The Bias-Variance Tradeoff in Data-Driven Optimization: A Local Misspecification Perspective

Haixiang Lan[1]*, Luofeng Liao[1]*, Adam N. Elmachtoub[1],
Christian Kroer[1], Henry Lam[1], Haofeng Zhang[1,2]
[1]Department of Industrial Engineering and Operations Research, Columbia University
[2]Morgan Stanley
{hl3725,ll3530,ae2516,ck2945,khl2114,hz2553}@columbia.edu

Abstract

Data-driven stochastic optimization is ubiquitous in machine learning and operational decision-making problems. Sample average approximation (SAA) and model-based approaches such as estimate-then-optimize (ETO) or integrated estimation-optimization (IEO) are all popular, with model-based approaches being able to circumvent some of the issues with SAA in complex context-dependent problems. Yet the relative performance of these methods is poorly understood, with most results confined to the dichotomous cases of the model-based approach being either well-specified or misspecified. We develop the first results that allow for a more granular analysis of the relative performance of these methods under a local misspecification setting, which models the scenario where the model-based approach is nearly well-specified. By leveraging tools from contiguity theory in statistics, we show that there is a bias-variance tradeoff between SAA, IEO, and ETO under local misspecification, and that the relative importance of the bias and the variance depends on the degree of local misspecification. Moreover, we derive explicit expressions for the decision bias, which allows us to characterize (un)impactful misspecification directions, and provide further geometric understanding of the variance.

1 Introduction

Data-driven stochastic optimization arises ubiquitously in machine learning and operational decision-making problems. Generally, this problem takes the form $\operatorname{argmin}_{\boldsymbol{w}} \mathbb{E}_Q\left[c(\boldsymbol{w}, \boldsymbol{z})\right]$ where $\boldsymbol{w}$ represents the decision that we aim to optimize, and $\boldsymbol{z}$ is a random variable (or vector) drawn from an unknown distribution Q. The non-linear cost function c is given and can represent a loss function in machine learning, financial portfolio losses, or resource allocation costs. In this paper, we focus on the setting where the expectation $\mathbb{E}_Q$ is unknown, but we observe data from Q.

A natural approach is to use empirical optimization, known as sample average approximation (SAA) [Shapiro et al., 2021], which approximates the unknown expectation with the empirical counterpart from the data. This approach is straightforward, but may not be suitable for complex scenarios in constrained and contextual optimization, when one needs to obtain a feature-dependent decision (i.e., a decision as a "function" of the features) and maintain feasibility [Hu et al., 2022]. In such cases, model-based approaches provide a workable alternative. A model-based approach fits a parametric distribution class to the data, say $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, and this fitted distribution is then injected into the downstream optimization to obtain a decision. Just as in standard machine-learning problems, this parametrization helps maintain generalizability from supervised data.

*Equal Contribution and Corresponding Authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Our focus in this work is the statistical performance of model-based approaches for data-driven optimization with nonlinear cost objectives, as compared to SAA. More specifically, we study the question of *how to fit* the data into the parametric distribution models. There have been two major methods proposed in the literature: *Estimation-then-Optimize (ETO)* and *Integrated Estimation-Optimization (IEO)*. ETO separates the fitting step from the downstream optimization, by simply fitting $P_{\boldsymbol{\theta}}$ via maximum likelihood estimation (MLE). IEO, on the other hand, integrates the downstream objective with the estimation process, by selecting the distribution parameter $\boldsymbol{\theta}$ that minimizes the empirical expected cost. Conceptually, ETO can readily leverage existing machine learning tools and fits the model disjointly from the downstream decision task, while IEO attempts to take into account the downstream decision task (in many cases with a nontrivial additional computational expense). Intuitively, then, IEO should have better statistical performance than ETO in terms of the ultimate objective value of the chosen decision.

Our main goal is to dissect the bias-variance tradeoff between ETO, IEO, and also SAA, especially under the setting of *model misspecification*. In particular, our study reveals not only the variance arising from data noise, but also the bias of the resulting decision elicited under model misspecification. This allows us to gain insight on how the direction and amount of misspecification impacts decision quality. More concretely, a well-specified model means that in the estimation-optimization pipeline, the chosen parametric class $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ contains the ground-truth distribution Q – a case that is rarely seen in reality. In other words, model misspecification arises generically, the question only being by how much. Unfortunately, the theoretical understanding of the statistical performance among the various estimation-optimization approaches, especially in relation to this misspecification, has been rather limited. Elmachtoub et al. [2023] compare these approaches in large-sample regimes via stochastic dominance, but their analysis divides into the cases of well-specification and misspecification, each with different asymptotic scaling. Unfortunately, there is no smooth transition in between that captures the impact of varying the misspecification amount. Elmachtoub et al. [2025] attempt to address this issue by deriving finite-sample bounds that depend on the sample size and misspecification amount.

In this paper, we remedy the shortcomings in the above literature by leveraging the notion of *local misspecification*, originated from contiguity theory in statistics [Le Cam and Yang, 2000, Copas and Eguchi, 2001, Andrews et al., 2020], to derive large-sample results in relation to both the amount and *direction* of misspecification. Our results explicitly show the decision bias and variance, and its resulting regret, that arises from misspecification. This allows us to smoothly compare ETO, IEO and SAA. We show the following results. When model misspecification is severe relative to the data noise level, SAA performs better than IEO, and IEO performs better than ETO, in terms of both bias and regret. This matches the intuition described previously that IEO should outperform ETO by integrating the estimation-optimization pipeline. On the other hand, when the misspecification amount is mild, the performance ordering is reversed, which generalizes previous similar findings in Elmachtoub et al. [2023] that focused on zero misspecification. Most importantly, in the most relevant case where the misspecification is roughly similar to the data noise, which we call the *balanced* case, the ordering of the methods exhibits a bias-variance tradeoff: SAA performs the best on bias, whereas ETO performs the best on variance, and IEO is in the middle for both metrics. This defies a universal performance ordering, but also points to the need for a deep understanding of the characteristics of the bias term in relation to the misspecification direction. Table 1 summarizes our performance ordering findings.

	mild ($\alpha > 1/2$)			balanced ($\alpha = 1/2$)			severe ($0 < \alpha < 1/2$)		
	bias	variance	regret	bias	variance	regret	bias	variance	regret
ETO	≈ 0	best	best	worst	best	depends	worst	≈ 0	worst
IEO	≈ 0	middle	middle	middle	middle	depends	middle	≈ 0	middle
SAA	≈ 0	worst	worst	best (≈ 0)	worst	depends	best (≈ 0)	≈ 0	best

Table 1: Summary of our results on performance orderings. "≈ 0" means asymptotically negligible. α is a parameter that signals the misspecification amount relative to the data noise level and will be detailed later.

Our next contribution is to provide an explicit formula for the bias attributed to model misspecification, which allows us to gain insights into how the bias is impacted by the misspecification direction. We went beyond the classical local minimax theory by showing the *non-regularity* of the ETO and

IEO estimators, which is rarely seen in standard statistical literature [van der Vaart, 2000]. In the severely misspecified regime, where there is no available contiguity theory tools, we develop a novel technique to characterize and compare the asymptotics of the three estimators. We further identify sufficient conditions on *approximately impactless* misspecification directions – model misspecification directions that result in bias that is first-order negligible compared to the data noise variance. In general, this direction is orthogonal to the difference between the influence functions of solutions obtained from the considered estimation-optimization pipeline and SAA. Here, influence functions are interpreted as the gradients with respect to the underlying data distribution, and they appear not only in the bias but also variance comparisons. This characterization in particular suggests how SAA is always (and naturally) the best in terms of bias (see Table 1), but also how the biases for IEO and ETO magnify when the obtained solution has a different influence function from that of SAA. Moreover, to enhance the transparency of our characterization, we show that a sufficient condition for being approximately impactless is to be in the linear span of the score function of the parametric model. While this latter condition is imposed purely on the parametric model (i.e., not the downstream optimization), it already shows the intriguing phenomenon where model misspecification could be insignificant in impacting the performance of the ultimate decision.

1.1 Related Works

Data-Driven Stochastic Optimization. Data-driven optimization, a cornerstone in machine learning and operations research, addresses problems where decision are informed from optimization problems with parameters or distributions learned from data. Existing popular methods include SAA [Shapiro et al., 2021] and distributionally robust optimization (DRO) [Delage and Ye, 2010]. Recently, there has been a growing interest in an integrated framework that combines predictive modeling of unknown parameters with downstream optimization tasks [Kao et al., 2009, Donti et al., 2017]. When the cost function is linear, Elmachtoub and Grigas [2022] propose a "Smart-Predict-then-Optimize" (SPO) approach that integrates prediction with optimization to improve decision-making. Recent literature explore further properties of the SPO approach [Mandi et al., 2020, Blondel et al., 2020, Ho-Nguyen and Kılınç-Karzan, 2022, Liu and Grigas, 2022, Liu et al., 2023, El Balghiti et al., 2023]. Hu et al. [2022] further compare the performances of different data-driven approaches in the linear cost function setting. In the context of non-linear cost functions, Grigas et al. [2021] propose an integrated approach tailored to discrete distributions, and Lam [2021] compares SAA with DRO and Bayesian extensions.

Local Misspecification. Model misspecification is extensively studied in the statistics and econometrics and machine learning literature [Marsden et al., 2021]. In this paper we focus on local misspecification, where the magnitude of misspecification vanishes as the sample size grows. Newey [1985] analyzes the asymptotic power properties of the generalized method of moments tests under a sequence of local misspecified alternatives. Kang and Schafer [2007] design a doubly robust procedure to estimate the population mean under the local misspecified models with incomplete data. Copas and Eguchi [2001, 2005] discuss the impact of local misspecification on the sensitivity of likelihood-based statistical inference under the asymptotic framework. Local misspecification is also discussed in robust estimation [Kitamura et al., 2013, Armstrong et al., 2023], causal inference [Conley et al., 2012, Fan et al., 2022], econometrics [Bugni et al., 2012, Andrews et al., 2017, 2020, Bugni and Ura, 2019, Armstrong and Kolesár, 2021, Bonhomme and Weidner, 2022, Candelaria and Zhang, 2024] and reinforcement learning [Dong et al., 2023]. To the best of our knowledge, we are the first to study the impact of local misspecification in data-driven optimization.

2 Settings and Methodologies

2.1 Data-Driven Stochastic Optimization

Consider a data-driven optimization problem in the following form:

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

where Ω is an open subset in $\mathbb{R}^{d_w}$, $\boldsymbol{z} \in \mathcal{Z} \subset \mathbb{R}^{d_z}$ is the uncertain parameter with unknown data-generating distribution Q, $c(\cdot, \cdot)$ is a known *non-linear* cost function, and $v_0(\cdot)$ is the expectation of the cost function under ground-truth distribution Q under a decision $\boldsymbol{w}$. We are given independent

and identically distributed (i.i.d.) data $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ drawn from Q, and the goal is to approximate the optimal decision $\boldsymbol{w}^*$ using the data.

In model-based approaches, we use a parametric distribution family $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{d_\theta}$ is the model parameter and Θ is an open subset of $\mathbb{R}^{d_\theta}$. To explain further, we define the *oracle* solution $\boldsymbol{w}_{\boldsymbol{\theta}}$ by

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

where $v(\boldsymbol{w}, \boldsymbol{\theta})$ is the expected cost function under the distribution $P_{\boldsymbol{\theta}}$. Depending on the choice of the model, the ground-truth distribution Q may or may not be in the parametric family $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. We say $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is well-specified if there exists $\boldsymbol{\theta}_0 \in \boldsymbol{\theta}$ such that $P_{\boldsymbol{\theta}_0} = Q$. We say $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is misspecified if it is not well-specified.

Notations. We denote $\mathbb{E}_{\tilde{P}}[\cdot]$ and $\text{var}_{\tilde{P}}$ as the expectation and variance under the distribution $\tilde{P}$. We denote $\mathbb{E}_{\boldsymbol{\theta}}[\cdot]$ as $\mathbb{E}_{P_{\boldsymbol{\theta}}}[\cdot]$ and $\text{var}_{\boldsymbol{\theta}}(\cdot) = \text{var}_{P_{\boldsymbol{\theta}}}(\cdot)$ in the parametric case. For a symmetric matrix $\boldsymbol{A}$, we write $\boldsymbol{A} \succeq \boldsymbol{0}$ if it is positive semi-definite and $\boldsymbol{A} \succ \boldsymbol{0}$ if it is positive definite. For two symmetric matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, we write $\boldsymbol{A}_1 \succeq \boldsymbol{A}_2$ if $\boldsymbol{A}_1 - \boldsymbol{A}_2 \succeq \boldsymbol{0}$ and $\boldsymbol{A}_1 \succ \boldsymbol{A}_2$ if $\boldsymbol{A}_1 - \boldsymbol{A}_2 \succ \boldsymbol{0}$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, we define the column span of $\boldsymbol{A}$ as $\text{col}(\boldsymbol{A}) = \{\boldsymbol{A}\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^n\}$. For a vector $\boldsymbol{x} \in \mathbb{R}^n$ and a positive semi-definite matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, we define the matrix-induced norm $\|\boldsymbol{x}\|_{\boldsymbol{A}} := \sqrt{\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}}$. We define $\xrightarrow{P^n}$ as convergence in distribution under the measure P^n. Precisely, $X_n \xrightarrow{P^n} X$ if $P^n(X_n \leq t) \to \mathbb{P}(X \leq t)$ for all continuous points of the distribution of X where $\mathbb{P}$ denotes the distribution of X. For a sequence of random variables $\{X_n\}_{n=1}^\infty$, we say $X_n = O_{P^n}(1)$ if it is stochastically bounded under the probability measure P^n, i.e., for all $\delta > 0$ there exists M and $N \in \mathbb{N}$ such that for all $n \geq N$, $P^n(|X_n| > M) \leq \delta$. We say $X_n = o_{P^n}(1)$ if it converges to zero in probability under the probability measure P^n, i.e., $X_n = o_{P^n}(1)$ if for all $\varepsilon > 0$, $\lim_{n \to \infty} P^n(|X_n| > \varepsilon) = 0$. More generally, we denote $X_n = O_{P^n}(a_n)$ if $X_n / a_n = O_{P^n}(1)$ and we denote $X_n = o_{P^n}(a_n)$ if $X_n / a_n = o_{P^n}(1)$. For two random variables X_1, X_2 with distribution $\mathbb{P}_1$ and $\mathbb{P}_2$, we say X_1 is (first-order) stochastically dominated by X_2, denoted as $X_1 \preceq_{\text{st}} X_2$, if for all $x \in \mathbb{R}$, it satisfies that $\mathbb{P}_1(X_1 > x) \leq \mathbb{P}_2(X_2 > x)$.

2.2 Three Data-Driven Methods

We consider three popular approaches for solving (1) in a data-driven fashion.

Sample Average Approximation (SAA). SAA simply replaces $\mathbb{E}_Q$ in (1) with its empirical counterpart. More precisely, we solve $\hat{\boldsymbol{w}}^{\text{SAA}} := \operatorname{argmin}_{\boldsymbol{w} \in \Omega} \left\{ \hat{v}_0(\boldsymbol{w}) := \frac{1}{n} \sum_{i=1}^n c(\boldsymbol{w}, \boldsymbol{z}_i) \right\}$.

Estimate-Then-Optimize (ETO). ETO uses maximum likelihood estimation (MLE) to infer $\boldsymbol{\theta}$ by solving $\hat{\boldsymbol{\theta}}^{\text{ETO}} = \sup_{\boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^n \log p_{\boldsymbol{\theta}}(\boldsymbol{z}_i)$ Here $p_{\boldsymbol{\theta}}$ is the probability density or mass function. Then we plug $\hat{\boldsymbol{\theta}}^{\text{ETO}}$ into (2): $\hat{\boldsymbol{w}}^{\text{ETO}} := \boldsymbol{w}_{\hat{\boldsymbol{\theta}}^{\text{ETO}}} = \operatorname{argmin}_{\boldsymbol{w} \in \Omega} v(\boldsymbol{w}, \boldsymbol{\theta}^{\text{ETO}})$.

Integrated-Estimation-Optimization (IEO). IEO selects the $\boldsymbol{\theta}$ that performs best on the empirical cost function $\hat{v}_0(\cdot)$ evaluated at $\boldsymbol{w}_{\boldsymbol{\theta}}$. More precisely, we solve $\inf_{\boldsymbol{\theta} \in \Theta} \hat{v}_0(\boldsymbol{w}_{\boldsymbol{\theta}})$ and get a solution $\hat{\boldsymbol{\theta}}^{\text{IEO}}$. Then we use the plug-in estimator $\hat{\boldsymbol{w}}^{\text{IEO}} := \boldsymbol{w}_{\hat{\boldsymbol{\theta}}^{\text{IEO}}} = \operatorname{argmin}_{\boldsymbol{w} \in \Omega} v(\boldsymbol{w}, \boldsymbol{\theta}^{\text{IEO}})$.

Among the three methods, SAA is model-free while ETO and IEO are model-based. ETO separates estimation (via MLE) with downstream optimization, while IEO integrates the latter into the estimation process. Our primary focus is to statistically compare these three data-to-decision pipelines in the so-called locally misspecified regime, which we discuss next.

Throughout the paper we assume certain classical technical assumptions on the cost c and the distribution $P_{\boldsymbol{\theta}}$ to ensure the asymptotic normality of certain M-estimators under $P_{\boldsymbol{\theta}_0}$ and the Cramer-Rao lower bound. In particular, they require relevant population minimizers of $\min_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{\theta}_0}[c(\boldsymbol{w}, \boldsymbol{z})]$, $\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}_0}[-\log p_{\boldsymbol{\theta}}(z)]$ and $\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}_0}[c(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z})]$ are uniquely attained in the interior of the parameter spaces. See Assumptions 3 and 4 in the Appendix for precise statements.

2.3 Local Misspecification

We first explain local misspecification intuitively before providing formal definitions. Recall the ground-truth data generating distribution Q, and our parametric distribution family (i.e., model)

$\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. At a high level, we assume that for a finite data size n, Q could deviate from the model in a certain "direction". However, as n is sufficiently large, we expect to have a more accurate model, and Q will approach a distribution in $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, which we denote as $P_{\boldsymbol{\theta}_0}$ for some $\boldsymbol{\theta}_0 \in \Theta$. In other words, $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is misspecified to Q in a "local" sense, and such misspecification will ultimately vanish. From now on, we use $P_{\boldsymbol{\theta}_0}$ to represent this distribution. Note that neither $P_{\boldsymbol{\theta}_0}$ nor $\boldsymbol{\theta}_0$ are known in practice.

To introduce the local misspecification regime, we first formally define a notion called local perturbation for describing the deviation between two distributions. We work with a general form of local perturbation [van der Vaart, 2000, Fan et al., 2022, Duchi and Ruan, 2021, Duchi, 2021] where the ground truth distribution Q is related to $P_{\boldsymbol{\theta}_0}$ through a general tilt the distribution (we will provide many classical examples in Appendix A.2):

Definition 1 (Local Perturbation)**.** Consider a scalar function $u(\boldsymbol{z}) : \mathcal{Z} \to \mathbb{R}$ with zero mean $\mathbb{E}_{\boldsymbol{\theta}_0}[u] = 0$ and finite second order moment $\mathbb{E}_{\boldsymbol{\theta}_0}[u^2]$. We define a tilted distribution Q_t for $t \in \mathbb{R}$ with probability density (mass) function q_t with respect to the dominated measure (note that Q_t is not necessarily in the parametric family $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$) with $q_0 = p_{\boldsymbol{\theta}_0}$. We further assume for all t, q_t is differentiable for almost every $\boldsymbol{z}$, as well as the quadratic mean differentiability condition:

$$\int \left(\sqrt{q_t} - \sqrt{p_{\boldsymbol{\theta}_0}} - \frac{1}{2} t u \sqrt{p_{\boldsymbol{\theta}_0}} \right)^2 d\boldsymbol{z} = o(t^2).$$

Note that when $t = 0$, $q_0 = p_{\boldsymbol{\theta}_0}$.

(a) First direction

(b) Second direction

Figure 1: Local Misspecification

The local perturbation in Definition 1 is standard in classical asymptotic statistics [van der Vaart, 2000] and consists of two crucial elements, the scalar function $u(\boldsymbol{z})$ and the real value t. Intuitively, we can think of t as the degree of perturbation and the function $u(\boldsymbol{z})$ as a certain direction of perturbation. Figure 1 presents such a geometric interpretation. The parametric family $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ could be viewed as a "curve" in the distribution space. Each point on the curve corresponds to a distribution in the parametric family. If we fix the value t and let the direction $u(\boldsymbol{z})$ range over all possible directions, then Q will lie within a neighborhood of radius t around this curve. In this sense, the perturbation quantity $tu(\boldsymbol{z})$ acts like the "vector" pointing from $P_{\boldsymbol{\theta}_0}$ to Q where the "length" is t and the "direction" is given by the vector $u(\boldsymbol{z})$. Below we will discuss the "local" case where the radius t vanishes as the sample size goes to infinity. In particular, in Figure 1 (b), the "direction" $u(\boldsymbol{z})$ is tangent to the curve. We will discuss this special case in the latter part of this paper (Theorem 4).

Local misspecification refers to the situation where, as the sample size n increases, the sequence Q_t, where t depends on n, approaches the model. When the ground-truth Q_t lies outside the parametric family, this misspecification adds errors to the standard inference errors on the order $\Theta(1/\sqrt{n})$, and when these two error levels coincide at the same order, we call this scenario *balanced* misspecification. More broadly, we consider local misspecification with $t = \Theta(1/n^\alpha)$ where $\alpha \in (0, \infty)$, which leads to the following definition. For n i.i.d. data $\{\boldsymbol{z}_i\}_{i=1}^n$ sampled from Q, let $Q^n := Q^{\otimes n}$ be the n-fold product measure of Q denoting their joint distribution, and analogously for $P_{\boldsymbol{\theta}_0}$.

Definition 2 (Three Local Misspecification Regimes)**.** Let $P^n = P_{\boldsymbol{\theta}_0}^{\otimes n}$. The tilted distribution Q_t is defined in Definition 1. Suppose $\alpha > 0$ and the joint distribution of the n i.i.d. data is $Q^n := Q_{1/n^\alpha}^{\otimes n}$.

1. (Mild). We call the case when $\alpha > 1/2$ the mildly misspecified regime.
2. (Balanced). We call the case when $\alpha = 1/2$ the balanced misspecified regime.
3. (Severe). We call the case when $0 < \alpha < 1/2$ the severely misspecified regime.

We note that the local misspecification regime described above should not be taken literally to imply that the data-generating process depends on the sample size. Instead, it is an information theoretic device to analyze and compare the local behavior of estimators in situations where the influence of misspecification is comparable to the order of the statistical error. In other words, we are interested in the more realistic setting where both misspecification and statistical error are small at a similar level, instead of assuming vanishing statistical error but fixed misspecification as in previous work [Elmachtoub et al., 2023, 2025].

3 Main Results

We derive theoretical results to compare the asymptotic performances of the three methods, SAA, ETO, and IEO, that encompass the three local misspecification regimes in Definition 2. We first list out several standard assumptions. We define $\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z}) = \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}_0}(\boldsymbol{z})$ as the score function at $\boldsymbol{\theta}_0$ mapping from $\mathcal{Z} \to \mathbb{R}^{d_\theta}$. Recall that $v(\boldsymbol{w}, \boldsymbol{\theta}) = \int c(\boldsymbol{w}, \boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z}) d\boldsymbol{z}$.

Assumption 1 (Smoothness)**.** *Assume that*

1. *The function $v(\boldsymbol{w}, \boldsymbol{\theta})$ is twice continuously differentiable with respect to $(\boldsymbol{w}, \boldsymbol{\theta})$ at $(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{\theta}_0)$ with a Hessian matrix, denoted by $\begin{bmatrix} \boldsymbol{V} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma}^\top & * \end{bmatrix}$, where $*$ denotes a matrix that is not of interest. Assume $\boldsymbol{\Sigma} \in \mathbb{R}^{d_w \times d_\theta}$ is full-rank and $\boldsymbol{V} \in \mathbb{R}^{d_w \times d_w}$ is invertible.*
2. *The function $\boldsymbol{\theta} \mapsto \boldsymbol{w}_{\boldsymbol{\theta}}$ is well-defined on a neighborhood of $\boldsymbol{\theta}_0$, twice continuously differentiable at $\boldsymbol{\theta}_0$ with a full-rank gradient matrix $\nabla_{\boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \in \mathbb{R}^{d_\theta \times d_w}$.*
3. *The Fisher information matrix $\boldsymbol{I} := \mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z}) \boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top] \in \mathbb{R}^{d_\theta \times d_\theta}$ is well-defined and invertible.*

Note that the matrices above are fixed quantities and are not related to whether the model is well-specified or misspecified. These matrices are critical for characterizing the sensitivity of the target stochastic optimization problem. We also define $\boldsymbol{\Phi} = \nabla^2_{\boldsymbol{\theta}_1 \boldsymbol{\theta}_1} \int c(\boldsymbol{w}_{\boldsymbol{\theta}_1}, \boldsymbol{z}) p_{\boldsymbol{\theta}_0}(\boldsymbol{z}) d\boldsymbol{z}|_{\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0}$. The following lemma provides closed-form expressions for the gradient $\nabla_{\boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{\theta}}$ and matrices $\boldsymbol{\Sigma}$, $\boldsymbol{\Phi}$.

Lemma 1. *Under Assumption 1, it holds that*

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = -\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}, \quad \boldsymbol{\Sigma} = \mathbb{E}_{\boldsymbol{\theta}_0}\left[\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z}) \boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top\right], \quad \boldsymbol{\Phi} = \boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1} \boldsymbol{\Sigma}.$$

Next, we introduce the influence function which is a key ingredient in our derived formulas. Originating from robust statistics [Hampel, 1974], it is the functional derivative of an estimator with respect to the data distribution. In our context, this refers to the derivative of the *decision* obtained from the estimation-optimization pipeline. Specifically, the influence functions for SAA, IEO and ETO are respectively

$$\begin{aligned} \mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z}) &= -\boldsymbol{V}^{-1} \nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z}), \\ \mathrm{IF}^{\mathrm{IEO}}(\boldsymbol{z}) &= \boldsymbol{V}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Phi}^{-1} \nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z}), \\ \mathrm{IF}^{\mathrm{ETO}}(\boldsymbol{z}) &= -\boldsymbol{V}^{-1} \boldsymbol{\Sigma} \boldsymbol{I}^{-1} \boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z}), \end{aligned}$$

all of which are $\mathbb{R}^{d_w}$-valued. Regarding notations, $\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})$ is the gradient of the map $\boldsymbol{w} \mapsto c(\boldsymbol{w}, \boldsymbol{z})$ at $\boldsymbol{w} = \boldsymbol{w}_{\boldsymbol{\theta}_0}$, and $\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})$ is the gradient of the map $\boldsymbol{\theta} \mapsto c(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Finally, we introduce *regret* as the criterion to evaluate the quality of a decision $\boldsymbol{w}$. Since we conduct local asymptotic analysis, the definition of regret is slightly different from the classical definition in asymptotic or finite-sample analysis [Lam, 2021, Elmachtoub and Grigas, 2022], as it needs to account for the changing sample size. We define v_n and $\boldsymbol{w}_n^*$ as follows:

$$\boldsymbol{w}_n^* := \operatorname*{argmin}_{\boldsymbol{w} \in \Omega} v_n(\boldsymbol{w}) := \mathbb{E}_{Q^n}\left[\frac{1}{n} \sum_{i=1}^n c(\boldsymbol{w}, \boldsymbol{z}_i)\right].$$

In the local misspecification setting, when the sample size is n, the data distribution is given by Q^n. Hence, $v_n(\boldsymbol{w})$ represents the ground-truth expected cost, and $\boldsymbol{w}_n^*$ is interpreted as the corresponding optimal solution at sample size n.

Definition 3 (Regret)**.** For any distribution Q^n and any $\boldsymbol{w} \in \Omega$, the regret of $\boldsymbol{w}$ at sample size n is defined as

$$R_{Q^n}(\boldsymbol{w}) = v_n(\boldsymbol{w}) - v_n(\boldsymbol{w}_n^*).$$

In the rest of this section, we conduct a comprehensive analysis on the regrets using the three estimation-optimization methods, for the three misspecification regimes introduced in Definition 2.

3.1 Balanced Misspecification

To state our main results, we define, for $\square \in \{\text{SAA}, \text{IEO}, \text{ETO}\}$,

$$\begin{aligned} N^{\square} &:= N(0, \text{var}(\text{IF}^{\square}(\boldsymbol{z}))), \\ \boldsymbol{b}^{\square} &:= \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})(\text{IF}^{\square}(\boldsymbol{z}) - \text{IF}^{\text{SAA}}(\boldsymbol{z}))] \in \mathbb{R}^{d_w}, \\ R^{\square} &:= \tfrac{1}{2} \left\| \boldsymbol{b}^{\square} \right\|_{\boldsymbol{V}}^2 \in \mathbb{R}, \end{aligned}$$

where $N^{\square}$ is the normal distribution with zero mean and covariance matrix $\text{var}(\text{IF}^{\square}(\boldsymbol{z}))$. Note that unless otherwise specified, $\text{var}(\cdot)$ should always be interpreted as $\text{var}_{P_{\boldsymbol{\theta}_0}}(\cdot)$.

Theorem 1 (Asymptotics under Balanced Misspecification)**.** *Suppose Assumptions 1, 3, and 4 hold. In the balanced regime in Definition 2, under Q^n, for $\square \in \{\text{SAA}, \text{ETO}, \text{IEO}\}$,*

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_n^*) &\overset{Q^n}{\to} N^{\square} + \boldsymbol{b}^{\square}, \\ n R_{Q^n}(\hat{\boldsymbol{w}}^{\square}) &\overset{Q^n}{\to} \frac{1}{2} \left(N^{\square} + \boldsymbol{b}^{\square} \right)^{\top} \boldsymbol{V} \left(N^{\square} + \boldsymbol{b}^{\square} \right). \end{aligned}$$

In terms of bias, $0 = \left\| \boldsymbol{b}^{\text{SAA}} \right\|_{\boldsymbol{V}} \leq \left\| \boldsymbol{b}^{\text{IEO}} \right\|_{\boldsymbol{V}} \leq \left\| \boldsymbol{b}^{\text{ETO}} \right\|_{\boldsymbol{V}}$*. In terms of variance,* $\text{var}\left(N^{\text{SAA}}\right) \geq \text{var}\left(N^{\text{IEO}}\right) \geq \text{var}\left(N^{\text{ETO}}\right)$.

Theorem 1 states that when $\alpha = 1/2$, i.e., the degree of misspecification is of the same order as the statistical error, the gap between the data-driven and optimal decisions is asymptotically normal. Moreover, this normal has mean zero for SAA (note $\boldsymbol{b}^{\text{SAA}} = \boldsymbol{0}$ and $R^{\text{SAA}} = 0$), but generally nonzero for ETO and IEO. More importantly, in the asymptotic limit, $\boldsymbol{b}^{\square}$ represents the bias coming from model misspecification as it involves u, while $N^{\square}$ captures the data noise variability. We highlight that the dependence on u in $\boldsymbol{b}$ directly implies that the ETO and IEO estimator are *non-regular* in the sense of van der Vaart [2000]. The theorem shows that in terms of bias, SAA generally outperforms IEO which in turn outperforms ETO. On the the hand, the ordering is reversed for variance. As a result there is no universal ordering for the overall error in general. The next theorem lifts further to compare the regrets of SAA, ETO and IEO under the balanced regime.

Theorem 2 (Regret Comparisons under Balanced Misspecification)**.** *Let*

$$\mathbb{G}^{\square} := \frac{1}{2}(N^{\square} + \boldsymbol{b}^{\square})^{\top} \boldsymbol{V} \left(N^{\square} + \boldsymbol{b}^{\square} \right)$$

denote the limiting regret distribution of $\square \in \{\text{SAA}, \text{ETO}, \text{IEO}\}$ in Theorem 1. Then $\mathbb{E}[\mathbb{G}^{\square}] = \mathbb{E}[\frac{1}{2}(N^{\square})^{\top} \boldsymbol{V} N^{\square}] + \frac{1}{2}(\boldsymbol{b}^{\square})^{\top} \boldsymbol{V} \boldsymbol{b}^{\square}$*. Moreover,*

$$\begin{aligned} \mathbb{E}\left[\frac{1}{2} \left(N^{\text{ETO}} \right)^{\top} \boldsymbol{V} N^{\text{ETO}} \right] \leq \mathbb{E}\left[\frac{1}{2} \left(N^{\text{IEO}} \right)^{\top} \boldsymbol{V} N^{\text{IEO}} \right] \leq \mathbb{E}\left[\frac{1}{2} \left(N^{\text{SAA}} \right)^{\top} \boldsymbol{V} N^{\text{SAA}} \right], \\ \frac{1}{2} \left(\boldsymbol{b}^{\text{SAA}} \right)^{\top} \boldsymbol{V} \boldsymbol{b}^{\text{SAA}} \leq \frac{1}{2} \left(\boldsymbol{b}^{\text{IEO}} \right)^{\top} \boldsymbol{V} \boldsymbol{b}^{\text{IEO}} \leq \frac{1}{2} \left(\boldsymbol{b}^{\text{ETO}} \right)^{\top} \boldsymbol{V} \boldsymbol{b}^{\text{ETO}}. \end{aligned}$$

Like Theorem 1, while Theorem 2 shows a lack of universal ordering for regrets, it depicts a decomposition of the asymptotic distribution of the regret into two parts where two opposite orderings emerge. In particular, it suggests that while ETO is best in terms of variance, and SAA best in terms of bias, IEO is in between and could potentially induce a lower decision error compared to the other two methods.

Another important insight from Theorems 1 and 2 regards the explicit form of the bias and variance. For this, let us introduce the analogous results for severe and mild misspecification regimes and discuss the formulas along the way.

3.2 Severe Misspecification

We first formally describe the $O(1/\sqrt{n})$ order of the statistical error via the following assumptions borrowed from Fang et al. [2023]. The assumption is natural because it says the empirical part deviates from the expected part at the rate $O(1/\sqrt{n})$. .

Assumption 2 (Statistical Error Order)**.** *For i.i.d. $\{\boldsymbol{z}_i\}_{i=1}^n$ with joint distribution Q^n, let*

$$\boldsymbol{\theta}_n^{\mathrm{KL}} := \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{Q^n}\left[\frac{1}{n}\sum_{i=1}^n \log p_{\boldsymbol{\theta}}(\boldsymbol{z}_i)\right],$$

$$\boldsymbol{\theta}_n^{*} := \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{Q^n}\left[\frac{1}{n}\sum_{i=1}^n c(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z}_i)\right].$$

Assume that $\left\|\hat{\boldsymbol{w}}^{\mathrm{ETO}} - \boldsymbol{w}_{\boldsymbol{\theta}_n^{\mathrm{KL}}}\right\|$, $\left\|\hat{\boldsymbol{w}}^{\mathrm{IEO}} - \boldsymbol{w}_{\boldsymbol{\theta}_n^{*}}\right\|$ and $\left\|\hat{\boldsymbol{w}}^{\mathrm{SAA}} - \boldsymbol{w}_n^{*}\right\|$ are all of order $O_{Q^n}(1/\sqrt{n})$. Moreover, assume that the matrix $\nabla^2_{\boldsymbol{w}\boldsymbol{w}} v_n(\boldsymbol{w}_n^*) \to \boldsymbol{V}$ as $n \to \infty$.

Theorem 3 (Asymptotics under Severe Misspecification)**.** *Suppose Assumptions 1, 2, 3 and 4 hold. In the severely misspecified case, under Q^n, for $\square \in \{\mathrm{SAA}, \mathrm{ETO}, \mathrm{IEO}\}$,*

$$n^{\alpha}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_n^*) \xrightarrow{p} \boldsymbol{b}^{\square},$$

$$n^{2\alpha} R_{Q^n}(\hat{\boldsymbol{w}}^{\square}) \xrightarrow{p} R^{\square} = \frac{1}{2}\|\boldsymbol{b}^{\square}\|^2_{\boldsymbol{V}}.$$

In terms of variance, $0 = \operatorname{var}\left(\boldsymbol{b}^{\mathrm{SAA}}\right) = \operatorname{var}\left(\boldsymbol{b}^{\mathrm{IEO}}\right) = \operatorname{var}\left(\boldsymbol{b}^{\mathrm{ETO}}\right)$. The comparison of bias has the same form as the regret (stated in the next theorem).

Theorem 4 (Bias/Regret Comparisons under Severe Misspecification)**.** *Under the same setting as in Theorem 3, we have $0 = \left\|\boldsymbol{b}^{\mathrm{SAA}}\right\|_{\boldsymbol{V}} \le \left\|\boldsymbol{b}^{\mathrm{IEO}}\right\|_{\boldsymbol{V}} \le \left\|\boldsymbol{b}^{\mathrm{ETO}}\right\|_{\boldsymbol{V}}$ and $0 = R^{\mathrm{SAA}} \le R^{\mathrm{IEO}} \le R^{\mathrm{ETO}}$.*

Theorems 3 and 4 stipulate that, once the degree of misspecification is larger than the statistical error ($0 < \alpha < 1/2$), SAA will dominate ETO which will further dominate IEO. This is because in this regime only the bias surfaces, and this ordering is in line with the bias ordering in Theorems 1 and 2.

In both the balanced and the severe misspecification cases, the bias term can be significant relative to the variance. In all cases, the bias has the form $\boldsymbol{b}^{\square} = \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})(\mathrm{IF}^{\square}(\boldsymbol{z}) - \mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z}))]$, an inner product between the misspecification direction and the difference of influence functions between the considered estimation-optimization pipeline and SAA. Note that the latter difference is always zero for SAA, which coincides with the model-free nature of SAA that elicits zero bias. On the other hand, for either IEO or ETO, the bias effect can be minimized if the misspecification direction is orthogonal to the influence function difference. While this characterization is generally opaque, the following provides a more manageable sufficient condition.

Theorem 5 (Approximately Impactless Misspecification Direction)**.** *Let the assumptions in Theorem 3 hold. If $u(\cdot) \in \left\{\boldsymbol{\beta}^\top \boldsymbol{s}_{\boldsymbol{\theta}_0}(\cdot) : \boldsymbol{\beta} \in \mathbb{R}^{d_{\boldsymbol{\theta}}}\right\}$, then $\boldsymbol{b}^{\mathrm{ETO}}$ (and thus $\boldsymbol{b}^{\mathrm{IEO}}$ and $\boldsymbol{b}^{\mathrm{SAA}}$) $= \boldsymbol{0}$.*

Theorem 5 states that if the misspecification direction is in the linear span of the score function of the imposed model at $\boldsymbol{\theta}_0$, the asymptotic bias is zero for all methods. Figure 1 (b) illustrates such a direction, where $u(\boldsymbol{z})$ is tangential to the model $\{P_\theta\}$ at $P_{\boldsymbol{\theta}_0}$. In this case, the interesting direction of misspecification $u(z)$ aligns with the parametric information and couples the influence function of ETO and SAA. As a result, ETO can induce zero decision bias by merely conducting MLE to infer $\boldsymbol{\theta}$, even though the model is misspecified.

Note that the condition in Theorem 5 depends only on the parametric model, but not the downstream optimization problem. Nonetheless, it already allows us to understand and provide examples where a model misspecification can be significant, namely shooting outside the parametric model, yet the impact on the bias of the resulting decision is negligible. We provide an explicit example as follows.

Example 1. Consider the distribution family $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}\}$ to be normal distributions with variance 1, $N(\boldsymbol{\theta}, 1)$, where $\boldsymbol{\theta}_0 = 0$. We define the tilted distribution $Q_t(\boldsymbol{z})$ with density $q_t(\boldsymbol{z}) \propto (1+t\boldsymbol{z})_+ e^{-\boldsymbol{z}^2/2}$. In this case, $Q_t(\boldsymbol{z})$ satisfies Definition 1 and the conditions in Theorem 5, but $Q_t \notin \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}\}$.

In Example 1, the parametric family is a normal location family, while at $\boldsymbol{\theta}_0 = 0$, the perturbed distribution family Q_t is never normally distributed for all $t > 0$. The direction of perturbation

(misspecification) at $\boldsymbol{\theta}_0$ here is $u(\boldsymbol{z}) = \boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z}) = \boldsymbol{z}$. In other words, even if the ground truth distribution of uncertain parameters is complicated, model-based approaches under a simplified and misspecified parametric family can still be employed with a satisfying decision regret performance.

3.3 Mild Misspecification

Finally, we establish results under mild misspecification.

Theorem 6 (Asymptotics and Comparisons under Mild Misspecification)**.** *Suppose Assumptions 1, 3 and 4 hold. In the mildly misspecified case, under Q^n, for $\square \in \{\text{SAA}, \text{ETO}, \text{IEO}\}$,*

$$\sqrt{n}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_n^*) \overset{Q^n}{\to} N^{\square} ,$$
$$nR_{Q^n}(\hat{\boldsymbol{w}}^{\square}) \overset{Q^n}{\to} \frac{1}{2}(N^{\square})^{\top} \boldsymbol{V} N^{\square}.$$

Moreover, in terms of bias, all three estimators have asymptotically zero biases. In terms of variance, $\text{var}\left(N^{\text{SAA}}\right) \geq \text{var}\left(N^{\text{IEO}}\right) \geq \text{var}\left(N^{\text{ETO}}\right)$*. In terms of regret, it holds that*

$$\frac{1}{2}(N^{\text{ETO}})^{\top} \boldsymbol{V} N^{\text{ETO}} \preceq_{\text{st}} \frac{1}{2}(N^{\text{IEO}})^{\top} \boldsymbol{V} N^{\text{IEO}} \preceq_{\text{st}} \frac{1}{2}(N^{\text{SAA}})^{\top} \boldsymbol{V} N^{\text{SAA}} .$$

Theorem 6 shows that the obtained solutions (consequently also the regret) exhibit asymptotic behaviors in accordance with the universal ordering of ETO best, then IEO, and then SAA. In this regime, the bias is negligible, while the variance is the dominant term, and the regret distribution is related to the variance. The phenomenon also holds in the well-specified regime stated as follows.

Proposition 1 (Asymptotics under Well-Specification[Elmachtoub et al., 2023])**.** *In the well-specified case where $Q = P_{\boldsymbol{\theta}_0}$, under Assumptions 1, 3 and 4, for $\square \in \{\text{SAA}, \text{ETO}, \text{IEO}\}$, we have*

$$\sqrt{n}\left(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}^*\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \text{IF}^{\square}(\boldsymbol{z}_i) + o_{P^n}(1).$$

Moreover, $\text{var}\left(\text{IF}^{\text{SAA}}(\boldsymbol{z})\right) \geq \text{var}\left(\text{IF}^{\text{IEO}}(\boldsymbol{z})\right) \geq \text{var}\left(\text{IF}^{\text{ETO}}(\boldsymbol{z})\right)$*. If $d_{\boldsymbol{\theta}} = d_{\boldsymbol{w}}$ and $\boldsymbol{\Sigma}$ is a square and full-rank matrix, then* $\text{var}\left(\text{IF}^{\text{SAA}}(\boldsymbol{z})\right) = \text{var}\left(\text{IF}^{\text{IEO}}(\boldsymbol{z})\right)$.

4 Numerical Experiments

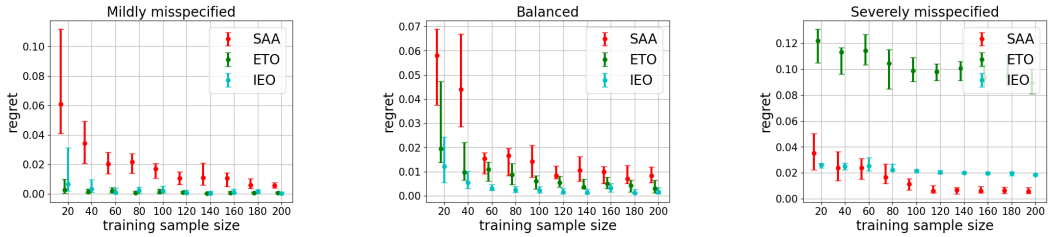
In this section, we validate our findings by conducting numerical experiments on the newsvendor problem, a classic example in operations research with non-linear cost objectives. We show and compare the performances of the three data-driven methods in the finite-sample regimes under different local misspecification settings, including different directions and degrees of misspecification. The experimental results in the finite-sample regime are consistent with our asymptotic comparisons. All computations were carried out on a personal desktop computer without GPU acceleration.

The newsvendor problem has the objective function $c(\boldsymbol{w}, \boldsymbol{z}) = \boldsymbol{a}^{\top} (\boldsymbol{w} - \boldsymbol{z})^{+} + \boldsymbol{d}^{\top} (\boldsymbol{z} - \boldsymbol{w})^{+}$, where for each $j \in [d_z]$: (1) $z^{(j)}$ is the customers' random demand of product j; (2) $w^{(j)}$ is the decision variable, the ordering quantity for product j; (3) $a^{(j)}$ is the holding cost for product j; (4) $d^{(j)}$ is the backlogging cost for product j. We assume the random demand for each product are independent and the holding cost and backlogging cost is uniform among all products by setting $a^{(j)} = 5$ and $d^{(j)} = 1$ for all $j \in [d_z]$.

We describe the local misspecified setting by using the framework of Example 5 and building a model and generating a random demand dataset as follows. We denote the training dataset as $\left\{z_i^{(j)}\right\}_{i=1}^{n}$, where n is the training sample size. The model assumes that the demand for each product $j \in [d_z]$ is normally distributed with the distribution $N(j\theta, 1)$ where θ is unknown and needs to be learned. We first describe the well-specified setting, where the demand distribution for product j is $N(3j, 1)$. In this case, the probability density function of the random demand of each product is $p_j(z^{(j)}) \propto \exp(-(z^{(j)} - 3j)^2/2)$. To describe the local misspecification, we need to specify the direction and degree of misspecification, i.e., the expression of $u(\boldsymbol{z})$ and α in Section 2.3. We set (1) $\alpha = 0.1$ to denote the severely misspecified setting, (2) $\alpha = 0.5$ to denote the balanced

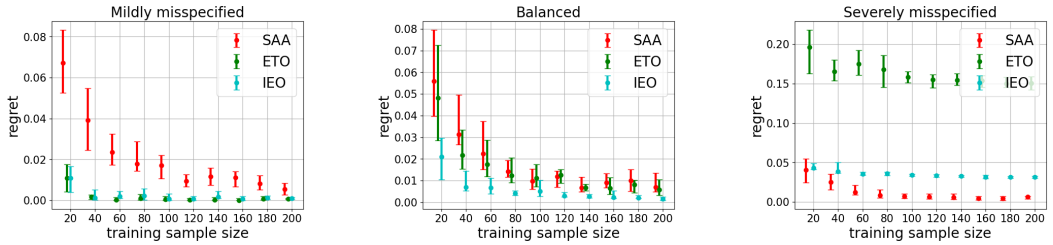
setting and $\alpha = 2$ to denote the mildly misspecified setting. We discuss two types of directions: (1) $u(\boldsymbol{z}) = \prod_{j=1}^{d_z} \left(z^{(j)}\right)^2$; (2) $u(\boldsymbol{z}) = \prod_{j=1}^{d_z} \left(z^{(j)} - 3j\right)^2 /2$.

We show experimental results in Figure 2 - Figure 3 to support our theoretical results in Section 3, using the mean, median, 25-th quantile, 75-th quantile and histograms of the regret. When $u(\boldsymbol{z}) = \prod_{j=1}^{d_z} \left(z^{(j)}\right)^2$ and $u(\boldsymbol{z}) = \prod_{j=1}^{d_z} \left(z^{(j)} - 3j\right)^2 /2$, in the mildly specified case, ETO has a lower regret than IEO, and IEO has a lower regret than SAA. However, in the severely misspecified regime, the ordering of the three methods flips. This is consistent with our theoretical comparison results in Theorems 6 and 4. In the balanced regime, experimental results show that IEO has the lowest regret among the three methods. This is also consistent with the theoretical insight in Theorem 2 that IEO has the advantage of achieving bias-variance trade-off in terms of the decisions and regrets.



(a) Mildly Misspecified (b) Balanced (c) Severely Misspecified

Figure 2: The direction of misspecification satisfies $u(\boldsymbol{z}) = \prod_{j=1}^{d_z} \left(z^{(j)}\right)^2$.



(a) Mildly Misspecified (b) Balanced (c) Severely Misspecified

Figure 3: The direction of misspecification satisfies $u(\boldsymbol{z}) = \prod_{j=1}^{d_z} \left(z^{(j)} - 3j\right)^2 /2$.

5 Conclusions and Discussions

In this paper, we present the first results on analyzing local model misspecification in data-driven optimization. Our framework captures scenarios where the model-based approach is nearly well-specified, moving beyond the existing dichotomy of well-specified and misspecified models. We conduct a detailed analysis of the relative performances of SAA, ETO, and IEO, providing insights into their bias, variance, and regret. By classifying local misspecification into three regimes, our analysis illustrates how varying degrees of misspecification impacts performance. In particular, we show that in the balanced misspecification case, ETO exhibits the best variance, SAA exhibits the best bias, while IEO entails a bias-variance tradeoff that can potentially result in lower overall decision errors than both ETO and SAA. Additionally, we derive closed-form expressions for decision bias and variance. From this, we show how the orthogonality between the misspecification direction and the difference of influence functions can lead to bias cancellation, and provide more transparent sufficient condition for such phenomenon in relation to the tangentiality on the score function. Technically, we leverage and generalize tools from contiguity theory in statistics to establish the performance orderings and the clean, interpretable bias and variance expressions. Future research directions include extending our framework to contextual or constrained optimization problems, where challenges like feasibility guarantees and model complexity become increasingly significant.

Acknowledgments and Disclosure of Funding

We gratefully acknowledge support from the National Science Foundation grant IIS-2238960, the Office of Naval Research awards N00014-22-1-2530 and N00014-23-1-2374, 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

I. Andrews, M. Gentzkow, and J. M. Shapiro. Measuring the sensitivity of parameter estimates to estimation moments. *The Quarterly Journal of Economics*, 132(4):1553–1592, 2017.

I. Andrews, M. Gentzkow, and J. M. Shapiro. On the informativeness of descriptive statistics for structural estimates. *Econometrica*, 88(6):2231–2258, 2020.

T. B. Armstrong and M. Kolesár. Sensitivity analysis using approximate moment condition models. *Quantitative Economics*, 12(1):77–108, 2021.

T. B. Armstrong, P. Kline, and L. Sun. Adapting to misspecification. *arXiv preprint arXiv:2305.14265*, 2023.

S. Asmussen and P. W. Glynn. *Stochastic simulation: algorithms and analysis*, volume 57. Springer, 2007.

S. Barratt. On the differentiability of the solution to convex optimization problems. *arXiv preprint arXiv:1804.05098*, 2018.

P. J. Bickel and K. A. Doksum. *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. Chapman and Hall/CRC, 2015.

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

S. Bonhomme and M. Weidner. Minimizing sensitivity to model misspecification. *Quantitative Economics*, 13(3):945–981, 2022. doi: 10.3982/QE1930.

F. A. Bugni and T. Ura. Inference in dynamic discrete choice problems under local misspecification. *Quantitative Economics*, 10(1):67–103, 2019.

F. A. Bugni, I. A. Canay, and P. Guggenberger. Distortions of asymptotic confidence size in locally misspecified moment inequality models. *Econometrica*, 80(4):1741–1768, 2012. doi: 10.3982/ECTA9604.

L. E. Candelaria and Y. Zhang. Robust inference in locally misspecified bipartite networks. *arXiv preprint arXiv:2403.13725*, 2024.

T. G. Conley, C. B. Hansen, and P. E. Rossi. Plausibly exogenous. *Review of Economics and Statistics*, 94(1):260–272, 2012.

J. Copas and S. Eguchi. Local sensitivity approximations for selectivity bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 63(4):871–895, 2001.

J. Copas and S. Eguchi. Local model uncertainty and incomplete-data bias (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(4):459–513, 2005.

E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

K. Dong, Y. Flet-Berliac, A. Nie, and E. Brunskill. Model-based offline reinforcement learning with local misspecification. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

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.

J. Duchi. A few notes on contiguity, asymptotics, and local asymptotic normality, 2021. Accessed: 2024-12-16.

J. C. Duchi and F. Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1):21–48, 2021.

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, H. Lam, H. Zhang, and Y. Zhao. Estimate-then-optimize versus integrated-estimation-optimization versus sample average approximation: a stochastic dominance perspective. *arXiv preprint arXiv:2304.06833*, 2023.

A. N. Elmachtoub, H. Lam, H. Lan, and H. Zhang. Dissecting the impact of model misspecification in data-driven optimization. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2025.

J. Fan, K. Imai, I. Lee, H. Liu, Y. Ning, and X. Yang. Optimal covariate balancing conditions in propensity score estimation. *Journal of Business & Economic Statistics*, 41(1):97–110, 2022.

Z. Fang, A. Santos, A. M. Shaikh, and A. Torgovitsky. Inference for large-scale linear systems with known coefficients. *Econometrica*, 91(1):299–327, 2023.

P. Glasserman. *Monte Carlo methods in financial engineering*, volume 53. Springer, 2004.

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

F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

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.

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

J. D. Y. Kang and J. L. Schafer. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4): 523–539, 2007. doi: 10.1214/07-STS227.

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

Y. Kitamura, T. Otsu, and K. Evdokimov. Robustness, infinitesimal neighborhoods, and moment restrictions. *Econometrica*, 81(3):1185–1201, 2013.

H. Lam. On the impossibility of statistically improving empirical optimization: A second-order stochastic dominance perspective. *arXiv preprint arXiv:2105.13419*, 2021.

L. M. Le Cam and G. L. Yang. *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media, 2000.

P. L'Ecuyer. A unified view of the ipa, sf, and lr gradient estimation techniques. *Management Science*, 36(11):1364–1383, 1990.

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.

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. Marsden, J. Duchi, and G. Valiant. Misspecification in prediction problems and robustness via improper learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2161–2169. PMLR, 2021.

W. K. Newey. Generalized method of moments specification testing. *Journal of econometrics*, 29(3): 229–256, 1985.

A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

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

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New York, NY, 1996. ISBN 978-0-387-94640-2. doi: 10.1007/978-1-4757-2545-2.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope on the local misspecification analysis of three date-driven stochasitic optimization methods.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The contextual and constrained case can be future research directions.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: For each theoretical result, the paper provides the full set of assumptions and a complete (and correct) proof (some in the appendix).

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper has introduced the experiment setup and problem parameters.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides open access to the data and code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper has provided experimental setting and details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper has provided statistical significance of the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper has provided information on the computer resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper is theoretical and neutrally demonstrates the math theorems.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: the core method development in this research does not involve LLMs as any important, original, or non-standard components

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Additional Examples and Techniques Details

A.1 Examples of Data-Driven Optimization in Operations Research wtih Nonlinear Cost Objectives

We now give two canonical examples of stochastic optimization problems in operations research with non-linear cost objectives.

Example 2 (Multi-Product Newsvendor Problem)**.** The newsvendor problem has the objective function $c(\boldsymbol{w}, \boldsymbol{z}) = \boldsymbol{a}^\top (\boldsymbol{w} - \boldsymbol{z})^+ + \boldsymbol{d}^\top (\boldsymbol{z} - \boldsymbol{w})^+$, where for each $j \in [d_z]$: (1) $z^{(j)}$ is the customers' random demand of product j, ; (2) $w^{(j)}$ is the decision variable, the ordering quantity for product j; (3) $a^{(j)}$ is the holding cost for product j; (4) $d^{(j)}$ is the backlogging cost for product j and (5) the goal is to minimize the expected total cost.

Consider another classical problem in operations research, the portfolio optimization problem [Kallus and Mao, 2022, Grigas et al., 2021, Elmachtoub et al., 2023].

Example 3 (Portfolio Optimization)**.** Let $d_w = d_z + 1$ and denote the cost function as $c(\boldsymbol{w}, \boldsymbol{z}) = \gamma \left(\boldsymbol{w}^\top (\boldsymbol{z}, -1)\right)^2 + \exp\left(-\boldsymbol{w}^\top (\boldsymbol{z}, 0)\right)$. The decision $\boldsymbol{w}$ satisfies $\left(w^{(1)}, w^{(2)}, ..., w^{(d_w - 1)}\right) \in \mathbb{R}^{d_w - 1}$, denoting the investment fraction on products $1, 2, ...d_z$(i.e., $d_w - 1$) and $w^{(d_w)}$ is an auxiliary decision variable. The first component represents the risk (variance) of the portfolio and the second component represents the exponential utility of the portfolio.

A.2 Further Examples of Local Misspecification

Example 4 (Parametric Perturbation: Quadratic Mean Differentiability (QMD) family)**.** Suppose $P^n = P_{\boldsymbol{\theta}_0}^{\otimes n}$ for some fixed $\boldsymbol{\theta}_0$. Consider a sequence of vectors in $\mathbb{R}^{d_\theta}$, say $\{\boldsymbol{h}_n\}_{n=1}^\infty$. Suppose the joint distribution of $\{\boldsymbol{z}_i\}_{i=1}^n$, Q^n, is also in the parametric family, but is of the form $P_{\boldsymbol{\theta}_0 + \boldsymbol{h}_n}^{\otimes n}$. If there exists a score function $\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(\boldsymbol{z}) : \mathcal{Z} \to \mathbb{R}^{d_\theta}$ with $\mathbb{E}_{\boldsymbol{\theta}_0}[\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_0}(\boldsymbol{z})] = \mathbf{0}$ such that

$$\int \left(\sqrt{p_{\boldsymbol{\theta}_0 + \boldsymbol{h}_n}} - \sqrt{p_{\boldsymbol{\theta}_0}} - \frac{1}{2} \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_0}^\top \boldsymbol{h}_n \sqrt{p_{\boldsymbol{\theta}_0}} \right)^2 d\boldsymbol{z} = o(\|\boldsymbol{h}_n\|^2), \boldsymbol{h}_n \to \mathbf{0}.$$

In particular, in our framework, we focus on the case where $\boldsymbol{h}_n = \boldsymbol{h}/n^\alpha$ for a fixed vector $\boldsymbol{h}$. When $\alpha = 1/2$, van der Vaart [2000] shows that the likelihood ratio between Q^n and P^n satisfies:

$$\log \frac{dQ^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)}{dP^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \boldsymbol{h}^\top \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_0}(\boldsymbol{z}_i) - \frac{1}{2} \boldsymbol{h}^\top \boldsymbol{I} \boldsymbol{h} + o_{P^n}(\|\boldsymbol{h}\|^2),$$

where $\boldsymbol{I} := \mathbb{E}_{\boldsymbol{\theta}_0}[\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_0} \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_0}^\top]$ denotes the Fisher information. In other words, under P^n

$$\log \frac{dQ^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)}{dP^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)} \overset{P^n}{\to} N(-\frac{1}{2} \boldsymbol{h}^\top \boldsymbol{I} \boldsymbol{h}, \boldsymbol{h}^\top \boldsymbol{I} \boldsymbol{h}),$$

where the limiting distribution is a Gaussian distribution with mean $-\frac{1}{2} \boldsymbol{h}^\top \boldsymbol{I} \boldsymbol{h}$ and variance $\boldsymbol{h}^\top \boldsymbol{I} \boldsymbol{h}$.

In the previous example, the ground truth distribution Q is still in $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ but is in the local neighbourhood of $P_{\boldsymbol{\theta}_0}$. The more common and interesting examples are when $Q \notin \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ as discussed in examples below.

Example 5 (Semi-parametric Local Perturbation: Part I)**.** Suppose $P_{\boldsymbol{\theta}_0}$ is a given distribution, and $u(\boldsymbol{z}) : \mathcal{Z} \to \mathbb{R}$ is an unobserved random variable with $\mathbb{E}_{\boldsymbol{\theta}_0}[u] = 0$ and a finite variance $\mathbb{E}_{\boldsymbol{\theta}_0}[u^2]$. For a scalar t in a neighborhood of zero, we define the tilted distribution of $P_{\boldsymbol{\theta}_0}$, called Q_t, as

$$dQ_t(\boldsymbol{z}) = \frac{\exp(tu(\boldsymbol{z}))}{C_t} dP_{\boldsymbol{\theta}_0}(\boldsymbol{z})$$

where $C_t = \int \exp(tu(\boldsymbol{z})) dP_{\boldsymbol{\theta}_0}(\boldsymbol{z}) < \infty$ is a normalization constant. Clearly $Q_{t=0} = P_{\boldsymbol{\theta}_0}$.

Lemma 2 (Log Likelihood Ratio Property in Example 5)**.** *Under Definition 2, when $\alpha = 1/2$, i.e., $Q^n = Q_{1/\sqrt{n}}^{\otimes n}$, the log-likelihood ratio between Q^n and P^n satisfies:*

$$\log \frac{dQ^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)}{dP^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)} = \frac{1}{\sqrt{n}} \sum_{i=1}^n u(\boldsymbol{z}_i) - \frac{1}{2} \mathbb{E}_{\boldsymbol{\theta}_0}[u^2] + o_{P^n}(1) \ .$$

This implies, under P^n, $\log \frac{dQ^n(\boldsymbol{z}_1,...,\boldsymbol{z}_n)}{dP^n(\boldsymbol{z}_1,...,\boldsymbol{z}_n)} \xrightarrow{P^n} N(-\frac{1}{2}\mathbb{E}_{\boldsymbol{\theta}_0}[u^2], \mathbb{E}_{\boldsymbol{\theta}_0}[u^2])$.

Example 6 (Semi-parametric Local Perturbation: Part II)**.** Consider the random variable $u(\boldsymbol{z}) : \mathcal{Z} \to \mathbb{R}$ with a zero mean, $\mathbb{E}_{\boldsymbol{\theta}_0}[u] = 0$, and finite second moment, say $\mathbb{E}_{\boldsymbol{\theta}_0}[u^2]$. Now we define the tilted distribution:

$$dQ_t(\boldsymbol{z}) = \frac{[1 + tu(\boldsymbol{z})]_+}{C_t} dP_{\boldsymbol{\theta}_0}(\boldsymbol{z}) \text{ where } C_t = \int [1 + tu(\boldsymbol{z})]_+ \, dP_{\boldsymbol{\theta}_0}(\boldsymbol{z}).$$

In particular, in our framework we focus on the case where $Q = Q_{1/n^\alpha}$ and $Q^n := Q_{1/n^\alpha}^{\otimes n}$. When $\alpha = 1/2$, by Duchi [2021], the log likelihood ratio satisfies

$$\log \frac{dQ^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)}{dP^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)} = \frac{1}{\sqrt{n}} \sum_{i=1}^n u(\boldsymbol{z}_i) - \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta}_0}[u^2] + o_{P^n}(1).$$

In other words, under P^n,

$$\log \frac{dQ^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)}{dP^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)} \xrightarrow{P^n} N(-\frac{1}{2}\mathbb{E}_{\boldsymbol{\theta}_0}[u^2], \mathbb{E}_{\boldsymbol{\theta}_0}[u^2]).$$

Example 7 (Semi-parametric Local Perturbation: Part III)**.** Consider the function $g : \mathbb{R} \to [-1, 1]$ be any three-times continuously differentiable function where $g(x) = x$ for $x \in [-1/2, 1/2]$ and $g' \geq 0$ and the first three derivatives of g are bounded. Consider the random variable $u(\boldsymbol{z}) : \mathcal{Z} \to \mathbb{R}$ with a zero mean $\mathbb{E}_{\boldsymbol{\theta}_0}[u] = \mathbf{0}$ and finite second moment, say $\mathbb{E}_{\boldsymbol{\theta}_0}[u^2]$. Now, for $\boldsymbol{t} \in \mathbb{R}$, we define the tilted distribution

$$dQ_t(\boldsymbol{z}) = \frac{1 + g(tu(\boldsymbol{z}))}{C_t} dP_{\boldsymbol{\theta}_0}(\boldsymbol{z}) \text{ where } C_t = 1 + \int g(tu(\boldsymbol{z})) dP_{\boldsymbol{\theta}_0}(\boldsymbol{z}).$$

In particular, in our framework we focus on the case where $Q = Q_{1/n^\alpha}$ and $Q^n := Q_{1/n^\alpha}^{\otimes n}$. When $\alpha = 1/2$, by Duchi and Ruan [2021], the log likelihood ratio satisfies the following Property:

$$\log \frac{dQ^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)}{dP^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)} = \frac{1}{\sqrt{n}} \sum_{i=1}^n u(\boldsymbol{z}_i) - \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta}_0}[u^2] + o_{P^n}(1).$$

In other words, under P^n,

$$\log \frac{dQ^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)}{dP^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)} \xrightarrow{P^n} N(-\frac{1}{2}\mathbb{E}_{\boldsymbol{\theta}_0}[u^2], \mathbb{E}_{\boldsymbol{\theta}_0}[u^2]).$$

Example 8 (Semi-parametric Local Perturbation: Part IV (QMD Family))**.** Consider a scalar function $u(\boldsymbol{z}) : \mathcal{Z} \to \mathbb{R}$ with zero mean $\mathbb{E}_{\boldsymbol{\theta}_0}[u] = 0$ and finite second order moment $\mathbb{E}_{\boldsymbol{\theta}_0}[u^2]$. We define a tilted distribution Q_t for $t \in \mathbb{R}$ with probability density (mass) function q_t with respect to the dominated measure (note that Q_t is not necessarily in the parametric family $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$) with $q_{\mathbf{0}} = p_{\boldsymbol{\theta}_0}$. We further assume the quadratic mean differentiability

$$\int \left(\sqrt{q_t} - \sqrt{p_{\boldsymbol{\theta}_0}} - \frac{1}{2} tu \sqrt{p_{\boldsymbol{\theta}_0}} \right)^2 d\boldsymbol{z} = o(t^2).$$

Note that when $q_0 = p_{\boldsymbol{\theta}_0}$. In particular, in our framework we focus on the case where $Q = Q_{1/n^\alpha}$ and $Q^n := Q_{1/n^\alpha}^{\otimes n}$. When $\alpha = 1/2$, by Duchi [2021], we have

$$\log \frac{dQ^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)}{dP^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)} = \frac{1}{\sqrt{n}} \sum_{i=1}^n u(\boldsymbol{z}_i) - \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta}_0}[u^2] + o_{P^n}(1).$$

Note that Example 8 is the most general version and includes Example 6-7 as particular examples under some mild assumptions.

A.3 Additional Technical Details

We introduce standard regularity assumptions for general M-estimation problems in asymptotic statistics [van der Vaart, 2000], which include our SAA, ETO, and IEO methods as examples.

Assumption 3 (Regularity Assumptions for M-estimation)**.** *Suppose the i.i.d. random variables $\{\boldsymbol{z}_i\}_{i=1}^n$ follows a distribution Q. Suppose the function $\boldsymbol{z} \to m_\zeta(\boldsymbol{z})$ is measurable with respect to $\boldsymbol{z}$ for all ζ and*

1. $\sup_\zeta \left| \frac{1}{n}\sum_{i=1}^n m_\zeta(\boldsymbol{z}_i) - \mathbb{E}_Q\left[m_\zeta(\boldsymbol{z})\right] \right| \xrightarrow{p} 0$,
2. *there exists $\zeta^* = \operatorname{argmax}_\zeta \mathbb{E}_Q\left[m_\zeta(\boldsymbol{z})\right]$, for all $\varepsilon > 0$, $\sup_{\zeta:\|\zeta-\zeta^*\|\geq\varepsilon} \mathbb{E}_Q\left[m_\zeta(\boldsymbol{z})\right] < \mathbb{E}_Q\left[m_{\zeta^*}(\boldsymbol{z})\right]$,*
3. *the mapping $\zeta \to m_\zeta(\boldsymbol{z})$ is differentiable at ζ^* for Q-almost every $\boldsymbol{z}$ with derivative $\nabla_\zeta m_{\zeta^*}(\boldsymbol{z})$ and such that for every ζ_1 and ζ_2 in a neighbourhood of ζ^* and a measurable function K with $\mathbb{E}_Q[K(\boldsymbol{z})^2] < \infty$*
$$|m_{\zeta_1}(\boldsymbol{z}) - m_{\zeta_2}(\boldsymbol{z})| \leq K(\boldsymbol{z}) \left\|\zeta_1 - \zeta_2\right\|.$$
4. *assume that the mapping $\zeta \to \mathbb{E}_Q[m_\zeta(\boldsymbol{z})]$ admits a second-order Taylor expansion at a point of maximum ζ^* with nonsigular symmetric second order matrix V_{ζ^*}.*

If the random sequence $\hat{\zeta}_n$ satisfies $\frac{1}{n}\sum_{i=1}^n m_{\hat{\zeta}_n}(\boldsymbol{z}_i) = \sup_\zeta \sum_{i=1}^n m_{\hat{\zeta}}(\boldsymbol{z}_i)$, then $\hat{\zeta}_n \xrightarrow{p} \zeta^*$ and

$$\sqrt{n}\left(\hat{\zeta}_n - \zeta^*\right) = -V_{\zeta^*}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla_\zeta m_{\zeta^*}(\boldsymbol{z}_i) + o_Q(1).$$

Throughout this paper, we assume Assumption 3 holds.

- For SAA, consider $m_\zeta(\boldsymbol{z}) = -c(\boldsymbol{w}, \boldsymbol{z})$ with the parameter $\zeta = \boldsymbol{w}$.
- For ETO, consider $m_\zeta(\boldsymbol{z}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{z})$ with parameter $\zeta = \boldsymbol{\theta}$.
- For IEO, consider $m_\zeta(\boldsymbol{z}) = c(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z})$ with $\zeta = \boldsymbol{\theta}$.

When we say Assumption 3 holds, it means that Assumption 3 holds for the corresponding $m_\zeta(\boldsymbol{z})$ in SAA, ETO, and IEO.

Assumption 4 (Interchangeability)**.** *For any $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{w} \in \Omega$,*

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

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

The interchangeability condition in Assumption 4 is a standard assumption in the Cramer-Rao bound [Bickel and Doksum, 2015]. A standard route to check the interchangeability condition is to use the dominated convergence theorem. For instance, we provide a way to check the first interchange equation. If $p_{\boldsymbol{\theta}}(\boldsymbol{z})$ is continuously differentiable with respect to $\boldsymbol{\theta}$, and there exists a real-valued function $q(\boldsymbol{z})$ such that $\int \nabla_{\boldsymbol{w}} c(\boldsymbol{w}, \boldsymbol{z})^\top q(\boldsymbol{z}) d\boldsymbol{z} < +\infty$ and $\|\nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{z})\|_\infty \leq q(\boldsymbol{z})$, then we have $\nabla_{\boldsymbol{\theta}} \int \nabla_{\boldsymbol{w}} c(\boldsymbol{w}, \boldsymbol{z})^\top p_{\boldsymbol{\theta}}(\boldsymbol{z}) d\boldsymbol{z} = \int \nabla_{\boldsymbol{w}} c(\boldsymbol{w}, \boldsymbol{z})^\top \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{z}) d\boldsymbol{z}$. Other sufficient conditions (more delicate but still based on the dominated convergence theorem) can be found in L'Ecuyer [1990], Asmussen and Glynn [2007], Glasserman [2004].

Next, we present some auxiliary lemmas that are helpful for deriving our theorems.

The first is a classic lemma in asymptotic statistics, called Le Cam's third lemma (Example 6.7 in van der Vaart [2000]).

Lemma 3 (Le Cam's third lemma)**.** *Let P^n and Q^n be sequences of probability measures on measurable spaces $(\Omega_n, \mathcal{F}_n)$ and let X_n be a sequence of random vectors. Suppose that*

$$\left(X_n, \log \frac{dQ^n}{dP^n}\right) \xrightarrow{P^n} N\left(\begin{pmatrix} \boldsymbol{\mu} \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\tau} \\ \boldsymbol{\tau}^\top & \sigma^2 \end{pmatrix}\right),$$

then

$$X_n \overset{Q^n}{\to} N(\boldsymbol{\mu} + \boldsymbol{\tau}, \boldsymbol{\Sigma}).$$

We now state a auxiliary lemma about the directional differentiability of the optimal solutions to stochastic optimization problems.

Lemma 4 (Directional differentiability of optimal solutions: Part I)**.** *Consider the distribution $Q_t(\boldsymbol{z})$ in Definition 1. Let*

$$\boldsymbol{w}_t := \operatorname*{argmin}_{\boldsymbol{w}\in\Omega} \mathbb{E}_{Q_t}\left[c(\boldsymbol{w}, \boldsymbol{z})\right].$$

Then under Assumptions 1, 3, and 4,

$$\lim_{t\to 0} \frac{1}{t}\left(\boldsymbol{w}_t - \boldsymbol{w}_0\right) = \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})].$$

Equipped with the lemma above, we can get the convergence of $n^\alpha\left(\boldsymbol{w}_n^* - \boldsymbol{w}_{\boldsymbol{\theta}_0}\right)$ under the three locally misspecified regimes:

$$\lim_{n\to\infty} n^\alpha\left(\boldsymbol{w}_n^* - \boldsymbol{w}_{\boldsymbol{\theta}_0}\right) = \lim_{t\to 0} \frac{1}{t}\left(\boldsymbol{w}_t - \boldsymbol{w}_0\right) = \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})].$$

Proof of Lemma 4. Note that $\boldsymbol{w}_n^*$ is the minimizer of $\mathbb{E}\left[c(\boldsymbol{w}, \boldsymbol{z})\right]$ under Q^n while $\boldsymbol{w}_{\boldsymbol{\theta}_0}$ under P^n. We will use the directional differentiablity of optimal solution to derive this fact.

We denote $v(\boldsymbol{w}, Q_t)$ as $\mathbb{E}_{Q_t}[c(\boldsymbol{w}, \boldsymbol{z})]$, $\mathcal{G}(\boldsymbol{w}, t) := \nabla_{\boldsymbol{w}} v(\boldsymbol{w}, Q_t)$ and $\boldsymbol{w}_t := \operatorname{argmin} v(\boldsymbol{w}, Q_t)$. Note that $\mathcal{G}(\boldsymbol{w}_t, t) = 0$ for all t. By implicit function theorem,

$$\begin{aligned}
\lim_{t\to 0} \frac{1}{t}\left(\boldsymbol{w}_t - \boldsymbol{w}_0\right) &= -[\nabla_{\boldsymbol{w}}\mathcal{G}(\boldsymbol{w}_{\boldsymbol{\theta}_0}, 0)]^{-1}\frac{\partial}{\partial t}\nabla_{\boldsymbol{w}} v(\boldsymbol{w}_{\boldsymbol{\theta}_0}, Q_t)|_{t=0} \\
&= -\nabla_{\boldsymbol{w}\boldsymbol{w}}\mathbb{E}_{\boldsymbol{\theta}_0}[c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})]\frac{\partial}{\partial t}\nabla_{\boldsymbol{w}} v(\boldsymbol{w}_{\boldsymbol{\theta}_0}, Q_t)|_{t=0} \\
&= -\boldsymbol{V}^{-1}\frac{\partial}{\partial t}\int \nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z}) dQ_t(\boldsymbol{z})|_{t=0} \\
&= -\boldsymbol{V}^{-1}\int \nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\frac{\partial}{\partial t} dQ_t(\boldsymbol{z})|_{t=0}.
\end{aligned}$$

From Definition 1, we know that for almost every $\boldsymbol{z}$, $\frac{\partial}{\partial t}\log q_t(\boldsymbol{z})|_{t=0} = u(\boldsymbol{z})$. Hence, we have for almost every $\boldsymbol{z}$,

$$\frac{\partial}{\partial t} q_t(\boldsymbol{z}) \Big|_{t=0} = q_0(\boldsymbol{z})u(\boldsymbol{z}). \tag{3}$$

In conclusion,

$$\begin{aligned}
&-\boldsymbol{V}^{-1}\int \nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\frac{\partial}{\partial t} dQ_t(\boldsymbol{z})|_{t=0} \\
&= -\boldsymbol{V}^{-1}\int \nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\left\{q_0(\boldsymbol{z})u(\boldsymbol{z})\right\} d\boldsymbol{z}
\end{aligned}$$

Since

$$\int \nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\left[\int q_0(\boldsymbol{z})u(\boldsymbol{z})d\boldsymbol{z}\right] q_0(\boldsymbol{z})d\boldsymbol{z} = \left[\int q_0(\boldsymbol{z})u(\boldsymbol{z})d\boldsymbol{z}\right]\nabla_{\boldsymbol{w}}\mathbb{E}_{\boldsymbol{\theta}_0}[c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})] = 0$$

and

$$\int \nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z}) q_0(\boldsymbol{z})u(\boldsymbol{z})d\boldsymbol{z} = \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})],$$

we have

$$-\boldsymbol{V}^{-1}\int \nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\frac{\partial}{\partial t} dQ_t(\boldsymbol{z})|_{t=0} = -\boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})] = \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})].$$

Therefore,

$$\lim_{n\to\infty}\sqrt{n}(\boldsymbol{w}_n^*-\boldsymbol{w}_{\boldsymbol{\theta}_0})=\lim_{t\to 0}\frac{1}{t}\left(\boldsymbol{w}_t-\boldsymbol{w}_0\right)=\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})].$$

More generally, under severely and mildly specified regime, we have further

$$\lim_{n\to\infty} n^{\alpha}(\boldsymbol{w}_n^*-\boldsymbol{w}_{\boldsymbol{\theta}_0})=\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})].$$

□

We note that Lemma 4 holds for Example 5. To be more specific,

$$q_t(\boldsymbol{z})=\frac{\exp(tu(\boldsymbol{z}))}{C_t}q_0(\boldsymbol{z}) \text{ where } C_t=\int\exp(tu(\boldsymbol{z}))q_0(\boldsymbol{z})d\boldsymbol{z}.$$

Therefore, the derivative of $q_t(\boldsymbol{z})$ with respect to t is

$$\begin{aligned}\frac{\partial}{\partial t}q_t(\boldsymbol{z})=&\frac{q_0(\boldsymbol{z})\exp(tu(\boldsymbol{z}))u(\boldsymbol{z})\left[\int\exp(tu(\boldsymbol{z}))q_0(\boldsymbol{z})d\boldsymbol{z}\right]}{\left[\int\exp(tu(\boldsymbol{z}))q_0(\boldsymbol{z})d\boldsymbol{z}\right]^2}\\&-\frac{\left[\int\exp(tu(\boldsymbol{z}))q_0(\boldsymbol{z})u(\boldsymbol{z})d\boldsymbol{z}\right]q_0(\boldsymbol{z})\exp(tu(\boldsymbol{z}))}{\left[\int\exp(tu(\boldsymbol{z}))q_0(\boldsymbol{z})d\boldsymbol{z}\right]^2}.\end{aligned}$$

At $t=0$, since $\mathbb{E}_{\boldsymbol{\theta}_0}[u]=0$, we have for almost every $\boldsymbol{z}$,

$$\frac{\partial}{\partial t}q_t(\boldsymbol{z})\Big|_{t=0}=q_0(\boldsymbol{z})u(\boldsymbol{z})-\left[\int q_0(\boldsymbol{z})u(\boldsymbol{z})d\boldsymbol{z}\right]q_0(\boldsymbol{z})=q_0(z)u(z).$$

It is also possible to extend the result to other examples under additional regularity assumptions.

For Example 6, the result still holds. Recall that

$$q_t(\boldsymbol{z})=\frac{[1+tu(\boldsymbol{z})]_+}{C_t}q_0(\boldsymbol{z}),\qquad C_t=\int[1+tu(\boldsymbol{z})]_+\,q_0(\boldsymbol{z})d\boldsymbol{z}.$$

Hence,

$$\begin{aligned}\frac{\partial}{\partial t}q_t(\boldsymbol{z})|_{t=0}&=q_0(\boldsymbol{z})\frac{C_t\frac{\partial}{\partial t}[1+tu(\boldsymbol{z})]_+-[1+tu(\boldsymbol{z})]_+\frac{\partial}{\partial t}C_t}{C_t^2}\Big|_{t=0}\\&=q_0(\boldsymbol{z})\left(\frac{\partial}{\partial t}[1+tu(\boldsymbol{z})]_+\Big|_{t=0}-\frac{\partial}{\partial t}C_t\Big|_{t=0}\right)\\&=q_0(\boldsymbol{z})\left(u(\boldsymbol{z})\mathbb{1}\{1+tu(\boldsymbol{z})\geq 0\}\Big|_{t=0}-\int u(\boldsymbol{z})\mathbb{1}\{1+tu(\boldsymbol{z})\geq 0\}\Big|_{t=0}q_0(\boldsymbol{z})d\boldsymbol{z}\right)\\&=q_0(\boldsymbol{z})\left(u(\boldsymbol{z})-\int u(\boldsymbol{z})q_0(\boldsymbol{z})d\boldsymbol{z}\right).\end{aligned}$$

The result is the same as (3) and the conclusion of Lemma 4 still holds.

For Example 7, the result still holds. Recall that

$$q_t(\boldsymbol{z})=\frac{1+g(tu(\boldsymbol{z}))}{C_t}q_0(\boldsymbol{z}),\qquad C_t=\int(1+g(tu(\boldsymbol{z})))\,q_0(\boldsymbol{z})d\boldsymbol{z}.$$

Hence, by noting that $g'(0)=1$,

$$\begin{aligned}\frac{\partial}{\partial t}q_t(\boldsymbol{z})|_{t=0}&=q_0(\boldsymbol{z})\frac{C_t\frac{\partial}{\partial t}(1+g(tu(\boldsymbol{z})))-(1+g(tu(\boldsymbol{z})))\frac{\partial}{\partial t}C_t}{C_t^2}\Big|_{t=0}\\&=q_0(\boldsymbol{z})\left(\frac{\partial}{\partial t}(1+g(tu(\boldsymbol{z})))\Big|_{t=0}-\frac{\partial}{\partial t}C_t\Big|_{t=0}\right)\\&=q_0(\boldsymbol{z})\left(u(\boldsymbol{z})g'(0)-\int u(\boldsymbol{z})g'(0)q_0(\boldsymbol{z})d\boldsymbol{z}\right)\\&=q_0(\boldsymbol{z})\left(u(\boldsymbol{z})-\int u(\boldsymbol{z})q_0(\boldsymbol{z})d\boldsymbol{z}\right).\end{aligned}$$

The result is the same as (3) and the conclusion of Lemma 4 still holds.

We provide another auxiliary lemma similar to Lemma 4.

Lemma 5 (Directional differentiability of optimal solutions: Part II)**.** *Consider the distribution Q_t in Definition 1 where $Q_0 = P_{\boldsymbol{\theta}_0}$. We denote*

$$\boldsymbol{\theta}_t^{\mathrm{KL}} := \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{Q_t}[\log p_{\boldsymbol{\theta}}(\boldsymbol{z})],$$
$$\boldsymbol{\theta}_t^{*} := \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{Q_t}[c(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z})].$$

Then under Assumptions 1, 3, and 4, we have

$$\nabla_t \boldsymbol{\theta}_t^{\mathrm{KL}} := \lim_{t\to 0} \frac{1}{t}\left(\boldsymbol{\theta}_t^{\mathrm{KL}} - \boldsymbol{\theta}_0\right) = \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{ETO}}(\boldsymbol{z})],$$
$$\nabla_t \boldsymbol{\theta}_t^{*} := \lim_{t\to 0} \frac{1}{t}\left(\boldsymbol{\theta}_t^{*} - \boldsymbol{\theta}_0\right) = \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{IEO}}(\boldsymbol{z})].$$

Proof of Lemma 5. We denote $v^{\mathrm{KL}}(\boldsymbol{\theta}, Q_t)$ as $\mathbb{E}_{Q_t}[\log p_{\boldsymbol{\theta}}(\boldsymbol{z})]$, $\mathcal{G}^{\mathrm{KL}}(\boldsymbol{\theta}, t) := \nabla_{\boldsymbol{\theta}} v^{\mathrm{KL}}(\boldsymbol{\theta}, Q_t)$ and $\boldsymbol{\theta}_t^{\mathrm{KL}} := \operatorname{argmin} v^{\mathrm{KL}}(\boldsymbol{\theta}, Q_t)$. Note that $\mathcal{G}^{\mathrm{KL}}(\boldsymbol{\theta}_t^{\mathrm{KL}}, t) = 0$ for all t. By implicit function theorem,

$$\begin{aligned}
\lim_{t\to 0} \frac{1}{t}\left(\boldsymbol{\theta}_t^{\mathrm{KL}} - \boldsymbol{\theta}_0\right) &= -[\nabla_{\boldsymbol{\theta}} \mathcal{G}^{\mathrm{KL}}(\boldsymbol{\theta}_0, 0)]^{-1} \frac{\partial}{\partial t} \nabla_{\boldsymbol{\theta}} v^{\mathrm{KL}}(\boldsymbol{\theta}_0, Q_t)|_{t=0} \\
&= -\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}_0}[\log p_{\boldsymbol{\theta}_0}(\boldsymbol{z})]^{-1} \frac{\partial}{\partial t} \nabla_{\boldsymbol{\theta}} v^{\mathrm{KL}}(\boldsymbol{\theta}_0, Q_t)|_{t=0} \\
&= \boldsymbol{I}^{-1} \frac{\partial}{\partial t} \int \boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z}) dQ_t(\boldsymbol{z})|_{t=0} \\
&= \boldsymbol{I}^{-1} \int \boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z}) \frac{\partial}{\partial t} dQ_t(\boldsymbol{z})|_{t=0}.
\end{aligned}$$

At $t = 0$, by (3),

$$\frac{\partial}{\partial t} q_t(\boldsymbol{z}) \Big|_{t=0} = q_0(\boldsymbol{z}) u(\boldsymbol{z}).$$

In conclusion,

$$\begin{aligned}
&\boldsymbol{I}^{-1} \int \boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z}) \frac{\partial}{\partial t} dQ_t(\boldsymbol{z})|_{t=0} \\
=&\boldsymbol{I}^{-1} \int \boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z}) \left\{q_0(\boldsymbol{z}) h^{\top} u(\boldsymbol{z})\right\} d\boldsymbol{z}.
\end{aligned}$$

Since

$$\int \boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z}) \left[\int q_0(\boldsymbol{z}) u(\boldsymbol{z}) d\boldsymbol{z}\right] q_0(\boldsymbol{z}) d\boldsymbol{z} = \left[\int q_0(\boldsymbol{z}) u(\boldsymbol{z}) d\boldsymbol{z}\right] \mathbb{E}_{\boldsymbol{\theta}_0}\left[\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})\right] = \mathbf{0}$$

and

$$\int \boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z}) q_0(\boldsymbol{z}) u(\boldsymbol{z}) d\boldsymbol{z} = \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z}) \boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})],$$

we have

$$\lim_{t\to 0} \frac{1}{t}\left(\boldsymbol{\theta}_t^{\mathrm{KL}} - \boldsymbol{\theta}_0\right) = \boldsymbol{I}^{-1} \int \boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z}) \frac{\partial}{\partial t} dQ_t(\boldsymbol{z})|_{t=0} = \boldsymbol{I}^{-1} \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z}) \boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})].$$

Similarly, we denote $v^*(\boldsymbol{\theta}, Q_t)$ as $\mathbb{E}_{Q_t}[c(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z})]$, $\mathcal{G}^*(\boldsymbol{\theta}, t) := \nabla_{\boldsymbol{\theta}} v^*(\boldsymbol{\theta}, Q_t)$ and $\boldsymbol{\theta}_t^* := \operatorname{argmin} v^*(\boldsymbol{\theta}, Q_t)$. Note that $\mathcal{G}^*(\boldsymbol{\theta}_t^*, t) = 0$ for all t. By implicit function theorem,

$$\begin{aligned}
\lim_{t\to 0} \frac{1}{t}\left(\boldsymbol{\theta}_t^{*} - \boldsymbol{\theta}_0\right) &= -[\nabla_{\boldsymbol{\theta}} \mathcal{G}^{*}(\boldsymbol{\theta}_0, 0)]^{-1} \frac{\partial}{\partial t} \nabla_{\boldsymbol{\theta}} v^{*}(\boldsymbol{\theta}_0, Q_t)|_{t=0} \\
&= -\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}_0}[c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})]^{-1} \frac{\partial}{\partial t} \nabla_{\boldsymbol{\theta}} v^{*}(\boldsymbol{\theta}_0, Q_t)|_{t=0} \\
&= -\boldsymbol{\Phi}^{-1} \frac{\partial}{\partial t} \int \nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z}) dQ_t(\boldsymbol{z})|_{t=0} \\
&= -\boldsymbol{\Phi}^{-1} \int \nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z}) \frac{\partial}{\partial t} dQ_t(\boldsymbol{z})|_{t=0}.
\end{aligned}$$

At $t = 0$, by (3),

$$\frac{\partial}{\partial t} q_t(\boldsymbol{z}) \Big|_{t=0} = q_0(\boldsymbol{z}) u(\boldsymbol{z}).$$

In conclusion,

$$\begin{aligned} &- \boldsymbol{\Phi}^{-1} \int \nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z}) \frac{\partial}{\partial t} dQ_t(\boldsymbol{z})|_{t=0} \\ = &- \boldsymbol{\Phi}^{-1} \int \nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z}) \left\{ q_0(\boldsymbol{z}) u(\boldsymbol{z}) \right\} d\boldsymbol{z}. \end{aligned}$$

Since

$$\int \nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z}) q_0(\boldsymbol{z}) u(\boldsymbol{z}) d\boldsymbol{z} = \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z}) \nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})],$$

we have

$$\lim_{t \to 0} \frac{1}{t} \left(\boldsymbol{\theta}_t^* - \boldsymbol{\theta}_0 \right) = -\boldsymbol{\Phi}^{-1} \int \nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z}) \frac{\partial}{\partial t} dQ_t(\boldsymbol{z})|_{t=0} = -\boldsymbol{\Phi}^{-1} \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z}) \nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})].$$

□

We remark that Lemma 5 also holds for Q_t in Example 6 and 7.

B Proofs

In this section, we supplement the proof of the results in this paper.

Proof of Theorem 6. We first notice the fact that, in the mildly misspecified regime, by defining $h_n = 1/(n^{\alpha - 1/2}) = o(1)$, we have

$$\log \frac{dQ^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)}{dP^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_n u(\boldsymbol{z}_i) - \frac{1}{2} \mathbb{E}_{\boldsymbol{\theta}_0}[u^2] h_n^2 + o_{P^n}(h_n) = o_{P^n}(1).$$

In the mild misspecified case, under P^n, we have a joint central limit theorem

$$\begin{bmatrix} \sqrt{n}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_{\boldsymbol{\theta}_0}) \\ \log \frac{dQ^n}{dP^n} \end{bmatrix} \overset{P^n}{\to} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \text{var}_{\boldsymbol{\theta}_0}(\text{IF}^{\square}(\boldsymbol{z})) & 0 \\ 0 & 0 \end{bmatrix} \right).$$

Using LeCam's third lemma, we change the measure from P^n to Q^n and get that under Q^n,

$$\sqrt{n}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_{\boldsymbol{\theta}_0}) \overset{Q^n}{\to} N(0, \text{var}_{\boldsymbol{\theta}_0}(\text{IF}^{\square}(\boldsymbol{z}))).$$

Using the same technique,

$$\begin{aligned} n^{\alpha}(\boldsymbol{w}_n^* - \boldsymbol{w}_{\boldsymbol{\theta}_0}) &\to \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z}) \text{IF}^{\text{SAA}}(\boldsymbol{z})], \\ \sqrt{n}(\boldsymbol{w}_n^* - \boldsymbol{w}_{\boldsymbol{\theta}_0}) &\to \mathbf{0}. \end{aligned}$$

In conclusion,

$$\sqrt{n}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_n^*) = \sqrt{n}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_{\boldsymbol{\theta}_0}) - \sqrt{n}(\boldsymbol{w}_n^* - \boldsymbol{w}_{\boldsymbol{\theta}_0}) \overset{Q^n}{\to} N(0, \text{var}_{\boldsymbol{\theta}_0}(\text{IF}^{\square}(\boldsymbol{z}))).$$

Let us now consider the regret. We use Taylor expansion of the regret with respect to $\boldsymbol{w}$ at $\boldsymbol{w}_n^*$ and note that $\nabla_{\boldsymbol{w}} v_n(\boldsymbol{w}_n^*) = 0$ for every n,

$$\begin{aligned} v_n(\hat{\boldsymbol{w}}^{\square}) - v_n(\boldsymbol{w}_n^*) &= \frac{1}{2} (\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_n^*)^{\top} \nabla_{\boldsymbol{w}\boldsymbol{w}} v_n(\boldsymbol{w}_n^*) (\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_n^*) + o_{Q^n}(\left\| \hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_n^* \right\|^2), \\ n(v_n(\hat{\boldsymbol{w}}^{\square}) - v_n(\boldsymbol{w}_n^*)) &= \frac{1}{2} \sqrt{n}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_n^*)^{\top} \nabla_{\boldsymbol{w}\boldsymbol{w}} v_n(\boldsymbol{w}_n^*) \sqrt{n}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_n^*) + o_{Q^n}(1). \end{aligned}$$

By Assumption 2 that $\nabla_{\boldsymbol{w}\boldsymbol{w}} v_n(\boldsymbol{w}_n^*) \to \boldsymbol{V}$, the function $f : \Omega \to \mathbb{R}$ with $f(\cdot) := \frac{1}{2}(\cdot)^{\top} \boldsymbol{V}(\cdot)$ and function sequence $f_n : \Omega \to \mathbb{R}$ with $f_n(\cdot) := \frac{1}{2}(\cdot)^{\top} \nabla_{\boldsymbol{w}\boldsymbol{w}} v_n(\boldsymbol{w}_n^*)(\cdot)$ satisfy: for all sequence $\{\boldsymbol{w}_n\}_{n=1}^{\infty}$, if $\boldsymbol{w}_n \to \boldsymbol{w}$ for some $\boldsymbol{w} \in \Omega$, then $f_n(\boldsymbol{w}_n) \to f(\boldsymbol{w})$ since continuity is preserved under multiplication. Using the extended continuous mapping theorem (Theorem 1.11.1 in van der Vaart and Wellner [1996]), we have under Q^n,

$$n(v_n(\hat{\boldsymbol{w}}^{\square}) - v_n(\boldsymbol{w}_n^*)) \overset{Q^n}{\to} \frac{1}{2} N^{\square} \boldsymbol{V} N^{\square}.$$

Moreover, ETO is stochastically dominated by IEO and IEO is stochastically dominated by SAA. □

Proof of Proposition 1. The asymptotic normality result is directly from van der Vaart [2000] by noting Lemma 1.

The asymptotic normality of SAA is by Proposition 2A of Elmachtoub et al. [2023]. For ETO and IEO, Proposition 2B and 2C of Elmachtoub et al. [2023] shows that

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\text{ETO}} - \boldsymbol{\theta}_0\right) \xrightarrow{P^n} N(\mathbf{0}, \boldsymbol{I}^{-1}),$$
$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\text{IEO}} - \boldsymbol{\theta}_0\right) \xrightarrow{P^n} N(\mathbf{0}, \boldsymbol{\Phi}^{-1} \operatorname{var}_{\boldsymbol{\theta}_0}\left(\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\right) \boldsymbol{\Phi}^{-1}).$$

Regarding the notation, $\operatorname{var}_{\boldsymbol{\theta}_0}\left(\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\right))$ is the variance of the random gradient $\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, under the distribution $P_{\boldsymbol{\theta}_0}$. Note that the subscript $\boldsymbol{\theta}_0$ under the variance is not a variable here. Using the delta method, we have

$$\begin{aligned}
\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\text{ETO}} - \boldsymbol{\theta}_0\right) &\xrightarrow{P^n} N(\mathbf{0}, \nabla_{\boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{\theta}_0}^\top \boldsymbol{I}^{-1} \nabla_{\boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{\theta}_0}) = N(\mathbf{0}, \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\boldsymbol{I}^{-1}\boldsymbol{\Sigma}^\top\boldsymbol{V}^{-1}) = N(0, \operatorname{var}_{\boldsymbol{\theta}_0}(\text{IF}^{\text{ETO}}(\boldsymbol{z}))),\\
\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\text{IEO}} - \boldsymbol{\theta}_0\right) &\xrightarrow{P^n} N(\mathbf{0}, \nabla_{\boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{\theta}_0}^\top \boldsymbol{\Phi}^{-1} \operatorname{var}_{\boldsymbol{\theta}_0}\left(\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\right)\boldsymbol{\Phi}^{-1}\nabla_{\boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{\theta}_0})\\
&= N(\mathbf{0}, \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{-1} \operatorname{var}_{\boldsymbol{\theta}_0}\left(\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\right)\boldsymbol{\Phi}^{-1}\boldsymbol{\Sigma}^\top\boldsymbol{V}^{-1})\\
&= N(0, \operatorname{var}_{\boldsymbol{\theta}_0}(\text{IF}^{\text{IEO}}(\boldsymbol{z}))).
\end{aligned}$$

The inequality $\operatorname{var}_{\boldsymbol{\theta}_0}(\text{IF}^{\text{ETO}}(\boldsymbol{z})) \le \operatorname{var}_{\boldsymbol{\theta}_0}(\text{IF}^{\text{IEO}}(\boldsymbol{z})) \le \operatorname{var}_{\boldsymbol{\theta}_0}(\text{IF}^{\text{SAA}}(\boldsymbol{z}))$ is from Theorem 2 of Elmachtoub et al. [2023]. □

Proof of Theorem 3. We use a different decomposition framework this time. We recall

$$\begin{aligned}
\boldsymbol{\theta}_t^{\text{KL}} &:= \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{Q_t}[\log p_{\boldsymbol{\theta}}(\boldsymbol{z})],\\
\boldsymbol{\theta}_t^{*} &:= \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{Q_t}[c(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z})],\\
\boldsymbol{w}_t^{*} &:= \operatorname*{argmin}_{\boldsymbol{w}\in\Omega} \mathbb{E}_{Q_t}[c(\boldsymbol{w}, \boldsymbol{z})].
\end{aligned}$$

We denote $t_n := 1/n^\alpha$. Note that here $\boldsymbol{w}_{t_n}^* = \boldsymbol{w}_n^*$ but generally $\boldsymbol{w}_{\boldsymbol{\theta}_{t_n}^{\text{KL}}} \neq \boldsymbol{w}_n^*$, $\boldsymbol{w}_{\boldsymbol{\theta}_{t_n}^*} \neq \boldsymbol{w}_n^*$. In this case,

$$\begin{aligned}
\hat{\boldsymbol{w}}^{\text{ETO}} - \boldsymbol{w}_n^* &= (\hat{\boldsymbol{w}}^{\text{ETO}} - \boldsymbol{w}_{\boldsymbol{\theta}_{t_n}^{\text{KL}}}) + (\boldsymbol{w}_{\boldsymbol{\theta}_{t_n}^{\text{KL}}} - \boldsymbol{w}_{\boldsymbol{\theta}_0}) - (\boldsymbol{w}_n^* - \boldsymbol{w}_{\boldsymbol{\theta}_0}),\\
\hat{\boldsymbol{w}}^{\text{IEO}} - \boldsymbol{w}_n^* &= (\hat{\boldsymbol{w}}^{\text{IEO}} - \boldsymbol{w}_{\boldsymbol{\theta}_{t_n}^{*}}) + (\boldsymbol{w}_{\boldsymbol{\theta}_{t_n}^{*}} - \boldsymbol{w}_{\boldsymbol{\theta}_0}) - (\boldsymbol{w}_n^* - \boldsymbol{w}_{\boldsymbol{\theta}_0}),\\
\hat{\boldsymbol{w}}^{\text{SAA}} - \boldsymbol{w}_n^* &= (\hat{\boldsymbol{w}}^{\text{SAA}} - \boldsymbol{w}_{t_n}^*) + (\boldsymbol{w}_{t_n}^* - \boldsymbol{w}_{\boldsymbol{\theta}_0}) - (\boldsymbol{w}_n^* - \boldsymbol{w}_{\boldsymbol{\theta}_0}).
\end{aligned}$$

We already show in Lemma 4 that

$$n^\alpha(\boldsymbol{w}_n^* - \boldsymbol{w}_{\boldsymbol{\theta}_0}) \to \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\text{IF}^{\text{SAA}}(\boldsymbol{z})].$$

Next we give a limit of the middle term, using Taylor expansion. For SAA, the middle term is equal to the third term. For ETO and IEO, $\boldsymbol{w}_{\boldsymbol{\theta}_0} = \boldsymbol{w}_{\boldsymbol{\theta}_t^*}|_{t=0}$ and $\boldsymbol{w}_{\boldsymbol{\theta}_0} = \boldsymbol{w}_{\boldsymbol{\theta}_t^{\text{KL}}}|_{t=0}$.

$$\begin{aligned}
\boldsymbol{w}_{\boldsymbol{\theta}_t^*} - \boldsymbol{w}_{\boldsymbol{\theta}_0} &:= \nabla_t \boldsymbol{w}_{\boldsymbol{\theta}_t^*} + o(t) = \nabla_{\boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{\theta}}^\top \nabla_t \boldsymbol{\theta}_t^* + o(t),\\
\boldsymbol{w}_{\boldsymbol{\theta}_t^{\text{KL}}} - \boldsymbol{w}_{\boldsymbol{\theta}_0} &:= \nabla_t \boldsymbol{w}_{\boldsymbol{\theta}_t^{\text{KL}}} + o(t) = \nabla_{\boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{\theta}}^\top \nabla_t \boldsymbol{\theta}_t^{\text{KL}} + o(t).
\end{aligned}$$

By Lemma 5, we can get $\nabla_t \boldsymbol{\theta}_t^*$ and $\nabla_t \boldsymbol{\theta}_t^{\text{KL}}$ at $t = 0$:

$$\begin{aligned}
\nabla_t \boldsymbol{\theta}_t^{\text{KL}} &= \boldsymbol{I}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})],\\
\nabla_t \boldsymbol{\theta}_t^{*} &= -\boldsymbol{\Phi}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})].
\end{aligned}$$

Moreover,

$$\begin{aligned}
\nabla_t \boldsymbol{w}_{\boldsymbol{\theta}_t^{\text{KL}}} &= \nabla_{\boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{\theta}}^\top \nabla_t \boldsymbol{\theta}_t^{\text{KL}} = -\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\boldsymbol{I}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}(u(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})) = \mathbb{E}_{\boldsymbol{\theta}_0}(u(\boldsymbol{z})\text{IF}^{\text{ETO}}(\boldsymbol{z})),\\
\nabla_t \boldsymbol{w}_{\boldsymbol{\theta}_t^{*}} &= \nabla_{\boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{\theta}}^\top \nabla_t \boldsymbol{\theta}_t^{*} = \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{\theta}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})] = \mathbb{E}_{\boldsymbol{\theta}_0}(u(\boldsymbol{z})\text{IF}^{\text{IEO}}(\boldsymbol{z})).
\end{aligned}$$

Finally, for the middle term,

$$n^\alpha(\boldsymbol{w}_{\boldsymbol{\theta}^*_{t_n}} - \boldsymbol{w}_{\boldsymbol{\theta}_0}) \to \mathbb{E}_{\boldsymbol{\theta}_0}(u(\boldsymbol{z})\text{IF}^{\text{IEO}}(\boldsymbol{z})),$$
$$n^\alpha(\boldsymbol{w}_{\boldsymbol{\theta}^{\text{KL}}_{t_n}} - \boldsymbol{w}_{\boldsymbol{\theta}_0}) \to \mathbb{E}_{\boldsymbol{\theta}_0}(u(\boldsymbol{z})\text{IF}^{\text{ETO}}(\boldsymbol{z})),$$
$$n^\alpha(\boldsymbol{w}^*_{t_n} - \boldsymbol{w}_{\boldsymbol{\theta}_0}) \to \mathbb{E}_{\boldsymbol{\theta}_0}(u(\boldsymbol{z})\text{IF}^{\text{SAA}}(\boldsymbol{z})).$$

For the first term, under Assumption 2 $(\hat{\boldsymbol{w}}^{\text{ETO}} - \boldsymbol{w}_{\boldsymbol{\theta}^{\text{KL}}_{t_n}})$, $(\hat{\boldsymbol{w}}^{\text{IEO}} - \boldsymbol{w}_{\boldsymbol{\theta}^*_{t_n}})$ and $(\hat{\boldsymbol{w}}^{\text{SAA}} - \boldsymbol{w}^*_{t_n})$ are all $O_{Q^n}(1/\sqrt{n})$, then

$$n^\alpha(\hat{\boldsymbol{w}}^{\text{ETO}} - \boldsymbol{w}_{\boldsymbol{\theta}^{\text{KL}}_{t_n}}) \xrightarrow{p} 0,$$
$$n^\alpha(\hat{\boldsymbol{w}}^{\text{IEO}} - \boldsymbol{w}_{\boldsymbol{\theta}^*_{t_n}}) \xrightarrow{p} 0,$$
$$n^\alpha(\hat{\boldsymbol{w}}^{\text{SAA}} - \boldsymbol{w}^*_{t_n}) \xrightarrow{p} 0.$$

When we multiply n^α, the term shrinks in probability to 0. In conclusion,

$$n^\alpha(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}^*_n) \xrightarrow{p} \boldsymbol{b}^{\square}.$$

Let us now consider the regret. We use Taylor expansion of the regret with respect to $\boldsymbol{w}$ at $\boldsymbol{w}^*_n$ and note that $\nabla_{\boldsymbol{w}} v_n(\boldsymbol{w}^*_n) = 0$ for every n,

$$v_n(\hat{\boldsymbol{w}}^{\square}) - v_n(\boldsymbol{w}^*_n) = \frac{1}{2}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}^*_n)^\top \nabla_{\boldsymbol{ww}} v_n(\boldsymbol{w}^*_n)(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}^*_n) + o_{Q^n}(\left\|\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}^*_n\right\|^2),$$
$$n^{2\alpha}(v_n(\hat{\boldsymbol{w}}^{\square}) - v_n(\boldsymbol{w}^*_n)) = \frac{1}{2} n^\alpha(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}^*_n)^\top \nabla_{\boldsymbol{ww}} v_n(\boldsymbol{w}^*_n) n^\alpha(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}^*_n) + o_{Q^n}(1).$$

By Assumption 2 that $\nabla_{\boldsymbol{ww}} v_n(\boldsymbol{w}^*_n) \to \boldsymbol{V}$, the function $f : \Omega \to \mathbb{R}$ with $f(\cdot) := \frac{1}{2}(\cdot)^\top \boldsymbol{V}(\cdot)$ and function sequence $f_n : \Omega \to \mathbb{R}$ with $f_n(\cdot) := \frac{1}{2}(\cdot)^\top \nabla_{\boldsymbol{ww}} v_n(\boldsymbol{w}^*_n)(\cdot)$ satisfy: for all sequence $\{\boldsymbol{w}_n\}_{n=1}^\infty$, if $\boldsymbol{w}_n \to \boldsymbol{w}$ for some $\boldsymbol{w} \in \Omega$, then $f_n(\boldsymbol{w}_n) \to f(\boldsymbol{w})$ since continuity is preserved under multiplication. Using the extended continuous mapping theorem (Theorem 1.11.1 in van der Vaart and Wellner [1996]), we have under Q^n,

$$n^{2\alpha}(v_n(\hat{\boldsymbol{w}}^{\square}) - v_n(\boldsymbol{w}^*_n)) \xrightarrow{p} \frac{1}{2}\left(\boldsymbol{b}^{\square}\right)^\top \boldsymbol{V}\boldsymbol{b}^{\square}.$$

□

Proof of Theorem 4. Recall the influence function of SAA, ETO, IEO:

$$\text{IF}^{\text{SAA}}(\boldsymbol{z}) = -\boldsymbol{V}^{-1}\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z}),$$
$$\text{IF}^{\text{IEO}}(\boldsymbol{z}) = \boldsymbol{V}^{-1}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z}) = -\boldsymbol{V}^{-1} P_{\boldsymbol{\Sigma},\boldsymbol{V}} \nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z}),$$
$$\text{IF}^{\text{ETO}}(\boldsymbol{z}) = -\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\boldsymbol{I}^{-1}\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z}) = -\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top]\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z}).$$

For regret comparison, since $\boldsymbol{b}^{\text{SAA}} = \boldsymbol{0}$, we have $R^{\text{SAA}} = 0$. Also, $R^{\text{IEO}} \geq 0$ and $R^{\text{ETO}} \geq 0$.

By noting that $\boldsymbol{V} = \nabla_{\boldsymbol{ww}} \mathbb{E}_{\boldsymbol{\theta}_0}[c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})]$, we observe that

$$\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\text{IF}^{\text{IEO}}(\boldsymbol{z})]^\top \boldsymbol{V}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\text{IF}^{\text{IEO}}(\boldsymbol{z})] = \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\text{IF}^{\text{IEO}}(\boldsymbol{z})]^\top \boldsymbol{V}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\text{IF}^{\text{SAA}}(\boldsymbol{z})].$$

This is because

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\text{IF}^{\text{IEO}}(\boldsymbol{z})]^\top \boldsymbol{V}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\text{IF}^{\text{IEO}}(\boldsymbol{z})] \\
=&\left(\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})]^\top \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\right)\cdot \boldsymbol{V}\cdot \\
&\left(\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})]\right) \\
=&\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})]^\top \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})] \\
=&\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})]^\top \boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})] \\
=&\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\text{IF}^{\text{SAA}}(\boldsymbol{z})]^\top \boldsymbol{V}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\text{IF}^{\text{IEO}}(\boldsymbol{z})] \\
=&\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\text{IF}^{\text{IEO}}(\boldsymbol{z})]^\top \boldsymbol{V}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\text{IF}^{\text{SAA}}(\boldsymbol{z})].
\end{aligned}$$

Now let us prove $R^{\mathrm{ETO}} - R^{\mathrm{IEO}} \geq 0$.

$$
\begin{aligned}
&2R^{\mathrm{ETO}} \\
=&\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})(\mathrm{IF}^{\mathrm{ETO}}(\boldsymbol{z}) - \mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z}))]^\top \boldsymbol{V} \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})(\mathrm{IF}^{\mathrm{ETO}}(\boldsymbol{z}) - \mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z}))] \\
=&\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{ETO}}(\boldsymbol{z})]^\top \boldsymbol{V} \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{ETO}}(\boldsymbol{z})] \\
&- 2\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{ETO}}(\boldsymbol{z})]^\top \boldsymbol{V} \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})] + \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})]^\top \boldsymbol{V} \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})] \\
&2R^{\mathrm{IEO}} \\
=&\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})(\mathrm{IF}^{\mathrm{IEO}}(\boldsymbol{z}) - \mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z}))]^\top \boldsymbol{V} \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})(\mathrm{IF}^{\mathrm{IEO}}(\boldsymbol{z}) - \mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z}))] \\
=&\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{IEO}}(\boldsymbol{z})]^\top \boldsymbol{V} \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{IEO}}(\boldsymbol{z})] \\
&- 2\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{IEO}}(\boldsymbol{z})]^\top \boldsymbol{V} \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})] + \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})]^\top \boldsymbol{V} \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})] \\
=&-\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{IEO}}(\boldsymbol{z})]^\top \boldsymbol{V} \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})] + \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})]^\top \boldsymbol{V} \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})].
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&2R^{\mathrm{ETO}} - 2R^{\mathrm{IEO}} \\
=&\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{ETO}}(\boldsymbol{z})]^\top \boldsymbol{V} \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{ETO}}(\boldsymbol{z})] \\
&- 2\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{ETO}}(\boldsymbol{z})]^\top \boldsymbol{V} \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})] + \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{IEO}}(\boldsymbol{z})]^\top \boldsymbol{V} \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})] \\
=&\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})]^\top \boldsymbol{I}^{-1}\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\boldsymbol{I}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})] \\
&+ 2\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})]^\top \boldsymbol{I}^{-1}\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})] \\
&+ \mathbb{E}[u(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})]^\top \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})] \\
=&\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})]^\top \boldsymbol{I}^{-1}\left(\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\right)\boldsymbol{I}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})] \\
&+ 2\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})]^\top \boldsymbol{I}^{-1}\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})] \\
&+ \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})]^\top \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})] \\
=&\left\|\left(\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\right)^{1/2}\boldsymbol{I}^{-1}\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})] - \left(\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\right)^{-1/2}\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})]\right\|^2 \\
\geq&0.
\end{aligned}
$$

The last equality is from the fact that

$$
\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{x}^\top \boldsymbol{y} + \boldsymbol{y}^\top \boldsymbol{A}^{-1}\boldsymbol{y} = \left(\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{x} - \boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{y}\right)^\top \left(\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{x} - \boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{y}\right).
$$

In conclusion, we have

$$
R^{\mathrm{ETO}} \geq R^{\mathrm{IEO}} \geq R^{\mathrm{SAA}} = 0.
$$

By the definition of $\boldsymbol{b}^{\square}$ and $R^{\square}$, we know $\left\|\boldsymbol{b}^{\square}\right\|_{\boldsymbol{V}} = \sqrt{2R^{\square}}$. Hence, by the monotonicity of square root function, we have $0 = \left\|\boldsymbol{b}^{\mathrm{SAA}}\right\|_{\boldsymbol{V}} \leq \left\|\boldsymbol{b}^{\mathrm{IEO}}\right\|_{\boldsymbol{V}} \leq \left\|\boldsymbol{b}^{\mathrm{ETO}}\right\|_{\boldsymbol{V}}$. □

Proof of Theorem 5. Part (i): We note that when $u(\boldsymbol{z}) = \boldsymbol{\beta}^\top s_{\boldsymbol{\theta}_0}(\boldsymbol{z})$ for some $\boldsymbol{\beta} \in \mathbb{R}^{d_\theta}$,

$$
\begin{aligned}
&\boldsymbol{b}^{\mathrm{ETO}} \\
=&\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{ETO}}(\boldsymbol{z})] - \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})] \\
=&\boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top]\left(\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top]\right)^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})] - \boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})] \\
=&\boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top]\left(\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top]\right)^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\beta}^\top \boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})] - \boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\beta}^\top \boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})] \\
=&\boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top]\left(\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top]\right)^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top\boldsymbol{\beta}] - \boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top\boldsymbol{\beta}] \\
=&\boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top]\left(\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top]\right)^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top]\boldsymbol{\beta} - \boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top]\boldsymbol{\beta} \\
=&\boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top]\boldsymbol{\beta} - \boldsymbol{V}^{-1}\mathbb{E}_{\boldsymbol{\theta}_0}[\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top]\boldsymbol{\beta} \\
=&\boldsymbol{0}.
\end{aligned}
$$

□

To prove Theorem 1, we need a useful result here. When $\alpha = 1/2$, we can show that the log-likelihood ratio is asymptotically normal characterized by the mean and variance of the perturbation direction. This result is used to convert the asymptotics in P^n to Q^n by conducting a change of measure from P^n to Q^n, and also contributes to the overall asymptotically normal limit of the decision that encompasses the bias term. It will also be leveraged later to prove results in the mild misspecification case.

Lemma 6 (Log Likelihood Ratio Property in Definition 1[Duchi, 2021])**.** *Under Definition 2, when $\alpha = 1/2$, i.e., $Q^n = Q^{\otimes n}_{1/\sqrt{n}}$, the log-likelihood ratio between Q^n and P^n satisfies:*

$$\log \frac{dQ^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)}{dP^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} u(\boldsymbol{z}_i) - \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta}_0}[u^2] + o_{P^n}(1).$$

Proof of Theorem 1. By Lemma 2 and Proposition 1, under P^n, we have a joint central limit theorem

$$\begin{bmatrix} \sqrt{n}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_{\boldsymbol{\theta}_0}) \\ \log \frac{dQ^n}{dP^n} \end{bmatrix} \overset{P^n}{\to} N\left(\begin{bmatrix} 0 \\ -\frac{1}{2}\mathbb{E}_{\boldsymbol{\theta}_0}(u^2) \end{bmatrix}, \begin{bmatrix} \text{var}_{\boldsymbol{\theta}_0}(\text{IF}^{\square}(\boldsymbol{z})) & \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\text{IF}^{\square}(\boldsymbol{z})] \\ \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\text{IF}^{\square}(\boldsymbol{z})^\top] & \mathbb{E}_{\boldsymbol{\theta}_0}(u^2) \end{bmatrix} \right).$$

Using LeCam's third lemma, we change the measure from P^n to Q^n and get that under Q^n,

$$\sqrt{n}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_{\boldsymbol{\theta}_0}) \overset{Q^n}{\to} N(\mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\text{IF}^{\square}(\boldsymbol{z})], \text{var}_{\boldsymbol{\theta}_0}(\text{IF}^{\square}(\boldsymbol{z}))).$$

Next, by Lemma 4 (note that this is not a stochastic convergence but deterministic sequence convergence)

$$\sqrt{n}(\boldsymbol{w}_n^* - \boldsymbol{w}_{\boldsymbol{\theta}_0}) \to \mathbb{E}_{\boldsymbol{\theta}_0}[u(\boldsymbol{z})\text{IF}^{\text{SAA}}(\boldsymbol{z})].$$

In conclusion,

$$\sqrt{n}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_n^*) = \sqrt{n}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_{\boldsymbol{\theta}_0}) - \sqrt{n}(\boldsymbol{w}_n^* - \boldsymbol{w}_{\boldsymbol{\theta}_0}) \overset{Q^n}{\to} N(\boldsymbol{b}^{\square}, \text{var}_{\boldsymbol{\theta}_0}(\text{IF}^{\square}(\boldsymbol{z}))).$$

Let us now consider the regret. We use Taylor expansion of the regret with respect to $\boldsymbol{w}$ at $\boldsymbol{w}_n^*$ and note that $\nabla_{\boldsymbol{w}} v_n(\boldsymbol{w}_n^*) = 0$ for every n,

$$\begin{aligned} v_n(\hat{\boldsymbol{w}}^{\square}) - v_n(\boldsymbol{w}_n^*) &= \frac{1}{2}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_n^*)^\top \nabla_{\boldsymbol{ww}} v_n(\boldsymbol{w}_n^*)(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_n^*) + o_{Q^n}(\left\|\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_n^*\right\|^2), \\ n(v_n(\hat{\boldsymbol{w}}^{\square}) - v_n(\boldsymbol{w}_n^*)) &= \frac{1}{2}\sqrt{n}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_n^*)^\top \nabla_{\boldsymbol{ww}} v_n(\boldsymbol{w}_n^*)\sqrt{n}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_n^*) + o_{Q^n}(1). \end{aligned}$$

By Assumption 2 that $\nabla_{\boldsymbol{ww}} v_n(\boldsymbol{w}_n^*) \to \boldsymbol{V}$, the function $f : \Omega \to \mathbb{R}$ with $f(\cdot) := \frac{1}{2}(\cdot)^\top \boldsymbol{V}(\cdot)$ and function sequence $f_n : \Omega \to \mathbb{R}$ with $f_n(\cdot) := \frac{1}{2}(\cdot)^\top \nabla_{\boldsymbol{ww}} v_n(\boldsymbol{w}_n^*)(\cdot)$ satisfy: for all sequence $\{\boldsymbol{w}_n\}_{n=1}^{\infty}$, if $\boldsymbol{w}_n \to \boldsymbol{w}$ for some $\boldsymbol{w} \in \Omega$, then $f_n(\boldsymbol{w}_n) \to f(\boldsymbol{w})$ since continuity is preserved under mulptiplication. Using the extended continuous mapping theorem (Theorem 1.11.1 in van der Vaart and Wellner [1996]), we have under Q^n,

$$n(v_n(\hat{\boldsymbol{w}}^{\square}) - v_n(\boldsymbol{w}_n^*)) \overset{Q^n}{\to} \frac{1}{2} N^{\square} \boldsymbol{V} N^{\square}.$$

□

Proof of Theorem 2. Recall that:

$$\sqrt{n}(\hat{\boldsymbol{w}}^{\square} - \boldsymbol{w}_n^*) \overset{Q^n}{\to} N^{\square} := N(\boldsymbol{b}^{\square}, \text{var}_{\boldsymbol{\theta}_0}(\text{IF}^{\square}(\boldsymbol{z}))).$$

$$n(v_n(\hat{\boldsymbol{w}}^{\square}) - v_n(\boldsymbol{w}_n^*)) \overset{Q^n}{\to} \mathbb{G}^{\square} := \frac{1}{2}(N^{\square})^\top \boldsymbol{V} N^{\square}.$$

By denoting $\boldsymbol{b}^{\square}$ as $\mathbb{E}_{\boldsymbol{\theta}_0}(u(\boldsymbol{z})(\text{IF}^{\square}(\boldsymbol{z}) - \text{IF}^{\text{SAA}}(\boldsymbol{z})))$, we can rewrite $\mathbb{G}^{\square}$ as

$$\begin{aligned} &\mathbb{G}^{\square} \\ =&\frac{1}{2}\left(N(0, \text{var}_{\boldsymbol{\theta}_0}(\text{IF}^{\square}(\boldsymbol{z}))) - \boldsymbol{b}^{\square}\right)^\top \boldsymbol{V}\left(N(0, \text{var}_{\boldsymbol{\theta}_0}(\text{IF}^{\square}(\boldsymbol{z}))) - \boldsymbol{b}^{\square}\right) \\ =&\frac{1}{2}\left[N(0, \text{var}_{\boldsymbol{\theta}_0}(\text{IF}^{\square}(\boldsymbol{z})))^\top \boldsymbol{V} N(0, \text{var}_{\boldsymbol{\theta}_0}(\text{IF}^{\square}(\boldsymbol{z}))) - 2\left(\boldsymbol{b}^{\square}\right)^\top \boldsymbol{V} N(0, \text{var}_{\boldsymbol{\theta}_0}(\text{IF}^{\square}(\boldsymbol{z}))) + \left(\boldsymbol{b}^{\square}\right)^\top \boldsymbol{V}\boldsymbol{b}^{\square}\right]. \end{aligned}$$

By taking the expectation, the cross term is zero. Hence,

$$\mathbb{E}\left(\mathbb{G}^{\square}\right) = \frac{1}{2}\left[\mathbb{E}\left[N(0, \mathrm{var}_{\boldsymbol{\theta}_0}(\mathrm{IF}^{\square}(\boldsymbol{z})))^{\top}\boldsymbol{V}N(0, \mathrm{var}_{\boldsymbol{\theta}_0}(\mathrm{IF}^{\square}(\boldsymbol{z})))\right] + \left(\boldsymbol{b}^{\square}\right)^{\top}\boldsymbol{V}\boldsymbol{b}^{\square}\right].$$

Since $\mathrm{var}_{\boldsymbol{\theta}_0}(\mathrm{IF}^{\mathrm{ETO}}(\boldsymbol{z})) \leq \mathrm{var}_{\boldsymbol{\theta}_0}(\mathrm{IF}^{\mathrm{IEO}}(\boldsymbol{z})) \leq \mathrm{var}_{\boldsymbol{\theta}_0}(\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z}))$, we know the stochastic dominance of the SAA, IEO and ETO, and their corresponding expectation.

$$\begin{aligned}
&N(0, \mathrm{var}_{\boldsymbol{\theta}_0}(\mathrm{IF}^{\mathrm{ETO}}(\boldsymbol{z})))^{\top}\boldsymbol{V}N(0, \mathrm{var}_{\boldsymbol{\theta}_0}(\mathrm{IF}^{\mathrm{ETO}}(\boldsymbol{z})))\\
\preceq_{\mathrm{st}} &N(0, \mathrm{var}_{\boldsymbol{\theta}_0}(\mathrm{IF}^{\mathrm{IEO}}(\boldsymbol{z})))^{\top}\boldsymbol{V}N(0, \mathrm{var}_{\boldsymbol{\theta}_0}(\mathrm{IF}^{\mathrm{IEO}}(\boldsymbol{z})))\\
\preceq_{\mathrm{st}} &N(0, \mathrm{var}_{\boldsymbol{\theta}_0}(\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})))^{\top}\boldsymbol{V}N(0, \mathrm{var}_{\boldsymbol{\theta}_0}(\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})))
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[N(0, \mathrm{var}_{\boldsymbol{\theta}_0}(\mathrm{IF}^{\mathrm{ETO}}(\boldsymbol{z})))^{\top}\boldsymbol{V}N(0, \mathrm{var}_{\boldsymbol{\theta}_0}(\mathrm{IF}^{\mathrm{ETO}}(\boldsymbol{z})))\right]\\
\leq &\mathbb{E}\left[N(0, \mathrm{var}_{\boldsymbol{\theta}_0}(\mathrm{IF}^{\mathrm{IEO}}(\boldsymbol{z})))^{\top}\boldsymbol{V}N(0, \mathrm{var}_{\boldsymbol{\theta}_0}(\mathrm{IF}^{\mathrm{IEO}}(\boldsymbol{z})))\right]\\
\leq &\mathbb{E}\left[N(0, \mathrm{var}_{\boldsymbol{\theta}_0}(\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})))^{\top}\boldsymbol{V}N(0, \mathrm{var}_{\boldsymbol{\theta}_0}(\mathrm{IF}^{\mathrm{SAA}}(\boldsymbol{z})))\right].
\end{aligned}$$

From pervious analysis, we already know

$$\left(\boldsymbol{b}^{\mathrm{ETO}}\right)^{\top}\boldsymbol{V}\boldsymbol{b}^{\mathrm{ETO}} \geq \left(\boldsymbol{b}^{\mathrm{IEO}}\right)^{\top}\boldsymbol{V}\boldsymbol{b}^{\mathrm{IEO}} \geq \left(\boldsymbol{b}^{\mathrm{SAA}}\right)^{\top}\boldsymbol{V}\boldsymbol{b}^{\mathrm{SAA}}.$$

Therefore, $\mathbb{E}(\mathbb{G}^{\square})$ consist of two terms. For the first term, ETO is less than IEO, and IEO is less than SAA. For the second term, the direction is flipped. □

Proposition 1 was essentially established by Elmachtoub et al. [2023], but here, we express the asymptotic behaviors of solutions more explicitly in terms of influence functions. Moreover, these more explicit expressions arise from a new projection interpretation of influence functions that allows us to describe the performances geometrically, providing another perspective different from Elmachtoub et al. [2023].

To this end, let $\boldsymbol{P}$ be the projection matrix onto the column span of $\boldsymbol{\Sigma}$ with respect to the norm $\|\boldsymbol{x}\|_{\boldsymbol{V}^{-1}}$, i.e.,

$$\boldsymbol{P}\boldsymbol{x} = \underset{\boldsymbol{y}:\boldsymbol{y}\in\mathrm{col}(\boldsymbol{\Sigma})}{\mathrm{argmin}} \|\boldsymbol{y}-\boldsymbol{x}\|^2_{\boldsymbol{V}^{-1}},$$

which has a closed-form expression $\boldsymbol{P} = \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}^{\top}\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}^{\top}\boldsymbol{V}^{-1}$. Second, define the functional $\mathcal{T} : L_2(P_{\boldsymbol{\theta}_0})^{d_w} \to L_2(P_{\boldsymbol{\theta}_0})^{d_w}$ as the projection operator onto the linear function subspace $\{\boldsymbol{A}\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z}) : \boldsymbol{A} \in \mathbb{R}^{d_w\times d_\theta}\}$, i.e., for a square integrable function $\boldsymbol{f}(\boldsymbol{z}) : \mathcal{Z} \to \mathbb{R}^{d_w}$,

$$\mathcal{T}\boldsymbol{f} = \underset{\boldsymbol{g}:\boldsymbol{g}=\boldsymbol{A}\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})}{\mathrm{argmin}} \int \|\boldsymbol{f}(\boldsymbol{z})-\boldsymbol{g}(\boldsymbol{z})\|^2 p_{\boldsymbol{\theta}_0}(\boldsymbol{z})d\boldsymbol{z}.$$

Theorem 7. *Under Assumptions 1, 3 and 4, the influence functions of* IEO *and* ETO *have the following projection interpretation.*

1. $\mathrm{IF}^{\mathrm{IEO}}(\boldsymbol{z}) = -\boldsymbol{V}^{-1}\boldsymbol{P}\nabla_{\boldsymbol{w}}c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})$,
2. $\mathrm{IF}^{\mathrm{ETO}}(\boldsymbol{z}) = -\boldsymbol{V}^{-1}\mathcal{T}\nabla_{\boldsymbol{w}}c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})$.

The above theorem points out that the influence functions of IEO and ETO are essentially projections of that of SAA, either in vector or function spaces, shedding light on the ordering of their variances by the contraction properties of projections.

Proof of Theorem 7. The fact that

$$\mathrm{IF}^{\mathrm{IEO}}(\boldsymbol{z}) = -\boldsymbol{V}^{-1}\boldsymbol{P}\nabla_{\boldsymbol{w}}c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})$$

is because

$$\begin{aligned}
\mathrm{IF}^{\mathrm{IEO}}(\boldsymbol{z}) &= \boldsymbol{V}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{-1}\nabla_{\boldsymbol{\theta}}c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\\
&= \boldsymbol{V}^{-1}\boldsymbol{\Sigma}(\nabla_{\boldsymbol{\theta}}\boldsymbol{w}_{\boldsymbol{\theta}_0}\boldsymbol{V}\nabla_{\boldsymbol{\theta}}\boldsymbol{w}_{\boldsymbol{\theta}_0}^{\top})^{-1}\nabla_{\boldsymbol{\theta}}\boldsymbol{w}_{\boldsymbol{\theta}_0}\nabla_{\boldsymbol{w}}c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\\
&= \boldsymbol{V}^{-1}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^{\top}\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{\Sigma})^{-1}\left(-\boldsymbol{\Sigma}^{\top}\boldsymbol{V}^{-1}\right)\nabla_{\boldsymbol{w}}c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\\
&= -\boldsymbol{V}^{-1}\boldsymbol{P}_{\boldsymbol{\Sigma},\boldsymbol{V}}\nabla_{\boldsymbol{w}}c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z}).
\end{aligned}$$

We then show the relationship between $\text{IF}^{\text{ETO}}(\boldsymbol{z})$ and $\text{IF}^{\text{SAA}}(\boldsymbol{z})$. Let $\mathcal{T}: L_2(p_{\boldsymbol{\theta}_0})^{d_w} \to L_2(p_{\boldsymbol{\theta}_0})^{d_\theta}$ be the projection matrix on the linear function subspace $\{\boldsymbol{A}\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z}) : \boldsymbol{A} \in \mathbb{R}^{d_w \times d_\theta}\}$, i.e., for general function $\boldsymbol{f}(\boldsymbol{z}) : \mathcal{Z} \to \mathbb{R}^{d_\theta}$,

$$\mathcal{T}\boldsymbol{f} = \operatorname*{argmin}_{\boldsymbol{g}:\boldsymbol{g}=\boldsymbol{A}\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})} \int \|\boldsymbol{f}(\boldsymbol{z}) - \boldsymbol{g}(\boldsymbol{z})\|^2 p_{\boldsymbol{\theta}_0}(\boldsymbol{z}) d\boldsymbol{z}.$$

The influence function of ETO is also a projection, i.e.,

$$\text{IF}^{\text{ETO}}(\boldsymbol{z}) = \boldsymbol{V}^{-1}\mathcal{T}\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z}).$$

The reason is as follows. For ETO, it suffices to prove the following fact:

$$\mathcal{T}\boldsymbol{f} = \mathbb{E}_{\boldsymbol{\theta}_0}(\boldsymbol{f}\boldsymbol{s}_{\boldsymbol{\theta}_0}^\top)\boldsymbol{I}^{-1}\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})$$

since $\boldsymbol{\Sigma} = \mathbb{E}[\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})]$. To prove the fact, we need to show that $\boldsymbol{A}^* = \mathbb{E}_{\boldsymbol{\theta}_0}(\boldsymbol{f}(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top)\boldsymbol{I}^{-1}$ is the minimizer of the optimization problem

$$\min_{\boldsymbol{A}\in\mathbb{R}^{d_w\times d_\theta}} \int \|\boldsymbol{f}(\boldsymbol{z}) - \boldsymbol{A}\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})\|^2 p_{\boldsymbol{\theta}_0}(\boldsymbol{z}) d\boldsymbol{z}.$$

Since this is essentially a quadratic optimization problem, the stationary point is the global minimum. Denote the objective function $h(\boldsymbol{A})$ and we require $\nabla_{\boldsymbol{A}} h(\boldsymbol{A}^*) = 0$. In other words, for all $\tilde{i}, \tilde{j}$, $\partial h(\boldsymbol{A})/\partial A_{\tilde{i},\tilde{j}} = 0$. For simplicity, we write $p_{\boldsymbol{\theta}_0}(\boldsymbol{z})$ as $p(\boldsymbol{z})$ and $\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})$ as $s(\boldsymbol{z})$. Note that

$$\begin{aligned} h(\boldsymbol{A}) &= \int_{\boldsymbol{z}\in\mathcal{Z}} \sum_{i=1}^{d_w} (f_i(\boldsymbol{z}) - \sum_{j=1}^{d_\theta} A_{ij} s_j(\boldsymbol{z}))^2 p(\boldsymbol{z}) d\boldsymbol{z} \\ &= \sum_{i=1}^{d_w} \int_{\boldsymbol{z}\in\mathcal{Z}} \left[f_i(\boldsymbol{z})^2 + (\sum_{j=1}^{d_\theta} A_{ij} s_j(\boldsymbol{z}))^2 - 2 f_i(\boldsymbol{z}) \sum_{j=1}^{d_\theta} A_{ij} s_j(\boldsymbol{z}) \right] p(\boldsymbol{z}) d\boldsymbol{z} \end{aligned}$$

We have

$$\partial h(\boldsymbol{A}^*)/\partial A_{\tilde{i},\tilde{j}} = \int_{\boldsymbol{z}\in\mathcal{Z}} \left[-2 f_{\tilde{i}}(\boldsymbol{z}) s_{\tilde{j}}(\boldsymbol{z}) + \left[2 \sum_{j=1}^{d_\theta} A_{\tilde{i}j} s_j(\boldsymbol{z}) \right] \right] p(\boldsymbol{z}) d\boldsymbol{z} = 0.$$

For all $\tilde{i}, \tilde{j}$, we have

$$\int_{\boldsymbol{z}\in\mathcal{Z}} f_{\tilde{i}}(\boldsymbol{z}) s_{\tilde{j}}(\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} = \int_{\boldsymbol{z}\in\mathcal{Z}} \sum_{j=1}^{d_\theta} A_{\tilde{i}j} s_j(\boldsymbol{z}) s_{\tilde{j}}(\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}.$$

Writing in a matrix form, the left hand side is $\mathbb{E}_{\boldsymbol{\theta}_0}(\boldsymbol{f}(\boldsymbol{z})s(\boldsymbol{z})^\top)$. The write hand side is $\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{A}^* s(\boldsymbol{z}) s(\boldsymbol{z})^\top] = \boldsymbol{A}^* \mathbb{E}_{\boldsymbol{\theta}_0}(s(\boldsymbol{z})s(\boldsymbol{z})^\top) = \boldsymbol{A}^* \boldsymbol{I}$. In conclusion, $\boldsymbol{A}^* = \mathbb{E}(\boldsymbol{f}(\boldsymbol{z})s(\boldsymbol{z})^\top)\boldsymbol{I}^{-1}$ and $\mathcal{T}\boldsymbol{f} = \mathbb{E}(\boldsymbol{f}(\boldsymbol{z})\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})^\top)\boldsymbol{I}^{-1}\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z})$. □

Proof of Lemma 1. The first identity follows from

$$\begin{aligned} \boldsymbol{\Sigma}^\top &= \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{w}} v(\boldsymbol{w}, \boldsymbol{\theta})|_{\boldsymbol{w}=\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ &= \nabla_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{w}} c(\boldsymbol{w}, \boldsymbol{z})]|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0, \boldsymbol{w}=\boldsymbol{w}_{\boldsymbol{\theta}_0}} \\ &= \nabla_{\boldsymbol{\theta}} \int \nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z}) p_{\boldsymbol{\theta}_0}(\boldsymbol{z}) d\boldsymbol{z} \\ &= \int \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}_0}(\boldsymbol{z}) \nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})^\top d\boldsymbol{z} \\ &= \int \left(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}_0}(\boldsymbol{z})\right) p_{\boldsymbol{\theta}_0}(\boldsymbol{z}) \left(\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\right)^\top d\boldsymbol{z} \\ &= \mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{z}) \left(\nabla_{\boldsymbol{w}} c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\right)^\top]. \end{aligned}$$

For the second identity, by implicit function theorem and applying Barratt [2018], we can prove the first identity

$$\begin{aligned}
\mathbf{0} &= \nabla_{\boldsymbol{ww}} v(\boldsymbol{w}, \boldsymbol{\theta}_0)|_{\boldsymbol{w}=\boldsymbol{w}_{\boldsymbol{\theta}_0}} \left(\nabla_{\boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)^\top + \nabla_{\boldsymbol{w}} \nabla_{\boldsymbol{\theta}} v(\boldsymbol{w}, \boldsymbol{\theta})|_{\boldsymbol{w}=\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{\theta}=\boldsymbol{\theta}_0}, \\
\Rightarrow \nabla_{\boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} &= -\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{w}} v(\boldsymbol{w}, \boldsymbol{\theta})|_{\boldsymbol{w}=\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{\theta}=\boldsymbol{\theta}_0} \cdot \nabla_{\boldsymbol{ww}} v(\boldsymbol{w}, \boldsymbol{\theta}_0)^{-1}|_{\boldsymbol{w}=\boldsymbol{w}_{\boldsymbol{\theta}_0}}, \\
&= -\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}.
\end{aligned}$$

The third identity follows since

$$\begin{aligned}
\boldsymbol{\Phi} &= \nabla_{\boldsymbol{\theta\theta}} \mathbb{E}_{\boldsymbol{\theta}_0} \left[c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\right] \\
&= \nabla_{\boldsymbol{\theta}} \left(\nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{\theta}_0} \left[c(\boldsymbol{w}_{\boldsymbol{\theta}}, \boldsymbol{z})\right] \nabla_{\boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{\theta}}^\top\right)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
&= \nabla_{\boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{\theta}} \nabla_{\boldsymbol{ww}} \mathbb{E}_{\boldsymbol{\theta}_0} \left[c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\right] \nabla_{\boldsymbol{\theta}} \boldsymbol{w}_{\boldsymbol{\theta}}^\top \\
&= \left(-\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\right) \boldsymbol{V} \left(-\boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1}\right)^\top \\
&= \boldsymbol{\Sigma}^\top \boldsymbol{V}^{-1} \boldsymbol{\Sigma}
\end{aligned}$$

by noting that $\nabla_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{\theta}_0} \left[c(\boldsymbol{w}_{\boldsymbol{\theta}_0}, \boldsymbol{z})\right] = \mathbf{0}$ since $\boldsymbol{w}_{\boldsymbol{\theta}_0}$ is the minimizer of the function $\boldsymbol{w} \to \mathbb{E}_{\boldsymbol{\theta}_0} \left[c(\boldsymbol{w}, \boldsymbol{z})\right]$. □

Proof of Lemma 2.

$$\log \frac{dQ^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)}{dP^n(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)} = \log \prod_{i=1}^n \frac{\exp(u(\boldsymbol{z}_i)/\sqrt{n})}{C_{1/\sqrt{n}}} = \frac{1}{\sqrt{n}} \sum_{i=1}^n u(\boldsymbol{z}_i) - n \log C_{1/\sqrt{n}}.$$

It now suffices to show that $n \log C_{1/\sqrt{n}} = \frac{1}{2} \mathbb{E}_{\boldsymbol{\theta}_0}[u^2] + o_{P^n}(1)$. From the definition of C_t, we know

$$C_t = \int \exp(tu(\boldsymbol{z})) dP_{\boldsymbol{\theta}_0}(\boldsymbol{z}).$$

Taking the derivative, we have

$$(C_t)' |_{t=0} = \int \exp(tu(\boldsymbol{z}))u(\boldsymbol{z}) dP_{\boldsymbol{\theta}_0}(\boldsymbol{z})|_{t=0} = \mathbb{E}_{\boldsymbol{\theta}_0}(u(\boldsymbol{z})) = 0.$$

Taking the second order derivative, we have

$$(C_t)'' |_{t=0} = \int \exp(tu(\boldsymbol{z}))u(\boldsymbol{z})u(\boldsymbol{z}) dP_{\boldsymbol{\theta}_0}(\boldsymbol{z})|_{t=0} = \mathbb{E}_{\boldsymbol{\theta}_0}(u^2).$$

By Talor expansion, we have

$$C_t = 1 + \frac{1}{2} \mathbb{E}_{\boldsymbol{\theta}_0}[u^2] t^2 + o(t^2)$$

In conclusion,

$$\begin{aligned}
n \log C_{1/\sqrt{n}} = n \log \left(1 + \frac{1}{2} \frac{1}{\sqrt{n}} \mathbb{E}_{\boldsymbol{\theta}_0}[u^2] \frac{1}{\sqrt{n}} + o\left(\frac{1}{n}\right)\right) \quad &= n \left(\frac{1}{2} \frac{1}{\sqrt{n}} \mathbb{E}_{\boldsymbol{\theta}_0}[u^2] \frac{1}{\sqrt{n}} + o\left(\frac{1}{n}\right)\right) \\
= \frac{1}{2} \mathbb{E}_{\boldsymbol{\theta}_0}[u^2] + o(1).
\end{aligned}$$

□