm{KL}}(\boldsymbol{z})| &\le K_1^{\mathrm{ETO}}(\boldsymbol{z})\|\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{KL}}\|, \quad \forall \boldsymbol{z} \in \mathcal{Z}, \\ \|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{z}) - \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \log P^{\mathrm{KL}}(\boldsymbol{z})\| &\le K_2^{\mathrm{ETO}}(\boldsymbol{z})\|\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{KL}}\|, \quad \forall \boldsymbol{z} \in \mathcal{Z}. \end{aligned}$$

Moreover, there exists a constant $c_3^{\mathrm{ETO}} \geq 0$ *and a nonnegative function* $K_3^{\mathrm{ETO}} : \mathcal{Z} \to \mathbb{R}$ *such that for any* $\boldsymbol{z} \in \mathcal{Z}$*:*

$$\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{z}) \leq K_3^{\mathrm{ETO}}(\boldsymbol{z})\boldsymbol{I} \text{ and } \|K_3^{\mathrm{ETO}}(\boldsymbol{z})\|_4 \leq c_3^{\mathrm{ETO}}.$$

Let $\Sigma^{\mathrm{ETO}} = \mathbb{E}[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})^\top \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})]$ *and* $V^{\mathrm{ETO}} = \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbb{E}[\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})]$*. Assume that there exist constants* $\lambda_1^{\mathrm{ETO}} > 0$ *and* $\lambda_2^{\mathrm{ETO}} > 0$ *such that* $\lambda_{\min}(\Sigma^{\mathrm{ETO}}) \geq \lambda_1^{\mathrm{ETO}}$ *and* $\lambda_{\min}(V^{\mathrm{ETO}}) \geq \lambda_2^{\mathrm{ETO}}$*. Moreover, assume that there exists a constant* $c_4^{\mathrm{ETO}} > 0$ *such that*

$$\|\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})\|_4 \leq c_4^{\mathrm{ETO}} \sqrt{q}.$$

Assumption 7.B (Regularity conditions for IEO)**.** *The function* $c(\boldsymbol{w}_{\boldsymbol{\theta}}, \cdot)$ *is twice differentiable with respect to* $\boldsymbol{\theta}$ *and there exist constant* $\mu^{\mathrm{IEO}} > 0, c_1^{\mathrm{IEO}} > 0, c_2^{\mathrm{IEO}} > 0$ *and two nonnegative functions* $K_1^{\mathrm{IEO}}, K_2^{\mathrm{IEO}} : \mathcal{Z} \to \mathbb{R}$ *with* $\|K_1^{\mathrm{IEO}}(\boldsymbol{z})\|_9 \leq c_1^{\mathrm{IEO}}$ *and* $\|K_2^{\mathrm{IEO}}(\boldsymbol{z})\|_4 \leq c_2^{\mathrm{IEO}}$ *such that for any* $\boldsymbol{\theta} \in \Theta$*,*

$$\begin{aligned} v_0(\boldsymbol{w}_{\boldsymbol{\theta}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) &\geq \mu^{\mathrm{IEO}} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2, \\ |c(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z}) - c(\boldsymbol{w}_{\boldsymbol{\theta}}^*, \boldsymbol{z})| &\leq K_1^{\mathrm{IEO}}(\boldsymbol{z}) \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|, \quad \forall \boldsymbol{z} \in \mathcal{Z}, \\ \|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z}) - \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}^*}, \boldsymbol{z})\| &\leq K_2^{\mathrm{IEO}}(\boldsymbol{z}) \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|, \quad \forall \boldsymbol{z} \in \mathcal{Z}. \end{aligned}$$

Moreover, there exists a constant $c_3^{\mathrm{IEO}} \geq 0$ *and a nonnegative function* $K_3^{\mathrm{IEO}} : \mathcal{Z} \to \mathbb{R}$ *such that for any* $\boldsymbol{z} \in \mathcal{Z}$*:*

$$\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}^*}, \boldsymbol{z}) \leq K_3^{\mathrm{IEO}}(\boldsymbol{z})\boldsymbol{I} \text{ and } \|K_3^{\mathrm{IEO}}(\boldsymbol{z})\|_4 \leq c_3^{\mathrm{IEO}}.$$

Let $\Sigma^{\mathrm{IEO}} = \mathbb{E}[\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}^*}, \boldsymbol{z})^\top \nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}^*}, \boldsymbol{z})]$ *and* $V^{\mathrm{IEO}} = \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbb{E}[c(\boldsymbol{w}_{\boldsymbol{\theta}^*}, \boldsymbol{z})]$*. Assume that there exist constants* $\lambda_1^{\mathrm{IEO}} > 0$ *and* $\lambda_2^{\mathrm{IEO}} > 0$ *such that* $\lambda_{\min}(\Sigma^{\mathrm{IEO}}) \geq \lambda_1^{\mathrm{IEO}}$ *and* $\lambda_{\min}(V^{\mathrm{IEO}}) \geq \lambda_2^{\mathrm{IEO}}$*. Moreover, assume that there exists a constant* $c_4^{\mathrm{IEO}} > 0$ *such that*

$$\|\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}^*}, \boldsymbol{z})\|_4 \leq c_4^{\mathrm{IEO}} \sqrt{q}.$$

We give some remarks on the twice differentiability in the assumptions, e.g. in Assumption 7.B. (1) The twice differentiability of $c(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z})$ holds for many applications. However, even if sometimes $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z})$ does not exist, Berry-Esseen bounds can still be established under some Lipschitz continuous gradient assumptions (Theorem 3.2 in Shao and Zhang (2022)). (2) Alternatively, we can smooth the original loss function by some approximated surrogate functions that are twice differentiable and satisfy Assumption 7.B. The surrogate functions can approximate the original function arbitrarily well by Lemma C.1 in Wang et al. (2018).

By Theorem 5.7 in Van der Vaart (2000) and Theorem 5.23 in Van der Vaart (2000), which are standard results of M-estimation theory, we have the consistency and asymptotic normality result for IEO, and ETO as follows.

Proposition 3.A (Consistency and asymptotic normality for ETO, Propositions 1.B and 2.B in Elmachtoub et al. (2023))**.** *Under Assumption 6.A,* $\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} \xrightarrow{P} \boldsymbol{\theta}^{\mathrm{KL}}$*. Under Assumption 6.A and 7.A,* $\sqrt{n}(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}})$ *converges in distribution to* $\mathbb{N}_1^{\mathrm{ETO}}$*, which is a normal distribution with mean zero and covariance matrix*

$$\mathrm{var}_P(\mathbb{N}_1^{\mathrm{ETO}}) := (\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbb{E}_P[\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})])^{-1} \mathrm{var}_P(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z}))(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbb{E}_P[\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})])^{-1} \tag{5}$$

where $\mathrm{var}_P(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z}))$ *is the covariance matrix of* $\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})$ *under* P*. Moreover, when* $\boldsymbol{\theta}^{\mathrm{KL}}$ *corresponds to the ground-truth* P*, i.e.,* $P_{\boldsymbol{\theta}^{\mathrm{KL}}} = P$*, the covariance matrix* (5) *is simplified to the inverse Fisher information* $\mathcal{I}_{\boldsymbol{\theta}^{\mathrm{KL}}}^{-1}$*, that is,*

$$(5) = \mathcal{I}_{\boldsymbol{\theta}^{\mathrm{KL}}}^{-1} = \left(\mathbb{E}_P[(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z}))^\top \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})]\right)^{-1}.$$

Proposition 3.B (Consistency and asymptotic normality for IEO, Propositions 1.C and 2.C in Elmachtoub et al. (2023))**.** *Under Assumption 6.B,* $\hat{\boldsymbol{\theta}}^{\mathrm{IEO}} \xrightarrow{P} \boldsymbol{\theta}^*$*. Under Assumption 6.B and 7.B,* $\sqrt{n}(\hat{\boldsymbol{\theta}}^{\mathrm{IEO}} - \boldsymbol{\theta}^*)$ *converges in distribution to* $\mathbb{N}_1^{\mathrm{IEO}}$*, which is a normal distribution with mean zero and covariance matrix*

$$\mathrm{var}_P(\mathbb{N}_1^{\mathrm{IEO}}) := \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})^{-1} \mathrm{var}_P(\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}^*}, \boldsymbol{z})) \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})^{-1}$$

where $\mathrm{var}_P(\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}^*}, \boldsymbol{z}))$ *is the covariance matrix of the cost gradient* $\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}^*}, \boldsymbol{z})$ *under* P*.*

We list out the following assumption adopted from Elmachtoub et al. (2023).

Assumption 8 (Smoothness and gradient-expectation interchangeability)**.** *Suppose that:*

1. *$v(\boldsymbol{w}, \boldsymbol{\theta})$ is twice differentiable with respect to $(\boldsymbol{w}, \boldsymbol{\theta})$.*
2. *The optimal solution $\boldsymbol{w}_{\boldsymbol{\theta}}$ to the oracle problem (2) satisfies that $\boldsymbol{w}_{\boldsymbol{\theta}}$ is twice differentiable with respect to $\boldsymbol{\theta}$.*
3. *Any involved operations of integration (expectation) and differentiation can be interchanged. Specifically, for any $\boldsymbol{\theta} \in \Theta$,*

$$\nabla_{\boldsymbol{\theta}} \int \nabla_{\boldsymbol{w}} c(\boldsymbol{w}, \boldsymbol{z})^\top p_{\boldsymbol{\theta}}(\boldsymbol{z}) d\boldsymbol{z} = \int \nabla_{\boldsymbol{w}} c(\boldsymbol{w}, \boldsymbol{z})^\top \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{z}) d\boldsymbol{z},$$

$$\int \nabla_{\boldsymbol{w}} c(\boldsymbol{w}, \boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z}) d\boldsymbol{z} = \nabla_{\boldsymbol{w}} \int c(\boldsymbol{w}, \boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z}) d\boldsymbol{z}.$$

It is well-known that Berry-Essen-type bounds provide a finite-sample version of the asymptotic normality. Below is an existing theorem of Berry-Esseen bounds for general M-estimation in statistical theory. It enhances the classical asymptotic results (e.g., Proposition 3) by providing a finite-sample performance guarantee.

Lemma 8 (Berry-Esseen bounds for M-estimators, Theorem 3.1 in Shao and Zhang (2022))**.** *Suppose the i.i.d. random variables $\boldsymbol{z}, \boldsymbol{z}_1, ..., \boldsymbol{z}_n$ follows the distribution P. The parameter space of ζ is a subset of $\mathbb{R}^{d_\zeta}$. Suppose that $m(\zeta, \boldsymbol{z})$ is a measurable function of $\boldsymbol{z}$. Let $M(\zeta) := \mathbb{E}_P(m(\zeta, \boldsymbol{z}))$ and*

$$\zeta^* := \operatorname*{argmin}_{\zeta} M(\zeta), \qquad \hat{\zeta}_n = \operatorname*{argmin}_{\zeta} \frac{1}{n} \sum_{i=1}^{n} m(\zeta, \boldsymbol{z}_i).$$

Moreover, suppose

1. *the function $m(\zeta, \cdot)$ is twice differentiable with respect to ζ and there exist constants $\mu > 0$, $c_1 > 0$, $c_2 > 0$ and two nonnegative functions $K_1, K_2 : \mathcal{Z} \to \mathbb{R}$ with $\|K_1(\boldsymbol{z})\|_9 \le c_1$ and $\|K_2(\boldsymbol{z})\|_4 \le c_2$ such that for any ζ,*

$$\begin{aligned} M(\zeta) - M(\zeta^*) &\ge \mu \|\zeta - \zeta^*\|^2, \\ |m(\zeta, \boldsymbol{z}) - m(\zeta^*, \boldsymbol{z})| &\le K_1(\boldsymbol{z}) \|\zeta - \zeta^*\|, \ \forall \boldsymbol{z} \in \mathcal{Z}, \\ \|\nabla_{\zeta\zeta} m(\zeta, \boldsymbol{z}) - \nabla_{\zeta\zeta} m(\zeta^*, \boldsymbol{z})\| &\le K_2(\boldsymbol{z}) \|\zeta - \zeta^*\|, \ \forall \boldsymbol{z} \in \mathcal{Z}. \end{aligned}$$

   Moreover, there exists a constant $c_3 \ge 0$ and a nonnegative function $K_3 : \mathcal{Z} \to \mathbb{R}$ such that for any $\boldsymbol{z} \in \mathcal{Z}$,

$$\nabla_{\zeta\zeta} m(\zeta^*, \boldsymbol{z}) \le K_3(\boldsymbol{z}) \boldsymbol{I} \text{ and } \|K_3(\boldsymbol{z})\|_4 \le c_3.$$

2. *Let $\Sigma = \mathbb{E}[\nabla_\zeta m(\zeta^*, \boldsymbol{z})^\top \nabla_\zeta m(\zeta^*, \boldsymbol{z})]$ and $V = \nabla_{\zeta\zeta} \mathbb{E}[m(\zeta^*, \boldsymbol{z})]$. Assume that there exist constants $\lambda_1 > 0$ and $\lambda_2 > 0$ such that $\lambda_{\min}(\Sigma) \ge \lambda_1$ and $\lambda_{\min}(V) \ge \lambda_2$. Moreover, assume that there exists a constant $c_4 > 0$ such that $\|\nabla_\zeta m(\zeta^*, \boldsymbol{z})\|_4 \le c_4 \sqrt{d_\zeta}$.*

Under the assumption above, let $\mathcal{A}$ denote all convex sets of $\mathbb{R}^{d_\zeta}$, and we have

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}(\sqrt{n}(\hat{\zeta}_n - \zeta^*) \in A) - \mathbb{P}(V^{-1} \Sigma^{\frac{1}{2}} \boldsymbol{Y}_0 \in A) \right| \le C d_\zeta^{\frac{9}{4}} n^{-\frac{1}{2}},$$

where $C > 0$ is a constant depending only on $c_1, c_2, c_3, c_4, \mu, \lambda_1, \lambda_2$,

Note that since

$$V^{-1} \Sigma^{\frac{1}{2}} \boldsymbol{Y}_0 \overset{d}{=} N(0, V^{-1} \Sigma^{\frac{1}{2}} (V^{-1} \Sigma^{\frac{1}{2}})^\top) \overset{d}{=} N(0, V^{-1} \Sigma V^{-1}) \overset{d}{=} (V^{-1} \Sigma V^{-1})^{\frac{1}{2}} \boldsymbol{Y}_0,$$

the above result will also hold if we replace $\mathbb{P}(V^{-1} \Sigma^{\frac{1}{2}} \boldsymbol{Y}_0 \in A)$ by $\mathbb{P}((V^{-1} \Sigma V^{-1})^{\frac{1}{2}} \boldsymbol{Y}_0 \in A)$.

We apply the above Berry-Esseen bounds for M-estimators (Lemma 8) directly to ETO and IEO, which leads to finite-sample performance guarantees in terms of the parameter estimation error.

Proposition 4.A (Berry-Esseen bounds for ETO)**.** *Under Assumption 6.A and 7.A, the following holds for any convex set A:*

$$|\mathbb{P}(\sqrt{n}(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}}) \in A) - \mathbb{P}(\boldsymbol{M}_1^{\mathrm{ETO}}\boldsymbol{Y}_0 \in A)| \le C^{\mathrm{ETO}} q^{\frac{9}{4}} n^{-\frac{1}{2}} =: C_{n,q}^{\mathrm{ETO}},$$

where $C^{\mathrm{ETO}} > 0$ is a constant depending only on $c_1^{\mathrm{ETO}}, c_2^{\mathrm{ETO}}, c_3^{\mathrm{ETO}}, c_4^{\mathrm{ETO}}, \mu^{\mathrm{ETO}}, \lambda_1^{\mathrm{ETO}}, \lambda_2^{\mathrm{ETO}}$.

Proposition 4.B (Berry-Esseen bounds for IEO)**.** *Under Assumption 6.B and 7.B, the following holds for any convex set A:*

$$|\mathbb{P}(\sqrt{n}(\hat{\boldsymbol{\theta}}^{\mathrm{IEO}} - \boldsymbol{\theta}^{*}) \in A) - \mathbb{P}(\boldsymbol{M}_1^{\mathrm{IEO}}\boldsymbol{Y}_0 \in A)| \le C^{\mathrm{IEO}} q^{\frac{9}{4}} n^{-\frac{1}{2}} =: C_{n,q}^{\mathrm{IEO}},$$

where $C^{\mathrm{IEO}} > 0$ is a constant depending only on $c_1^{\mathrm{IEO}}, c_2^{\mathrm{IEO}}, c_3^{\mathrm{IEO}}, c_4^{\mathrm{IEO}}, \mu^{\mathrm{IEO}}, \lambda_1^{\mathrm{IEO}}, \lambda_2^{\mathrm{IEO}}$.

Proof of Proposition 4. For ETO, consider $m(\zeta, \boldsymbol{z}) = -\log p_{\boldsymbol{\theta}}(\boldsymbol{z})$ with parameter $\zeta = \boldsymbol{\theta}$ and apply Lemma 8.

For IEO, consider $m(\zeta, \boldsymbol{z}) = c(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z})$ with parameter $\zeta = \boldsymbol{\theta}$ and apply Lemma 8. □

We give some remarks on Proposition 4. First, the inequalities in Proposition 4 clearly hold for all sets whose complements are convex. Second, Proposition 4 can be viewed as a finite-sample extension of Proposition 3 by showing the bound for the finite-sample error to the corresponding normal distributions, which is $O(q^{\frac{9}{4}} n^{-\frac{1}{2}})$. Finally, in general we have $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}^{\mathrm{KL}}$, $\boldsymbol{w}_{\boldsymbol{\theta}^*} \neq \boldsymbol{w}^*$, $\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}} \neq \boldsymbol{w}^*$ (unless the model is well-specified). Therefore, the centered parameters in Propositions 4.A and 4.B are generally different. This implies that we must tackle the difference of the centered parameters in addition to the distribution error, thus making the comparisons more delicate.

B.2 Proofs in Section 4.1

Proof of Theorem 2. For the zeroth-order term, this result has been established in Elmachtoub et al. (2023). Next, we examine the first-order and second-order terms for ETO and IEO, respectively.

We use Taylor expansion of $R(\boldsymbol{w}_{\boldsymbol{\theta}})$ at $\boldsymbol{\theta}$ ($\boldsymbol{\theta} = \boldsymbol{\theta}^{\mathrm{KL}}$ or $\boldsymbol{\theta}^*$), so that

$$\begin{aligned}
&R(\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}) \\
=&v_0(\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}) - v_0(\boldsymbol{w}^*) \\
=&v_0(\boldsymbol{w}_{\hat{\boldsymbol{\theta}}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}}) + v_0(\boldsymbol{w}_{\boldsymbol{\theta}}) - v_0(\boldsymbol{w}^*) \\
=&v_0(\boldsymbol{w}_{\boldsymbol{\theta}}) - v_0(\boldsymbol{w}^*) + \nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2).
\end{aligned}$$

Then we apply the higher-order delta method.

For IEO, we particularly have $\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) = 0$ by the first-order optimality condition. Therefore,

$$\begin{aligned}
R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) =&v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) - v_0(\boldsymbol{w}^*) + \frac{1}{2}(\hat{\boldsymbol{\theta}}^{\mathrm{IEO}} - \boldsymbol{\theta}^*)^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})(\hat{\boldsymbol{\theta}}^{\mathrm{IEO}} - \boldsymbol{\theta}^*) + o_P(\|\hat{\boldsymbol{\theta}}^{\mathrm{IEO}} - \boldsymbol{\theta}^*\|^2) \\
=&\kappa_0^{\mathrm{IEO}} + \frac{1}{2}(\hat{\boldsymbol{\theta}}^{\mathrm{IEO}} - \boldsymbol{\theta}^*)^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})(\hat{\boldsymbol{\theta}}^{\mathrm{IEO}} - \boldsymbol{\theta}^*) + o_P(\|\hat{\boldsymbol{\theta}}^{\mathrm{IEO}} - \boldsymbol{\theta}^*\|^2).
\end{aligned}$$

Proposition 3.B shows that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^{\mathrm{IEO}} - \boldsymbol{\theta}^{\mathrm{KL}}) \xrightarrow{d} \boldsymbol{M}_1^{\mathrm{IEO}}\boldsymbol{Y}_0.$$

Thus we have

$$\begin{aligned}
\sqrt{n}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - \kappa_0^{\mathrm{IEO}}) &\xrightarrow{P} 0. \\
n(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - \kappa_0^{\mathrm{IEO}}) &\xrightarrow{d} \mathbb{G}^{\mathrm{IEO}}.
\end{aligned}$$

For ETO, in general $\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) \neq 0$ and thus this term does not vanish.

$$\begin{aligned}
R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) =&v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) - v_0(\boldsymbol{w}^*) + \nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}}) \\
&+ \frac{1}{2}(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}}) + o_P(\|\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}}\|^2).
\end{aligned}$$

Proposition 3.A shows that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}}) \xrightarrow{d} \boldsymbol{M}_1^{\mathrm{ETO}} \boldsymbol{Y}_0.$$

By the continuous mapping theorem,

$$\sqrt{n}\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}}) \xrightarrow{d} \nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})^\top \boldsymbol{M}_1^{\mathrm{ETO}} \boldsymbol{Y}_0,$$

$$\frac{1}{2} n(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}}) \xrightarrow{d} \mathbb{G}^{\mathrm{ETO}}.$$

Hence we conclude that

$$\sqrt{n}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) - \kappa_0^{\mathrm{ETO}}) \xrightarrow{d} \nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})^\top \boldsymbol{M}_1^{\mathrm{ETO}} \boldsymbol{Y}_0$$

and

$$n(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) - \kappa_0^{\mathrm{ETO}} - \nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}})) \xrightarrow{d} \mathbb{G}^{\mathrm{ETO}}.$$

In addition, for any non-decreasing convex function $u : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[u(\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})^\top \boldsymbol{M}_1^{\mathrm{ETO}} \boldsymbol{Y}_0)] \geq u(\mathbb{E}(\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})^\top \boldsymbol{M}_1^{\mathrm{ETO}} \boldsymbol{Y}_0)) = 0.$$

Thus we have $\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})^\top \boldsymbol{M}_1^{\mathrm{ETO}} \boldsymbol{Y}_0 \succeq_{\text{s-st}} 0$.

□

Proof of Theorem 3. By the definition of $\mathbb{G}^{\mathrm{IEO}}$ and $\mathbb{G}^{\mathrm{ETO}}$,

$$\begin{aligned}
\mathbb{G}^{\mathrm{IEO}} \overset{d}{=} & \frac{1}{2} \boldsymbol{Y}_0^\top (\boldsymbol{M}_1^{\mathrm{IEO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \boldsymbol{M}_1^{\mathrm{IEO}} \boldsymbol{Y}_0 \\
= & \frac{1}{2} \boldsymbol{Y}_0^\top (\boldsymbol{M}_1^{\mathrm{ETO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) \boldsymbol{M}_1^{\mathrm{ETO}} \boldsymbol{Y}_0 \\
& + \frac{1}{2} \boldsymbol{Y}_0^\top \Big((\tilde{\boldsymbol{M}}_1^{\mathrm{ETO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{ETO}} - (\boldsymbol{M}_1^{\mathrm{ETO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) \boldsymbol{M}_1^{\mathrm{ETO}} \Big) \boldsymbol{Y}_0 \\
& + \frac{1}{2} \boldsymbol{Y}_0^\top \Big((\tilde{\boldsymbol{M}}_1^{\mathrm{IEO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{IEO}} - (\tilde{\boldsymbol{M}}_1^{\mathrm{ETO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{ETO}} \Big) \boldsymbol{Y}_0 \\
& + \frac{1}{2} \boldsymbol{Y}_0^\top \Big((\boldsymbol{M}_1^{\mathrm{IEO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \boldsymbol{M}_1^{\mathrm{IEO}} - (\tilde{\boldsymbol{M}}_1^{\mathrm{IEO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{IEO}} \Big) \boldsymbol{Y}_0 \\
\overset{d}{=} & \mathbb{G}^{\mathrm{ETO}} + \frac{1}{2} \boldsymbol{Y}_0^\top \Big(\Delta_1 + (\tilde{\boldsymbol{M}}_1^{\mathrm{IEO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{IEO}} - (\tilde{\boldsymbol{M}}_1^{\mathrm{ETO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{ETO}} + \Delta_2 \Big) \boldsymbol{Y}_0,
\end{aligned}$$

where

$$\begin{aligned}
\Delta_1 &= (\tilde{\boldsymbol{M}}_1^{\mathrm{ETO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{ETO}} - (\boldsymbol{M}_1^{\mathrm{ETO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) \boldsymbol{M}_1^{\mathrm{ETO}}, \\
\Delta_2 &= (\boldsymbol{M}_1^{\mathrm{IEO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \boldsymbol{M}_1^{\mathrm{IEO}} - (\tilde{\boldsymbol{M}}_1^{\mathrm{IEO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{IEO}}.
\end{aligned}$$

Note that in the proof of Lemma 1 in Elmachtoub et al. (2023), it has been shown that $(\tilde{\boldsymbol{M}}_1^{\mathrm{IEO}})^2 \geq (\tilde{\boldsymbol{M}}_1^{\mathrm{ETO}})^2$ by viewing P^{KL} as the ground-truth distribution. By Lemma 10, we have

$$(\tilde{\boldsymbol{M}}_1^{\mathrm{IEO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{IEO}} \geq (\tilde{\boldsymbol{M}}_1^{\mathrm{ETO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{ETO}}.$$

Hence,

$$\frac{1}{2} \boldsymbol{Y}_0^\top \Big((\tilde{\boldsymbol{M}}_1^{\mathrm{IEO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{IEO}} - (\tilde{\boldsymbol{M}}_1^{\mathrm{ETO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{ETO}} \Big) \boldsymbol{Y}_0 \geq 0$$

Moreover, by Assumption 4, we claim that

$$\|\Delta_1\| \lesssim B_0, \|\Delta_2\| \lesssim B_0.$$

To see this, we note the following facts. By the local Lipschitz continuity of matrix inversion, Assumption 4 implies

$$\begin{aligned}
\big\| (\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbb{E}_P[\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})]))^{-1} - (\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbb{E}_{P^{\mathrm{KL}}}[\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})]))^{-1} \big\| &\lesssim B_0, \\
\big\| (\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}))^{-1} - (\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}, P^{\mathrm{KL}}))^{-1} \big\| &\lesssim B_0
\end{aligned}$$

where $\lesssim$ hides the local Lipschitz constant. By using Lemma 9 and Assumption 4, we have

$$\begin{aligned}
&\| (\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbb{E}_P[\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})])^{-1} \operatorname{var}_P(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z}))(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbb{E}_P[\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})])^{-1} \\
&- (\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbb{E}_{P^{\mathrm{KL}}}[\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})])^{-1} \operatorname{var}_{P^{\mathrm{KL}}}(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z}))(\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbb{E}_{P^{\mathrm{KL}}}[\log p_{\boldsymbol{\theta}^{\mathrm{KL}}}(\boldsymbol{z})])^{-1} \| \lesssim B_0, \\
&\|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})^{-1} \operatorname{var}_P(\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}^*}, \boldsymbol{z}))\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})^{-1} \\
&- \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}, P^{\mathrm{KL}})^{-1} \operatorname{var}_{P^{\mathrm{KL}}}(\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}, \boldsymbol{z}))\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}, P^{\mathrm{KL}})^{-1} \| \lesssim B_0
\end{aligned}$$

In other words,

$$\|(\boldsymbol{M}_1^{\mathrm{ETO}})^2 - (\tilde{\boldsymbol{M}}_1^{\mathrm{ETO}})^2\| \lesssim B_0, \|(\boldsymbol{M}_1^{\mathrm{IEO}})^2 - (\tilde{\boldsymbol{M}}_1^{\mathrm{IEO}})^2\| \lesssim B_0.$$

By the local Lipschitz continuity of matrix square root, we also have

$$\|\boldsymbol{M}_1^{\mathrm{ETO}} - \tilde{\boldsymbol{M}}_1^{\mathrm{ETO}}\| \lesssim B_0, \|\boldsymbol{M}_1^{\mathrm{IEO}} - \tilde{\boldsymbol{M}}_1^{\mathrm{IEO}}\| \lesssim B_0.$$

By using Lemma 9 and Assumption 4 again, we have

$$\|(\tilde{\boldsymbol{M}}_1^{\mathrm{ETO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{ETO}} - (\boldsymbol{M}_1^{\mathrm{ETO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) \boldsymbol{M}_1^{\mathrm{ETO}}\| \lesssim B_0$$

$$\|(\boldsymbol{M}_1^{\mathrm{IEO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \boldsymbol{M}_1^{\mathrm{IEO}} - (\tilde{\boldsymbol{M}}_1^{\mathrm{IEO}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \tilde{\boldsymbol{M}}_1^{\mathrm{IEO}}\| \lesssim B_0$$

Hence, we have $\|\Delta\| = \|\Delta_1 + \Delta_2\| \leq \|\Delta_1\| + \|\Delta_2\| \lesssim B_0$. We thus conclude that there exists a problem-dependent constant C_{mis} such that $\|\Delta\| \leq C_{\mathrm{mis}} B_0$. □

Lemma 9. *Suppose Q_1, Q_2, Q_3, Q_1', Q_2', Q_3' are square matrices. Then we have*

$$\|Q_1Q_2Q_3 - Q_1'Q_2'Q_3'\| \lesssim \left(\max_i \|Q_i - Q_i'\|\right) \left(\max_i (\|Q_i\|, \|Q_i'\|)\right)^2,$$

$$\|Q_1Q_2Q_3 - Q_1'Q_2Q_3'\| \lesssim \left(\max_i \|Q_i - Q_i'\|\right) \left(\max_i (\|Q_i\|, \|Q_i'\|)\right)^2.$$

Proof of Lemma 9. Note that

$$\begin{aligned}
&\|Q_1Q_2Q_3 - Q_1'Q_2'Q_3'\| \\
\leq &\|(Q_1 - Q_1')Q_2Q_3 + Q_1'(Q_2 - Q_2')Q_3 + Q_1'Q_2'(Q_3 - Q_3')\| \\
\leq &\|Q_1 - Q_1'\|\|Q_2\|\|Q_3\| + \|Q_1'\|\|Q_2 - Q_2'\|\|Q_3\| + \|Q_1'\|\|Q_2'\|\|Q_3 - Q_3'\| \\
\leq &3 \left(\max_i \|Q_i - Q_i'\|\right) \left(\max_i (\|Q_i\|, \|Q_i'\|)\right)^2
\end{aligned}$$

The other inequality can be derived similarly. □

Lemma 10. *Suppose Q_1, Q_2, and Q_3 are all positive definite matrices. If $Q_1^2 \leq Q_2^2$, then we have*

$$Q_1Q_3Q_1 \leq Q_2Q_3Q_2.$$

Proof of Lemma 10. We note that $Q_1Q_1 \leq Q_2Q_2$ implies that $Q_2^{-1}Q_1Q_1Q_2^{-1} \leq I$ so we have

$$\|Q_2^{-1}Q_1Q_1Q_2^{-1}\|_{\mathrm{op}} \leq 1$$

where $\|\cdot\|_{\mathrm{op}}$ is the operator norm of the matrix and thus $\|Q_1Q_2^{-1}\|_{\mathrm{op}}^2 = \|(Q_1Q_2^{-1})^\top Q_1Q_2^{-1}\|_{\mathrm{op}} = \|Q_2^{-1}Q_1Q_1Q_2^{-1}\|_{\mathrm{op}} \leq 1$. This shows that all eigenvalues of $Q_1Q_2^{-1}$ are less than 1. Since $Q_3^{\frac{1}{2}}Q_1Q_2^{-1}Q_3^{-\frac{1}{2}}$ is similar to $Q_1Q_2^{-1}$, all the eigenvalues of $Q_3^{\frac{1}{2}}Q_1Q_2^{-1}Q_3^{-\frac{1}{2}}$ are the same as $Q_1Q_2^{-1}$ (all less than 1), which implies that

$$\|Q_3^{\frac{1}{2}}Q_1Q_2^{-1}Q_3^{-\frac{1}{2}}\|_{\mathrm{op}} \leq 1.$$

Taking the transpose, we also have

$$\|Q_3^{-\frac{1}{2}}Q_2^{-1}Q_1Q_3^{\frac{1}{2}}\|_{\mathrm{op}} \leq 1.$$

Hence we have

$$\|Q_3^{-\frac{1}{2}}Q_2^{-1}Q_1Q_3^{\frac{1}{2}}Q_3^{\frac{1}{2}}Q_1Q_2^{-1}Q_3^{-\frac{1}{2}}\|_{\mathrm{op}} \le \|Q_3^{-\frac{1}{2}}Q_2^{-1}Q_1Q_3^{\frac{1}{2}}\|_{\mathrm{op}}\|Q_3^{\frac{1}{2}}Q_1Q_2^{-1}Q_3^{-\frac{1}{2}}\|_{\mathrm{op}} \le 1.$$

This implies that

$$Q_3^{-\frac{1}{2}}Q_2^{-1}Q_1Q_3^{\frac{1}{2}}Q_3^{\frac{1}{2}}Q_1Q_2^{-1}Q_3^{-\frac{1}{2}} \le I$$

and thus

$$Q_1Q_3Q_1 \le Q_2Q_3Q_2.$$

□

Proof of Corollary 1. Theorem 3 implies that

$$\begin{aligned}
&Z^{\mathrm{IEO}} - Z^{\mathrm{ETO}} \\
=&\frac{1}{2}\boldsymbol{Y}_0^\top\Big((\boldsymbol{M}_1^{\mathrm{IEO}})^\top\nabla_{\boldsymbol{\theta\theta}}v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})\boldsymbol{M}_1^{\mathrm{IEO}} - (\boldsymbol{M}_1^{\mathrm{ETO}})^\top\nabla_{\boldsymbol{\theta\theta}}v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}}\Big)\boldsymbol{Y}_0 \\
\ge&\frac{1}{2}\boldsymbol{Y}_0^\top\Big((\tilde{\boldsymbol{M}}_1^{\mathrm{IEO}})^\top\nabla_{\boldsymbol{\theta\theta}}v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})\tilde{\boldsymbol{M}}_1^{\mathrm{IEO}} - (\tilde{\boldsymbol{M}}_1^{\mathrm{ETO}})^\top\nabla_{\boldsymbol{\theta\theta}}v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})\tilde{\boldsymbol{M}}_1^{\mathrm{ETO}} - C_{\mathrm{mis}}B_0\boldsymbol{I}\Big)\boldsymbol{Y}_0 \\
\ge&\frac{1}{2}\boldsymbol{Y}_0^\top\Big(\tau_3\boldsymbol{I} - C_{\mathrm{mis}}B_0\boldsymbol{I}\Big)\boldsymbol{Y}_0 \\
=&\frac{1}{2}\Big(\tau_3 - C_{\mathrm{mis}}B_0\Big)\boldsymbol{Y}_0^\top\boldsymbol{Y}_0.
\end{aligned}$$

Hence $\tau_1 \ge \tau_3 - C_{\mathrm{mis}}B_0$.

If $\tau_3 - C_{\mathrm{mis}}B_0 \ge 0$, we have $\tau_1 \ge \tau_3 - C_{\mathrm{mis}}B_0 \ge 0$ and $Z^{\mathrm{IEO}} \ge Z^{\mathrm{ETO}}$. Hence, $\mathbb{G}^{\mathrm{ETO}}$ is first-order stochastically dominated by $\mathbb{G}^{\mathrm{IEO}}$. □

B.3 Proofs in Section 4.2

We first introduce a simple result on the tail upper bound of a multivariate Gaussian distribution.

Lemma 11. *Suppose that $\boldsymbol{X} \in \mathbb{R}^q$ is a multivariate Gaussian distribution $\boldsymbol{X} \overset{d}{=} \boldsymbol{M}^\top\boldsymbol{Y}_0$ where $\boldsymbol{M}$ is a square matrix and $\boldsymbol{Y}_0$ is the standard Gaussian distribution. Let $\boldsymbol{v} \in \mathbb{R}^q$ be a vector. For $t \ge 0$, we have*

$$\mathbb{P}(\|\boldsymbol{X}\| \ge t) \le 2q\exp\left(-\frac{t^2}{2q\|\boldsymbol{M}\|^2}\right).$$

$$\mathbb{P}(\boldsymbol{v}^\top\boldsymbol{X} \ge t) \le \exp\left(-\frac{t^2}{2\|\boldsymbol{M}\boldsymbol{v}\|^2}\right).$$

For any $s_1 \le s_2$, we have

$$\mathbb{P}(s_1 \le \boldsymbol{v}^\top\boldsymbol{X} \le s_2) \le \frac{1}{\sqrt{2\pi}}\frac{s_2 - s_1}{\|\boldsymbol{M}\boldsymbol{v}\|^2}.$$

Proof of Lemma 11. Note that

$$\begin{aligned}
\mathbb{P}(\|\boldsymbol{X}\| \ge t) =&\mathbb{P}(\|\boldsymbol{M}^\top\boldsymbol{Y}_0\| \ge t) \\
\le&\mathbb{P}(\|\boldsymbol{M}\|\|\boldsymbol{Y}_0\| \ge t) \\
\le&\mathbb{P}\left(\|\boldsymbol{Y}_0\|^2 \ge \frac{t^2}{\|\boldsymbol{M}\|^2}\right) \\
\le&q\mathbb{P}\left((Y_0^{(1)})^2 \ge \frac{t^2}{q\|\boldsymbol{M}\|^2}\right) \\
\le&2q\exp\left(-\frac{t^2}{2q\|\boldsymbol{M}\|^2}\right).
\end{aligned}$$

Note that $\boldsymbol{v}^\top \boldsymbol{X} \stackrel{d}{=} N(0, \boldsymbol{v}^\top \boldsymbol{M}^\top \boldsymbol{M}\boldsymbol{v})$ so $\frac{\boldsymbol{v}^\top \boldsymbol{X}}{\|\boldsymbol{M}\boldsymbol{v}\|} \stackrel{d}{=} N(0,1)$ and thus

$$\mathbb{P}(\boldsymbol{v}^\top \boldsymbol{X} \geq t) = \mathbb{P}\left(\frac{\boldsymbol{v}^\top \boldsymbol{X}}{\|\boldsymbol{M}\boldsymbol{v}\|} \geq \frac{t}{\|\boldsymbol{M}\boldsymbol{v}\|}\right) \leq \exp\left(-\frac{t^2}{2\|\boldsymbol{M}\boldsymbol{v}\|^2}\right).$$

$$\mathbb{P}(s_1 \leq \boldsymbol{v}^\top \boldsymbol{X} \leq s_2) = \mathbb{P}\left(\frac{s_1}{\|\boldsymbol{M}\boldsymbol{v}\|} \leq \frac{\boldsymbol{v}^\top \boldsymbol{X}}{\|\boldsymbol{M}\boldsymbol{v}\|} \leq \frac{s_2}{\|\boldsymbol{M}\boldsymbol{v}\|}\right) \leq \frac{1}{\sqrt{2\pi}} \frac{s_2 - s_1}{\|\boldsymbol{M}\boldsymbol{v}\|^2}.$$

□

We provide proofs for Proposition 1.

Proof of Proposition 1.A. Since $\|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}})\| \leq L_2$, we have

$$v_0(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) \leq \nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}}) + \frac{L_2}{2}\left\|\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}}\right\|^2.$$

For ETO, it holds that $\sqrt{n}(v_0(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})) \xrightarrow{d} \nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\mathbb{N}_1^{\mathrm{ETO}}$. Let $\boldsymbol{X}_n := \sqrt{n}(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}})$ and $\boldsymbol{X} := \mathbb{N}_1^{\mathrm{ETO}}$. Let f be the function with $f(\cdot) = \nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\cdot)$. Let $Y_n = v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) - \nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}})$ with $\sqrt{n}Y_n \leq L_2 \|\boldsymbol{X}_n\|^2 / (2\sqrt{n})$. We have

$$\begin{aligned}
&\sup_{t\geq 0}\left|\mathbb{P}(\sqrt{n}(v_0(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})) \geq t) - \mathbb{P}(\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\mathbb{N}_1^{\mathrm{ETO}} \geq t)\right| \\
=&\sup_{t\geq 0}\left|\mathbb{P}(f(\boldsymbol{X}_n) + \sqrt{n}Y_n \geq t) - \mathbb{P}(f(\boldsymbol{X}) \geq t)\right| \\
\leq&\sup_{t\geq 0}\left|\mathbb{P}(f(\boldsymbol{X}_n) + \sqrt{n}Y_n \geq t) - \mathbb{P}(f(\boldsymbol{X}_n) \geq t)\right| + \sup_{t\geq 0}\left|\mathbb{P}(f(\boldsymbol{X}_n) \geq t) - \mathbb{P}(f(\boldsymbol{X}) \geq t)\right|.
\end{aligned}$$

On one hand,

$$\sup_{t\geq 0}|\mathbb{P}(f(\boldsymbol{X}_n) \geq t) - \mathbb{P}(f(\boldsymbol{X}) \geq t)| = \sup_{t\geq 0}\left|\mathbb{P}(\boldsymbol{X}_n \in \underbrace{f^{-1}([t,\infty))}_{\text{convex}}) - \mathbb{P}(\boldsymbol{X} \in \underbrace{f^{-1}([t,\infty))}_{\text{convex}})\right| \leq C_{n,q}^{\mathrm{ETO}}.$$

On the other hand,

$$\begin{aligned}
&|\mathbb{P}(f(\boldsymbol{X}_n) + \sqrt{n}Y_n \geq t) - \mathbb{P}(f(\boldsymbol{X}_n) \geq t)| \\
\leq&\mathbb{P}(f(\boldsymbol{X}_n) + nY_n \geq t, f(\boldsymbol{X}_n) < t) + \mathbb{P}(f(\boldsymbol{X}_n) + \sqrt{n}Y_n < t, f(\boldsymbol{X}_n) \geq t) \\
\leq&2\mathbb{P}\left(|\sqrt{n}Y_n| > |t - f(\boldsymbol{X}_n)|)\right) \\
\leq&2\mathbb{P}\left(\frac{L_2}{2\sqrt{n}}\|\boldsymbol{X}_n\|^2 > |t - f(\boldsymbol{X}_n)|\right).
\end{aligned}$$

Note that for any $\varepsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{L_2}{2\sqrt{n}}\|\boldsymbol{X}_n\|^2 > |t - f(\boldsymbol{X}_n)|\right) \\
=&\mathbb{P}\left(\frac{L_2}{2\sqrt{n}}\|\boldsymbol{X}_n\|^2 > |t - f(\boldsymbol{X}_n)|, |t - f(\boldsymbol{X}_n)| > \varepsilon\right) + \mathbb{P}\left(\frac{L_2}{2\sqrt{n}}\|\boldsymbol{X}_n\|^2 > |t - f(\boldsymbol{X}_n)|, |t - f(\boldsymbol{X}_n)| \leq \varepsilon\right) \\
\leq&\mathbb{P}\left(\frac{L_2}{2\sqrt{n}}\|\boldsymbol{X}_n\|^2 > \varepsilon\right) + \mathbb{P}\left(|t - f(\boldsymbol{X}_n)| \leq \varepsilon\right).
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{P}(|t - f(\boldsymbol{X}_n)| \leq \varepsilon) &= \mathbb{P}(\boldsymbol{X}_n \in f^{-1}(-\infty, t+\varepsilon)) - \mathbb{P}(\boldsymbol{X}_n \in f^{-1}(-\infty, t-\varepsilon)) \\
&\leq \mathbb{P}(\boldsymbol{X} \in f^{-1}(-\infty, t+\varepsilon)) - \mathbb{P}(\boldsymbol{X} \in f^{-1}(-\infty, t-\varepsilon)) + 2C_{n,q}^{\mathrm{ETO}} \\
&= \mathbb{P}(t - \varepsilon \leq f(\boldsymbol{X}) \leq t + \varepsilon) + 2C_{n,q}^{\mathrm{ETO}} \\
&\leq \frac{2\varepsilon}{\sqrt{2\pi}\left\|\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}}\right\|} + 2C_{n,q}^{\mathrm{ETO}}.
\end{aligned}$$

where the last inequality follows from Lemma 11.

Moreover, by Lemma 11,

$$\begin{aligned} P\left(\frac{L_2}{2\sqrt{n}}\|\boldsymbol{X}_n\|^2 > \varepsilon\right) &\le P\left(\frac{L_2}{2\sqrt{n}}\|\boldsymbol{X}\|^2 > \varepsilon\right) + C_{n,q}^{\mathrm{ETO}} \\ &\le 2q\exp\left(-\frac{\sqrt{n}\varepsilon}{q\left\|\boldsymbol{M}_1^{\mathrm{ETO}}\right\|^2 L_2}\right) + C_{n,q}^{\mathrm{ETO}}. \end{aligned}$$

Taking $\varepsilon = \left\|\boldsymbol{M}_1^{\mathrm{ETO}}\right\|^2 L_2 q n^{-\frac{1}{2}} \log n$, we have

$$P\left(\frac{L_2}{2\sqrt{n}}\|\boldsymbol{X}_n\|^2 > \varepsilon\right) \le 2qn^{-1} + C_{n,q}^{\mathrm{ETO}}$$

and

$$\mathbb{P}\left(\frac{L_2}{2\sqrt{n}}\|\boldsymbol{X}_n\|^2 > |t - f(\boldsymbol{X}_n)|\right) \lesssim \frac{\left\|\boldsymbol{M}_1^{\mathrm{ETO}}\right\|^2 L_2}{\left\|\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}}\right\|} q n^{-\frac{1}{2}} \log n + qn^{-1} + C_{n,q}^{\mathrm{ETO}}$$

Combining things together, we have

$$\begin{aligned} &\sup_{t\ge 0}\left|\mathbb{P}(\sqrt{n}(v_0(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})) \ge t) - \mathbb{P}(\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\mathbb{N}_1^{\mathrm{ETO}} \ge t)\right| \\ \lesssim &\frac{\left\|\boldsymbol{M}_1^{\mathrm{ETO}}\right\|^2 L_2}{\left\|\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}}\right\|} q n^{-\frac{1}{2}} \log n + qn^{-1} + C_{n,q}^{\mathrm{ETO}}. \end{aligned}$$

□

Proof of Proposition 1.B. Since $\|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}})\| \le L_2$, we have

$$v_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \le \frac{L_2}{2}\left\|\hat{\boldsymbol{\theta}}^{\mathrm{IEO}} - \boldsymbol{\theta}^*\right\|^2.$$

For IEO, it holds that $\sqrt{n}(v_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})) \xrightarrow{P} 0$. Let $\boldsymbol{X}_n := \sqrt{n}(\hat{\boldsymbol{\theta}}^{\mathrm{IEO}} - \boldsymbol{\theta}^*)$ and $\boldsymbol{X} := \mathbb{N}_1^{\mathrm{IEO}}$. Following a similar argument as in Proposition 1.A, we obtain that

$$\begin{aligned} \mathbb{P}(\sqrt{n}(v_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})) \ge t) &\le \mathbb{P}(\sqrt{n}\frac{L_2}{2}\left\|\hat{\boldsymbol{\theta}}^{\mathrm{IEO}} - \boldsymbol{\theta}^*\right\|^2 \ge t) \\ &= \mathbb{P}(\|\boldsymbol{X}_n\|^2 \ge \frac{2\sqrt{n}t}{L_2}) \\ &\le \mathbb{P}(\|\boldsymbol{X}\|^2 \ge \frac{2\sqrt{n}t}{L_2}) + C_{n,q}^{\mathrm{IEO}} \\ &\le 2q\exp\left(-\frac{\sqrt{n}t}{qL_2\left\|\boldsymbol{M}_1^{\mathrm{IEO}}\right\|^2}\right) + C_{n,q}^{\mathrm{IEO}}. \end{aligned}$$

□

Next, without loss of generality, we mainly focus on Proposition 2.B, since the techniques are almost the same for ETO in Proposition 2.A. We first introduce the following two technical lemmas in optimization and probability literature.

Lemma 12 (Lemma 1.2.4 in Nesterov (2013))**.** *For a twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, if it has M-Lipschitz Hessian, i.e.,*

$$\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \le M\|\boldsymbol{x} - \boldsymbol{y}\|,$$

then

$$\left|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top(\boldsymbol{y} - \boldsymbol{x}) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^\top\nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})\right| \le \frac{M}{6}\|\boldsymbol{y} - \boldsymbol{x}\|^3.$$

Lemma 13 (Quadratic forms of random variables (Muirhead, 1982)). *Let $\boldsymbol{X}$ be $q \times 1$ random vector with mean value $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then for a $q \times q$ symmetric nonsingular matrix $\boldsymbol{A}$, the quadratic form $Q(\boldsymbol{X}) := \boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}$ has the following representation:*

$$Q(\boldsymbol{X}) = \boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} = \sum_{j=1}^{q} \lambda_j (U_j + b_j)^2,$$

for some $\lambda_1, ..., \lambda_q \in \mathbb{R}$, some fixed vector $\boldsymbol{b} \in \mathbb{R}^q$ and some random vector $\boldsymbol{U}$ with $\mathbb{E}(\boldsymbol{U}) = 0$ and $\mathrm{cov}(\boldsymbol{U}) = \boldsymbol{I}$.

Proof of Lemma 13. By denoting $\boldsymbol{Y} = \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{X}$ and $\boldsymbol{Z} = \boldsymbol{Y} - \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\mu}$, we have

$$\begin{aligned} \mathbb{E}(\boldsymbol{Y}) = \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\mu}, \qquad & \mathrm{cov}(Y) = \boldsymbol{\Sigma}^{-\frac{1}{2}} \mathrm{cov}(\boldsymbol{X}) \boldsymbol{\Sigma}^{-\frac{1}{2}} = \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-\frac{1}{2}} = \boldsymbol{I}, \\ \mathbb{E}(\boldsymbol{Z}) = \boldsymbol{0}, \qquad & \mathrm{cov}(\boldsymbol{Z}) = \boldsymbol{I}, \end{aligned}$$

and

$$Q(\boldsymbol{X}) = \boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} = \boldsymbol{Y}^\top \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{A} \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{Y} = (\boldsymbol{Z} + \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\mu}) \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{A} \boldsymbol{\Sigma}^{\frac{1}{2}} (\boldsymbol{Z} + \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\mu}).$$

Let $\boldsymbol{P}$ be a $q \times q$ orthonormal matrix that diagonalizes $\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{A} \boldsymbol{\Sigma}^{\frac{1}{2}}$ with

$$\boldsymbol{P}^\top \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{A} \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{P} = \mathrm{diag}(\lambda_1, ..., \lambda_q), \qquad \boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I},$$

where $\lambda_1, ..., \lambda_q$ are the eigenvalues of $\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{A} \boldsymbol{\Sigma}^{\frac{1}{2}}$. Letting $\boldsymbol{U} = \boldsymbol{P}^\top \boldsymbol{Z}$, we have

$$\boldsymbol{Z} = \boldsymbol{P}^\top \boldsymbol{U}, \qquad , \mathbb{E}(\boldsymbol{U}) = \boldsymbol{0}, \qquad \mathrm{cov}(\boldsymbol{U}) = \boldsymbol{I}.$$

Then by setting $\boldsymbol{b} = \boldsymbol{P}^\top \Sigma^{-\frac{1}{2}} \boldsymbol{\mu}$,

$$\begin{aligned} Q(\boldsymbol{X}) &= (\boldsymbol{Z} + \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\mu}) \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{A} \boldsymbol{\Sigma}^{\frac{1}{2}} (\boldsymbol{Z} + \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\mu}) = (\boldsymbol{U} + \boldsymbol{b})^\top \boldsymbol{P}^\top \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{A} \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{P} (\boldsymbol{U} + \boldsymbol{b}) \\ &= (\boldsymbol{U} + \boldsymbol{b})^\top \mathrm{diag}(\lambda_1, ..., \lambda_q)(\boldsymbol{U} + \boldsymbol{b}) = \sum_{j=1}^{q} \lambda_j (U_j + b_j)^2. \end{aligned}$$

□

In particular, when $\boldsymbol{X}$ is Gaussian, then $Q(\boldsymbol{X})$ is a linear combination of independent chi-square variables when $\boldsymbol{\mu} = \boldsymbol{0}$ and of noncentral chi-square variables when $\boldsymbol{\mu} \neq \boldsymbol{0}$. More generally, for the general quadratic form with nonsingular symmetric matrix $\boldsymbol{A}$,

$$\begin{aligned} Q(\boldsymbol{X}) &= \frac{1}{2} \boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} + \boldsymbol{d}^\top \boldsymbol{X} + \boldsymbol{c} = \frac{1}{2} (\boldsymbol{x} - \boldsymbol{A}^{-1}\boldsymbol{d})^\top \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{A}^{-1}\boldsymbol{d}) + \boldsymbol{c} - \frac{1}{2} \boldsymbol{d}^\top \boldsymbol{A}^{-1} \boldsymbol{d} \\ &= \sum_{j=1}^{q} \lambda_j (U_j + b_j)^2 + \boldsymbol{c} - \frac{1}{2} \boldsymbol{d}^\top \boldsymbol{A}^{-1} \boldsymbol{d}, \end{aligned}$$

for some $\lambda_1, ..., \lambda_q \in \mathbb{R}$, some fixed vector $\boldsymbol{b} \in \mathbb{R}^q$ and some random vector $\boldsymbol{U}$ with $\mathbb{E}(\boldsymbol{U}) = 0$ and $\mathrm{cov}(\boldsymbol{U}) = \boldsymbol{I}$.

Proof of Proposition 2.B. For simplicity, let $\boldsymbol{X}_n := \sqrt{n}(\hat{\boldsymbol{\theta}}^{\mathrm{IEO}} - \boldsymbol{\theta}^*)$, $\boldsymbol{X} := \mathbb{N}_1^{\mathrm{IEO}}$ and denote the quadratic function $f : \mathbb{R}^q \to \mathbb{R}$, $f(\boldsymbol{x}) = \frac{1}{2} \boldsymbol{x}^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \boldsymbol{x}$. In this case, $\mathbb{G}^{\mathrm{IEO}} = f(\boldsymbol{X})$. Before going to the proof, we point our three facts.

Claim 1. For all $\boldsymbol{\theta} \in \Theta$, v_0 satisfies:

$$\left| v_0(\boldsymbol{w}_{\boldsymbol{\theta}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) - \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right| \leq \frac{L_1}{6} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^3. \tag{6}$$

The reason is as follows. By Assumption 7.B,

$$\|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z}) - \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}, \boldsymbol{z})\| \leq L_1 \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|, \quad \forall \boldsymbol{z} \in \mathcal{Z},$$

By Lemma 12, we can get (6).

Claim 2. We have $\mathbb{E}\|\boldsymbol{X}\|_2^r < \infty$ for all $r \geq 2$. To see this, For each fixed vector $\boldsymbol{x} \in \mathbb{R}^q$, and any $2 \leq r < \infty$, by Höder's inequality, we have $\|\boldsymbol{x}\|_2 \leq q^{\frac{1}{2}-1/r}\|\boldsymbol{x}\|_r$. Hence,

$$\mathbb{E}[\|\boldsymbol{X}\|_2^r] \leq \mathbb{E}[(q^{\frac{1}{2}-1/r}\|\boldsymbol{X}\|_r)^r] = q^{r/2-1}\mathbb{E}\|\boldsymbol{X}\|_r^r = q^{r/2-1}\mathbb{E}[\sum_{i=1}^q |\boldsymbol{X}_i|^r] \leq q^{r/2-1}\sum_{i=1}^q \mathbb{E}|\boldsymbol{X}_i|^r$$
$$\leq q^{r/2-1}\sum_{i=1}^q (K_i\sqrt{r})^r \leq q^{r/2}(\max_i K_i\sqrt{r})^r = (\max_{i\in[q]} K_i\sqrt{qr})^r \asymp (qr)^{r/2},$$

where the second last inequality holds because each component of $\boldsymbol{X}$, say $\boldsymbol{X}^{(i)}$, is a Gaussian distribution with variance proxy K_i, satisfying the moment property (Proposition 2.5.2 in Vershynin (2018)).

Claim 3. From Lemma 13, $\mathbb{G}^{\text{IEO}}$ is the linear combination of independent Chi-squared distributions with degree of freedom 1. Moreover, $\mathbb{G}^{\text{IEO}}$ has bounded density when $q \geq 2$. More precisely,

$$\mathbb{G}^{\text{IEO}} = \sum_{i=1}^q \lambda_i \chi^2(1),$$

where $\lambda_1, ..., \lambda_q \geq 0$ are eigenvalues of the matrix $\frac{1}{2}\boldsymbol{M}_1^{\text{IEO}}\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})\boldsymbol{M}_1^{\text{IEO}}$. When $q = 1$, $\mathbb{G}^{\text{IEO}}$ is a scaled version of the chi-square distribution with a degree of freedom 1. When $q \geq 2$, by Theorem 1 of Bobkov et al. (2020), the density of $\mathbb{G}^{\text{IEO}}$, say $p_{\mathbb{G}^{\text{IEO}}}(x)$, satisfies

$$\frac{1}{4e^2\sqrt{2\pi}}\left(\left(\sum_{i=1}^q \lambda_i^2\right)\left(\sum_{i=2}^q \lambda_i^2\right)\right)^{-\frac{1}{4}} \leq \sup_{x>0} p_{\mathbb{G}^{\text{IEO}}}(x) \leq \frac{2}{\sqrt{\pi}}\left(\left(\sum_{i=1}^q \lambda_i^2\right)\left(\sum_{i=2}^q \lambda_i^2\right)\right)^{-\frac{1}{4}}.$$

From the assumption, we define Y_n by

$$Y_n := v_0(\boldsymbol{w}_{\hat{\boldsymbol{\theta}}^{\text{IEO}}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) - \frac{1}{2}(\hat{\boldsymbol{\theta}}^{\text{IEO}} - \boldsymbol{\theta}^*)^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})(\hat{\boldsymbol{\theta}}^{\text{IEO}} - \boldsymbol{\theta}^*) \leq \frac{L_1}{6}\|\hat{\boldsymbol{\theta}}^{\text{IEO}} - \boldsymbol{\theta}^*\|^3,$$

with $|nY_n| \leq \frac{L_1}{6\sqrt{n}}\|\boldsymbol{X}_n\|^3$. For all $t \geq 0$,

$$\begin{aligned}
&\sup_{t\geq 0}\left|\mathbb{P}(nv_0(\hat{\boldsymbol{w}}^{\text{IEO}}) - nv_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \geq t) - \mathbb{P}(\mathbb{G}^{\text{IEO}} \geq t)\right| \\
=&\sup_{t\geq 0}\left|P\left(\frac{n}{2}(\hat{\boldsymbol{\theta}}^{\text{IEO}} - \boldsymbol{\theta}^*)^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})(\hat{\boldsymbol{\theta}}^{\text{IEO}} - \boldsymbol{\theta}^*) + nY_n \leq t\right) - \mathbb{P}(\mathbb{G}^{\text{IEO}} \geq t)\right| \\
=&\sup_{t\geq 0}|\mathbb{P}(f(\boldsymbol{X}_n) + nY_n \geq t) - \mathbb{P}(f(\boldsymbol{X}) \geq t)| \\
=&\sup_{t\geq 0}|\mathbb{P}(f(\boldsymbol{X}_n) + nY_n \geq t) - \mathbb{P}(f(\boldsymbol{X}_n) \geq t)| + \sup_{t\geq 0}|\mathbb{P}(f(\boldsymbol{X}_n) \geq t) - \mathbb{P}(f(\boldsymbol{X}) \geq t)|
\end{aligned}$$

From Proposition 4.B,

$$\sup_{t\geq 0}|\mathbb{P}(f(\boldsymbol{X}_n) \geq t) - \mathbb{P}(f(\boldsymbol{X}) \geq t)| = \sup_{t\geq 0}\left|\mathbb{P}(\boldsymbol{X}_n \in \underbrace{f^{-1}((-\infty, t])}_{\text{convex}}) - \mathbb{P}(\boldsymbol{X} \in \underbrace{f^{-1}((-\infty, t])}_{\text{convex}})\right| \leq C_{n,q}^{\text{IEO}}.$$

On the other hand,

$$\begin{aligned}
&|\mathbb{P}(f(\boldsymbol{X}_n) + nY_n \geq t) - \mathbb{P}(f(\boldsymbol{X}_n) \geq t)| \\
\leq&\mathbb{P}(f(\boldsymbol{X}_n) + nY_n \geq t, f(\boldsymbol{X}_n) < t) + \mathbb{P}(f(\boldsymbol{X}_n) + nY_n < t, f(\boldsymbol{X}_n) \geq t) \\
\leq&2\mathbb{P}\left(|nY_n| > |t - f(\boldsymbol{X}_n)|\right) \\
\leq&2\mathbb{P}\left(\frac{L_1}{6\sqrt{n}}\|\boldsymbol{X}_n\|^3 > |t - f(\boldsymbol{X}_n)|\right).
\end{aligned}$$

Note that for any $\varepsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{L_1}{6\sqrt{n}}\|\boldsymbol{X}_n\|^3 > |t - f(\boldsymbol{X}_n)|\right) \\
=&\mathbb{P}\left(\frac{L_1}{6\sqrt{n}}\|\boldsymbol{X}_n\|^3 > |t - f(\boldsymbol{X}_n)|, |t - f(\boldsymbol{X}_n)| > \varepsilon\right) + \mathbb{P}\left(\frac{L_1}{6\sqrt{n}}\|\boldsymbol{X}_n\|^3 > |t - f(\boldsymbol{X}_n)|, |t - f(\boldsymbol{X}_n)| \le \varepsilon\right) \\
\le&\mathbb{P}\left(\frac{L_1}{6\sqrt{n}}\|\boldsymbol{X}_n\|^3 > \varepsilon\right) + \mathbb{P}\left(|t - f(\boldsymbol{X}_n)| \le \varepsilon\right).
\end{aligned}$$

When $q \ge 2$, from Proposition 4.B and the fact that $\mathbb{G}^{\mathrm{IEO}}$ has a bounded density

$$\begin{aligned}
\mathbb{P}(|t - f(\boldsymbol{X}_n)| \le \varepsilon) &= \mathbb{P}(\boldsymbol{X}_n \in f^{-1}(-\infty, t+\varepsilon)) - \mathbb{P}(\boldsymbol{X}_n \in f^{-1}(-\infty, t-\varepsilon)) \\
&\lesssim \mathbb{P}(\boldsymbol{X} \in f^{-1}(-\infty, t+\varepsilon)) - \mathbb{P}(\boldsymbol{X} \in f^{-1}(-\infty, t-\varepsilon)) + C_{n,q}^{\mathrm{IEO}} \\
&= \mathbb{P}(t - \varepsilon \le \mathbb{G}^{\mathrm{IEO}} \le t + \varepsilon) + C_{n,q}^{\mathrm{IEO}} \\
&\lesssim \left(\left(\sum_{i=1}^{q} \lambda_i^2\right)\left(\sum_{i=2}^{q} \lambda_i^2\right)\right)^{-\frac{1}{4}} \varepsilon + C_{n,q}^{\mathrm{IEO}}.
\end{aligned}$$

Moreover, by Lemma 11,

$$\begin{aligned}
\mathbb{P}\left(\frac{L_1}{6\sqrt{n}}\|\boldsymbol{X}_n\|^3 > \varepsilon\right) &\le \mathbb{P}\left(\frac{L_1}{6\sqrt{n}}\|\boldsymbol{X}\|^3 > \varepsilon\right) + C_{n,q}^{\mathrm{IEO}} \\
&\le 2q \exp\left(-\frac{1}{2q\left\|\boldsymbol{M}_1^{\mathrm{IEO}}\right\|^2}\left(\frac{6\sqrt{n}\varepsilon}{L_1}\right)^{2/3}\right) + C_{n,q}^{\mathrm{IEO}}.
\end{aligned}$$

Taking $\varepsilon = \frac{L_1}{6\sqrt{n}}(2q\left\|\boldsymbol{M}_1^{\mathrm{IEO}}\right\|^2 \log n)^{3/2}$, we have

$$P\left(\frac{L_1}{6\sqrt{n}}\|\boldsymbol{X}_n\|^3 > \varepsilon\right) \le 2qn^{-1} + C_{n,q}^{\mathrm{IEO}}.$$

Hence,

$$\mathbb{P}\left(\frac{L_1}{6\sqrt{n}}\|\boldsymbol{X}_n\|^3 > |t - f(\boldsymbol{X}_n)|\right) \lesssim \left(\left(\sum_{i=1}^{q} \lambda_i^2\right)\left(\sum_{i=2}^{q} \lambda_i^2\right)\right)^{-\frac{1}{4}} L_1\left\|\boldsymbol{M}_1^{\mathrm{IEO}}\right\|^3 q^{\frac{3}{2}}(\log n)^{\frac{3}{2}} n^{-\frac{1}{2}} + qn^{-1} + C_{n,q}^{\mathrm{IEO}}.$$

Combining things together, we have

$$\begin{aligned}
&\sup_{t \ge 0} |\mathbb{P}(nv_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - nv_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \ge t) - \mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \ge t)| \\
\lesssim &\left(\left(\sum_{i=1}^{q} \lambda_i^2\right)\left(\sum_{i=2}^{q} \lambda_i^2\right)\right)^{-\frac{1}{4}} L_1\left\|\boldsymbol{M}_1^{\mathrm{IEO}}\right\|^3 q^{3/2}(\log n)^{\frac{3}{2}} n^{-\frac{1}{2}} + qn^{-1} + C_{n,q}^{\mathrm{IEO}},
\end{aligned}$$

where "$\lesssim$" hides the constants that is independent of n and q.

For $q = 1$, we have $\mathbb{G}^{\mathrm{IEO}} = \lambda_1 \chi^2(1)$. For any $t, \varepsilon > 0$,

$$\mathbb{P}(t - \varepsilon \le \mathbb{G}^{\mathrm{IEO}} \le t + \varepsilon) = \mathbb{P}\left(\frac{1}{\lambda_1}(t - \varepsilon) \le \chi^2(1) \le \frac{1}{\lambda_1}(t + \varepsilon)\right) = 2\mathbb{P}\left(N(0,1) \in \left[\sqrt{\frac{(t-\varepsilon)_+}{\lambda_1}}, \sqrt{\frac{t+\varepsilon}{\lambda_1}}\right]\right).$$

If $t < \varepsilon$,

$$\mathbb{P}\left(N(0,1) \in \left[\sqrt{\frac{(t-\varepsilon)_+}{\lambda_1}}, \sqrt{\frac{t+\varepsilon}{\lambda_1}}\right]\right) \le \mathbb{P}\left(0 \le N(0,1) \le \sqrt{\frac{2\varepsilon}{\lambda_1}}\right) \lesssim \sqrt{\varepsilon/\lambda_1}.$$

If $t \geq \varepsilon$, $\sqrt{(t+\varepsilon)/\lambda_1} - \sqrt{(t-\varepsilon)/\lambda_1} \leq \sqrt{2\varepsilon/\lambda_1}$. Hence,

$$\mathbb{P}\left(N(0,1) \in \left[\sqrt{\frac{t-\varepsilon}{\lambda_1}}, \sqrt{\frac{t+\varepsilon}{\lambda_1}}\right]\right) \leq \mathbb{P}(0 \leq N(0,1) \leq \sqrt{\frac{2\varepsilon}{\lambda_1}}) \lesssim \sqrt{\varepsilon/\lambda_1}.$$

Taking $\varepsilon = \frac{L_1}{6\sqrt{n}}(2\left\|\boldsymbol{M}_1^{\mathrm{IEO}}\right\|^2 \log n)^{3/2}$ (where $q=1$), we have

$$\mathbb{P}\left(\frac{L_1}{6\sqrt{n}}\|\boldsymbol{X}_n\|^3 > |t - f(\boldsymbol{X}_n)|\right) \lesssim \frac{1}{\sqrt{\lambda_1}} L_1^{\frac{1}{2}} \left\|\boldsymbol{M}_1^{\mathrm{IEO}}\right\|^{\frac{3}{2}} (\log n)^{\frac{3}{4}} n^{-\frac{1}{4}} + n^{-1} + C_{n,1}^{\mathrm{IEO}}$$

Combining things together, we have

$$\sup_{t\geq 0} |\mathbb{P}(nv_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - nv_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \geq t) - \mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \geq t)| \lesssim \frac{1}{\sqrt{\lambda_1}} L_1^{\frac{1}{2}} \left\|\boldsymbol{M}_1^{\mathrm{IEO}}\right\|^{\frac{3}{2}} (\log n)^{\frac{3}{4}} n^{-\frac{1}{4}} + n^{-1} + C_{n,1}^{\mathrm{IEO}}.$$

□

We can immediately prove Corollary 2 by Proposition 2.B.

Proof of Corollary 2. We show the result for $q \geq 2$. The case of $q = 1$ can be proven similarly. By Proposition 2.B, there exists a problem dependent C_{prob} such that

$$\mathbb{P}(nv_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - nv_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \leq t) \geq \mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \leq t) - C_{\mathrm{prob}}(\log n)^{\frac{3}{2}} n^{-\frac{1}{2}}.$$

When n is larger than a threshold such that $C_{\mathrm{prob}}(\log n)^{\frac{3}{2}} n^{-\frac{1}{2}} \leq \varepsilon/2$, since $\mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \leq F^{-1}_{\mathbb{G}^{\mathrm{IEO}}}(1-\varepsilon/2)) = 1-\varepsilon/2$, we have

$$\mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \leq F^{-1}_{\mathbb{G}^{\mathrm{IEO}}}(1-\varepsilon/2)) - C_{\mathrm{prob}}(\log n)^{\frac{3}{2}} n^{-\frac{1}{2}} = 1 - \varepsilon/2 - C_{\mathrm{prob}}(\log n)^{\frac{3}{2}} n^{-\frac{1}{2}} \geq 1 - \varepsilon.$$

In conclusion, when n satisfies $C_{\mathrm{prob}}(\log n)^{\frac{3}{2}} n^{-\frac{1}{2}} \leq \varepsilon/2$, with probability at least $1-\varepsilon$, $R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \leq \kappa_0^{\mathrm{IEO}} + \frac{F^{-1}_{\mathbb{G}^{\mathrm{IEO}}}(1-\varepsilon/2)}{n}$. □

We can use the same techniques of Proposition 2.B to build the proof of Proposition 2.A.

Proof of Proposition 2.A. The idea is similar to Proposition 2.B. First, we note by Assumption 5 and Lemma 12 that for all $\boldsymbol{\theta} \in \Theta$, v_0 satisfies:

$$|v_0(\boldsymbol{w}_{\boldsymbol{\theta}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) - \nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{KL}}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{KL}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{KL}})| \leq \frac{L_1}{6}\|\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{KL}}\|^3. \quad (7)$$

For simplicity, let $\boldsymbol{X}_n := \sqrt{n}(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}})$, $\boldsymbol{X} := \mathbb{N}_1^{\mathrm{ETO}}$ and denote the quadratic function $f : \mathbb{R}^q \to \mathbb{R}$, $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{x}$. By the assumption, we know that $f(\boldsymbol{X})$ is a convex function and has a convex sub-level set. The remaining analysis is almost the same by setting

$$\begin{aligned} Y_n :=& v_0(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) - \nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{KL}}) - \frac{1}{2}(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}})^\top \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}}) \\ \leq& \frac{L_1}{6}\|\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}}\|^3, \end{aligned}$$

with $|nY_n| \leq \frac{L_1}{6\sqrt{n}}\|\boldsymbol{X}_n\|^3$. The remaining analysis is exactly the same as Proposition 2.B.

□

B.4 Proofs in Section 4.3

Proof of Theorem 4. Case 1: $t \le \kappa_0^{\mathrm{IEO}}$. In this case, $R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge R(\boldsymbol{w}_{\boldsymbol{\theta}^*}) = \kappa_0^{\mathrm{IEO}} \ge t$ and $R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge R(\boldsymbol{w}_{\boldsymbol{\theta}^*}) = \kappa_0^{\mathrm{IEO}} \ge t$ as any realization of $\hat{\boldsymbol{w}}^{\mathrm{IEO}}$ or $\hat{\boldsymbol{w}}^{\mathrm{ETO}}$ should have larger regret than $\boldsymbol{w}_{\boldsymbol{\theta}^*}$. So

$$\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge t) = 1 - 1 = 0.$$

Case 2. For any $t > \kappa_0^{\mathrm{IEO}}$, by Proposition 1.B,

$$\begin{aligned}\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge t) &= \mathbb{P}\left(v_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*}) \ge t - \kappa_0^{\mathrm{IEO}}\right)\\ &= \mathbb{P}\left(\sqrt{n}\left(v_0(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})\right) \ge \sqrt{n}(t - \kappa_0^{\mathrm{IEO}})\right)\\ &\le G^{\mathrm{IEO}}_{n,q,\sqrt{n}(t-\kappa_0^{\mathrm{IEO}})}.\end{aligned}$$

On the other hand, by Proposition 1.A,

$$\begin{aligned}&\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge t) - 1\\ =& -\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \le t)\\ =& -\mathbb{P}(v_0(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) \le t - \kappa_0^{\mathrm{ETO}})\\ =& -\mathbb{P}(\sqrt{n}(v_0(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) - v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})) \le \sqrt{n}(t - \kappa_0^{\mathrm{ETO}}))\\ \ge& -\mathbb{P}(\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\mathbb{N}_1^{\mathrm{ETO}} \le \sqrt{n}(t - \kappa_0^{\mathrm{ETO}})) - G^{\mathrm{ETO}}_{n,q}.\end{aligned}$$

Thus,

$$\begin{aligned}&\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge t)\\ \le& -1 + \mathbb{P}(\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\mathbb{N}_1^{\mathrm{ETO}} \le \sqrt{n}(t - \kappa_0^{\mathrm{ETO}})) + G^{\mathrm{IEO}}_{n,q,\sqrt{n}(t-\kappa_0^{\mathrm{IEO}})} + G^{\mathrm{ETO}}_{n,q}\end{aligned}$$

or

$$\begin{aligned}&\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge t)\\ \ge& 1 - \mathbb{P}(\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\mathbb{N}_1^{\mathrm{ETO}} \le \sqrt{n}(t - \kappa_0^{\mathrm{ETO}})) - G^{\mathrm{IEO}}_{n,q,\sqrt{n}(t-\kappa_0^{\mathrm{IEO}})} - G^{\mathrm{ETO}}_{n,q}\end{aligned}$$

We further discuss the case when (2) $\kappa_0^{\mathrm{IEO}} < t < \kappa_0^{\mathrm{ETO}}$, (3) $t = \kappa_0^{\mathrm{ETO}}$, (4) $t > \kappa_0^{\mathrm{ETO}}$.

When $\kappa_0^{\mathrm{IEO}} < t < \kappa_0^{\mathrm{ETO}}$, we have $\mathbb{P}(\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\mathbb{N}_1^{\mathrm{ETO}} \le \sqrt{n}(t - \kappa_0^{\mathrm{ETO}})) \le \exp\left(-\frac{n(\kappa_0^{\mathrm{ETO}} - t)^2}{2\left\|\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}}\right\|^2}\right)$ by Lemma 11. So

$$\begin{aligned}&\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge t)\\ \ge& 1 - \exp\left(-\frac{n(\kappa_0^{\mathrm{ETO}} - t)^2}{2\left\|\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}}\right\|^2}\right) - G^{\mathrm{IEO}}_{n,q,\sqrt{n}(t-\kappa_0^{\mathrm{IEO}})} - G^{\mathrm{ETO}}_{n,q}\end{aligned}$$

When $t = \kappa_0^{\mathrm{ETO}}$, we have $\mathbb{P}(\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\mathbb{N}_1^{\mathrm{ETO}} \le \sqrt{n}(t - \kappa_0^{\mathrm{ETO}})) = \frac{1}{2}$. So

$$\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge t) \ge \frac{1}{2} - G^{\mathrm{IEO}}_{n,q,\sqrt{n}(t-\kappa_0^{\mathrm{IEO}})} - G^{\mathrm{ETO}}_{n,q}$$

When $t > \kappa_0^{\mathrm{ETO}}$, we have $1 - \mathbb{P}(\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\mathbb{N}_1^{\mathrm{ETO}} \le \sqrt{n}(t - \kappa_0^{\mathrm{ETO}})) \ge 0$. So

$$\begin{aligned}&\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge t)\\ \ge& -G^{\mathrm{IEO}}_{n,q,\sqrt{n}(t-\kappa_0^{\mathrm{IEO}})} - G^{\mathrm{ETO}}_{n,q}.\end{aligned}$$

□

Proof of Theorem 5. In the following analysis, we define $t' := t - \kappa_0^{\mathrm{IEO}}$.

Case 1: $t \le \kappa_0^{\mathrm{IEO}}$. In this case, $R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge R(\boldsymbol{w}_{\boldsymbol{\theta}^*}) = \kappa_0^{\mathrm{IEO}} \ge t$ and $R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge R(\boldsymbol{w}_{\boldsymbol{\theta}^*}) = \kappa_0^{\mathrm{IEO}} \ge t$ as any realization of $\hat{\boldsymbol{w}}^{\mathrm{IEO}}$ or $\hat{\boldsymbol{w}}^{\mathrm{ETO}}$ should have larger regret than $\boldsymbol{w}_{\boldsymbol{\theta}^*}$. So

$$\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge t) = 1 - 1 = 0.$$

Case 2: $t > \kappa_0^{\mathrm{IEO}} + \frac{\tau_6+\tau_1}{\tau_1}\delta$, i.e., $t' > \frac{\tau_6+\tau_1}{\tau_1}\delta$. In this case, we have that $(1 + \frac{\tau_1}{\tau_6})(nt' - n\delta) > nt'$. Suppose $\delta > 0$ and consider any $0 < \varepsilon < \frac{\tau_6+\tau_1}{\tau_1}t' - \delta$.

We first observe that the Berry-Esseen bound implies the following inequality. Since $\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\cdot)$ is an affine function:

$$\sup_{t\ge 0} |\mathbb{P}(\sqrt{n}\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}}) \ge t) - \mathbb{P}(\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\mathbb{N}_1^{\mathrm{ETO}} \ge t)| \le C_{n,q}^{\mathrm{ETO}}.$$

For any sample size n and any given $\varepsilon > 0$, we have

$$\begin{aligned}
&\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge t)\\
=&\mathbb{P}(nR(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \ge nt) - \mathbb{P}(nR(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \ge nt)\\
=&\mathbb{P}(n(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) - R(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})) \ge nt - n\kappa_0^{\mathrm{ETO}}) - \mathbb{P}(n(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - R(\boldsymbol{w}_{\boldsymbol{\theta}^*})) \ge nt - n\kappa_0^{\mathrm{IEO}})\\
=&\mathbb{P}(n(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) - R(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})) \ge nt' - n\delta) - \mathbb{P}(n(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - R(\boldsymbol{w}_{\boldsymbol{\theta}^*})) \ge nt')\\
&\qquad (\text{where } \delta = \kappa_0^{\mathrm{ETO}} - \kappa_0^{\mathrm{IEO}} \ge 0 \ t' := t - \kappa_0^{\mathrm{IEO}})\\
\le&\mathbb{P}(n\left(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) - \kappa_0^{\mathrm{ETO}} - \nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}})\right) \ge nt' - n\delta - n\varepsilon)\\
&+ \mathbb{P}(n\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})(\hat{\boldsymbol{\theta}}^{\mathrm{ETO}} - \boldsymbol{\theta}^{\mathrm{KL}})) \ge n\varepsilon)\\
&- \mathbb{P}(n(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - R(\boldsymbol{w}_{\boldsymbol{\theta}^*})) \ge nt')\\
\le&\mathbb{P}(\mathbb{G}^{\mathrm{ETO}} \ge nt' - n\delta - n\varepsilon) + \mathbb{P}(\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}}\boldsymbol{Y}_0 \ge \sqrt{n}\varepsilon)\\
&- \mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \ge nt') + D_{n,q}^{\mathrm{ETO}} + D_{n,q}^{\mathrm{IEO}} + C_{n,q}^{\mathrm{ETO}}\\
\le&\mathbb{P}(\mathbb{G}^{\mathrm{ETO}} \ge nt' - n\delta - n\varepsilon) - \mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \ge nt') + D_{n,q}^{\mathrm{ETO}} + D_{n,q}^{\mathrm{IEO}} + C_{n,q}^{\mathrm{ETO}} + E_n^{\delta,\varepsilon}
\end{aligned}$$

where, by Lemma 11,

$$E_n^{\delta,\varepsilon} = \exp\left(-\frac{n\varepsilon^2}{2\|\nabla_{\boldsymbol{\theta}} v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}}\|^2}\right).$$

Therefore, we only need to provide the bounds for

$$\mathbb{P}(\mathbb{G}^{\mathrm{ETO}} \ge nt' - n\delta - n\varepsilon) - \mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \ge nt').$$

To achieve this, we can apply Theorem 3 to claim that for all $s \ge 0$,

$$\mathbb{P}(\mathbb{G}^{\mathrm{ETO}} \ge s) \le \mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \ge (1 + \frac{\tau_1}{\tau_6})s) \le \mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \ge s).$$

To see this, we note that

$$\begin{aligned}
&\mathbb{P}(\mathbb{G}^{\mathrm{ETO}} \geq s) \\
=&\mathbb{P}\Big(\frac{1}{2}\boldsymbol{Y}_0^\top(\boldsymbol{M}_1^{\mathrm{ETO}})^\top\nabla_{\boldsymbol{\theta\theta}}v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}}\boldsymbol{Y}_0 \geq s\Big) \\
=&\mathbb{P}\Big(\frac{1}{2}\boldsymbol{Y}_0^\top(\boldsymbol{M}_1^{\mathrm{ETO}})^\top\nabla_{\boldsymbol{\theta\theta}}v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}}\boldsymbol{Y}_0 \geq s \text{ and } \frac{1}{2}\tau_6\boldsymbol{Y}_0^\top\boldsymbol{Y}_0 \geq s\Big) \\
&\qquad\qquad\text{(since } \tau_6 \text{ is the largest eigenvalue of } (\boldsymbol{M}_1^{\mathrm{ETO}})^\top\nabla_{\boldsymbol{\theta\theta}}v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}}) \\
=&\mathbb{P}\Big(\frac{1}{2}\boldsymbol{Y}_0^\top(\boldsymbol{M}_1^{\mathrm{ETO}})^\top\nabla_{\boldsymbol{\theta\theta}}v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}}\boldsymbol{Y}_0 \geq s \text{ and } \frac{1}{2}\tau_1\boldsymbol{Y}_0^\top\boldsymbol{Y}_0 \geq \frac{\tau_1}{\tau_6}s\Big) \\
=&\mathbb{P}\Big(\frac{1}{2}\boldsymbol{Y}_0^\top(\boldsymbol{M}_1^{\mathrm{ETO}})^\top\nabla_{\boldsymbol{\theta\theta}}v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}}\boldsymbol{Y}_0 \geq s \\
&\quad\text{and } \frac{1}{2}\boldsymbol{Y}_0^\top((\boldsymbol{M}_1^{\mathrm{IEO}})^\top\nabla_{\boldsymbol{\theta\theta}}v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})\boldsymbol{M}_1^{\mathrm{IEO}} - (\boldsymbol{M}_1^{\mathrm{ETO}})^\top\nabla_{\boldsymbol{\theta\theta}}v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}})\boldsymbol{Y}_0 \geq \frac{\tau_1}{\tau_6}s\Big) \\
&\text{(since } \tau_1 \text{ is the smallest eigenvalue of } ((\boldsymbol{M}_1^{\mathrm{IEO}})^\top\nabla_{\boldsymbol{\theta\theta}}v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})\boldsymbol{M}_1^{\mathrm{IEO}} - (\boldsymbol{M}_1^{\mathrm{ETO}})^\top\nabla_{\boldsymbol{\theta\theta}}v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}})\boldsymbol{M}_1^{\mathrm{ETO}})) \\
\leq&\mathbb{P}\Big(\frac{1}{2}\boldsymbol{Y}_0^\top(\boldsymbol{M}_1^{\mathrm{IEO}})^\top\nabla_{\boldsymbol{\theta\theta}}v_0(\boldsymbol{w}_{\boldsymbol{\theta}^*})\boldsymbol{M}_1^{\mathrm{IEO}}\boldsymbol{Y}_0 \geq (1+\frac{\tau_1}{\tau_6})s\Big) \\
=&\mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \geq (1+\frac{\tau_1}{\tau_6})s).
\end{aligned}$$

This shows that

$$\begin{aligned}
&\mathbb{P}(\mathbb{G}^{\mathrm{ETO}} \geq nt' - n\delta - n\varepsilon) - \mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \geq nt') \\
\leq&\mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \geq (1+\frac{\tau_1}{\tau_6})(nt' - n\delta - n\varepsilon)) - \mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \geq nt') \\
=&-\mathbb{P}(nt' \leq \mathbb{G}^{\mathrm{IEO}} \leq (1+\frac{\tau_1}{\tau_6})(nt' - n\delta - n\varepsilon))
\end{aligned}$$

In conclusion,

$$\begin{aligned}
&\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \geq t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \geq t) \\
\leq&\mathbb{P}(\mathbb{G}^{\mathrm{ETO}} \geq nt - n\kappa_0^{\mathrm{IEO}} - n\delta - n\varepsilon) - \mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \geq nt - n\kappa_0^{\mathrm{IEO}}) + D_{n,q}^{\mathrm{ETO}} + D_{n,q}^{\mathrm{IEO}} + C_{n,q}^{\mathrm{ETO}} + E_n^{\delta,\varepsilon} \\
\leq&-\mathbb{P}(nt - n\kappa_0^{\mathrm{IEO}} \leq \mathbb{G}^{\mathrm{IEO}} \leq (1+\frac{\tau_1}{\tau_6})(nt - n\kappa_0^{\mathrm{IEO}} - n\delta - n\varepsilon)) + D_{n,q}^{\mathrm{ETO}} + D_{n,q}^{\mathrm{IEO}} + C_{n,q}^{\mathrm{ETO}} + E_n^{\delta,\varepsilon}.
\end{aligned}$$

Finally, we remark that when $\delta = 0$, $\theta^* = \theta^{\mathrm{KL}}$, $\nabla_{\boldsymbol{\theta}}v_0(\boldsymbol{w}_{\boldsymbol{\theta}^{\mathrm{KL}}}) = 0$, and $\kappa_0^{\mathrm{ETO}} = \kappa_0^{\mathrm{IEO}}$. In this case, there is no need to introduce ε, and we have similarly

$$\begin{aligned}
&\mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) \geq t) - \mathbb{P}(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) \geq t) \\
=&\mathbb{P}(n(R(\hat{\boldsymbol{w}}^{\mathrm{ETO}}) - R(\boldsymbol{w}_{\boldsymbol{\theta}^*})) \geq nt') - \mathbb{P}(n(R(\hat{\boldsymbol{w}}^{\mathrm{IEO}}) - R(\boldsymbol{w}_{\boldsymbol{\theta}^*})) \geq nt') \\
\leq&\mathbb{P}(\mathbb{G}^{\mathrm{ETO}} \geq nt') - \mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \geq nt') + D_{n,q}^{\mathrm{ETO}} + D_{n,q}^{\mathrm{IEO}} \\
\leq&\mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \geq (1+\frac{\tau_1}{\tau_6})nt') - \mathbb{P}(\mathbb{G}^{\mathrm{IEO}} \geq nt') + D_{n,q}^{\mathrm{ETO}} + D_{n,q}^{\mathrm{IEO}} \\
=&-\mathbb{P}(nt' \leq \mathbb{G}^{\mathrm{IEO}} \leq (1+\frac{\tau_1}{\tau_6})nt') + D_{n,q}^{\mathrm{ETO}} + D_{n,q}^{\mathrm{IEO}}.
\end{aligned}$$

□