Orthogonal Statistical Learning with Self-Concordant Loss

Lang Liu LIU16@UW.EDU
Department of Statistics, University of Washington

Carlos Cinelli CINELLI@UW.EDU
Department of Statistics, University of Washington

Zaid Harchaoui ZAID@UW.EDU
Department of Statistics, University of Washington

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Orthogonal statistical learning and double machine learning have emerged as general frameworks for two-stage statistical prediction in the presence of a nuisance component. We establish non-asymptotic bounds on the excess risk of orthogonal statistical learning methods with a loss function satisfying a self-concordance property. Our bounds improve upon existing bounds by a dimension factor while lifting the assumption of strong convexity. We illustrate the results with examples from multiple treatment effect estimation and generalized partially linear modeling.

Keywords: Orthogonal statistical learning, self-concordance, effective dimension, sandwich covariance, excess risk

1. Introduction

As statistical machine learning impacts several domain applications of major importance to the planet and society, ranging from healthcare to the environment, sophisticated approaches to estimation, proceeding in multiple stages, are being developed to overcome confounding factors and to address high-dimensional nuisance parameters (Peters et al., 2017). Orthogonal statistical learning (OSL), and its statistical estimation predecessor double machine learning (DML), have emerged as general frameworks for two-stage statistical machine learning in the presence of a nuisance component (Mackey et al., 2018; Liu et al., 2021; Nekipelov et al., 2022).

The power of this framework can be illustrated on the task of assessing the causal effect of a treatment on an outcome of interest. Let $Z := (Y, D, X)$ be a vector of observed variables, where $Y \in \mathbb{R}$ is the outcome, $D \in \{0, 1\}$ is the treatment, and $X \in \mathbb{R}^p$ is a vector of features. Denote by $Y(\mathrm{d})$ the potential outcome of Y when the treatment variable D is set (by intervention) to be $\mathrm{d} \in \{0, 1\}$. Our goal is to estimate the average treatment effect (ATE) of D on Y, defined as $\theta_0 := \mathbb{E}[Y(1) - Y(0)]$.

If the treatment assignment D is conditionally ignorable (unconfounded) given X; or, equivalently, if the set of features X satisfy the "backdoor" (adjustment) criterion for estimating the causal effect of D on Y (see Figure 1 for an illustrative causal diagram), a well known identification result in the causal inference literature is that the ATE θ_0 can be identified as a functional of the conditional expectation function (CEF) of the outcome (Rosenbaum and Rubin, 1983; Pearl, 2009; Shpitser et al., 2010; Imbens and Rubin, 2015; Hernán and Robins, 2020). To be more concrete, we obtain that $\theta_0 = \mathbb{E}\left[\mathbb{E}[Y \mid D = 1, X] - \mathbb{E}[Y \mid D = 0, X]\right]$. Note that, in order to estimate θ_0, which is a scalar, we may need to learn the potentially infinite dimensional nuisance $g := (g_0, g_1)$

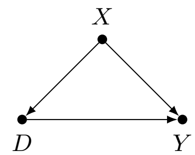

Figure 1: Causal diagram in which X satisfies the backdoor criterion for the causal effect of D on Y. Here, unconfoundedness holds conditional on X, that is, $Y(\mathrm{d}) \perp\!\!\!\perp D \mid X$.

where $g_d := \mathbb{E}[Y \mid D = \mathrm{d}, X]$. This type of challenge, where in order to learn about the target of inference, one needs to estimate many quantities that are not of primary interest, is the one OSL and DML both seek to address.

We work in the framework of OSL and state our results in terms of excess risk in the spirit of statistical learning theory. Formally, let $\mathcal{D} := \{Z_1, \dots, Z_{2n}\}$ be an i.i.d. sample of size $2n$ from an unknown distribution $\mathbb{P}$ on $\mathcal{Z}$. We are interested in learning parameters from the model $\mathcal{M}_{\theta,g}$ equipped with some loss function $\ell(\theta, g; z)$, where $\theta \in \Theta \subset \mathbb{R}^d$ is the *target parameter* and $g \in (\mathcal{G}, \|\cdot\|_{\mathcal{G}})$ is the *nuisance parameter* which may be infinite dimensional. Define the *population risk* at (θ, g) as $L(\theta, g) := \mathbb{E}_{Z\sim\mathbb{P}}[\ell(\theta, g; Z)]$. We will assume throughout that ℓ is three times differentiable w.r.t. θ and twice differentiable w.r.t. g.

Following Foster and Syrgkanis (2020), we assume that there exists a true nuisance parameter $g_0 \in \mathcal{G}$. Without access to g_0, we aim to learn an estimator $\hat{\theta}$ that minimizes the *excess risk*,

$$\mathcal{E}(\theta, g_0) := L(\theta, g_0) - \inf_{\theta\in\Theta} L(\theta, g_0). \tag{1}$$

We assume that the infimum in the excess risk is attainable at a minimizer $\theta_\star$ and the Hessian of $L(\cdot, g_0)$ at $\theta_\star$ is invertible. Consequently, we can rewrite (1) as

$$\mathcal{E}(\theta, g_0) = L(\theta, g_0) - L(\theta_\star, g_0).$$

We focus on the two-stage learning procedure with sample splitting from Foster and Syrgkanis (2020); see also Chernozhukov et al. (2018). Denote by $\mathcal{D}_1 := \{Z_i\}_{i=1}^n$ the first sample split, and by $\mathcal{D}_2 := \{Z_i\}_{i=n+1}^{2n}$ the second sample split. The estimator we study in this paper is constructed from the following algorithm.

OSL Meta-Algorithm

- *Nuisance parameter.* The first stage learning algorithm takes $\mathcal{D}_2$ as input and outputs an estimator $\hat{g}$.
- *Target parameter.* The second stage learning algorithm solves the minimization problem

$$\min_{\theta\in\Theta} L_n(\theta, \hat{g}) := \frac{1}{n}\sum_{i=1}^n \ell(\theta, \hat{g}; Z_i) \tag{2}$$

and outputs an estimator $\hat{\theta}$.

The main contribution of this paper is establishing non-asymptotic guarantees on the excess risk $\mathcal{E}(\hat{\theta}, g_0)$ for the OSL estimator $\hat{\theta}$ under a uniform self-concordance assumption, allowing the dimension of the target parameter to grow at the rate $d = O(n^{1/2})$. In particular, Theorems 4 and 5 derive novel non-asymptotic bounds for the excess risk and characterize its convergence as $n \to \infty$, both in a "fast" and "slow" regime. Compared to previous work, such as (Foster and Syrgkanis, 2020), these new bounds depend on the "effective dimension" as defined by the trace of the sandwich covariance matrix, and recover guarantees that were only available to supervised learning without a nuisance parameter. Effectively, this improves prior bounds on the excess risk at least by a factor of d in a wide range of eigendecay regimes.

In what follows, Section 2 provides the main definitions, assumptions, and establishes the main results of this paper. Section 3 provides further discussions on the converge rate and how our work relates to existing literature. Section 4 examines concrete examples such as treatment effect estimation in a partially linear model and semi-parametric logistic regression. Finally, Section 5 offers some concluding remarks. The full proofs are collected in the Appendix sections.

2. Main Results

We first introduce the notation and some key definitions. We then present all the assumptions required by our analysis. Finally, we summarize our main results and their proof sketches.

2.1. Preliminaries

Notation. Let $S(\theta, g; z) := \nabla_\theta \ell(\theta, g; z)$ be the gradient at z and $H(\theta, g; z) := \nabla_\theta^2 \ell(\theta, g, z)$ be the Hessian at z. We also call $S(\theta, g; z)$ the score at z which is named after the *likelihood score* in maximum likelihood estimation. Their population counterparts are $S(\theta, g) := \mathbb{E}_{Z\sim\mathbb{P}}[S(\theta, g; Z)]$ and $H(\theta, g) := \mathbb{E}_{Z\sim\mathbb{P}}[H(\theta, g; Z)]$. We assume standard regularity assumptions so that $S(\theta, g) = \nabla_\theta L(\theta, g)$ and $H(\theta, g) = \nabla_\theta^2 L(\theta, g)$. Moreover, we let $G(\theta, g) := \mathbb{C}\mathrm{ov}_{Z\sim\mathbb{P}}(S(\theta, g; Z))$ be the covariance matrix of the score $S(\theta, g; Z)$. For simplicity of the notation, we let $S_\star := S(\theta_\star, g_0)$, $G_\star := G(\theta_\star, g_0)$, and $H_\star := H(\theta_\star, g_0)$. We define their empirical quantities as $S_n(\theta, g) := \frac{1}{n}\sum_{i=1}^n S(\theta, g; Z_i)$, $H_n(\theta, g) := \frac{1}{n}\sum_{i=1}^n H(\theta, g; Z_i)$, and

$$G_n(\theta, g) := \frac{1}{n}\sum_{i=1}^n [S(\theta, g; Z_i) - S(\theta, g)][S(\theta, g; Z_i) - S(\theta, g)]^\top.$$

Our analysis is local to a *Dikin ellipsoid* at $\theta_\star$ of radius $\bar{r}_1 := r_1\sqrt{\lambda_{\min}(H_*)}$ and a ball at g_0 of radius r_2, i.e.,

$$\Theta_{\bar{r}_1}(\theta_\star) := \{\theta \in \Theta : \|\theta - \theta_\star\|_{H_\star} < \bar{r}_1\} \quad \text{and} \quad \mathcal{G}_{r_2}(g_0) := \{g \in \mathcal{G} : \|g - g_0\|_{\mathcal{G}} < r_2\},$$

where, given a positive semi-definite matrix J, we let $\|x\|_J := \left\|J^{1/2}x\right\|_2 = \sqrt{x^\top J x}$.

Effective dimension. The quantity that plays a central role in our analysis is the *profile effective dimension* defined as follows. The term *profile* is used in the same sense as in the profile likelihood literature; see, e.g., Murphy and Van der Vaart (2000).

Definition 1 *We define the profile effective dimension to be*

$$\bar{d}_\star := \sup_{g \in \mathcal{G}_{r_2}(g_0)} \mathbf{Tr}(H_\star^{-1/2} G(\theta_\star, g) H_\star^{-1/2}). \tag{3}$$

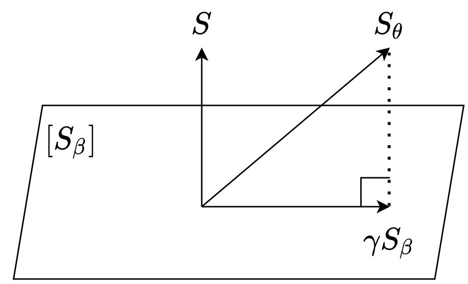

Figure 2: Illustration of the orthogonal score by projection.

When the model is well-specified, we have $H_\star = G_\star$ and thus $\bar{d}_* \approx d$. When the model is mis-specified, it corresponds to the mismatch between the covariance matrix G_* and the Hessian matrix H_*. It can be either as small as a constant or as large as exponential in d depending on the eigendecays of G_* and H_*; see Section 3 for more details.

Self-concordance. We shall use the notion of *self-concordance* from convex optimization. Self-concordance was introduced to analyze the interior-point and Newton-type convex optimization algorithms (Nesterov and Nemirovskii, 1994). Bach (2010) introduced a modified version, which we call the *pseudo self-concordance*, to derive non-asymptotic bounds for the logistic regression. We focus here on the pseudo self-concordance. For a functional F mapping from a vector space $\mathcal{F}$ to $\mathbb{R}$, we define the *derivative operator* D as $\mathrm{D}F(f)[h] := \frac{\mathrm{d}}{\mathrm{d}t}F(f+th)|_{t=0}$ for $f, h \in \mathcal{F}$.

Definition 2 *Let $\mathcal{X} \subset \mathbb{R}^d$ be open and $f : \mathcal{X} \to \mathbb{R}$ be a closed convex function. We say f is pseudo self-concordant with parameter R on $\mathcal{X}$ if*

$$\left|\mathrm{D}^3 f(x)[u,u,u]\right| \le R \left\|u\right\|_2 \mathrm{D}^2 f(x)[u,u], \quad \text{for all } x \in \mathcal{X}, u \in \mathbb{R}^d.$$

Neyman orthogonality. We use *Neyman orthogonality* (Neyman, 1959, 1979) to obtain a fast rate for the excess risk. The intuition behind it is that we want the risk to be insensitive to perturbations in the nuisance g so that a good estimate $\hat{\theta}$ can be obtained even if $\hat{g}$ is of poor quality.

Definition 3 *We say the population risk L is Neyman orthogonal at $(\theta_\star, g_0)$ over $\Theta' \times \mathcal{G}'$ if*

$$\mathrm{D}_g \mathrm{D}_\theta L(\theta_\star, g_0)[\theta - \theta_\star, g - g_0] = 0, \quad \text{for all } \theta \in \Theta', g \in \mathcal{G}'. \tag{4}$$

Since (4) also implies that $\mathrm{D}_g S(\theta_\star, g_0)[g - g_0] = 0$ for all $g \in \mathcal{G}'$, we will also say the score S is Neyman orthogonal at $(\theta_\star, g_0)$.

When g is parametrized by a finite-dimensional vector β, we can obtain a Neyman orthogonal score by projection. Let $L(\theta, \beta)$ be some population risk which may not be Neyman orthogonal. We project $S_\theta := \nabla_\theta L(\theta, \beta)$ onto the space spanned by $S_\beta := \nabla_\beta L(\theta, \beta)$ and obtain $S := S_\theta - \gamma S_\beta$ where $\gamma := [\nabla_\theta \nabla_\beta L(\theta, \beta)][\nabla_\beta^2 L(\theta, \beta)^{-1}]$. It can be shown that S is Neyman orthogonal at $(\theta_\star, \beta_0)$. This procedure is illustrated in Figure 2. Now, to get a population risk that satisfies Neyman orthogonality, it suffices to take the integral of S w.r.t. θ.

2.2. Assumptions

Since our analysis is local to neighborhoods of $\theta_\star$ and g_0, our first assumption localizes the estimator $\hat{\theta}$ and $\hat{g}$ to such neighborhoods.

Assumption 1 (Localization) *Let $r_1, r_2 > 0$ be constants and $\bar{r}_1 := r_1\sqrt{\lambda_{\min}(H_\star)} > 0$. There exists a function $N_{\bar{r}_1, r_2} : [0,1] \to \mathbb{N}_+$ and such that for any $\delta \in (0,1)$ we have, with probability at least $1-\delta$, $\hat{\theta} \in \Theta_{\bar{r}_1}(\theta_\star)$ and $\hat{g} \in \mathcal{G}_{r_2}(g_0)$ for all $n \geq N_{\bar{r}_1, r_2}(\delta)$.*

The localization assumption is necessary to avoid a global strong convexity assumption which is assumed by Foster and Syrgkanis (2020). In order to control the empirical score, we assume that the normalized score at $\theta_\star$ is sub-Gaussian uniformly over $\mathcal{G}_{r_2}(g_0)$.

Assumption 2 (Score sub-Gaussianity) *There exists a constant $K_1 > 0$ such that, for every $g \in \mathcal{G}_{r_2}(g_0)$, we have $\left\|G(\theta_\star, g)^{-1/2}[S(\theta_\star, g; Z) - S(\theta_\star, g)]\right\|_{\psi_2} \leq K_1$, where $\|\cdot\|_{\psi_2}$ is the sub-Gaussian norm defined in Appendix C.*

Another quantity that we need to control is $S(\theta_\star, \hat{g})^\top(\hat{\theta} - \theta_\star) = \mathrm{D}_\theta L(\theta_\star, \hat{g})[\hat{\theta} - \theta_\star]$. Note that $S(\theta_\star, g_0) = 0$ by the first order optimality condition. Hence, we may control it with a smoothness assumption on the population risk.

Assumption 3a *For all $\theta \in \Theta_{\bar{r}_1}(\theta_\star)$ and $g, \bar{g} \in \mathcal{G}_{r_2}(g_0)$, it holds that*

$$|\mathrm{D}_g \mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0]| \leq \beta_1 \left\|\theta - \theta_\star\right\|_{H_\star} \left\|g - g_0\right\|_{\mathcal{G}}$$

for some constant $\beta_1 > 0$.

As we will show in Section 2, this assumption will lead to a slow rate which scales as $O(n^{-1} + \|\hat{g} - g_0\|_{\mathcal{G}}^2)$. If $S(\theta_\star, g)$ is insensitive to g around g_0, we can obtain a faster rate $O(n^{-1} + \|\hat{g} - g_0\|_{\mathcal{G}}^4)$. This insensitivity can be characterized by the *Neyman orthogonality* and higher order smoothness.

Assumption 3b *The population risk L is* Neyman orthogonal *at $(\theta_\star, g_0)$ over $\Theta_{\bar{r}_1}(\theta_\star) \times \mathcal{G}_{r_2}(g_0)$. Moreover, it holds for some constant $\beta_2 > 0$ that*

$$\left|\mathrm{D}_g^2 \mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0, g - g_0]\right| \leq \beta_2 \left\|\theta - \theta_\star\right\|_{H_\star} \left\|g - g_0\right\|_{\mathcal{G}}^2$$

for all $\theta \in \Theta_{\bar{r}_1}(\theta_\star)$ and $g, \bar{g} \in \mathcal{G}_{r_2}(g_0)$.

To facilitate the control of the empirical Hessian, we use the pseudo self-concordance as in Definition 2, which allows us to relate $H_n(\theta, g)$ and $H(\theta, g)$ to $H_n(\theta_\star, g)$ and $H(\theta_\star, g)$, respectively.

Assumption 4 (Uniform pseudo self-concordance) *For any $z \in \mathcal{Z}$ and $g \in \mathcal{G}_{r_2}(g_0)$, $\ell(\cdot, g; z)$ is pseudo self-concordant with parameter R on $\Theta_{\bar{r}_1}(\theta_\star)$. Consequently, for any $g \in \mathcal{G}_{r_2}(g_0)$, $L(\cdot, g)$ is pseudo self-concordant with parameter R on $\Theta_{\bar{r}_1}(\theta_\star)$.*

Since $\hat{\theta}$ is random, we assume that the Hessian satisfies the Bernstein condition so that we can use a covering number argument to relate $H_n(\hat{\theta}, g)$ to $H(\theta_\star, g)$. Due to the variability in $\hat{g}$, the Bernstein condition is satisfied uniformly over a neighborhood of g_0 and we also assume the stability of $H(\theta_\star, g)$ around g_0.

Assumption 5 *For any $\theta \in \Theta_{\bar{r}_1}(\theta_\star)$ and $g \in \mathcal{G}_{r_2}(g_0)$, the centered sandwich Hessian*

$$H(\theta, g)^{-1/2} H(\theta, g; Z) H(\theta, g)^{-1/2} - I_d$$

satisfies a Bernstein condition with parameter K_2 and

$$\sigma_H^2 := \sup_{\theta \in \Theta_{\bar{r}_1}(\theta_\star), g \in \mathcal{G}_{r_2}(g_0)} \left| \mathbb{V}\mathrm{ar}\left(H(\theta, g)^{-1/2} H(\theta, g; Z) H(\theta, g)^{-1/2} \right) \right|_2 < \infty,$$

where, for a matrix $J \in \mathbb{R}^{d \times d}$, we define $|J|_2 := \max\{\lambda_{\max}(J), |\lambda_{\min}(J)|\}$ and $\mathbb{V}\mathrm{ar}(J) := \mathbb{E}[JJ^\top] - \mathbb{E}[J]\,\mathbb{E}[J]^\top$. Moreover, there exist constants κ and $\mathcal{K}$ depending on r_2 such that

$$\kappa H_\star \preceq H(\theta_\star, g) \preceq \mathcal{K} H_\star, \quad \textit{for all } g \in \mathcal{G}_{r_2}(g_0). \tag{5}$$

2.3. Main Results

We now present our main results. We will discuss our results in more detail in Section 3. The first result is a fast rate of convergence for the excess risk assuming Neyman orthogonality.

Theorem 4 (Fast rate) *Under Assumptions 1, 2, 3b, 4, and 5, the OSL estimator $\hat{\theta}$ has excess risk, with probability at least $1 - \delta$,*

$$\mathcal{E}(\hat{\theta}, g_0) \lesssim \frac{e^{3Rr_1}}{\kappa^2} \left[\frac{K_1^2 \log(1/\delta) \bar{d}_\star}{n} + \beta_2^2 \|\hat{g} - g_0\|_{\mathcal{G}}^4 \right] \tag{6}$$

whenever $n \geq \max\{N_{\bar{r}_1, r_2}(\delta/5), 16(K_2^2 + 2\sigma_H^2)[\log(20d/\delta) + d\log(3Rr_1/\log 2)]^2\}$, where $\lesssim$ hides an absolute constant.

When Neyman orthogonality fails to hold, we have a similar bound with $\|\hat{g} - g_0\|_{\mathcal{G}}^4$ being replaced by $\|\hat{g} - g_0\|_{\mathcal{G}}^2$.

Theorem 5 (Slow rate) *Under Assumptions 1, 2, 3a, 4, and 5, the OSL estimator $\hat{\theta}$ has excess risk, with probability at least $1 - \delta$,*

$$\mathcal{E}(\hat{\theta}, g_0) \lesssim \frac{e^{3Rr_1}}{\kappa^2} \left[\frac{K_1^2 \log(1/\delta) \bar{d}_\star}{n} + \beta_1^2 \|\hat{g} - g_0\|_{\mathcal{G}}^2 \right] \tag{7}$$

whenever $n \geq \max\{N_{\bar{r}_1, r_2}(\delta/5), 16(K_2^2 + 2\sigma_H^2)[\log(20d/\delta) + d\log(3Rr_1/\log 2)]^2\}$, where $\lesssim$ hides an absolute constant.

The detailed proofs of Theorems 4 and 5 are deferred to Appendix B. On a high level, the proofs proceed as follows. To begin with, due to Assumption 1, we can dedicate our analysis to the case when $\hat{\theta} \in \Theta_{\bar{r}_1}(\theta_\star)$. By Taylor's theorem,

$$\mathcal{E}(\hat{\theta}, g_0) := L(\hat{\theta}, g_0) - L(\theta_\star, g_0) = S(\theta_\star, g_0)^\top (\hat{\theta} - \theta_\star) + \frac{1}{2} \left\| \hat{\theta} - \theta_\star \right\|_{H(\bar{\theta}, g_0)}^2$$

for some $\bar{\theta} \in \mathrm{Conv}\{\hat{\theta}, \theta_\star\}$. By the first order orthogonality condition, it holds that $S(\theta_\star, g_0) = 0$. For the second term, it follows from the property of the pseudo self-concordance (Assumption 4) that

$$\left\| \hat{\theta} - \theta_\star \right\|_{H(\bar{\theta}, g_0)}^2 \leq e^{R\|\bar{\theta} - \theta_\star\|_2} \left\| \hat{\theta} - \theta_\star \right\|_{H_\star}^2 \leq e^{Rr_1} \left\| \hat{\theta} - \theta_\star \right\|_{H_\star}^2.$$

It now remains to control $\left\|\hat{\theta} - \theta_\star\right\|_{H_\star}^2$.

By Taylor's theorem again, it holds that

$$L_n(\hat{\theta}, \hat{g}) - L_n(\theta_\star, \hat{g}) = S_n(\theta_\star, \hat{g})^\top (\hat{\theta} - \theta_\star) + \frac{1}{2} \left\|\hat{\theta} - \theta_\star\right\|_{H_n(\bar{\theta}', \hat{g})}^2, \tag{8}$$

where $\bar{\theta}' \in \mathrm{Conv}\{\hat{\theta}, \theta_\star\}$. By the optimality of $\hat{\theta}$, we have $L_n(\hat{\theta}, \hat{g}) - L_n(\theta_\star, \hat{g}) \le 0$. We then lower bound the right hand side of (8). According to Assumption 1, we have $\hat{g} \in \mathcal{G}_{r_2}(g_0)$ with high probability when n is sufficiently large. The following Lemma 6, which is a direct consequence of the independence between $\{Z_i\}_{i=1}^n$ and $\hat{g}$, allows us to work with fixed $g \in \mathcal{G}_{r_2}(g_0)$ instead of the random estimator $\hat{g}$.

Lemma 6 *Let $\mathcal{A}(\hat{g}, \{Z_i\}_{i=1}^n)$ be some event regarding $\hat{g}$ and $\{Z_i\}_{i=1}^n$. Let $\mathcal{G}' \subset \mathcal{G}$. If there exists $\delta \in (0,1)$ such that $\mathbb{P}(\mathcal{A}(g, \{Z_i\}_{i=1}^n)) \ge 1 - \delta$ for all fixed $g \in \mathcal{G}'$, then $\mathbb{P}(\mathcal{A}(\hat{g}, \{Z_i\}_{i=1}^n)) \ge (1-\delta)\,\mathbb{P}(\hat{g} \in \mathcal{G}')$.*

Now we focus on the score term $S_n(\theta_\star, g)^\top(\hat{\theta} - \theta_\star)$ in (8) with $\hat{g}$ replaced by a fixed $g \in \mathcal{G}_{r_2}(g_0)$. We split it into two terms

$$[S_n(\theta_\star, g) - S(\theta_\star, g)]^\top (\hat{\theta} - \theta_\star) + S(\theta_\star, g)^\top (\hat{\theta} - \theta_\star). \tag{9}$$

The first term in (9) can be controlled using the sub-Gaussianity of the score. Recall that

$$\bar{d}_\star := \sup_{g \in \mathcal{G}_{r_2}(g_0)} \mathbf{Tr}(H_\star^{-1/2} G(\theta_\star, g) H_\star^{-1/2}).$$

Proposition 7 *Under Assumption 2, it holds for any fixed $g \in \mathcal{G}_{r_2}(g_0)$ that, with probability at least $1 - \delta$,*

$$\|S_n(\theta_\star, g) - S(\theta_\star, g)\|_{H_\star^{-1}}^2 \lesssim \frac{K_1^2 \log(1/\delta) \bar{d}_\star}{n}.$$

We handle the second term in (9) by Neyman orthogonality and smoothness assumptions.

Lemma 8 *Under Assumption 3b, it holds that*

$$S(\theta_\star, g)^\top (\theta - \theta_\star) \ge -\frac{\beta_2}{2} \|\theta - \theta_\star\|_{H_\star} \|g - g_0\|_{\mathcal{G}}^2, \quad \textit{for all } \theta \in \Theta_{\bar{r}_1}(\theta_0) \textit{ and } g \in \mathcal{G}_{r_2}(g_0).$$

By Proposition 7 and Lemma 8, we have

$$\begin{aligned} S_n(\theta_\star, g)^\top (\hat{\theta} - \theta_\star) &\ge - \|S_n(\theta_\star, g) - S(\theta_\star, g)\|_{H_\star^{-1}} \left\|\hat{\theta} - \theta_\star\right\|_{H_\star} + S(\theta_\star, g)^\top (\hat{\theta} - \theta_\star) \\ &\gtrsim -\sqrt{\frac{K_1^2 \log(1/\delta) \bar{d}_\star}{n}} \left\|\hat{\theta} - \theta_\star\right\|_{H_\star} - \beta_2 \left\|\hat{\theta} - \theta_\star\right\|_{H_\star} \|\hat{g} - g_0\|_{\mathcal{G}}^2. \end{aligned} \tag{10}$$

For the Hessian term, with $\hat{g}$ replaced by g, $\|\hat{\theta} - \theta_\star\|_{H_n(\bar{\theta}, g)}$ in (8), we control it using pseudo self-concordance and a covering number argument.

Table 1: In its simplest version (e.g., ignoring the effect of $\hat{g}$), our bound scales as $O(d_\star/n)$ where $d_* := \mathbf{Tr}(H_*^{-1/2} G_* H_*^{-1/2})$ is the effective dimension, while the bound of Foster and Syrgkanis (2020) scales as $O(d'/n)$ where $d' := d^2/\lambda_{\min}(H_\star)$. We compare the two in different regimes of eigendecays of G_* and H_* assuming they share the same eigenvectors.

	Eigendecay		**Dimension Dependency**		**Ratio**
	$G_\star$	$H_\star$	$d_\star$	d'	$d'/d_\star$
Poly-Poly	$i^{-\alpha}$	$i^{-\beta}$	$d^{(\beta-\alpha+1)\vee 0}$	$d^{\beta+2}$	$d^{(\alpha+1)\wedge(\beta+2)}$
Poly-Exp	$i^{-\alpha}$	$e^{-\nu i}$	$d^{-(\alpha-1)\vee 1} e^{\nu d}$	$d^2 e^{\nu d}$	$d^{1\wedge(3-\alpha)}$
Exp-Poly	$e^{-\mu i}$	$i^{-\beta}$	1	$d^{\beta+2}$	$d^{\beta+2}$
			d if $\mu=\nu$		$de^{\nu d}$ if $\mu=\nu$
Exp-Exp	$e^{-\mu i}$	$e^{-\nu i}$	1 if $\mu>\nu$	$d^2 e^{\nu d}$	$d^2 e^{\nu d}$ if $\mu>\nu$
			$e^{(\nu-\mu)d}$ if $\mu<\nu$		$d^2 e^{\mu d}$ if $\mu<\nu$

Proposition 9 *Under Assumptions 4 and 5, it holds, with probability at least $1-\delta$, that*

$$\frac{\kappa}{4e^{Rr_1}} H_\star \preceq H_n(\theta, g) \preceq 3\mathcal{K} e^{Rr_1} H_\star, \quad \textit{for all } \theta \in \Theta_{\bar{r}_1}(\theta_\star), g \in \mathcal{G}_{r_2}(g_0),$$

whenever $n \geq 16(K_2^2 + 2\sigma_H^2)[\log(4d/\delta) + d\log(3Rr_1/\log 2)]^2$.

As a consequence of Proposition 9, we have

$$\frac{1}{2}\left\|\hat{\theta}-\theta_\star\right\|^2_{H_n(\bar{\theta},g)} \gtrsim \frac{\kappa}{e^{Rr_1}}\left\|\hat{\theta}-\theta_\star\right\|^2_{H_\star}. \tag{11}$$

Putting together (8), (10) and (11) leads to an upper bound on $\|\hat{\theta}-\theta_\star\|_{H_\star}$ and thus an upper bound on the excess risk $\mathcal{E}(\hat{\theta}, g_0)$.

When Neyman orthogonality fails to hold, we can replace Lemma 8 by the following lemma and repeat the above steps to obtain the slow rate.

Lemma 10 *Under Assumption 3a, it holds that*

$$S(\theta_\star, g)^\top(\theta-\theta_\star) \geq -\beta_1 \|\theta-\theta_\star\|_{H_\star} \|g-g_0\|_{\mathcal{G}}, \quad \textit{for all } \theta\in\Theta_{\bar{r}_1}(\theta_0) \textit{ and } g \in \mathcal{G}_{r_2}(g_0).$$

3. Discussion

Convergence rate and effective dimension. There are two terms in the bounds (6) and (7). In the case of no nuisance parameter, the second term involving $\|\hat{g}-g\|_{\mathcal{G}}$ vanishes. As for the first term, the profile effective dimension $\bar{d}_\star$ simplifies to $d_\star := \mathbf{Tr}(H_\star^{-1/2} G_\star H_\star^{-1/2})$ as shown in Appendix A. This coincides with the result from Ostrovskii and Bach (2021) on generalized linear models, i.e., the loss is given by $\ell(\theta; Z) := \ell(Y, X^\top\theta)$. Under a well-specified model, the effective dimension $d_\star$ becomes d, recovering the same rate $O(d/n)$ as in classical parametric least-squares regression (see, e.g. Bach, 2021, Proposition 3.5). When the model is misspecified, the effective dimension measures the mismatch between the covariance matrix $G_\star$ and the Hessian matrix $H_\star$. This quantity is related to the *sandwich covariance* in statistics (Wakefield, 2013, Sec. 6.7).

To facilitate its understanding, we summarize the effective dimension $d_\star$ in Table 1 under different regimes of eigendecay, assuming that $G_\star$ and $H_\star$ share the same eigenvectors. Table 1 shows that the dimension dependence can be better than $O(d)$ when the spectrum of $G_\star$ decays faster than the one of $H_\star$. In particular, it is most favorable when the spectrum of $G_\star$ decays as $e^{-\mu i}$ and the one of $H_\star$ decays as $i^{-\beta}$.

In the case when the nuisance parameter needs to be estimated, we pay the price of not knowing the true nuisance in both of the two terms. In this first term, we have $\bar{d}_\star$ rather than $d_\star$ which is the maximum effective dimension in a neighborhood of g_0. As for the second term, the estimator $\hat{g}$ will typically have a rate of convergence $\|\hat{g} - g_0\|_{\mathcal{G}} = O(n^{-\varphi})$ with $\varphi < 1/2$ in high dimensions (Chernozhukov et al., 2018, Section 1). As a result, the term $\|\hat{g} - g_0\|_{\mathcal{G}}^2$ has a dominating effect in the bound (7) which is slower than $O(n^{-1})$. If Neyman orthogonality holds as assumed in Theorem 4, we do not pay this price in the fast rate (6) as long as $\varphi \geq 1/4$. Note that Neyman orthogonality is only used in Lemma 8 to control $S(\theta_\star, g)^\top(\theta - \theta_\star)$. If $|\mathrm{D}_g \mathrm{D}_\theta L(\theta_\star, g_0)[\hat{\theta} - \theta_\star, \hat{g} - g_0]|$ does not vanish but decays as $O(r_n)$, then the second term in (6) will read $O(r_n^2 + \|\hat{g} - g_0\|_{\mathcal{G}}^4)$.

Orthogonal statistical learning and double machine learning. Our work lies in the framework of orthogonal statistical learning. Under a strong convexity assumption and a Neyman orthogonality assumption on the population risk, Foster and Syrgkanis (2020) obtain the rate

$$\mathcal{E}(\hat{\theta}, g_0) \lesssim O\left(\frac{d^2}{n\lambda^2} + \frac{d}{\lambda^2}\|\hat{g} - g_0\|_{\mathcal{G}}^4\right), \quad \text{for all } n \geq 1, \tag{12}$$

where λ is the infimum of $\lambda_{\min}(H(\theta, g_0))$ over a neighborhood of $\theta_\star$ (see Foster and Syrgkanis, 2020, Theorems 1 and 3). Our results improve on theirs in several ways. When $\bar{d}_\star$ is at most proportional to the dimension d, our results improve the excess risk bound by at least a factor of d. Our bounds also remove the explicit dependence on the minimum eigenvalue λ, owing to our tail assumptions 2 and 5 on the normalized score and the Hessian. However, our bounds may depend on λ implicitly through, e.g., the sub-Gaussian parameter K_1. This dependency contributes at most a factor of λ^{-1} for applications considered in Section 4. Hence, to be more concrete, we compare $d_\star$ with $d^2/\lambda_{\min}(H_\star)$ in different eigendecay regimes in Table 1. For instance, when the spectrum of $G_\star$ decays as $e^{-\mu i}$ and the one of $H_\star$ decays as $i^{-\beta}$, our bound gives a rate $O(n^{-1})$ while theirs gives a rate $O(d^{\beta+2}/n)$.

Chernozhukov et al. (2018) recently proposed a set of methods based on Neyman orthogonal scores and cross-fitting, denoted by double or debiased machine learning (DML), to the classical problem of semi-parametric inference. There is an abundant literature on semi-parametric estimation in mathematical statistics (Xia and Härdle, 2006; Wellner and Zhang, 2007) and machine learning (Smola et al., 1998; Rakotomamonjy et al., 2005; Mackey et al., 2018; Bertail et al., 2021) and we refer to classical books for a bibliography (Bickel et al., 1998; Ruppert et al., 2003; Tsiatis, 2006; Kosorok, 2008; Van der Laan and Rose, 2011). In Chernozhukov et al. (2018), the authors establish the asymptotic normality of their estimators when the dimension of the target parameter is kept fixed. In this work, we provide *non-asymptotic* guarantees in terms of excess risk for DML under self-concordance, allowing the dimension of the target parameter to grow at the rate $d = O(n^{1/2})$. In a recent work (Nekipelov et al., 2022), regularized estimators with sparsity-inducing regularization are analyzed in terms of parameter recovery under restricted convexity assumptions.

4. Applications and Examples

4.1. Treatment Effect Estimation

Let us revisit the problem of treatment effect estimation under the assumption of unconfoundedness, as presented in the introduction. Before we had a binary treatment case, and our target of inference was a one dimensional parameter. Here, to better fit our framework, we consider a vector of predictors $D := (D^k)_{k=1}^d \in \mathbb{R}^d$, under partially linear CEF of the following form:

$$\mathbb{E}[Y \mid D, X] = \theta_0^\top D + \gamma_0(X).$$

Note that, by targeting multiple coefficients $\theta \in \Theta \subset \mathbb{R}^d$, we can model not only multiple treatments, but also heterogeneous treatment effects across different binary groups, as well as other non-linear effects, by performing nonlinear transformations of our original treatment variable. To illustrate, suppose T is the original treatment and there is a finite dimensional feature map $D := \phi(T) = [\phi_1(T), \ldots, \phi_d(T)]$ such that $\mathbb{E}[Y \mid T, X] = \theta_0^\top \phi(T) + \gamma_0(X)$. Under the the assumption of uncounfoundedness conditional on X, the ATE of setting T to t_1 versus t_0 is then given by:

$$\mathbb{E}[Y(T = t_1) - Y(T = t_0)] = \theta_0^\top (\phi(t_1) - \phi(t_0)).$$

Heterogeneous effects could be estimated in a similar manner. Letting $G = [G_1, \ldots, G_d]$ denote indicators for d subgroups, and letting $T \in \{0, 1\}$ denote the binary treatment indicator, we can define the covariates $D := TG$. With this flexibility in mind, we now examine the partially linear model in the context of our framework.

Multiple target coefficients in a partially linear model. Let the "target" predictors be $D := (D^k)_{k=1}^d \in \mathbb{R}^d$. Consider the model

$$\begin{aligned} D &= \alpha_0(X) + U \\ Y &= \theta_0^\top D + \gamma_0(X) + V = \zeta_0(X) + \theta_0^\top U + V, \end{aligned}$$

where $\alpha_0 : \mathbb{R}^p \to \mathbb{R}^d$, $\mathbb{E}[U \mid X] \overset{\text{a.s.}}{=} 0$ and $\mathbb{E}[V \mid D, X] \overset{\text{a.s.}}{=} 0$ are the residuals. Moreover, U has a non-singular covariance Σ_u and V is independent of D and X with variance $\sigma_v^2 > 0$. We reparametrize the model by $g = (\zeta, \alpha)$ and work with the loss

$$\ell(\theta, g; Z) := [Y - \zeta(X) - \theta^\top (D - \alpha(X))]^2.$$

Since $\mathbb{E}[UV] = \mathbb{E}[(D - \alpha_0(X))V] = 0$, we have

$$\begin{aligned} L(\theta, g) &:= \mathbb{E}[[Y - \zeta(X) - \theta^\top (D - \alpha(X))]^2] \\ &= \mathbb{E}\left[\Big(\zeta_0(X) - \zeta(X) - \theta^\top (\alpha_0(X) - \alpha(X))\Big)^2\right] + \|\theta - \theta_0\|_{\Sigma_u}^2 + \sigma_v^2. \end{aligned}$$

This implies that the population risk L at g_0 has a unique minimizer $\theta_\star = \theta_0$.

Now suppose that U is bounded (i.e., $\|U\|_2 \le M$), V is sub-Gaussian with parameter $\|V\|_{\psi_2}$, and $\|\cdot\|_{\mathcal{G}}$ is chosen as the sup-norm, i.e., $\|g\|_{\mathcal{G}} = \sup_x \sqrt{\|\alpha(x)\|^2 + \zeta^2(x)}$. Let us verify the assumptions in Section 2.2 for this model. For Assumption 2, we have

$$S(\theta_\star, g; Z) = 2(\alpha_0(X) - \alpha(X) + U) \left[(\alpha_0(X) - \alpha(X))^\top \theta_\star - (\zeta_0(X) - \zeta(X) + V)\right],$$

$$S(\theta_\star, g) = 2\,\mathbb{E}\left[(\alpha_0(X) - \alpha(X))(\alpha_0(X) - \alpha(X))^\top \theta_\star - (\alpha_0(X) - \alpha(X))(\zeta_0(X) - \zeta(X))\right],$$

and $G(\theta_\star, g) \succeq 4\sigma_v^2 \Sigma_u$. Note that $\|\alpha_0(X) - \alpha(X)\| \le \|g - g_0\|_{\mathcal{G}} \le r_2$, $\|\zeta_0(X) - \zeta(X)\| \le \|g - g_0\|_{\mathcal{G}} \le r_2$, $\|U\|_2 \le M$, and V is sub-Gaussian. Hence, it follows from Lemmas 19,20 and 21 that the normalized score is sub-Gaussian with sub-Gaussian norm

$$K_1 \lesssim \frac{(r_2 + M)[r_2(\|\theta_\star\|_2 + 1) + \|V\|_{\psi_2}]}{\sigma_v \sqrt{\lambda_{\min}(\Sigma_u)}}.$$

For Assumption 3b, it holds that, for any $\bar{\theta} \in \Theta$,

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

which verifies the Neyman orthogonality. Moreover, we have $H_\star = \nabla_\theta^2 L(\theta_\star, g_0) = 2\Sigma_u$ and

$$\begin{aligned}
&\left|\mathrm{D}_g^2 \mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0, g - g_0]\right| \\
&= \left|4(\theta - \theta_\star)^\top \mathbb{E}\left\{(\alpha_0(X) - \alpha(X))\left[(\alpha_0(X) - \alpha(X))^\top \theta_\star - (\zeta_0(X) - \zeta(X))\right]\right\}\right| \\
&\lesssim \frac{\|\theta_\star\|_2 + 1}{\sqrt{\lambda_{\min}(\Sigma_u)}} \|\theta - \theta_\star\|_{H_\star} \|g - g_0\|_{\mathcal{G}}^2 .
\end{aligned}$$

In other words, Assumption 3b holds true with $\beta_2 \lesssim (\|\theta_\star\|_2 + 1)/\sqrt{\lambda_{\min}(\Sigma_u)}$.

For Assumption 4, both the loss ℓ and the population risk L are pseudo self-concordant with arbitrary parameter $R \ge 0$ since their third derivatives w.r.t. θ are zero. For Assumption 5, we have

$$\begin{aligned}
H(\theta, g; Z) &= 2(\alpha_0(X) - \alpha(X) + U)(\alpha_0(X) - \alpha(X) + U)^\top, \\
H(\theta, g) &= 2\,\mathbb{E}[(\alpha_0(X) - \alpha(X) + U)(\alpha_0(X) - \alpha(X) + U)^\top] \succeq 2\Sigma_u.
\end{aligned}$$

Note that $H(\theta, g)^{-1/2} H(\theta, g; Z) H(\theta, g)^{-1/2} - I_d$ has mean-zero and satisfies

$$\left|H(\theta, g)^{-1/2} H(\theta, g; Z) H(\theta, g)^{-1/2} - I_d\right|_2 \lesssim \left\|H(\theta, g)^{-1}\right\|_2 [\|g - g_0\|_{\mathcal{G}}^2 + \|U\|_2^2] \lesssim \frac{r_2^2 + M^2}{\lambda_{\min}(\Sigma_u)}.$$

Hence, it follows from Wainwright (2019, Equation 6.30) that

$$H(\theta, g)^{-1/2} H(\theta, g; Z) H(\theta, g)^{-1/2} - I_d$$

satisfies the Bernstein condition with parameter $K_2 \lesssim (r_2^2 + M^2)/\lambda_{\min}(\Sigma_u)$. Moreover, $\sigma_H^2 \lesssim (r_2^2 + M^2)^2/\lambda_{\min}^2(\Sigma_u)$. For the stability (5), we have $H_\star = 2\Sigma_u$ and

$$2\Sigma_u \preceq H(\theta_\star, g) = 2\,\mathbb{E}[(\alpha_0(X) - \alpha(X) + U)(\alpha_0(X) - \alpha(X) + U)^\top] \preceq 2[r_2^2 I_d + \Sigma_u].$$

Thus, the stability holds with $\kappa = 1$ and $\mathcal{K} = 1 + r_2^2/\lambda_{\min}(\Sigma_u)$. To summarize, invoking Theorem 4 gives the following risk bound up to a constant factor:

$$\frac{(r_2 + M)^2[r_2(\|\theta_\star\|_2 + 1) + \|V\|_{\psi_2}]^2}{\sigma_v^2 \lambda_{\min}(\Sigma_u)} \frac{\bar{d}_\star}{n} \log(1/\delta) + \frac{(\|\theta_\star\|_2 + 1)^2}{\lambda_{\min}(\Sigma_u)} \|\hat{g} - g_0\|_{\mathcal{G}}^4 . \tag{13}$$

Remark 11 *As a comparison, assuming $\|U\|_2 \le M$, $\|V\|_2 \le M'$, and $R := \sup_{\theta \in \Theta} \|\theta\|_2 \vee 1 < \infty$, Theorems 1 and 3 of Foster and Syrgkanis (2020) yield the bound*

$$\frac{\bar{K}^2}{\lambda_{\min}(\Sigma_u)^2} \frac{d^2}{n} \log(1/\delta) + \frac{R\bar{K}d}{\lambda_{\min}(\Sigma_u)^2} \|\hat{g} - g_0\|_{\mathcal{G}}^4 ,$$

where $\bar{K} := (r_2 + M)[r_2(R + 1) + M' + RM]$. Our result not only requires less stringent assumptions but also improves their result by a factor of $d/\lambda_{\min}(\Sigma_u)$ when $\bar{d}_\star \lesssim d$.

4.2. Semi-Parametric Logistic Regression

We consider a semi-parametric logistic regression model to illustrate the usefulness of the pseudo self-concordance assumption.

Let $Z := (X, W, Y)$ where $X \in \mathbb{R}^d$, $W \in \mathcal{W}$, and $Y \in \{-1, 1\}$. Consider the model

$$\mathbb{P}(Y = 1 \mid X, W) = \sigma\left(\theta_0^\top X + g_0(W)\right),$$

where $\sigma(u) := (1 + e^{-u})^{-1}$. It is clear that

$$\mathbb{E}[Y \mid X, W] = \mathbb{P}(Y = 1 \mid X, W) - \mathbb{P}(Y = -1 \mid X, W) = 2\sigma(\theta_0^\top X + g_0(W)) - 1.$$

The logistic loss is defined as

$$\ell(\theta, g; Z) := \log\left(1 + \exp\left(-Y(\theta^\top X + g(W))\right)\right).$$

It can be shown that

$$S(\theta, g; Z) = \left[\sigma(\theta^\top X + g(W)) - \frac{1}{2} - \frac{Y}{2}\right] X$$

and

$$\begin{aligned}
S(\theta, g) &= \mathbb{E}[\mathbb{E}[S(\theta, g; Z) \mid X, W]] = \mathbb{E}\left\{X\left[\sigma(\theta^\top X + g(W)) - \sigma(\theta_0^\top X + g_0(W))\right]\right\} \\
H(\theta, g) &= \mathbb{E}\left[\sigma(\theta^\top X + g(W))[1 - \sigma(\theta^\top X + g(W))]XX^\top\right] \\
G(\theta, g) &= \mathbb{E}\left\{[\sigma(\theta^\top X + g(W)) - \sigma(\theta_0^\top X + g_0(W))]^2 XX^\top\right\} - S(\theta, g)S(\theta, g)^\top + H(\theta_0, g_0).
\end{aligned}$$

Assume that $H_\star := H(\theta_0, g_0)$ is non-singular. The population risk $L(\theta, g_0)$ is minimized at $\theta_\star = \theta_0$.

Suppose that X is bounded (i.e., $\|X\|_2 \le M$), $r_1 \le \lambda_{\min}(H_\star)/M^3$, and $r_2 \le \lambda_{\min}(H_\star)/M^2$. By the non-singularity of $H_\star$, the covariance $G(\theta, g)$ is non-singular for all $\theta \in \Theta$ and $g \in \mathcal{G}$. Let us verify the assumptions in Section 2.2 for this model. For Assumption 2, it follows directly from Lemmas 19 and 21 that the normalized score is sub-Gaussian with sub-Gaussian norm $K_1 \lesssim M/\sqrt{\lambda_{\min}(H_\star)}$. Assumption 3a holds true with $\beta_1 := M/(4\sqrt{\lambda_{\min}(H_\star)})$ since

$$\begin{aligned}
&\left|\mathrm{D}_g \mathrm{D}_\theta L(\theta_\star, \bar{g})[\theta - \theta_\star, g - g_0]\right| \\
&= \left|\mathbb{E}\left[\sigma(\theta_\star^\top X + \bar{g}(W))[1 - \sigma(\theta_\star^\top X + \bar{g}(W))]X^\top(\theta - \theta_\star)(g(W) - g_0(W))\right]\right| \\
&\le \frac{M}{4\sqrt{\lambda_{\min}(H_\star)}} \|\theta - \theta_\star\|_{H_\star} \|g - g_0\|_{\mathcal{G}}.
\end{aligned}$$

For Assumption 4, we have, with $a := \theta^\top x + g(w)$,

$$\begin{aligned}
\left|\mathrm{D}_\theta^3 \ell(\theta, g; z)[u, u, u]\right| &= \left|\sigma(a)[1 - \sigma(a)][1 - 2\sigma(a)](u^\top x)^3\right| \\
&\le \left|\sigma(a)[1 - \sigma(a)] \|u\|_2 \|x\|_2 (u^\top x)^2\right| \\
&\le M \|u\|_2 \mathrm{D}_\theta^2 \ell(\theta, g; z)[u, u],
\end{aligned}$$

which implies that $\ell(\cdot, g; z)$ is pseudo self-concordance with parameter $R = M$. The pseudo self-concordance of $L(\theta, g)$ can be verified similarly.

For Assumption 5, we first show that $H(\theta, g)$ is non-singular on $\Theta_{\bar{r}_1}(\theta_\star) \times \mathcal{G}_{r_2}(g_0)$. In fact, with $A := \theta^\top X + g(W)$ and $A_0 := \theta_0^\top X + g_0(W)$, we have

$$\begin{aligned}
|H(\theta, g) - H_\star|_2 &\leq \mathbb{E}\left[|\sigma(A) - \sigma(A_0)| \left|XX^\top\right|_2\right] \\
&\leq \frac{1}{4}\,\mathbb{E}\left[[|\theta^\top X - \theta_0^\top X| + |g(W) - g_0(W)|] \left|XX^\top\right|_2\right] \\
&\leq \frac{1}{4}\left(r_1 M^3 + r_2 M^2\right), \quad \text{for all } (\theta, g) \in \Theta_{\bar{r}_1}(\theta_\star) \times \mathcal{G}_{r_2}(g_0).
\end{aligned}$$

This yields that

$$\left|H_\star^{-1/2} H(\theta, g) H_\star^{-1/2} - I_d\right|_2 \leq \frac{1}{4\lambda_{\min}(H_\star)}\left(r_1 M^3 + r_2 M^2\right) \leq \frac{1}{2},$$

and thus $H(\theta, g) \succeq I_d/2$ for all $(\theta, g) \in \Theta_{\bar{r}_1}(\theta_\star) \times \mathcal{G}_{r_2}(g_0)$. Analogously, we can show that

$$\left|H_\star^{-1/2} H(\theta_\star, g) H_\star^{-1/2} - I_d\right|_2 \leq \frac{r_2 M^2}{4\lambda_{\min}(H_\star)} \leq \frac{1}{4},$$

and thus the stability (5) holds true with $\kappa = 3/4$ and $\mathcal{K} = 5/4$. As for the Bernstein condition, we note that

$$|H(\theta, g; Z)|_2 = \left|\sigma(\theta^\top X + g(W))[1 - \sigma(\theta^\top X + g(W))]XX^\top\right|_2 \leq \frac{M^2}{4}.$$

It follows that

$$\left|H(\theta, g)^{-1/2} H(\theta, g; Z) H(\theta, g)^{-1/2} - I_d\right|_2 \leq \frac{M^2}{2\lambda_{\min}(H(\theta, g))} \leq M^2.$$

Due to (Wainwright, 2019, Equation 6.30), $H(\theta, g)^{-1/2} H(\theta, g; Z) H(\theta, g)^{-1/2} - I_d$ satisfies the Bernstein condition with parameter $K_2 \lesssim M^2$. Moreover, $\sigma_H^2 \lesssim M^4$. To summarize, invoking Theorem 5 gives the following risk bound up to a constant factor:

$$\frac{M^2}{\lambda_{\min}(H_\star)}\left[\frac{\bar{d}_\star}{n}\log\left(1/\delta\right) + \|\hat{g} - g_0\|_{\mathcal{G}}^2\right].$$

Remark 12 *Since the semi-parametric logistic loss does not satisfy the Neyman orthogonality, the results of Foster and Syrgkanis (2020) do not directly apply here.*

5. Conclusion

We established non-asymptotic guarantees in terms of the excess risk for the orthogonal statistical learning under pseudo self-concordance, allowing the dimension of the target parameter to grow at the rate $d = O(n^{1/2})$. The dimension dependency in our bound is characterized by the effective dimension—the trace of the sandwich covariance matrix—which recovers existing results in supervised learning without the nuisance parameter. Compared with previous work (Foster and Syrgkanis, 2020), our results improve on the excess risk bound at least by a factor of d in a wide range of eigendecay regimes. The extension of our theoretical analysis to handle sparse regularization is an interesting venue for future work.

Acknowledgments

The authors would like to thank Jon Wellner for fruitful discussions. L. Liu is supported by NSF CCF-2019844 and NSF DMS-2023166. Z. Harchaoui is supported by NSF CCF-2019844, NSF DMS-2134012, NSF DMS-2023166, CIFAR-LMB, and faculty research awards. Part of this work was done while Z. Harchaoui was visiting the Simons Institute for the Theory of Computing.

References

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4, 2010.

Francis Bach. *Learning Theory from First Principles*. Online version, 2021.

Patrice Bertail, Stephan Clémençon, Yannick Guyonvarch, and Nathan Noiry. Learning from biased data: A semi-parametric approach. In *ICML*, 2021.

Peter J Bickel, Chris AJ Klaassen, Ya'acov Ritov, and Jon A Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, 1998.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), 2018.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv preprint*, 2020.

Miguel A Hernán and James M Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Michael R Kosorok. *Introduction to Empirical Processes and Semiparametric Inference.* Springer, 2008.

Molei Liu, Yi Zhang, and Doudou Zhou. Double/debiased machine learning for logistic partially linear model. *The Econometrics Journal*, 24(3), 2021.

Lester Mackey, Vasilis Syrgkanis, and Ilias Zadik. Orthogonal machine learning: Power and limitations. In *ICML*, 2018.

Susan A Murphy and Aad W Van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95(450), 2000.

Denis Nekipelov, Vira Semenova, and Vasilis Syrgkanis. Regularised orthogonal machine learning for nonlinear semiparametric models. *The Econometrics Journal*, 25(1), 2022.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics, 1994.

Jerzy Neyman. Optimal asymptotic tests of composite hypotheses. *Probability and Statistics*, 1959.

Jerzy Neyman. $c(\alpha)$ tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A*, 41 (1/2), 1979.

Dmitrii M Ostrovskii and Francis Bach. Finite-sample analysis of M-estimators using self-concordance. *Electronic Journal of Statistics*, 15(1), 2021.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

Alain Rakotomamonjy, Stéphane Canu, and Alex Smola. Frames, reproducing kernels, regularization and learning. *Journal of Machine Learning Research*, 6(9), 2005.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 1983.

David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric Regression*. Cambridge University Press, 2003.

Ilya Shpitser, Tyler VanderWeele, and James M Robins. On the validity of covariate adjustment for estimating causal effects. In *UAI*, 2010.

Alex Smola, Thilo T Frieß, and Bernhard Schölkopf. Semiparametric support vector and linear programming machines. In *NIPS*, 1998.

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

Mark J Van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Jon Wakefield. *Bayesian and Frequentist Regression Methods*. Springer, 2013.

Jon A Wellner and Ying Zhang. Two likelihood-based semiparametric estimation methods for panel count data with covariates. *The Annals of Statistics*, 35(5), 2007.

Yingcun Xia and Wolfgang Härdle. Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis*, 97(5), 2006.

The Appendix is organized as follows. For simplicity, we first prove in Appendix A the main results assuming that the true nuisance parameter g_0 is known. We then prove in Appendix B the main results presented in Section 2. The technical tools used in the proofs are reviewed and developed in Appendix C.

Appendix A. Risk Bound with Known Nuisance Parameter

In this section, we assume that the true nuisance parameter g_0 is known and control the excess risk. The proofs in this section are inspired by and extend those from Ostrovskii and Bach (2021). We denote by $\hat{\theta}_0$ the minimizer of the empirical risk $L_n(\theta, g_0)$. Our analysis is local to $\theta_\star$, in other words, we make the following assumption on $\hat{\theta}_0$. Recall that $\Theta_r(\theta_\star) := \{\theta \in \Theta : \|\theta - \theta_\star\|_{H_\star} < r\}$.

Assumption 6 *Let $r_0 > 0$ be a constant and $\tilde{r}_0 := \min\{r_0, r_0\sqrt{\lambda_{\min}(H_\star)}\} > 0$. There exists a function $N_{\tilde{r}_0} : [0,1] \to \mathbb{N}_+$ such that for any $\delta \in (0,1)$ we have, with probability at least $1-\delta$, $\hat{\theta}_0 \in \Theta_{\tilde{r}_0}(\theta_\star)$ for all $n \geq N_{\tilde{r}_0}(\delta)$.*

Control of the score. In order to control the score, we assume that the normalized score at $\theta_\star$ is sub-Gaussian.

Assumption 7 *The normalized score at θ_0 is sub-Gaussian, i.e., there exists a constant $K_{1,0} > 0$ such that*

$$\left\| G_\star^{-1/2}[S(\theta_\star, g_0; Z) - S(\theta_\star, g_0)] \right\|_{\psi_2} \leq K_{1,0}.$$

Recall that $d_\star := \mathbf{Tr}(\Omega_\star)$ and $\Omega_\star := H_\star^{-1/2} G_\star H_\star^{-1/2}$.

Proposition 13 *Under Assumption 7, it holds that, with probability at least $1-\delta$,*

$$\|S_n(\theta_\star, g_0)\|_{H_\star^{-1}}^2 \lesssim \frac{1}{n}[d_\star + K_{1,0}^2 \log(e/\delta) \|\Omega_\star\|_2],$$

where $\lesssim$ hides an absolute constant.

Proof By the first order optimality condition, we have $S(\theta_\star, g_0) = 0$. As a result,

$$X := \sqrt{n} G_\star^{-1/2} S_n(\theta_\star, g_0; Z)$$

is an isotropic random vector. Moreover, it follows from Lemma 22 that $\|X\|_{\psi_2} \lesssim K_{1,0}$. Define $J := G_\star^{1/2} H_\star^{-1} G_\star^{1/2}/n$. Then we have

$$\|S_n(\theta_\star, g_0)\|_{H_\star^{-1}}^2 = \|X\|_J^2 .$$

Invoking Theorem 23 yields the claim. ■

Control of the Hessian. In order to control the Hessian, we use the pseudo self-concordance as in Definition 2.

Assumption 8 *For any $z \in \mathcal{Z}$, $\ell(\theta, g_0; z)$ is pseudo self-concordant on $\Theta_{\tilde{r}_0}(\theta_\star)$, i.e.,*

$$\left|\mathrm{D}_\theta^3 \ell(\theta, g_0; z)[u,u,u]\right| \leq R_0 \left\|u\right\|_2 \mathrm{D}_\theta^2 \ell(\theta, g_0; z)[u,u], \quad \textit{for all } \theta \in \Theta_{\tilde{r}_0}(\theta_\star), u \in \mathbb{R}^d.$$

Moreover, $L(\theta, g_0)$ is pseudo self-concordant on $\Theta_{\tilde{r}_0}(\theta_\star)$.

We also assume that the Hessian $H(\theta, g_0; Z)$ satisfies the Bernstein condition uniformly over $\Theta_{\tilde{r}_0}(\theta_\star)$.

Assumption 9 *For any $\theta \in \Theta_{\tilde{r}_0}(\theta_\star)$, the centered sandwich Hessian*

$$H(\theta, g_0)^{-1/2} H(\theta, g_0; Z) H(\theta, g_0)^{-1/2} - I_d$$

satisfies a Bernstein condition with parameter $K_{2,0}$. Moreover,

$$\sigma_{H,0}^2 := \sup_{\theta \in \Theta_{\tilde{r}_0}(\theta_\star)} \left|\mathbb{V}\mathrm{ar}(H(\theta, g_0)^{-1/2} H(\theta, g_0; Z) H(\theta, g_0)^{-1/2})\right|_2 < \infty.$$

Proposition 14 *Under Assumption 8, for any $\theta \in \Theta_{\tilde{r}_0}(\theta_\star)$, we have*

$$e^{-R_0 r_0} H_\star \preceq H(\theta, g_0) \preceq e^{R_0 r_0} H_\star. \tag{14}$$

Moreover, if Assumption 9 holds true, then, with probability at least $1-\delta$, we have

$$\frac{1}{4e^{R_0 r_0}} H_\star \preceq H_n(\theta, g_0) \preceq 3e^{R_0 r_0} H_\star, \quad \textit{for all } \theta \in \Theta_{\tilde{r}_0}(\theta_\star), \tag{15}$$

whenever $n \geq 16(K_{2,0}^2 + 2\sigma_{H,0}^2)\left[\log(4d/\delta) + d\log(3r_0 R_0/\log 2)\right]^2$.

Proof According to Assumption 8 and Proposition 25, we have

$$e^{-R_0\|\theta-\theta_\star\|_2} H_\star \preceq H(\theta, g_0) \preceq e^{R_0\|\theta-\theta_\star\|_2} H_\star.$$

Hence, the claim (14) follows from $\left\|\theta - \theta_\star\right\|_2 \leq \tilde{r}_0/\sqrt{\lambda_{\min}(H_\star)} \leq r_0$. As for (15), we prove it in the following steps.

Step 1. Let $\epsilon = \sqrt{\lambda_{\min}(H_\star)}\log 2/R_0$. Take an ϵ-covering $\mathcal{N}_\epsilon$ of $\Theta_{\tilde{r}_0}(\theta_\star)$ w.r.t. $\left\|\cdot\right\|_{H_\star}$, and let $\pi(\theta)$ be the projection of θ onto $\mathcal{N}_\epsilon$. By the self-concordance of $\ell(\cdot, g_0; z)$ (Assumption 8), we have, for all $\theta \in \Theta_{\tilde{r}_0}(\theta_\star)$,

$$e^{-R_0 r} H(\pi(\theta), g_0; Z) \preceq H(\theta, g_0; Z) \preceq e^{R_0 r} H(\pi(\theta), g_0; Z),$$

where $r := \left\|\theta - \pi(\theta)\right\|_2 \leq \epsilon/\sqrt{\lambda_{\min}(H_\star)} = \log 2/R_0$. It then follows that

$$\frac{1}{2} H(\pi(\theta), g_0; Z_i) \preceq H(\theta, g_0; Z_i) \preceq 2H(\pi(\theta), g_0; Z_i), \quad \text{for all } \theta \in \Theta_{\tilde{r}_0}(\theta_\star) \text{ and } i \in [n],$$

which yields

$$\frac{1}{2}H_n(\pi(\theta), g_0) \preceq H_n(\theta, g_0) \preceq 2H_n(\pi(\theta), g_0), \quad \text{for all } \theta \in \Theta_{\tilde{r}_0}(\theta_\star). \tag{16}$$

Step 2. By Theorem 24, for each $\theta \in \Theta_{\tilde{r}_0}(\theta_\star)$, it holds that, with probability at least $1-\delta$,

$$\left| H(\theta, g_0)^{-1/2} H_n(\theta, g_0) H(\theta, g_0)^{-1/2} - I_d \right|_2 \leq \frac{1}{2},$$

or equivalently,

$$\frac{1}{2}H(\theta, g_0) \preceq H_n(\theta, g_0) \preceq \frac{3}{2}H(\theta, g_0)$$

whenever $n \geq 16(K_{2,0}^2 + 2\sigma_{H,0}^2)\log^2(2d/\delta)$. Since $|\mathcal{N}_\epsilon| \leq (3\tilde{r}_0/\epsilon)^d$ (Ostrovskii and Bach, 2021), by a union bound, we get, with probability at least $1-\delta/2$,

$$\frac{1}{2}H(\pi(\theta), g_0) \preceq H_n(\pi(\theta), g_0) \preceq \frac{3}{2}H(\pi(\theta), g_0), \quad \text{for all } \theta \in \Theta_{\tilde{r}_0}(\theta_\star), \tag{17}$$

whenever $n \geq 16(K_{2,0}^2 + 2\sigma_{H,0}^2)\left[\log(4d/\delta) + d\log(3r_0R_0/\log 2)\right]^2$. Hence, the statement (15) follows from (14), (16), and (17). ■

Control of the excess risk. The next theorem shows that the excess risk is upper bounded by $d_\star/n$ up to a constant factor.

Theorem 15 *Under Assumptions 6-9, with probability at least* $1-\delta$, *the excess risk of* $\hat{\theta}_0$ *satisfies*

$$\mathcal{E}(\hat{\theta}_0, g_0) \lesssim K_{1,0}^2 e^{3R_0r_0}\log(1/\delta)\frac{d_\star}{n} \tag{18}$$

whenever $n \geq \max\{N_{\tilde{r}_0}(\delta/3), 16(K_{2,0}^2 + 2\sigma_{H,0}^2)\left[\log(12d/\delta) + d\log(3r_0R_0/\log 2)\right]^2\}$.

Proof We start by defining three events. Let

$$\begin{aligned}
\mathcal{A} &:= \left\{\hat{\theta}_0 \in \Theta_{\tilde{r}_0}(\theta_\star)\right\} \\
\mathcal{B} &:= \left\{\frac{1}{4e^{R_0r_0}}H_\star \preceq H_n(\theta, g_0) \preceq 3e^{R_0r_0}H_\star, \quad \text{for all } \theta \in \Theta_{\tilde{r}_0}(\theta_\star)\right\} \\
\mathcal{C} &:= \left\{\|S_n(\theta_\star, g_0)\|_{H_\star^{-1}} \lesssim \sqrt{\frac{d_\star + K_{1,0}^2\log(3e/\delta)\|\Omega_\star\|_2}{n}}\right\}.
\end{aligned}$$

In the following, we let

$$n \geq \max\{N_{\tilde{r}_0}(\delta/3), 16(K_{2,0}^2 + 2\sigma_{H,0}^2)\left[\log(12d/\delta) + d\log(3r_0R_0/\log 2)\right]^2\}.$$

According to Assumption 6, we have $\mathbb{P}(\mathcal{A}) \geq 1-\delta/3$. By Proposition 14, it holds that $\mathbb{P}(\mathcal{B}) \geq 1-\delta/3$. Finally, it follows from Proposition 13 that $\mathbb{P}(\mathcal{C}) \geq 1-\delta/3$.

Now, we prove the upper bound (18) on the event $\mathcal{ABC}$. By Taylor's theorem,

$$\mathcal{E}(\hat{\theta}_0, g_0) := L(\hat{\theta}_0, g_0) - L(\theta_\star, g_0) = S(\theta_\star, g_0)^\top(\hat{\theta}_0 - \theta_\star) + \frac{1}{2}\left\|\hat{\theta}_0 - \theta_\star\right\|^2_{H(\bar{\theta}, g_0)}$$

for some $\bar{\theta} \in \mathrm{Conv}\{\hat{\theta}_0, \theta_\star\} \subset \Theta_{\tilde{r}_0}(\theta_\star)$. According to (14), it holds that

$$\left\|\hat{\theta}_0 - \theta_\star\right\|^2_{H(\bar{\theta}, g_0)} \le e^{R_0 r_0}\left\|\hat{\theta}_0 - \theta_\star\right\|^2_{H_\star}.$$

By the first order optimality condition, we have $S(\theta_\star, g_0) = 0$. As a result,

$$\mathcal{E}(\hat{\theta}_0, g_0) \le \frac{1}{2}e^{R_0 r_0}\left\|\hat{\theta}_0 - \theta_\star\right\|^2_{H_\star}.$$

It then suffices to upper bound $\left\|\hat{\theta}_0 - \theta_\star\right\|_{H_\star}$.

Note that, by Taylor's theorem,

$$L_n(\hat{\theta}_0, g_0) - L_n(\theta_\star, g_0) = S_n(\theta_\star, g_0)^\top(\hat{\theta}_0 - \theta_\star) + \frac{1}{2}\left\|\hat{\theta}_0 - \theta_\star\right\|^2_{H_n(\bar{\theta}, g_0)}$$

for some $\bar{\theta} \in \mathrm{Conv}\{\hat{\theta}_0, \theta_\star\} \subset \Theta_{\tilde{r}_0}(\theta_\star)$. On the event $\mathcal{B}$, it holds that

$$\left\|\hat{\theta}_0 - \theta_\star\right\|^2_{H_n(\bar{\theta}, g_0)} \ge \frac{1}{4e^{R_0 r_0}}\left\|\hat{\theta}_0 - \theta_\star\right\|^2_{H_\star}.$$

Moreover, by the Cauchy-Schwarz inequality,

$$S_n(\theta_\star, g_0)^\top(\hat{\theta}_0 - \theta_\star) \ge -\left\|S_n(\theta_\star, g_0)\right\|_{H_\star^{-1}}\left\|\hat{\theta}_0 - \theta_\star\right\|_{H_\star}.$$

On the event $\mathcal{C}$, we get

$$\left\|S_n(\theta_\star, g_0)\right\|_{H_\star^{-1}} \lesssim \frac{\sqrt{d_\star + K_{1,0}^2 \log(3e/\delta)\left\|\Omega_\star\right\|_2}}{\sqrt{n}}.$$

Due to the optimality of $\hat{\theta}_0$, we also have $L_n(\hat{\theta}_0, g_0) - L_n(\theta_\star, g_0) \le 0$. Consequently,

$$\frac{1}{4e^{R_0 r_0}}\left\|\hat{\theta}_0 - \theta_\star\right\|^2_{H_\star} \lesssim \frac{\sqrt{d_\star + K_{1,0}^2 \log(3e/\delta)\left\|\Omega_\star\right\|_2}}{\sqrt{n}}\left\|\hat{\theta}_0 - \theta_\star\right\|_{H_\star}.$$

It then follows that

$$\mathcal{E}(\hat{\theta}_0, g_0) \le \frac{e^{R_0 r_0}}{2}\left\|\hat{\theta}_0 - \theta_\star\right\|^2_{H_\star} \lesssim \frac{d_\star + K_{1,0}^2 \log(3e/\delta)\left\|\Omega_\star\right\|_2}{e^{-3R_0 r_0} n} \lesssim K_{1,0}^2 e^{3R_0 r_0}\log(1/\delta)\frac{d_\star}{n}.$$

Therefore, the claim (18) holds with probability at least

$$\mathbb{P}(\mathcal{ABC}) = 1 - \mathbb{P}(\mathcal{A}^c \cup \mathcal{B}^c \cup \mathcal{C}^c) \ge 1 - \mathbb{P}(\mathcal{A}^c) - \mathbb{P}(\mathcal{B}^c) - \mathbb{P}(\mathcal{C}^c) \ge 1 - \delta.$$

■

Remark 16 *Our results generalize the results (Ostrovskii and Bach, 2021, Theorem 4.1) which were developed for parametric linear models, i.e., when the loss is given by $\ell(\theta; Z) := \ell(Y, X^\top\theta)$. The paper of Ostrovskii and Bach (2021) relies heavily on the special structure of the Hessian. Our results apply to a broader class of models owing to the matrix Bernstein inequality.*

Appendix B. Proof of Theorem 4

We then consider the case when the true nuisance parameter g_0 is unknown and estimated from a separate sample as in the OSL meta-algorithm. Again, our analysis is local to both $\theta_\star$ and g_0. The independence between $\hat{g}$ and the sample $\{Z_i\}_{i=1}^n$ greatly simplifies our analysis according to Lemma 6 in Section 2.

Proof [Proof of Lemma 6] By the independence between $\hat{g}$ and $\{Z_i\}_{i=1}^n$, we have

$$\mathbb{E}[\mathbb{1}\{\mathcal{A}(\hat{g}, \{Z_i\}_{i=1}^n)\} \mid \hat{g}](g) = \mathbb{E}[\mathbb{1}\{\mathcal{A}(g, \{Z_i\}_{i=1}^n)\}] \geq 1 - \delta, \quad \text{for any } g \in \mathcal{G}'.$$

By the tower property of the conditional expectation,

$$\begin{aligned}\mathbb{P}(\mathcal{A}(\hat{g}, \{Z_i\}_{i=1}^n)) &= \mathbb{E}\left[\mathbb{E}[\mathbb{1}\{\mathcal{A}(\hat{g}, \{Z_i\}_{i=1}^n)\} \mid \hat{g}]\right] \\ &\geq \mathbb{E}\left[\mathbb{E}[\mathbb{1}\{\mathcal{A}(\hat{g}, \{Z_i\}_{i=1}^n)\} \mid \hat{g}]\mathbb{1}\{\hat{g} \in \mathcal{G}'\}\right] \geq (1-\delta)\,\mathbb{P}(\hat{g} \in \mathcal{G}').\end{aligned}$$

■

Control of the score. Recall that $\bar{d}_\star := \sup_{g \in \mathcal{G}_{r_2}(g_0)} \mathbf{Tr}(H_\star^{-1/2} G(\theta_\star, g) H_\star^{-1/2})$.

Proof [Proof of Proposition 7] Define $W := \sqrt{n} G(\theta_\star, g)^{-1/2} \bar{S}_n(\theta_\star, g)$ where

$$\bar{S}_n(\theta_\star, g) := S_n(\theta_\star, g) - S(\theta_\star, g).$$

It is straightforward to check that W is isotropic. Moreover, it follows from Lemma 22 that $\|W\|_{\psi_2} \leq K_1$. Let $J := G(\theta_\star, g)^{1/2} H_\star^{-1} G(\theta_\star, g)^{1/2}$. By Theorem 23, we have, with probability at least $1 - \delta$,

$$\|W\|_J^2 \lesssim K_1^2 \log(e/\delta)\, \mathbf{Tr}(J) \leq K_1^2 \log(e/\delta) \bar{d}_\star.$$

The statement then follows from the fact that

$$\|W\|_J^2 = W^\top J W = n \bar{S}_n(\theta_\star, g)^\top H_\star^{-1} \bar{S}_n(\theta_\star, g) = n \left\|S_n(\theta_\star, g) - S(\theta_\star, g)\right\|_{H_\star^{-1}}^2.$$

■

Proof [Proof of Lemma 8] By Taylor's theorem,

$$\begin{aligned} S(\theta_\star, g)^\top (\hat{\theta} - \theta_\star) &= \mathrm{D}_\theta L(\theta_\star, g)[\hat{\theta} - \theta_\star] \\ &= \mathrm{D}_\theta L(\theta_\star, g_0)[\hat{\theta} - \theta_0] + \mathrm{D}_g \mathrm{D}_\theta L(\theta_\star, g_0)[\hat{\theta} - \theta_\star, g - g_0] + \\ &\quad \frac{1}{2} \mathrm{D}_g^2 \mathrm{D}_\theta L(\theta_\star, \bar{g})[\hat{\theta} - \theta_\star, g - g_0, g - g_0] \\ &\geq -\frac{\beta_2}{2} \left\|\hat{\theta} - \theta_\star\right\|_{H_\star} \|g - g_0\|_{\mathcal{G}}^2, \end{aligned}$$

where the last inequality follows from the first order optimality condition and Assumption 3b. ■

Proof [Proof of Lemma 10] By Taylor's theorem,

$$\begin{aligned} S(\theta_\star, g)^\top (\hat{\theta} - \theta_\star) &= \mathrm{D}_\theta L(\theta_\star, g)[\hat{\theta} - \theta_\star] \\ &= \mathrm{D}_\theta L(\theta_\star, g_0)[\hat{\theta} - \theta_0] + \mathrm{D}_g \mathrm{D}_\theta L(\theta_\star, \bar{g})[\hat{\theta} - \theta_\star, g - g_0] \\ &\geq -\beta_1 \left\|\hat{\theta} - \theta_\star\right\|_{H_\star} \|g - g_0\|_{\mathcal{G}}, \end{aligned}$$

where the last inequality follows from the first order optimality condition and Assumption 3a. ■

Control of the Hessian. We then prove Proposition 9.

Proof [Proof Proposition 9] Fix an arbitrary $g \in \mathcal{G}_{r_2}(g_0)$. *Step 1.* Invoking Proposition 25 leads to

$$e^{-Rr} H(\theta_\star, g) \preceq H(\theta, g) \preceq e^{Rr} H(\theta_\star, g), \quad \text{for all } \theta \in \Theta_{\bar{r}_1}(\theta_\star),$$

where $r := \|\theta - \theta_\star\|_2 \le \bar{r}_1/\sqrt{\lambda_{\min}(H_\star)} \le r_1$. Consequently, by (5), we have, for all $\theta \in \Theta_{\bar{r}_1}(\theta_\star)$,

$$\kappa e^{-Rr_1} H_\star \preceq e^{-Rr_1} H(\theta_\star, g) \preceq H(\theta, g) \preceq e^{Rr_1} H(\theta_\star, g) \preceq \mathcal{K} e^{Rr_1} H_\star. \tag{19}$$

Step 2. Let $\epsilon := \sqrt{\lambda_{\min}(H_\star)} \log 2/R$. Take an ϵ-covering $\mathcal{N}_\epsilon$ of $\Theta_{\bar{r}_1}(\theta_\star)$ w.r.t. $\|\cdot\|_{H_\star}$, and let $\pi(\theta)$ be the projection of θ onto $\mathcal{N}_\epsilon$. By Assumption 4, we have, for all $\theta \in \Theta_{\bar{r}_1}(\theta_\star)$,

$$e^{-Rr'} H(\pi(\theta), g; Z) \preceq H(\theta, g; Z) \preceq e^{Rr'} H(\pi(\theta), g; Z),$$

where $r' := \|\theta - \pi(\theta)\|_2 \le \epsilon/\sqrt{\lambda_{\min}(H_\star)} = \log 2/R$. This implies that

$$\frac{1}{2} H(\pi(\theta), g; Z_i) \preceq H_n(\theta, g; Z_i) \preceq 2H_n(\pi(\theta), g; Z_i), \quad \text{for all } \theta \in \Theta_{\bar{r}_1}(\theta_\star) \text{ and } i \in [n].$$

Hence,

$$\frac{1}{2} H_n(\pi(\theta), g) \preceq H_n(\theta, g) \preceq 2H_n(\pi(\theta), g), \quad \text{for all } \theta \in \Theta_{\bar{r}_1}(\theta_\star). \tag{20}$$

Step 3. By Theorem 24, for each $\theta \in \Theta_{\bar{r}_1}(\theta_\star)$, it holds that, with probability at least $1 - \delta$,

$$\frac{1}{2} H(\theta, g) \preceq H_n(\theta, g) \preceq \frac{3}{2} H(\theta, g)$$

whenever $n \ge 16(K_2^2 + 2\sigma_H^2) \log^2(2d/\delta)$. Since $|\mathcal{N}_\epsilon| \le (3\bar{r}_1/\epsilon)^d$ (Ostrovskii and Bach, 2021), by a union bound, we get, with probability at least $1 - \delta/2$,

$$\frac{1}{2} H(\pi(\theta), g) \preceq H_n(\pi(\theta), g) \preceq \frac{3}{2} H(\pi(\theta), g), \quad \text{for all } \theta \in \Theta_{\bar{r}_1}(\theta_\star), \tag{21}$$

whenever $n \ge 16(K_2^2 + 2\sigma_H^2)[\log(4d/\delta) + d\log(3Rr_1/\log 2)]^2$. Hence, the claim follows from (19), (20), and (21). ■

Control of the excess risk. Now we are ready to prove Theorem 4.

Proof [Proof of Theorem 4] Fix an arbitrary $g \in \mathcal{G}_{r_2}(g_0)$. We start by defining three events. Let

$$\begin{aligned}
\mathcal{A} &:= \left\{\hat{\theta} \in \Theta_{\bar{r}_1}(\theta_\star), \hat{g} \in \mathcal{G}_{r_2}(g_0)\right\} \\
\mathcal{B}(g) &:= \left\{\frac{\kappa}{4e^{Rr_1}} H_\star \preceq H_n(\theta, g) \preceq 3\mathcal{K} e^{Rr_1} H_\star, \quad \text{for all } \theta \in \Theta_{\bar{r}_1}(\theta_\star)\right\} \\
\mathcal{C}(g) &:= \left\{\|S_n(\theta_\star, g) - S(\theta_\star, g)\|_{H_\star^{-1}} \lesssim \sqrt{\frac{K_1^2 \log(5e/\delta) \bar{d}_\star}{n}}\right\}.
\end{aligned}$$

In the following, we let

$$n \ge \max\{N_{\bar{r}_1, r_2}(\delta/5), 16(K_2^2 + 2\sigma_H^2)[\log(20d/\delta) + d\log(3Rr_1/\log 2)]^2\}.$$

According to Assumption 1, we have $\mathbb{P}(\mathcal{A}) \geq 1 - \delta/5$. By Propositions 7 and 9, it holds that $\mathbb{P}(\mathcal{C}(g)) \geq 1 - \delta/5$ and $\mathbb{P}(\mathcal{B}(g)) \geq 1 - \delta/5$, respectively. Since $g \in \mathcal{G}_{r_2}(g_0)$ is arbitrary, it follows from Lemma 6 that

$$\mathbb{P}(\mathcal{B}(\hat{g})) \geq (1 - \delta/5)\, \mathbb{P}(\hat{g} \in \mathcal{G}_{r_2}(g_0)) \geq 1 - 2\delta/5.$$

Similarly, $\mathbb{P}(\mathcal{C}(\hat{g})) \geq 1 - 2\delta/5$.

Now, we prove the upper bound (18) on the event $\mathcal{A}\mathcal{B}(\hat{g})\mathcal{C}(\hat{g})$. By Taylor's theorem,

$$\mathcal{E}(\hat{\theta}, g_0) := L(\hat{\theta}, g_0) - L(\theta_\star, g_0) = S(\theta_\star, g_0)^\top (\hat{\theta} - \theta_\star) + \frac{1}{2} \left\| \hat{\theta} - \theta_\star \right\|^2_{H(\bar{\theta}, g_0)}$$

for some $\bar{\theta} \in \mathrm{Conv}\{\hat{\theta}, \theta_\star\} \subset \Theta_{r_1}(\theta_\star)$. According to (14), it holds that

$$\left\| \hat{\theta} - \theta_\star \right\|^2_{H(\bar{\theta}, g_0)} \leq e^{Rr_1} \left\| \hat{\theta} - \theta_\star \right\|^2_{H_\star}.$$

By the first order optimality condition, we have $S(\theta_\star, g_0) = 0$. As a result,

$$\mathcal{E}(\hat{\theta}, g_0) \leq \frac{e^{Rr_1}}{2} \left\| \hat{\theta} - \theta_\star \right\|^2_{H_\star}.$$

It then suffices to upper bound $\left\| \hat{\theta} - \theta_\star \right\|_{H_\star}$.

Note that, by Taylor's theorem,

$$L_n(\hat{\theta}, \hat{g}) - L_n(\theta_\star, \hat{g}) = S_n(\theta_\star, \hat{g})^\top (\hat{\theta} - \theta_\star) + \frac{1}{2} \left\| \hat{\theta} - \theta_\star \right\|^2_{H_n(\bar{\theta}, \hat{g})}$$

for some $\bar{\theta} \in \mathrm{Conv}\{\hat{\theta}, \theta_\star\} \subset \Theta_{\bar{r}_1}(\theta_\star)$. On the event $\mathcal{B}(\hat{g})$, it holds that

$$\left\| \hat{\theta} - \theta_\star \right\|^2_{H_n(\bar{\theta}, \hat{g})} \geq \frac{\kappa}{4e^{Rr_1}} \left\| \hat{\theta} - \theta_\star \right\|^2_{H_\star}.$$

Moreover,

$$\begin{aligned}
&S_n(\theta_\star, \hat{g})^\top (\hat{\theta} - \theta_\star) \\
&= \bar{S}_n(\theta_\star, \hat{g})^\top (\hat{\theta} - \theta_\star) + S(\theta_\star, \hat{g})^\top (\hat{\theta} - \theta_\star) \\
&\geq \bar{S}_n(\theta_\star, \hat{g})^\top (\hat{\theta} - \theta_\star) - \frac{\beta_2}{2} \left\| \hat{\theta} - \theta_\star \right\|_{H_\star} \|\hat{g} - g_0\|^2_{\mathcal{G}}, \quad \text{by Lemma 8} \\
&\geq - \left\| \bar{S}_n(\theta_\star, \hat{g}) \right\|_{H_\star^{-1}} \left\| \hat{\theta} - \theta_\star \right\|_{H_\star} - \frac{\beta_2}{2} \left\| \hat{\theta} - \theta_\star \right\|_{H_\star} \|\hat{g} - g_0\|^2_{\mathcal{G}}, \quad \text{by the Cauchy-Schwarz} \\
&\gtrsim - \left\| \hat{\theta} - \theta_\star \right\|_{H_\star} \left[\sqrt{\frac{K_1^2 \log(5e/\delta) \bar{d}_\star}{n}} + \frac{\beta_2}{2} \|\hat{g} - g_0\|^2_{\mathcal{G}} \right], \quad \text{by the event } \mathcal{C}(\hat{g}).
\end{aligned}$$

Due to the optimality of $\hat{\theta}$, we also have $L_n(\hat{\theta}, \hat{g}) - L_n(\theta_\star, \hat{g}) \leq 0$. Consequently,

$$\frac{\kappa}{4e^{Rr_1}} \left\| \hat{\theta} - \theta_\star \right\|^2_{H_\star} \lesssim \left[\sqrt{\frac{K_1^2 \log(5e/\delta) \bar{d}_\star}{n}} + \frac{\beta_2}{2} \|\hat{g} - g_0\|^2_{\mathcal{G}} \right] \left\| \hat{\theta} - \theta_\star \right\|_{H_\star}.$$

It then follows that

$$\mathcal{E}(\hat{\theta}, g_0) \leq \frac{e^{Rr_1}}{2} \left\| \hat{\theta} - \theta_\star \right\|_{H_\star}^2$$
$$\lesssim \frac{e^{3Rr_1}}{\kappa^2} \left[\frac{K_1^2 \log(1/\delta) \bar{d}_\star}{n} + \beta_2^2 \left\| \hat{g} - g_0 \right\|_{\mathcal{G}}^4 \right].$$

Therefore, the claim holds with probability at least

$$\mathbb{P}(\mathcal{A}\mathcal{B}(\hat{g})\mathcal{C}(\hat{g})) = 1 - \mathbb{P}(\mathcal{A}^c \cup \mathcal{B}(\hat{g})^c \cup \mathcal{C}(\hat{g})^c) \geq 1 - \mathbb{P}(\mathcal{A}^c) - \mathbb{P}(\mathcal{B}(\hat{g})^c) - \mathbb{P}(\mathcal{C}(\hat{g})^c) \geq 1 - \delta.$$

■

Proof [Proof of Theorem 5] The proof is largely similar to the one of Theorem 4. The only difference is that Lemma 8 is replaced by Lemma 10 at places where it is invoked. ■

Appendix C. Technical Tools

In this section, we give the precise definitions of sub-Gaussian random vectors (Vershynin, 2018, Chapter 3.4) and the matrix Bernstein condition (Wainwright, 2019, Chapter 6.4). We then review and prove some key results that are used in our analysis. Finally, we recall a proposition for the pseudo self-concordance.

Definition 17 (Sub-Gaussian vector) *Let $S \in \mathbb{R}^d$ be a mean-zero random vector. We say S is sub-Gaussian if $\langle S, s \rangle$ is sub-Gaussian for every $s \in \mathbb{R}^d$. Moreover, we define the sub-Gaussian norm of S as*

$$\|S\|_{\psi_2} := \sup_{\|s\|_2 = 1} \|\langle S, s \rangle\|_{\psi_2}.$$

Note that $\|\cdot\|_{\psi_2}$ is a norm and satisfies, e.g., the triangle inequality.

Definition 18 (Matrix Bernstein condition) *Let $H \in \mathbb{R}^{d \times d}$ be a zero-mean symmetric random matrix. We say H satisfies a Bernstein condition with parameter $b > 0$ if, for all $j \geq 3$,*

$$\mathbb{E}[H^j] \preceq \frac{1}{2} j! b^{j-2} \, \mathbb{V}\mathrm{ar}(H).$$

Lemma 19 *If $X \in \mathbb{R}^d$ is a bounded random vector with $\|X\|_2 \leq_{a.s.} M < \infty$, then X is a sub-Gaussian random vector with*

$$\|X\|_{\psi_2} \lesssim M \quad \textit{and} \quad \|X - \mathbb{E}[X]\|_{\psi_2} \lesssim M.$$

Proof For any $x \in \mathbb{R}^d$, we have $\langle X, x \rangle \leq \|X\|_2 \|x\| \leq M$. Hence, by definition, X is a sub-Gaussian random vector with

$$\|X\|_{\psi_2} = \sup_{\|x\|=1} \|\langle X, x \rangle\|_{\psi_2} \lesssim M.$$

Moreover, since $\|\cdot\|_{\psi_2}$ is a norm, we have

$$\|X - \mathbb{E}[X]\|_{\psi_2} \leq \|X\|_{\psi_2} + \|\mathbb{E}[X]\|_{\psi_2}.$$

Note that $\|a\|_{\psi_2} \lesssim \|a\|_2$ for a constant vector a. Hence, we have $\|X - \mathbb{E}[X]\|_{\psi_2} \lesssim M$. ■

Lemma 20 *Let $X \in \mathbb{R}^d$ be a random vector with $\|X\|_2 \leq_{a.s.} M < \infty$ and $Y \in \mathbb{R}$ be a sub-Gaussian random variable. Then XY is a sub-Gaussian random vector with*

$$\|XY\|_{\psi_2} \leq M \|Y\|_{\psi_2}.$$

Proof By the definition of sub-Gaussian random variables, we have

$$\mathbb{E}[\exp(Y^2/\|Y\|_{\psi_2}^2)] \leq 2. \tag{22}$$

It then follows that, for any $\|x\| = 1$,

$$\mathbb{E}[\exp((x^\top X)^2 Y^2/M^2 \|Y\|_{\psi_2}^2)] \leq \mathbb{E}[\exp(\|X\|_2^2 Y^2/M^2 \|Y\|_{\psi_2}^2)] \leq 2,$$

and thus $\|XY\|_{\psi_2} \leq M \|Y\|_{\psi_2}$. ■

Lemma 21 *Let $X \in \mathbb{R}^d$ be a sub-Gaussian random vector and $A \in \mathbb{R}^{d\times d}$ be a fixed matrix. Then AX is a sub-Gaussian random vector with*

$$\|AX\|_{\psi_2} \leq \|A\|_2 \|X\|_{\psi_2}.$$

Proof Take an arbitrary $\|x\|_2 = 1$. It holds that

$$\begin{aligned}
\mathbb{E}[\exp(\lambda x^\top AX)] &= \mathbb{E}\left[\exp\left(\lambda \left\|A^\top x\right\|_2 \left(\frac{A^\top x}{\|A^\top x\|_2}\right)^\top X\right)\right] \\
&\leq \exp\left(\left\|A^\top x\right\|_2^2 \|X\|_{\psi_2}^2 \lambda^2\right), \quad \text{by the sub-Gaussianity of } X \\
&\leq \exp\left(\|A\|_2^2 \|X\|_{\psi_2}^2 \lambda^2\right).
\end{aligned}$$

Hence, we obtain $\|AX\|_{\psi_2} \leq \|A\|_2 \|X\|_{\psi_2}$. ■

The sum of i.i.d. sub-Gaussian vectors is also sub-Gaussian according to the following lemma.

Lemma 22 (Vershynin (2018), Lemma 5.9) *Let $X_1, \ldots, X_n$ be i.i.d. random vectors, then we have $\|\sum_{i=1}^n X_i\|_{\psi_2}^2 \lesssim \sum_{i=1}^n \|X_i\|_{\psi_2}^2$.*

We call a random vector $X \in \mathbb{R}^d$ isotropic if $\mathbb{E}[X] = 0$ and $\mathbb{E}[XX^\top] = I_d$. The following theorem is a tail bound for quadratic forms of isotropic sub-Gaussian random vectors.

Theorem 23 (Ostrovskii and Bach (2021), Theorem A.1) *Let $X \in \mathbb{R}^d$ be an isotropic random vector with $\|X\|_{\psi_2} \le K$, and let $J \in \mathbb{R}^{d\times d}$ be positive semi-definite. Then, with probability at least $1-\delta$, it holds that*

$$\|X\|_J^2 - Tr(J) \lesssim K^2 \left(\|J\|_2 \sqrt{\log(e/\delta)} + \|J\|_\infty \log(1/\delta) \right). \tag{23}$$

A zero-mean symmetric random matrix Q is said to be sub-Gaussian with parameter V if $\mathbb{E}[e^{\lambda Q}] \preceq e^{\lambda^2 V/2}$ for all $\lambda \in \mathbb{R}$. The next theorem is the Bernstein bound for random matrices.

Theorem 24 (Wainwright (2019), Theorem 6.17) *Let $\{Q_i\}_{i=1}^n$ be a sequence of zero-mean independent symmetric random matrices that satisfies the Bernstein condition with parameter $b > 0$. Then, for all $\delta > 0$, it holds that*

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n Q_i \right|_2 \ge \delta \right) \le 2\,\mathbf{Rank}\left(\sum_{i=1}^n \mathbb{V}\mathrm{ar}(Q_i) \right) \exp\left\{ -\frac{n\delta^2}{2(\sigma^2 + b\delta)} \right\}, \tag{24}$$

where $\sigma^2 := \frac{1}{n} \left|\sum_{i=1}^n \mathbb{V}\mathrm{ar}(Q_i)\right|_2$.

One advantage of pseudo self-concordance is that we can relate the Hessian at y to the Hessian at x in terms of the norm $\|y-x\|_2$.

Proposition 25 (Bach (2010), Proposition 1) *Assume that f is pseudo self-concordant with parameter R on $\mathcal{X}$. For any $y \in \mathcal{X}$, we have*

$$e^{-R\|y-x\|_2} \nabla^2 f(x) \preceq \nabla^2 f(y) \preceq e^{R\|y-x\|_2} \nabla^2 f(x).$$