Stochastic Gradients under Nuisances

Facheng Yu* **Ronak Mehta** **Alex Luedtke** **Zaid Harchaoui**

University of Washington

Abstract

Stochastic gradient optimization is the dominant learning paradigm for a variety of scenarios, from classical supervised learning to modern self-supervised learning. We consider stochastic gradient algorithms for learning problems whose objectives rely on unknown nuisance parameters, and establish non-asymptotic convergence guarantees. Our results show that, while the presence of a nuisance can alter the optimum and upset the optimization trajectory, the classical stochastic gradient algorithm may still converge under appropriate conditions, such as Neyman orthogonality. Moreover, even when Neyman orthogonality is not satisfied, we show that an algorithm variant with approximately orthogonalized updates (with an approximately orthogonalized gradient oracle) may achieve similar convergence rates. Examples from orthogonal statistical learning/double machine learning and causal inference are discussed.

1 Introduction

Machine learning, statistics, and causal inference rely on risk minimization problems of the form

$$\min_{\theta \in \Theta} \left[L_0(\theta) := \mathbb{E}_{Z \sim \mathbb{P}} \left[\ell_0(\theta; Z) \right] \right], \tag{1}$$

where $\Theta \subseteq \mathbb{R}^d$ is a parameter space, Z is a $\mathcal{Z}$-valued random variable, and $\ell_0 : \Theta \times \mathcal{Z} \to \mathbb{R}$ is a loss function. The quantity $\ell_0(\theta; z)$ describes the performance of a model parametrized by $\theta \in \Theta$ on a test example $z \in \mathcal{Z}$. Given only an oracle that provides a stochastic gradient estimate of the objective (1), practitioners are able to train models ranging from linear functions on tabular data to billion-parameter neural networks on vision and language data.

The success of stochastic gradient descent (SGD) algorithms [Amari, 1993, Bottou and Le Cun, 2005, Bottou and Bousquet, 2007, Ward et al., 2020] has motivated an abundance of work on their theoretical properties under various algorithmic and risk conditions, such as class separability [Soudry et al., 2018], random reshuffling [Gürbüzbalaban et al., 2021], decomposable objectives [Schmidt et al., 2017, Vaswani et al., 2019], quantization noise [Gorbunov et al., 2020], and noise dominance [Sclocchi and Wyart, 2024]. This success has been fueled by machine learning and AI software libraries such as JAX, PyTorch, TensorFlow, and others, which offer a wide range of SGD variants, as long as a loss function can be clearly specified. The gradient is then evaluated automatically on a mini-batch of datapoints and used for stochastic updates.

Though powerful, this recipe takes one thing for granted: that the learner can always compute the risk (or an unbiased estimate thereof). Indeed, many complex learning problems rely on a risk function that is only partially specified up to a class

$$\mathcal{L} := \{ L(\cdot, g) : g \in \mathcal{G} \}, \tag{2}$$

where $\mathcal{G}$ is a possibly infinite-dimensional set and $L : \Theta \times \mathcal{G} \to \mathbb{R}$ is a function of both the target parameter $\theta \in \Theta$ and an unknown *nuisance parameter* $g \in \mathcal{G}$.

*email: fachengy@uw.edu.

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

This framework originates from semiparametric estimation and inference [Levit, 1979, Linnik, 2008, Bickel et al., 1993, Van der Vaart, 2000], where the risk is a Kullback-Leibler (KL) divergence and g provides information about the true data-generating distribution $\mathbb{P}$, but is not of primary scientific interest. However, the partially specified loss formulation from (2) is not limited to semiparametric estimation and inference problems. This framework connects to many areas of interest, including profile likelihood based learning [Murphy and and, 2000, Pavlichin et al., 2019, Hao and Orlitsky, 2019] and distributionally robust learning [Shapiro, 2017, Levy et al., 2020, Mehta et al., 2024].

For instance, profile likelihood based learning reduces (2) by applying a pointwise minimum over $g \in \mathcal{G}$ to then construct a problem that can be solved in $\theta \in \Theta$. Another example arises in applications with distribution shifts, for which g represents an unknown test data distribution that may differ from the one from which the training data were drawn. Distributionally robust learning reduces (2) by instead taking a pointwise maximum over $\mathcal{G}$ and solving the resulting problem. Although the pointwise minimum and maximum are natural reductions, it is often the case that there is a "true" $g_0 \in \mathcal{G}$ and the loss class is reduced by first estimating g_0 with auxiliary data to produce some $\hat{g}$, which we refer to as *double/debiased machine learning*, or DML, following Chernozhukov et al. [2018a]. The problem (1) is then thought to be derived via $L_0(\theta) \equiv L(\theta, g_0)$ in this case (see examples in Sec. 2). This is the focus of this paper.

Despite the prominence of SGD and DML individually, the convergence guarantees of SGD to recover the risk minimizer with a misspecified nuisance parameter remain unknown. Indeed, after producing $\hat{g}$, the user typically solves a (full batch) empirical risk minimization problem, i.e. minimizing a sample average approximation of $L(\cdot, \hat{g})$. In this paper, we aim to fill this gap by proving convergence guarantees on the sequence $(\theta^{(n)})_{n\geq 1}$ generated by updates of the form

$$\theta^{(n)} = \theta^{(n-1)} - \eta S(\theta^{(n-1)}, \hat{g}; Z_n), \tag{3}$$

where $\eta > 0$ is a learning rate, $\mathcal{D}_n := (Z_i)_{i=1}^n$ is a stream of independent data drawn from $\mathbb{P}$, $\hat{g}$ is a nuisance parameter estimate, and $S : \Theta \times \mathcal{G} \times \mathcal{Z} \to \mathbb{R}^d$ is a stochastic gradient oracle satisfying of $\mathbb{E}_{Z\sim\mathbb{P}}[S(\theta, g; Z)] = \nabla_\theta L(\theta, g)$ for all $(\theta, g) \in \Theta \times \mathcal{G}$. In particular, when $\mathcal{G}$ lies within a Banach space equipped with norm $\|\cdot\|_{\mathcal{G}}$, we wish to compare $\theta^{(n)}$ to

$$\theta_\star = \arg\min_{\theta\in\Theta} L(\theta, g_0), \tag{4}$$

given conditions on the degree of misspecification $\|\hat{g} - g_0\|_{\mathcal{G}}$ and (approximate) *Neyman orthogonality* of the risk L [Neyman, 1959].

Intuitively, Neyman orthogonal classes of objectives are instances of (2) whose curvature with respect to θ is insensitive to the choice of g (see Sec. 2 for the formal description). When Neyman orthogonality is satisfied, the double machine learning framework is also known as *orthogonal statistical learning (OSL)* [Zadik et al., 2018, Liu et al., 2022, Foster and Syrgkanis, 2023]. In addition to the obvious computational considerations, we argue that the SGD perspective in this paper also sheds light on the methodological opportunities in DML/OSL. Indeed, while loss functions are typically specified by the chosen architecture, Neyman orthogonality is often achieved by specialized analytic calculations on the part of the user. Although this property is generally seen as a second-order property of the loss, it can also be viewed as a first-order property of the gradient oracle S. As we detail in Sec. 3, it may be easier and more aligned with the spirit of modern machine learning, to craft Neyman orthogonal gradient oracles instead of losses.

Contributions. We prove the first theoretical convergence guarantees for SGD under an unknown nuisance model. We find that $\theta^{(n)}$ converges linearly to a neighborhood of $\theta_\star$—the optimum in the well-specified case—with a radius that has a fourth-power (resp. squared) dependence on $\|\hat{g} - g_0\|_{\mathcal{G}}$ when Neyman orthogonality is (resp. is not) satisfied. Our analysis can also apply to two-stream settings in which the nuisance parameter is learned online alongside the target. We further analyze a new algorithm, called orthogonalized SGD (OSGD), wherein the gradient oracle of a possibly non-orthogonal loss can be iteratively made orthogonal using an "approximately orthogonalized" gradient oracle, which is based on a separate estimation procedure. This algorithm enjoys a convergence guarantee that interpolates between the $\|\hat{g} - g_0\|_{\mathcal{G}}^4$ (nuisance insensitive) and $\|\hat{g} - g_0\|_{\mathcal{G}}^2$ (nuisance sensitive) regimes depending on the quality of the orthogonalizing operator.

We provide an introduction to the OSL/DML setting in Sec. 2. The SGD and OSGD algorithms are described and analyzed in Sec. 3. We discuss related work in Sec. 4; additional discussion can be found in Appx. F. All proofs and numerical illustrations can be found in the Appendix.

2 Orthogonal Statistical Learning

We first introduce various examples of risk functions in the form of (2), then formally introduce Neyman orthogonality and its implications. As is common in learning settings, the risk will be in the form of an expectation,

$$L(\theta, g) = \mathbb{E}_{Z\sim\mathbb{P}}\left[\ell(\theta, g; Z)\right],$$

where $\ell : \Theta \times \mathcal{G} \times \mathcal{Z} \to \mathbb{R}$ is an instance-level loss function. Various assumptions used in the analysis in Sec. 3 (e.g. convexity) may be placed on either the loss ℓ or the risk L. In each example, we provide the structure of the data point Z, the set $\mathcal{G}$, and the loss ℓ, and the true $g_0 \in \mathcal{G}$ to fully specify the problem. Here, we interpret "true" to mean that g_0 is a parameter of the data-generating distribution (e.g. a propensity score in causal inference), or that g_0 satisfies a cost-minimizing or utility-maximizing criterion (as in the profile likelihood or distributional robustness examples from Sec. 1).

Example	$\ell(\theta, g; z)$	g_0
PLM	$\frac{1}{2}(y - g_Y(w) - \langle\theta, x - g_X(w)\rangle)^2$	$(\mathbb{E}_{\mathbb{P}}[Y \mid W], \mathbb{E}_{\mathbb{P}}[X \mid W])$
CATE	$\frac{1}{2}\left(g^{(1)}(x) - g^{(0)}(x) + \frac{(w - g^{\text{prop}}(x))(y - g^{(w)}(x))}{g^{\text{prop}}(x)(1-g^{\text{prop}}(x))} - \langle\theta, x\rangle\right)^2$	$(\mathbb{E}_{\mathbb{P}}[Y \mid W=1, X], \mathbb{E}_{\mathbb{P}}[X \mid W=0, X], \mathbb{E}_{\mathbb{P}}[W \mid X])$
CRR	$-\left[\mu_g^{(1)}(z)\log p_\theta(x) + \mu_g^{(0)}(z)\log(1 - p_\theta(x))\right]$	$(\mathbb{E}_{\mathbb{P}}[Y \mid W=1, X], \mathbb{E}_{\mathbb{P}}[X \mid W=0, X], \mathbb{E}_{\mathbb{P}}[W \mid X])$

Table 1: Examples of Neyman Orthogonal Losses.

Example 1 (Partially Linear Model)**.** Let $Z = (X, Y, W) \sim \mathbb{P}$, where X is an $\mathbb{R}^d$-valued input, Y is a real-valued outcome, and W is a $\mathcal{W}$-valued control or confounder. The space $\mathcal{G}$ is a nonparametric class containing functions of the form

$$g = (g_Y, g_X) : \mathcal{W} \to \mathbb{R} \times \mathbb{R}^d.$$

Following the construction of Robinson [1988], this g is supplied to the loss

$$\ell_{\text{PLM}}(\theta, g; z) = \frac{1}{2}(y - g_Y(w) - \langle\theta, x - g_X(w)\rangle)^2.$$

To ensure $\theta_\star$ can be interpreted via the projection of $\mathbb{E}_{\mathbb{P}}[Y|X, W]$ onto partially linear additive functions, the true nuisance is given by $g_0 = (g_{0,X}, g_{0,Y})$, where

$$g_{0,Y}(w) := \mathbb{E}_{\mathbb{P}}\left[Y \mid W = w\right] \text{ and } g_{0,X}(w) := \mathbb{E}_{\mathbb{P}}\left[X \mid W = w\right].$$

The next example concerns a quantity widely studied in causal inference [Kennedy, 2023].

Example 2 (Conditional Average Treatment Effect)**.** We observe $Z = (X, Y, W) \sim \mathbb{P}$, where W is a binary treatment assignment. The functions in $\mathcal{G}$ are of the form

$$g = (g^{(0)}, g^{(1)}, g^{\text{prop}}) : \mathbb{R}^d \to \mathbb{R} \times \mathbb{R} \times (0, 1),$$

and are evaluated (see van der Laan and Luedtke [2014, Thm. 1]) at the loss

$$\ell_{\text{CATE}}(\theta, g; z) = \frac{1}{2}\left(g^{(1)}(x) - g^{(0)}(x) + \frac{w - g^{\text{prop}}(x)}{g^{\text{prop}}(x)(1 - g^{\text{prop}}(x))}(y - g^{(w)}(x)) - \langle\theta, x\rangle\right)^2.$$

For $g_0 = (g_0^{(0)}, g_0^{(1)}, g_0^{\text{prop}})$ nuisance functions $g_0^{(0)}$ and $g_0^{(1)}$ represent the outcome regressions

$$g_0^{(0)}(x) := \mathbb{E}_{\mathbb{P}}[Y \mid W = 1, X = x] \text{ and } g_0^{(1)}(x) := \mathbb{E}_{\mathbb{P}}[Y \mid W = 0, X = x],$$

whereas $g_0^{\text{prop}}(x) := \mathbb{E}_{\mathbb{P}}[W \mid X = x]$ denotes the propensity score. The minimizer $\theta_\star$ indexes a projection of the conditional average treatment effect $g_0^{(1)} - g_0^{(0)}$ onto linear functions.

Finally, we maintain the data structure from the previous example, but consider a loss corresponding to a different target parameter according to van der Laan et al. [2024].

Example 3 (Conditional Relative Risk)**.** We retain all components of the previous example, changing only the loss and assuming that the outcome Y is binary/non-negative. First, consider the "label" function

$$\mu_g^{(s)}(z) = g^{(s)}(x) + \frac{\mathbb{1}(w = s)}{s g^{\text{prop}}(x) + (1 - s)(1 - g^{\text{prop}}(x))}(y - g^{(s)}(x)),$$

where $\mathbb{1}(\cdot)$ denotes the indicator function, and the logit-linear predictor $p_\theta(x) = e^{\langle \theta, x \rangle}/(1 + e^{\langle \theta, x \rangle})$. To obtain a linear approximation of the log-relative risk $\log(g_0^{(1)}/g_0^{(0)})$, we employ the cross entropy-type loss function

$$\ell_{\text{CRR}}(\theta, g; z) = -\left[\mu_g^{(1)}(z) \log p_\theta(x) + \mu_g^{(0)}(z) \log(1 - p_\theta(x))\right].$$

While the choices of the loss function in Examples 2 and 3 might look opaque to readers outside of causal inference and statistics, they are carefully designed to be *Neyman orthogonal*. To motivate its definition, notice that, invariably, g_0 is unknown to the user. In DML, the user may produce or access some $\hat{g} \in \mathcal{G}$, which is an estimate of g_0 based on independent training data other than the stream $(Z_i)_{i=1}^n$ used to produce $\theta^{(n)}$. It is of clear interest how stochastic optimization algorithms (and their resulting minimizers) behave in light of the misspecification of g_0, and what precise theoretical conditions govern this behavior. Moreover, as we demonstrate in Sec. 3, these same conditions can be used to analyze procedures for which the user may access additional data to progressively improve the estimate $\hat{g}$ and learn $\theta_\star$ simultaneously. We now formally introduce Neyman orthogonality, and by extension, the orthogonal statistical learning (OSL) variant of DML.

Neyman Orthogonality. For a definition that accounts for a possibly infinite-dimensional function class $\mathcal{G}$, we introduce the *directional derivative*, or equivalently, the *derivative operator*.

Definition 1 (Derivative Operator)**.** *For a functional F mapping from a vector space $\mathcal{F}$ to $\mathbb{R}$, we define the (directional) derivative operator* D *as* $\mathrm{D}F(f)[h] := \frac{\mathrm{d}}{\mathrm{d}t} F(f + th) \mid_{t=0}$ *for any $f, h \in \mathcal{F}$. For a vector-valued $F : \mathcal{F} \mapsto \mathbb{R}^d$, this derivative operator can be generalized by taking derivatives coordinate-wise. We define the second-order derivative as* $\mathrm{D}^2 F(f)[h, h'] := \mathrm{D}(\mathrm{D}F(f)[h])[h']$ *for $h, h' \in \mathcal{F}$ and higher-order derivatives similarly. For functionals of multiple variables $F : \mathcal{F} \times \mathcal{G} \to \mathbb{R}$, we use the subscript notation* $\mathrm{D}_f F(f, g)[h]$ *to indicate the directional derivative of $f \mapsto F(f, g)$ with $g \in \mathcal{G}$ fixed.*

We denote by $S_\theta(\theta, g; z) = \nabla_\theta \ell(\theta, g; z)$ the gradient of the loss function *w.r.t.* the target parameter $\theta \in \Theta$. Borrowing terminology from statistics, we call this the *score*, whether ℓ is based on a likelihood or not.[2] This constitutes one particular example of a stochastic gradient oracle S used in (3). Overloading notation, the *population gradient oracle* is defined as $S_\theta(\theta, g) = \mathbb{E}_{Z \sim \mathbb{P}}[\nabla_\theta \ell(\theta, g; Z)]$.

Definition 2 (Neyman Orthogonality)**.** *The population gradient oracle S_θ is* Neyman orthogonal *at $(\theta_\star, g_0)$ over $\mathcal{G}' \subseteq \mathcal{G}$ if*

$$\mathrm{D}_g S_\theta(\theta_\star, g_0)[g - g_0] = 0 \quad \textit{for all } g \in \mathcal{G}'. \tag{5}$$

For $\Theta' \subseteq \Theta$, the population loss L is Neyman orthogonal *at $(\theta_\star, g_0)$ over $\Theta' \times \mathcal{G}'$ if*

$$\mathrm{D}_g \mathrm{D}_\theta L(\theta_\star, g_0)[\theta - \theta_\star, g - g_0] = 0 \quad \textit{for all } (\theta, g) \in \Theta' \times \mathcal{G}'. \tag{6}$$

[2] This notion of (Fisher) score differs from the "score" used in score-based generative modeling [Song et al., 2021]. If ℓ is based on a log-likelihood, then S_θ is the gradient *w.r.t.* the parameter $\theta \in \Theta$, *not* the input $z \in \mathcal{Z}$.

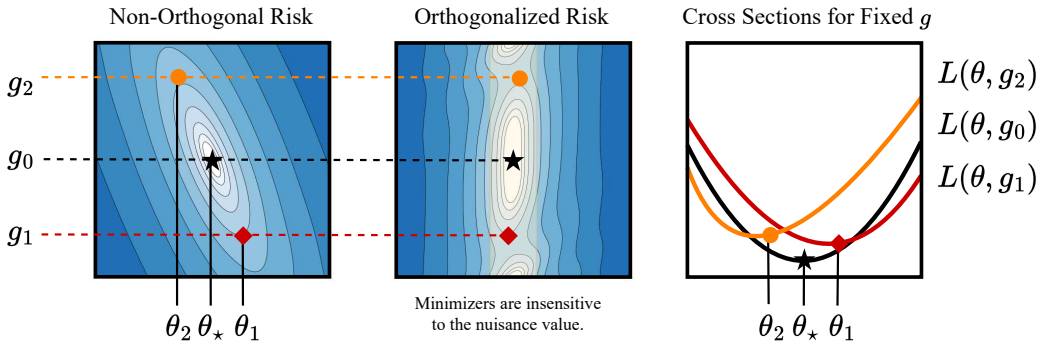

Figure 1: **Illustration of Neyman Orthogonalization.** The first two panels are contour plots of the risk function $L(\theta, g)$, where θ varies on the x-axis and g varies on the y-axis. For the orthogonalized risk (center) the contours are approximately axis-aligned. The right plot shows the cross sections of the non-orthogonal risk when fixing $g = g_0, g_1, g_2$. Due to non-orthogonality, the minimizers θ_1 and θ_2 shown in the first and third plots may drift significantly from $\theta_\star$. In contrast, the minimizers in the center plot are less sensitive to the choice of g.

In Definition 2, we allow $\Theta' \times \mathcal{G}' \subseteq \Theta \times \mathcal{G}$ to be a proper subset, which not only provides a weaker condition, but also accounts for localization-style arguments. Moreover, since $\mathrm{D}_\theta L(\theta_\star, g_0)[\theta - \theta_\star] = \langle S_\theta(\theta_\star, g_0), \theta - \theta_\star \rangle$, if the population risk L satisfies (5), then (6) holds for any target parameter class $\Theta' \subseteq \mathbb{R}^d$. As mentioned above, the risk functions in Examples 1–3 are all Neyman orthogonal at their respective value of $(\theta_\star, g_0)$. In the next section, we will discuss a procedure to make a non-orthogonal gradient oracle "approximately" orthogonal. We illustrate the intended outcome intuitively in Fig. 1 by comparing a generic loss and its orthogonalized counterpart.

3 Stochastic Gradient Optimization

In this section, we propose two stochastic gradient algorithms, which rely on different choices of the stochastic gradient oracle S used in (3). The first is the familiar stochastic gradient oracle that provides a sample estimate of the gradient $\nabla L(\cdot, \hat{g})$ for a fixed estimate $\hat{g}$. The second employs an *approximately orthogonalized gradient oracle*, or OSGD oracle, to achieve a notion of approximate Neyman orthogonality (in a manner we make precise in this section). We analyze the first algorithm under both non-orthogonal and orthogonal settings, achieving an illustrative breakdown of "nuisance sensitive" and "nuisance insensitive" regimes regarding the theoretical convergence guarantee. For the OSGD algorithm, we prove a convergence guarantee that interpolates between the two regimes, depending on the accuracy of the oracle.

Notation and Assumptions. For readers' convenience, a table of all the notation we introduce throughout the paper is collected in Appx. A. We maintain the prototypical bias/variance conditions on the stochastic gradient oracle S, that is, an unbiasedness condition and a second-moment growth condition (see Asm. 3(d)). To dispel confusion, note that by "unbiased", we mean specifically that $\mathbb{E}_{Z \sim \mathbb{P}}[S(\theta, g; Z)] = \nabla_\theta L(\theta, g)$ for all (θ, g), as opposed to the "bias" of replacing $\hat{g}$ with g_0, a terminology sometimes used in DML/OSL. Our analysis will rely partly on the initial distance r of the nuisance estimate $\hat{g}$ to g_0 in $\mathcal{G}$, which defines the ball

$$\mathcal{G}_r(g_0) = \{g \in \mathcal{G} : \|g - g_0\|_{\mathcal{G}} \leq r\}. \tag{7}$$

Various assumptions on the risk will be required to hold only locally, that is, within $\mathcal{G}_r(g_0)$ as opposed to the entire linear space $\mathcal{G}$. Thus, the assumptions become weaker as the estimate $\hat{g}$ improves.

Assumption 3. *The following conditions hold:*

(a) Differentiability: *For any $(z, g) \in \mathcal{Z} \times \mathcal{G}$, $\theta \mapsto \ell(\theta, g; z)$ is twice continuously differentiable. For any $(\theta, g), (\bar{\theta}, \bar{g}) \in \Theta \times \mathcal{G}$, (i) $\mathrm{D}_g^2 \mathrm{D}_\theta \ell(\bar{\theta}, \bar{g}; z)[\theta - \theta_\star, g - g_0, g - g_0]$ exists and is continuous, (ii) $\mathrm{D}_\theta L(\bar{\theta}, \bar{g})[\theta - \theta_\star]$ and $\mathrm{D}_g L(\bar{\theta}, \bar{g})[g - g_0]$ exist,*

and (iii) $\mathbb{E}_{\mathbb{P}}\left[\mathrm{D}_\theta \ell(\bar{\theta}, \bar{g}; Z)[\theta - \theta_\star]\right] = \mathrm{D}_\theta L(\bar{\theta}, \bar{g})[\theta - \theta_\star]$, $\mathbb{E}_{\mathbb{P}}\left[\mathrm{D}_g \ell(\bar{\theta}, \bar{g}; Z)[g - g_0]\right] = \mathrm{D}_g L(\bar{\theta}, \bar{g})[g - g_0]$.

(b) First-order optimality: *The pair* $(\theta_\star, g_0)$ *satisfies* $S_\theta(\theta_\star, g_0) = 0$.

(c) Smoothness and strong convexity: *There exist constants* $M \geq \mu > 0$ *such that for all* $g \in \mathcal{G}_r(g_0)$, *the population risk* $L(\cdot, g)$ *is* M*-smooth and* μ*-strongly convex for* $\theta \in \Theta$.

(d) Second-moment growth: *There exist constants* $K_1, \kappa_1 \geq 0$ *such that*

$$\mathbb{E}_{Z\sim\mathbb{P}}[\|S_\theta(\theta, g; Z) - S_\theta(\theta, g)\|_2^2] \leq K_1 + \kappa_1 \|\theta - \theta_\star\|_2^2 \quad \forall \theta \in \Theta, g \in \mathcal{G}_r(g_0).$$

(e) Second-order smoothness: *There exists a constant* $\alpha_1 \geq 0$ *such that*

$$\left|\mathrm{D}_g \mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0]\right| \leq \alpha_1 \|\theta - \theta_\star\|_2 \|g - g_0\|_{\mathcal{G}} \quad \forall \theta \in \Theta, g, \bar{g} \in \mathcal{G}_r(g_0).$$

Asm. 3 does not require Neyman orthogonality at $(\theta_\star, g_0)$. Instead, Asm. 3(a) is a standard differentiability condition. Asm. 3(b) and (c) implies that $\theta_\star$ is a unique global minimizer. Asm. 3(d) generalizes the uniformly bounded second moment condition in stochastic optimization (e.g. Cutler et al. [2023]) by adding a quadratic form in θ, which allows us to consider an unbounded feasible set Θ and more loss classes. Finally, Asm. 3(d) and (e) can be satisfied when the Hessian of the population risk is a bounded operator. Usually, K_1, κ_1, and α_1 would depend on the initial nuisance estimation distance r. We provide in Appx. B estimates of the constants in Asm. 3 for each motivating example. We proceed to the main results regarding the convergence of SGD and OSGD.

Stochastic Gradient Algorithm. Here, we use the standard single-sample stochastic gradient estimate $S = S_\theta$ in (3). This leads to the update

$$\theta^{(n)} = \theta^{(n-1)} - \eta S_\theta(\theta^{(n-1)}, \hat{g}; Z_n), \quad \theta^{(0)} \in \Theta. \tag{8}$$

While the SGD procedure can be easily extended to using a batch of unbiased gradient estimates, we keep our single-observation construction to highlight the most important aspects of the analysis. In order to achieve quantitative guarantees in the Neyman orthogonal setting, which essentially removes certain second-order terms that include θ and g, we will consider the following higher-order condition in some cases.

Assumption 4 (Higher-Order Smoothness)**.** *The risk* L *satisfies Definition 2 at* $(\theta_\star, g_0)$, *and there exists some constant* $\beta_1 > 0$ *such that*

$$\left|\mathrm{D}_g^2 \mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0, g - g_0]\right| \leq \beta_1 \|\theta - \theta_\star\|_2 \|g - g_0\|_{\mathcal{G}}^2 \quad \forall \theta \in \Theta, g, \bar{g} \in \mathcal{G}_r(g_0).$$

When satisfied, Asm. 4 results in the nuisance insensitivity alluded to at the beginning of this section. Notice that Neyman orthogonality is not necessary to construct a stochastic optimizer, and it is still possible to obtain a nuisance sensitive rate under only Asm. 3. We demonstrate this in Thm. 1.

Theorem 1. *Define* $\mathcal{D}_n = (Z_1, \ldots, Z_n)$, *sampled from the product measure* $\mathbb{P}^n$. *Suppose that Asm. 3 holds,* $\hat{g} \in \mathcal{G}_r(g_0)$ *is estimated independently of* $\mathcal{D}_n$, *and* $\theta^{(0)}, \ldots, \theta^{(n)} \in \Theta$ *almost surely. The iterates of* (8) *satisfy:*

1. Nuisance sensitive: *If* $\eta \leq \mu/2(M\mu + \kappa_1)$, *then*

$$\mathbb{E}_{\mathcal{D}_n\sim\mathbb{P}^n}[\|\theta^{(n)} - \theta_\star\|_2^2] \leq \left(1 - \frac{\mu\eta}{2}\right)^n \|\theta^{(0)} - \theta_\star\|_2^2 + \frac{2\alpha_1^2}{\mu^2}\|\hat{g} - g_0\|_{\mathcal{G}}^2 + \frac{4K_1\eta}{\mu}.$$

2. Nuisance insensitive: *If Asm. 4 also holds, then, for* $\eta \leq \mu/2(M\mu + \kappa_1)$,

$$\mathbb{E}_{\mathcal{D}_n\sim\mathbb{P}^n}[\|\theta^{(n)} - \theta_\star\|_2^2] \leq \left(1 - \frac{\mu\eta}{2}\right)^n \|\theta^{(0)} - \theta_\star\|_2^2 + \frac{\beta_1^2}{2\mu^2}\|\hat{g} - g_0\|_{\mathcal{G}}^4 + \frac{4K_1\eta}{\mu}.$$

Note that the assumption that the iterates remain in Θ is satisfied in common cases. It is satisfied trivially for the first two examples in Sec. 2 because $\Theta = \mathbb{R}^d$. Another case is when the loss decomposes into the sum of a G-Lipschitz continuous component and the ℓ_2^2-norm regularizer, i.e.

$\ell(\theta, g; z) = h(\theta, g; z) + \frac{\mu}{2}\|\theta\|_2^2$. Then, the iterates and the optimum remain in $\{\theta : \|\theta\|_2 \leq G/\mu\}$ (see, e.g., Mehta et al. [2023, Appx. C]), so Definition 2 can be restricted to this compact set.

Thm. 1 states that SGD converges linearly to a ball around $\theta_\star$ with a radius that depends on the bias (due to the replacement of g_0 with $\hat{g}$) and the variance due to gradient noise. Moreover, the variance component decays proportionally to the learning rate η. Under Asm. 4, the bias component can have a significantly more favorable scaling with the error in the nuisance estimate $\|\hat{g} - g_0\|_{\mathcal{G}}$—specifically, $\|\hat{g} - g_0\|_{\mathcal{G}}^4$ instead of $\|\hat{g} - g_0\|_{\mathcal{G}}^2$. A similar breakdown into two regimes of the bias scaling occurs in the works of both Foster and Syrgkanis [2023] and Liu et al. [2022] under Asm. 4 (called "slow rate" and "fast rate" there). Importantly, their bounds are based on an exact, offline empirical risk minimization procedure for a fixed training set, i.e. they provide excess risk bounds on the quantity $L(\hat{\theta}_n, g_0) - L(\theta_\star, g_0)$, where

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, \hat{g}; Z_i).$$

In contrast, Thm. 1 accounts for both the expected distance to optimum and the interplay between bias incurred by $\hat{g}$ and the progress achieved at each step. In particular, even when using a constant learning rate, the bias does not accrue on each iterate and is in fact constant in n. When $\hat{\theta}_n$ is designed to be doubly robust, using SGD can achieve *double robustness*; see Appx. F.5 for an example.

Orthogonalized Stochastic Gradient Algorithm. Given the marked improvement in the rate of decay of the bias term when an orthogonal loss is used, it is clearly beneficial to do so when possible. We now describe how we can induce orthogonality by adjusting the stochastic gradient oracle using the solution of an auxiliary problem.

The construction of Neyman orthogonal losses has historically been motivated in semiparametric theory and statistical learning as a means to build efficient – minimum asymptotic variance – *full batch* statistical estimators [Tsiatis, 2006, Van der Vaart, 2000, Foster and Syrgkanis, 2023, Chernozhukov et al., 2018b]. The approach we follow is inspired by the construction reviewed in Chernozhukov et al. [2018a, Section 2.2]; see also Luedtke [2024]. We also give an intuitive explanation based on least-squares estimation, instead of the usual differential/information geometry one.

While our construction holds in general spaces, let us first consider the illustrative case when $\mathcal{G} = \mathbb{R}^k$. At the true parameters $(\theta_\star, g_0)$, consider the problem of finding the best predictor of the $\mathbb{R}^d$-valued target variable $S_\theta(\theta_\star, g_0; Z) = \nabla_\theta \ell(\theta_\star, g_0; Z)$ given the $\mathbb{R}^k$-valued predictor $\nabla_g \ell(\theta_\star, g_0; Z)$ variable in the space $\mathcal{L}(\mathcal{G}, \Theta)$ containing all continuous and linear operators from $\mathcal{G}$ to Θ:

$$\Gamma_0 = \arg\min_{\Gamma \in \mathcal{L}(\mathcal{G}, \Theta)} \mathbb{E}_{\mathbb{P}} \left[\|S_\theta(\theta_\star, g_0; Z) - \Gamma \nabla_g \ell(\theta_\star, g_0; Z)\|_2^2 \right]. \tag{9}$$

In the special case where $\ell(\theta, g; z) = -\log p_{\theta,g}(z)$ for a density $p_{\theta,g}$ on $\mathcal{Z}$ that governs the random variable Z, the projection direction solving (9) can be shown to satisfy $\Gamma_0 = H_{\theta g}^\top H_{gg}^{-1}$, where $H_{\theta g} = \nabla_g S_\theta(\theta_\star, g_0) \in \mathbb{R}^{k \times d}$ is the transposed Jacobian and $H_{gg} = \nabla_g^2 L(\theta_\star, g_0) \in \mathbb{R}^{k \times k}$ is the Hessian. The prediction $\Gamma_0 \nabla_g \ell(\theta_\star, g_0; Z)$ accounts for the covariance between $\nabla_\theta \ell(\theta_\star, g_0; Z)$ (the gradient *w.r.t.* θ) and $\nabla_g \ell(\theta_\star, g_0; Z)$ (the gradient *w.r.t.* g). It stands to reason that as $\theta \to \theta_\star$, the random vector

$$S(\theta, g_0; Z) := S_\theta(\theta, g_0; Z) - \Gamma_0 \nabla_g \ell(\theta, g_0; Z) = S_\theta(\theta, g_0; Z) - H_{\theta g}^\top H_{gg}^{-1} \nabla_g \ell(\theta, g_0; Z) \tag{10}$$

would be less sensitive to perturbations of g_0, as the component of $S_\theta(\theta, g_0; Z)$ that is predictable through changes in g_0 is subtracted out. Furthermore, if we are aware that the expectation of S is made zero at $\theta_\star$, then a stochastic gradient scheme based on (10) could conceivably achieve a nuisance insensitive rate guarantee in lieu of Thm. 1. From a variance reduction viewpoint, the correction term in (10) subtracts the regression of the θ gradient of the loss on the g "gradient" of the loss. By the law of total variance, the variance of the gradient reduces and improves the trajectory of stochastic optimization; see Appx. F.4 for more details. This variational description (10) hints at how such an operator can be computed algorithmically, instead of the historical approach of deriving the operator via calculation by hand on case by case basis.

Supported by this illustration, we define a generalization that will provide a modified stochastic gradient oracle to use for optimization purposes. Without assuming that ℓ is a negative log-likelihood, we generalize the formulas for $\nabla_g \ell(\theta, g; z) \in \mathbb{R}^k$, $H_{\theta g} \in \mathbb{R}^{k \times d}$ and $H_{gg} \in \mathbb{R}^{k \times k}$ for when $\mathcal{G} \equiv (\mathcal{G}, \langle \cdot, \cdot \rangle_{\mathcal{G}})$ is an infinite-dimensional Hilbert space. Under regularity conditions on the directional derivatives of L, we have that $\nabla_g \ell(\theta, g; z) \in \mathcal{G}$ for all $z \in \mathcal{Z}$, $H_{\theta g} = (H_{\theta g}^{(1)}, \ldots, H_{\theta g}^{(d)}) \in \mathcal{G}^d$, and $H_{gg} : \mathcal{G} \to \mathcal{G}$ is a bounded and self-adjoint operator. The formal details of their construction are contained in Appx. D. Just as in (10), we may consider the operator $\Gamma_0 : \mathcal{G} \to \mathbb{R}^d$, defined elementwise by $[\Gamma_0 g]_j = \langle H_{\theta g}^{(j)}, H_{gg}^{-1} g \rangle_{\mathcal{G}}$, where the invertibility of H_{gg} is satisfied by our assumptions preceding Thm. 3. As shown in (9), the orthogonalizing Γ_0 is defined by both the true nuisance g_0 and the target $\theta_\star$, where g_0 can usually be learned as some conditional expectation and $\theta_\star$ can be learned by our proposed methods. We then construct the central object of the upcoming Thm. 3: the *Neyman orthogonalized (NO) gradient oracle*

$$S_{\text{no}}(\theta, g; z) = S_\theta(\theta, g; z) - \Gamma_0 \nabla_g \ell(\theta, g; z). \tag{11}$$

Lemma 2. *Suppose that Asm. 3(a) holds and* $\mathrm{D}_g^2 L(\theta_\star, g_0)[\cdot, \cdot] : \mathcal{G} \times \mathcal{G} \mapsto \mathbb{R}$ *is a bounded and symmetric bilinear form. Then the NO gradient oracle* $S_{no}(\theta, g; z)$ *is Neyman orthogonal at* (θ_0, g_0).

We refer readers to Lem. 15 for the proof. In this context, we refer to the operator Γ_0 as the "orthogonalizing operator". As a natural sanity check, we note that for a risk function that is already Neyman orthogonal at $(\theta_\star, g_0)$, the NO score S_{no} is exactly equal to score function S_θ itself since $\Gamma_0 = 0$. To construct S_{no} for the non-orthogonal loss, we provide the following example in partially linear model where the corresponding derivations of Γ_0 and S_{no} are included in Appx. B.1.2.

Example 4 (Partially Linear Model)**.** In addition to Example 1, suppose that $Z = (X, Y, W) \sim \mathbb{P}$ satisfies

$$Y = \langle \theta_\star, X \rangle + g_0(W) + \epsilon,$$

where $\theta_\star \in \mathbb{R}^d$ is the true parameter, $g_0 : \mathcal{W} \mapsto \mathbb{R}$ is the true nuisance function, and $\mathbb{E}_{\mathbb{P}}[\epsilon \mid X, W] = 0$. The space $\mathcal{G} \in L_2(\mathbb{P})$ with inner product $\langle g_1, g_2 \rangle_{\mathcal{G}} = \mathbb{E}_{\mathbb{P}}[g_1(W) g_2(W)]$ for any $g_1, g_2 \in \mathcal{G}$ is a nonparametric class containing functions of the form

$$g : \mathcal{W} \to \mathbb{R}.$$

Consider the following non-orthogonal squared loss function:

$$\tilde{\ell}_{\text{PLM}}(\theta, g; z) = \frac{1}{2}[y - g(w) - \langle \theta, x \rangle]^2.$$

The orthogonalizing operator for this non-orthogonal loss is

$$\Gamma_0 : g \mapsto \mathbb{E}_{\mathbb{P}}[\mathbb{E}_{\mathbb{P}}[X \mid W] g(W)],$$

and the NO gradient oracle is obtained as

$$S_{\text{no}}(\theta, g; z) = -(y - g(w) - \langle \theta, x \rangle)(x - \mathbb{E}_{\mathbb{P}}[X \mid W = w]).$$

Motivated by the advantage of a Neyman orthogonal score, we now construct our OSGD algorithm using an estimated the NO score S_{no}. While Γ_0 (like g_0) is unknown to the user in general, using an arbitrary estimate $\hat{\Gamma}$, we can define the *estimated NO score* $\hat{S}_{\text{no}}$ oracle via

$$\hat{S}_{\text{no}}(\theta, g; z) = S_\theta(\theta, g; z) - \hat{\Gamma} \nabla_g \ell(\theta, g; z). \tag{12}$$

Usually, one can obtain such an estimate $\hat{\Gamma}$ using the same data stream of $\hat{g}$; we discuss possible strategies in Appx. F.3. Finally, using $\hat{S}_{\text{no}}$ as the stochastic gradient oracle S in (3), we derive the OSGD update

$$\theta^{(n)} = \theta^{(n-1)} - \eta \hat{S}_{\text{no}}(\theta^{(n-1)}, \hat{g}; Z_n), \quad \theta^{(0)} \in \Theta. \tag{13}$$

To measure the quality of $\hat{\Gamma}$ in our analysis, we use the Frobenius norm $\|\Gamma\|_{\text{Fro}}^2 = \sum_{j=1}^d \|\Gamma^{(j)}\|_{\text{op}}^2$ where $\Gamma : \mathcal{G} \to \mathbb{R}^d$, $\Gamma^{(j)} : g \mapsto [\Gamma g]_j$ and $\|\cdot\|_{\text{op}}$ denotes the usual operator norm for linear functionals. As an example, by the uniqueness of Riesz representations, $\|\Gamma_0\|_{\text{Fro}}^2 = \sum_{j=1}^d \|H_{gg}^{-1} H_{\theta g}^{(j)}\|_{\mathcal{G}}^2$.

Using this modified oracle (12) requires similar assumptions to those used in Thm. 1. For ease of presentation, we defer the formal assumption statement to Appx. E, but note that the result depends on the constants $(\mu_{\text{no}}, M_{\text{no}}, \alpha_2, \beta_2, K_2)$, which are exactly analogous to $(\mu, M, \alpha_1, \beta_1, K_1)$ from Asm. 3.

Theorem 3. *Consider the setting of Thm. 1, with the addition of Asm. 6. When* $\|\hat{\Gamma} - \Gamma_0\|_{\text{Fro}} < \mu_{no}/(4\alpha_1)$ *and*

$$\eta \leq \frac{\mu_{no} - 4\alpha_1 \|\hat{\Gamma} - \Gamma_0\|_{\text{Fro}}}{12 M_{no}^2 - 3\mu_{no}^2/2 + 4(\kappa_1 + \kappa_2 \|\hat{\Gamma}\|_{\text{Fro}}^2)},$$

the iterates of (13) *satisfy:*

$$\mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}^n}[\|\theta^{(n)} - \theta_\star\|_2^2] \leq \left(1 - \frac{\mu_{no}\eta}{2}\right)^n \|\theta^{(0)} - \theta_\star\|_2^2 + \frac{4(K_1 + K_2\|\hat{\Gamma}\|_{\text{Fro}}^2)\eta}{\mu_{no}} + \frac{3}{\mu_{no}^2}\left(\beta_2^2 \|\hat{g} - g_0\|_{\mathcal{G}}^4 + 4\alpha_2^2 \|\hat{g} - g_0\|_{\mathcal{G}}^2 \cdot \|\hat{\Gamma} - \Gamma_0\|_{\text{Fro}}^2\right). \tag{14}$$

Compared with Thm. 1, Thm. 3 shows that OSGD can outperform the nuisance sensitive rate through the correction term $\|\hat{g} - g_0\|_{\mathcal{G}}^2 \cdot \|\hat{\Gamma} - \Gamma_0\|_{\text{Fro}}^2$, and can align with the nuisance insensitive rate when $\|\hat{\Gamma} - \Gamma_0\|_{\text{Fro}}$ is of the order $\mathcal{O}(\|\hat{g} - g_0\|_{\mathcal{G}})$. With slightly different assumptions, Thm. 3 can further simplified – see Appx. E for details.

Interleaving Target and Nuisance Estimation. The results seen thus far have considered for simplicity the estimate $\hat{g}$ to be a fixed element of $\mathcal{G}$, and included terms that depend on the discrepancy $\|\hat{g} - g_0\|_{\mathcal{G}}$. Part of the convenience of these results is that if $\hat{g} \equiv \hat{g}^{(m)}$ is the result of a learning procedure with m independent data points, then statistical bounds on $\|\hat{g}^{(m)} - g_0\|_{\mathcal{G}}$ (either in expectation or high probability, depending on the situation) can be plugged in to quantify the bias. While the results naturally account for full batch learning procedures, they are also amenable to analyzing staggered procedures in which two data sources are queried to estimate $\theta_\star$ and g_0, respectively. To our knowledge, this is the first theoretical analysis of such an orthogonal stochastic learning method.

To be precise, suppose that we update the nuisance estimator for m times, leading to the sequence $\hat{g}^{(1)}, \ldots, \hat{g}^{(m)}$ on a stream of $\mathcal{W}$-valued data $W_1, \ldots, W_m$, sampled i.i.d. from a probability measure $\mathbb{Q}$. We define $\theta^{(0,n)} = \theta^{(0)} \in \Theta$, and for the update of $\hat{g}^{(i)}$ for $1 \leq i \leq m$, we define $\theta^{(i,0)} = \theta^{(i-1,n)}$ and produce the sequence $\theta^{(i,1)}, \ldots, \theta^{(i,n)}$ using n steps of the SGD update (8) initialized at $\theta^{(i,0)}$. Consider, for example, the case in which $\mathcal{G}$ is a reproducing kernel Hilbert space (RKHS) with kernel $k(\cdot, \cdot)$. With the assumption that the eigenvalues $(\lambda_j)_{j \geq 1}$ of covariance operator $\mathbb{E}_{\mathbb{Q}}[k(W, \cdot) \otimes k(W, \cdot)]$ decay polynomially at order $j^{-\alpha}$, the nonparametric stochastic gradient algorithm of Dieuleveut and Bach [2016] satisfies $\mathbb{E}_{\mathbb{Q}^m}\left[\|\hat{g}^{(m)} - g_0\|_{\mathcal{G}}^2\right] = \mathcal{O}(m^{-(2\alpha-1)/(2\alpha)})$. This leads to the following nuisance sensitive rate for a non-Neyman orthogonal loss, by Prop. 22:

$$\mathbb{E}_{\mathbb{P}^{mn} \otimes \mathbb{Q}^m}[\|\theta^{(m,n)} - \theta_\star\|_2^2] = \mathcal{O}\left((1 - \mu\eta/2)^{mn} + m^{-\frac{2\alpha-1}{2\alpha}} + n^{-1} + \eta\right).$$

As another example, suppose that, in addition, we can estimate $\hat{\Gamma} \equiv \hat{\Gamma}^{(m)}$ using the nonparametric stochastic gradient algorithm of Dieuleveut and Bach [2016] and using the same data stream $(W_1, \ldots, W_m)$. If there are high probability bounds for $\|\hat{g}^{(m)} - g_0\|_{\mathcal{G}}^2$ and $\|\hat{\Gamma}^{(m)} - \Gamma_0\|_{\text{Fro}}^2$ of the same order as $\mathcal{O}(m^{-(2\alpha-1)/(2\alpha)})$ and $\|\theta^{(m,n)} - \theta_\star\|_2^2$ decays as described in Thm. 3, then we have in Prop. 23 that $\|\theta^{(m,n)} - \theta_\star\|_2^2 = \mathcal{O}_p\left((1 - \mu\eta/2)^{mn} + m^{-(2\alpha-1)/\alpha} + n^{-1} + \eta\right)$ where the $O_p(m^{-(2\alpha-1)/\alpha})$ nuisance bias term decays quadratically faster than the one for a non-Neyman orthogonal loss. We refer the reader to Appx. F.3 for further details of this analysis.

4 Related Work

We summarize in this section our discussion of the related work. Additional discussions, as well as calculations supporting them, can be found in Appx. F. Possible extensions to SGD variants such as SGD with momentum, averaged SGD, and Adam, are explored in Appx. H.

From an optimization perspective, it is helpful to know how our convergence bounds perform in the idealized case of a known nuisance, which is equivalent to (1). In this case, Thm. 1 gives the convergence rate $\mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}^n}[\|\theta^{(n)} - \theta_\star\|_2^2] = \mathcal{O}((1 - \mu\eta/2)^n + \eta)$, which aligns with the non-asymptotic SGD convergence rates, in mean-square error [Bach and Moulines, 2011] and in high-probability [Cutler et al., 2023]. Our result requires a smaller learning rate $\eta < \mu/2(M\mu + 2\kappa_1)$ when compared to the requirement $\eta < 1/(2M)$ from Cutler et al. [2023]. This is entirely due to our bounded moment assumption (see Asm. 3(d)), which contrasts with a uniform boundedness assumption over all $\Theta \times \mathcal{G}_r(g_0)$. In addition, when the uniform moment bound holds true, κ_1 becomes zero, and our learning rate requirement becomes $\eta < 1/(2M)$.

The comparison with unbiased SGD, and biased SGD, respectively, is also valuable. In the biased SGD literature, the "bias" refers to the fact that $\mathbb{E}_{Z\sim\mathbb{P}}[S(\theta, \hat{g}; Z)] \neq \nabla_\theta L(\theta, g_0)$ in general. The convergence radius then depends on the average value of $\|\mathbb{E}_{Z\sim\mathbb{P}}\left[S(\theta^{(n)}, \hat{g}; Z)\right] - \nabla_\theta L(\theta^{(n)}, g_0)\|_2^2$. Results along this line result in a radius that may not scale with η; see Demidovich et al. [2023, Thm. 3]. Although this form of bias may be related to $\|\hat{g} - g_0\|_{\mathcal{G}}$ under Lipschitzness conditions on the oracle, it is unclear how to effectively incorporate Neyman orthogonality into these general-purpose approaches. Our approach naturally leverages Neyman orthogonality whenever it holds.

In the general case of an unknown nuisance, Foster and Syrgkanis [2023], Chernozhukov et al. [2018b] consider full batch learning methods based on analytically crafted Neyman orthogonal risk functions in various scenarios. For regression functionals, the procedure from Chernozhukov et al. [2022] using random forests or neural networks can ensure that the bias term $\|\hat{g} - g_0\|_{\mathcal{G}}^2$ is asymptotically negligible for large samples, in the sense that classical statistical confidence sets for $\theta_\star$ are asymptotically valid. These papers are focused on algorithm-independent statistical properties.

Our work fills this gap, by providing non-asymptotic convergence guarantees for stochastic gradient algorithms under unknown nuisances. Moreover, the modified stochastic gradient oracle moreover offers a flexible solution to deal with general risk functions. If deriving an orthogonalized risk by hand is difficult or impossible, then the strategy we propose can be applied, and Thm. 3 demonstrates that, when the learning rate η is set appropriately, the convergence rate using the modified stochastic gradient oracle can be improved to

$$\mathcal{O}\Big(\Big(1 - \frac{\mu_{\text{no}}\eta}{2}\Big)^n + \underbrace{\|\hat{g} - g_0\|_{\mathcal{G}}^4 + \|\hat{g} - g_0\|_{\mathcal{G}}^2 \cdot \|\hat{\Gamma} - \Gamma_0\|_{\text{Fro}}^2}_{\text{improvement over } \|\hat{g} - g_0\|_{\mathcal{G}}^2} + \eta\Big).$$

When we have the true orthogonalizing Γ_0, the improved rate recovers the nuisance insensitive one from Thm. 1. Besides, when $\hat{g}$ converges but $\|\hat{\Gamma} - \Gamma_0\|_{\mathcal{G}}^2 = \mathcal{O}_p(1)$, the improved rate resembles the nuisance sensitive rate of Thm. 1, plus a $\mathcal{O}(\eta)$ bias term. Thus, the quality of the estimated orthogonalizing operator governs how the optimization interpolates between these two rates.

Having understood the performance of SGD when using an estimated orthogonalizing operator, one question is how to compute or approximate such an operator. Luedtke [2024] recently demonstrated that an orthogonalizing operator can be derived using algorithmic/reverse mode functional differentiation in many interesting cases. This can also be effective in our stochastic setting. In Sec. 3, using least-squares regression as an illustration, we developed a control variate [Johnstone and Velleman, 1985] interpretation of the variance reduction. This viewpoint offers another venue to develop approximate orthogonalizing operators.

Conclusion. We established non-asymptotic convergence guarantees for SGD algorithms under nuisances. We showed how the Neyman orthogonality of the loss function can mitigate the sensitivity of SGD algorithms to the effect of nuisances, and obtained results that align with recent ones from the DML/OSL literature in the batch setting. We also presented an iteratively orthogonalized SGD algorithm, whose convergence rate aligns with the rate in the nuisance insensitive regime. Extensions to hypothesis testing and reinforcement learning are interesting venues for future work.

Acknowledgments. The authors would like to thank L. Liu, V. Roulet, and J. Wellner for valuable comments and suggestions. This work was supported by NSF DMS-2023166, DMS-2134012, DMS-2210216, DMS-2502281, CCF-2019844, NIH, and IARPA 2022-22072200003. Part of this work was performed while R. Mehta and Z. Harchaoui were visiting the Simons Institute for the Theory of Computing, and A. Luedtke was visiting the Institute of Statistical Mathematics, and with the University of Washington.

Broader Impact. This work lies at the intersection of machine learning, mathematical optimization, learning theory, and mathematical statistics. While there are many applications of stochastic gradient optimization for practitioners, this particular work is of a theoretical nature and does not have any immediate positive or negative societal impact.

References

S. Amari. Backpropagation and Stochastic Gradient Descent Method. *Neurocomputing*, 1993.

F. Bach and E. Moulines. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. In *NeurIPS*, 2011.

P. J. Bickel, C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press Baltimore, 1993.

M. Bonvini and E. H. Kennedy. Fast Convergence Rates for Dose-Response Estimation, 2022.

L. Bottou and O. Bousquet. The Tradeoffs of Large Scale Learning. In *NeurIPS*, 2007.

L. Bottou and Y. Le Cun. On-Line Learning for Very Large Data Sets. *Applied Stochastic Models in Business and Industry*, 2005.

S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

M. Carone, A. R. Luedtke, and M. J. van Der Laan. Toward Computerized Efficient Estimation in Infinite-Dimensional Models. *Journal of the American Statistical Association*, 2019.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, 2018a.

V. Chernozhukov, D. N. Nekipelov, V. Semenova, and V. Syrgkanis. Plug-In Regularized Estimation of High-Dimensional Parameters in Nonlinear Semiparametric Models. Technical report, Cemmap Working Paper, 2018b.

V. Chernozhukov, W. K. Newey, and R. Singh. Automatic Debiased Machine Learning of Causal and Structural Effects. *Econometrica*, 2022.

V. Chernozhukov, W. K. Newey, V. Quintas-Martinez, and V. Syrgkanis. Automatic Debiased Machine Learning via Riesz Regression, 2024.

J. Clore, K. Cios, J. DeShazo, and B. Strack. Diabetes 130-US Hospitals for Years 1999-2008. UCI Machine Learning Repository, 2014.

J. Cutler, D. Drusvyatskiy, and Z. Harchaoui. Stochastic Optimization under Distributional Drift. *Journal of Machine Learning Research*, 2023.

A. Défossez and F. Bach. Averaged Least-Mean-Squares: Bias-Variance Trade-Offs and Optimal Sampling Distributions. In *AISTATS*, 2015.

A. Défossez, L. Bottou, F. Bach, and N. Usunier. A Simple Convergence Proof of Adam and Adagrad. *A Simple Convergence Proof of Adam and Adagrad*, 2022.

Y. Demidovich, G. Malinovsky, I. Sokolov, and P. Richtárik. A Guide Through the Zoo of Biased SGD. In *NeurIPS*, 2023.

A. Dieuleveut and F. Bach. Nonparametric Stochastic Approximation with Large Step-Sizes. *The Annals of Statistics*, 2016.

T. S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach*. Academic press, 2014.

D. J. Foster and V. Syrgkanis. Orthogonal Statistical Learning. *The Annals of Statistics*, 2023.

E. Gorbunov, F. Hanzely, and P. Richtarik. A Unified Theory of SGD: Variance Reduction, Sampling, Quantization and Coordinate Descent. In *AISTATS*, 2020.

R. M. Gower, M. Schmidt, F. Bach, and P. Richtárik. Variance-Reduced Methods for Machine Learning. *Proceedings of the IEEE*, 2020.

C. Graham and D. Talay. *Stochastic Simulation and Monte Carlo Methods*. Springer Berlin, Heidelberg, 2013.

M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo. Why Random Reshuffling Beats Stochastic Gradient Descent. *Mathematical Programming*, 2021.

Y. Hao and A. Orlitsky. The Broad Optimality of Profile Maximum Likelihood. In *NeurIPS*, 2019.

H. Ichimura and W. K. Newey. The Influence Function of Semiparametric Estimators. *Quantitative Economics*, 2022.

I. M. Johnstone and P. F. Velleman. Efficient Scores, Variance Decompositions, and Monte Carlo Swindles. *Journal of the American Statistical Association*, 1985.

M. Jordan, Y. Wang, and A. Zhou. Empirical Gateaux Derivatives for Causal Inference. In *NeurIPS*, 2022.

E. H. Kennedy. Towards Optimal Doubly Robust Estimation of Heterogeneous Causal Effects. *Electronic Journal of Statistics*, 2023.

M. J. Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer, 2003.

B. Y. Levit. Infinite-Dimensional Informational Inequalities. *Theory of Probability & Its Applications*, 1979.

D. Levy, Y. Carmon, J. Duchi, and A. Sidford. Large-Scale Methods for Distributionally Robust Optimization. In *NeurIPS*, 2020.

X. Li, M. Liu, and F. Orabona. On the Last Iterate Convergence of Momentum Methods. In *ALT*, 2022.

J. V. Linnik. *Statistical Problems with Nuisance Parameters*. American Mathematical Society, 2008.

L. Liu, C. Cinelli, and Z. Harchaoui. Orthogonal Statistical Learning with Self-Concordant Loss. In *COLT*, 2022.

A. Luedtke. Simplifying Debiased Inference via Automatic Differentiation and Probabilistic Programming, 2024.

A. Luedtke and I. Chung. One-Step Estimation of Differentiable Hilbert-Valued Parameters. *The Annals of Statistics*, 2024.

R. Mehta, V. Roulet, K. Pillutla, L. Liu, and Z. Harchaoui. Stochastic Optimization for Spectral Risk Measures. In *AISTATS*, 2023.

R. Mehta, V. Roulet, K. Pillutla, and Z. Harchaoui. Distributionally Robust Optimization with Bias and Variance Reduction. In *ICLR*, 2024.

S. A. Murphy and A. W. V. D. V. and. On Profile Likelihood. *Journal of the American Statistical Association*, 2000.

W. K. Newey. The Asymptotic Variance of Semiparametric Estimators. *Econometrica*, 1994.

J. Neyman. Optimal Asymptotic Tests for Composite Hypotheses. *Probability and Statistics*, 1959.

J. Neyman. C (α) Tests and Their Use. *Sankhyā: The Indian Journal of Statistics, Series A*, 1979.

X. Nie and S. Wager. Quasi-Oracle Estimation of Heterogeneous Treatment Effects. *Biometrika*, 2021.

D. S. Pavlichin, J. Jiao, and T. Weissman. Approximate Profile Maximum Likelihood. *Journal of Machine Learning Research*, 2019.

J. Pfanzagl. *Asymptotic Expansions for General Statistical Models*. Springer-Verlag Berlin Heidelberg, 1985.

A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. In *NeurIPS*, 2007.

J. Robins, L. Li, E. Tchetgen, A. van der Vaart, et al. Higher Order Influence Functions and Minimax Estimation of Nonlinear Functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*. Institute of Mathematical Statistics, 2008.

J. M. Robins and A. Rotnitzky. Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association*, 1995.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, 1994.

P. M. Robinson. Root-N-Consistent Semiparametric Regression. *Econometrica*, 1988.

A. Rotnitzky, E. Smucler, and J. M. Robins. Characterization of Parameters with a Mixed Bias Property. *Biometrika*, 2021.

D. B. Rubin. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 1974.

M. Schmidt, N. Le Roux, and F. Bach. Minimizing Finite Sums with the Stochastic Average Gradient. *Mathematical Programming*, 2017.

A. Sclocchi and M. Wyart. On the Different Regimes of Stochastic Gradient Descent. *PNAS*, 2024.

A. Shapiro. Distributionally Robust Stochastic Programming. *SIAM Journal on Optimization*, 2017.

C. Shi, D. Blei, and V. Veitch. Adapting Neural Networks for the Estimation of Treatment Effects. In *NeurIPS*, 2019.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*, 2021.

D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The Implicit Bias of Gradient Descent on Separable Data. *Journal of Machine Learning Research*, 2018.

A. A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006.

L. van der Laan, M. Carone, and A. Luedtke. Combining T-learning and DR-Learning: A Framework for Oracle-Efficient Estimation of Causal Contrasts, 2024.

L. van der Laan, A. Bibaut, N. Kallus, and A. Luedtke. Automatic Debiased Machine Learning for Smooth Functionals of Nonparametric M-Estimands, 2025.

M. J. van der Laan and A. R. Luedtke. Targeted Learning of an Optimal Dynamic Treatment, and Statistical Inference for Its Mean Outcome. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 2014.

M. J. van der Laan, S. Rose, et al. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

S. Vaswani, A. Mishkin, I. Laradji, M. Schmidt, G. Gidel, and S. Lacoste-Julien. Painless Stochastic Gradient: Interpolation, Line-Search, and Convergence Rates. In *NeurIPS*, 2019.

R. Ward, X. Wu, and L. Bottou. Adagrad Stepsizes: Sharp Convergence over Nonconvex Landscapes. *Journal of Machine Learning Research*, 2020.

J. M. Wooldridge. Specification Testing and Quasi-Maximum-Likelihood Estimation. *Journal of Econometrics*, 1991.

I. Zadik, L. Mackey, and V. Syrgkanis. Orthogonal Machine Learning: Power and Limitations. In *ICML*, 2018.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Our claims are theoretical convergence guarantees for various optimization algorithms. The results are included in Sec. 3 and the proofs are written in the appendix.

   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: A detailed discussion including limitations, relationships to other work, and future work is contained in Sec. 4.

   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: Assumptions are given before the statement of each result and proofs are contained 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: While this work is primarily theoretical, we provide code that reproduces our numerical illustrations.

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 code is provided in github.

   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: Experiments are included in Appx. G.

   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: [NA]

   Justification: Experiments of this nature are not included in our paper.

   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: [NA]

   Justification: Our numerical illustration is not computationally prohibitive, and can run on an instance of Google Colab.

   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: There are no ethical violations, to the authors' knowledge.

   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: [Yes]

    Justification: A "Broader Impact" statement is included before the references.

    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: We do not provide any models or datasets in this paper.

    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: We do not use any third party data/models that may incur licensing issues.

    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: [Yes]

    Justification: Our code is documented in notebook format.

    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: Experiments of this nature are not included in this paper.

    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: Experiments of this nature are not included in this paper.

    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: No substantive part of this research involved the use of large language models.

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.

Appendix

Table of Contents

A Notation

Symbol	Description
$\Theta \subseteq \mathbb{R}^d$	Finite-dimensional parameter class.
$(\mathcal{G}, \|\cdot\|_{\mathcal{G}})$	Possibly infinite-dimensional nuisance space.
$(\mathcal{G}, \langle\cdot,\cdot\rangle_{\mathcal{G}})$	The nuisance space as a Hilbert space.
$\mathbb{P}$	The unknown distribution of interest.
$Z \in \mathcal{Z}$	The random variable under $\mathbb{P}$.
$\theta_\star$	The target of interest.
g_0	The true nuisance parameter.
$\ell(\theta, g; z)$	The prespecified loss function.
$L(\theta, g)$	The population loss $\mathbb{E}_{Z\sim\mathbb{P}}[\ell(\theta, g; Z)]$.
$S_\theta(\theta, g; z)$	The score function $\nabla_\theta \ell(\theta, g; z)$.
$S_\theta(\theta, g)$	The population score $\mathbb{E}_{Z\sim\mathbb{P}}[\nabla_\theta \ell(\theta, g; z)]$.
$S_{\text{no}}(\theta, g; z)$	The Neyman orthogonalized score.
$S_{\text{no}}(\theta, g)$	The population Neyman orthogonalized score $\mathbb{E}_{Z\sim\mathbb{P}}[S_{\text{no}}(\theta, g; Z)]$.
$(\nabla_\theta, \nabla_g)$	The gradient *w.r.t.* θ and g
$(\mathrm{D}_\theta, \mathrm{D}_g)$	The derivative operator *w.r.t.* θ and g.
$H_{\theta g}$	The transposed Jacobian defined by $\nabla_g S_\theta(\theta_\star, g_0) \in \mathcal{G}^d$
H_{gg}	The nuisance Hessian operator defined by $\nabla_g^2 L(\theta_\star, g_0)$
Γ_0	Linear operator defined by $[\Gamma_0 g]_j = \langle H_{\theta g}^{(j)}, H_{gg}^{-1} g\rangle_{\mathcal{G}}$.
μ	The strong convexity constant of L.
M	The smoothness constant of L.
(K_1, κ_1)	Constants to bound the second moment of $S_\theta(\theta, g; Z)$.
(α_1, α_2)	The second order smoothness constant of L.
β_1	The higher order smoothness constant of a Neyman orthogonal L.
μ_{no}	The strong convexity constant of $\nabla_\theta S_{\text{no}}(\theta_\star, g_0)$.
M_{no}	The smoothness constant of $\nabla_\theta S_{\text{no}}(\theta_\star, g_0)$.
(K_2, κ_2)	Constants to bound the second moment of $S_{\text{no}}(\theta, g; Z)$.
β_2	The higher order smoothness constant of S_{no}.
η	The learning rate of stochastic optimization.
n	The iteration of stochastic gradient.
m	The iteration of nuisance estimation.

Table 2: Notation used throughout the paper.

B Detailed Examples

In this section, we describe in detail how the three examples in Sec. 2 from the main text satisfy Asm. 3 and Asm. 4. We first talk about the partially linear model (PLM) in Appx. B.1, and then introduce the conditional averaged treatment effect (CATE) based on the potential outcomes framework in Appx. B.1. Under the same framework, finally we talk about the conditional relative risk (CRR) in Appx. B.3. In addition, we also study a non-orthogonal loss usually used for PLM in Appx. B.1.2 and an unrestricted loss function for CATE in Appx. B.2.1. The constants for all examples are concluded in Tab. 3 and proofs of lemmas in this section are provided in Appx. B.4.

B.1 Partially Linear Model

B.1.1 Orthogonal Loss

We revisit Example 1 from the main text where we consider the target of interest as a solution of a partially linear model. Let $Z = (X, Y, W)$, where X is an $\mathbb{R}^d$-valued input, Y is a real-valued outcome, and W is a $\mathcal{W}$-valued control or confounder. The space $\mathcal{G}$ is a nonparametric class containing functions of the form

$$g = (g_Y, g_X) : \mathcal{W} \to \mathbb{R} \times \mathbb{R}^d.$$

Following the construction of Robinson [1988], this g is supplied to the loss

$$\ell(\theta, g; z) = \frac{1}{2}[y - g_Y(w) - \langle \theta, x - g_X(w) \rangle]^2. \tag{15}$$

To ensure $\theta_\star$ can be interpreted via the projection of $\mathbb{E}_\mathbb{P}[Y|X, W]$ onto partially linear additive functions, the true nuisance is given by $g_0 = (g_{0,X}, g_{0,Y})$, where

$$g_{0,Y}(w) := \mathbb{E}_\mathbb{P}\left[Y \mid W = w\right] \text{ and } g_{0,X}(w) := \mathbb{E}_\mathbb{P}\left[X \mid W = w\right].$$

We define the residual ϵ at $(\theta_\star, g_0)$ as

$$\epsilon = Y - g_{0,Y}(W) - \langle \theta_\star, X - g_{0,X}(W) \rangle.$$

Lemma 4. *Let $\tilde{Y} = Y - g_{0,Y}(w)$ and $\tilde{X} = X - g_{0,X}(w)$. We assume the following conditions:*

(a) $\lambda_{\min}(\mathbb{E}_\mathbb{P}[\tilde{X}\tilde{X}^\top]) \geq \lambda_0$ for some constant $\lambda_0 > 0$.

(b) $\|\tilde{X}\|_2 \leq C_X$ a.s. and $\mathbb{E}_\mathbb{P}\left[\epsilon^4\right] \leq \sigma^4$ for some constants $C_X, \sigma > 0$.

Then Asm. 3 and Asm. 4 are satisfied. The target $\theta_\star$ is the minimizer of the squared loss:

$$\theta_\star = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}_\mathbb{P}[(\tilde{Y} - \tilde{X}^\top \theta)^2].$$

The proof of Lem. 4 is provided in Appx. B.4.1.

B.1.2 Non-orthogonal Loss

Suppose that the outcome Y is generated under the partially linear model:

$$Y = \langle \theta_0, X \rangle + g_0(W) + \epsilon, \tag{16}$$

where $\theta_0 \in \mathbb{R}^d$ is the true parameter, $g_0 : \mathcal{W} \mapsto \mathbb{R}$ is the true nuisance function and $\mathbb{E}_\mathbb{P}\left[\epsilon \mid X, W\right] = 0$. The space $\mathcal{G}$ is a nonparametric class containing functions of the form

$$g : \mathcal{W} \mapsto \mathbb{R}.$$

We can also consider the following non-orthogonal squared loss function:

$$\ell(\theta, g; z) = \frac{1}{2}[y - g(w) - \langle \theta, x \rangle]^2. \tag{17}$$

Example	μ	M	K_1	κ_1	α_1	β_1
(1) Orthogonal PLM	λ_0	$C_X^2 + r^2$	$18C_X^2\sigma^2 + \mathcal{O}(r^2)$	$18C_X^4 + \mathcal{O}(r^2)$	$2(1 + \\|\theta_\star\\|_2)r$	$2(1 + \\|\theta_\star\\|_2)$
(2) Non-Orthogonal PLM	λ_0	C_X^2	$6C_X^2(\sigma^2 + 2r^2)$	$2C_X^4$	C_X	–
(3) Unrestricted CATE	$c_0^2\lambda_0$	$C_X^2(1 + r^2)$	$12C_X^2(\sigma^2 + 4(C_X\\|\theta_\star\\|_2 + C_\tau)^2) + \mathcal{O}(r^2)$	$27C_X^4 + \mathcal{O}(r^2)$	$\mathcal{O}(r)$	$C_X(1 + 4(C_X\\|\theta_\star\\|_2 + C_\tau))$
(4) Restricted CATE	λ_0	C_X^2	$27C_X^2(\sigma^2 + 2C^2) + \mathcal{O}(r^2)$	$3C_X^4$	$2C_X(2r + 3)r$	$4c_0^{-2}C_X(1 + r)$
(5) CRR	$C^2\delta\lambda_0$	$C_X^2(1 + 2\delta^{-1}r^2)$	$24C_X^2(\delta^{-2} + 4c_0^{-1}C_Y^2) + \mathcal{O}(r^2)$	$6C_X^4(1 + 6(\delta^{-2} + 4c_0^{-1}C_Y^2)) + \mathcal{O}(r^2)$	$2C_X(c_0^{-1} + 1)r$	$4c_0^{-2}C_X(1 + r)$

Table 3: Constants for All Examples.

We define the residual ϵ at $(\theta_\star, g_0)$ as

$$\epsilon = Y - g_0(W) - \langle \theta_\star, X \rangle.$$

Lemma 5. *We assume the following conditions:*

(a) $\lambda_{\min}(\mathbb{E}_{\mathbb{P}}\left[XX^\top\right]) \geq \lambda_0$ *for some constant* $\lambda_0 > 0$.

(b) $\|X\|_\infty \leq C_X$ *a.s. and* $\mathbb{E}_{\mathbb{P}}\left[\epsilon^2\right] \leq \sigma^2$ *for some constants* $C_X, \sigma > 0$.

Then Asm. 3 is satisfied and the target $\theta_\star$ *is the true parameter, i.e.,* $\theta_\star = \theta_0$.

The proof of Lem. 5 is provided in Appx. B.4.2.

Orthogonalization. We can perform our orthogonalization method to obtain the Neyman orthogonal gradient oracle for this non-orthogonal loss. For any $h_1, h_2 \in \mathcal{G}$, we define the inner product of $\mathcal{G}$ as

$$\langle h_1, h_2 \rangle_{\mathcal{G}} = \mathbb{E}_{\mathbb{P}}\left[h_1(W)h_2(W)\right]. \tag{18}$$

For any $(\theta, g, z) \in \Theta \times G \times \mathcal{Z}$ By Definition 1 the derivative of non-orthogonal loss (17) along the direction of h_1 is given by

$$\mathrm{D}_g \ell(\theta, g; z)[h_1] = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}[y - (g + th_1)(w) - \langle \theta, x \rangle]^2\right) = -(y - g(w) - \langle \theta, x \rangle)h_1(w). \tag{19}$$

Do derivative on $\mathrm{D}_g \ell(\theta, g; z)[h_1]$ along the direction of h_2 and we have

$$\mathrm{D}_g^2 \ell(\theta, g; z)[h_1, h_2] = \frac{\mathrm{d}}{\mathrm{d}t}\left(-(y - (g + th_2)(w) - \langle \theta, x \rangle)h_1(w)\right) = h_1(w)h_2(w), \tag{20}$$

which implies

$$\mathrm{D}_g^2 L(\theta_\star, g_0)[h_1, h_2] = \mathbb{E}_{\mathbb{P}}\left[\mathrm{D}_g^2 \ell(\theta_\star, g_0; Z)[h_1, h_2]\right] = \mathbb{E}_{\mathbb{P}}\left[h_1(W)h_2(W)\right].$$

By the definition in (84), we have $H_{gg} = \mathbf{I}$ the identity operator. In addition, do derivative on the score along the direction of $h \in \mathcal{G}$ and we have

$$\mathrm{D}_g S_\theta(\theta, g; z)[h] = \frac{\mathrm{d}}{\mathrm{d}t}\left(-(y - (g + th)(w) - \langle \theta, x \rangle)x\right) = h(w)x,$$

which implies that

$$\mathrm{D}_g S_\theta(\theta, g)[h] = \mathbb{E}_{\mathbb{P}}\left[S_\theta(\theta, g; Z)[h]\right] = \mathbb{E}_{\mathbb{P}}\left[h(W)\mathbb{E}_{\mathbb{P}}\left[X \mid W\right]\right].$$

By the definition in (83), we have $H_{\theta g} = \mathbb{E}_{\mathbb{P}}\left[X \mid W\right]$. Thus, by (85) we have

$$\Gamma_0 : g \mapsto \langle \mathbb{E}_{\mathbb{P}}\left[X \mid W\right], g \rangle_{\mathcal{G}} = \mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\mathbb{P}}\left[X \mid W\right] g(W)\right]. \tag{21}$$

Thus, the Neyman orthogonalized gradient oracle defined in (86) is given by

$$\begin{aligned} S_{\text{no}}(\theta, g; z) &= S_\theta(\theta, g; z) - \mathrm{D}_g \ell(\theta, g; z)[\mathbb{E}_{\mathbb{P}}\left[X \mid W = w\right]] \\ &= -(y - g(w) - \langle \theta, x \rangle)(x - \mathbb{E}_{\mathbb{P}}\left[X \mid W = w\right]). \end{aligned} \tag{22}$$

Lemma 6. *Consider the bounded linear operator* $\hat{\Gamma} : \mathcal{G} \mapsto \mathbb{R}^d$ *such that* $[\hat{\Gamma} g]_j = \langle \hat{\gamma}^{(j)}, g \rangle_{\mathcal{G}}, \forall g \in \mathcal{G}$ *for some* $\hat{\gamma}^{(j)} \in \mathcal{G}, j = 1, \ldots, d$. *Let* $\tilde{Y} = Y - g_{0,Y}(w)$ *and* $\tilde{X} = X - g_{0,X}(w)$. *We assume the following conditions:*

(a) $\lambda_{\min}(\mathbb{E}_{\mathbb{P}}[\tilde{X}\tilde{X}^\top]) \geq \lambda_0$ *for some constant* $\lambda_0 > 0$.

(b) $\|\tilde{X}\|_2 \leq C_X$ *a.s. and* $\mathbb{E}_{\mathbb{P}}\left[\epsilon^4\right] \leq \sigma^4$ *for some constants* $C_X, \sigma > 0$.

Then Asm. 6 is satisfied.

The proof of Lem. 6 is provided in Appx. B.4.3.

B.2 Conditional Averaged Treatment Effect

We now introduce examples in causal inference which are established based on the potential outcomes framework. The potential outcomes framework [Rubin, 1974] has been widely used in causal inference. Let $Z = (W, X, Y) \in \{0, 1\} \times \mathbb{R}^d \times \mathbb{R}$ under some distribution $\mathbb{P}$. We posit the existence of potential outcomes $Y(1), Y(0) \in \mathbb{R}$. The conditional averaged treatment effect (CATE) is then defined as

$$\tau_0(x) = \mathbb{E}_{\mathbb{P}}[Y(1) - Y(0) \mid X = x].$$

To identify $\tau_0(x)$ and the following causal assumptions are required:

Assumption 5. *The following conditions hold:*

(a) (consistency) $Y = Y(W)$.

(b) (unconfoundedness) $Y(w) \perp W \mid X$ *for all* $w \in \{0, 1\}$.

(c) (positive overlap) $c_0 \leq \mathbb{P}(W = 1 \mid X) \leq 1 - c_0$ *a.s. for some* $c_0 > 0$.

Under Asm. 5, τ_0 can be identified by observed data since

$$\begin{aligned}\tau_0(x) &= \mathbb{E}_{\mathbb{P}}[Y(1) - Y(0) \mid X = x] \\ &= \mathbb{E}_{\mathbb{P}}[Y(1) \mid W = 1, X = x] - \mathbb{E}_{\mathbb{P}}[Y(0) \mid W = 0, X = x] \\ &= \mathbb{E}_{\mathbb{P}}[Y \mid W = 1, X = x] - \mathbb{E}_{\mathbb{P}}[Y \mid W = 0, X = x].\end{aligned}$$

B.2.1 Unrestricted Nuisance

We observe $Z = (X, Y, W)$, where W is a binary treatment assignment. The functions in $\mathcal{G}$ are of the form

$$g = (g^{\text{out}}, g^{\text{prop}}) : \mathbb{R}^d \to \mathbb{R} \times \mathbb{R},$$

and are evaluated (see Nie and Wager [2021, Eq. (2)]) at the loss

$$\ell(\theta, g; z) = \frac{1}{2}\left(y - g^{\text{out}}(x) - (w - g^{\text{prop}}(x))\langle \theta, x \rangle\right)^2. \tag{23}$$

For $g_0 = (g_0^{\text{out}}, g_0^{\text{prop}})$ nuisance functions g_0^{out} and g_0^{prop} represent the outcome regression and the propensity score, respectively:

$$g_0^{\text{out}}(x) := \mathbb{E}_{\mathbb{P}}[Y \mid X = x] \text{ and } g_0^{\text{prop}}(x) := \mathbb{E}_{\mathbb{P}}[W \mid X = x].$$

We define the residual ϵ under the true model as

$$\epsilon = Y - g_0^{\text{out}}(X) - \left(W - g_0^{\text{prop}}(X)\right)\tau_0(X).$$

Lemma 7. *We assume Asm. 5 and the following conditions hold:*

(a) $\lambda_{min}(\mathbb{E}_{\mathbb{P}}\left[XX^\top\right]) \geq \lambda_0$ *for some constant* $\lambda_0 > 0$.

(b) $\|X\|_2 \leq C_X$ *and* $|\tau_0(X)| \leq C_\tau$ *a.s. and* $\mathbb{E}_{\mathbb{P}}\left[\epsilon^4\right] \leq \sigma^4$ *for some constants* $C_X, C_\tau, \sigma > 0$.

Then Asm. 3 and Asm. 4 are satisfied. The target $\theta_\star$ *is the minimizer of the squared loss:*

$$\theta_\star = \underset{\theta \in \mathbb{R}^d}{\arg\min}\, \mathbb{E}_{\mathbb{P}}\left[(W - g_0^{prop}(X))^2(\tau_0(X) - X^\top\theta)^2\right].$$

The proof of Lem. 7 is provided in Appx. B.4.4.

B.2.2 Restricted Nuisance

We observe $Z = (X, Y, W)$, where W is a binary treatment assignment. Here we restrict the propensity model as $g^{\text{prop}} : \mathbb{R}^d \mapsto (0, 1)$. The functions in $\mathcal{G}$ are of the form

$$g = (g^{(0)}, g^{(1)}, g^{\text{prop}}) : \mathbb{R}^d \to \mathbb{R} \times \mathbb{R} \times (0, 1),$$

and are evaluated (see van der Laan and Luedtke [2014, Thm. 1]) at the loss

$$\ell(\theta, g; z) = \frac{1}{2}\left(g^{(1)}(x) - g^{(0)}(x) + \frac{w - g^{\text{prop}}(x)}{g^{\text{prop}}(x)(1 - g^{\text{prop}}(x))}(y - g^{(w)}(x)) - \langle \theta, x \rangle\right)^2. \tag{24}$$

This loss also appears in Foster and Syrgkanis [2023, Eq. (23)]. For $g_0 = (g_0^{(0)}, g_0^{(1)}, g_0^{\text{prop}})$ nuisance functions $g_0^{(0)}$ and $g_0^{(1)}$ represent the outcome regressions

$$g_0^{(0)}(x) := \mathbb{E}_{\mathbb{P}}[Y \mid W = 1, X = x] \text{ and } g_0^{(1)}(x) := \mathbb{E}_{\mathbb{P}}[Y \mid W = 0, X = x].$$

We define the residual ϵ as

$$\epsilon = \frac{W - g_0^{\text{prop}}(X)}{g_0^{\text{prop}}(X)(1 - g_0^{\text{prop}}(X))}(Y - g_0^{(W)}(X)).$$

Lemma 8. *We assume Asm. 5 and the following conditions hold:*

(a) $\lambda_{min}(\mathbb{E}_{\mathbb{P}}\left[XX^\top\right]) \geq \lambda_0$ *for some constant* $\lambda_0 > 0$.

(b) $\mathbb{E}_{\mathbb{P}}\left[\epsilon^2\right] \leq \sigma^2$, $\|X\|_2 \leq C_X$, *and* $\left|Y - g_0^{(w)}(X)\right| \leq C_Y, w = 0, 1$ *a.s. for some constants* $\sigma, C_X, C_Y > 0$.

Then Asm. 3 and Asm. 4 are satisfied. The target $\theta_\star$ *is the minimizer of the squared loss:*

$$\theta_\star = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{\mathbb{P}}\left[(\tau_0(X) - X^\top \theta)^2\right].$$

The proof of Lem. 8 is provided in Appx. B.4.5.

B.3 Conditional Relative Risk

We retain all components of the previous example, changing only the loss and assuming that the outcome Y is binary/non-negative. First, consider the "label" function

$$\mu_g^{(s)}(z) = g^{(s)}(x) + \frac{\mathbb{1}(w = s)}{s g^{\text{prop}}(x) + (1 - s)(1 - g^{\text{prop}}(x))}(y - g^{(s)}(x)),$$

where $\mathbb{1}(\cdot)$ denotes the indicator function, and the log-linear predictor $p_\theta(x) = e^{\langle \theta, x \rangle}/(1 + e^{\langle \theta, x \rangle})$. Following Example 2 in van der Laan et al. [2024], we then employ the cross entropy-type loss function

$$\ell(\theta, g; z) = -\left[\mu_g^{(1)}(z) \log p_\theta(x) + \mu_g^{(0)}(z) \log(1 - p_\theta(x))\right]. \tag{25}$$

Lemma 9. *We assume the following conditions:*

(a) $\lambda_{min}(\mathbb{E}_{\mathbb{P}}\left[XX^\top\right]) \geq \lambda_0$ *for some constant* $\lambda_0 > 0$.

(b) $\|X\|_2 \leq C_X$ *and* $Y(w) - g_0^{(w)}(X) \leq C_Y, w = 0, 1$ *a.s. for some constants* $C_X, C_Y > 0$.

(c) $\delta \leq g_0^{(0)}(X) + g_0^{(1)}(X) \leq \delta^{-1}$ *a.s. for some constant* $\delta > 0$.

Then Asm. 3 and Asm. 4 are satisfied. The target $\theta_\star$ *is the minimizer of the weighted cross entropy loss:*

$$\theta_\star = \arg\min_{\theta \in \mathbb{R}^d} -\mathbb{E}_{\mathbb{P}}\left[g_0^{(1)}(X) \log p_\theta(X) + g_0^{(0)}(X) \log(1 - p_\theta(X))\right].$$

The proof of Lem. 9 is provided in Appx. B.4.6.

B.4 Proofs

B.4.1 Proof of Lemma 4

Proof. We consider the following loss:

$$\ell(\theta, g; z) = \frac{1}{2}(y - g_Y(w) - \langle \theta, x - g_X(w)\rangle)^2,$$

with the corresponding risk function defined as

$$L(\theta, g) = \frac{1}{2}\mathbb{E}_{\mathbb{P}}\left[(Y - g_Y(W) - \langle \theta, X - g_X(W)\rangle)^2\right].$$

Let $\tilde{Y} = Y - g_{0,Y}(w)$ and $\tilde{X} = X - g_{0,X}(w)$. By definition, the target $\theta_\star$ is the minimizer of the squared loss:

$$\theta_\star = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{\mathbb{P}}\left[(\tilde{Y} - \tilde{X}^\top \theta)^2\right] = \mathbb{E}_{\mathbb{P}}\left[\tilde{X}\tilde{X}^\top\right]^{-1} \mathbb{E}_{\mathbb{P}}\left[\tilde{Y}\tilde{X}\right]. \tag{26}$$

Differentiating $\ell(\theta, g; z)$ with respect to θ, we obtain the gradient and Hessian *w.r.t.* θ as

$$\begin{aligned} S_\theta(\theta, g; z) &= -(y - g_Y(w) - \langle \theta, x - g_X(w)\rangle)(x - g_X(w)), \\ H_{\theta\theta}(\theta, g; z) &= (x - g_X(w))(x - g_X(w))^\top. \end{aligned}$$

The expected gradient and expected Hessian are then obtained as

$$\begin{aligned} S_\theta(\theta, g) &= -\mathbb{E}_{\mathbb{P}}\left[(Y - g_Y(W) - \langle \theta, X - g_X(W)\rangle)(X - g_X(W))\right], \\ H_{\theta\theta}(\theta, g) &= \mathbb{E}_{\mathbb{P}}\left[(X - g_X(W))(X - g_X(W))^\top\right]. \end{aligned}$$

We consider the nuisance neighborhood such that for $g \in \mathcal{G}_r\,(g_0)$,

$$\|g - g_0\|_{\mathcal{G}} := \max\left\{\mathbb{E}_{\mathbb{P}}\left[\|g_X(W) - g_{0,X}(W)\|_2^4\right]^{\frac{1}{4}}, \mathbb{E}_{\mathbb{P}}\left[(g_Y(W) - g_{0,Y}(W))^4\right]^{\frac{1}{4}}\right\} \le r. \tag{27}$$

We now verify that the loss function ℓ satisfies Asm. 3.

(a) We assume that $g_X(w) : \mathcal{W} \mapsto \mathbb{R}^d$ and $g_Y(w) : \mathcal{W} \mapsto \mathbb{R}$ are continuous functions, thus Asm. 3(a) is satisfied.

(b) By (26), it follows from KKT conditions that

$$S_\theta(\theta_\star, g_0) = -\mathbb{E}_{\mathbb{P}}\left[(\tilde{Y} - \langle \theta_\star, \tilde{X}\rangle)\tilde{X}\right] = 0. \tag{28}$$

(c) Since $\mathbb{E}_{\mathbb{P}}[\tilde{X} \mid W] = 0$ and $\mathbb{E}_{\mathbb{P}}[\tilde{Y} \mid W] = 0$, we have

$$H_{\theta\theta}(\theta, g) = \mathbb{E}_{\mathbb{P}}\left[\tilde{X}\tilde{X}^\top\right] + \mathbb{E}_{\mathbb{P}}\left[(g_X(W) - g_{0,X}(W))(g_X(W) - g_{0,X}(W))^\top\right].$$

For any $g \in \mathcal{G}_r$, when $\lambda_{\min}(\mathbb{E}_{\mathbb{P}}[\tilde{X}\tilde{X}^\top]) \ge \lambda_0$ and $\|\tilde{X}\|_2 \le C_X$ *a.s.* , we have

$$\lambda_0 \mathbf{I} \preccurlyeq H_{\theta\theta}(\theta, g) \preccurlyeq (C_X^2 + r^2)\mathbf{I} \implies \mu = \lambda_0 \text{ and } M = C_X^2 + r^2. \tag{29}$$

(d) Consider the Taylor expansion around $\theta_\star$, we have

$$S_\theta(\theta, g; Z) - S_\theta(\theta, g) = S_\theta(\theta_\star, g; Z) - S_\theta(\theta_\star, g) + (H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g))(\theta - \theta_\star).$$

Let $\epsilon = \tilde{Y} - \langle \theta_\star, \tilde{X}\rangle$. Note that $X - g_X(W) = \tilde{X} - (g_X - g_{0,X})(W)$ and

$$Y - g_Y(W) - \langle \theta_\star, X - g_X(W)\rangle = \epsilon - (g_Y - g_{0,Y})(W) + \langle \theta_\star, (g_X - g_{0,X})(W)\rangle. \tag{30}$$

Since $\mathbb{E}_{\mathbb{P}}[\epsilon \mid W] = 0$, $\mathbb{E}_{\mathbb{P}}[\tilde{X} \mid W] = 0$ by definition and $\mathbb{E}_{\mathbb{P}}[\epsilon\tilde{X}] = 0$ by (28), then for any $g \in \mathcal{G}_r(g_0)$,

$$\begin{aligned}
&\|S_\theta(\theta_\star, g)\|_2 = \|\mathbb{E}_{\mathbb{P}}\left[(g_Y - g_{0,Y})(g_X - g_{0,X})(W) - \langle\theta_\star, (g_X - g_{0,X})(W)\rangle(g_X - g_{0,X})(W)\right]\|_2 \\
&\quad \le \left(\mathbb{E}_{\mathbb{P}}\left[((g_Y - g_{0,Y})(W))^2\right]\mathbb{E}_{\mathbb{P}}\left[\|(g_X - g_{0,X})(W)\|_2^2\right]\right)^{1/2} + \mathbb{E}_{\mathbb{P}}\left[\|(g_X - g_{0,X})(W)\|_2^2\right]\|\theta_\star\|_2 \\
&\quad \le r^2(1 + \|\theta_\star\|).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\|S_\theta(\theta_\star, g; Z)\|_2^2 &\le \left(\epsilon - (g_Y - g_{0,Y})(W) + \langle\theta_\star, (g_X - g_{0,X})(W)\rangle\right)^2 \|\tilde{X} - (g_X - g_{0,X})(W)\|_2^2 \\
&\le 3\left(\epsilon^2 + ((g_Y - g_{0,Y})(W))^2 + \|(g_X - g_{0,X})(W)\|_2^2\|\theta_\star\|_2^2\right)\left(C_X + \|(g_X - g_{0,X})(W)\|_2\right)^2 \\
&\le 6\left(\epsilon^2 + ((g_Y - g_{0,Y})(W))^2 + \|(g_X - g_{0,X})(W)\|_2^2\|\theta_\star\|_2^2\right)\left(C_X^2 + \|(g_X - g_{0,X})(W)\|_2^2\right),
\end{aligned}$$

which implies that for $g \in \mathcal{G}_r(g_0)$, when $\mathbb{E}_{\mathbb{P}}[\epsilon^4] \le \sigma^4$,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}}[\|S_\theta(\theta_\star, g; Z)\|_2^2] \\
&\quad \le 6C_X^2\left(\mathbb{E}_{\mathbb{P}}\left[\epsilon^2\right] + \mathbb{E}_{\mathbb{P}}\left[((g_Y - g_{0,Y})(W))^2\right] + \mathbb{E}_{\mathbb{P}}\left[\|(g_X - g_{0,X})(W)\|_2^2\right]\|\theta_\star\|_2^2\right) \\
&\qquad + 6\mathbb{E}_{\mathbb{P}}\left[\epsilon^2\|(g_X - g_{0,X})(W)\|_2^2\right] + 6\mathbb{E}_{\mathbb{P}}\left[((g_Y - g_{0,Y})(W))^2\|(g_X - g_{0,X})(W)\|_2^2\right] \\
&\qquad + 6\mathbb{E}_{\mathbb{P}}\left[\|(g_X - g_{0,X})(W)\|_2^4\right]\|\theta_\star\|_2^2 \\
&\quad \le 6C_X^2(\sigma^2 + r^2 + r^2\|\theta_\star\|_2^2) + 6r^4\|\theta_\star\|_2^2 + 6\left(\mathbb{E}_{\mathbb{P}}\left[\epsilon^4\right]\mathbb{E}_{\mathbb{P}}\left[\|(g_X - g_{0,X})(W)\|_2^4\right]\right)^{1/2} \\
&\qquad + 6\left(\mathbb{E}_{\mathbb{P}}\left[((g_Y - g_{0,Y})(W))^4\right]\mathbb{E}_{\mathbb{P}}\left[\|(g_X - g_{0,X})(W)\|_2^4\right]\right)^{1/2} \\
&\quad \le 6C_X^2\sigma^2 + 6\left\{\sigma^2 + C_X^2(1 + \|\theta_\star\|_2^2)\right\}r^2 + 6(1 + \|\theta_\star\|_2^2)r^4.
\end{aligned}$$

Thus, for any $g \in \mathcal{G}_r(g_0)$,

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}[\|S_\theta(\theta_\star, g; Z) - S_\theta(\theta_\star, g)\|_2^2] &\le 2\mathbb{E}_{\mathbb{P}}\left[\|S_\theta(\theta_\star, g; Z)\|_2^2\right] + 2\|S_\theta(\theta_\star, g)\|_2^2 \\
&\le 12C_X^2\sigma^2 + \left\{12\sigma^2 + 2(1 + \|\theta_\star\|_2) + 12C_X^2(1 + \|\theta_\star\|_2^2)\right\}r^2 + 12(1 + \|\theta_\star\|_2^2)r^4 \\
&= 12C_X^2\sigma^2 + \mathcal{O}(r^2).
\end{aligned}$$

On the other hand, since

$$\begin{aligned}
\|H_{\theta\theta}(\theta_\star, g; Z)\|_2 &= \|(\tilde{X} - (g_X - g_{0,X})(W))(\tilde{X} - (g_X - g_{0,X})(W))^\top\|_2 \\
&\le \|\tilde{X} - (g_X - g_{0,X})(W)\|_2^2 \\
&\le 2\|\tilde{X}\|_2^2 + 2\|(g_X - g_{0,X})(W)\|_2^2 \le 2C_X^2 + 2\|(g_X - g_{0,X})(W)\|_2^2,
\end{aligned}$$

by (29) we have

$$\begin{aligned}
\|H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g)\|_2 &\le \|H_{\theta\theta}(\theta_\star, g; Z)\|_2 + \|H_{\theta\theta}(\theta_\star, g)\|_2 \\
&\le 3C_X^2 + r^2 + 2\|(g_X - g_{0,X})(W)\|_2^2,
\end{aligned}$$

which implies that

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}[\|(H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g))(\theta - \theta_\star)\|_2^2] &\le \mathbb{E}_{\mathbb{P}}\left[(3C_X^2 + r^2 + 2\|(g_X - g_{0,X})(W)\|_2^2)^2\right]\|\theta - \theta_\star\|_2^2 \\
&= (9C_X^4 + \mathcal{O}(r^2))\|\theta - \theta_\star\|_2^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[\|S_\theta(\theta, g; Z) - S_\theta(\theta, g)\|_2^2\right] \le& 2\mathbb{E}_{\mathbb{P}}\left[\|S_\theta(\theta_\star, g; Z) - S_\theta(\theta_\star, g)\|_2^2\right] \\
&+ 2\mathbb{E}_{\mathbb{P}}\left[\|(H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g))(\theta - \theta_\star)\|_2^2\right] \\
\le& 24C_X^2\sigma^2 + \mathcal{O}(r^2) + (18C_X^4 + \mathcal{O}(r^2))\|\theta - \theta_\star\|_2^2,
\end{aligned}$$

which implies

$$K_1 = 24C_X^2\sigma^2 + \mathcal{O}(r^2) \text{ and } \kappa_1 = 18C_X^4 + \mathcal{O}(r^2). \tag{31}$$

(e) For any $\theta \in \Theta$ and $g, \bar{g} \in \mathcal{G}_r(g_0)$, by (30) we have

$$\begin{aligned}
&\mathrm{D}_g\mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0] \\
&\quad = \mathbb{E}_{\mathbb{P}}\left[(-(g_Y - g_{0,Y})(W) + \langle \theta_\star, (g_X - g_{0,X})(W)\rangle)\langle \theta - \theta_\star, (\bar{g}_X - g_{0,X})(W)\rangle\right] \\
&\qquad + \mathbb{E}_{\mathbb{P}}\left[(-(\bar{g}_Y - g_{0,Y})(W) + \langle \theta_\star, (\bar{g}_X - g_{0,X})(W)\rangle)\langle \theta - \theta_\star, (g_X - g_{0,X})(W)\rangle\right].
\end{aligned}$$

Since $\bar{g} \in \mathcal{G}_r(g_0)$,

$$\begin{aligned}
&|\mathbb{E}_{\mathbb{P}}\left[(-(g_Y - g_{0,Y})(W) + \langle \theta_\star, (g_X - g_{0,X})(W)\rangle)\langle \theta - \theta_\star, (\bar{g}_X - g_{0,X})(W)\rangle\right]| \\
&\leq \mathbb{E}_{\mathbb{P}}\left[|(g_Y - g_{0,Y})(W)|\, \|(\bar{g}_X - g_{0,X})(W)\|_2\right] \|\theta - \theta_\star\|_2 \\
&\quad + \|\theta_\star\|_2 \mathbb{E}_{\mathbb{P}}\left[\|(g_X - g_{0,X})(W)\|_2 \|(\bar{g}_X - g_{0,X})(W)\|_2\right] \|\theta - \theta_\star\|_2 \\
&\leq \mathbb{E}_{\mathbb{P}}\left[\|(\bar{g}_X - g_{0,X})(W)\|_2^2\right]^{1/2} \mathbb{E}_{\mathbb{P}}\left[((g_Y - g_{0,Y})(W))^2\right]^{1/2} \|\theta - \theta_\star\|_2 \\
&\quad + \|\theta_\star\|_2 \mathbb{E}_{\mathbb{P}}\left[\|(\bar{g}_X - g_{0,X})(W)\|_2^2\right]^{1/2} \mathbb{E}_{\mathbb{P}}\left[\|(g_X - g_{0,X})(W)\|_2^2\right]^{1/2} \|\theta - \theta_\star\|_2 \\
&\leq (1 + \|\theta_\star\|_2) r \|g - g_0\|_{\mathcal{G}} \|\theta - \theta_\star\|_2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&|\mathbb{E}_{\mathbb{P}}\left[(-(\bar{g}_Y - g_{0,Y})(W) + \langle \theta_\star, (\bar{g}_X - g_{0,X})(W)\rangle)\langle \theta - \theta_\star, (g_X - g_{0,X})(W)\rangle\right]| \\
&\leq (1 + \|\theta_\star\|_2) r \|g - g_0\|_{\mathcal{G}} \|\theta - \theta_\star\|_2.
\end{aligned}$$

Thus,

$$\begin{aligned}
&|\mathrm{D}_g\mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0]| \\
&\quad \leq |\mathbb{E}_{\mathbb{P}}\left[(-(g_Y - g_{0,Y})(W) + \langle \theta_\star, (g_X - g_{0,X})(W)\rangle)\langle \theta - \theta_\star, (\bar{g}_X - g_{0,X})(W)\rangle\right]| \\
&\qquad + |\mathbb{E}_{\mathbb{P}}\left[(-(\bar{g}_Y - g_{0,Y})(W) + \langle \theta_\star, (\bar{g}_X - g_{0,X})(W)\rangle)\langle \theta - \theta_\star, (g_X - g_{0,X})(W)\rangle\right]| \\
&\quad \leq 2(1 + \|\theta_\star\|_2) r \|g - g_0\|_{\mathcal{G}} \|\theta - \theta_\star\|_2.
\end{aligned}$$

which implies

$$\alpha_1 = 2(1 + \|\theta_\star\|_2) r. \tag{32}$$

In addition, the risk L is Neyman orthogonal at $(\theta_\star, g_0)$ since

$$\mathrm{D}_g\mathrm{D}_\theta L(\theta_\star, g_0)[\theta - \theta_\star, g - g_0] = 0. \tag{33}$$

Note that

$$\begin{aligned}
&\mathrm{D}_g^2\mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0, g - g_0] \\
&\quad = 2\mathbb{E}_{\mathbb{P}}\left[(-(g_Y - g_{0,Y})(W) + \langle \theta_\star, (g_X - g_{0,X})(W)\rangle)\langle \theta - \theta_\star, (g_X - g_{0,X})(W)\rangle\right].
\end{aligned}$$

By identical proof of (32), we have that L satisfies Asm. 4 since

$$\mathrm{D}_g^2\mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0, g - g_0] \leq 2(1 + \|\theta_\star\|_2)\|g - g_0\|_{\mathcal{G}}^2 \|\theta - \theta_\star\|_2,$$

which implies

$$\beta_1 = 2(1 + \|\theta_\star\|_2). \tag{34}$$

□

B.4.2 Proof of Lemma 5

Proof. We consider the following loss:

$$\ell(\theta, g; z) = \frac{1}{2}(y - g(w) - \langle \theta, x\rangle)^2,$$

with the corresponding risk function defined as

$$L(\theta, g) = \frac{1}{2}\mathbb{E}_{\mathbb{P}}\left[(Y - g(W) - \langle\theta, X\rangle)^2\right]$$

Under the true nuisance, the target is the minimizer of the following squared loss:

$$\theta_\star = \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{2}\mathbb{E}_{\mathbb{P}}\left[(Y - g_0(W) - \langle\theta, X\rangle)^2\right].$$

Since $\epsilon = Y - g_0(W) - \langle\theta_0, X\rangle$ satisfies $\mathbb{E}_{\mathbb{P}}[\epsilon \mid X, W] = 0$ under the true model, by bias-variance decomposition, we have

$$\theta_\star = \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{2}\mathbb{E}_{\mathbb{P}}\left[(\langle\theta_0, X\rangle - \langle\theta, X\rangle)^2\right] = \theta_0. \tag{35}$$

Differentiating $\ell(\theta, g; z)$ with respect to θ, we obtain the gradient and Hessian *w.r.t.* θ as

$$S_\theta(\theta, g; z) = -(y - g(w) - \langle x, \theta\rangle)x \text{ and } H_{\theta\theta}(\theta, g; z) = xx^\top.$$

The expected gradient and expected Hessian are then obtained as

$$S_\theta(\theta, g) = -\mathbb{E}_{\mathbb{P}}\left[(Y - g(W) - \langle X, \theta\rangle)X\right] \text{ and } H_{\theta\theta}(\theta, g) = \mathbb{E}_{\mathbb{P}}\left[XX^\top\right].$$

We consider the nuisance neighborhood such that for $g \in \mathcal{G}_r(g_0)$,

$$\|g - g_0\|_{\mathcal{G}} := \mathbb{E}_{\mathbb{P}}\left[(g(W) - g_0(W))^2\right]^{1/2} \le r. \tag{36}$$

We now verify that the loss function ℓ satisfies Asm. 3.

(a) We assume that $g : \mathcal{W} \mapsto \mathbb{R}$ is continuous, thus Asm. 3(a) is satisfied.

(b) Since $\theta_\star = \theta_0$ by (35), we have

$$S_\theta(\theta_\star, g_0) = -\mathbb{E}_{\mathbb{P}}[\epsilon X] = 0. \tag{37}$$

(c) When $\lambda_{\min}(\mathbb{E}_{\mathbb{P}}\left[XX^\top\right]) \ge \lambda_0 > 0$ and $\|X\|_2 \le C_X$ *a.s.* , $L(\theta, g)$ is λ_0-strongly convex and C_X^2-smooth since

$$\lambda_0\mathbf{I} \preccurlyeq H_{\theta\theta}(\theta, g) \preccurlyeq C_X^2\mathbf{I} \implies \mu = \lambda_0 \text{ and } M = C_X^2. \tag{38}$$

(d) Consider the Taylor expansion around $\theta_\star$, we have

$$S_\theta(\theta, g; Z) - S_\theta(\theta, g) = S_\theta(\theta_\star, g; Z) - S_\theta(\theta_\star, g) + (H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g))(\theta - \theta_\star).$$

Since $S_\theta(\theta_\star, g; Z) = ((g - g_0)(w) - \epsilon)X$ and $\|X\|_2 \le C_X$ *a.s.* , we have

$$\begin{aligned}\|S_\theta(\theta_\star, g; Z) - S_\theta(\theta_\star, g)\|_2 &= \|((g - g_0)(W) - \epsilon)X - \mathbb{E}_{\mathbb{P}}\left[((g - g_0)(W))X\right]\|_2 \\ &\le C_X\left(|(g - g_0)(W)| + \mathbb{E}_{\mathbb{P}}\left[|(g - g_0)(W)|\right] + |\epsilon|\right).\end{aligned}$$

On the other hand,

$$H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g) = XX^\top - \mathbb{E}\left[XX^\top\right] \preccurlyeq 2C_X^2\mathbf{I},$$

which implies that

$$\|(H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g))(\theta - \theta_\star)\|_2 \le 2C_X^2\|\theta - \theta_\star\|_2.$$

For $g \in \mathcal{G}_r(g_0)$, when $\mathbb{E}_{\mathbb{P}}\left[\epsilon^2\right] \le \sigma^2$ we have

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[\|S_\theta(\theta, g; Z) - S_\theta(\theta, g)\|_2^2\right] \le & 2\mathbb{E}_{\mathbb{P}}\left[\|S_\theta(\theta_\star, g; Z) - S_\theta(\theta_\star, g)\|_2^2\right] \\
& + 2\mathbb{E}_{\mathbb{P}}\left[\|(H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g))(\theta - \theta_\star)\|_2^2\right] \\
\le & 2C_X^2\mathbb{E}_{\mathbb{P}}\left[\left(|(g - g_0)(W)| + \mathbb{E}_{\mathbb{P}}\left[(g - g_0)(W)\right] + |\epsilon|\right)^2\right] \\
& + 2C_X^4\|\theta - \theta_\star\|_2^2 \\
\le & 6C_X^2\left(2\mathbb{E}_{\mathbb{P}}\left[((g - g_0)(W))^2\right] + \mathbb{E}_{\mathbb{P}}\left[\epsilon^2\right]\right) + 2C_X^4\|\theta - \theta_\star\|_2^2 \\
\le & 6C_X^2\left(2r^2 + \sigma^2\right) + 2C_X^4\|\theta - \theta_\star\|_2^2,
\end{aligned}$$

which implies that

$$K_1 = 6C_X^2\left(2r^2 + \sigma^2\right) \text{ and } \kappa_1 = 2C_X^4. \tag{39}$$

(e) For any $\theta \in \Theta$ and $g, \bar{g} \in \mathcal{G}_r(g_0)$, we have

$$\begin{aligned}
|\mathrm{D}_g\mathrm{D}_\theta L(\theta, \bar{g})[\theta - \theta_\star, g - g_0]| &= |\mathbb{E}_{\mathbb{P}}\left[(g - g_0)(W)\langle X, \theta - \theta_\star\rangle\right]| \\
&\le \mathbb{E}_{\mathbb{P}}\left[|(g - g_0)(W)\langle X, \theta - \theta_\star\rangle|\right] \\
&\le C_X\|\theta - \theta_\star\|_2\mathbb{E}_{\mathbb{P}}\left[((g - g_0)(W))^2\right]^{1/2},
\end{aligned}$$

which implies that

$$\alpha_1 = C_X. \tag{40}$$

□

B.4.3 Proof of Lemma 6

Proof. We consider the following loss:

$$\ell(\theta, g; z) = \frac{1}{2}(y - g(w) - \langle\theta, x\rangle)^2,$$

with the corresponding risk function defined as

$$L(\theta, g) = \frac{1}{2}\mathbb{E}_{\mathbb{P}}\left[(Y - g(W) - \langle\theta, X\rangle)^2\right].$$

First by the same proof as Appx. B.4.2, we have $\theta_\star = \theta_0$. Define the inner product $\langle\cdot,\cdot\rangle_{\mathcal{G}}$ as (18) and define the norm $\|\cdot\|_{\mathcal{G}}$ such that $\|g\|_{\mathcal{G}}^2 = \langle g, g\rangle_{\mathcal{G}} \forall g \in \mathcal{G}$. Consider a uniformly bounded neighborhood $\mathcal{G}_r(g_0)$ such that

$$\mathcal{G}_r(g_0) = \{g \in \mathcal{G} : |g(W) - g_0(W)| \le r \textit{ almost surely}\}. \tag{41}$$

The NO gradient oracle for this non-orthogonal loss is derived as (22) such that

$$S_{\text{no}}(\theta, g; z) = -(y - g(w) - \langle\theta, x\rangle)(x - \mathbb{E}\left[X \mid W = w\right]).$$

We now verify that Asm. 6 is satisfied.

(a) Since $\epsilon = Y - g_0(W) - \langle\theta_0, X\rangle$ satisfies $\mathbb{E}_{\mathbb{P}}\left[\epsilon \mid X, W\right] = 0$ under the true model, by (22) we first have

$$\begin{aligned}
S_{\text{no}}(\theta_\star, g_0) &= \mathbb{E}_{\mathbb{P}}\left[S_{\text{no}}(\theta_\star, g_0; Z)\right] \\
&= -\mathbb{E}_{\mathbb{P}}\left[\epsilon(X - \mathbb{E}\left[X \mid W\right])\right] \\
&= -\mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\mathbb{P}}\left[\epsilon \mid X, W\right](X - \mathbb{E}\left[X \mid W\right])\right] = 0.
\end{aligned} \tag{42}$$

Let $\gamma_0^{(j)} = H_{gg}^{-1}H_{\theta g}^{(j)}$ for $j = 1, \ldots, d$. By (85), we have $[\Gamma_0 g]_j = \langle\gamma_0^{(j)}, g\rangle_{\mathcal{G}}, \forall g \in \mathcal{G}$. Thus, by (82) we have

$$[(\hat{\Gamma} - \Gamma_0)\nabla_g L(\theta_\star, g_0)]_j = \langle\hat{\gamma}^{(j)} - \gamma_0^{(j)}, \nabla_g L(\theta_\star, g_0)\rangle_{\mathcal{G}} = \mathrm{D}_g L(\theta_\star, g_0)[\hat{\gamma}^{(j)} - \gamma_0^{(j)}],$$

which, by (19), implies that

$$[(\hat{\Gamma} - \Gamma_0)\nabla_g L(\theta_\star, g_0)]_j = -\mathbb{E}_{\mathbb{P}}\left[\epsilon[\hat{\Gamma} - \Gamma_0]_j(W)\right] = \mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\mathbb{P}}\left[\epsilon \mid W\right][\hat{\Gamma} - \Gamma_0]_j(W)\right] = 0. \quad (43)$$

Thus, Asm. 6(a) holds true due to (42) and (43).

(b) By (22), for any $(\theta, g) \in \Theta \times \mathcal{G}$,

$$\begin{aligned}\nabla_\theta S_{\text{no}}(\theta, g) &= \mathbb{E}_{\mathbb{P}}\left[X(X - \mathbb{E}_{\mathbb{P}}\left[X \mid W\right])^\top\right] \\ &= \mathbb{E}_{\mathbb{P}}\left[XX^\top\right] - \mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\mathbb{P}}\left[X \mid W\right]\mathbb{E}_{\mathbb{P}}\left[X \mid W\right]^\top\right] \\ &= \mathbb{E}_{\mathbb{P}}\left[\tilde{X}\tilde{X}^\top\right], \end{aligned} \quad (44)$$

which implies that

$$\lambda_{\min}\nabla_\theta S_{\text{no}}(\theta, g) + \nabla_\theta S_{\text{no}}(\theta, g)^\top = \lambda_{\min} 2\mathbb{E}_{\mathbb{P}}\left[\tilde{X}\tilde{X}^\top\right] \geq 2\lambda_0.$$

Thus, Asm. 6(b) holds true for $\mu_{\text{no}} = \lambda_0$.

(c) For any $(\theta, g, \bar{g}) \in \Theta \times \mathcal{G} \times \mathcal{G}_r(g_0)$, by (19),

$$\begin{aligned}&\mathbb{E}_{\mathbb{P}}\left[(\mathrm{D}_g L(\theta, \bar{g}; Z)[g] - \mathrm{D}_g L(\theta, \bar{g})[g])^2\right] \\ &\leq \mathbb{E}_{\mathbb{P}}\left[(\mathrm{D}_g L(\theta, \bar{g}; Z)[g])^2\right] \\ &= \mathbb{E}_{\mathbb{P}}\left[(Y - \bar{g}(W) - \langle\theta, X\rangle)^2 (g(W))^2\right] \\ &= \mathbb{E}_{\mathbb{P}}\left[(\epsilon - (\bar{g} - g_0)(W) - \langle\theta - \theta_\star, X\rangle)^2 (g(W))^2\right] \\ &\leq 3\mathbb{E}_{\mathbb{P}}\left[\epsilon^2 (g(W))^2\right] + 3\mathbb{E}_{\mathbb{P}}\left[(\bar{g} - g_0)(W))^2 (g(W))^2\right] + 3\mathbb{E}_{\mathbb{P}}\left[\langle\theta - \theta_\star, X\rangle^2 (g(W))^2\right].\end{aligned}$$

Assume that $\mathbb{E}_{\mathbb{P}}\left[\epsilon^2 \mid W\right] \leq \sigma^2$ and $\|X\|_\infty \leq C_X$ *a.s.* . By (41) we have

$$\mathbb{E}_{\mathbb{P}}\left[(\mathrm{D}_g L(\theta, \bar{g}; Z)[g] - \mathrm{D}_g L(\theta, \bar{g})[g])^2\right] \leq 3(\sigma^2 + r^2 + C_X^2\|\theta - \theta_\star\|_2^2)\|g\|_{\mathcal{G}}^2. \quad (45)$$

Thus, Asm. 6(c) holds true for $K_2 = 3(\sigma^2 + r^2)$ and $\kappa_2 = 3C_X^2$.

(d) For any $(\theta, \bar{g}, g_1, g_2) \in \Theta \times \mathcal{G}_r(g_0) \times \mathcal{G} \times \mathcal{G}$, by (20),

$$\begin{aligned}\left|\mathrm{D}_g^2 L(\theta, g)[g_1, g_2]\right| &= \left|\mathbb{E}_{\mathbb{P}}\left[g_1(W)g_2(W)\right]\right| \\ &\leq \mathbb{E}_{\mathbb{P}}\left[(g_1(W))^2\right]^{1/2}\mathbb{E}_{\mathbb{P}}\left[(g_2(W))^2\right]^{1/2} = \|g_1\|_{\mathcal{G}}\|g_2\|_{\mathcal{G}}. \end{aligned} \quad (46)$$

In addition, for any $(\theta, \bar{\theta}, g) \in \Theta \times \Theta \times \mathcal{G}$, by (19) we have

$$\begin{aligned}\left|\mathrm{D}_\theta \mathrm{D}_g L(\bar{\theta}, g_0)[g, \theta - \theta_\star]\right| &= \left|\mathrm{D}_{\bar{\theta}}\mathbb{E}_{\mathbb{P}}\left[(Y - g_0(W) - \langle\bar{\theta}, X\rangle)g(W)\right][\theta - \theta_\star]\right| \\ &= \left|\mathbb{E}_{\mathbb{P}}\left[\langle\theta - \theta_\star, X\rangle g(W)\right]\right| \\ &\leq \mathbb{E}_{\mathbb{P}}\left[\langle\theta - \theta_\star, X\rangle^2\right]^{1/2}\mathbb{E}_{\mathbb{P}}\left[(g(W))^2\right]^{1/2} \\ &\leq C_X\|\theta - \theta_\star\|_2\|g\|_{\mathcal{G}}. \end{aligned} \quad (47)$$

Thus, Asm. 6(d) holds true for $\alpha_2 = 1$ due to (46) and $\alpha_1 = C_X$ due to (47).

(e) Note that

$$S_{\text{no}}(\theta, g; z) = -(y - g(w) - \langle\theta, x\rangle)(x - \mathbb{E}_{\mathbb{P}}\left[X \mid W = w\right]),$$

which implies that for any $g_1, g_2 \in \mathcal{G}$,

$$D_g^2 S_{\text{no}}(\theta, g; z)[g_1, g_2] = 0. \quad (48)$$

Thus, Asm. 6(e) holds true for $\beta_2 = 0$. □

B.4.4 Proof of Lemma 7

Proof. We consider the following loss:

$$\ell(\theta, g; z) = \frac{1}{2}\left(y - g^{\text{out}}(x) - (w - g^{\text{prop}}(x))\langle\theta, x\rangle\right)^2,$$

with the corresponding risk function defined as

$$L(\theta, g) = \frac{1}{2}\mathbb{E}_{\mathbb{P}}\left[\left(Y - g^{\text{out}}(X) - (W - g^{\text{prop}}(X))\langle\theta, X\rangle\right)^2\right].$$

Note that $\epsilon = Y - g_0^{\text{out}}(X) - \left(W - g_0^{\text{prop}}(X)\right)\tau_0(X)$. Under Asm. 5, we have $\mathbb{E}_{\mathbb{P}}\left[\epsilon \mid W, X\right] = 0$, which implies that

$$\begin{aligned} L(\theta, g_0) &= \frac{1}{2}\mathbb{E}_{\mathbb{P}}\left[\left(\epsilon + \left(W - g_0^{\text{prop}}(X)\right)(\tau_0(X) - \langle\theta, X\rangle)\right)^2\right] \\ &= \frac{1}{2}\mathbb{E}_{\mathbb{P}}\left[\left(W - g_0^{\text{prop}}(X)\right)^2(\tau_0(X) - \langle\theta, X\rangle)^2\right] + \frac{1}{2}\mathbb{E}_{\mathbb{P}}\left[\epsilon^2\right]. \end{aligned}$$

Thus, the target is the minimizer of the following squared loss:

$$\theta_\star = \underset{\theta\in\mathbb{R}^d}{\arg\min}\,\mathbb{E}_{\mathbb{P}}\left[\left(W - g_0^{\text{prop}}(X)\right)^2(\tau_0(X) - \langle\theta, X\rangle)^2\right]. \tag{49}$$

Differentiating $\ell(\theta, g; z)$ with respect to θ, we obtain the gradient and Hessian *w.r.t.* θ as

$$\begin{aligned} S_\theta(\theta, g; z) &= -\left(y - g^{\text{out}}(x) - (w - g^{\text{prop}}(x))\langle\theta, x\rangle\right)(w - g^{\text{prop}}(x))\,x, \\ H_{\theta\theta}(\theta, g; z) &= (w - g^{\text{prop}}(x))^2\,xx^\top. \end{aligned}$$

The expected gradient and expected Hessian are then obtained as

$$\begin{aligned} S_\theta(\theta, g) &= -\mathbb{E}_{\mathbb{P}}\left[\left(Y - g^{\text{out}}(X) - (W - g^{\text{prop}}(X))\langle\theta, X\rangle\right)(W - g^{\text{prop}}(X))\,X\right], \\ H_{\theta\theta}(\theta, g) &= \mathbb{E}_{\mathbb{P}}\left[\left(g_0^{\text{prop}}(1 - g_0^{\text{prop}})(X) + ((g^{\text{prop}} - g_0^{\text{prop}})(X))^2\right)XX^\top\right]. \end{aligned}$$

We consider the nuisance neighborhood such that for $g \in \mathcal{G}_r\,(g_0)$,

$$\|g - g_0\|_{\mathcal{G}} := \max\left\{\mathbb{E}_{\mathbb{P}}\left[(g^{\text{out}}(X) - g_0^{\text{out}}(X))^4\right]^{\frac{1}{4}}, \mathbb{E}_{\mathbb{P}}\left[(g^{\text{out}}(X) - g_0^{\text{out}}(X))^4\right]^{\frac{1}{4}}\right\} \le r.$$

We now verify that the loss function ℓ satisfies Asm. 3.

(a) We assume that $g^{\text{out}} : \mathbb{R}^d \mapsto \mathbb{R}$ and $g^{\text{prop}} : \mathbb{R}^d \mapsto \mathbb{R}$ are continuous, thus Asm. 3(a) is satisfied.

(b) Since $\theta_\star$ is a global minimizer of (49), we have

$$S_\theta(\theta_\star, g_0) = 0. \tag{50}$$

(c) We assume that $c_0 \le g_0^{\text{prop}}(X) \le 1 - c_0$ *a.s.* for some $c_0 > 0$. When $\lambda_{\min}(\mathbb{E}_{\mathbb{P}}\left[XX^\top\right]) \ge \lambda_0 > 0$ and $\|X\|_2 \le C_X$ *a.s.* , we have

$$c_0^2\lambda_0\mathbf{I} \preccurlyeq H_{\theta\theta}(\theta, g) \preccurlyeq (1 + r^2)C_X^2\mathbf{I} \implies \mu = c_0^2\lambda_0 \text{ and } M = (1 + r^2)C_X^2. \tag{51}$$

(d) Consider the Taylor expansion around $\theta_\star$, we have

$$S_\theta(\theta, g; Z) - S_\theta(\theta, g) = S_\theta(\theta_\star, g; Z) - S_\theta(\theta_\star, g) + (H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g))(\theta - \theta_\star).$$

Note that

$$\begin{aligned} S_\theta(\theta_\star, g; Z) &= -\left(Y - g^{\text{out}}(X) - (W - g^{\text{prop}}(X))\langle\theta_\star, X\rangle\right)(W - g^{\text{prop}}(X))\,X \\ &= -\left(\epsilon - (g^{\text{out}} - g_0^{\text{out}})(X) + (g^{\text{prop}} - g_0^{\text{prop}})(X)\tau_0(X)\right)(W - g^{\text{prop}}(X))\,X \\ &\quad + (W - g^{\text{prop}}(X))^2(\langle\theta_\star, X\rangle - \tau_0(X))X. \end{aligned}$$

We assume that $\tau_0 : \mathbb{R}^d \mapsto \mathbb{R}$ is continuous. Then when $\|X\|_2 \le C_X$ *a.s.* , $|\tau_0(X)| \le C_\tau$ for some $C_\tau > 0$. It follows that

$$\begin{aligned}\|S_\theta(\theta_\star, g; Z)\|_2 \le& C_X \left|\left(\epsilon - (g^{\text{out}} - g_0^{\text{out}})(X) + (g^{\text{prop}} - g_0^{\text{prop}})(X)\tau_0(X)\right)(W - g^{\text{prop}}(X))\right| \\ &+ C_X \left(W - g^{\text{prop}}(X)\right)^2 |\langle \theta_\star, X\rangle - \tau_0(X)| \\ \le& C_X \left(|\epsilon| + \left|(g^{\text{out}} - g_0^{\text{out}})(X)\right| + C_\tau \left|(g^{\text{prop}} - g_0^{\text{prop}})(X)\right|\right)|W - g^{\text{prop}}(X)| \\ &+ C_X(C_X\|\theta_\star\|_2 + C_\tau)\left(W - g^{\text{prop}}(X)\right)^2.\end{aligned}$$

Since $\mathbb{E}_{\mathbb{P}}\left[(W - g^{\text{prop}}(X))^2 \mid X\right] = g_0^{\text{prop}}(1 - g_0^{\text{prop}})(X) + ((g^{\text{prop}} - g_0^{\text{prop}})(X))^2$ and $(W - g^{\text{prop}}(X))^2 \le 2 + 2((g^{\text{prop}} - g_0^{\text{prop}})(X))^2$, we have

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\left[\|S_\theta(\theta_\star, g; Z)\|_2^2\right] \le& 4C_X^2\mathbb{E}_{\mathbb{P}}\left[\epsilon^2\left(1 + ((g^{\text{prop}} - g_0^{\text{prop}})(X))^2\right)\right] \\ &+ 4C_X^2\mathbb{E}_{\mathbb{P}}\left[((g^{\text{out}} - g_0^{\text{out}})(X))^2\left(1 + ((g^{\text{prop}} - g_0^{\text{prop}})(X))^2\right)\right] \\ &+ 4C_X^2C_\tau^2\mathbb{E}_{\mathbb{P}}\left[((g^{\text{prop}} - g_0^{\text{prop}})(X))^2\left(1 + ((g^{\text{prop}} - g_0^{\text{prop}})(X))^2\right)\right] \\ &+ 4C_X^2(C_X\|\theta_\star\|_2 + C_\tau)^2\mathbb{E}_{\mathbb{P}}\left[4 + 4((g^{\text{prop}} - g_0^{\text{prop}})(X))^4\right].\end{aligned}$$

When $\mathbb{E}_{\mathbb{P}}\left[\epsilon^4\right] \le \sigma^4$, by Hölder inequality,

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\left[\|S_\theta(\theta_\star, g; Z)\|_2^2\right] \le& 4C_X^2(\sigma^2 + 4(C_X\|\theta_\star\|_2 + C_\tau)^2) \\ &+ 4C_X^2(1 + \sigma^2 + C_\tau^2)r^2 + 4C_X^2(1 + C_\tau^2 + 4(C_X\|\theta_\star\|_2 + C_\tau)^2)r^4.\end{aligned}$$

Similarly, use the fact that $\mathbb{E}_{\mathbb{P}}\left[\epsilon \mid W, X\right] = 0$ and by the stationary condition of (49), we have

$$\begin{aligned}S_\theta(\theta_\star, g) =& -\mathbb{E}_{\mathbb{P}}\left[\left((g^{\text{out}} - g_0^{\text{out}})(X) - (g^{\text{prop}} - g_0^{\text{prop}})(X)\tau_0(X)\right)\left(g^{\text{prop}} - g_0^{\text{prop}}\right)(X)X\right] \\ &+ \mathbb{E}_{\mathbb{P}}\left[\left((W - g_0^{\text{prop}}(X))^2 + ((g^{\text{prop}} - g_0^{\text{prop}})(X))^2\right)(\langle\theta_\star, X\rangle - \tau_0(X))X\right] \\ =& -\mathbb{E}_{\mathbb{P}}\left[\left((g^{\text{out}} - g_0^{\text{out}})(X) - (g^{\text{prop}} - g_0^{\text{prop}})(X)\tau_0(X)\right)\left(g^{\text{prop}} - g_0^{\text{prop}}\right)(X)X\right] \\ &+ \mathbb{E}_{\mathbb{P}}\left[((g^{\text{prop}} - g_0^{\text{prop}})(X))^2(\langle\theta_\star, X\rangle - \tau_0(X))X\right],\end{aligned}$$

which implies

$$\|S_\theta(\theta_\star, g)\|_2^2 \le 3C_X^2(1 + C_\tau^2 + (C_X\|\theta_\star\|_2 + C_\tau)^2)r^4.$$

On the other hand,

$$\begin{aligned}H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g) &\preccurlyeq \left(W - g^{\text{prop}}(X)\right)^2 XX^\top + (1 + r^2)C_X^2\mathbf{I} \\ &\preccurlyeq C_X^2(3 + r^2 + 2((g^{\text{prop}} - g_0^{\text{prop}})(X))^2)\mathbf{I},\end{aligned}$$

which implies that

$$\mathbb{E}_{\mathbb{P}}\left[\|(H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g))(\theta - \theta_\star)\|_2^2\right] \le (9C_X^4 + \mathcal{O}(r^2))\|\theta - \theta_\star\|_2^2.$$

Thus,

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\left[\|S_\theta(\theta, g; Z) - S_\theta(\theta, g)\|_2^2\right] \le& 3\mathbb{E}_{\mathbb{P}}\left[\|S_\theta(\theta_\star, g; Z)\|_2^2\right] + 3\|S_\theta(\theta_\star, g)\|_2^2 \\ &+ 3\mathbb{E}_{\mathbb{P}}\left[\|(H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g))(\theta - \theta_\star)\|_2^2\right],\end{aligned}$$

which implies

$$K_1 = 12C_X^2(\sigma^2 + 4(C_X\|\theta_\star\|_2 + C_\tau)^2) + \mathcal{O}(r^2) \text{ and } \kappa_1 = 27C_X^4 + \mathcal{O}(r^2). \tag{52}$$

(e) For any $\theta \in \Theta$ and $g, \bar{g} \in \mathcal{G}_r(g_0)$, we have

$$\begin{aligned}&\mathrm{D}_g\mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0] \\ &\quad= -\mathbb{E}_{\mathbb{P}}\left[\left((\bar{g}^{\text{out}} - g_0^{\text{out}})(g^{\text{prop}} - g_0^{\text{prop}})(X) + (g^{\text{out}} - g_0^{\text{out}})(\bar{g}^{\text{prop}} - g_0^{\text{prop}})(X)\right)\langle X, \theta - \theta_\star\rangle\right] \\ &\qquad+ 2\mathbb{E}_{\mathbb{P}}\left[\tau_0(X)(g^{\text{prop}} - g_0^{\text{prop}})(\bar{g}^{\text{prop}} - g_0^{\text{prop}})(X)\langle X, \theta - \theta_\star\rangle\right] \\ &\qquad+ 2\mathbb{E}_{\mathbb{P}}\left[(g^{\text{prop}} - g_0^{\text{prop}})(\bar{g}^{\text{prop}} - g_0^{\text{prop}})(X)(\langle\theta_\star, X\rangle - \tau_0(X))\langle X, \theta - \theta_\star\rangle\right].\end{aligned}$$

Thus,

$$
\begin{aligned}
&|\mathrm{D}_g\mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0]| \\
&\quad \le C_X\|\theta - \theta_\star\|_2 \mathbb{E}_{\mathbb{P}}\left[\left|(\bar{g}^{\text{out}} - g_0^{\text{out}})(g^{\text{prop}} - g_0^{\text{prop}})(X)\right| + \left|(g^{\text{out}} - g_0^{\text{out}})(\bar{g}^{\text{prop}} - g_0^{\text{prop}})(X)\right|\right] \\
&\qquad + 2C_X(2C_\tau + C_X\|\theta_\star\|_2)\|\theta - \theta_\star\|_2 \mathbb{E}_{\mathbb{P}}\left[\left|(g^{\text{prop}} - g_0^{\text{prop}})(\bar{g}^{\text{prop}} - g_0^{\text{prop}})(X)\right|\right] \\
&\quad \le C_X(1 + 4(C_X\|\theta_\star\|_2 + C_\tau))r\|\theta - \theta_\star\|_2\|g - g_0\|_{\mathcal{G}},
\end{aligned}
$$

which implies

$$
\alpha_1 = C_X(1 + 4(C_X\|\theta_\star\|_2 + C_\tau))r. \tag{53}
$$

In addition, $L(\theta, g)$ is Neyman orthogonal at $(\theta_\star, g_0)$ since

$$
\mathrm{D}_g\mathrm{D}_\theta L(\theta_\star, g_0)[\theta - \theta_\star, g - g_0] = 0.
$$

Since for any $\theta \in \Theta$ and $g, \bar{g} \in \mathcal{G}_r\,(g_0)$,

$$
\begin{aligned}
&\mathrm{D}_g^2\mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0, g - g_0] \\
&\quad = -\mathbb{E}_{\mathbb{P}}\left[\left((g^{\text{out}} - g_0^{\text{out}})(g^{\text{prop}} - g_0^{\text{prop}})(X) + (g^{\text{out}} - g_0^{\text{out}})(g^{\text{prop}} - g_0^{\text{prop}})(X)\right)\langle X, \theta - \theta_\star\rangle\right] \\
&\qquad + 2\mathbb{E}_{\mathbb{P}}\left[\tau_0(X)((g^{\text{prop}} - g_0^{\text{prop}})(X))^2\langle X, \theta - \theta_\star\rangle\right] \\
&\qquad + 2\mathbb{E}_{\mathbb{P}}\left[((g^{\text{prop}} - g_0^{\text{prop}})(X))^2(\langle\theta_\star, X\rangle - \tau_0(X))\langle X, \theta - \theta_\star\rangle\right].
\end{aligned}
$$

Similarly, we can show that

$$
\begin{aligned}
&|\mathrm{D}_g^2\mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0, g - g_0]| \\
&\qquad \le C_X(1 + 4(C_X\|\theta_\star\|_2 + C_\tau))\|\theta - \theta_\star\|_2\|g - g_0\|_{\mathcal{G}}^2,
\end{aligned}
$$

which implies that

$$
\beta_1 = C_X(1 + 4(C_X\|\theta_\star\|_2 + C_\tau)). \tag{54}
$$

□

B.4.5 Proof of Lemma 8

Proof. Let

$$
\phi(g; z) = g^{(1)}(x) - g^{(0)}(x) + \frac{w - g^{\text{prop}}(x)}{g^{\text{prop}}(x)(1 - g^{\text{prop}}(x))}(y - g^{(w)}(x)).
$$

We consider the following loss:

$$
\ell(\theta, g; z) = \frac{1}{2}\left(\phi(g; z) - \langle\theta, x\rangle\right)^2,
$$

with the corresponding risk function defined as

$$
L(\theta, g) = \frac{1}{2}\mathbb{E}_{\mathbb{P}}\left[\left(\phi(g; z) - \langle\theta, x\rangle\right)^2\right].
$$

We define the residual ϵ as

$$
\epsilon = \frac{W - g_0^{\text{prop}}(X)}{g_0^{\text{prop}}(X)(1 - g_0^{\text{prop}}(X))}(Y - g_0^{(W)}(X)).
$$

Under Asm. 5, we have $\mathbb{E}_{\mathbb{P}}\left[\epsilon \mid W, X\right] = 0$. Since $\tau_0(x) = g_0^{(1)}(x) - g_0^{(0)}(x)$, we have

$$
\begin{aligned}
L(\theta, g_0) &= \frac{1}{2}\mathbb{E}_{\mathbb{P}}\left[\left(\epsilon + \tau_0(X) - \langle\theta, X\rangle\right)^2\right] \\
&= \frac{1}{2}\mathbb{E}_{\mathbb{P}}\left[\left(\tau_0(X) - \langle\theta, X\rangle\right)^2\right] + \frac{1}{2}\mathbb{E}_{\mathbb{P}}\left[\epsilon^2\right].
\end{aligned}
$$

Thus, the target is the minimizer of the following squared loss:

$$\theta_\star = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{\mathbb{P}} \left[(\tau_0(X) - \langle \theta, X \rangle)^2 \right]. \tag{55}$$

Differentiating $\ell(\theta, g; z)$ with respect to θ, we obtain the gradient and Hessian *w.r.t.* θ as

$$\begin{aligned} S_\theta(\theta, g; z) &= -\left(\phi(g; z) - \langle \theta, x \rangle\right) x, \\ H_{\theta\theta}(\theta, g; z) &= x x^\top. \end{aligned}$$

The expected gradient and expected Hessian are then obtained as

$$\begin{aligned} S_\theta(\theta, g) &= -\mathbb{E}_{\mathbb{P}} \left[\left(\phi(g; Z) - \langle \theta, X \rangle\right) X \right], \\ H_{\theta\theta}(\theta, g) &= \mathbb{E}_{\mathbb{P}} \left[X X^\top \right]. \end{aligned}$$

We consider the nuisance neighborhood such that for $g \in \mathcal{G}_r(g_0)$,

$$\|g - g_0\|_{\mathcal{G}} := \max \left\{ \mathbb{E}_{\mathbb{P}} \left[\left(\frac{(g^{(w)} - g_0^{(w)})(X)}{g^{(w)}(1 - g^{(w)})(X)} \right)^4 \right]^{\frac{1}{4}}, \mathbb{E}_{\mathbb{P}} \left[\left(\frac{(g^{\text{prop}} - g_0^{\text{prop}})(X)}{g^{\text{prop}}(1 - g^{\text{prop}})(X)} \right)^4 \right]^{\frac{1}{4}} \right\} \leq r.$$

We now verify that the loss function ℓ satisfies Asm. 3.

(a) We assume that $g^{(w)} : \mathbb{R}^d \mapsto \mathbb{R}$, $w = 0, 1$, and $g^{\text{prop}} : \mathbb{R}^d \mapsto (0, 1)$ are continuous, thus Asm. 3(a) is satisfied.

(b) Since $\theta_\star$ is a global minimizer of (55), we have

$$S_\theta(\theta_\star, g_0) = 0. \tag{56}$$

(c) When $\lambda_{\min}(\mathbb{E}_{\mathbb{P}} \left[X X^\top \right]) \geq \lambda_0 > 0$ and $\|X\|_2 \leq C_X$ *a.s.* , we have

$$\lambda_0 \mathbf{I} \preccurlyeq H_{\theta\theta}(\theta, g) \preccurlyeq C_X^2 \mathbf{I} \implies \mu = \lambda_0 \text{ and } M = C_X^2. \tag{57}$$

(d) Consider the Taylor expansion around $\theta_\star$, we have

$$S_\theta(\theta, g; Z) - S_\theta(\theta, g) = S_\theta(\theta_\star, g; Z) - S_\theta(\theta_\star, g) + (H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g))(\theta - \theta_\star).$$

Let $\tau = g^{(1)} - g^{(0)}$ and

$$\begin{aligned} \psi(g; x) &= \frac{1}{g^{\text{prop}}(1 - g^{\text{prop}})(x)} - \frac{1}{g_0^{\text{prop}}(1 - g_0^{\text{prop}})(x)} \\ &= \frac{(g^{\text{prop}} - g_0^{\text{prop}})(x)}{g^{\text{prop}}(1 - g^{\text{prop}})(x)} \cdot \frac{(g^{\text{prop}} + g_0^{\text{prop}})(x) - 1}{g_0^{\text{prop}}(1 - g_0^{\text{prop}})(x)}. \end{aligned}$$

Under Asm. 5, we have

$$|\psi(g; X)| \leq 2 c_0^{-2} \left| \frac{(g^{\text{prop}} - g_0^{\text{prop}})(X)}{g^{\text{prop}}(1 - g^{\text{prop}})(X)} \right|.$$

We can decompose $S_\theta(\theta_\star, g; Z)$ as

$$S_\theta(\theta_\star, g; Z) = I_1 + I_2 + I_3 + I_4 + I_5 + I_6 + I_7 + I_8 + I_9,$$

where

$$
\begin{aligned}
I_1 &= -\left((\tau-\tau_0)(X)+\tau_0(X)-\langle\theta_\star,X\rangle\right)X,\\
I_2 &= -\psi(g;X)(W-g_0^{\text{prop}}(X))(Y-g_0^{(W)}(X))X,\\
I_3 &= \psi(g;X)(W-g_0^{\text{prop}}(X))(g^{(W)}-g_0^{(W)})(X)X,\\
I_4 &= \psi(g;X)(g^{\text{prop}}-g_0^{\text{prop}})(X)(Y-g_0^{(W)}(X))X,\\
I_5 &= -\psi(g;X)(g^{\text{prop}}-g_0^{\text{prop}})(g^{(W)}-g_0^{(W)})(X)X,\\
I_6 &= -\frac{(W-g_0^{\text{prop}}(X))(Y-g_0^{(W)}(X))}{g_0^{\text{prop}}(1-g_0^{\text{prop}})(X)}X=-\epsilon X,\\
I_7 &= \frac{(W-g_0^{\text{prop}}(X))(g^{(W)}-g_0^{(W)})(X)}{g_0^{\text{prop}}(1-g_0^{\text{prop}})(X)}X,\\
I_8 &= \frac{(g^{\text{prop}}-g_0^{\text{prop}})(X)(Y-g_0^{(W)}(X))}{g_0^{\text{prop}}(1-g_0^{\text{prop}})(X)}X,\\
I_9 &= -\frac{(g^{\text{prop}}-g_0^{\text{prop}})(g^{(W)}-g_0^{(W)})(X)}{g_0^{\text{prop}}(1-g_0^{\text{prop}})(X)}X.
\end{aligned}
$$

For I_1, when $\|X\|_2 \le C_X$ *a.s.* , we have $|\tau_0(X)-\langle\theta_\star,X\rangle| \le C$ for some $C>0$ and

$$\|I_1\|_2 \le C_X(|(\tau-\tau_0)(X)|+C),$$

which implies

$$\mathbb{E}_{\mathbb{P}}\left[\|I_1\|_2^2\right] \le 2C_X^2(2r^2+C^2).$$

For I_2, when $\left|Y-g_0^{(W)}(X)\right| \le C_Y$ *a.s.* , we have

$$\|I_2\|_2 \le 4c_0^{-2}C_XC_Y\left|\frac{(g^{\text{prop}}-g_0^{\text{prop}})(X)}{g^{\text{prop}}(1-g^{\text{prop}})(X)}\right|,$$

which implies

$$\mathbb{E}_{\mathbb{P}}\left[\|I_2\|_2^2\right] \le 16c_0^{-4}(C_XC_Y)^2r^2.$$

For I_3,

$$\|I_3\|_2 \le 4c_0^{-2}C_X\left|\frac{(g^{\text{prop}}-g_0^{\text{prop}})(X)}{g^{\text{prop}}(1-g^{\text{prop}})(X)}\right|\left|(g^{(W)}-g_0^{(W)})(X)\right|,$$

which implies

$$\mathbb{E}_{\mathbb{P}}\left[\|I_3\|_2^2\right] \le 16c_0^{-4}C_X^2r^4.$$

For I_4, since $\left|g^{\text{prop}}-g_0^{\text{prop}}\right| \le 2$,

$$\|I_4\|_2 \le 4c_0^{-2}C_XC_Y\left|\frac{((g^{\text{prop}}-g_0^{\text{prop}})(X))}{g^{\text{prop}}(1-g^{\text{prop}})(X)}\right|,$$

which implies

$$\mathbb{E}_{\mathbb{P}}\left[\|I_4\|_2^2\right] \le 16c_0^{-4}(C_XC_Y)^2r^2.$$

For I_5,

$$\|I_5\|_2 \le 4c_0^{-2}C_X\left|\frac{(g^{\text{prop}}-g_0^{\text{prop}})(X)}{g^{\text{prop}}(1-g^{\text{prop}})(X)}\right|\left|(g^{(W)}-g_0^{(W)})(X)\right|,$$

which implies

$$\mathbb{E}_{\mathbb{P}}\left[\|I_5\|_2^2\right] \leq 16c_0^{-4}(C_X C_Y)^2 r^4.$$

For I_6,

$$\|I_6\|_2 \leq C_X \left|\epsilon\right|.$$

When $\mathbb{E}_{\mathbb{P}}\left[\epsilon^2\right] \leq \sigma^2$, we have

$$\mathbb{E}_{\mathbb{P}}\left[\|I_6\|_2^2\right] \leq C_X^2 \sigma^2.$$

For I_7,

$$\|I_7\|_2 \leq 2c_0^{-2} C_X \left|(g^{(W)} - g_0^{(W)})(X)\right|,$$

which implies

$$\mathbb{E}_{\mathbb{P}}\left[\|I_7\|_2^2\right] \leq 4c_0^{-4} C_X^2 r^2.$$

For I_8,

$$\|I_8\|_2 \leq c_0^{-2} C_X C_Y \left|(g^{\text{prop}} - g_0^{\text{prop}})(X)\right|,$$

which implies

$$\mathbb{E}_{\mathbb{P}}\left[\|I_8\|_2^2\right] \leq c_0^{-4}(C_X C_Y)^2 r^2.$$

For I_9,

$$\|I_9\|_2 \leq c_0^{-2} C_X \left|(g^{\text{prop}} - g_0^{\text{prop}})(X)\right| \left|(g^{(W)} - g_0^{(W)})(X)\right|,$$

which implies

$$\mathbb{E}_{\mathbb{P}}\left[\|I_9\|_2^2\right] \leq c_0^{-4} C_X^2 r^4.$$

By Cauchy-Schwarz inequality, it follows that

$$\mathbb{E}_{\mathbb{P}}\left[\|S_\theta(\theta_\star, g; Z)\|_2^2\right] \leq 9C_X^2(2C^2 + \sigma^2) + \mathcal{O}(r^2).$$

Similarly, we have

$$\|S_\theta(\theta_\star, g)\|_2^2 \leq 9C_X^2(2C^2 + \sigma^2) + \mathcal{O}(r^2).$$

Since $H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g) \preccurlyeq C_X^2 \mathbf{I}$, we have

$$\|(H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g))(\theta - \theta_\star)\|_2^2 \leq C_X^4 \|\theta - \theta_\star\|_2^2.$$

Thus,

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\left[\|S_\theta(\theta, g; Z) - S_\theta(\theta, g)\|_2^2\right] \leq & 3\mathbb{E}_{\mathbb{P}}\left[\|S_\theta(\theta_\star, g; Z)\|_2^2\right] + 3\|S_\theta(\theta_\star, g)\|_2^2 \\ & + 3\mathbb{E}_{\mathbb{P}}\left[\|(H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g))(\theta - \theta_\star)\|_2^2\right] \\ = & 27C_X^2(2C^2 + \sigma^2) + \mathcal{O}(r^2) + 3C_X^4\|\theta - \theta_\star\|_2^2,\end{aligned}$$

which implies

$$K_1 = 27C_X^2(2C^2 + \sigma^2) + \mathcal{O}(r^2) \text{ and } \kappa_1 = 3C_X^4. \tag{58}$$

(e) Since $Y = W(Y(1) - Y(0)) + Y(0)$ and $g^{(W)}(X) = W(g^{(1)}(X) - g^{(0)}(X)) + g^{(0)}(X)$, $\phi(g; Z)$ can be written as

$$\begin{aligned}
\phi(g; Z) &= \tau_0(X) + \left(1 - \frac{W(W - g^{\text{prop}}(X))}{g^{\text{prop}}(X)(1 - g^{\text{prop}}(X))}\right)(g^{(1)}(X) - g^{(0)}(X) - \tau_0(X)) \\
&\quad + \frac{W - g^{\text{prop}}(X)}{g^{\text{prop}}(1 - g^{\text{prop}})(X)}(Y(0) - g^{(0)}(X)) + \frac{W(W - g^{\text{prop}}(X))}{g^{\text{prop}}(1 - g^{\text{prop}})(X)}(Y(1) - Y(0) - \tau_0(X)) \\
&= \tau_0(X) + \frac{g^{\text{prop}}(X) - W}{g^{\text{prop}}(X)}\left(g^{(1)} - g_0^{(1)}\right)(X) + \frac{g^{\text{prop}}(X) - W}{1 - g^{\text{prop}}(X)}(g^{(0)} - g_0^{(0)})(X) \\
&\quad + \frac{W - g^{\text{prop}}(X)}{g^{\text{prop}}(X)(1 - g^{\text{prop}}(X))}(Y(0) - g_0^{(0)}(X)) + \frac{W}{g^{\text{prop}}(X)}(Y(1) - Y(0) - \tau_0(X)).
\end{aligned}$$

Under Asm. 5, we have

$$\mathbb{E}_{\mathbb{P}}[\phi(g; Z) \mid X] = \tau_0(X) + \frac{g^{\text{prop}} - g_0^{\text{prop}}}{g^{\text{prop}}}(g^{(1)} - g_0^{(1)})(X) + \frac{g^{\text{prop}} - g_0^{\text{prop}}}{1 - g^{\text{prop}}}(g^{(0)} - g_0^{(0)})(X).$$

Thus, for $\tau = g^{(1)} - g^{(0)}$ and for any $\theta \in \Theta$ and $g, \bar{g} \in \mathcal{G}_r(g_0)$ such that $\bar{g} = tg + (1 - t)g_0$ for some $t \in (0, 1)$, we have

$$\begin{aligned}
&\mathrm{D}_g \mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0] \\
&= -\mathrm{D}_g \mathbb{E}_{\mathbb{P}}\left[(\mathbb{E}_{\mathbb{P}}[\phi(g; Z) \mid X] + \langle \theta_\star, X\rangle)\langle X, \theta - \theta_\star\rangle\right][g - g_0] \\
&= -\mathbb{E}_{\mathbb{P}}\left[\langle X, \theta - \theta_\star\rangle \mathrm{D}_g \mathbb{E}_{\mathbb{P}}[\phi(\bar{g}; Z) \mid X][g - g_0]\right] \\
&= -\mathbb{E}_{\mathbb{P}}\left[\langle X, \theta - \theta_\star\rangle \left(\frac{g_0^{\text{prop}}}{(\bar{g}^{\text{prop}})^2}(\bar{g}^{(1)} - g_0^{(1)})(g^{\text{prop}} - g_0^{\text{prop}})(X) - \frac{\bar{g}^{\text{prop}} - g_0^{\text{prop}}}{\bar{g}^{\text{prop}}}(g^{(1)} - g_0^{(1)})(X)\right)\right] \\
&\quad - \mathbb{E}_{\mathbb{P}}\left[\langle X, \theta - \theta_\star\rangle \left(\frac{(1 - g_0^{\text{prop}})(g^{\text{prop}} - g_0^{\text{prop}})(\bar{g}^{(0)} - g_0^{(0)})}{(1 - \bar{g}^{\text{prop}})^2} - \frac{(\bar{g}^{\text{prop}} - g_0^{\text{prop}})(g^{(0)} - g_0^{(0)})}{1 - \bar{g}^{\text{prop}}}\right)(X)\right].
\end{aligned}$$

Since $(a + b)^4 \le 8a^4 + 8b^4$ for $a, b \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{g_0^{\text{prop}}(X)^4}{g^{\text{prop}}(X)^4}\right] &= \mathbb{E}\left[\frac{\left(g_0^{\text{prop}}(X) - g^{\text{prop}}(X) + g^{\text{prop}}(X)\right)^4}{g^{\text{prop}}(X)^4}\right] \\
&\le 8\mathbb{E}\left[\frac{\left(g_0^{\text{prop}}(X) - g^{\text{prop}}(X)\right)^4 + g^{\text{prop}}(X)^4}{g^{\text{prop}}(X)^4}\right] \le 8r^4 + 8. \qquad (59)
\end{aligned}$$

Similarly, we have

$$\mathbb{E}\left[\frac{(1 - g_0^{\text{prop}}(X))^4}{(1 - g^{\text{prop}}(X))^4}\right] \le 8r^4 + 8.$$

It is easy to show that

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbb{P}}\left[\langle X, \theta - \theta_\star\rangle \frac{g_0^{\text{prop}}}{(\bar{g}^{\text{prop}})^2}(\bar{g}^{(1)} - g_0^{(1)})(g^{\text{prop}} - g_0^{\text{prop}})(X)\right]\right| \\
&\le C_X(8r^4 + 8)^{\frac{1}{4}} r\|\theta - \theta_\star\|_2 \|g - g_0\|_{\mathcal{G}},
\end{aligned}$$

and

$$\left|\mathbb{E}_{\mathbb{P}}\left[\langle X, \theta - \theta_\star\rangle \frac{\bar{g}^{\text{prop}} - g_0^{\text{prop}}}{\bar{g}^{\text{prop}}}(g^{(1)} - g_0^{(1)})(X)\right]\right| \le C_X r\|\theta - \theta_\star\|_2 \|g - g_0\|_{\mathcal{G}}.$$

Thus,

$$\begin{aligned}
|\mathrm{D}_g \mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0]| &\le 2((8r^4 + 8)^{\frac{1}{4}} + 1) C_X r\|\theta - \theta_\star\|_2 \|g - g_0\|_{\mathcal{G}} \\
&\le 2(2(r + 1) + 1) C_X r\|\theta - \theta_\star\|_2 \|g - g_0\|_{\mathcal{G}},
\end{aligned}$$

which implies

$$\alpha_1 = 2C_X(2r+3)r. \tag{60}$$

In addition, $L(\theta, g)$ is Neyman orthogonal at $(\theta_\star, g_0)$ since

$$\mathrm{D}_g\mathrm{D}_\theta L(\theta_\star, g_0)[\theta - \theta_\star, g - g_0] = 0.$$

We have the higher-order derivative such that for any $\theta \in \Theta$ and $g, \bar{g} \in \mathcal{G}_r(g_0)$,

$$\begin{aligned}
&\mathrm{D}_g^2\mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0, g - g_0] \\
&\quad = 2\mathbb{E}_{\mathbb{P}}\left[\left(\frac{g_0^{\text{prop}}(g^{\text{prop}} - g_0^{\text{prop}})^2}{(\bar{g}^{\text{prop}})^3}\left(\bar{g}^{(1)}(X) - g_0^{(1)}(X)\right)\right)\langle X, \theta - \theta_\star\rangle\right] \\
&\qquad - 2\mathbb{E}_{\mathbb{P}}\left[\left(\frac{g_0^{\text{prop}}(g^{\text{prop}} - g_0^{\text{prop}})}{(\bar{g}^{\text{prop}})^2}\left(g^{(1)}(X) - g_0^{(1)}(X)\right)\right)\langle X, \theta - \theta_\star\rangle\right] \\
&\qquad - 2\mathbb{E}_{\mathbb{P}}\left[\left(\frac{(1 - g_0^{\text{prop}})(g^{\text{prop}} - g_0^{\text{prop}})^2}{(1 - \bar{g}^{\text{prop}})^3}\left(\bar{g}^{(0)}(X) - g_0^{(0)}(X)\right)\right)\langle X, \theta - \theta_\star\rangle\right] \\
&\qquad - 2\mathbb{E}_{\mathbb{P}}\left[\left(\frac{(1 - g_0^{\text{prop}})(g^{\text{prop}} - g_0^{\text{prop}})}{(1 - \bar{g}^{\text{prop}})^2}\left(g^{(0)}(X) - g_0^{(0)}(X)\right)\right)\langle X, \theta - \theta_\star\rangle\right].
\end{aligned}$$

Note that $\bar{g}^{\text{prop}} = tg^{\text{prop}} + (1-t)g_0^{\text{prop}}$ for some $t \in (0,1)$ by Taylor's theorem. Then

$$\begin{aligned}
\frac{g^{\text{prop}}(X)}{\bar{g}^{\text{prop}}(X)} &= \frac{g^{\text{prop}}(X)}{tg^{\text{prop}}(X) + (1-t)g_0^{\text{prop}}(X)} \\
&= \frac{1}{t + (1-t)(g_0^{\text{prop}}/g^{\text{prop}})(X)} \leq \frac{1}{t + (1-t)g_0^{\text{prop}}(X)} \leq c_0^{-1}.
\end{aligned} \tag{61}$$

Thus,

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbb{P}}\left[\left(\frac{g_0^{\text{prop}}(g^{\text{prop}} - g_0^{\text{prop}})^2}{(\bar{g}^{\text{prop}})^3}\left(\bar{g}^{(1)}(X) - g_0^{(1)}(X)\right)\right)\langle X, \theta - \theta_\star\rangle\right]\right| \\
&\leq \mathbb{E}_{\mathbb{P}}\left[\frac{(g^{\text{prop}})^4}{(\bar{g}^{\text{prop}})^4}\frac{(g^{\text{prop}}(X) - g_0^{\text{prop}}(X))^4}{(g^{\text{prop}})^4}\right]^{1/2}\mathbb{E}_{\mathbb{P}}\left[\left(\frac{\bar{g}^{(1)}(X) - g_0^{(1)}(X)}{\bar{g}^{\text{prop}}}\right)^2\right]^{1/2} C_X\|\theta - \theta_\star\|_2 \\
&\leq c_0^{-2}C_X r\|\theta - \theta_\star\|_2\|g - g_0\|_{\mathcal{G}}^2.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbb{P}}\left[\left(\frac{g_0^{\text{prop}}(g^{\text{prop}} - g_0^{\text{prop}})}{(\bar{g}^{\text{prop}})^2}\left(g^{(1)}(X) - g_0^{(1)}(X)\right)\right)\langle X, \theta - \theta_\star\rangle\right]\right| \\
&\leq \mathbb{E}_{\mathbb{P}}\left[\left|\frac{(g^{\text{prop}})^2}{(\bar{g}^{\text{prop}})^2}\frac{(g^{\text{prop}} - g_0^{\text{prop}})}{g^{\text{prop}}}\frac{(g^{(1)} - g_0^{(1)})}{g^{\text{prop}}}\right|(X)\right] C_X\|\theta - \theta_\star\|_2 \leq c_0^{-2}C_X\|\theta - \theta_\star\|_2\|g - g_0\|_{\mathcal{G}}^2.
\end{aligned}$$

Then we can show that

$$\left|\mathrm{D}_g^2\mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0, g - g_0]\right| \leq 4c_0^{-2}C_X(1+r)\|\theta - \theta_\star\|_2\|g - g_0\|_{\mathcal{G}}^2.$$

which implies that

$$\beta_1 = 4c_0^{-2}C_X(1+r). \tag{62}$$

□

B.4.6 Proof of Lemma 9

Proof. Define

$$\mu_g^{(s)}(z) = g^{(s)}(x) + \frac{\mathbb{1}(w = s)}{sg^{\text{prop}}(x) + (1-s)(1 - g^{\text{prop}}(x))}(y - g^{(s)}(x)),$$

where $\mathbb{1}(\cdot)$ denotes the indicator function, and the log-linear predictor $p_\theta(x) = e^{\langle\theta,x\rangle}/(1+e^{\langle\theta,x\rangle})$. Under Asm. 5,

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}[\mu_g^{(s)}(Z) \mid X] &= \mathbb{E}_{\mathbb{P}}\left[g^{(s)}(X) + \frac{\mathbb{1}(W=s)}{sg^{\text{prop}}(X)+(1-s)(1-g^{\text{prop}}(X))}(Y(s)-g^{(s)}(X)) \mid X\right] \\ &= g^{(s)}(X) + \frac{sg_0^{\text{prop}}(X)+(1-s)(1-g_0^{\text{prop}}(X))}{sg^{\text{prop}}(X)+(1-s)(1-g^{\text{prop}}(X))}(g_0^{(s)}(X)-g^{(s)}(X)) \\ &= g_0^{(s)}(X) + \frac{(2s-1)(g^{\text{prop}}(X)-g_0^{\text{prop}}(X))}{sg^{\text{prop}}(X)+(1-s)(1-g^{\text{prop}}(X))}(g^{(s)}(X)-g_0^{(s)}(X)) =: f^{(s)}(g;X). \end{aligned} \tag{63}$$

We consider the following loss:

$$\ell(\theta, g; z) = -\left[\mu_g^{(1)}(z)\log p_\theta(x) + \mu_g^{(0)}(z)\log(1-p_\theta(x))\right].$$

with the corresponding risk function defined as

$$L(\theta, g) = -\mathbb{E}_{\mathbb{P}}\left[\mu_g^{(1)}(Z)\log p_\theta(x) + \mu_g^{(0)}(Z)\log(1-p_\theta(X))\right].$$

By (63), we have $\mathbb{E}_{\mathbb{P}}[\mu_{g_0}^{(s)}(Z) \mid X] = g_0^{(s)}(X)$. Thus, the target is the minimizer of the following squared loss:

$$\theta_\star = \underset{\theta\in\mathbb{R}^d}{\arg\min} -\mathbb{E}_{\mathbb{P}}\left[g_0^{(1)}(X)\log p_\theta(x) + g_0^{(0)}(X)\log(1-p_\theta(X))\right]. \tag{64}$$

Since $\nabla_\theta p_\theta(x) = p_\theta(x)(1-p_\theta(x))x$, we have

$$\begin{aligned}\nabla_\theta \log p_\theta(x) &= \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)} = (1-p_\theta(x))x, \\ \nabla_\theta \log(1-p_\theta(x)) &= -\frac{\nabla_\theta p_\theta(x)}{1-p_\theta(x)} = -p_\theta(x)x.\end{aligned}$$

Differentiating $\ell(\theta, g; z)$ with respect to θ, we obtain the gradient and Hessian *w.r.t.* θ as

$$\begin{aligned}S_\theta(\theta, g; z) &= -\left[\mu_g^{(1)}(z)(1-p_\theta(x))x - \mu_g^{(0)}(z)p_\theta(x)x\right], \\ H_{\theta\theta}(\theta, g; z) &= (\mu_g^{(1)}(z)+\mu_g^{(0)}(z))p_\theta(x)(1-p_\theta(x))xx^\top.\end{aligned}$$

The expected gradient and expected Hessian are then obtained as

$$\begin{aligned}S_\theta(\theta, g) &= -\mathbb{E}_{\mathbb{P}}\left[f^{(1)}(g;X)(1-p_\theta(X))X - f^{(0)}(g;X)p_\theta(X)X\right], \\ H_{\theta\theta}(\theta, g) &= \mathbb{E}_{\mathbb{P}}\left[(f^{(1)}(g;X)+f^{(0)}(g;X))p_\theta(X)(1-p_\theta(X))XX^\top\right].\end{aligned}$$

We consider the nuisance neighborhood such that for $g \in \mathcal{G}_r(g_0)$,

$$\|g-g_0\|_{\mathcal{G}} := \max\left\{\mathbb{E}_{\mathbb{P}}\left[\left(\frac{(g^{(w)}-g_0^{(w)})(X)}{g^{(w)}(1-g^{(w)})(X)}\right)^4\right]^{\frac{1}{4}}, \mathbb{E}_{\mathbb{P}}\left[\left(\frac{(g^{\text{prop}}-g_0^{\text{prop}})(X)}{g^{\text{prop}}(1-g^{\text{prop}})(X)}\right)^4\right]^{\frac{1}{4}}\right\} \le r.$$

We assume that $\delta \le g_0^{(0)}(X)+g_0^{(1)}(X) \le \delta^{-1}$ for $\delta > 0$. In addition, we assume that for $g \in \mathcal{G}_r(g_0)$, $f^{(1)}(g;X)+f^{(0)}(g;X) \ge \delta$ *a.s.* . Note that

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\left[|f^{(1)}(g;X)+f^{(0)}(g;X)|\right] &\le \delta + \sum_{s=\{0,1\}}\mathbb{E}_{\mathbb{P}}\left[\left|\frac{g^{\text{prop}}(X)-g_0^{\text{prop}}(X)}{g^{\text{prop}}(1-g^{\text{prop}}(X))}(g^{(s)}(X)-g_0^{(s)}(X))\right|\right] \\ &\le \delta + 2r^2.\end{aligned}$$

We now verify that the loss function ℓ satisfies Asm. 3.

(a) We assume that $g^{(w)}: \mathbb{R}^d \mapsto \mathbb{R}$, $w = 0, 1$, and $g^{\text{prop}}: \mathbb{R}^d \mapsto (0,1)$ are continuous, thus Asm. 3(a) is satisfied.

(b) Since $\theta_\star$ is a global minimizer of (64), we have

$$S_\theta(\theta_\star, g_0) = 0. \tag{65}$$

(c) We assume that Θ is bounded such that $C \le p_\theta(X) \le 1 - C$ *a.s.* for some $C > 0$. When $\lambda_{\min}(\mathbb{E}_{\mathbb{P}}\left[XX^\top\right]) \ge \lambda_0 > 0$ and $\|X\|_2 \le C_X$ *a.s.* , we have

$$C^2\delta\lambda_0\mathbf{I} \preccurlyeq H_{\theta\theta}(\theta, g) \preccurlyeq C_X^2(1 + 2\delta^{-1}r^2)\mathbf{I} \implies \mu = C^2\delta\lambda_0 \text{ and } M = C_X^2(1 + 2\delta^{-1}r^2). \tag{66}$$

(d) Consider the Taylor expansion around $\theta_\star$, we have

$$S_\theta(\theta, g; Z) - S_\theta(\theta, g) = S_\theta(\theta_\star, g; Z) - S_\theta(\theta_\star, g) + (H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g))(\theta - \theta_\star).$$

Note that

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[\|S_\theta(\theta_\star, g; Z) - S_4(\theta_\star, g_4)\|_2^2\right] &\le \mathbb{E}_{\mathbb{P}}\left[\|S_\theta(\theta_\star, g; Z)\|_2^2\right] \\
&\le C_X^2\mathbb{E}\left[\left(\left|\mu_g^{(1)}(Z)\right| + \left|\mu_g^{(0)}(Z)\right|\right)^2\right] \\
&\le 2C_X^2\mathbb{E}\left[(\mu_g^{(1)}(Z))^2 + (\mu_g^{(1)}(Z))^2\right].
\end{aligned}$$

For $s = 1$, when $Y(1) - g_0^{(1)}(X) \le C_Y$ *a.s.* for $C_Y > 0$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}[\mu_g^{(1)}(Z)^2] &= \mathbb{E}_{\mathbb{P}}\left[\left(g_0^{(1)}(X) + \frac{g^{\text{prop}}(X) - W}{g^{\text{prop}}(X)}(g^{(1)} - g_0^{(1)})(X) + \frac{W}{g^{\text{prop}}(X)}\left(Y - g_0^{(1)}(X)\right)\right)^2\right] \\
&\le 3\delta^{-2} + 3\mathbb{E}_{\mathbb{P}}\left[\frac{(g^{\text{prop}}(X) - W)^2}{(g^{\text{prop}}(X))^2}\left((g^{(1)} - g_0^{(1)})(X)\right)^2\right] + 3C_Y^2\mathbb{E}_{\mathbb{P}}\left[\frac{g_0^{\text{prop}}(X)}{(g^{\text{prop}}(X))^2}\right] \\
&\le 3\delta^{-2} + 12\mathbb{E}_{\mathbb{P}}\left[\left(\frac{(g^{(1)} - g_0^{(1)})(X)}{g^{\text{prop}}(X)}\right)^2\right] + 3c_0^{-1}C_Y^2\mathbb{E}_{\mathbb{P}}\left[\frac{(g_0^{\text{prop}}(X))^2}{(g^{\text{prop}}(X))^2}\right] \\
&\le 3\delta^{-2} + 12r^2 + 3c_0^{-1}C_Y^2(8r^4 + 8)^{1/2} \le 3(\delta^{-2} + 4c_0^{-1}C_Y^2) + \mathcal{O}(r^2).
\end{aligned}$$

Thus,

$$\mathbb{E}_{\mathbb{P}}\left[\|S_\theta(\theta_\star, g; Z) - S_4(\theta_\star, g_4)\|_2^2\right] \le 12C_X^2(\delta^{-2} + 4c_0^{-1}C_Y^2) + \mathcal{O}(r^2).$$

Since

$$H_{\theta\theta}(\theta, g; Z) - H_{\theta\theta}(\theta, g) \preccurlyeq C_X^2(\mu_g^{(1)}(Z) + \mu_g^{(0)}(Z) + 1 + 2\delta^{-1}r^2)\mathbf{I},$$

we have

$$\|(H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g))\|_2^2 \le 3C_X^4((\mu_g^{(1)}(Z))^2 + (\mu_g^{(0)}(Z))^2 + (1 + 2\delta^{-1}r^2)^2).$$

Similarly, we can show that

$$\mathbb{E}_{\mathbb{P}}\left[\|(H_{\theta\theta}(\theta_\star, g; Z) - H_{\theta\theta}(\theta_\star, g))\|_2^2\right] \le 3C_X^4(1 + 6(\delta^{-2} + 4c_0^{-1}C_Y^2)) + \mathcal{O}(r^2).$$

It follows that

$$K_1 = 24C_X^2(\delta^{-2} + 4c_0^{-1}C_Y^2) + \mathcal{O}(r^2) \text{ and } \kappa_1 = 6C_X^4(1 + 6(\delta^{-2} + 4c_0^{-1}C_Y^2)) + \mathcal{O}(r^2). \tag{67}$$

(e) For $s = 1$, we have

$$\begin{aligned}
\mathrm{D}_g f^{(1)}(\bar{g}; X)[g - g_0] =& \frac{g_0^{\text{prop}}(X)}{(\bar{g}^{\text{prop}}(X))^2}(\bar{g}^{(1)}(X) - g_0^{(1)}(X))(g^{\text{prop}}(X) - g_0^{\text{prop}}(X)) \\
&+ \frac{\bar{g}^{\text{prop}}(X) - g_0^{\text{prop}}(X)}{\bar{g}^{\text{prop}}(X)}(g^{(1)}(X) - g_0^{(1)}(X)).
\end{aligned}$$

Similarly, for $s = 0$,

$$\begin{aligned}\mathrm{D}_g f^{(0)}(\bar{g}; X)[g - g_0] =& \frac{1 - g_0^{\text{prop}}(X)}{(1 - \bar{g}^{\text{prop}}(X))^2}(\bar{g}^{(0)}(X) - g_0^{(0)}(X))(g^{\text{prop}}(X) - g_0^{\text{prop}}(X)) \\ & - \frac{\bar{g}^{\text{prop}}(X) - g_0^{\text{prop}}(X)}{1 - \bar{g}^{\text{prop}}(X)}(g^{(0)}(X) - g_0^{(0)}(X)).\end{aligned}$$

For any $\bar{g}, g \in \mathcal{G}_r(g_0)$ such that $\bar{g} = tg + (1-t)g_0$ for some $t \in (0,1)$, we have

$$\begin{aligned}&\mathrm{D}_g \mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0] \\ &= -\mathbb{E}_{\mathbb{P}}\left[\mathrm{D}_g(f^{(1)}(\bar{g}; X)(1 - p_\theta(X)) - f^{(0)}(\bar{g}; X)p_\theta(X))[g - g_0]\langle X, \theta - \theta_\star\rangle\right] \\ &= -\mathbb{E}_{\mathbb{P}}\left[\frac{g_0^{\text{prop}}(1 - p_\theta)}{(\bar{g}^{\text{prop}})^2}(\bar{g}^{(1)} - g_0^{(1)})(g^{\text{prop}} - g_0^{\text{prop}})(X)\langle X, \theta - \theta_\star\rangle\right] \\ &\quad - \mathbb{E}_{\mathbb{P}}\left[\frac{(1 - g_0^{\text{prop}})p_\theta}{(1 - \bar{g}^{\text{prop}})^2}(\bar{g}^{(0)} - g_0^{(0)})(g^{\text{prop}} - g_0^{\text{prop}})(X)\langle X, \theta - \theta_\star\rangle\right] \\ &\quad - \mathbb{E}_{\mathbb{P}}\left[\frac{(\bar{g}^{\text{prop}} - g_0^{\text{prop}})(1 - p_\theta)}{\bar{g}^{\text{prop}}}(g^{(1)} - g_0^{(1)})(X)\langle X, \theta - \theta_\star\rangle\right] \\ &\quad + \mathbb{E}_{\mathbb{P}}\left[\frac{(\bar{g}^{\text{prop}} - g_0^{\text{prop}})p_\theta}{1 - \bar{g}^{\text{prop}}}(g^{(0)} - g_0^{(0)})(X)\langle X, \theta - \theta_\star\rangle\right].\end{aligned}$$

Note that by (61),

$$\begin{aligned}&\left|\mathbb{E}_{\mathbb{P}}\left[\frac{g_0^{\text{prop}}(1 - p_\theta)}{(\bar{g}^{\text{prop}})^2}(\bar{g}^{(1)} - g_0^{(1)})(g^{\text{prop}} - g_0^{\text{prop}})(X)\langle X, \theta - \theta_\star\rangle\right]\right| \\ &\le \mathbb{E}_{\mathbb{P}}\left[\left|\frac{g^{\text{prop}}(X)}{\bar{g}^{\text{prop}}(X)}\frac{(\bar{g}^{(1)} - g_0^{(1)})(X)}{g^{\text{prop}}(X)}\frac{(g^{\text{prop}} - g_0^{\text{prop}})(X)}{g^{\text{prop}}(X)}\right|\right] C_X\|\theta - \theta_\star\|_2 \\ &\le c_0^{-1} C_X r\|\theta - \theta_\star\|_2\|g - g_0\|_{\mathcal{G}}.\end{aligned}$$

In addition,

$$\begin{aligned}&\left|\mathbb{E}_{\mathbb{P}}\left[\frac{(\bar{g}^{\text{prop}} - g_0^{\text{prop}})(1 - p_\theta)}{\bar{g}^{\text{prop}}}(g^{(1)} - g_0^{(1)})(X)\langle X, \theta - \theta_\star\rangle\right]\right| \\ &\le \mathbb{E}_{\mathbb{P}}\left[\left|\frac{(\bar{g}^{\text{prop}} - g_0^{\text{prop}})}{\bar{g}^{\text{prop}}}(g^{(1)} - g_0^{(1)})(X)\right|\right] C_X\|\theta - \theta_\star\|_2 \le C_X r\|\theta - \theta_\star\|_2\|g - g_0\|_{\mathcal{G}}.\end{aligned}$$

Thus, it is easy to show that

$$|\mathrm{D}_g \mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0]| \le 2(c_0^{-1} + 1)C_X r\|\theta - \theta_\star\|_2\|g - g_0\|_{\mathcal{G}},$$

which implies

$$\alpha_1 = 2(c_0^{-1} + 1)C_X r. \tag{68}$$

In addition, $L(\theta, g)$ is Neyman orthogonal at $(\theta_\star, g_0)$ since

$$\mathrm{D}_g \mathrm{D}_\theta L(\theta_\star, g_0)[\theta - \theta_\star, g - g_0] = 0.$$

Now we compute the the higher-order derivative. For $s = 1$, we have

$$\begin{aligned}\mathrm{D}_g^2 f^{(1)}(\bar{g}; X)[g - g_0, g - g_0] =& -\frac{2g_0^{\text{prop}}(X)}{(\bar{g}^{\text{prop}}(X))^3}(\bar{g}^{(1)}(X) - g_0^{(1)}(X))(g^{\text{prop}}(X) - g_0^{\text{prop}}(X))^2 \\ & + \frac{2g_0^{\text{prop}}(X)}{(\bar{g}^{\text{prop}}(X))^2}(g^{(1)}(X) - g_0^{(1)}(X))(g^{\text{prop}}(X) - g_0^{\text{prop}}(X)).\end{aligned}$$

Similarly, for $s = 0$,

$$\begin{aligned}\mathrm{D}_g^2 f^{(0)}(\bar{g}; X)[g - g_0, g - g_0] =& 2\frac{1 - g_0^{\text{prop}}(X)}{(1 - \bar{g}^{\text{prop}}(X))^3}(\bar{g}^{(0)}(X) - g_0^{(0)}(X))(g^{\text{prop}}(X) - g_0^{\text{prop}}(X))^2 \\ &+ 2\frac{1 - g_0^{\text{prop}}(X)}{(1 - \bar{g}^{\text{prop}}(X))^2}(g^{(0)}(X) - g_0^{(0)}(X))(g^{\text{prop}}(X) - g_0^{\text{prop}}(X)).\end{aligned}$$

Then for any $\theta \in \Theta$ and $g, \bar{g} \in \mathcal{G}_r(g_0)$ such that $\bar{g} = tg + (1-t)g_0$ for some $t \in (0,1)$,

$$\begin{aligned}&\mathrm{D}_g^2 \mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0, g - g_0] \\ &\quad= -\mathbb{E}_{\mathbb{P}}\left[\mathrm{D}_g^2(f^{(1)}(\bar{g}; X)(1 - p_\theta(X)) - f^{(0)}(\bar{g}; X)p_\theta(X))[g - g_0, g - g_0]\langle X, \theta - \theta_\star\rangle\right] \\ &\quad= 2\mathbb{E}_{\mathbb{P}}\left[\frac{g_0^{\text{prop}}(X)}{(\bar{g}^{\text{prop}}(X))^3}(\bar{g}^{(1)}(X) - g_0^{(1)}(X))(g^{\text{prop}}(X) - g_0^{\text{prop}}(X))^2\langle X, \theta - \theta_\star\rangle\right] \\ &\qquad- 2\mathbb{E}_{\mathbb{P}}\left[\frac{1 - g_0^{\text{prop}}(X)}{(1 - \bar{g}^{\text{prop}}(X))^3}(\bar{g}^{(0)}(X) - g_0^{(0)}(X))(g^{\text{prop}}(X) - g_0^{\text{prop}}(X))^2\langle X, \theta - \theta_\star\rangle\right] \\ &\qquad- 2\mathbb{E}_{\mathbb{P}}\left[\frac{g_0^{\text{prop}}(X)}{(\bar{g}^{\text{prop}}(X))^2}(g^{(1)}(X) - g_0^{(1)}(X))(g^{\text{prop}}(X) - g_0^{\text{prop}}(X))\langle X, \theta - \theta_\star\rangle\right] \\ &\qquad- 2\mathbb{E}_{\mathbb{P}}\left[\frac{1 - g_0^{\text{prop}}(X)}{(1 - \bar{g}^{\text{prop}}(X))^2}(g^{(0)}(X) - g_0^{(0)}(X))(g^{\text{prop}}(X) - g_0^{\text{prop}}(X))\langle X, \theta - \theta_\star\rangle\right].\end{aligned}$$

By (61), we have

$$\begin{aligned}&\left|\mathbb{E}_{\mathbb{P}}\left[\frac{g_0^{\text{prop}}(X)}{(\bar{g}^{\text{prop}}(X))^3}(\bar{g}^{(1)}(X) - g_0^{(1)}(X))(g^{\text{prop}}(X) - g_0^{\text{prop}}(X))^2\langle X, \theta - \theta_\star\rangle\right]\right| \\ &\leq \mathbb{E}_{\mathbb{P}}\left[\left|\frac{(g^{\text{prop}}(X))^2}{(\bar{g}^{\text{prop}}(X))^2}\frac{\bar{g}^{(1)}(X) - g_0^{(1)}(X)}{\bar{g}^{\text{prop}}(X)}\left(\frac{g^{\text{prop}}(X) - g_0^{\text{prop}}(X)}{g^{\text{prop}}(X)}\right)^2\right|\right] C_X\|\theta - \theta_\star\|_2 \\ &\leq c_0^{-2} C_X r\|\theta - \theta_\star\|_2\|g - g_0\|_{\mathcal{G}}^2.\end{aligned}$$

In addition,

$$\begin{aligned}&\left|\mathbb{E}_{\mathbb{P}}\left[\frac{g_0^{\text{prop}}(X)}{(\bar{g}^{\text{prop}}(X))^2}(g^{(1)}(X) - g_0^{(1)}(X))(g^{\text{prop}}(X) - g_0^{\text{prop}}(X))\langle X, \theta - \theta_\star\rangle\right]\right| \\ &\leq \mathbb{E}_{\mathbb{P}}\left[\frac{(g^{\text{prop}}(X))^2}{(\bar{g}^{\text{prop}}(X))^2}\frac{g^{(1)}(X) - g_0^{(1)}(X)}{g^{\text{prop}}(X)}\frac{g^{\text{prop}}(X) - g_0^{\text{prop}}(X)}{g^{\text{prop}}(X)}\right] C_X\|\theta - \theta_\star\|_2 \\ &\leq c_0^{-2} C_X\|\theta - \theta_\star\|_2\|g - g_0\|_{\mathcal{G}}^2.\end{aligned}$$

Together we have

$$\left|\mathrm{D}_g^2 \mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0, g - g_0]\right| \leq 4c_0^{-2} C_X(1 + r)\|\theta - \theta_\star\|_2\|g - g_0\|_{\mathcal{G}}^2,$$

which implies

$$\beta_1 = 4c_0^{-2} C_X(1 + r). \tag{69}$$

□

C Convergence Proofs for Stochastic Gradient

This section is dedicated to demonstrate the SGD convergence in Thm. 1 from Sec. 3 of the main text using Asm. 3 and Asm. 4. We first give an overview of the problem settings and the expected results with the proof outline in Appx. C.1. We then provide all the technical lemmas needed for Thm. 1 in Appx. C.2, and finally prove our first main result in Appx. C.3.

C.1 Overview

In this section, we demonstrate the convergence of SGD for a risk minimization problem with nuisance:

$$\theta_\star = \arg\min_{\theta\in\Theta} L(\theta, g_0),$$

where $g_0 \in \mathcal{G}$ is the true nuisance, $L(\theta, g) = \mathbb{E}_{Z\sim\mathbb{P}}[\ell(\theta, g; Z)]$, and ℓ is a prespecified loss function. We consider the stochastic gradient method for learning $\theta_\star$ when g_0 is unknown but an estimate $\hat{g}$ is accessible. Define $\mathcal{D}_n = (Z_1, \ldots, Z_n)$, sampled from the product measure $\mathbb{P}^n$. Recall the SGD $\theta^{(n)}$ defined as

$$\theta^{(n)} = \theta^{(n-1)} - \eta S_\theta(\theta^{(n-1)}, \hat{g}; Z_n). \tag{70}$$

Throughout the section, we take the following notations for simplicity:

$$\delta^{(n)} = \theta^{(n)} - \theta_\star, \tag{71}$$

$$S^{(n)} = S_\theta(\theta^{(n-1)}, \hat{g}; Z_n), \tag{72}$$

$$v^{(n)} = S^{(n)} - S_\theta(\theta^{(n-1)}, \hat{g}), \tag{73}$$

where $S_\theta(\theta, g; z) = \nabla_\theta \ell(\theta, g; z)$ is the gradient and $S_\theta(\theta, g) = \mathbb{E}_{Z\sim\mathbb{P}}[S_\theta(\theta, g; Z)]$ is the population gradient. We are interested in the mean squared error using an estimated nuisance $\hat{g}$, and our results show that for non-Neyman orthogonal loss L, the error $\delta^{(n)}$ satisfies

$$\mathbb{E}_{\mathcal{D}_n\sim\mathbb{P}^n}[\|\delta^{(n)}\|_2^2] \lesssim \left(1 - \frac{\mu\eta}{2}\right)^n \|\delta^{(0)}\|_2^2 + \|\hat{g} - g_0\|_{\mathcal{G}}^2 + \eta, \tag{74}$$

where the nuisance estimator $\hat{g}$ would lead to a bias of order $\mathcal{O}(\|\hat{g} - g_0\|_{\mathcal{G}}^2)$ for the SGD convergence. If L is Neyman orthogonal, this bias introduced by the nuisance estimation would be further removed, resulting in the following convergence

$$\mathbb{E}_{\mathcal{D}_n\sim\mathbb{P}^n}[\|\delta^{(n)}\|_2^2] \lesssim \left(1 - \frac{\mu\eta}{2}\right)^n \|\delta^{(0)}\|_2^2 + \|\hat{g} - g_0\|_{\mathcal{G}}^4 + \eta. \tag{75}$$

Proof Outline. The proofs for both results (74) and (75) proceed through the following four steps:

1. Upper bound the excess risk $L(\theta^{(n)}, \hat{g}) - L(\theta_\star, \hat{g})$ in terms of the SGD improvement.
2. Lower bound $L(\theta^{(n)}, \hat{g}) - L(\theta_\star, \hat{g})$ using strong convexity and Neyman orthogonality.
3. Derive a recursive formula of $\mathbb{E}_{\mathcal{D}_n\sim\mathbb{P}^n}[\|\delta^{(n)}\|_2^2]$ from these bounds.
4. Perform the recursion and obtain the final result.

Follow these steps above, we provide technical lemma in Appx. C.2, and then prove our first main result Thm. 1 in Appx. C.3.

C.2 Technical Lemma

Lemma 10 (One-step improvement for SGD)**.** *Suppose that Asm. 3 holds. If $\eta < 1/M$, $\theta^{(n)} \in \Theta$, and $\hat{g} \in \mathcal{G}_r(g_0)$, it holds that*

$$2\eta(L(\theta^{(n)}, \hat{g}) - L(\theta_\star, \hat{g})) \le (1 - \mu\eta)\|\delta^{(n-1)}\|_2^2 - \|\delta^{(n)}\|_2^2 - 2\eta\langle v^{(n)}, \delta^{(n-1)}\rangle + \frac{\eta^2}{1 - M\eta}\|v^{(n)}\|_2^2.$$

Proof. We first define the η^{-1}-strongly convex function f_n as

$$f_n(u) = \langle S^{(n)}, u - \theta^{(n-1)} \rangle + \frac{1}{2\eta}\|u - \theta^{(n-1)}\|_2^2. \tag{76}$$

Note that

$$\arg\min_{u\in\mathbb{R}^d} f_n(u) = \arg\min_{u\in\mathbb{R}^d} \|u - (\theta^{(n-1)} - \eta S^{(n)})\|_2^2,$$

which implies that $\theta^{(n)} = \theta^{(n-1)} - \eta S^{(n)}$ is the global minimizer of (76) and $\nabla_\theta f_n(\theta^{(n)}) = 0$. Then

$$\begin{aligned} f_n(\theta_\star) &\geq f_n(\theta^{(n)}) + \langle \nabla_\theta f_n(\theta^{(n)}), -\delta^{(n)} \rangle + \frac{1}{2\eta}\|\delta^{(n)}\|_2^2 \\ &= f_n(\theta^{(n)}) + \frac{1}{2\eta}\|\delta^{(n)}\|_2^2. \end{aligned} \tag{77}$$

Since $L(\cdot, \hat{g})$ is μ-strongly convex and $f_n(\theta_\star) = \langle S^{(n)}, -\delta^{(n-1)} \rangle + (2\eta)^{-1}\|\delta^{(n-1)}\|_2^2$, we have that

$$\begin{aligned} L(\theta_\star, \hat{g}) &\geq L(\theta^{(n-1)}, \hat{g}) + \langle S_\theta(\theta^{(n-1)}, \hat{g}), -\delta^{(n-1)} \rangle + \frac{\mu}{2}\|\delta^{(n-1)}\|_2^2 \\ &= L(\theta^{(n-1)}, \hat{g}) + \langle S^{(n)}, -\delta^{(n-1)} \rangle + \langle v^{(n)}, \delta^{(n-1)} \rangle + \frac{\mu}{2}\|\delta^{(n-1)}\|_2^2 \\ &= L(\theta^{(n-1)}, \hat{g}) + f_n(\theta_\star) + \langle v^{(n)}, \delta^{(n-1)} \rangle + \left(\frac{\mu}{2} - \frac{1}{2\eta}\right)\|\delta^{(n-1)}\|_2^2. \end{aligned}$$

Together with (77), it follows that

$$\begin{aligned} L(\theta^{(n-1)}, \hat{g}) + f_n(\theta^{(n)}) \leq & L(\theta_\star, \hat{g}) - \langle v^{(n)}, \delta^{(n-1)} \rangle \\ & + \left(\frac{1}{2\eta} - \frac{\mu}{2}\right)\|\delta^{(n-1)}\|_2^2 - \frac{1}{2\eta}\|\delta^{(n)}\|_2^2. \end{aligned} \tag{78}$$

Since $L(\cdot, \hat{g})$ is M-smooth and $f_n\left(\theta^{(n)}\right) = \langle S^{(n)}, \theta^{(n)} - \theta^{(n-1)} \rangle + (2\eta)^{-1}\|\theta^{(n)} - \theta^{(n-1)}\|_2^2$, we have that

$$\begin{aligned} L(\theta^{(n)}, \hat{g}) &\leq L(\theta^{(n-1)}, \hat{g}) + \langle S(\theta^{(n-1)}, \hat{g}), \theta^{(n)} - \theta^{(n-1)} \rangle + \frac{M}{2}\|\theta^{(n)} - \theta^{(n-1)}\|_2^2 \\ &= L(\theta^{(n-1)}, \hat{g}) + \langle S^{(n)}, \theta^{(n)} - \theta^{(n-1)} \rangle + \frac{M}{2}\|\theta^{(n)} - \theta^{(n-1)}\|_2^2 - \langle v^{(n)}, \theta^{(n)} - \theta^{(n-1)} \rangle \\ &= L(\theta^{(n-1)}, \hat{g}) + f_n(\theta^{(n)}) + \left(\frac{M}{2} - \frac{1}{2\eta}\right)\|\theta^{(n)} - \theta^{(n-1)}\|_2^2 - \langle v^{(n)}, \theta^{(n)} - \theta^{(n-1)} \rangle. \end{aligned}$$

By (78), it follows that

$$\begin{aligned} L(\theta^{(n)}, \hat{g}) \leq & L(\theta_\star, \hat{g}) - \langle v^{(n)}, \delta^{(n-1)} \rangle + \left(\frac{1}{2\eta} - \frac{\mu}{2}\right)\|\delta^{(n-1)}\|_2^2 - \frac{1}{2\eta}\|\delta^{(n)}\|_2^2 \\ & + \left(\frac{M}{2} - \frac{1}{2\eta}\right)\|\theta^{(n)} - \theta^{(n-1)}\|_2^2 - \langle v^{(n)}, \theta^{(n)} - \theta^{(n-1)} \rangle, \end{aligned}$$

which implies that

$$\begin{aligned} L(\theta^{(n)}, \hat{g}) - L(\theta_\star, \hat{g}) \leq & \left(\frac{1}{2\eta} - \frac{\mu}{2}\right)\|\delta^{(n-1)}\|_2^2 - \frac{1}{2\eta}\|\delta^{(n)}\|_2^2 - \langle v^{(n)}, \delta^{(n-1)} \rangle \\ & + \left(\frac{M}{2} - \frac{1}{2\eta}\right)\|\theta^{(n)} - \theta^{(n-1)}\|_2^2 - \langle v^{(n)}, \theta^{(n)} - \theta^{(n-1)} \rangle. \end{aligned} \tag{79}$$

For any $\omega > 0$, by Cauchy-Schwarz inequality and Young's inequality, we have

$$-\langle v^{(n)}, \theta^{(n)} - \theta^{(n-1)} \rangle \leq \frac{\omega}{2}\|v^{(n)}\|_2^2 + \frac{1}{2\omega}\|\theta^{(n)} - \theta^{(n-1)}\|_2^2.$$

Take this into (79) and we have

$$L(\theta^{(n)},\hat{g})-L(\theta_\star,\hat{g})\leq\left(\frac{1}{2\eta}-\frac{\mu}{2}\right)\|\delta^{(n-1)}\|_2^2-\frac{1}{2\eta}\|\delta^{(n)}\|_2^2-\langle v^{(n)},\delta^{(n-1)}\rangle \\ +\left(\frac{M}{2}-\frac{1}{2\eta}+\frac{1}{2\omega}\right)\|\theta^{(n)}-\theta^{(n-1)}\|_2^2+\frac{\omega}{2}\|v^{(n)}\|_2^2.$$

When $\eta<1/M$, set $\frac{M}{2}-\frac{1}{2\eta}+\frac{1}{2\omega}=0$, i.e., set $\omega=1/(\eta^{-1}-M)$. It follows that

$$L(\theta^{(n)},\hat{g})-L(\theta_\star,\hat{g})\leq\left(\frac{1}{2\eta}-\frac{\mu}{2}\right)\|\delta^{(n-1)}\|_2^2-\frac{1}{2\eta}\|\delta^{(n)}\|_2^2-\langle v^{(n)},\delta^{(n-1)}\rangle \\ +\frac{\eta}{2(1-M\eta)}\|v^{(n)}\|_2^2.$$

We complete the proof by multiplying both sides of the inequality by 2η. □

Lemma 11. *Suppose that Asm. 3 holds. If $\eta<1/M$, $\theta^{(n)}\in\Theta$, and $\hat{g}\in\mathcal{G}_r(g_0)$, it holds that*

$$\|\delta^{(n)}\|_2^2\leq(1-\mu\eta)\|\delta^{(n-1)}\|_2^2+\frac{\alpha_1^2\eta}{\mu}\|\hat{g}-g_0\|_{\mathcal{G}}^2+2\eta\langle v^{(n)},\delta^{(n-1)}\rangle+\frac{\eta^2\|v^{(n)}\|_2^2}{1-M\eta}.$$

Proof. Under Asm. 3,

$$\begin{aligned}D_\theta L(\theta_\star,\hat{g})[\delta^{(n)}]&=D_\theta L(\theta_\star,g_0)[\delta^{(n)}]+D_gD_\theta L(\theta_\star,\bar{g})[\delta^{(n)},\hat{g}-g_0]\\&=D_gD_\theta L(\theta_\star,\bar{g})[\delta^{(n)},\hat{g}-g_0]\geq-\alpha_1\|\delta^{(n)}\|_2\|\hat{g}-g_0\|_{\mathcal{G}}.\end{aligned}\tag{80}$$

Since $L(\cdot,\hat{g})$ is μ-strongly convex,

$$\begin{aligned}L(\theta^{(n)},\hat{g})-L(\theta_\star,\hat{g})&\geq\langle S(\theta_\star,\hat{g}),\delta^{(n)}\rangle+\frac{\mu}{2}\|\delta^{(n)}\|_2^2\\&=D_\theta L(\theta_\star,\hat{g})[\delta^{(n)}]+\frac{\mu}{2}\|\delta^{(n)}\|_2^2.\end{aligned}$$

By (80), it follows that

$$L(\theta^{(n)},\hat{g})-L(\theta_\star,\hat{g})\geq-\alpha_1\|\delta^{(n)}\|_2\|\hat{g}-g_0\|_{\mathcal{G}}+\frac{\mu}{2}\|\delta^{(n)}\|_2^2.$$

Together with Lemma 10, we have

$$2\eta\left(-\alpha_1\|\delta^{(n)}\|_2\|\hat{g}-g_0\|_{\mathcal{G}}+\frac{\mu}{2}\|\delta^{(n)}\|_2^2\right)\leq(1-\mu\eta)\|\delta^{(n-1)}\|_2^2 \\ -\|\delta^{(n)}\|_2^2-2\eta\langle v^{(n)},\delta^{(n-1)}\rangle+\frac{\eta^2\|v^{(n)}\|_2^2}{1-M\eta}.$$

Rearranging it, we have

$$(1+\mu\eta)\|\delta^{(n)}\|_2^2\leq(1-\mu\eta)\|\delta^{(n-1)}\|_2^2 \\ +2\eta\alpha_1\|\delta^{(n)}\|_2\|\hat{g}-g_0\|_{\mathcal{G}}-2\eta\langle v^{(n)},\delta^{(n-1)}\rangle+\frac{\eta^2\|v^{(n)}\|_2^2}{1-M\eta}.\tag{81}$$

By Young's inequality,

$$2\eta\alpha_1\|\delta^{(n)}\|_2\|\hat{g}-g_0\|_{\mathcal{G}}\leq\eta\alpha_1\left(\frac{\mu}{\alpha_1}\|\delta^{(n)}\|_2^2+\frac{\alpha_1}{\mu}\|\hat{g}-g_0\|_{\mathcal{G}}^2\right).$$

Take this into (81) and we have

$$\|\delta^{(n)}\|_2^2\leq(1-\mu\eta)\|\delta^{(n-1)}\|_2^2+\eta\alpha_1^2\mu^{-1}\|\hat{g}-g_0\|_{\mathcal{G}}^2-2\eta\langle v^{(n)},\delta^{(n-1)}\rangle+\frac{\eta^2\|v^{(n)}\|_2^2}{1-M\eta}.$$

□

Corollary 12. *Suppose that Asm. 3 holds. If $\eta < 1/M$, $\theta^{(n)} \in \Theta$, and $\hat{g} \in \mathcal{G}_r(g_0)$, it holds that*

$$\mathbb{E}_{Z_n \sim \mathbb{P}}\left[\|\delta^{(n)}\|_2^2\right] \leq \left(1 - \mu\eta + \frac{\kappa_1 \eta^2}{1 - M\eta}\right) \|\delta^{(n-1)}\|_2^2 + \frac{\alpha_1^2 \eta}{\mu} \|\hat{g} - g_0\|_{\mathcal{G}}^2 + \frac{K_1 \eta^2}{1 - M\eta}.$$

Proof. Note that $\mathbb{E}_{Z_n \sim \mathbb{P}}\left[\langle v^{(n)}, \delta^{(n-1)} \rangle\right] = 0$. By Lem. 11, we have

$$\mathbb{E}_{Z_n \sim \mathbb{P}}\left[\|\delta^{(n)}\|_2^2\right] \leq (1 - \mu\eta) \|\delta^{(n-1)}\|_2^2 + \frac{\alpha_1^2 \eta}{\mu} \|\hat{g} - g_0\|_{\mathcal{G}}^2 + \frac{\eta^2}{1 - M\eta} \mathbb{E}_{Z_n \sim \mathbb{P}}\left[\|v^{(n)}\|_2^2\right].$$

Under Asm. 3, $\mathbb{E}_{Z_n \sim \mathbb{P}}\left[\|v^{(n)}\|_2^2\right] \leq K_1 + \kappa_1 \|\delta^{(n)}\|_2^2$, and it follows that

$$\mathbb{E}_{Z_n \sim \mathbb{P}}\left[\|\delta^{(n)}\|_2^2\right] \leq \left(1 - \mu\eta + \frac{\kappa_1 \eta^2}{1 - M\eta}\right) \|\delta^{(n-1)}\|_2^2 + \frac{\alpha_1^2 \eta}{\mu} \|\hat{g} - g_0\|_{\mathcal{G}}^2 + \frac{K_1 \eta^2}{1 - M\eta}.$$

□

Lemma 13. *Suppose that Asm. 3 and Asm. 4 hold. If $\eta < 1/M$, $\theta^{(n)} \in \Theta$, and $\hat{g} \in \mathcal{G}_r(g_0)$, it holds that*

$$\|\delta^{(n)}\|_2^2 \leq (1 - \mu\eta) \|\delta^{(n-1)}\|_2^2 + \frac{\beta_1^2 \eta}{4\mu} \|\hat{g} - g_0\|_{\mathcal{G}}^4 - 2\eta \langle v^{(n)}, \delta^{(n-1)} \rangle + \frac{\eta^2 \|v^{(n)}\|_2^2}{1 - M\eta}.$$

Proof. Under Asm. 3 and Asm. 4,

$$\begin{aligned} D_\theta L(\theta_\star, \hat{g})[\delta^{(n)}] =& D_\theta L(\theta_\star, g_0)[\delta^{(n)}] + D_g D_\theta L(\theta_\star, g_0)[\delta^{(n)}, \hat{g} - g_0] \\ &+ \frac{1}{2} D_g^2 D_\theta L(\theta_\star, \bar{g})[\delta^{(n)}, \hat{g} - g_0, \hat{g} - g_0] \\ =& \frac{1}{2} D_g^2 D_\theta L(\theta_\star, \bar{g})[\delta^{(n)}, \hat{g} - g_0, \hat{g} - g_0] \geq -\frac{\beta_1}{2} \|\delta^{(n)}\|_2 \|\hat{g} - g_0\|_{\mathcal{G}}^2. \end{aligned}$$

The rest of the proof is similar to Lem. 11. □

Corollary 14. *Suppose that Asm. 3 and Asm. 4 hold. If $\eta < 1/M$, $\theta^{(n)} \in \Theta$, and $\hat{g} \in \mathcal{G}_r(g_0)$, it holds that*

$$\mathbb{E}_{Z_n \sim \mathbb{P}}\left[\|\delta^{(n)}\|_2^2\right] \leq \left(1 - \mu\eta + \frac{\kappa_1 \eta^2}{1 - M\eta}\right) \|\delta^{(n-1)}\|_2^2 + \frac{\beta_1^2 \eta}{4\mu} \|\hat{g} - g_0\|_{\mathcal{G}}^4 + \frac{K_1 \eta^2}{1 - M\eta}.$$

Proof. The proof is similar to Cor. 12 using Lem. 13. □

C.3 Proof of Theorem 1

Proof. Let $c(\eta) = \mu - \kappa_1 \eta / (1 - M\eta)$. When $\eta < \mu / (M\mu + \kappa_1)$, we have

$$1 - \mu\eta + \frac{\kappa_1 \eta^2}{1 - M\eta} = 1 - \left(\mu - \frac{\kappa_1 \eta}{1 - M\eta}\right) \eta = 1 - c(\eta)\eta < 1.$$

Under Asm. 3 and by Cor. 12, we have that

$$\begin{aligned} \mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}^n}\left[\|\delta^{(n)}\|_2^2\right] =& \mathbb{E}_{\mathcal{D}_{n-1} \sim \mathbb{P}^{n-1}}\left[\mathbb{E}_{Z_n \sim \mathbb{P}}\left[\|\delta^{(n)}\|_2^2\right]\right] \\ \leq& (1 - c(\eta)\eta)\, \mathbb{E}_{\mathcal{D}_{n-1} \sim \mathbb{P}^{n-1}}\left[\|\delta^{(n-1)}\|_2^2\right] + \frac{\alpha_1^2 \eta}{\mu} \|\hat{g} - g_0\|_{\mathcal{G}}^2 + \frac{K_1 \eta^2}{1 - M\eta} \\ \leq& (1 - c(\eta)\eta)^2\, \mathbb{E}_{\mathcal{D}_{n-2} \sim \mathbb{P}^{n-2}}\left[\|\delta^{(n-1)}\|_2^2\right] \\ &+ \{1 + (1 - c(\eta)\eta)\} \frac{\alpha_1^2 \eta}{\mu} \|\hat{g} - g_0\|_{\mathcal{G}}^2 + \{1 + (1 - c(\eta)\eta)\} \frac{K_1 \eta^2}{1 - M\eta}. \end{aligned}$$

By recursion, it follows that

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n\sim\mathbb{P}^n}\left[\|\delta^{(n)}\|_2^2\right] &\leq (1-c(\eta)\eta)^n\,\|\delta^{(0)}\|_2^2 + \frac{\alpha_1^2\eta}{\mu}\|\hat{g}-g_0\|_{\mathcal{G}}^2\sum_{i=0}^{n-1}(1-c(\eta)\eta)^i \\ &\quad + \frac{K_1\eta^2}{1-M\eta}\sum_{i=0}^{n-1}(1-c(\eta)\eta)^i \\ &\leq (1-c(\eta)\eta)^n\,\|\delta^{(0)}\|_2^2 + \frac{\alpha_1^2}{\mu c(\eta)}\|\hat{g}-g_0\|_{\mathcal{G}}^2 + \frac{K_1\eta}{c(\eta)(1-M\eta)}.\end{aligned}$$

If $\eta \leq \mu/2(M\mu+\kappa_1)$, we have $\mu/2 \leq c(\eta) \leq \mu$. Thus,

$$\mathbb{E}_{\mathcal{D}_n\sim\mathbb{P}^n}\left[\|\delta^{(n)}\|_2^2\right] \leq \left(1-\frac{\mu\eta}{2}\right)^n\|\delta^{(0)}\|_2^2 + \frac{2\alpha_1^2}{\mu^2}\|\hat{g}-g_0\|_{\mathcal{G}}^2 + \frac{4K_1\eta}{\mu}.$$

In addition, if Asm. 4 holds, then by Cor. 14 and using identical proof as above, it follows that for a Neyman orthogonal risk L,

$$\mathbb{E}_{\mathcal{D}_n\sim\mathbb{P}^n}\left[\|\delta^{(n)}\|_2^2\right] \leq \left(1-\frac{\mu\eta}{2}\right)^n\|\delta^{(0)}\|_2^2 + \frac{\beta_1^2}{2\mu^2}\|\hat{g}-g_0\|_{\mathcal{G}}^4 + \frac{4K_1\eta}{\mu}.$$

□

D Orthogonalization with respect to Nuisance

In this section, we establish our orthogonalization method for the possibly infinite-dimensional nuisance introduced in Sec. 3 of the main text. We demonstrate how we construct the orthogonalizing operator in Appx. D.1, and provide all the technical lemmas in Appx. D.2.

D.1 Orthogonalization via Riesz Representation

We consider $\mathcal{G} \equiv (\mathcal{G}, \langle\cdot,\cdot\rangle_{\mathcal{G}})$ as a possibly infinite-dimensional Hilbert space. Recall the *derivative operator* D_g defined as for any $h \in \mathcal{G}$,

$$\mathrm{D}_g \ell(\theta, g; z)[h] = \frac{\mathrm{d}}{\mathrm{d}t} \ell(\theta, g + th; z) \mid_{t=0} .$$

This derivative operator is also known as the Gateaux derivative. We posit the usual assumption as Jordan et al. [2022] that the derivative operator $\mathrm{D}_g \ell(\theta, g; z)$ is linear and continuous in $\mathcal{G}$ for any $(\theta, g, z) \in \Theta \times \mathcal{G}_r(g_0) \times \mathcal{Z}$. We also assume regularity conditions such that $\mathrm{D}_g \mathrm{D}_\theta \ell(\theta, g; z)$ is continuous and $\mathrm{D}_g \mathrm{D}_\theta \ell(\theta, g; z) = \mathrm{D}_\theta \mathrm{D}_g \ell(\theta, g; z)$ at any (θ, g, z).

Since $\mathrm{D}_g \ell(\theta, g; z)$ is linear and continuous, by the Riesz representation theorem, there uniquely exists some $\nabla_g \ell(\theta, g; z) \in \mathcal{G}$ such that for any $g \in \mathcal{G}$,

$$\mathrm{D}_g \ell(\theta, g; z)[g] = \langle \nabla_g \ell(\theta, g; z), g \rangle_{\mathcal{G}}. \tag{82}$$

Lem. 16 shows that the operator $\mathrm{D}_g S_\theta(\theta_\star, g_0)$ is linear and continuous. By Riesz representation theorem, we can define $H_{\theta g} = (H_{\theta g}^{(1)}, \ldots, H_{\theta g}^{(d)}) \in \mathcal{G}^d$ such that for all $g \in \mathcal{G}$,

$$\mathrm{D}_g S_\theta^{(j)}(\theta_\star, g_0)[g] = \langle H_{\theta g}^{(j)}, g \rangle_{\mathcal{G}} \text{ for any } j = 1, \ldots, n. \tag{83}$$

The Hessian operator $H_{gg} : \mathcal{G} \mapsto \mathcal{G}$ is defined as for any $g_1, g_2 \in \mathcal{G}$,

$$\mathrm{D}_g^2 L(\theta_\star, g_0)[g_1, g_2] = \langle H_{gg} g_1, g_2 \rangle_{\mathcal{G}}. \tag{84}$$

We will show in Lem. 17 that H_{gg} uniquely exists and is an self-adjoint and bounded linear operator when $\mathrm{D}_g^2 L(\theta_\star, g_0)$ is bounded and symmetric bilinear. Assuming that H_{gg} is invertible, we define the orthogonalizing operator as

$$\Gamma_0 : \mathcal{G} \mapsto \mathbb{R}^d, \quad [\Gamma_0 g]_j = \langle H_{\theta g}^{(j)}, H_{gg}^{-1} g \rangle_{\mathcal{G}}, \forall g \in \mathcal{G}. \tag{85}$$

We now construct the Neyman orthogonalized (NO) gradient oracle

$$S_{\text{no}}(\theta, g; z) = S_\theta(\theta, g; z) - \Gamma_0 \nabla_g \ell(\theta, g; z).$$

In addition, $\Gamma_0 \nabla_g \ell(\theta, g; z)$ can be written as the derivative in the sense that for each $j = 1, \ldots, d$,

$$\begin{aligned}
[\Gamma_0 \nabla_g \ell(\theta, g; z)]_j &= \langle H_{\theta g}^{(j)}, H_{gg}^{-1} \nabla_g \ell(\theta, g; z) \rangle_{\mathcal{G}} \\
&= \langle H_{gg}^{-1} H_{\theta g}^{(j)}, \nabla_g \ell(\theta, g; z) \rangle_{\mathcal{G}} = \mathrm{D}_g \ell(\theta, g; z)[H_{gg}^{-1} H_{\theta g}^{(j)}].
\end{aligned}$$

That is, the NO gradient oracle can be easily obtain by

$$S_{\text{no}}(\theta, g; z) = S_\theta(\theta, g; z) - \mathrm{D}_g \ell(\theta, g; z)[H_{gg}^{-1} H_{\theta g}]. \tag{86}$$

The following Lemma shows that $S_{\text{no}}(\theta, g; z)$ is Neyman orthogonal at $(\theta_\star, g_0)$.

Lemma 15. *$S_{no}(\theta, g; z)$ is a Neyman orthogonal score at (θ_0, g_0).*

Proof. Since $S_{\text{no}}(\theta, g; z) = S_\theta(\theta, g; z) - \mathrm{D}_g \ell(\theta, g; z)[H_{gg}^{-1} H_{\theta g}]$, for any $h \in \mathcal{G}$,

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}} \left[\mathrm{D}_g S_{\text{no}}(\theta_0, g_0; Z)[h] \right] &= \mathrm{D}_g S_\theta(\theta_0, g_0)[h] - \mathrm{D}_g^2 L(\theta_0, g_0)[H_{gg}^{-1} H_{\theta g}, h] \\
&= \langle H_{\theta g}, h \rangle_{\mathcal{G}} - \langle H_{gg} H_{gg}^{-1} H_{\theta g}, h \rangle_{\mathcal{G}} \\
&= \langle H_{\theta g}, h \rangle_{\mathcal{G}} - \langle H_{\theta g}, h \rangle_{\mathcal{G}} = 0,
\end{aligned}$$

which implies that $S_{\text{no}}(\theta, g; z)$ is Neyman orthogonal at (θ_0, g_0). □

D.2 Technical Lemma

Lemma 16. *$\mathrm{D}_g S_\theta(\theta, g; z) : \mathcal{G} \mapsto \mathbb{R}^d$ and $\mathrm{D}_g S_\theta(\theta, g) : \mathcal{G} \mapsto \mathbb{R}^d$ are linear and continuous in $\mathcal{G}$.*

Proof. The continuity of $\mathrm{D}_g S_\theta(\theta, g; z)$ and $\mathrm{D}_g S_\theta(\theta, g)$ follows from the continuity of $\mathrm{D}_g \mathrm{D}_\theta \ell(\theta, g; z)$. It suffices to prove that $\mathrm{D}_g S_\theta(\theta, g; z)$ is linear. For all $u \in \mathbb{R}^d$, $h_1, h_2 \in \mathcal{G}$,

$$\begin{aligned}
\langle u, \mathrm{D}_g S_\theta(\theta, g; z)[h_1 + h_2] \rangle &= \mathrm{D}_g \langle u, S_\theta(\theta, g; z) \rangle [h_1 + h_2] \\
&= \mathrm{D}_g \mathrm{D}_\theta \ell(\theta, g; z)[u, h_1 + h_2] \\
&= \mathrm{D}_\theta \mathrm{D}_g \ell(\theta, g; z)[h_1 + h_2, u] \\
&= \langle \nabla_\theta \left(\mathrm{D}_g \ell(\theta, g; z)[h_1 + h_2] \right), u \rangle \\
&= \langle \nabla_\theta \left(\mathrm{D}_g \ell(\theta, g; z)[h_1] + \mathrm{D}_g \ell(\theta, g; z)[h_2] \right), u \rangle \\
&= \mathrm{D}_\theta \mathrm{D}_g \ell(\theta, g; z)[h_1, u] + \mathrm{D}_\theta \mathrm{D}_g \ell(\theta, g; z)[h_2, u] \\
&= \mathrm{D}_g \mathrm{D}_\theta \ell(\theta, g; z)[u, h_1] + \mathrm{D}_g \mathrm{D}_\theta \ell(\theta, g; z)[u, h_2] \\
&= \langle u, \mathrm{D}_g S_\theta(\theta, g; z)[h_1] + \mathrm{D}_g S_\theta(\theta, g; z)[h_2] \rangle,
\end{aligned}$$

which implies that

$$\mathrm{D}_g S_\theta(\theta, g; z)[h_1 + h_2] = \mathrm{D}_g S_\theta(\theta, g; z)[h_1] + \mathrm{D}_g S_\theta(\theta, g; z)[h_2].$$

□

Lemma 17. *Suppose that $\mathrm{D}_g^2 L(\theta_\star, g_0)[\cdot, \cdot] : \mathcal{G} \times \mathcal{G} \mapsto \mathbb{R}$ is a bounded and symmetric bilinear form. Then $H_{gg} : \mathcal{G} \mapsto \mathcal{G}$ uniquely exists and is self-adjoint, bounded, and linear.*

Proof. Given $g_1, g_2 \in \mathcal{G}$, since $\mathrm{D}_g^2 L(\theta_\star, g_0)[g_1, \cdot]$ is a bounded linear map from $\mathcal{G}$ to $\mathbb{R}$, by Riesz representation theorem, for any $g_2 \in \mathcal{G}$, there uniquely exists some $Th_1 \in \mathcal{G}$ such that

$$\mathrm{D}_g^2 L(\theta_\star, g_0)[g_1, g_2] = \langle Tg_1, g_2 \rangle_{\mathcal{G}}.$$

Thus, we define the operator $T : \mathcal{G} \mapsto \mathcal{G}$. Note that $\mathrm{D}_g^2 L(\theta_\star, g_0)[\cdot, \cdot]$ is bilinear. For any $a, a' \in \mathbb{R}$, and any $g_1, g_1', g_2 \in \mathcal{G}$, we have

$$\begin{aligned}
\langle T(ag_1 + a'g_1'), g_2 \rangle_{\mathcal{G}} &= \mathrm{D}_g^2 L(\theta_\star, g_0)[ag_1 + a'g_1', g_2] \\
&= a\mathrm{D}_g^2 L(\theta_\star, g_0)[g_1, g_2] + a'\mathrm{D}_g^2 L(\theta_\star, g_0)[g_1', g_2] \\
&= a\langle Tg_1, g_2 \rangle_{\mathcal{G}} + a'\langle Tg_1', g_2 \rangle_{\mathcal{G}} \\
&= \langle aTg_1 + a'Tg_1', g_2 \rangle_{\mathcal{G}},
\end{aligned}$$

which implies T is a linear operator. To show T is bounded, suppose that the norm of the bilinear form $\mathrm{D}_g^2 L(\theta_\star, g_0)$ is bounded by B. Thus, for $Tg_1 \neq 0$,

$$\|Tg_1\|_{\mathcal{G}} = \langle Tg_1, \frac{Tg_1}{\|Tg_1\|_{\mathcal{G}}} \rangle_{\mathcal{G}} \leq \sup_{\|g_2\|_{\mathcal{G}}=1} \langle Tg_1, g_2 \rangle_{\mathcal{G}} \leq \sup_{\|g_2\|_{\mathcal{G}}=1} \left| \mathrm{D}_g^2 L(\theta_\star, g_0)[g_1, g_2] \right| \leq B\|g_1\|_{\mathcal{G}},$$

which implies T is bounded. Note that $\mathrm{D}_g^2 L(\theta_\star, g_0)[\cdot, \cdot]$ is symmetric, we have T being self-adjoint since

$$\langle Tg_1, g_2 \rangle_{\mathcal{G}} = \mathrm{D}_g^2 L(\theta_\star, g_0)[g_1, g_2] = \mathrm{D}_g^2 L(\theta_\star, g_0)[g_2, g_1] = \langle Tg_2, g_1 \rangle_{\mathcal{G}} = \langle g_1, Tg_2 \rangle_{\mathcal{G}}.$$

Finally, we show that T is unique. If there exists some $T' : \mathcal{G} \mapsto \mathcal{G}$ such that for any $g_1, g_2 \in \mathcal{G}$,

$$\langle Tg_1, g_2 \rangle_{\mathcal{G}} = \mathrm{D}_g^2 L(\theta_\star, g_0)[g_1, g_2] = \langle T'g_1, g_2 \rangle_{\mathcal{G}},$$

which implies

$$\langle (T - T')g_1, g_2 \rangle_{\mathcal{G}} = 0 \quad \forall g_1, g_2 \in \mathcal{G}.$$

That is, $T = T'$. We finish the proof by letting $H_{gg} = T$. □

E Convergence Proofs for Orthogonalized Stochastic Gradient

This section is dedicated to demonstrate the OSGD convergence in Thm. 3 of the main text. In Appx. E.1, we give an overview of the OSGD settings, the additional assumptions, and the expected results with the proof outline. We then provide all the technical lemmas in Appx. E.2, and finally prove Thm. 3 in Appx. E.3.

E.1 Overview

Following the same problem settings in Appx. C, we consider the orthogonalized SGD (OSGD) using the estimated NO score $\hat{S}_{\text{no}}$ oracle defined as

$$\hat{S}_{\text{no}}(\theta, g; z) = S_\theta(\theta, g; z) - \hat{\Gamma}\nabla_g \ell(\theta, g; z), \tag{87}$$

where $\hat{\Gamma}$ is an estimator for the orthogonalizing operator defined in (85). Specifically, we consider all continuous linear $\hat{\Gamma} : \mathcal{G} \mapsto \mathbb{R}^d$ for estimating the orthogonalizing operator Γ_0. By Riesz representation theorem, there exists some $\hat{\gamma}^{(j)} \in \mathcal{G}$ for $j = 1, \ldots, d$, such that

$$[\hat{\Gamma} g]_j = \langle \hat{\gamma}^{(j)}, g \rangle_{\mathcal{G}} \text{ for all } g \in \mathcal{G}.$$

For the orthogonalizing operator Γ_0, we define $\gamma_0^{(j)} = H_{gg}^{-1} H_{\theta g}^{(j)}, j = 1, \ldots, d$, such that

$$[\Gamma_0 g]_j = \langle \gamma_0^{(j)}, g \rangle_{\mathcal{G}} \text{ for all } g \in \mathcal{G}.$$

We focus on the OSGD defined below:

$$\theta^{(n)} = \theta^{(n-1)} - \eta \hat{S}_{\text{no}}(\theta^{(n-1)}, \hat{g}; Z_n), \quad \theta^{(0)} \in \Theta. \tag{88}$$

Throughout the section, we take the following notations for simplicity:

$$\delta_{\text{no}}^{(n)} = \theta^{(n)} - \theta_\star, \tag{89}$$

$$\hat{S}_{\text{no}}^{(n)} = \hat{S}_{\text{no}}(\theta^{(n-1)}, \hat{g}; Z_n), \tag{90}$$

$$v_{\text{no}}^{(n)} = \hat{S}_{\text{no}}^{(n)} - \hat{S}_{\text{no}}(\theta^{(n-1)}, \hat{g}). \tag{91}$$

Let $\nabla_g L(\theta, g) = \mathbb{E}_{Z\sim\mathbb{P}}[\nabla_g \ell(\theta, g; Z)]$ and $S_{\text{no}}(\theta, g) = \mathbb{E}_{Z\sim\mathbb{P}}[S_{\text{no}}(\theta, g; Z)]$. We need the following assumptions to establish the convergence result of the OSGD.

Assumption 6. *The following conditions hold:*

(a) First-order optimality: $S_{no}(\theta_\star, g_0) = 0$ *and* $(\hat{\Gamma} - \Gamma_0)\nabla_g L(\theta_\star, g_0) = 0$.

(b) Smoothness and strong convexity: *There exists some* $M_{no}, \mu_{no} > 0$ *such that for all* $\theta \in \Theta$ *and* $g \in \mathcal{G}_r(g_0)$, $\|\nabla_\theta S_{no}(\theta, g)\|_2 \le M_{no}$ *and*

$$\lambda_{\min}(\nabla_\theta S_{no}(\theta, g) + \nabla_\theta S_{no}(\theta, g)^\top) \ge 2\mu_{no}.$$

(c) Second-moment growth: *There exist constants* $K_2, \kappa_2 > 0$ *such that*

$$\mathbb{E}_{Z\sim\mathbb{P}}[(\mathrm{D}_g L(\theta, \bar{g}; Z)[g] - \mathrm{D}_g L(\theta, \bar{g})[g])^2] \le (K_2 + \kappa_2 \|\theta - \theta_\star\|_2^2)\|g\|_{\mathcal{G}}^2.$$

for all $\theta \in \Theta$, $\bar{g} \in \mathcal{G}_r(g_0)$, *and* $g \in \mathcal{G}$.

(d) Second-order smoothness: *There exists a constant* $\alpha_2 > 0$ *such that*

$$\left|\mathrm{D}_g^2 L(\theta, \bar{g})[g_1, g_2]\right| \le \alpha_2 \|g_1\|_{\mathcal{G}} \|g_2\|_{\mathcal{G}} \quad \forall \theta \in \Theta, \bar{g} \in \mathcal{G}_r(g_0), g_1, g_2 \in \mathcal{G},$$
$$\left|\mathrm{D}_\theta \mathrm{D}_g L(\bar{\theta}, g_0)[g, \theta - \theta_\star]\right| \le \alpha_1 \|\theta - \theta_\star\|_2 \|g\|_{\mathcal{G}} \quad \forall \theta, \bar{\theta} \in \Theta, g \in \mathcal{G}.$$

(e) Higher-order smoothness: *There exists a constants* $\beta_2 > 0$ *such that*

$$\|\mathrm{D}_g^2 S_{no}(\theta_\star, \bar{g})[g - g_0, g - g_0]\|_2 \le \beta_2 \|g - g_0\|_{\mathcal{G}}^2 \quad \forall g, \bar{g} \in \mathcal{G}_r(g_0).$$

Asm. 6(a) is necessary for the convergence of the OSGD to $\theta_\star$. When S_θ is Neyman orthogonal at $(\theta_\star, g_0)$, $\Gamma_0 = 0$ is accessible and thus, $S_{\mathrm{no}} = S_\theta$. When S_θ is non-orthogonal, Asm. 6(a) can be satisfied whenever $\nabla_g L(\theta_\star, g_0) = 0$, implying that $(\theta_\star, g_0)$ is a *local* minimizer of $L(\theta, g)$. Asm. 6(b) is related to the Schur complement of the population Hessian. Thus, the hypothetical objective relating to S_{no} inherits its strong convexity from that of the population risk L *w.r.t.* $(\theta, g) \in \Theta \times \mathcal{G}_r(g_0)$ when $\mathcal{G}$ is finite-dimensional; see Boyd and Vandenberghe [2004]. Asm. 6(c) and (d) are exactly analogous to Asm. 3(d) and (e), while Asm. 6(e) is analogous to Asm. 4.

With Asm. 6, we aim to show that the error $\delta_{\mathrm{no}}^{(n)}$ satisfies

$$\mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}^n}[\|\delta_{\mathrm{no}}^{(n)}\|_2^2] \lesssim \left(1 - \frac{\mu_{\mathrm{no}}\eta}{2}\right)^n \|\delta_{\mathrm{no}}^{(0)}\|_2^2 + \|\hat{g} - g_0\|_{\mathcal{G}}^4 + \|\hat{g} - g_0\|_{\mathcal{G}}^2 \cdot \|\hat{\Gamma} - \Gamma_0\|_{\mathrm{Fro}}^2 + \eta. \quad (92)$$

Proof Outline. The proof the (92) follows the following four steps:

1. Upper bound $\|I - \eta\nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g)\|_2$ *w.r.t.* the operator estimation error $\|\hat{\Gamma} - \Gamma_0\|_{\mathrm{Fro}}$.
2. Upper bound $\|\hat{S}_{\mathrm{no}}(\theta_\star, \hat{g})\|_2$ using Neyman orthogonality and the first order optimality.
3. Derive a recursive formula of $\mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}^n}[\|\delta_{\mathrm{no}}^{(n)}\|_2^2]$ from these bounds.
4. Perform the recursion and obtain the final result.

Follow these steps above, we provide technical lemma in Appx. E.2, and then prove our second main result Thm. 3 in Appx. E.3.

Alternatively, the intuition of step 1 also suggests that we should focus on $\hat{\Gamma}$ that lies in the neighborhood of Γ_0 such that $\|\hat{\Gamma} - \Gamma_0\|_{\mathrm{Fro}} \le R$ for a small $R > 0$. Then, instead of assuming Asm. 6(b), we can directly assume that for all $\theta \in \Theta$ and $g \in \mathcal{G}_r(g_0)$, $\|\nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g)\|_2 \le M_{\mathrm{no}}$ and

$$\lambda_{\min}(\nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g) + \nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g)^\top) \ge 2\mu_{\mathrm{no}}.$$

With this assumption, one can still show the same OSGD convergence rate by the identical proof while the constraint of the learning rate η will be simplified.

E.2 Technical Lemma

Lemma 18. *Given $\eta > 0$. For any $\omega > 0$ and $u, v \in \mathbb{R}^d$,*

$$\|u + \eta v\|_2^2 \le (1 + \eta\omega)\|u\|_2^2 + (\eta^2 + \eta\omega^{-1})\|v\|_2^2.$$

Proof. By definition,

$$\|u + \eta v\|_2^2 = \|u\|_2^2 + \eta^2\|v\|_2^2 + 2\eta\langle u, v\rangle \le \|u\|_2^2 + \eta^2\|v\|_2^2 + 2\eta\|u\|_2\|v\|_2.$$

By Young's inequality, for any $\omega > 0$,

$$2\|u\|_2\|v\|_2 \le \omega\|u\|_2^2 + \omega^{-1}\|v\|_2^2.$$

Thus,

$$\|u + \eta v\|_2^2 \le (1 + \eta\omega)\|u\|_2^2 + (\eta^2 + \eta\omega^{-1})\|v\|_2^2.$$

□

Lemma 19. *Suppose that Asm. 6 holds. For all $\theta \in \Theta$ and $g \in \mathcal{G}_r(g_0)$,*

$$\|I - \eta\nabla_\theta \hat{S}_{no}(\theta, g)\|_2^2 \le 1 - 2\eta(\mu_{no} - \alpha_1\|\hat{\Gamma} - \Gamma_0\|_{\mathrm{Fro}}) + 2\eta^2(M_{no}^2 + 2\alpha_1^2\|\hat{\Gamma} - \Gamma_0\|_{\mathrm{Fro}}^2).$$

Proof. Note that

$$\begin{aligned}
&(I - \eta\nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g))(I - \eta\nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g))^\top \\
&\quad = I - \eta\left(\nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g) + \nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g)^\top\right) + \eta^2 \nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g)\nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g)^\top.
\end{aligned}$$

Since $\nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g) = \nabla_\theta S_{\mathrm{no}}(\theta, g) - \nabla_\theta((\hat{\Gamma} - \Gamma_0)\nabla_g L(\theta, g))$, we have

$$\begin{aligned}
&\nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g) + \nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g)^\top \\
&\quad = \nabla_\theta S_{\mathrm{no}}(\theta, g) + \nabla_\theta S_{\mathrm{no}}(\theta, g)^\top - \nabla_\theta((\hat{\Gamma} - \Gamma_0)\nabla_g L(\theta, g)) - \nabla_\theta((\hat{\Gamma} - \Gamma_0)\nabla_g L(\theta, g))^\top \\
&\quad \succcurlyeq 2(\mu_{\mathrm{no}} - \|\nabla_\theta((\hat{\Gamma} - \Gamma_0)\nabla_g L(\theta, g))\|_2)\mathbf{I}.
\end{aligned}$$

We now bound $\|\nabla_\theta(\hat{\Gamma} - \Gamma_0)\nabla_g L(\theta, g)\|_2$. For each $j = 1, \ldots, d$, for any $\theta \in \mathbb{R}^d$,

$$\begin{aligned}
\left|\nabla_\theta((\hat{\Gamma} - \Gamma_0)^{(j)}\nabla_g L(\theta, g))(\theta - \theta_\star)\right| &= \left|\mathrm{D}_\theta \langle \hat{\gamma}^{(j)} - \gamma_0^{(j)}, \nabla_g L(\theta, g)\rangle_{\mathcal{G}}[\theta - \theta_\star]\right| \\
&= \left|\mathrm{D}_\theta \mathrm{D}_g L(\theta, g)[\hat{\gamma}^{(j)} - \gamma_0^{(j)}, \theta - \theta_\star]\right| \\
&\leq \alpha_1 \|\hat{\gamma}^{(j)} - \gamma_0^{(j)}\|_{\mathcal{G}} \|\theta - \theta_\star\|_2.
\end{aligned}$$

Thus,

$$\|\nabla_\theta((\hat{\Gamma} - \Gamma_0)\nabla_g L(\theta, g))(\theta - \theta_\star)\|_2 \leq \alpha_1 \|\hat{\Gamma} - \Gamma_0\|_{\mathrm{Fro}} \|\theta - \theta_\star\|_2,$$

which implies that $\|\nabla_\theta(\hat{\Gamma} - \Gamma_0)\nabla_g L(\theta, g)\|_2 \leq \alpha_1 \|\hat{\Gamma} - \Gamma_0\|_{\mathrm{Fro}}$ and

$$\nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g) + \nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g)^\top \geq 2(\mu_{\mathrm{no}} - \alpha_1 \|\hat{\Gamma} - \Gamma_0\|_{\mathrm{Fro}})\mathbf{I}.$$

Additionally, we have

$$\begin{aligned}
\|\nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g)\nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g)^\top\|_2 &\leq \|\nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g)\|_2^2 \\
&\leq 2\|\nabla_\theta S_{\mathrm{no}}(\theta, g)\|_2^2 + 2\|\nabla_\theta((\hat{\Gamma} - \Gamma_0)\nabla_g L(\theta, g))\|_2^2 \\
&\leq 2M_{\mathrm{no}}^2 + 2\alpha_1^2 \|\hat{\Gamma} - \Gamma_0\|_{\mathrm{Fro}}^2.
\end{aligned}$$

In conclusion, we have

$$\begin{aligned}
&(I - \eta\nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g))(I - \eta\nabla_\theta \hat{S}_{\mathrm{no}}(\theta, g))^\top \\
&\quad \preccurlyeq \left(1 - 2\eta(\mu_{\mathrm{no}} - \alpha_1 \|\hat{\Gamma} - \Gamma_0\|_{\mathrm{Fro}}) + 2\eta^2(M_{\mathrm{no}}^2 + 2\alpha_1^2 \|\hat{\Gamma} - \Gamma_0\|_{\mathrm{Fro}}^2)\right)\mathbf{I}.
\end{aligned}$$

□

Lemma 20. *Suppose that Asm. 6 holds. When $\hat{g} \in \mathcal{G}_r(g_0)$,*

$$\|\hat{S}_{no}(\theta_\star, \hat{g})\|_2^2 \leq \frac{\beta_2^2}{2}\|\hat{g} - g_0\|_{\mathcal{G}}^4 + 2\alpha_2^2 \|\hat{\Gamma} - \Gamma_0\|_{\mathrm{Fro}}^2 \cdot \|\hat{g} - g_0\|_{\mathcal{G}}^2.$$

Proof. Note that

$$\begin{aligned}
\|\hat{S}_{\mathrm{no}}(\theta_\star, \hat{g})\|_2^2 &= \|S_{\mathrm{no}}(\theta_\star, \hat{g}) - (\hat{\Gamma} - \Gamma_0)\nabla_g L(\theta_\star, \hat{g})\|_2^2 \\
&\leq 2\|S_{\mathrm{no}}(\theta_\star, \hat{g})\|_2^2 + 2\|(\hat{\Gamma} - \Gamma_0)\nabla_g L(\theta_\star, \hat{g})\|_2^2 \\
&= 2\|S_{\mathrm{no}}(\theta_\star, \hat{g})\|_2^2 + 2\sum_{j=1}^d \langle \nabla_g L(\theta_\star, \hat{g}), \hat{\gamma}^{(j)} - \gamma_0^{(j)}\rangle_{\mathcal{G}}^2 \\
&= 2\|S_{\mathrm{no}}(\theta_\star, \hat{g})\|_2^2 + 2\sum_{j=1}^d (\mathrm{D}_g L(\theta_\star, \hat{g})[\hat{\gamma}^{(j)} - \gamma_0^{(j)}])^2.
\end{aligned}$$

Since $S_{\mathrm{no}}(\theta_\star, g_0) = 0$ and S_{no} is Neyman orthogonal at $(\theta_\star, g_0)$, we have for some $\bar{g} \in \mathcal{G}_r(g_0)$,

$$\begin{aligned}
S_{\mathrm{no}}(\theta_\star, \hat{g}) &= S_{\mathrm{no}}(\theta_\star, g_0) + \mathrm{D}_g S_{\mathrm{no}}(\theta_\star, g_0)[\hat{g} - g_0] + \frac{1}{2}\mathrm{D}_g^2 S_{\mathrm{no}}(\theta_\star, \bar{g})[\hat{g} - g_0] \\
&= \frac{1}{2}\mathrm{D}_g^2 S_{\mathrm{no}}(\theta_\star, \bar{g})[\hat{g} - g_0],
\end{aligned}$$

which implies

$$\|S_{\text{no}}(\theta_\star, \hat{g})\|_2 \leq \frac{\beta_2}{2}\|\hat{g} - g_0\|_{\mathcal{G}}^2.$$

Similarly, since $(\hat{\Gamma}^{(j)} - \Gamma_0)\nabla_g L(\theta_\star, g_0) = 0$, we have for some $\bar{g}' \in \mathcal{G}_r(g_0)$,

$$\begin{aligned}
\mathrm{D}_g L(\theta_\star, \hat{g})[\hat{\gamma}^{(j)} - \gamma_0^{(j)}] &= \mathrm{D}_g L(\theta_\star, g_0)[\hat{\gamma}^{(j)} - \gamma_0^{(j)}] + \mathrm{D}_g^2 L(\theta_\star, \bar{g}')[\hat{\gamma}^{(j)} - \gamma_0^{(j)}] \\
&= (\hat{\Gamma}^{(j)} - \Gamma_0)\nabla_g L(\theta_\star, g_0) + \mathrm{D}_g^2 L(\theta_\star, \bar{g}')[\hat{\gamma}^{(j)} - \gamma_0^{(j)}] \\
&= \mathrm{D}_g^2 L(\theta_\star, \bar{g}')[\hat{\gamma}^{(j)} - \gamma_0^{(j)}],
\end{aligned}$$

which implies

$$\left|\mathrm{D}_g L(\theta_\star, \hat{g})[\hat{\gamma}^{(j)} - \gamma_0^{(j)}, \hat{g} - g_0]\right| \leq \alpha_2 \|\hat{\gamma}^{(j)} - \gamma_0^{(j)}\|_{\mathcal{G}} \|\hat{g} - g_0\|_{\mathcal{G}}.$$

In conclusion,

$$\|\hat{S}_{\text{no}}(\theta_\star, \hat{g})\|_2^2 \leq \frac{\beta_2^2}{2}\|\hat{g} - g_0\|_{\mathcal{G}}^4 + 2\alpha_2^2 \|\hat{\Gamma} - \Gamma_0\|_{\text{Fro}}^2 \cdot \|\hat{g} - g_0\|_{\mathcal{G}}^2.$$

□

Lemma 21. *Suppose that Asm. 3 and Asm. 6 holds. Then*

$$\mathbb{E}_{Z_n \sim \mathbb{P}}\left[\|v_{no}^{(n)}\|_2^2\right] \leq 2(K_1 + K_2\|\hat{\Gamma}\|_{\text{Fro}}^2) + 2(\kappa_1 + \kappa_2\|\hat{\Gamma}\|_{\text{Fro}}^2)\|\delta_{no}^{n-1}\|_2^2.$$

Proof. Note that $v_{\text{no}}^{(n)} = \hat{S}_{\text{no}}(\theta^{(n-1)}, \hat{g}; Z_n) - \hat{S}_{\text{no}}(\theta^{(n-1)}, \hat{g})$.

$$\begin{aligned}
\mathbb{E}_{Z_n \sim \mathbb{P}}\left[\|v_{\text{no}}^{(n)}\|_2^2\right] &\leq 2\mathbb{E}_{Z_n \sim \mathbb{P}}\left[\|S_\theta(\theta^{(n-1)}, \hat{g}; Z_n) - S_\theta(\theta^{(n-1)}, \hat{g})\|_2^2\right] \\
&\quad + 2\mathbb{E}_{Z_n \sim \mathbb{P}}\left[\|\hat{\Gamma}\nabla_g \ell(\theta^{(n-1)}, \hat{g}; Z_n) - \hat{\Gamma}\nabla_g L(\theta^{(n-1)}, \hat{g})\|_2^2\right].
\end{aligned}$$

By Asm. 3,

$$\mathbb{E}_{Z_n \sim \mathbb{P}}\left[\|S_\theta(\theta^{(n-1)}, \hat{g}; Z_n) - S_\theta(\theta^{(n-1)}, \hat{g})\|_2^2\right] \leq K_1 + \kappa_1\|\delta_{\text{no}}^{n-1}\|_2^2.$$

Since

$$\begin{aligned}
&\mathbb{E}_{Z_n \sim \mathbb{P}}\left[\|\hat{\Gamma}\nabla_g \ell(\theta^{(n-1)}, \hat{g}; Z_n) - \hat{\Gamma}\nabla_g L(\theta^{(n-1)}, \hat{g})\|_2^2\right] \\
&= \sum_{j=1}^d \mathbb{E}_{Z_n \sim \mathbb{P}}\left[(\mathrm{D}_g \ell(\theta^{(n-1)}, \hat{g}; Z_n)[\hat{\gamma}^{(j)}] - \mathrm{D}_g L(\theta^{(n-1)}, \hat{g})[\hat{\gamma}^{(j)}])^2\right] \\
&\leq \sum_{j=1}^d (K_2 + \kappa_2\|\theta - \theta_\star\|_2^2)\|\hat{\gamma}^{(j)}\|_{\mathcal{G}}^2 = (K_2 + \kappa_2\|\theta - \theta_\star\|_2^2)\|\hat{\Gamma}\|_{\text{Fro}}^2.
\end{aligned}$$

In conclusion,

$$\mathbb{E}_{Z_n \sim \mathbb{P}}\left[\|v_{\text{no}}^{(n)}\|_2^2\right] \leq 2(K_1 + K_2\|\hat{\Gamma}\|_{\text{Fro}}^2) + 2(\kappa_1 + \kappa_2\|\hat{\Gamma}\|_{\text{Fro}}^2)\|\delta_{\text{no}}^{n-1}\|_2^2.$$

□

E.3 Proof of Theorem 3

Proof. Since $\theta^{(n)} = \theta^{(n-1)} - \eta\hat{S}_{\text{no}}^{(n)}$, by Taylor's theorem we have that for some $\bar{\theta}^{(n-1)}$,

$$\begin{aligned}
\delta_{\text{no}}^{(n)} &= \delta_{\text{no}}^{(n-1)} - \eta(\hat{S}_{\text{no}}(\theta^{(n-1)}, \hat{g}) - \hat{S}_{\text{no}}(\theta_\star, \hat{g})) - \eta\hat{S}_{\text{no}}(\theta_\star, \hat{g}) - \eta v_{\text{no}}^{(n)} \\
&= (I - \eta\nabla_\theta \hat{S}_{\text{no}}(\bar{\theta}^{(n-1)}, \hat{g}))\delta_{\text{no}}^{(n-1)} - \eta\hat{S}_{\text{no}}(\theta_\star, \hat{g}) - \eta v_{\text{no}}^{(n)}.
\end{aligned}$$

Note that $\mathbb{E}_{Z_n\sim\mathbb{P}}\left[v_{\text{no}}^{(n)}\right]=0$. Take the expectation of the squared norm of both sides *w.r.t.* Z_n and we have

$$\mathbb{E}_{Z_n\sim\mathbb{P}}\left[\|\delta_{\text{no}}^{(n)}\|_2^2\right]=\|(I-\eta\nabla_\theta\hat{S}_{\text{no}}(\bar{\theta}^{(n-1)},\hat{g}))\delta_{\text{no}}^{(n-1)}-\eta\hat{S}_{\text{no}}(\theta_\star,\hat{g})\|_2^2+\eta^2\mathbb{E}_{Z_n\sim\mathbb{P}}\left[\|v_{\text{no}}^{(n)}\|_2^2\right].$$

By Lem. 18, Lem. 19, and Lem. 20, for any $\omega>0$,

$$\begin{aligned}
&\|(I-\eta\nabla_\theta\hat{S}_{\text{no}}(\bar{\theta}^{(n-1)},\hat{g}))\delta_{\text{no}}^{(n-1)}-\eta\hat{S}_{\text{no}}(\theta_\star,\hat{g})\|_2^2\\
&\quad\le(1+\eta\omega)\|(I-\eta\nabla_\theta\hat{S}_{\text{no}}(\bar{\theta}^{(n-1)},\hat{g}))\delta_{\text{no}}^{(n-1)}\|_2^2+(\eta^2+\eta\omega^{-1})\|\hat{S}_{\text{no}}(\theta_\star,\hat{g})\|_2^2\\
&\quad\le(1+\eta\omega)\left(1-2\eta(\mu_{\text{no}}-\alpha_1\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}})+2\eta^2(M_{\text{no}}^2+2\alpha_1^2\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}}^2)\right)\|\delta_{\text{no}}^{(n-1)}\|_2^2\\
&\qquad+(\eta^2+\eta\omega^{-1})\left(\frac{\beta_2^2}{2}\|\hat{g}-g_0\|_{\mathcal{G}}^4+2\alpha_2^2\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}}^2\cdot\|\hat{g}-g_0\|_{\mathcal{G}}^2\right).
\end{aligned}$$

Set $\omega=\mu_{\text{no}}$. For $\eta\le 2/\mu_{\text{no}}$, we have

$$\begin{aligned}
&(1+\eta\omega)\left(1-2\eta(\mu_{\text{no}}-\alpha_1\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}})+2\eta^2(M_{\text{no}}^2+2\alpha_1^2\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}}^2)\right)\\
&=1-(\mu_{\text{no}}-2\alpha_1\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}})\eta-2(\mu_{\text{no}}^2-\mu_{\text{no}}\alpha_1\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}})\eta^2\\
&\quad+2(1+\eta\mu_{\text{no}})(M_{\text{no}}^2+2\alpha_1^2\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}}^2)\eta^2\\
&\le 1-(\mu_{\text{no}}-2\alpha_1\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}})\eta\\
&\quad+\left(6M_{\text{no}}^2-2\mu_{\text{no}}^2+2\mu_{\text{no}}\alpha_1\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}}+12\alpha_1^2\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}}^2\right)\eta^2\\
&=:1-b(\eta)\eta,
\end{aligned}$$

where

$$b(\eta)=\mu_{\text{no}}-2\alpha_1\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}}-(6M_{\text{no}}^2-2\mu_{\text{no}}^2+2\mu_{\text{no}}\alpha_1\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}}+12\alpha_1^2\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}}^2)\eta.$$

By Lem. 21,

$$\mathbb{E}_{Z_n\sim\mathbb{P}}\left[\|v_{\text{no}}^{(n)}\|_2^2\right]\le 2(K_1+K_2\|\hat{\Gamma}\|_{\text{Fro}}^2)+2(\kappa_1+\kappa_2\|\hat{\Gamma}\|_{\text{Fro}}^2)\|\delta_{\text{no}}^{n-1}\|_2^2.$$

It follows that

$$\begin{aligned}
\mathbb{E}_{Z_n\sim\mathbb{P}}[\|\delta_{\text{no}}^{(n)}\|_2^2]&\le(1-b(\eta)\eta+2(\kappa_1+\kappa_2\|\hat{\Gamma}\|_{\text{Fro}}^2)\eta^2)\|\delta_{\text{no}}^{(n-1)}\|_2^2\\
&\quad+\frac{3\eta}{\mu_{\text{no}}}\left(\frac{\beta_2^2}{2}\|\hat{g}-g_0\|_{\mathcal{G}}^4+2\alpha_2^2\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}}^2\cdot\|\hat{g}-g_0\|_{\mathcal{G}}^2\right)+2(K_1+K_2\|\hat{\Gamma}\|_{\text{Fro}}^2)\eta^2.
\end{aligned}$$

Thus, it is clear that when $\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}}\le\mu_{\text{no}}/(4\alpha_1)$ and the learning rate satisfies

$$\begin{aligned}
\eta&\le\min\left\{\frac{2}{\mu_{\text{no}}},\frac{\mu_{\text{no}}-4\alpha_1\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}}}{12M_{\text{no}}^2-3\mu_{\text{no}}^2/2+4(\kappa_1+\kappa_2\|\hat{\Gamma}\|_{\text{Fro}}^2)}\right\}\\
&=\frac{\mu_{\text{no}}-4\alpha_1\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}}}{12M_{\text{no}}^2-3\mu_{\text{no}}^2/2+4(\kappa_1+\kappa_2\|\hat{\Gamma}\|_{\text{Fro}}^2)},
\end{aligned}\tag{93}$$

we have $1-\mu_{\text{no}}\eta/2\ge 0$ and

$$1-b(\eta)\eta+2(\kappa_1+\kappa_2\|\hat{\Gamma}\|_{\text{Fro}}^2)\eta^2\le 1-\frac{\mu_{\text{no}}\eta}{2}.$$

When η satisfies (93), it follows that

$$\begin{aligned}
\mathbb{E}_{Z_n\sim\mathbb{P}}[\|\delta_{\text{no}}^{(n)}\|_2^2]&\le\left(1-\frac{\mu_{\text{no}}\eta}{2}\right)\|\delta_{\text{no}}^{(n-1)}\|_2^2\\
&\quad+\frac{3\eta}{\mu_{\text{no}}}\left(\frac{\beta_2^2}{2}\|\hat{g}-g_0\|_{\mathcal{G}}^4+2\alpha_2^2\|\hat{\Gamma}-\Gamma_0\|_{\text{Fro}}^2\cdot\|\hat{g}-g_0\|_{\mathcal{G}}^2\right)+2(K_1+K_2\|\hat{\Gamma}\|_{\text{Fro}}^2)\eta^2.
\end{aligned}$$

Finally, perform the same recursion in Appx. C.3 and we have

$$\mathbb{E}_{\mathcal{D}_n\sim\mathbb{P}^n}[\|\delta_{\mathrm{no}}^{(n)}\|_2^2] \le \left(1-\frac{\mu_{\mathrm{no}}\eta}{2}\right)^n \|\delta_{\mathrm{no}}^{(0)}\|_2^2 + \frac{4(K_1+K_2\|\hat{\Gamma}\|_{\mathrm{Fro}}^2)\eta}{\mu_{\mathrm{no}}} + \frac{3}{\mu_{\mathrm{no}}^2}\left(\beta_2^2\|\hat{g}-g_0\|_{\mathcal{G}}^4 + 4\alpha_2^2\|\hat{\Gamma}-\Gamma_0\|_{\mathrm{Fro}}^2\cdot\|\hat{g}-g_0\|_{\mathcal{G}}^2\right).$$

□

F Detailed Discussion

This section provides details on comparisons and remarks following the statements of the main results in the main text, and details on the discussions summarized in Sec. 4 from the main text. In Appx. F.1, we compare our results to generic state-of-the-art results on biased SGD, that is SGD with errors in the stochastic gradients. In Appx. F.2, we discuss different orthogonalization method in orthogonal statistical learning. In Appx. F.3, we discuss how to interleave the target and nuisance estimation. In Appx. F.4, we describe the connection between our orthogonalized gradient to variance reduction method in the Monte Carlo estimation literature. In Appx. F.5, we discuss the double robustness of SGD for dose-response estimation.

F.1 Comparison to Biased SGD

There are several ways to think about the bias induced by using an imperfect estimate $\hat{g}$ as opposed to the true nuisance $g_0 \in \mathcal{G}$. For the sake of discussion, we will define $L(\cdot, g_0)$ and $L(\cdot, \hat{g})$ as the "original objective" and "shifted objective", respectively. Accordingly, we will call $\theta_\star$ the "original minimizer" and denote by

$$\hat{\theta}_\star = \arg\min_{\theta \in \Theta} L(\theta, \hat{g}). \tag{94}$$

the "shifted minimizer". The bias can be measured in terms of (i) the error $\|\hat{\theta}_\star - \theta_\star\|_2^2$ between the original and shifted minimizers, (ii) the uniform error $\sup_\theta |L(\theta, g_0) - L(\theta, \hat{g})|$ between the original and shifted objectives, and (iii) some summary of the gradient bias

$$\mathbb{E}_{Z_t \sim \mathbb{P}} \left[S(\theta^{(t-1)}, \hat{g}; Z_t) \right] - \nabla_\theta L(\theta^{(t)}, g_0), \tag{95}$$

of the oracle S (a vector-valued quantity) for step $t = 1, \ldots, n$ of the algorithm. Whether one appeals to (i) or (ii) depends on whether the convergence guarantees are stated in terms of iterate convergence or function value convergence; because we analyze convergence of iterates, our discussion will cover (i) and (iii).

On (i), one applies the decomposition

$$\|\theta^{(n)} - \theta_\star\|_2^2 \leq 2\|\theta^{(n)} - \hat{\theta}_\star\|_2^2 + 2\|\hat{\theta}_\star - \theta_\star\|_2^2,$$

and plugs an analysis of unbiased SGD from the current literature for the $\|\theta^{(n)} - \hat{\theta}_\star\|_2^2$ term. The purpose of this substitution is to check how our theoretical results align with the known results on *unbiased SGD*.

Bach and Moulines [2011, Thm. 1] show that for constant learning rate $\eta = \mathcal{O}(\mu/M^2)$, the iterate $\theta^{(n)}$ satisfies

$$\mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}^n}[\|\theta^{(n)} - \theta_\star\|_2^2] \lesssim \exp\left(-\frac{\mu\eta n}{2}\right) \|\theta^{(0)} - \theta_\star\|_2^2 + \frac{K_1 \eta}{\mu} + \|\hat{\theta}_\star - \theta_\star\|_2^2. \tag{96}$$

Cutler et al. [2023, Thm. 3] demonstrate that with the learning rate $\eta = \mathcal{O}(1/M)$, the iterates would satisfy the following bound:

$$\mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}^n}[\|\theta^{(n)} - \theta_\star\|_2^2] \lesssim \left(1 - \frac{\mu\eta}{2}\right)^n \|\theta^{(0)} - \theta_\star\|_2^2 + \frac{K_1 \eta}{\mu} + \|\hat{\theta}_\star - \theta_\star\|_2^2 \tag{97}$$

In addition, Cutler et al. [2023, Thm. 6] provide the high probability bound of $\theta^{(n)}$ that for $\eta = \mathcal{O}(1/M)$, with probability at least $1 - \delta$,

$$\|\theta^{(n)} - \theta_\star\|_2^2 \lesssim \left(1 - \frac{\mu\eta}{2}\right)^n \|\theta^{(0)} - \theta_\star\|_2^2 + \frac{K_1 \eta}{\mu} \log\left(\frac{e}{\delta}\right) + \|\hat{\theta}_\star - \theta_\star\|_2^2 \tag{98}$$

All of these bounds essentially agree, as we may apply $(1 - \mu\eta/2) \leq \exp(-\mu\eta/2)$. In comparison to our Thm. 1, our bias term is stated directly in terms of the nuisance error $\|\hat{g} - g_0\|_{\mathcal{G}}^2$. This can be viewed as a refinement of the less transparent bias measurement $\|\hat{\theta}_\star - \theta_\star\|_2^2$. Moreover, although (96)–(98) are of the same order as our results in Thm. 1 when the true nuisance g_0 is

available, all of the three bounds above require $\kappa_1 = 0$ in Asm. 3(d). In this case, provide (97) and (98) use a learning rate of the order $\mathcal{O}(1/M)$ (whereas the learning rate of (96) encounters an additional condition number M/μ). Our learning rate recovers $\mathcal{O}(1/M)$ when $\kappa_1 = 0$, and adapts via the setting $\eta = \mathcal{O}(\mu/(M\mu + \kappa_1))$ when $\kappa_1 > 0$. Finally, the high probability bound (98) requires a stronger assumption in the sense that $S_\theta(\theta, g; Z) - S_\theta(\theta, g)$ is sub-Gaussian with uniform parameter $K_1/2$ for all $\theta \in \Theta$ and $g \in \mathcal{G}_r(g_0)$.

Returning to the bias in the stochastic gradient oracle (95), this case is handled quite generally in Demidovich et al. [2023]. Their "ABC assumption" considers constants $A, B, C, b, c \geq 0$ such that the inequalities

$$\langle \nabla L(\theta, g_0), \mathbb{E}_{Z\sim\mathbb{P}}[S_\theta(\theta, \hat{g}; Z)] \rangle \geq b\|\nabla L(\theta, g_0)\|_2^2 + c \tag{99}$$

$$\mathbb{E}_{Z\sim\mathbb{P}}\|S_\theta(\theta, \hat{g}; Z)\|_2^2 \leq 2A\left(L(\theta, g_0) - L(\theta_\star, g_0)\right) + B\|\nabla L(\theta, g_0)\|_2^2 + C \tag{100}$$

$$A + M\left(B + 1 - 2b\right) < \frac{\mu}{2} \tag{101}$$

hold for all $\theta \in \mathbb{R}^d$ (where the expectations are conditional on any randomness in $\hat{g}$).[3] The bias is really captured in the first of the three inequalities, whereas the third inequality places conditions on the parameters of the problem that are not in the hands of the algorithm user. By strong convexity, our Asm. 3(d) satisfies (100) with $A = \kappa_1/\mu$, $B = 0$, $C = K_1$. The resulting convergence guarantee [Demidovich et al., 2023, Thm. 5] gives

$$\mathbb{E}_{\mathcal{D}_n\sim\mathbb{P}^n}[\|\theta^{(n)} - \theta_\star\|_2^2] \lesssim \left(1 - \frac{\mu\eta}{2}\right)^n \|\theta^{(0)} - \theta_\star\|_2^2 + \frac{\eta K_1}{\mu} + \frac{K_1 + c}{\mu^2}.$$

for a learning rate set as

$$\eta \leq \min\left\{\frac{2}{\mu}, \frac{\mu - 2(\kappa_1/\mu + M(1-2b))}{\kappa_1}\right\}.$$

We would like to check how our theoretical results compare with the above application of generic results on *biased SGD*. In our setting, the condition (101) reads $\kappa_1/\mu + M(1-2b) < \mu/2$. We assume neither this nor (99), and in our Thm. 1, replace the $\frac{K_1+c}{\mu^2}$ term with a bias term that depends directly on the nuisance error $\|\hat{g} - g_0\|_{\mathcal{G}}^2$, either in the nuisance sensitive regime, or in the nuisance insensitive regime.

F.2 Discussion of Full-sample Orthogonal Statistical Learning and Related Methods

Comparison of Orthogonalizing Operators. Constructing orthogonal losses or scores has been widely studied in semiparametric inference, hypothesis testing, and machine learning. In semiparametric statistics, such constructions often rely on the efficient influence function, which characterizes the asymptotic efficient estimation bound; see Bickel et al. [1993, Ch. 3], Tsiatis [2006, Ch. 4], Van der Vaart [2000, Ch. 25], Luedtke and Chung [2024]. In hypothesis testing, orthogonal scores were used by Neyman [1959, 1979] and Ferguson [2014] to guarantee the local unbiasedness of specific tests based on the likelihood with finite-dimensional nuisance. In machine learning, the construction of orthogonal scores was latter extended to non-likelihood losses in Wooldridge [1991] and Liu et al. [2022], which aligns with our construction limited to the finite-dimensional nuisance case. Recent work of orthogonalization in machine learning with infinite-dimensional nuisance relies on the approach named *concentrating-out* [Newey, 1994, Chernozhukov et al., 2018a]. However, although all these constructions produce Neyman orthogonal losses or scores, none of them consider the stochastic design. Our work is complementary to these, providing non-asymptotic guarantees for stochastic optimization.

Although these constructions might lead to different orthogonal scores, they can be the same at both the target and the true nuisance. Specifically, when ℓ is the negative log-likelihood and $\mathcal{G} = \mathbb{R}^k$, the concentrating-out approach and our NO gradient oracle S_{no} both produce the efficient score in the semiparametric theory literature; see Newey [1994, Page 1359], Van der Vaart [2000, Ch. 25.4], and Tsiatis [2006, Def. 8]. This identity can happen for infinite-dimensional nuisances as well. As

[3] The third inequality is actually $A + M\left(B + 1 - 2b\right) < \mu$, but the constant $(1/2)$ to make the resulting bound more comparable, in that their bound can only improve over ours for the stronger inequality.

an example, consider the partial linear model from Appx. B.1.2, where the non-orthogonal loss is defined as

$$\ell(\theta, g; z) = \frac{1}{2}(y - g(w) - \langle \theta, x \rangle)^2.$$

Chernozhukov et al. [2018a, Sec. 2.2.2] showed that the concentrating-out approach would produce an orthogonal score under the concentrated-out nuisance $\varphi_0(\theta) = Z \mapsto \mathbb{E}_{\mathbb{P}}[Y - \langle \theta, X \rangle \mid W]$ as

$$S(\theta_\star, \varphi_0(\theta_\star); Z) = -(X - \mathbb{E}_{\mathbb{P}}[X \mid W])(Y - \mathbb{E}_{\mathbb{P}}[Y \mid W] - \langle \theta_\star, X - \mathbb{E}_{\mathbb{P}}[X \mid W] \rangle).$$

On the other hand, it is easy to verify that $H_{gg} = \mathbf{I}$, $H_{\theta g} = \mathbb{E}_{\mathbb{P}}[X \mid W]$, and $\Gamma_0 \nabla_g \ell(\theta, g; z) = \mathrm{D}_g \ell(\theta, g; z)[H_{gg}^{-1} H_{\theta g}]$, which implies that our orthogonal gradient oracle S_{no} in (11) has the same form under the target $\theta_\star$ and true nuisance $g_0(W) = \mathbb{E}_{\mathbb{P}}[Y - \langle \theta_\star, X \rangle \mid W]$:

$$S_{\text{no}}(\theta_\star, g_0; Z) = -(X - \mathbb{E}_{\mathbb{P}}[X \mid W])(Y - \mathbb{E}_{\mathbb{P}}[Y \mid W] - \langle \theta_\star, X - \mathbb{E}_{\mathbb{P}}[X \mid W] \rangle).$$

Comparison with Debiased Machine Learning. In machine learning, debiasing typically refers to reducing the impact of model selection error on the parameter or quantity of interest. In particular, mitigating the bias introduced by nuisance estimation is known as *debiased machine learning* (DML), which has been recently studied by van der Laan et al. [2011], Shi et al. [2019], Chernozhukov et al. [2024], van der Laan et al. [2025]. Some of the calculations used by DML estimators have been shown to be amenable to computerization, simplifying their construction [Carone et al., 2019, Ichimura and Newey, 2022, Luedtke, 2024]. Statistical learning methods that use debiasing are also called *orthogonal statistical learning* (OSL) and have been studied in Foster and Syrgkanis [2023], Liu et al. [2022], Zadik et al. [2018]. While the earlier studies focus on the empirical risk minimization, our paper provide a stochastic approximation method in DML/OSL and establish the convergence rate of the debiased estimation.

To strengthen the debiasing effect, one possible approach is to consider the higher-order Neyman orthogonality. If the loss function satisfies the k-th order orthogonality at $(\theta_\star, g_0)$, Zadik et al. [2018, Cor. 4] show that we only need the nuisance estimator to converge at rate $\mathcal{O}_p(n^{-\frac{1}{2(k+1)}})$ to have the nuisance effect in the order of $\mathcal{O}_p(n^{-1})$, which aligns with the nuisance insensitive rate in Thm. 1, where $k = 1$ and the nuisance effect $\|\hat{g} - g_0\|_{\mathcal{G}}^4 = \mathcal{O}_p(n^{-1})$ when $\|\hat{g} - g_0\|_{\mathcal{G}} = \mathcal{O}_p(n^{-1/4})$. Similar improvements in sensitivity to nuisance estimation rates have been developed previously using higher-order influence functions [Pfanzagl, 1985, Robins et al., 2008].

For a range of problems, debiasing methods often lead to cross-product estimations consisting of two nuisance estimators [Rotnitzky et al., 2021]. Such remainders frequently result from orthogonalization procedures used in missing data problems and causal inference problems [Robins et al., 1994, Robins and Rotnitzky, 1995, Laan and Robins, 2003]. Chernozhukov et al. [2024] consider cases where $Z = (W, Y)$, $g_0(W) = \mathbb{E}_{\mathbb{P}}[Y \mid W]$, and the target can be written as the averaged moment of the form

$$\theta_\star = \mathbb{E}_{\mathbb{P}}[m(g_0; W)],$$

where $\mathbb{E}[m(g; W)] : \mathcal{G} \times \mathcal{W} \mapsto \mathbb{R}^d$ is a continuous linear functional of $g : \mathcal{W} \mapsto \mathbb{R}$. By Riesz representation theorem, there uniquely exists a Riesz representer (RR) $g_0^{\text{RR}} \in \mathcal{G}$ such that $\mathbb{E}_{\mathbb{P}}[m(g; W)] = \mathbb{E}_{\mathbb{P}}[g_0^{\text{RR}}(W) g(W)]$. Then the debiased score for estimating $\theta_\star$ is defined as

$$S(\theta, g; Z) = m(g; W) + g_0^{\text{RR}}(W)(Y - g(W)) - \theta.$$

The debiasing effect on the nuisance turns out depending on the cross-product $\|\hat{g}^{\text{RR}} - g_0^{\text{RR}}\|_{\mathcal{G}} \cdot \|\hat{g} - g_0\|_{\mathcal{G}}$. Specifically, Chernozhukov et al. [2024, Asm. 4] requires the cross product to be in the order $\mathcal{O}_p(n^{-1/2})$ to construct $\mathcal{O}_p(n^{-1})$ consistent target estimator. This aligns with the cross-product $\|\hat{g} - g_0\|_{\mathcal{G}} \cdot \|\hat{\Gamma} - \Gamma_0\|_{\text{Fro}}$ in Thm. 3 where the same requirement needs to be satisfied to obtain a $\mathcal{O}_p(n^{-1/2})$ consistent estimator. However, Thm. 3 also has a second, non-cross-product remainder $\|\hat{g} - g_0\|_{\mathcal{G}}^4$ that will only be small if $\hat{g}$ approximates g_0, making it so that our consistency guarantee is robust to misspecification of $\hat{\Gamma}$, but not to misspecification of $\hat{g}$.

F.3 Discussion of Interleaving Target and Nuisance Estimation

To propose the interleaving approach, we consider the case where we learn the nuisance from the $\mathcal{W}$-valued data $W = (U, V)$ from a probability measure $\mathbb{Q}$. We assume that the true nuisance g_0 satisfies $g_0 : \mathcal{U} \mapsto \mathbb{R}$ and is the minimizer of the mean squared error over $\mathcal{G}$:

$$g_0 = \arg\min_{g \in \mathcal{G}} \mathbb{E}_{\mathbb{Q}} \left[(g(U) - V)^2 \right].$$

Suppose that we observe another data stream $W_1, \ldots, W_m$ sampled i.i.d. from $\mathbb{Q}$, and that $\mathcal{S}_m = \{W_1, \ldots, W_m\}$ is independent of the parameter stream $\mathcal{D}_n$. We define the sigma algebra $\mathcal{H}_m = \sigma(\mathcal{S}_m), m \geq 1$ as the nuisance filtration and the sigma algebra $\mathcal{F}_{m,t} = \sigma(\mathcal{S}_m \cup \mathcal{D}_{(m-1)n+t}), 0 \leq t \leq n$ as the parameter filtration. We assume that there are two stochastic processes $\hat{g}^{(m)}, m \geq 1$ adapted to $\mathcal{H}_m$ and $\theta^{(m,t)}, 0 \leq t \leq n$ adapted to $\mathcal{F}_{m,t}$, to which we refer as the nuisance estimator and the parameter estimator, respectively. Intuitively, this means that the nuisance estimator $\hat{g}^{(m)}$ can be updated now instead of being the fixed $\hat{g}$, and the parameter estimator $\theta^{(m,t)}$ can be updated n times between every two nuisance updates. Specifically, we use SGD as the parameter estimator. We define $\theta^{(0,n)} = \theta^{(0)} \in \Theta$ and $\theta^{(i,0)} = \theta^{(i-1,n)}$ for $1 \leq i \leq m$, and produce the sequence $\theta^{(i,1)}, \ldots, \theta^{(i,n)}$ using n steps of the SGD update (8) initialized at $\theta^{(i,0)}$.

Under Non-orthogonality. Consider the case that $\mathcal{G}$ is a reproducing kernel Hilbert space (RKHS) with kernel $k(\cdot, \cdot)$. To obtain a sequence of nuisance estimator $\hat{g}^{(m)}$ on $\mathcal{H}_m$, one possible approach is to adopt the non-parametric stochastic approximation. With the assumption that the eigenvalues $(\lambda_j)_{j \geq 1}$ of covariance operator $\mathbb{E}_{\mathbb{Q}}[k(W, \cdot) \otimes k(W, \cdot)]$ decay polynomially at order $j^{-\alpha}$, Dieuleveut and Bach [2016, Cor. 3] suggests that the non-parametric stochastic approximation $\hat{g}^{(m)}$ satisfies for some $C > 0$,

$$\xi_m := \mathbb{E}_{\mathcal{S}_m \sim \mathbb{Q}^m} \left[\|\hat{g}^{(m)} - g_0\|_{\mathcal{G}}^2 \right] \leq C m^{-\frac{2\alpha - 1}{2\alpha}}. \tag{102}$$

This leads to the following nuisance sensitive rate for non-Neyman orthogonal losses.

Proposition 22. *Suppose that $\hat{g}^{(m)}$ satisfies (102) and that $\hat{g}^{(m)} \in \mathcal{G}_r(g_0)$ and $\theta^{(m,t)} \in \Theta$ almost surely for all $m \geq 1$ and $0 \leq t \leq n$. Under the same conditions to Thm. 1, it holds that*

$$\mathbb{E}_{\mathcal{D}_{mn} \cup \mathcal{S}_m \sim \mathbb{P}^{mn} \otimes \mathbb{Q}^m} [\|\theta^{(m,n)} - \theta_\star\|_2^2] \lesssim \left(1 - \frac{\mu\eta}{2}\right)^{mn} \|\theta^{(0)} - \theta_\star\|_2^2 + m \exp\left(-\frac{\mu\eta n m}{4}\right) + (m^{-\frac{2\alpha-1}{2\alpha}} + \eta)((\eta n)^{-1} + 1).$$

In addition, when $(\eta n)^{-1} = \mathcal{O}(1)$, it holds that

$$\mathbb{E}_{\mathcal{D}_{mn} \cup \mathcal{S}_m \sim \mathbb{P}^{mn} \otimes \mathbb{Q}^m} [\|\theta^{(m,n)} - \theta_\star\|_2^2] \lesssim \left(1 - \frac{\mu\eta}{2}\right)^{mn} \|\theta^{(0)} - \theta_\star\|_2^2 + m^{-\frac{2\alpha-1}{2\alpha}} + n^{-1} + \eta.$$

The proof is provided in Appx. F.6. Prop. 22 demonstrates that interleaving the target and nuisance estimation allows $\eta \asymp n^{-1}$ since the nuisance update iterations guarantees the shrinking of the term $(1 - \mu\eta/2)^{mn}$ in this case. This is an improvement to Thm. 1 where η should satisfy $(\eta n)^{-1} = o(1)$ to ensure $(1 - \mu\eta/2)^n$ shrinking to zero.

Under Orthogonalized SGD. To establish a similar probability bound for OSGD, we assume that the orthogonalizing operator Γ_0 can be written as the minimizer of the following program:

$$\Gamma_0 = \arg\min_{\Gamma \in \mathcal{G}_*^d} \mathbb{E}_{\mathbb{P}} \left[\|S_\theta(\theta_\star, g_0; Z) - \Gamma \nabla_g \ell(\theta_\star, g_0; Z)\|_2^2 \right],$$

where $\mathcal{G}_*$ is the dual space of $\mathcal{G}$. When d is fixed, we assume that Γ_0 can be estimated (coordinate-wisely) from the data stream $\mathcal{S}_m$ using the stochastic approximation of Dieuleveut and Bach [2016], which leads to a sequence of operator estimators $\hat{\Gamma}^{(m)}, m \geq 1$. For any $s > 0$, we define the

following events for $i = 0, 1, \ldots, m$,

$$\mathcal{A}_i(s) = \left\{ \|\hat{g}^{(i)} - g_0\|_{\mathcal{G}}^2 \leq C s^{-1} i^{-\frac{2\alpha-1}{2\alpha}} \right\} \text{ and } \mathcal{B}_i(s) = \left\{ \|\hat{\Gamma}^{(i)} - \Gamma_0\|_{\text{Fro}}^2 \leq C s^{-1} i^{-\frac{2\alpha-1}{2\alpha}} \right\}.$$

We assume that for some constant $c \geq 1$ the nuisance estimator $\hat{g}^{(i)}$ satisfies

$$\mathbb{E}_{\mathcal{S}_i} \left[\|\hat{g}^{(i)} - g_0\|_{\mathcal{G}}^{2c} \mid \mathcal{A}_{i-1}(s^{1/c}), \ldots, \mathcal{A}_1(s^{1/c}) \right] \leq C^c i^{-\frac{(2\alpha-1)c}{2\alpha}}. \tag{103}$$

Additionally, we assume that $\hat{\Gamma}^{(i)}$ decays in the same rate such that

$$\mathbb{E}_{\mathcal{S}_i} \left[\|\hat{\Gamma}^{(i)} - \Gamma_0\|_{\text{Fro}}^{2c} \mid \mathcal{B}_{i-1}(s^{1/c}), \ldots, \mathcal{B}_1(s^{1/c}) \right] \leq C^c i^{-\frac{(2\alpha-1)c}{2\alpha}}. \tag{104}$$

With all the assumptions above, it is possible to provide a convergence bound of $\|\theta^{(m,n)} - \theta_\star\|_2^2$ in probability. The following proposition shows that estimations from $\mathcal{S}_m$ using OSGD contribute to a nuisance insensitive rate of $\mathcal{O}(m^{-\frac{2\alpha-1}{\alpha}})$, compared to the nuisance sensitive rate $\mathcal{O}(m^{-\frac{2\alpha-1}{2\alpha}})$ in Prop. 22 for non-Neyman orthogonal losses.

Proposition 23. *Suppose that* $\{\hat{g}^{(m)}, m \geq 1\}$ *satisfies* (103)*, and that* $\{\hat{\Gamma}^{(m)}, m \geq 1\}$ *satisfies* (104)*. Assume that* $\theta^{(m,t)} \in \Theta$ *almost surely for all* $m \geq 1$ *and* $0 \leq t \leq n$*. For any* $s \geq 0$*, define* $\delta(s) = \mathcal{O}(ms)$ *as* (109)*. Under the same conditions to Thm. 1, with probability at least* $1 - \delta(s)$*, it holds that*

$$\begin{aligned}\|\theta^{(m,n)} - \theta_\star\|_2^2 \lesssim\ & s^{-1} \left(1 - \frac{\mu\eta}{2}\right)^{mn} \|\theta^{(0)} - \theta_\star\|_2^2 \\ & + s^{-1} \left(m \exp\left(-\frac{\mu\eta nm}{4}\right) + (s^{-2/c} m^{-\frac{2\alpha-1}{\alpha}} + \eta)((\eta n)^{-1} + 1) \right).\end{aligned}$$

In addition, when $(\eta n)^{-1} = \mathcal{O}(1)$*, with probability at least* $1 - \delta(s)$*, it holds that*

$$\|\theta^{(m,n)} - \theta_\star\|_2^2 \lesssim s^{-1} \left(1 - \frac{\mu\eta}{2}\right)^{mn} \|\theta^{(0)} - \theta_\star\|_2^2 + s^{-1} \left(s^{-2/c} m^{-\frac{2\alpha-1}{\alpha}} + n^{-1} + \eta \right).$$

We refer the reader to Appx. F.7 for the proof.

F.4 Interpretation as Control Variate for Variance Reduction

The regression equation (9), which provides an alternate characterization of the orthogonalized stochastic gradient oracle in the case of negative log-likelihood losses, yields an interesting connection to the Monte Carlo estimation literature. Variance reduction techniques (or "swindles") are used in problems such as estimating the mean or variance of a statistic via Monte Carlo simulation. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with expectation denoted by $\mathbb{E}$ and an unknown vector-valued target $v \in \mathbb{R}^d$. We have $\hat{v} : \Omega \to \mathbb{R}^d$, where we interpret $\hat{v}$ as a (not necessarily unbiased) sample estimate of v. Several variance reduction techniques fall into the category of *control variates* [Graham and Talay, 2013], where a random variable $\hat{u} : \Omega \to \mathbb{R}^k$ with known expectations $u = \mathbb{E}[\hat{u}]$ and a matrix $\Gamma \in \mathbb{R}^{d \times k}$ are used in the variance-reduced estimator

$$\tilde{v} = \hat{v} - \Gamma(\hat{u} - u).$$

A mean squared error decomposition yields the identity

$$\begin{aligned}\mathbb{E}\|\tilde{v} - v\|_2^2 &= \mathbb{E}\|\hat{v} - v\|_2^2 - 2\mathbb{E}\langle \hat{v} - v, \Gamma(\hat{u} - u)\rangle + \mathbb{E}\|\Gamma(\hat{u} - u)\|_2^2 \\ &= \mathbb{E}\|\hat{v} - v\|_2^2 - 2\mathbb{E}\langle \hat{v} - v, \Gamma(\hat{u} - u)\rangle + o(\|\Gamma\|_{\text{op}}),\end{aligned}$$

indicating that for sufficiently "small" Γ, $\tilde{v}$ provides an improved estimator if $\hat{v} - v$ and $\hat{u} - u$ have high (multiple) correlation. While in the Monte Carlo literature, $\hat{u}$ and Γ can be chosen optimally provided knowledge of the underlying data-generating mechanism, as Γ can be interpreted as the regression function of $\hat{v} - v$ on $\hat{u} - u$.[4] Outside of Monte Carlo simulation, this procedure can be applied more widely if the user chooses $\hat{u}$ and $\hat{\Gamma}$ based on intuition or limiting arguments.

[4] In the Monte Carlo settings, it often holds that $d = k$ and $\Gamma = \alpha I$ for some constant $\alpha \in \mathbb{R}$. Then, $\mathbb{E}\langle \hat{v} - v, \Gamma(\hat{u} - u)\rangle$ can be replaced by $\alpha\, \mathbf{Tr}(\mathbb{C}\text{ov}(\hat{v}, \hat{u}))$ and $o(\|\Gamma\|_{\text{op}})$ can be replaced by $o(\alpha)$.

In the stochastic optimization setting, v represents the true gradient of the objective at a particular parameter, while $\hat{v}$ represents a stochastic gradient estimate from an oracle. Variance reduction techniques have previously been applied in an incremental setting, in which a fixed data set of size n is provided at initialization, and the algorithm may only make multiple passes through this same data set [Gower et al., 2020]. Note that this differs from our fully stochastic setting, in which we receive a fresh sample Z_t on each iterate $t = 1, \ldots, n$. For negative log-likelihood losses, our orthogonalized oracle can be viewed in a similar light to control variate-based variance reduction methods (although in an infinite-dimensional setting). To summarize, we have from (9) that

$$\begin{aligned} v &= S_\theta(\theta_\star, g_0) \\ \hat{v} &= S_\theta(\theta_\star, \hat{g}) \\ u &= 0 && \text{(by Asm. 6)} \\ \hat{u} &= \nabla_g \ell(\theta_\star, g_0; Z) \\ \tilde{v} &= S_\theta(\theta_\star, g_0; Z) - \Gamma_0 \nabla_g \ell(\theta_\star, g_0; Z), \end{aligned}$$

using the idealized parameters. Using the approximations for $\theta \neq \theta_\star$, we have

$$\begin{aligned} v &= S_\theta(\theta, g_0) \\ \hat{v} &= S_\theta(\theta, \hat{g}) \\ u &\approx 0 && \text{(for } \theta \approx \theta_\star\text{)} \\ \hat{u} &= \nabla_g \ell(\theta, \hat{g}; Z) \\ \tilde{v} &= S_\theta(\theta, \hat{g}; Z) - \hat{\Gamma} \nabla_g \ell(\theta, \hat{g}; Z). \end{aligned}$$

In fact, using the derivative of the log likelihood in a control variate procedure has been explored in the simulation literature as early as Johnstone and Velleman [1985], as the correlation between a statistic and the score function has tight connections to the Cramér-Rao lower variance bound and exponential families. We emphasize, however, that our method does not require the loss to be of negative log-likelihood form nor any specific distributional knowledge to be applied.

F.5 Discussion of Double Robustness

We now study the double robustness of SGD for dose-response estimation as discussed in Bonvini and Kennedy [2022]. Consider estimating the effect of the continuous treatment $A \in \mathcal{A} \subset \mathbb{R}$ on the outcome $Y \in \mathcal{Y} \subset \mathbb{R}$, which is defined as $\mathbb{E}Y(a)$ (known as the dose-response function, DRF) under the potential outcomes framework. Under standard assumptions, the DRF takes the form

$$\theta_0(t) = \mathbb{E}\left[\mathbb{E}[Y \mid A = t, X]\right] = \int \mathbb{E}[Y \mid A = t, X = x] \mathrm{d}\mathbb{P}(x),$$

where $X \in \mathcal{X} \subset \mathbb{R}^d$ is the measured confounders. Let $Z = (Y, A, X) \sim \mathbb{P}$ with density p. We take the following notations:

$$p(u) = \frac{\mathrm{d}}{\mathrm{d}u}\mathbb{P}(U \leq u), \pi(a \mid x) = \frac{p(a,x)}{p(x)}, \mu(a,x) = \mathbb{E}[Y \mid A = a, X = x], w(a,x) = \frac{p(a)}{\pi(a \mid x)}.$$

We can rewrite $\theta_0(t)$ equivalently as

$$\theta_0(t) = \mathbb{E}\left[\mu(t, X)\right] = \mathbb{E}\left[w(t, X) Y \mid A = t\right].$$

We also take the notations $\mathbb{P}(g(Z)) = \int g(z) \mathrm{d}\mathbb{P}(z)$, $\mathbb{P}_n(g(Z)) = n^{-1} \sum_{i=1}^n g(Z_i)$, $\|g\|_{L_2(\mathbb{P})} = [\mathbb{P}(g^2(Z))]^{1/2}$ to denote the $L_2(\mathbb{P})$ norm, and $\|g\|_{L_4(\mathbb{P})} = [\mathbb{P}(g^4(Z))]^{1/4}$ to denote the $L_4(\mathbb{P})$ norm. We now establish the procedure to estimate $\theta_0(t)$ as Algorithm 1 in Bonvini and Kennedy [2022] with slightly modification to apply SGD:

1. Observe *i.i.d.* samples $\{Z'_i\}_{i=1}^m$ for the nuisance estimation and *i.i.d.* samples $\{Z_i\}_{i=1}^n$ for the parameter estimation.

2. Estimate μ, w, and $m(a) = \mathbb{P}\mu(a, \cdot)$ using $\{Z'_i\}_{i=1}^m$ with $\hat{\mu}$, $\hat{w}$, and $\hat{m}(a) = \mathbb{P}_n(\hat{\mu}(a, \cdot))$, respectively.

3. Construct the pseudo-outcome

$$\hat{\varphi}(Z) = \hat{w}(A, X)Y - \hat{\mu}(A, X) + \hat{m}(A).$$

We also define the true nuisance as

$$\varphi_0(Z) = w(A, X)Y - \mu(A, X) + \int \mu(A, x)\mathrm{d}\mathbb{P}(x).$$

4. Define the loss function via a parametric function class $\mathcal{F}_\Theta = \{f_\theta : \mathcal{A} \mapsto \mathbb{R} \mid \theta \in \Theta \subset \mathbb{R}^d\}$ as

$$\ell(\theta, \varphi; z) = \frac{1}{2}(f_\theta(a) - \varphi(z))^2. \tag{105}$$

Define the stochastic gradient oracle as

$$S_\theta(\theta, \varphi; z) = (f_\theta(a) - \varphi(z))\nabla_\theta f_\theta(a).$$

5. Solve the optimization problem

$$\theta_\star = \arg\min_{\theta\in\Theta} \mathbb{E}\left[\ell(\theta, \varphi_0; Z)\right]$$

using SGD with the stochastic gradient $S_\theta(\theta, \hat{\varphi}; Z)$ by

$$\theta^{(n)} = \theta^{(n-1)} - \eta S_\theta(\theta^{(n-1)}, \hat{\varphi}; Z_{n-1}), \quad \theta^{(0)} \in \Theta. \tag{106}$$

As demonstrated in Bonvini and Kennedy [2022], this procedure would yield a doubly robust ERM estimator. In the following proposition, we claim that double robustness would be preserved if the SGD estimator is adopted instead.

Proposition 24. *Assume that* $\mathbb{E}\left[\|\nabla_\theta f_{\theta_\star}(A)\|_2^2\right]^{1/2} \le C_A$. *Suppose that Asm. 3 holds and* $\theta^{(0)}, \ldots, \theta^{(n)} \in \Theta$ *almost surely for* $\theta^{(n)}$ *in* (106). *If* $\eta \le \mu/2(M\mu + \kappa_1)$, *the iterates of* (106) *satisfy*

$$\mathbb{E}_{\mathcal{D}_n\sim\mathbb{P}^n}[\|\theta^{(n)} - \theta_\star\|_2^2] \lesssim \left(1 - \frac{\mu\eta}{2}\right)^n + \eta + \|w - \hat{w}\|_{L_4(\mathbb{P})}\|\mu - \hat{\mu}\|_{L_4(\mathbb{P})} + \max_{a\in\mathcal{A}} \left|(\mathbb{P}_n - \mathbb{P})\left\{\hat{\mu}(a, X)\right\}\right|^2.$$

Prop. 24 follows directly from the following two lemmas, Lem. 25 and Lem. 26, and the nuisance sensitive rate in Thm. 1. Whenever the empirical estimation $\max_{a\in\mathcal{A}} \left|(\mathbb{P}_n - \mathbb{P})\left\{\hat{\mu}(a, X)\right\}\right|^2$ shrinks, Prop. 24 suggests that the $\theta^{(n)}$ would converge to the the target parameter when either $\hat{w}$ or $\hat{\mu}$ is correctly specified.

Lemma 25. *Assume that* $\mathbb{E}\left[\|\nabla_\theta f_{\theta_\star}(A)\|_2^2\right]^{1/2} \le C_A$. *Then for the loss defined in* (105), *we have*

$$|\mathrm{D}_\varphi\mathrm{D}_\theta L(\theta_\star, \bar{\varphi})[\theta - \theta_\star, \hat{\varphi} - \varphi_0]| \le C_A\|\hat{\varphi} - \varphi_0\|_{\mathcal{G}}\|\theta - \theta_\star\|_2,$$

where $\|\hat{\varphi} - \varphi_0\|_{\mathcal{G}} = \mathbb{E}[\mathbb{E}[\hat{\varphi}(Z) - \varphi_0(Z) \mid A]^2]^{1/2} = \mathbb{E}[(\mathbb{E}[\hat{\varphi}(Z) \mid A] - \theta_0(A))^2]^{1/2}$.

Proof. Let $\hat{r}(t) = \mathbb{E}\left[\hat{\varphi}(Z) \mid A = t\right] - \theta_0(t)$. Note that

$$\begin{aligned}
\mathrm{D}_\varphi\mathrm{D}_\theta L(\theta_\star, \bar{\varphi})[\theta - \theta_\star, \hat{\varphi} - \varphi_0] &= -\mathbb{E}\left[(\hat{\varphi}(Z) - \varphi_0(Z))\langle\nabla_\theta f_{\theta_\star}(A), \theta - \theta_\star\rangle\right] \\
&= -\mathbb{E}\left[(\mathbb{E}\left[\hat{\varphi}(Z) \mid A\right] - \mathbb{E}\left[\varphi_0(Z) \mid A\right])\langle\nabla_\theta f_{\theta_\star}(A), \theta - \theta_\star\rangle\right] \\
&= -\mathbb{E}\left[(\mathbb{E}\left[\hat{\varphi}(Z) \mid A\right] - \theta_0(A))\langle\nabla_\theta f_{\theta_\star}(A), \theta - \theta_\star\rangle\right] \\
&= -\mathbb{E}\left[\hat{r}(A)\langle\nabla_\theta f_{\theta_\star}(A), \theta - \theta_\star\rangle\right].
\end{aligned}$$

Thus, by the assumption that $\mathbb{E}\left[\|\nabla_\theta f_{\theta_\star}(A)\|_2^2\right]^{1/2} \le C_A$,

$$\begin{aligned}|\mathrm{D}_\varphi \mathrm{D}_\theta L(\theta_\star, \bar{\varphi})[\theta - \theta_\star, \hat{\varphi} - \varphi_0]| &\le \mathbb{E}\left[|\hat{r}(A)|^2\right]^{1/2} \mathbb{E}\left[\|\nabla_\theta f_{\theta_\star}(A)\|_2^2\right]^{1/2} \|\theta - \theta_\star\|_2 \\ &\le C_A \mathbb{E}\left[|\hat{r}(A)|^2\right]^{1/2} \|\theta - \theta_\star\|_2.\end{aligned}$$

□

Lemma 26. *For the norm defined in Lem. 25, we have*

$$\|\hat{\varphi} - \varphi_0\|_{\mathcal{G}} \le \|w - \hat{w}\|_{L_4(\mathbb{P})}^{1/2} \|\mu - \hat{\mu}\|_{L_4(\mathbb{P})}^{1/2} + \max_{a \in \mathcal{A}} \left|(\mathbb{P}_n - \mathbb{P})\left\{\hat{\mu}(t, X)\right\}\right|.$$

Proof. Lemma 1 of Bonvini and Kennedy [2022] demonstrates that

$$|\hat{r}(t)| \le \|w - \hat{w}\|_t \|\mu - \hat{\mu}\|_t + \left|(\mathbb{P}_n - \mathbb{P})\left\{\hat{\mu}(t, X)\right\}\right|, \tag{107}$$

where $\|f\|_t^2 = \int f^2(z) \mathrm{d}\mathbb{P}(z \mid A = t)$. By (107), we have

$$\begin{aligned}\|\hat{\varphi} - \varphi_0\|_{\mathcal{G}} &\le \left\|\|w - \hat{w}\|_A \|\mu - \hat{\mu}\|_A + |(\mathbb{P}_n - \mathbb{P})_{t=A}\left\{\hat{\mu}(t, X)\right\}|\right\|_{L_2(\mathbb{P}_A)} \\ &\le \|w - \hat{w}\|_{L_4(\mathbb{P})}^{1/2} \|\mu - \hat{\mu}\|_{L_4(\mathbb{P})}^{1/2} + \max_{a \in \mathcal{A}} \left|(\mathbb{P}_n - \mathbb{P})\left\{\hat{\mu}(t, X)\right\}\right|.\end{aligned}$$

□

F.6 Proof of Proposition 22

Proof. For simplicity, we use the notation $\mathbb{E}_{m,n}$ to replace $\mathbb{E}_{\mathcal{D}_{mn} \cup \mathcal{S}_m \sim \mathbb{P}^{mn} \otimes \mathbb{Q}^m}$. Let $q_n = (1 - \mu\eta/2)^n$, $\delta^{(m,n)} = \theta^{(m,n)} - \theta_\star$, and $\delta^{(0)} = \delta^{(0,n)}$. Thus, by Thm. 1,

$$\begin{aligned}\mathbb{E}_{m,n}[\|\delta^{(m,n)}\|_2^2] &\le q_n \mathbb{E}_{m-1,n}[\|\delta^{(m,0)}\|_2^2] + \frac{2\alpha_1^2}{\mu^2}\xi_m + \frac{4K_1\eta}{\mu} \\ &= q_n \mathbb{E}_{m-1,n}[\|\delta^{(m-1,n)}\|_2^2] + \frac{2\alpha_1^2}{\mu^2}\xi_m + \frac{4K_1\eta}{\mu}.\end{aligned}$$

This recursive formula gives a complete bound for $\theta^{(m,n)}$ as

$$\mathbb{E}_{m,n}[\|\delta^{(m,n)}\|_2^2] \le q_n^m \|\delta^{(0)}\|_2^2 + \frac{2\alpha_1^2}{\mu^2} \sum_{i=1}^m q_n^{m-i} \xi_i + \frac{4K_1\eta}{\mu} \sum_{i=1}^m q_n^{m-i}.$$

By (96), we assume that $\xi_m \le C m^{-\frac{2\alpha-1}{2\alpha}}$ for some $C > 0$. Note that

$$q_n = \left(1 - \frac{\mu\eta}{2}\right)^n \le \exp\left(-\frac{\mu\eta n}{2}\right).$$

For the second term, when $q_n \in (0, 1)$ we have

$$\begin{aligned}\sum_{i=1}^m q_n^{m-i} \xi_i &= \sum_{i=1}^{\lfloor m/2 \rfloor} q_n^{m-i} \xi_i + \sum_{i=\lfloor m/2 \rfloor + 1}^m q_n^{m-i} \xi_i \\ &\le C \sum_{i=1}^{\lfloor m/2 \rfloor} q_n^{m-i} + C\left(\frac{m}{2}\right)^{-\frac{2\alpha-1}{2\alpha}} \sum_{i=\lfloor m/2 \rfloor + 1}^m q_n^{m-i} \\ &\le \frac{Cm}{2} q_n^{m/2} + \frac{C}{1 - q_n}\left(\frac{m}{2}\right)^{-\frac{2\alpha-1}{2\alpha}} \\ &\le \frac{Cm}{2} \exp\left(-\frac{\mu\eta n m}{4}\right) + \frac{C}{1 - q_n}\left(\frac{m}{2}\right)^{-\frac{2\alpha-1}{2\alpha}}.\end{aligned}$$

The last term is easy to bound since for $q_n \in (0,1)$,

$$\sum_{i=1}^{m} q_i^{m-i} = \sum_{i=0}^{m-1} q_n^i \leq \frac{1}{1-q_n}.$$

We claim that for some constant $c > 0$,

$$1 - q_n = 1 - \left(1 - \frac{\mu\eta}{2}\right)^n \geq c \min\left\{\frac{\mu\eta n}{2}, 1\right\}. \tag{108}$$

With (108), we have

$$\frac{1}{1-q_n} \leq c^{-1}\left(\frac{2}{\mu\eta n} + 1\right),$$

which implies that

$$\mathbb{E}_{m,n}[\|\delta^{(m,n)}\|_2^2] = \mathcal{O}\left(q_n^m \|\delta^{(0)}\|_2^2 + m\exp\left(-\frac{\mu\eta n m}{4}\right) + \left(m^{-\frac{2\alpha-1}{2\alpha}} + \eta\right)\left(\frac{1}{\eta n} + 1\right)\right).$$

When $(\eta n)^{-1} = \mathcal{O}(1)$, the bound above reduces to

$$\mathbb{E}_{m,n}[\|\delta^{(m,n)}\|_2^2] = \mathcal{O}\left(q_n^m \|\delta^{(0)}\|_2^2 + m^{-\frac{2\alpha-1}{2\alpha}} + n^{-1} + \eta\right).$$

We will finish the proof by showing (108). The key step is to show $1 - e^{-x} \geq c\min(x, 1)$ for all $x > 0$ and some constant $c > 0$.

Let $f(x) = 1 - e^{-x} - x/2$, for $x \in (0,1)$ we have

$$f'(x) = e^{-x} - \frac{1}{2}\begin{cases} > 0 \text{ for } x \in (0, \log 2), \\ = 0 \text{ for } x = \log 2, \\ < 0 \text{ for } x \in (\log 2, 1). \end{cases}$$

Thus, $f(x) \geq f(\log 2) = (1 - \log 2)/2 > 0$ for $x \in (0,1)$, which implies that $1 - e^{-x} > x/2$ for $x \in (0,1)$. Note that $1 - e^{-x} \geq 1 - e^{-1}$ for $x \geq 1$. Let $c = \min(2^{-1}, 1 - e^{-1})$. Then we have $1 - e^{-x} \geq c\min(x, 1)$.

It follows that

$$1 - q_n = 1 - \exp\left(-n\log\left(\frac{1}{1-\mu\eta/2}\right)\right) \geq c\min\left\{n\log\left(\frac{1}{1-\mu\eta/2}\right), 1\right\}.$$

Since $x - 1 \geq \log x$ for all $x > 0$, we have

$$\log(1 - \mu\eta/2) \leq 1 - \mu\eta/2 - 1 = -\mu\eta/2,$$

which implies that

$$n\log\left(\frac{1}{1-\mu\eta/2}\right) \geq \frac{\mu\eta n}{2}.$$

Thus, we complete the proof. □

F.7 Proof of Proposition 23

Proof. Given $s > 0$, we define $\mathcal{A}_i = \mathcal{A}_i(s^{1/c})$ and $\mathcal{B}_i = \mathcal{B}_i(s^{1/c})$ for $i = 0, \dots, m$ for simplicity. First, since $c \geq 1$, by (103) and Markov inequality, for $i = 1, \dots, m$,

$$\begin{aligned}
\mathbb{P}\left[\mathcal{A}_i \mid \mathcal{A}_{i-1}, \dots, \mathcal{A}_0\right] &= 1 - \mathbb{P}\left[\|\hat{g}^{(i)} - g_0\|_{\mathcal{G}}^2 \geq C s^{-1/c} i^{-\frac{2\alpha-1}{2\alpha}} \mid \mathcal{A}_{i-1}, \dots, \mathcal{A}_0\right] \\
&= 1 - \mathbb{P}\left[\|\hat{g}^{(i)} - g_0\|_{\mathcal{G}}^{2c} \geq C^c s^{-1} i^{-\frac{(2\alpha-1)c}{2\alpha}} \mid \mathcal{A}_{i-1}, \dots, \mathcal{A}_0\right] \\
&\geq 1 - \frac{\mathbb{E}_{\mathcal{S}_m}\left[\|\hat{g}^{(i)} - g_0\|_{\mathcal{G}}^{2c} \mid \mathcal{A}_{i-1}, \dots, \mathcal{A}_1\right]}{C^c s^{-1} i^{-\frac{(2\alpha-1)c}{2\alpha}}} \\
&\geq 1 - s.
\end{aligned}$$

We assume that $\mathbb{P}[\mathcal{A}_0] = \mathbb{P}[\mathcal{B}_0] = 1$, and we have

$$\begin{aligned}
\mathbb{P}\left[\mathcal{A}_m, \mathcal{A}_{i-1}, \dots, \mathcal{A}_1, \mathcal{A}_0\right] &= \mathbb{P}\left[\mathcal{A}_m \mid \mathcal{A}_{m-1}, \dots, \mathcal{A}_1\right] \dots \mathbb{P}\left[\mathcal{A}_1 \mid \mathcal{A}_0\right] \mathbb{P}\left[\mathcal{A}_0\right] \\
&\geq \prod_{i=1}^{m} (1 - s) = (1 - s)^m.
\end{aligned}$$

Similarly, we have

$$\mathbb{P}\left[\mathcal{B}_m, \mathcal{B}_{i-1}, \dots, \mathcal{B}_1\right] \geq (1 - s)^m.$$

we consider the conditional mean squared error of $\theta^{(m,n)}$ given $(\cap_{i=0}^m \mathcal{A}_i) \cap (\cap_{i=0}^m \mathcal{B}_i)$. By similar proof to Prop. 22, we can show that for some constant $C_1 > 0$,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_{mn} \cup \mathcal{S}_m}[\|\theta^{(m,n)} - \theta_\star\|_2^2 \mid (\cap_{i=0}^m \mathcal{A}_i) \cap (\cap_{i=0}^m \mathcal{B}_i)] &\leq C_1 \left(1 - \frac{\mu\eta}{2}\right)^{mn} \|\theta^{(0)} - \theta_\star\|_2^2 \\
&\quad + C_1 \left(m \exp\left(-\frac{\mu\eta n m}{4}\right) + (s^{-2/c} m^{-\frac{2\alpha-1}{\alpha}} + \eta)((\eta n)^{-1} + 1)\right).
\end{aligned}$$

We define the event of interest as

$$\mathcal{E}(s) = \left\{\|\theta^{(m,n)} - \theta_\star\|_2^2 \leq C_1 s^{-1} f_s(m, n)\right\},$$

where $f_s(m, n)$ is defined as

$$\begin{aligned}
f_s(m, n) &= \left(1 - \frac{\mu\eta}{2}\right)^{mn} \|\theta^{(0)} - \theta_\star\|_2^2 \\
&\quad + m \exp\left(-\frac{\mu\eta n m}{4}\right) + (s^{-2/c} m^{-\frac{2\alpha-1}{\alpha}} + \eta)((\eta n)^{-1} + 1).
\end{aligned}$$

By Markov inequality, we have

$$\begin{aligned}
\mathbb{P}\left[\mathcal{E}_3(s) \mid (\cap_{i=0}^m \mathcal{A}_i) \cap (\cap_{i=0}^m \mathcal{B}_i)\right] &\geq 1 - \frac{\mathbb{E}_{\mathcal{D}_{mn} \cup \mathcal{S}_m}[\|\theta^{(m,n)} - \theta_\star\|_2^2 \mid (\cap_{i=0}^m \mathcal{A}_i) \cap (\cap_{i=0}^m \mathcal{B}_i)]}{C_1 s^{-1} f(m, n)} \\
&\geq 1 - s.
\end{aligned}$$

Since

$$\begin{aligned}
\mathbb{P}\left[\mathcal{E}_3(s)^c\right] &= \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}_{\mathcal{E}_3(s)^c} \mathbb{1}_{(\cap_{i=0}^m \mathcal{A}_i) \cap (\cap_{i=0}^m \mathcal{B}_i)}\right] + \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}_{\mathcal{E}_3(s)^c} \mathbb{1}_{((\cap_{i=0}^m \mathcal{A}_i) \cap (\cap_{i=0}^m \mathcal{B}_i))^c}\right] \\
&\leq \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}_{\mathcal{E}_3(s)^c \cap (\cap_{i=0}^m \mathcal{A}_i) \cap (\cap_{i=0}^m \mathcal{B}_i)}\right] + \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}_{((\cap_{i=0}^m \mathcal{A}_i) \cap (\cap_{i=0}^m \mathcal{B}_i))^c}\right] \\
&= \mathbb{P}\left[\mathcal{E}_3(s)^c \cap (\cap_{i=0}^m \mathcal{A}_i) \cap (\cap_{i=0}^m \mathcal{B}_i)\right] + \mathbb{P}\left[((\cap_{i=0}^m \mathcal{A}_i) \cap (\cap_{i=0}^m \mathcal{B}_i))^c\right] \\
&\leq \mathbb{P}\left[\mathcal{E}_3(s)^c \mid (\cap_{i=0}^m \mathcal{A}_i) \cap (\cap_{i=0}^m \mathcal{B}_i)\right] + \mathbb{P}\left[(\cap_{i=0}^m \mathcal{A}_i)^c\right] + \mathbb{P}\left[(\cap_{i=0}^m \mathcal{B}_i)^c\right],
\end{aligned}$$

which implies that

$$\mathbb{P}\left[\mathcal{E}_3(s)\right] \geq 2(1 - s)^m - s - 1.$$

Define $\delta(s)$ as

$$\delta(s) = s + 2(1 - (1-s)^m) = \mathcal{O}(ms). \tag{109}$$

Then, with probability at least $1 - \delta(s)$, we have

$$\begin{aligned}\|\theta^{(m,n)} - \theta_\star\|_2^2 \lesssim\ & s^{-1}\left(1 - \frac{\mu\eta}{2}\right)^{mn} \|\theta^{(0)} - \theta_\star\|_2^2 \\ & + s^{-1}\left(m\exp\left(-\frac{\mu\eta n m}{4}\right) + (s^{-2/c} m^{-\frac{2\alpha-1}{\alpha}} + \eta)((\eta n)^{-1} + 1)\right).\end{aligned}$$

When $(\eta n)^{-1} = \mathcal{O}(1)$, it follows that

$$\|\theta^{(m,n)} - \theta_\star\|_2^2 \lesssim s^{-1}\left(1 - \frac{\mu\eta}{2}\right)^{mn} \|\theta^{(0)} - \theta_\star\|_2^2 + s^{-1}\left(s^{-2/c} m^{-\frac{2\alpha-1}{\alpha}} + n^{-1} + \eta\right).$$

□

G Numerical Experiments

This section provides numerical experiments of the proposed stochastic methods in this paper. In Appx. G.1, we design a numerical experiment to illustrate our orthogonalization method. In Appx. G.2, we design simulations based on a partially linear model. In Appx. G.3, we conduct a real data analysis with synthetic outcome to evaluate the performance of our methods. Code for reproduction can be found at https://fachengyu.github.io/.

G.1 Numerical Illustration

In this section, we design a numerical experiment to illustrate how our orthogonalization method effects the target estimation as shown in Fig. 1 from the main text.

Settings. Consider $\Theta \in R$ and $\mathcal{G} = \mathbb{R}$. Let $L(\theta, g)$ be a real-valued risk function defined as

$$L(\theta, g) := L(u) = \frac{1}{2}\langle u, Au\rangle + \lambda \sin^2(\langle u, Bu\rangle), \tag{110}$$

where $u = (\theta, g)^\top \in \mathbb{R}^2$, $\lambda = 0.02$ is the regularization parameter, and

$$A = \begin{pmatrix} 8 & 3 \\ 3 & 2 \end{pmatrix} \succ 0 \text{ and } B = \begin{pmatrix} 2 & -1 \\ -1 & 1.5 \end{pmatrix} \succ 0.$$

It is easy to see that $(0, 0)$ is the global minimizer of L since $L(\theta, g) \geq 0$. Let $q(u) = \langle u, Bu\rangle$. The gradient w.r.t. u is

$$\begin{aligned} \nabla_u L(u) &= Au + 4\lambda \sin(q(u)) \cos(q(u)) Bu \\ &= (A + 2\lambda \sin(2q(u))B)u. \end{aligned}$$

Since $A + 2\lambda \sin(2q(u))B \succcurlyeq A - 0.04B \succ 0$, it is clear that $(0, 0)$ is the only stationary point, implying that $(0, 0)$ is the only minimizer of L. Furthermore, we can obtain the Hessian w.r.t. u as

$$\nabla_u^2 L(u) = A + 2\lambda \sin(2q(u))B + 8\lambda \cos(2q(u))Bu(Bu)^\top, \tag{111}$$

which implies that $L(\cdot, g)$ is not convex in $\mathbb{R}$ given any $g \in \mathcal{G}_r(g_0)$. However, when Θ is a small neighborhood around zero, it is still possible to have $L(\cdot, g)$ strongly convex for in Θ given any $g \in \mathcal{G}_r(g_0)$.

Orthogonalization. To orthogonalize L, we first derive the orthogonal gradient oracle using (11), and then integral the oracle *w.r.t.* θ to obtain the orthogonalized loss L_{no}.

Let H be the Hessian at $(0, 0)$. By (111), we know that $H = A$, implying $H_{\theta g} = A_{12}$ and $H_{gg} = A_{22}$. Since the gradient w.r.t. θ satisfies

$$\begin{aligned} \nabla_\theta L(\theta, g) &= [1, 0](A + 2\lambda \sin(2q(u))B)u \\ &= (A_{11} + 2\lambda \sin(2q(\theta, g))B_{11})\theta + (A_{12} + 2\lambda \sin(2q(\theta, g))B_{12})g, \end{aligned}$$

and the gradient w.r.t. g satisfies

$$\begin{aligned} \nabla_g L(\theta, g) &= [0, 1](A + 2\lambda \sin(2q(u))B)u \\ &= (A_{21} + 2\lambda \sin(2q(\theta, g))B_{21})\theta + (A_{22} + 2\lambda \sin(2q(u))B_{22})g, \end{aligned}$$

follow the construction of (11) and we obtain the orthogonal gradient oracle as

$$\begin{aligned} S_{\text{no}}(\theta, g) &= \nabla_\theta L(\theta, g) - H_{\theta g} H_{gg}^{-1} \nabla_g L(\theta, g) \\ &= (a + 2b\lambda \sin(2q(\theta, g)))\theta + 2c\lambda \sin(2q(\theta, g))g, \end{aligned}$$

where $a = A_{11} - A_{12}A_{22}^{-1}A_{21}$, $b = B_{11} - A_{12}A_{22}^{-1}B_{21}$, and $c = B_{12} - A_{12}A_{22}^{-1}B_{22}$. Finally, we can integral $S_{\text{no}}(\theta, g)$ w.r.t. θ and recover the orthognalized loss L_{no} as

$$L_{\text{no}}(\theta, g) = \int_0^\theta S_{\text{no}}(s, g) \mathrm{d}s.$$

Numerical Computation. Usually, $S_{\text{no}}(s,g)$ contains a form of integral, which needs to be numerically computed. For the example introduced above, we can simplify $L_{\text{no}}(\theta,g)$ to stabilize the numerical computation. Note that $\nabla_\theta \sin^2(q(\theta,g)) = \sin(2q(\theta,g))(B_{11}\theta + B_{12}g)$. Then

$$\begin{aligned}2b\lambda\int_0^\theta \sin(2q(s,g))s\mathrm{d}s &= \frac{2b\lambda}{B_{11}}\left(\int_0^\theta \sin(2q(s,g))(B_{11}s+B_{12}g)\mathrm{d}s - B_{12}g\int_0^\theta \sin(2q(s,g))\mathrm{d}s\right)\\ &= \frac{2b\lambda}{B_{11}}\left(\sin^2(q(\theta,g)) - B_{12}g\int_0^\theta \sin(2q(s,g))\mathrm{d}s\right).\end{aligned}$$

It follows that the orthogonalized loss L_{no} admits the following form

$$L_{\text{no}}(\theta,g) = \frac{a}{2}\theta^2 + \frac{2b\lambda}{B_{11}}\sin^2(q(\theta,g)) + 2\left(c - \frac{B_{12}}{B_{11}}b\right)\lambda g\int_0^\theta \sin(2q(s,g))\mathrm{d}s,$$

which implies that only the integral of $\sin(2q(s,g))$ *w.r.t.* s needs to be computed.

G.2 Simulations

G.2.1 Data Generating Process

To demonstrate Thm. 1 and Thm. 3, we revisit the partially linear model and the corresponding orthogonal and non-orthogonal losses in Appx. B.1. Specifically, $(X,W,Y) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ satisfies the following partially linear model where the nonlinear function is determined by the distribution of $(W,U) \in \mathbb{R}^d \times \mathbb{R}$:

$$Y = \langle \theta_0, X\rangle + \alpha_0(W) + \epsilon, \tag{112}$$
$$U = \alpha_0(W) + \xi, \tag{113}$$

where $\theta_0 \in \mathbb{R}^d$ is the true parameter, $\alpha_0 : \mathcal{W} \mapsto \mathbb{R}$ is the true nonlinear function, ϵ and ξ are independent noises that satisfy $\mathbb{E}[\epsilon \mid X, W] = 0$ and $\mathbb{E}[\xi \mid W] = 0$. It is clear that $\alpha_0(W) = \mathbb{E}[U \mid W]$. In our simulations, we choose $d = 2$ and $\theta_0 = [-0.5 \;\; 1]^\top$.

To get samples for simulations, we first generate (X,W) under the Gaussian model

$$\begin{bmatrix} X \\ W \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_W \end{bmatrix}, \begin{bmatrix} (1+\delta)\mathbf{I}_2 & \lambda\mathbf{I}_2 \\ \lambda\mathbf{I}_2 & (1+\delta)\mathbf{I}_2 \end{bmatrix}\right), \tag{114}$$

where $\boldsymbol{\mu}_X = [1 \;\; 1]^\top$, $\boldsymbol{\mu}_W = [2 \;\; 2]^\top$, $\mathbf{I}_2 \in \mathbb{R}^{2\times 2}$ is the identity matrix, $\lambda \in [0,1]$ is used to control the correlation between X and W, and $\delta = 0.05$ is used to prevent the degeneration of the covariance matrix. For simplicity, we define the nonlinear function α_0 as

$$\alpha_0(w) = 0.5 \times \cos\left(\frac{w_1+w_2}{2}\right) + 0.5 \times \sin\left(\frac{w_1+w_2}{2}\right), \tag{115}$$

where $w = [w_1 \;\; w_2]^\top \in \mathbb{R}^2$. We then generate Y and U using independent Gaussian noises $\epsilon \sim \mathcal{N}(0,1)$ and $\xi \sim \mathcal{N}(0,1)$ based on (112) and (113), respectively.

G.2.2 Stochastic Gradient Oracles

To estimate the true parameter θ_0 using stochastic gradients, we need to design a correspond loss whose minimizer $\theta_\star$ is equal to θ_0. Based on Appx. B.1, there are two types of loss, the orthogonal loss and the non-orthogonal loss, available for this goal. We will derive the stochastic gradient oracle for these two losses and further derive the orthogonalized gradient oracle for the non-orthogonal loss.

Orthogonal loss. Recall the orthogonal loss in Appx. B.1.1:

$$\ell(\theta, g; z) = \frac{1}{2}[y - g_Y(w) - \langle \theta, x - g_X(w)\rangle]^2, \tag{116}$$

where $g = (g_Y, g_X) : \mathcal{W} \to \mathbb{R} \times \mathbb{R}^d$ and the norm $\|\cdot\|_{\mathcal{G}}$ is defined in (27). The true nuisance for this loss is $g_0 = (g_{0,X}, g_{0,Y})$, where $g_{0,Y}(w) := \mathbb{E}_{\mathbb{P}}[Y \mid W = w]$ and $g_{0,X}(w) := \mathbb{E}_{\mathbb{P}}[X \mid W = w]$.

In fact, the explicit expression for g_0 can be easily obtained as

$$g_{0,Y}(w) = \langle \theta_0, g_{0,X}(w) \rangle + \alpha_0(w), \tag{117}$$

$$g_{0,X}(w) = \mu_X + \frac{\lambda}{1.05}(w - \mu_W). \tag{118}$$

From (112) and (117), it is clear that

$$Y - g_{0,Y}(W) = \langle \theta_0, X - g_{0,X}(W) \rangle + \epsilon,$$

which implies that $\theta_\star = \theta_0$ by Lem. 4. The stochastic gradient oracle for the orthogonal loss (116) is then defined as

$$S_\theta(\theta, g; z) = -(y - g_Y(w) - \langle \theta, x - g_X(w) \rangle)(x - g_X(w)). \tag{119}$$

Non-orthogonal loss. We also provide the non-orthogonal loss in Appx. B.1.2 as

$$\ell(\theta, g; z) = \frac{1}{2}[y - g(w) - \langle \theta, x \rangle]^2, \tag{120}$$

where $g : \mathcal{W} \mapsto \mathbb{R}$ and the norm $\|\cdot\|_{\mathcal{G}}$ is now defined in (36). The true nuisance for this non-orthogonal loss satisfies

$$g_0(w) = \alpha_0(w) = \mathbb{E}\left[U \mid W = w\right]. \tag{121}$$

By Lem. 5, we have $\theta_\star = \theta_0$. The stochastic gradient oracle for the orthogonal loss (120) is then defined as

$$S_\theta(\theta, g; z) = -(y - g(w) - \langle \theta, x \rangle)x. \tag{122}$$

Orthogonalized gradient oracle. Since we perform orthogonalization on the non-orthogonal loss, we have $\theta_\star = \theta_0$ being the same target parameter. By (22) in Appx. B.1.2, the Neyman orthogonalized gradient oracle for this non-orthogonal loss (120) is given by

$$S_{\text{no}}(\theta, g; z) = -(y - g(w) - \langle \theta, x \rangle)(x - \mathbb{E}\left[X \mid W = w\right]). \tag{123}$$

G.2.3 Estimation Methods

Throughout the experiments, we estimate the nuisances and the orthogonalizing operator using full-batch data and stream data, respectively.

Nuisance estimation. Note that the true nuisances for the orthogonal loss and the non-orthogonal loss are conditional expectation given W. To conduct nonparametric regression, we use random Fourier feature (RFF) [Rahimi and Recht, 2007] using the kernel $w \mapsto \exp\left(-\gamma \cdot \|w\|_2^2\right)$ to generate a randomized feature map for W.

The nuisance estimation procedure for obtaining $\hat{g}^{(m)}$ using full batch data can be summarized as

1. Fit RFF sampler with 20 components using m *i.i.d.* samples from $P_{W|\lambda}$.
2. Fit Ridge regressions where the regularization parameter is set to be $0.01/m$. Specifically,
   - For the orthogonal loss, fit two Ridge regressions using m *i.i.d.* samples from the joint distribution $P_{X,W,Y|\lambda}$ and the fitted RFF sampler to coordinate-wisely estimate $\mathbb{E}[X \mid W]$. With the same data, fit one Ridge regression using the fitted RFF sampler to estimate $\mathbb{E}[Y \mid W]$.
   - For the non-orthogonal loss, fit one Ridge regression using m *i.i.d.* samples from the joint distribution $P_{X,W,Y|\lambda}$ and the fitted RFF sampler to estimate $\mathbb{E}[U \mid W]$.

To estimate nuisances using stream data, instead of fit a Ridge regression each time, we perform SGD for the Ridge regression loss. The procedure can be summarized as

1. Initialize RFF sampler with 20 components using n_0 *i.i.d.* samples $(W_i)_{i=1}^{n_0}$ from $P_{W|\lambda}$.
2. Perform SGD update once observing a mini-batch of *i.i.d.* samples from the joint distribution $P_{X,W,Y|\lambda}$ with size n_g. Specifically,

- For the orthogonal loss, perform two SGD with the Ridge loss for m iterations to estimate $\mathbb{E}[X \mid W]$ coordinate-wisely. With the same data perform another SGD with the Ridge loss for m iterations to estimate $\mathbb{E}[Y \mid W]$.
- For the non-orthogonal loss, perform one SGD with the Ridge loss for m iterations to estimate $\mathbb{E}[U \mid W]$.

Orthogonalizing operator estimation. To approximate the orthogonalizing operator Γ_0, it suffices to estimate $\mathbb{E}[X \mid W]$ by (21). To that end, we use the same method as the nuisance estimation. The orthogonalizing operator estimation procedure for obtaining $\hat{\Gamma}^{(k)}$ can be summarized as

1. Fit RFF sampler with 20 components using k *i.i.d.* samples $(W_i')_{i=1}^k$ from $P_{W|\lambda}$.
2. Fit two Ridge regressions with the regularization parameter being $0.01/k$ using the fitted RFF sampler and another k *i.i.d.* samples $(X_i', W_i')_{i=k}^{2k}$ to coordinate-wisely estimate $\mathbb{E}[X \mid W]$.

Target estimation. After the estimation of nuisances and orthogonalizing operator, we perform stochastic gradient descent (SGD) to estimate $\theta_\star$ using each of the three stochastic gradient oracles in (119), (122), and (123) on n *i.i.d.* samples drawn from the joint distribution $P_{X,W,U,Y}$. The learning rates of all the three SGDs are fixed during the training.

G.2.4 Simulation Results

Setup. For each nuisance estimation setting, we study three types of estimation methods for learning θ_0 established in this paper: (1) (orthogonal loss) obtain nuisance estimator $\hat{g}^{(m)} = (\hat{g}_Y^{(m)}, \hat{g}_X^{(m)})$ of (117) and (118) and then perform SGD to obtain $\theta^{(n)}$ using the gradient oracle (119) after plugging in $\hat{g}^{(m)}$; (2) (non-orthogonal loss) obtain the nuisance estimator $\hat{g}^{(m)} = \hat{\alpha}^{(m)}$ of (121) and then perform SGD to obtain $\theta^{(n)}$ using the gradient oracle (122) after plugging in $\hat{g}^{(m)}$; (3) (OSGD) obtain the nuisance estimator $\hat{g}^{(m)}$ of (121) and the orthogonalizing operator estimator $\hat{\Gamma}^{(k)}$ of (21), and then perform SGD to obtain $\theta^{(n)}$ using the gradient oracle (123) after plugging in $\hat{g}^{(m)}$ and $\hat{\Gamma}^{(k)}$. Each method is independently repeated 20 times. For nuisance estimated using stream data, we allow the procedure repeated by plugging in updated nuisance estimators and an updated operator estimator, where the nuisances get updated for 2000 iterations after every 2000 target SGD iterations.

Evaluation. We evaluate the performance of nuisance estimators using the corresponding norms defined in (27) and (36). Specifically, for method (1), we evaluate the nuisance estimation by

$$\|\hat{g}^{(m)} - g_0\|_{\mathcal{G}} = \max\left\{\mathbb{E}\left[\|\hat{g}_X^{(m)}(W) - g_{0,X}(W)\|_2^4\right]^{\frac{1}{4}}, \mathbb{E}\left[(\hat{g}_Y^{(m)}(W) - g_{0,Y}(W))^4\right]^{\frac{1}{4}}\right\}. \quad (124)$$

For method (2) and (3), we evaluate the nuisance estimation by

$$\|\hat{g}^{(m)} - g_0\|_{\mathcal{G}} = \mathbb{E}\left[\|\hat{\alpha}^{(m)}(W) - \alpha_0(W)\|_2^2\right]^{\frac{1}{2}}. \quad (125)$$

We evaluate $\hat{\Gamma}^{(k)} : g \mapsto \mathbb{E}[\hat{g}_X^{(k)}(W)g(W)]$ in method (3) using the Frobenius norm $\|\hat{\Gamma}^{(k)} - \Gamma_0\|_{\mathrm{Fro}}$, which is defined as

$$\|\hat{\Gamma}^{(k)} - \Gamma_0\|_{\mathrm{Fro}} = \mathbb{E}\left[\|\hat{g}_X^{(k)}(W) - \hat{g}_{0,X}(W)\|_2^2\right]^{\frac{1}{2}}. \quad (126)$$

Finally, we evaluate the target estimation using two kinds of criterion: (a) the relative error $\frac{\|\theta^{(n)} - \theta_0\|_2}{\|\theta_0\|_2}$, and (b) the risk $L(\theta^{(n)}, g_0) - L(\theta_\star, g_0)$ where $L(\theta, g) = \mathbb{E}[\ell(\theta, g; Z)]$. For method (1), $\ell(\theta, g; z)$ is the orthogonal loss defined in (116) while for method (2) and (3), $\ell(\theta, g; z)$ is the non-orthogonal loss defined in (120).

Results using nuisances fitted on full-batch data. We first estimate the target using prefitted nuisances and operator. The estimation errors of nuisances and the operator fitted on full-batch data are shown in Fig. 2, where all estimation converges when the sample size m increases and less samples are usually required to obtain the same error level when λ increases.

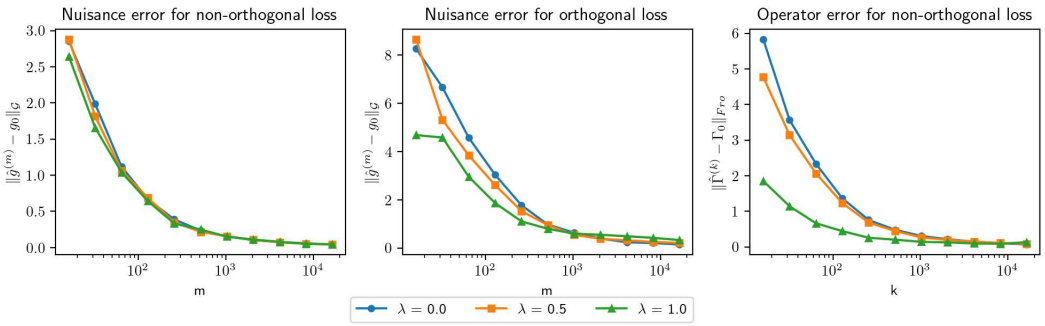

Figure 2: **The Nuisance and Orthogonalizing Operator Fitted on Full-Batch Simulated Data.** The y-axis measures the corresponding error defined in (124) - (126) and the x-axis displays the sample size of data used to estimate the nuisance and operator.

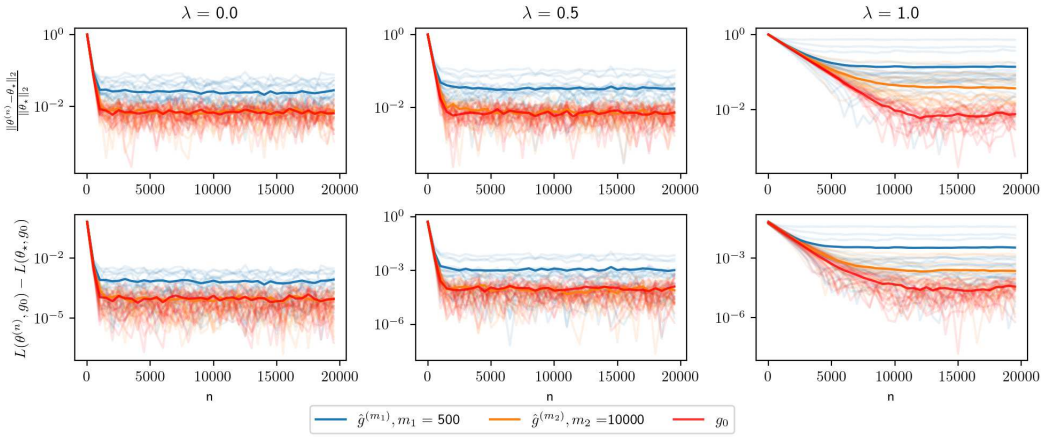

Figure 3: **SGD for Orthogonal Loss with the Nuisance Fitted on Full-Batch Simulated Data.** The x-axis represents the SGD iteration. **Top:** The y-axis measures the relative error. **Bottom:** The y-axis measures the risk.

The performances of SGDs using prefitted nuisances and stochastic gradient oracles (119), (122), and (123) are shown in Fig. 3, Fig. 4, and Fig. 5, respectively. These figures suggest that when λ increases, i.e., the correlation between X and W increases, usually more iterations are required to have SGD converged due to the difficulty of separating the effect of X from W. In addition, a well prefitted nuisance estimator would largely reduce the SGD estimation error, which aligns with Thm. 1. This improvement would be more obvious as λ increases. Fig. 5 also shows that either using a well estimated nuisance or a well estimated orthogonalizing operator can improve the OSGD performance, and that OSGD using both well prefitted nuisance and operator would perform nearly the same as OSGD using the true nuisance and the true operator.

Results using nuisances fitted on stream data. We then study the interleaving the nuisance and target estimations discussed in Appx. F.3. Here, Both the nuisance and the operator are learned using the same data stream and the results are shown in Fig. 6. Compared with Fig. 2, nuisances estimated using stream data usually has larger error and need more iterations to converge due to mini-batch, learning rate, and other tuning parameters.

The performances of SGDs by interleaving nuisance and target updates with stochastic gradient oracles (119), (122), and (123) are shown in Fig. 7, Fig. 8, and Fig. 9, respectively. For all the three stochastic gradients, when λ increases, the relative errors of the target SGD always get larger and their convergence rates become slower. There are obvious errors for SGDs using gradient oracles (119) and (122) in Fig. 7 and Fig. 8 since nuisances are not well estimated. However, OSGD

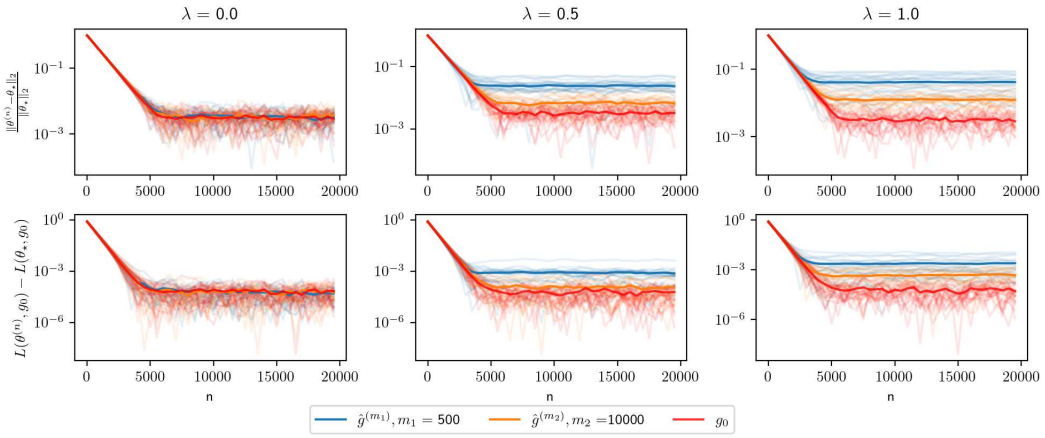

Figure 4: **SGD for Non-Orthogonal Loss with the Nuisance Fitted on Full-Batch Simulated Data.** The x-axis represents the SGD iteration. **Top:** The y-axis measures the relative error. **Bottom:** The y-axis measures the risk.

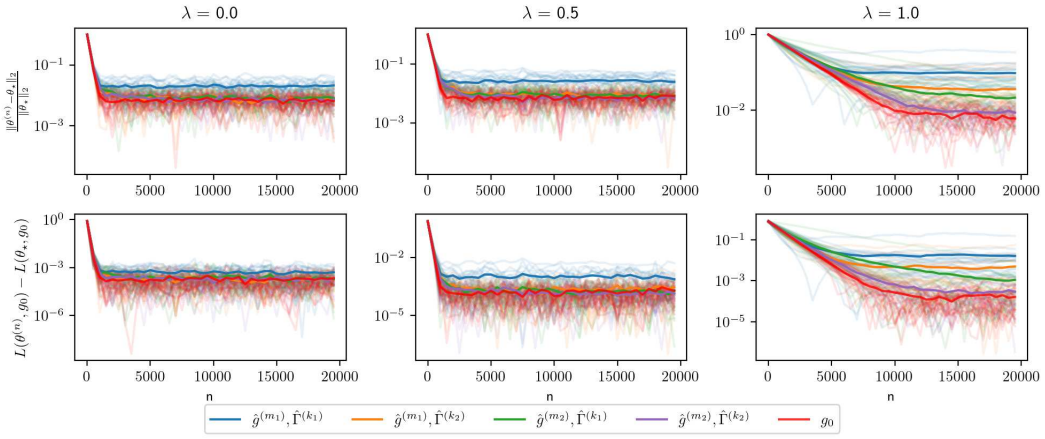

Figure 5: **OSGD with the Nuisance and Operator Fitted on Full-Batch Simulated Data.** Here, $m_1 = 500$, $m_2 = 10000$, $k_1 = 300$, $k_2 = 10000$. The x-axis represents the OSGD iteration. **Top:** The y-axis measures the relative error. **Bottom:** The y-axis measures the risk.

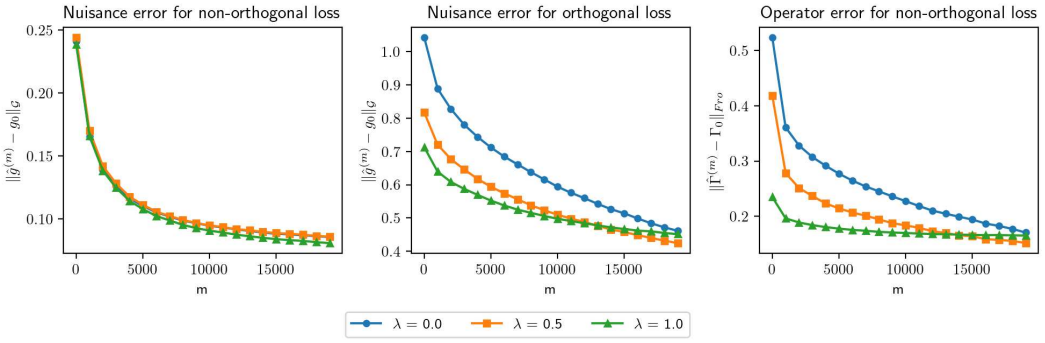

Figure 6: **Nuisance and Orthogonalizing Operator Fitted on Simulated Stream Data.** The y-axis measures the corresponding error defined in (124) - (126) and the x-axis displays the sample size of data used to estimate the nuisance and operator.

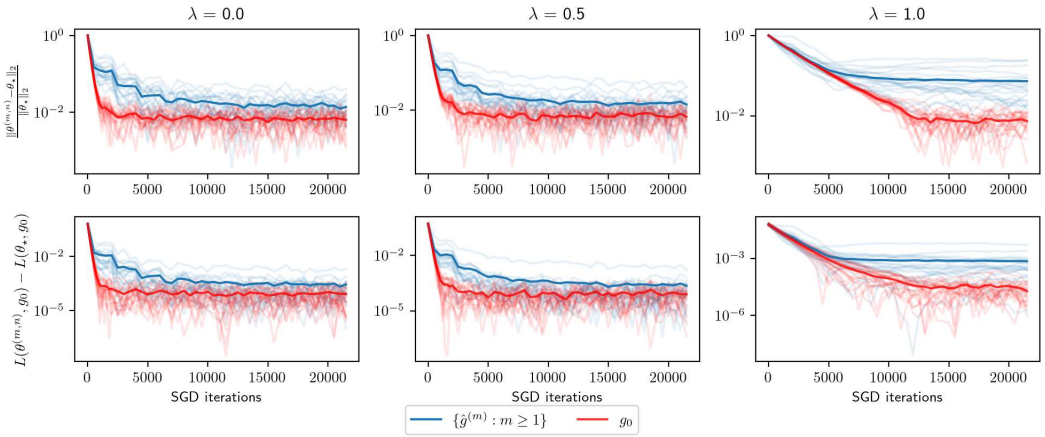

Figure 7: **SGD for Orthogonal Loss with the Nuisance Fitted on Simulated Stream Data.** The x-axis represents the SGD iteration. **Top:** The y-axis measures the relative error. **Bottom:** The y-axis measures the risk.

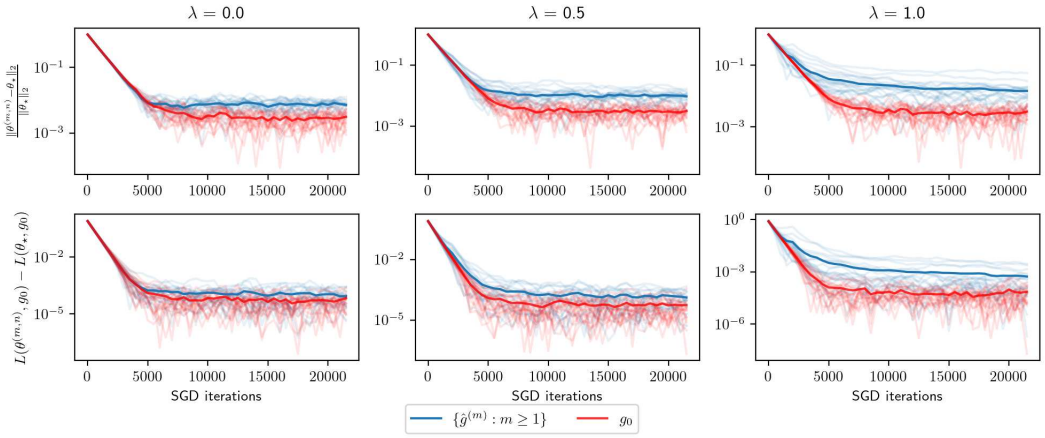

Figure 8: **SGD for Non-Orthogonal Loss with the Nuisance Fitted on Simulated Stream Data.** The x-axis represents the SGD iteration. **Top:** The y-axis measures the relative error. **Bottom:** The y-axis measures the risk.

performs perfectly as shown in Fig. 9, which verifies the analysis of Thm. 3 that using an estimated orthogonalizing operator would reduce the bias from nuisance estimation.

G.3 Real Data Analysis

We consider the Diabetes 130-Hospitals Dataset [Clore et al., 2014] as the real dataset example. We use six of these features as covariates, which are summarized in Tab. 4. We take the binary feature "change" as the input $X \in \{0, 1\}$ and take the rest five features as the control $W \in \mathbb{R}^5$.

G.3.1 Synthetic outcomes

To evaluate the performance of our proposed methods, we use the synthetic outcome instead of a real outcome to examine the performance of our proposed methods. Using the synthetic outcome is common in causal inference; see Nie and Wager [2021, Sec. 4.1]. In this real data analysis, we generate outcome according to the following partially linear model:

$$Y = \tilde{\theta}_0 \cdot X + \tilde{\alpha}_0(W) + 0.5 \times \epsilon, \tag{127}$$

$$U = \tilde{\alpha}_0(W) + 0.5 \times \xi, \tag{128}$$

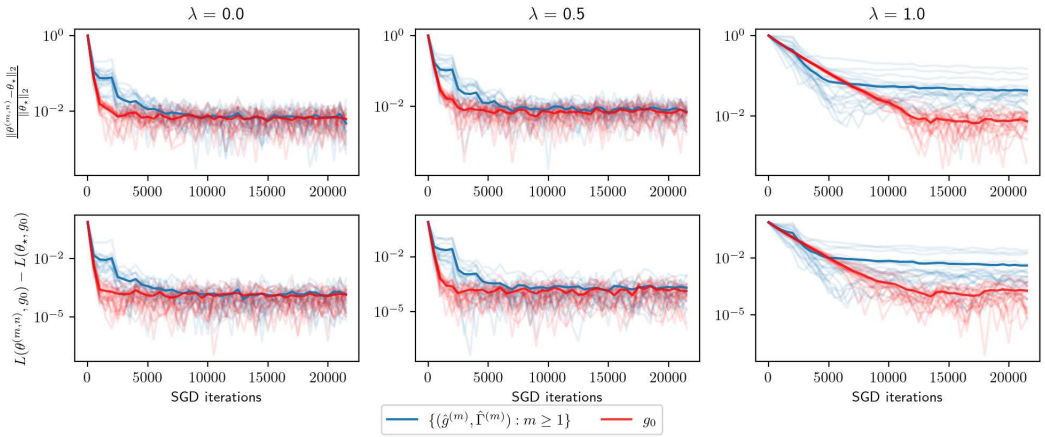

Figure 9: **OSGD with the Nuisance Fitted on Simulated Stream Data.** The x-axis represents the SGD iteration. **Top:** The y-axis measures the relative error. **Bottom:** The y-axis measures the risk.

Feature	Description
change	Indicates if there was a change in diabetic medications.
time_in_hospital	Integer number of days between admission and discharge.
num_lab_procedures	Integer number of lab tests performed during the encounter.
num_procedures	Integer number of procedures (other than lab tests) performed during the encounter.
num_medications	Integer number of distinct generic names administered during the encounter.
number_diagnoses	Integer number of diagnoses.

Table 4: Features used for real data analysis.

where $\tilde{\theta}_0 = -1$, $\epsilon \sim \mathcal{N}(0,1)$ and $\xi \sim \mathcal{N}(0,1)$ are independent noises, and $\tilde{\alpha}_0 : \mathbb{R}^5 \mapsto \mathbb{R}$ satisfies that for $w = (w_1, \ldots, w_5)$,

$$\tilde{\alpha}_0(w) = 0.5 \times \cos\left(5^{-1}\sum_{i=1}^{5} w_i\right) + 0.5 \times \sin\left(5^{-1}\sum_{i=1}^{5} w_i\right).$$

Similar to Appx. G.2.2, we have $\theta_\star = \tilde{\theta}_0$ in this case.

G.3.2 Real Data Results

Setup. We consider the same three stochastic gradient oracles as Appx. G.2.2 and the same two nuisance estimation methods as Appx. G.2.3 except that we use logistic regression on full batch data and SGD of the logistic loss on stream data for estimating $\mathbb{E}[X \mid W]$. The setup of SGD using prefitted nuisances for this real data analysis is the same as Appx. G.2.4. For nuisance estimated using stream data, we update nuisances for 100 iterations after every 500 target SGD iterations.

Evaluation. Since the true nuisances $\mathbb{E}[X \mid W]$ and $\mathbb{E}[Y \mid W]$ are unknown, we evaluate the performance of nuisance estimation $\hat{g}^{(m)} = (\hat{g}_Y^{(m)}, \hat{g}_X^{(m)})$ for the orthogonal loss by the mean squared error:

$$\mathrm{MSE}_1(\hat{g}^{(m)}) = \max\left\{\mathbb{E}_{\mathbb{P}}\left[(\hat{g}_Y^{(m)}(W) - Y)^2\right], \mathbb{E}_{\mathbb{P}}\left[(\hat{g}_X^{(m)}(W) - X)^2\right]\right\}. \tag{129}$$

We adopt the nuisance estimation error $\|\hat{g}^{(n)} - g_0\|_{\mathcal{G}}$ defined in (125) as the nuisance evaluation for non-orthogonal loss due to the synthetic outcome, where now $g_0 = \tilde{\alpha}_0$. For the operator estimation

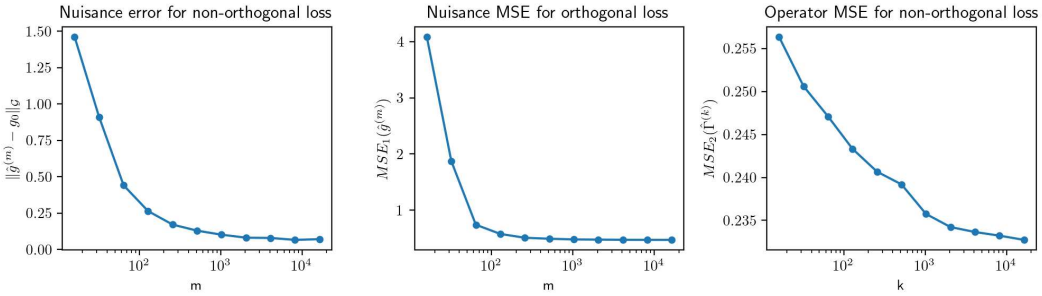

Figure 10: **Nuisance and Orthogonalizing Operator Fitted on Full-Batch Real Data.** The x-axis displays the sample size of data used to estimate the nuisance and operator. **Left.** The y-axis measure the nuisance error defined in (125). **Middle.** The y-axis measure the nuisance estimation MSE defined in (129). **Right.** The y-axis measure the operator estimation MSE defined in (130).

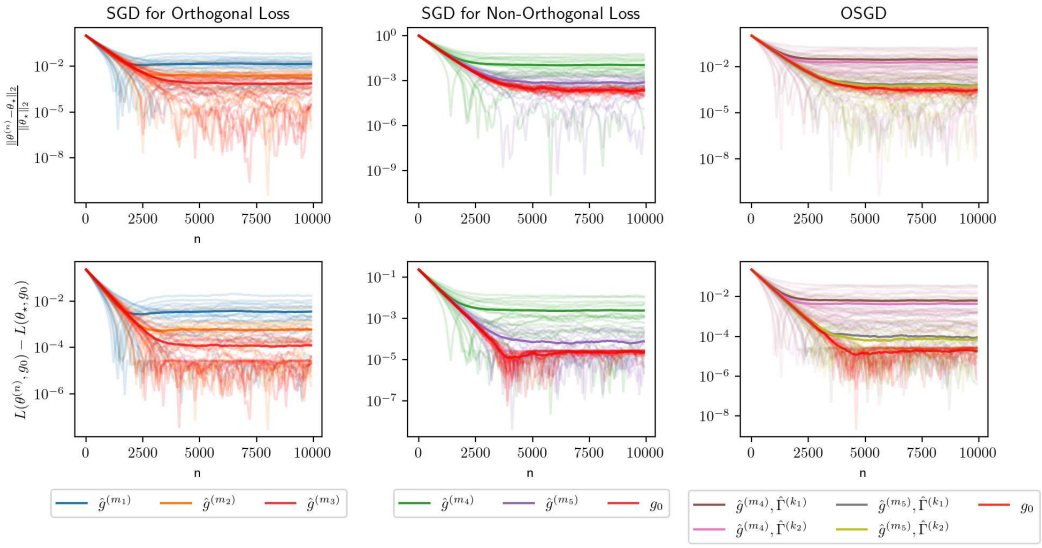

Figure 11: **Stochastic Gradients with Nuisance Fitted on Full-Batch Real Data.** Here, $m_1 = 32$, $m_2 = 64$, $m_3 = 128$, $m_4 = 8$, $m_5 = 128$, $k_1 = 32$ and $k_2 = 128$. The x-axis represents the SGD iteration using corresponding stochastic gradient. **Top:** The y-axis measures the relative error. **Bottom:** The y-axis measures the risk.

$\hat{\Gamma}^{(m)} : g \mapsto \mathbb{E}[\hat{g}_X^{(m)}(W)g(W)]$, evaluate its performance by the mean squared error:

$$\mathrm{MSE}_2(\hat{\Gamma}^{(m)}) = \mathbb{E}_{\mathbb{P}}\left[(\hat{g}_X^{(m)}(W) - X)^2\right]. \tag{130}$$

Results using nuisances fitted on full-batch data. We first estimate the target using prefitted nuisances and operator. The estimation errors of nuisances and the operator using full-batch real data are shown in Fig. 10, which suggests that the estimation of $\tilde{\alpha}_0$ converges to zero due to our design while there exists obvious bias for estimating the nuisance $(g_{0,X}, g_{0,Y})$ and the orthogonalizing operator Γ_0 possibly due to model misspecification.

The performances of SGDs using prefitted nuisances and stochastic gradient oracles (119), (122), and (123) are shown in Fig. 11. Overall, the relative error and the risk are small when well estimated nuisances are used. In addition, both relative errors and risks become smaller when we use more samples to estimate nuisances for the orthogonal loss and the non-orthogonal loss.

Results using nuisances fitted on stream data. We then estimate the target by interleaving the nuisance and target updates. Here, Both the nuisance and the operator are learned using the same

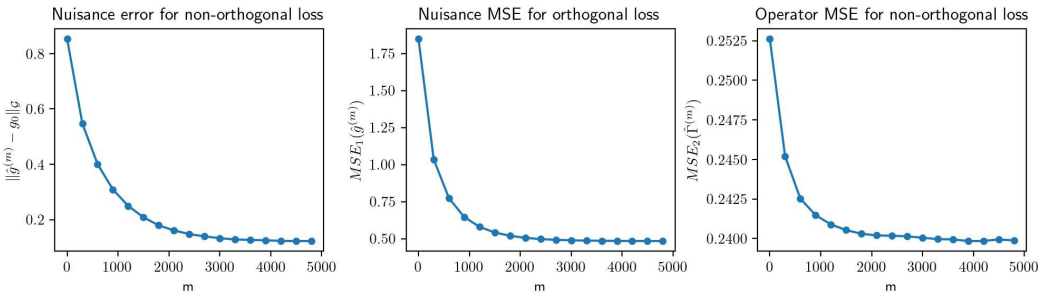

Figure 12: **Estimation Errors of Nuisance and Orthogonalizing Operator Fitted on Stream Data.** The x-axis displays the sample size of data used to estimate the nuisance and operator. **Left.** The y-axis measure the nuisance error defined in (125). **Middle.** The y-axis measure the nuisance estimation MSE defined in (129). **Right.** The y-axis measure the operator estimation MSE defined in (130).

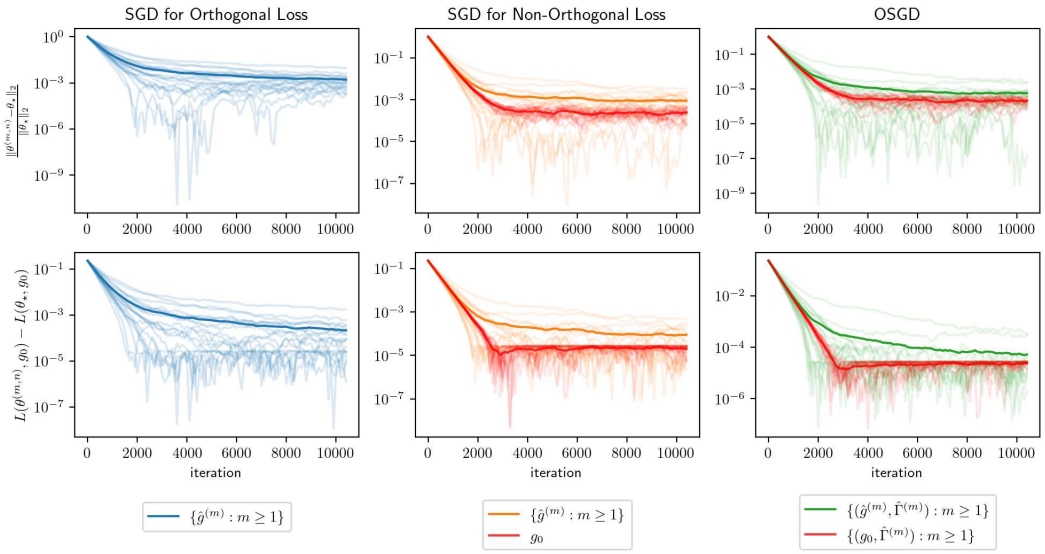

Figure 13: **Stochastic Gradients with Nuisance Fitted on Real Stream Data.** The x-axis represents the SGD iteration. **Top:** The y-axis measures the relative error. **Bottom:** The y-axis measures the risk.

data stream and the results are shown in Fig. 12. Compared with Fig. 10, nuisances estimated using stream data converges similarly.

The performances of SGDs by interleaving nuisance and target updates with stochastic gradient oracles (119), (122), and (123) are shown in Fig. 13. The figure on the left in Fig. 13 shows that the target estimation has small relative error using the estimated nuisance sequence $\{\hat{g}^{(m)} : m \geq 1\}$. The figure in the middle suggests that there is still some bias for the target estimation while this bias is negligible. The figure on the right shows the performance of OSGD, where the relative error of OSGD using the estimated nuisance sequence is similar to OSGD using the true nuisance, which aligns with Thm. 3.

H Extension to SGD Variants

In this section, we discuss strategies for analyzing other variants of SGD under nuisances. In Appx. H.1, we discuss the relationship between SGD with momentum and averaged SGD and provide a convergence analysis example of the averaged SGD. In Appx. H.2, we discuss Adam as a generalization of SGD with momentum and the difficulties to analyze the convergence rate of Adam.

H.1 SGD with Momentum and Averaged SGD

For the gradient oracle sequence $S^{(n)}$, SGD with momentum following the description of Li et al. [2022] can be expressed as

$$m^{(n+1)} = \beta_n m^{(n)} + S^{(n)} \text{ and } \bar{\theta}^{(n+1)} = \bar{\theta}^{(n)} - \alpha_n m^{(n)}, \tag{131}$$

where $\bar{\theta}^{(n)}$ is the SGD estimation sequence, $m^{(n)}$ is the momentum sequence, and $(\alpha_n)_{n\geq 0}$ and $(\beta_n)_{n\geq 0}$ can be any positive sequence. The following example shows that the averaged SGD is a special case of SGD with momentum.

Example 5 (Averaged SGD). Let $\beta_n = 1/n$ and $\alpha_n = \eta(1-\beta_{n+1})$ for all $n \geq 1$. The momentum updates implied by this sequence are

$$m^{(n+1)} = \frac{1}{n} m^{(n)} + S^{(n)} \text{ and } \bar{\theta}^{(n+1)} = \bar{\theta}^{(n)} - \eta \left(1 - \frac{1}{n+1}\right) m^{(n)},$$

which implies that $\bar{\theta}^{(n+1)}$ is the averaged SGD such that

$$\bar{\theta}^{(n+1)} = \frac{1}{n+1} \sum_{t=0}^{n} \theta^{(t)}. \tag{132}$$

Example 5 demonstrates that the convergence rate of SGD with momentum can be analyzed in the same way as averaged SGD. While it is not the focus of this paper, we provide a convergence result of the averaged SGD based on the analysis of Défossez and Bach [2015].

Proposition 27 (Convergence rate of averaged SGD)**.** *Consider the partially linear model and the non-orthogonal loss $\ell(\theta, g; z)$ in Appx. B.1.2. Define $\mathcal{D}_n = (Z_1, \ldots, Z_n)$, sampled from the product measure $\mathbb{P}^n$. Choose the gradient oracle $S^{(n)}$ to be the score $S_\theta(\theta, \hat{g}; Z_n)$ where $\hat{g}$ is estimated independently of $\mathcal{D}_n$. Let $\bar{\theta}^{(n)}$ be the averaged SGD defined in (132). Suppose the same assumptions as Lem. 5. If $0 < \eta < \eta_{\max}$, then*

$$\mathbb{E}_{\mathbb{P}}\left[\|\bar{\theta}^{(n)} - \theta_\star\|_2^2\right] \lesssim \frac{1}{n} + \|\hat{g} - g_0\|_{\mathcal{G}}^2,$$

where $\eta_{\max} = \sup\{\eta > 0 : tr\left(A^\top \mathbb{E}_{\mathbb{P}}[XX^\top]A\right) - \eta \mathbb{E}_{\mathbb{P}}\left[(X^\top A X)^2\right] > 0, \forall A \in \mathcal{S}(\mathbb{R}^d)\}$ and $\mathcal{S}(\mathbb{R}^d)$ is the set of all $d \times d$ symmetric matrices.

Before we prove Prop. 27, recall the example of non-orthogonal loss for the partially linear model in Appx. B.1.2, where $Z = (X, W, Y) \sim \mathbb{P}$ satisfies

$$Y = \langle \theta_0, X \rangle + g_0(W) + \epsilon, \quad \mathbb{E}_{\mathbb{P}}\left[\epsilon \mid X, W\right] = 0. \tag{133}$$

The target parameter $\theta_\star = \arg\min_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}}\left[\ell(\theta, g; Z)\right]$ where ℓ is the non-orthogonal loss defined as

$$\ell(\theta, g; z) = \frac{1}{2}[y - g(w) - \langle \theta, x \rangle]^2, \tag{134}$$

By Lem. 5, we have $\theta_\star = \theta_0$. The stochastic gradient oracle for this non-orthogonal loss is

$$S_\theta(\theta, g; z) = -X(y - g(w) - \langle \theta, x \rangle),$$

and the SGD iteration is defined by $\theta^{(0)} \in \Theta$ and

$$\theta^{(n)} = \theta^{(n-1)} - \eta S_\theta(\theta, \hat{g}; Z_{n-1}) = \theta^{(n-1)} + \eta X_{n-1}(Y_{n-1} - \hat{g}(W_{n-1}) - \langle \theta, X_{n-1} \rangle), \quad (135)$$

where $\hat{g} \in \mathcal{G}_r(g_0)$ is any nuisance estimator independent of $\{Z_i\}_{i=1}^n$. Note that (135) can be written as

$$\begin{aligned}
\theta^{(n)} - \theta_\star &= \left(I - \eta X_n X_n^\top\right) \theta^{(n-1)} + \eta X_n \left(Y_n - \hat{g}\left(W_n\right)\right) - \theta_\star \\
&= \left(I - \eta X_n X_n^\top\right) \left(\theta^{(n-1)} - \theta_\star\right) + \eta X_n \left(Y_n - \hat{g}\left(W_n\right) - X_n^\top \theta_\star\right) \\
&= \left(I - \eta X_n X_n^\top\right) \left(\theta^{(n-1)} - \theta_\star\right) + \eta X_n \epsilon_n - \eta X_n \left(\hat{g}\left(W_n\right) - g_0\left(W_n\right)\right).
\end{aligned}$$

Let $\beta^{(n)} = \theta^{(n)} - \theta_\star$, $r_n = \hat{g}(W_n) - g_0(W_n)$, and

$$M_{k,j} = \left(\prod_{i=k+1}^{j} \left(I - \eta X_i X_i^\top\right)\right)^\top \in \mathbb{R}^{d \times d}.$$

By recursion, we have

$$\begin{aligned}
\beta^{(n)} &= \left(I - \eta X_n X_n^\top\right) \beta^{(n-1)} + \eta X_n \epsilon_n - \eta X_n r_n \\
&= M_{0,n} \beta^{(0)} + \eta \sum_{k=1}^{n} M_{k,n} X_k \epsilon_k - \eta \sum_{k=1}^{n} M_{k,n} X_k r_k.
\end{aligned}$$

Let $\bar{\beta}^{(n)} = \bar{\theta}^{(n)} - \theta_\star = (n+1)^{-1} \sum_{j=0}^{n} \beta^{(j)}$, we have

$$\begin{aligned}
\bar{\beta}^{(n)} = &\frac{1}{n+1} \sum_{j=0}^{n} M_{0,j} \beta^{(0)} + \frac{\eta}{n+1} \sum_{k=1}^{n} \left(\sum_{j=k}^{n} M_{k,j}\right) \left(X_k \left(\epsilon_k - r_k\right) + \mathbb{E}\left[X_k r_k\right]\right) \\
&- \frac{\eta}{n+1} \sum_{k=1}^{n} \left(\sum_{j=k}^{n} M_{k,j}\right) \mathbb{E}\left[X_k r_k\right].
\end{aligned}$$

In the above formula, first two terms are usually interpreted as the bias term and the variance term under the true nuisance, respectively according to Défossez and Bach [2015], and the last term can be viewed as the error term caused by the nuisance estimation.

To analyze the bias term and the variance term, we adopt the notations of Défossez and Bach [2015] for matrices and operators. First, Define $H = \mathbb{E}_{\mathbb{P}}[XX^\top]$. Let H_L (resp. H_R) be the matrix operator representing left multiplication (resp. right multiplication) by H, and T be the linear operator such that for any square matrix $M \in \mathbb{R}^{d \times d}$, $TM = HM + MH - \eta \mathbb{E}_{\mathbb{P}}\left[(X^\top M X) X X^\top\right]$. Let $\rho = \max\{\|I - \eta H\|_{\text{op}}, \|I - \eta T\|_{\text{op}}\}$ where the operator norm $\|\cdot\|_{\text{op}}$ is defined as the largest singular value. Finally, let $\eta_{\max}$ be the same as in Prop. 27. With definitions above, the asymptotic covariances of the bias and the variance follow directly from Theorems 1 and 2 of Défossez and Bach [2015, Appx. 3].

Lemma 28 (Asymptotic covariance of the bias)**.** *Let* $\Xi_0 = \mathbb{E}_{\mathbb{P}}\left[\beta^{(0)} \beta^{(0)^T}\right]$. *If* $0 < \eta < \eta_{\max}$, *then*

$$\mathbb{E}_{\mathbb{P}}\left[B_n B_n^\top\right] = \frac{1}{n^2 \eta^2} \left(H_L^{-1} + H_R^{-1} - \eta I\right) \left(T^{-1} \Xi_0\right) + O\left(\frac{\rho^n}{n} \|\Xi_0\|_F\right),$$

where $B_n = \frac{1}{n+1} \sum_{j=0}^{n} M_{0,j} \beta^{(0)}$.

Lemma 29 (Asymptotic covariance of the variance)**.** *Let* $\Sigma_0 = \mathbb{V}\text{ar}(X_n(\epsilon_n - r_n) + \mathbb{E}[X_n r_n])$. *If* $0 < \eta < \eta_{\max}$, *then*

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[V_n V_n^\top\right] = &\frac{1}{n} \left(H_L^{-1} + H_R^{-1} - \eta I\right) T^{-1} \Sigma_0 \\
&- \frac{1}{\eta n^2} \left(H_L^{-1} + H_R^{-1} - \eta I\right) (I - \eta T) T^{-2} \Sigma_0 + O\left(\frac{\rho^n}{n} \|\Sigma_0\|_F\right),
\end{aligned}$$

where $V_n = \frac{\eta}{n+1}\sum_{k=1}^{n}\left(\sum_{j=k}^{n} M_{k,j}\right)\left(X_k\left(\epsilon_k - r_k\right) + \mathbb{E}\left[X_k r_k\right]\right)$.

In fact, the convergence rate of averaged SGD depends on tr $\left(B_n B_n^\top\right)$ and tr $\left(V_n V_n^\top\right)$. When $\rho < 1$, Lem. 28 demonstrate that tr $\left(B_n B_n^\top\right)$ is of order n^{-2}, while Lem. 29 shows that tr $\left(V_n V_n^\top\right)$ is of order n^{-1}, which is reasonable due to the randomness of the noise $\epsilon_k - r_k$.

For the error term, we can analyze it in a similar way to the bias term. Let $\Delta = \mathbb{E}_{\mathbb{P}}\left[X_n r_n\right]$ and

$$E_n = \frac{\eta}{n+1}\sum_{k=1}^{n}\left(\sum_{j=k}^{n} M_{k,j}\right)\Delta.$$

Note that by Jensen's inequality,

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\left[\|E_n\|_2^2\right] &\le \mathbb{E}_{\mathbb{P}}\left[\frac{\eta^2}{n}\sum_{k=1}^{n}\left\|\left(\sum_{j=k}^{n} M_{k,j}\right)\Delta\right\|_2^2\right] \\ &= \frac{\eta^2}{n}\sum_{k=1}^{n}\mathrm{tr}\left(\mathbb{E}_{\mathbb{P}}\left[\left(\sum_{j=k}^{n} M_{k,j}\Delta\right)\left(\sum_{j=k}^{n} M_{k,j}\Delta\right)^\top\right]\right).\end{aligned}$$

It is clear to see that the asymptotic covariance of $\Delta_{k,n} := \sum_{j=k}^{n} M_{k,j}\Delta$ is of the same form as the bias term in Lem. 28. Let $G_0 = \mathbb{E}_{\mathbb{P}}\left[\Delta\Delta^\top\right]$ and we have

$$\mathbb{E}_{\mathbb{P}}\left[\Delta_{k,n}\Delta_{k,n}^\top\right] = \frac{1}{\eta^2}\left(H_L^{-1} + H_R^{-1} - \eta I\right)\left(T^{-1}G_0\right) + O\left((n-k)\rho^{n-k}\|G_0\|_F\right).$$

Thus, the trace of the covariance summation over $k = 1, \ldots, n$ satisfies

$$\frac{\eta^2}{n}\sum_{k=1}^{n}\mathrm{tr}\left(\mathbb{E}_{\mathbb{P}}\left[\Delta_{k,n}\Delta_{k,n}^\top\right]\right) = \mathrm{tr}\left((H_L^{-1} + H_R^{-1} - \eta I)(T^{-1}G_0)\right) + O\left(\frac{\eta^2}{n}\sum_{k=0}^{n-1} k\rho^k\|G_0\|_F\right). \tag{136}$$

Gathering the bias term, the variance term, and the error term, we are now ready to proof Prop. 27.

Proof of Prop. 27. By Lemma 1 of Défossez and Bach [2015], $0 < \rho < 1$ when $0 < \eta < \eta_{\max}$. Suppose that $\|X\|_\infty \le C_X$ almost surely. By Jensen's inequality we have that

$$\|G_0\|_2 = \|\mathbb{E}_{\mathbb{P}}\left[X_n r_n\right]\|_2^2 \le C_X^2\mathbb{E}_{\mathbb{P}}\left[r_n\right]^2 \le C_X^2\mathbb{E}_{\mathbb{P}}\left[r_n^2\right] = C_X^2\|\hat{g} - g_0\|_{\mathcal{G}}^2.$$

Note that

$$\sum_{k=0}^{n-1} k\rho^k = \rho\frac{\mathrm{d}}{\mathrm{d}\rho}\left(\sum_{k=0}^{n-1}\rho^k\right) = \rho\frac{\mathrm{d}}{\mathrm{d}\rho}\left(\frac{1-\rho^n}{1-\rho}\right) = \frac{\rho - n\rho^n + (n-1)\rho^{n+1}}{(1-\rho)^2} = O(1).$$

By Lem. 28, Lem. 29, and (136), we have

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\left[\|\bar{\theta}^{(n)} - \theta_\star\|_2^2\right] &\lesssim \mathbb{E}_{\mathbb{P}}\left[\|B_n\|_2^2\right] + \mathbb{E}_{\mathbb{P}}\left[\|V_n\|_2^2\right] + \mathbb{E}_{\mathbb{P}}\left[\|E_n\|_2^2\right] \\ &\lesssim \mathrm{tr}\left(\mathbb{E}_{\mathbb{P}}\left[B_n B_n^\top\right]\right) + \mathrm{tr}\left(\mathbb{E}_{\mathbb{P}}\left[V_n V_n^\top\right]\right) + \frac{\eta^2}{n}\sum_{k=1}^{n}\mathrm{tr}\left(\mathbb{E}_{\mathbb{P}}\left[\Delta_{k,n}\Delta_{k,n}^\top\right]\right) \\ &\lesssim \frac{1}{n} + \|\hat{g} - g_0\|_{\mathcal{G}}^2.\end{aligned}$$

□

H.2 Adam

The primary updates for Adam under nuisance estimate $\hat{g}$ are given by the following recursive equations. Below, we let $i \in \{1, \ldots, d\}$ denote a particular dimension of the finite-dimensional parameter of interest. Following the description of Défossez et al. [2022], for the gradient oracle sequence $S^{(n)}$ the Adam generates the target estimator $\tilde{\theta}^{(n)}$ as below:

$$m_i^{(n)} = \beta_1 m_i^{(n-1)} + S_i^{(n)} \tag{137}$$

$$v_i^{(n)} = \beta_2 v_i^{(n-1)} + \left(S_i^{(n)}\right)^2 \tag{138}$$

$$\tilde{\theta}_i^{(n)} = \tilde{\theta}_i^{(n-1)} - \eta \frac{m_i^{(n)}}{\sqrt{\epsilon + v_i^{(n)}}}, \tag{139}$$

where $\beta_2 \in (0, 1]$, $\beta_1 \in [0, \beta_2]$ are the momentum and variance parameters, $m^{(n)}, v^{(n)} \in \mathbb{R}^d$ are the momentum and variance sequences, and $\epsilon > 0$ is a numerical stability parameter. Adam differs from the SGD with momentum by adding a variance sequence $v^{(n)}$. When $v^{(n)}$ is chosen to be a constant, then (137) and (139) would reduce to the special case of SGD with momentum where β_n and α_n are constant.

The analysis of Adam is often done in the case of smooth non-convex optimization, in which it is shown that the gradient of the objective tends to zero [Ward et al., 2020, Défossez et al., 2022]. Specifically, Défossez et al. [2022] consider a momentum-free Adam ($\beta_1 = 0$) to analyze the essential ingredients that differ from momentum: the variance pre-conditioning and element-wise updates, which suggests that under the true nuisance, i.e., $S^{(n)} = S_\theta(\theta^{(n)}, g_0; Z_n)$ for an *i.i.d.* sample $\{Z_i\}_{i=1}^n \sim \mathbb{P}^n$, the convergence result of Adam satisfies

$$\mathbb{E}_{\mathbb{P}^n}\left[\|S_\theta(\tilde{\theta}^{(n)}, g_0)\|_2^2\right] \lesssim \frac{1}{\sqrt{n}}\left(1 + \log\left(\frac{1}{\epsilon}\right)\right).$$

Note that this result is not comparable to our convergence criterion (in terms of iterations), which differs non-trivially from a stationarity or function value analysis. While the convergence of Adam without nuisance has been studied in the literature, it still remains unclear that how to analysis Adam under an estimated nuisance $\hat{g}$ and what should be the nuisance effect on the gradient norm criterion.