Orthogonal Statistical Learning

Dylan J. Foster
MIT
dylanf@mit.edu

Vasilis Syrgkanis
Microsoft Research, New England
vasy@microsoft.com

Abstract

We provide excess risk guarantees for statistical learning in a setting where the population risk with respect to which we evaluate the target model depends on an unknown model that must be to be estimated from data (a "nuisance model"). We analyze a two-stage sample splitting meta-algorithm that takes as input two arbitrary estimation algorithms: one for the target model and one for the nuisance model. We show that if the population risk satisfies a condition called *Neyman orthogonality*, the impact of the nuisance estimation error on the excess risk bound achieved by the meta-algorithm is of second order. Our theorem is agnostic to the particular algorithms used for the target and nuisance and only makes an assumption on their individual performance. This enables the use of a plethora of existing results from statistical learning and machine learning literature to give new guarantees for learning with a nuisance component. Moreover, by focusing on excess risk rather than parameter estimation, we can give guarantees under weaker assumptions than in previous works and accommodate the case where the target parameter belongs to a complex nonparametric class. We characterize conditions on the metric entropy such that *oracle rates*—rates of the same order as if we knew the nuisance model—are achieved. We also analyze the rates achieved by specific estimation algorithms such as variance-penalized empirical risk minimization, neural network estimation and sparse high-dimensional linear model estimation. We highlight the applicability of our results in four settings of central importance in the literature: 1) heterogeneous treatment effect estimation, 2) offline policy optimization, 3) domain adaptation, and 4) learning with missing data.

Contents

1 Introduction

Predictive models based on modern machine learning methods are becoming increasingly widespread in policy making, with applications in health care, education, law enforcement, and business decision making. Most problems that arise in policy making, such as attempting to predict counterfactual outcomes for different interventions or optimizing policies over such interventions, are not pure prediction problems, but rather are causal in nature. It is important to address the causal aspect of these problems and build models that have a causal interpretation.

A common paradigm in the search of causality is that to estimate a model with a causal interpretation from observational data—for example, data not collected via randomized trial or via a known treatment policy—one typically needs to estimate many other quantities that are not of primary interest, but that can be used to de-bias a purely predictive machine learning model by formulating an appropriate loss. Examples of such *nuisance parameters* include the propensity for taking an action under the current policy, which can be used to form unbiased estimates for the reward of new policies, but is typically unknown in datasets that do not come from controlled experiments.

To make matters more concrete, let us walk through an example for which certain variants have been well-studied in machine learning (e.g., Dudík et al. (2011); Swaminathan and Joachims (2015a); Nie and Wager (2017); Kallus and Zhou (2018)). Suppose a decision maker wants to estimate the causal effect of some treatment $T \in \{0, 1\}$ on an outcome Y as a function of a set of observable features X; the causal effect will be denoted as $\theta(X)$. Typically, one has access to data consisting of tuples (X_i, T_i, Y_i), where X_i is the observed feature for sample i, T_i is the treatment taken, and Y_i is the observed outcome. Such settings are often referred to as having *bandit* feedback, since we only observe the outcome for the treatment that was chosen. Due to the bandit nature of the problem, one needs to create unbiased estimates of the unobserved outcome. A standard approach is to use the so-called *doubly-robust* formula, which is a combination of direct regression and inverse propensity scoring: if we let $Y_i^{(t)}$ denote the potential outcome from treatment t in sample i, and let $m_0^{(t)}(x_i) := \mathbb{E}\left[Y_i^{(t)} \mid x_i\right]$ and $p_0^{(t)} := \mathbb{E}[1\{T = t\}|x_i]$, then the following is an unbiased estimator for each potential outcome:

$$\hat{Y}_i^{(t)} = m_0^{(t)}(x_i) + \frac{(Y_i - m_0^{(t)}(x_i)) \, 1\{T_i = t\}}{p_0^{(t)}(x_i)} \tag{1}$$

Given such an estimator, we can estimate the treatment effect by running a regression between the unbiased estimates and the features, i.e. solve $\min_{\theta \in \Theta_n} \sum_i \left(\hat{Y}^{(1)} - \hat{Y}^{(0)} - \theta(X_i)\right)^2$ over some model class Θ_n. In the population limit, with infinite samples, this corresponds to finding a model $\theta(x)$ that minimizes the population risk $\mathbb{E}\left[(\hat{Y}_i^{(1)} - \hat{Y}_i^{(0)} - \theta(X))^2\right]$. Similarly, if one is interested in policy optimization rather than estimating treatment effects, they could use these unbiased estimates to solve $\min_{\theta \in \Theta_n} \sum_i (\hat{Y}_i^{(0)} - \hat{Y}_i^{(1)}) \cdot \theta(X_i)$ over a policy space Θ_n of functions mapping features to $\{0, 1\}$. When dealing with observational data, the functions m_0 and p_0 are not known, and must be estimated if we wish to evaluate the proxy labels $\hat{Y}^{(t)}$. The goal of the learner is to find a model θ that achieves good population risk when evaluated at the true nuisance functions as opposed to the estimated, since only then the model has a causal interpretation.

This phenomenon is ubiquitous in causal inference and motivates us to formulate the abstract problem of statistical learning with a nuisance component: Given n i.i.d. examples from a distribution $\mathcal{D}$, a learner is interested in finding a *target model* $\widehat{\theta}_n \in \Theta_n$ so as to minimize a population risk function $L_\mathcal{D} : \Theta_n \times \mathcal{G}_n \to \mathbb{R}$. The population risk depends not just on the target model, but also on a *nuisance model* whose true value $g_0 \in \mathcal{G}_n$ is unknown to the learner. The goal of the learner is to

produce an estimate that has small *excess risk* evaluated at the unknown true nuisance model:

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - \inf_{\theta \in \Theta_n} L_{\mathcal{D}}(\theta, g_0). \tag{2}$$

Depending on the application such an excess risk bound can take different interpretations. For many settings, such as treatment effect estimation, it is closely related to mean squared error, while in policy optimization problems it may correspond to regret. Following the tradition of statistical learning theory (Vapnik, 1995; Bousquet et al., 2004), we make excess risk the primary focus of our work, independent of the interpretation. We develop algorithms and analysis tools that generically address (2), then apply these tools to a number of applications of interest.

The problem of statistical learning with a nuisance component is strongly connected to the problem of semiparametric inference (Robinson, 1988; Kosorok, 2008), where a true parameter θ_0 is the minimizer of a population risk that depends on unknown nuisance components. Our paper builds on a growing body of results on "double" or "debiased" machine learning in statistics and econometrics literature (Chernozhukov et al., 2017, 2018a,c,b) for addressing semiparametric inference problems. This line of research has focused on providing so-called "$\sqrt{n}$-consistent and asymptotically normal" estimates when the target parameter θ_0 is low-dimensional and nuisance parameters belong to a nonparametric class. Unlike the semiparametric inference problem, statistical learning with a nuisance component does not require a well-specified model, nor a unique minimizer of the population risk. Moreover, we do not ask for parameter recovery and asymptotic inference (i.e. asymptotically valid confidence intervals). Rather, we are content with an excess risk bound, regardless of whether there is an underlying true parameter to be identified. As a consequence, we provide guarantees even when the target model belongs to a large, potentially nonparametric class.

The case where the target parameter belongs to an arbitrary class has not been addressed at the level of generality we consider in the present work, but we mention some prior work that goes beyond the low-dimensional/parametric setup for special cases. Athey and Wager (2017) and Zhou et al. (2018) give guarantees based on metric entropy of the target class for the specific problem of treatment policy learning. For estimation of treatment effects, various nonparametric classes have been used for the target class on a fairly cases by case basis, including kernels (Nie and Wager, 2017), random forests (Athey et al., 2016; Oprescu et al., 2018; Friedberg et al., 2018), and high-dimensional linear models (Chernozhukov et al., 2017, 2018b). Our work unifies several of these papers into a single framework and our general results have implications and improve upon each of these directions (see Section 8 for a detailed comparison).

Our approach is to reduce the problem of statistical learning with a nuisance component to the standard formulation of statistical learning. Rather than directly analyzing particular algorithms and models from machine learning (e.g., regularized regression, gradient boosting, or neural network estimation) we assume a black-box guarantee for the excess risk in the case where a nuisance value $g \in \mathcal{G}_n$ is fixed. Our main theorem asks only for the existence of an algorithm $\mathrm{Alg}(\Theta_n, S; g)$ that, for any given nuisance model g and data set S, achieves good excess risk with respect to the population risk $L_D(\theta, g)$, i.e. with probability $1 - \delta$:

$$L_{\mathcal{D}}(\widehat{\theta}_n, g) - \inf_{\theta \in \Theta_n} L_{\mathcal{D}}(\theta, g) \le \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S, \delta; g). \tag{3}$$

Likewise, we assume the existence of a black-box algorithm $\mathrm{Alg}(\mathcal{G}_n, S)$ to estimate the nuisance component g_0 from the data, with the required estimation guarantee varying from problem to problem.[1]

---

[1]Our approach is conceptually related to Künzel et al. (2017), who studied meta-algorithms based on off-the-shelf learning procedures for the specific problem of estimating the conditional average treatment effect, albeit our results are technically very different.

Meta-Algorithm 1 (Two-Stage Estimation with Sample Splitting).
Input: *Sample set* $S = z_1, \ldots, z_n$.
- *Split* S *into subsets* $S^{(1)} = z_1, \ldots, z_{\lfloor n/2 \rfloor}$ *and* $S^{(2)} = S \smallsetminus S^{(1)}$.
- *Let* $\widehat{g}_n$ *be the output of* $\mathrm{Alg}(\mathcal{G}_n, S^{(1)})$.
- *Return* $\widehat{\theta}_n$, *the output of* $\mathrm{Alg}(\Theta_n, S^{(2)} ; \widehat{g}_n)$.

Given access to the two black-box algorithms, we analyze a simple sample splitting-based meta-algorithm for statistical learning with a nuisance component, presented as Meta-Algorithm 1. We can now state the main question addressed in this paper: *When is the excess risk achieved by sample splitting robust to nuisance component estimation error?*

In more technical terms, we seek to understand when the two-stage sample splitting estimation algorithm achieves an excess risk bound with respect to g_0, in spite of error in the estimator $\widehat{g}_n$ output by the first-stage algorithm. Robustness to nuisance estimation error allows the learner to use more complex models for nuisance estimation and—under certain conditions on the complexity of the target and nuisance model classes—to learn target models whose error is, up to lower order terms, as good as if the learner had known the true nuisance model. Such a guarantee is typically referred to as achieving an *oracle rate* in semiparametric inference.

Overview of results. We show that *Neyman orthogonality*, which has been used to prove oracle rates for inference in semiparametric models (Neyman, 1959, 1979; Chernozhukov et al., 2018a,b), is key to providing oracle rates for statistical learning with a nuisance component. We prove that if the population risk satisfies a functional analogue of Neyman orthogonality, then the estimation error of $\widehat{g}_n$ has a second order impact on the overall excess risk (relative to g_0) achieved by $\widehat{\theta}_n$. To gain some intuition, Neyman orthogonality is weaker condition than double robustness, albeit similar in flavor, (see e.g. Chernozhukov et al. (2016)) and is satisfied by both the treatment effect loss and the policy learning loss described in the introduction. In more detail, our extension of the Neyman orthogonality condition asks that a cross-functional derivative of the loss vanish to zero, when evaluated at the optimal target and nuisance models. Prior work on the classical notion of Neyman orthogonality provides a number of means through which to construct orthogonal losses whenever certain moment conditions are satisfied by the data generating process (Chernozhukov et al., 2018a, 2016, 2018b). Indeed, orthogonal losses can be constructed in settings including treatment effect estimation, policy learning, missing data problems, estimation of structural econometric problems and game theoretic models.

We identify two regimes of excess risk behavior:

1. When the population risk is strongly convex with respect to the prediction of the target model (e.g. the treatment effect estimation loss), then typically so-called *fast rates* (e.g. rates of order of $O(1/n)$ for parametric classes) are achievable had we known the true nuisance model. Letting $R_{\mathcal{G}_n}$ denote the estimation error of the nuisance component (root-mean-squared prediction error for most of our settings), then in the fast rate setting we show that orthogonality implies that the first stage error has an impact on the excess risk of the order of $R_{\mathcal{G}_n}^4$ (e.g. $n^{-1/4}$ RMSE rates for the nuisance suffice when the target is parametric).

2. Absent any assumption on the convexity of the population risk (e.g. the treatment policy optimization loss), then typically *slow rates* (e.g. rates of order $O(1/\sqrt{n})$ for parametric classes) are achievable had we known the true nuisance model. In this case the impact is

of nuisance estimation error is of the order $R_{\mathcal{G}_n}^2$ so, once again, $n^{-1/4}$ RMSE rates for the nuisance suffice when the target is parametric.

To extend the sufficient conditions above to arbitrary classes, we give conditions on the relative complexity of the target and nuisance classes—quantified via *metric entropy*—under which the sample splitting meta-algorithm achieves oracle rates (assuming the two black-box estimation algorithms are appropriately instantiated). This allows us to extend several prior works beyond the parametric regime to complex nonparametric target classes. Our technical results extends the works of Yang and Barron (1999); Rakhlin et al. (2017), which provide minimax optimal rates without nuisance components and utilize the technique of *aggregation* in designing optimal algorithms.

The flexibility of our approach allows us to instantiate the framework with any machine learning model and algorithm of interest for both nuisance and target model estimation, and to utilize the vast literature on generalization bounds in machine learning to establish data-dependent and dimension-independent rates for several classes of interests. For instance, our approach allows us to use recent work on size-independent generalization error of neural networks. We obtain sharp guarantees for these specific model classes and more as a consequence of a new analysis for empirical risk minimization with plug-in estimation of nuisance parameters in the presence of orthogonality. Our results on plugin empirical risk minimization extend the local Rademacher complexity analysis of generalization bounds (Koltchinskii and Panchenko, 2000; Bartlett et al., 2005), to account for the impact of the nuisance error. In the slow rate regime we also give a new analysis of *variance-penalized* empirical risk minimization, which allows us to recover and extend several prior results in the literature on policy learning. Our result improves upon the variance-penalized risk minimization approach of Maurer and Pontil (2009) by replacing the dependence on the metric entropy at a fixed approximation level with the critical radius, which is related to the entropy integral.

As a consequence of focusing on excess risk, we obtain oracle rates under weaker assumptions on the data generating process than in previous works. Notably, we obtain guarantees even when the target model is misspecified and the target parameters are not identifiable. For instance, for sparse high-dimensional linear classes, we obtain optimal prediction rates with no restricted eigenvalue assumptions.

We highlight the applicability of our results to four settings of primary importance in the literature: 1) estimation of heterogeneous treatment effects from observational data, 2) offline policy optimization, 3) domain adaptation, 4) learning with missing data. For each of these applications, our general theorems allow for the use of arbitrary estimators for the nuisance and target model classes and provide robustness to the nuisance estimation error.

2 Framework: Statistical Learning with a Nuisance Component

We work in a learning setting in which data instances belong to an abstract set $\mathcal{Z}$. We receive a sample set $S_n := z_1, \ldots, z_n$ where each z_t is drawn i.i.d. from an unknown distribution $\mathcal{D} \in \Delta(\mathcal{Z})$.

Define variable subsets $\mathcal{X} \subseteq \mathcal{W} \subset \mathcal{Z}$.[2] We focus on learning models that come from a *target model class* $\Theta_n : \mathcal{X} \to \mathcal{V}^{(2)}$ and *nuisance model class* $\mathcal{G}_n : \mathcal{W} \to \mathcal{V}^{(1)}$. Here $\mathcal{V}^{(1)}$ and $\mathcal{V}^{(2)}$ are finite dimensional vector spaces of dimension $K^{(1)}$ and $K^{(2)}$ respectively, equipped with norms $\|\cdot\|_{\mathcal{V}^{(1)}}$ and $\|\cdot\|_{\mathcal{V}^{(2)}}$. We use the subscript n for Θ_n and $\mathcal{G}_n$ to emphasize that the complexity of the classes may grow with n. Given an example $z_t \in S$, we write $w_t \in \mathcal{W}$ and $x_t \in \mathcal{X}$ to denote the subsets for z that act as arguments to the nuisance and target parameters respectively. For example, we may write

---

[2]The restriction $\mathcal{X} \subseteq \mathcal{W}$ is not strictly necessary but reduces notation.

$g(w_t)$ for $g \in \mathcal{G}_n$ or $\theta(x_t)$ for $\theta \in \Theta_n$. We assume that the function spaces Θ_n and $\mathcal{G}_n$ are equipped with norms $\|\cdot\|_{\Theta_n}$ and $\|\cdot\|_{\mathcal{G}_n}$ respectively.[3]

We measure performance of the target predictor through the real-valued *population loss functional* $L_{\mathcal{D}}(\theta, g)$, which maps a target predictor θ and nuisance predictor g to a loss. The subscript $\mathcal{D}$ in $L_{\mathcal{D}}$ denotes that the functional depends on the underlying distribution $\mathcal{D}$. For all of our applications $L_{\mathcal{D}}$ has the following structure, in line the classical statistical learning setting: First define a pointwise loss function $\ell(\theta, g; z)$, then define $L_{\mathcal{D}}(\theta, g) \coloneqq \mathbb{E}_{z \sim \mathcal{D}} \ell(\theta, g; z)$. Our general framework does not explicitly assume this structure, however.

Let $g_0 \in \mathcal{G}_n$ be the unknown true value for the nuisance parameter. Given the samples S, and without knowledge of g_0, we aim to produce a *target predictor* $\widehat{\theta}_n$ that minimizes the *excess risk* evaluated at g_0

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - \inf_{\theta \in \Theta_n} L_{\mathcal{D}}(\theta, g_0). \tag{4}$$

As discussed in the introduction, we will always produce such a predictor via the sample splitting meta-algorithm (Algorithm 1), which makes uses of a nuisance predictor $\widehat{g}_n$.

When the infimum in the excess risk is obtained we will use $\theta_n^\star$ to denote the corresponding minimizer, in which case the excess risk can be written as

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0).$$

We will sometimes use the notation θ_0 to refer to a particular target parameter with respect to which the second stage satisfies a *first-order condition*, i.e. $D_\theta L_{\mathcal{D}}(\theta_0, g_0)[\theta - \theta_0] = 0 \ \forall \theta \in \Theta_n$. If $\theta_0 \in \Theta_n$ and the population risk is convex then we can take $\theta_n^\star = \theta_0$ without loss of generality, but we do not assume this, and in general we do not assume existence of a such a parameter θ_0.

Notation. $\langle \cdot, \cdot \rangle$ denotes the standard inner product. $\|\cdot\|_p$ will denote the ℓ_p norm over $\mathbb{R}^d$ and $\|\cdot\|_\sigma$ will denote the spectral norm over $\mathbb{R}^{d_1 \times d_2}$. We let $\mathbb{E}$ and $\mathbb{P}$ denote expectation and probability respectively under a distribution that will be clear from context. For a subset $\mathcal{X}$ of a vector space $\mathrm{conv}(\mathcal{X})$ will denote the convex hull. For an element $x \in \mathcal{X}$, we define the star hull via

$$\mathrm{star}(\mathcal{X}, x) = \left\{ t \cdot x + (1 - t) \cdot x' \mid x' \in \mathcal{X}, t \in [0, 1] \right\}. \tag{5}$$

We also adopt the shorthand $\mathrm{star}(\mathcal{X}) \coloneqq \mathrm{star}(\mathcal{X}, 0)$.

Definition 1 (Directional Derivative). *Let $\mathcal{V}$ be a vector space of functions. For a functional $F : \mathcal{V} \to \mathbb{R}$, we define the derivative operator*

$$D_g F(g)[\nu] = \frac{d}{dt} F(g + t\nu) \big|_{t=0},$$

for a pair of functions $g, \nu \in \mathcal{V}$. Likewise, we define

$$D_g^k F(g)[\nu_1, \ldots, \nu_k] = \frac{\partial^k}{\partial t_1 \ldots \partial t_k} F(g + t_1 \nu_1 + \ldots + t_k \nu_k) \big|_{t_1 = \cdots = t_k = 0}.$$

When considering a functional in two arguments, e.g. $F(\theta, g)$, we will write $D_g F(\theta, g)$ and $D_\theta F(\theta, g)$ to make the argument with respect to which the derivative is taken explicit.

---

[3]Both norms typically take the form $\|f\|_{L_p(\mathcal{V};\mathcal{D})} = \left(\mathbb{E}_{z \sim \mathcal{D}} \|f(z)\|_{\mathcal{V}} \right)^{1/p}$ for functions $f : \mathcal{Z} \to \mathcal{V}$, where $\mathcal{V}$ is a vector space equipped with some norm $\|\cdot\|_{\mathcal{V}}$.

3   Orthogonal Statistical Learning

In this section we present our main results on orthogonal statistical learning, which state that under certain conditions on the loss function, the error due to estimation of the nuisance component g_0 has higher-order impact on the prediction error of the target component. The results in this section, which form the basis for all subsequent results, are *algorithm-independent*, and only involve assumptions on properties of the population risk $L_{\mathcal{D}}$. To emphasize the high level of generality, the results in this section invoke the learning algorithms in Algorithm 1 only through "rate" functions $\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, \ldots)$, $\mathrm{Rate}_{\mathcal{D}}(\Theta_n, \ldots)$ which respectively bound the estimation error of the first stage and the excess risk of the second stage.

Definition 2 (Algorithms and Rates)**.** *The first and second stage algorithms and corresponding rate functions are defined as follows:*

a) Nuisance Algorithm and Rate. *The first stage learning algorithm* $\mathrm{Alg}(\mathcal{G}_n, S)$, *when given a sample set S from distribution $\mathcal{D}$, outputs a predictor $\widehat{g}_n$ for which*

$$\|\widehat{g}_n - g_0\|_{\mathcal{G}_n} \leq \mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S, \delta)$$

with probability at least $1 - \delta$.

b) Target Algorithm and Rate. *Let $\widehat{\Theta}_n$ be some set with $\theta_n^\star \in \widehat{\Theta}_n$. The second stage learning algorithm* $\mathrm{Alg}(\Theta_n, S\,;\, g)$, *when given sample set S from distribution $\mathcal{D}$ and any $g \in \mathcal{G}$ outputs a predictor $\widehat{\theta}_n \in \widehat{\Theta}_n$ for which*

$$L_{\mathcal{D}}(\widehat{\theta}, g) - L_{\mathcal{D}}(\theta_n^\star, g) \leq \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S, \delta\,;\, g)$$

with probability at least $1 - \delta$.

We define worst-case rates $\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, n, \delta) \coloneqq \sup_{S:|S|=n} \mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S, \delta)$ *and* $\mathrm{Rate}_{\mathcal{D}}(\Theta_n, n, \delta\,;\, g) \coloneqq \sup_{S:|S|=n} \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S, \delta\,;\, g)$.

Observe that if one naively applies the algorithm for the target class using the nuisance predictor $\widehat{g}_n$ as a plug-in estimate for g_0, the rate stated in Definition 2 will only yield an excess risk bound of the form

$$L_{\mathcal{D}}(\widehat{\theta}_n, \widehat{g}_n) - L_{\mathcal{D}}(\theta_n^\star, \widehat{g}_n) \leq \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S, \delta\,;\, \widehat{g}_n). \tag{6}$$

This clearly does not match the desired bound (4), which involves only the true nuisance value g_0 and not the plug-in estimate $\widehat{g}_n$. The bulk of our work is to show that orthogonality may be used to correct this mismatch.

Note that Definition 2 allows the target predictor $\widehat{\theta}_n$ belongs to a class $\widehat{\Theta}_n$ which in general has $\widehat{\Theta}_n \neq \Theta_n$. This extra level of generality serves two purposes. First, it allows for refined analysis in the case where $\widehat{\Theta}_n \subset \Theta_n$, which is encountered when using algorithms based on regularization that do not a-priori constraints on, e.g., the norm of the class of predictors. Second, it permits the use of *improper prediction*, i.e. $\widehat{\Theta}_n \supset \Theta_n$, which in some cases is required to obtain fast rates for misspecified models (Audibert, 2008; Foster et al., 2018).

Recall that for a sample set $S = z_1, \ldots, z_n$, the empirical loss is defined via $L_S(\theta, g) = \frac{1}{n} \sum_{t=1}^n \ell(\theta, g\,;\, z_t)$. Many classical results from statistical learning can be applied to the double machine learning setting by minimizing the empirical loss with plug-in estimates for g_0, and we can simply cite these results to provide examples of $\mathrm{Rate}_{\mathcal{D}}$ for the target class Θ_n. Note however that this structure is not assumed by Definition 2, and we indeed consider algorithms that do not have this form.

Fast rates and slow rates. The rates presented in this section fall into two distinct categories, which we distinguish by referring to them as either *fast rates* or *slow rates*. The meaning of the word "fast" or "slow" here is two-fold: First, for fast rates, our assumptions on the loss imply that when the target class Θ_n is not too large (e.g. a parametric or VC-subgraph class) prediction error rates of order $O(1/n)$ are possible in the absence of nuisance parameters. For our slow rate results, the best prediction error rate that can be achieved is $O(1/\sqrt{n})$, even for small classes. This distinction is consistent with the usage of the term fast rate in statistical learning (Bousquet et al., 2004; Bartlett et al., 2005; Srebro et al., 2010), and we will see concrete examples of such rates for specific classes in later sections (Section 5, Section 6).

The second meaning "fast" versus "slow" refers to the first stage: When estimation error for the nuisance is of order ε, the impact on the second stage in our fast rate results is of order ε^4, while for our slow rate results the impact is of order ε^2.

The fast rate regime above—particularly, the ε^4-type dependence on the nuisance error—will be the more familiar of the two for readers accustomed to semiparametric inference (Chernozhukov et al., 2018a). While fast rates might seem like a strict improvement over slow rates, these results require stronger assumptions on the loss. Our results in Section 6 and Section 7 show that which setting is more favorable will in general depend on the precise relationship between the complexity of the target class and the nuisance parameter class.

3.1 Fast Rates Under Strong Convexity

We first present general conditions under which the sample splitting meta-algorithm obtains so-called fast rates for prediction. To present the conditions, we fix a representative $\theta_n^\star \in \arg\min_{\theta \in \Theta_n} L_{\mathcal{D}}(\theta, g_0)$. In general the minimizer may not be unique—indeed, by focusing on prediction we can provide guarantees even though parameter recovery is clearly impossible in this case. Thus, it is important to keep in mind that the single fixed representative $\theta_n^\star$ is used throughout all the assumptions stated in this section.

Our first assumption is the starting point for this work, and asserts that the population loss is *orthogonal* in the sense that the certain pathwise derivatives vanish.

Assumption 1 (Orthogonal Loss). *The population risk $L_{\mathcal{D}}$ is orthogonal:*

$$D_g D_\theta L_{\mathcal{D}}(\theta_n^\star, g_0)[\theta - \theta_n^\star, g - g_0] = 0 \quad \forall \theta \in \widehat{\Theta}_n, \forall g \in \mathcal{G}_n. \tag{7}$$

In addition to orthogonality, our main theorem for fast rates requires three additional assumptions, all of which are ubiquitous in results on fast rates for prediction in statistical learning. We require a first-order optimality condition for the target class, and require that the population risk is both strongly convex with respect to the target class and smooth.

Assumption 2 (First Order Optimality). *The infimum of the population risk satisfies the first-order optimality condition:*

$$D_\theta L_{\mathcal{D}}(\theta_n^\star, g_0)[\theta - \theta_n^\star] \geq 0 \quad \forall \theta \in \text{star}(\widehat{\Theta}_n, \theta_n^\star). \tag{8}$$

Remark 1. *The first-order condition is typically satisfied for models that are* well-specified, *meaning that there is some variable in z that identifies the target predictor θ_0. More generally, it suffices to "almost" satisfy the first-order condition, i.e. to replace (8) by the condition*

$$D_\theta L_{\mathcal{D}}(\theta_n^\star, g_0)[\theta - \theta_n^\star] \geq -o(\text{Rate}_{\mathcal{D}}(\Theta_n, n, \delta; \widehat{g}_n)). \tag{9}$$

The first-order condition is also satisfied whenever $\widehat{\Theta}_n$ is star-shaped around $\theta_n^\star$, i.e. $\text{star}(\widehat{\Theta}_n, \theta_n^\star) \subseteq \widehat{\Theta}_n$.

9

Assumption 3 (Strong Convexity in Prediction)**.** *The population risk $L_{\mathcal{D}}$ is strongly convex with respect to the prediction:*

$$D_\theta^2 L_{\mathcal{D}}(\bar{\theta}, g)[\theta - \theta_n^\star, \theta - \theta_n^\star] \geq \lambda \|\theta - \theta_n^\star\|_{\Theta_n}^2 - \kappa \|g - g_0\|_{\mathcal{G}_n}^4 \quad \forall \theta \in \widehat{\Theta}_n, \ \forall g \in \mathcal{G}_n, \ \forall \bar{\theta} \in \mathrm{star}(\widehat{\Theta}_n, \theta_n^\star).$$

Assumption 4 (Higher-Order Smoothness)**.** *There exist constants β_1, β_2, and β_3, such that the following derivative bounds hold:*

a) Second-order smoothness with respect to target. *For all $\theta \in \widehat{\Theta}_n$ and all $\bar{\theta} \in \mathrm{star}(\widehat{\Theta}_n, \theta_n^\star)$:*

$$D_\theta^2 L_{\mathcal{D}}(\bar{\theta}, g_0)[\theta - \theta_n^\star, \theta - \theta_n^\star] \leq \beta_1 \cdot \|\theta - \theta_n^\star\|_{\Theta_n}^2.$$

b) Higher-order smoothness. *For all $\theta \in \mathrm{star}(\widehat{\Theta}_n, \theta_n^\star)$, $g \in \mathcal{G}_n$, and $\bar{g} \in \mathrm{star}(\mathcal{G}_n, g_0)$:*

$$\left| D_g^2 D_\theta L_{\mathcal{D}}(\theta_n^\star, \bar{g})[\theta - \theta_n^\star, g - g_0, g - g_0] \right| \leq \beta_2 \cdot \|\theta - \theta_n^\star\|_{\Theta_n} \cdot \|g - g_0\|_{\mathcal{G}_n}^2.$$

Remark 2. *All of the conditions of Assumption 3 and Assumption 4 are easily satisfied whenever the population loss is obtained by applying the square loss or any other strongly convex and smooth link to the prediction of the target class. Concrete examples are given in Section 4.*

Remark 3. *Assumption 3 and Assumption 4 can be weakened so that we require the inequalities to hold only when evaluated at the predictors $\widehat{\theta}_n$ and $\widehat{g}_n$ output by the algorithm, instead of requiring them to hold for all $\theta \in \widehat{\Theta}_n$ and $g \in \mathcal{G}_n$.*

We now state our main theorem for fast rates.

Theorem 1. *Suppose that there is some $\theta_n^\star \in \arg\min_{\theta \in \Theta_n} L_{\mathcal{D}}(\theta, g_0)$ such that Assumption 1, Assumption 2, Assumption 3 and Assumption 4, are satisfied. Then the sample splitting meta-algorithm Algorithm 1 produces a predictor $\widehat{\theta}_n$ that guarantees that with probability at least $1 - \delta$,*

$$\left\|\widehat{\theta}_n - \theta_n^\star\right\|_{\Theta_n}^2 \leq \frac{4}{\lambda} \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta/2; \widehat{g}_n) + \frac{1}{\lambda}\left(\frac{\beta_2^2}{\lambda} + 2\kappa\right) \cdot \left(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2)\right)^4, \tag{10}$$

and

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0) \leq \frac{2\beta_1}{\lambda} \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta/2; \widehat{g}_n) + \frac{\beta_1}{2\lambda}\left(\frac{\beta_2^2}{\lambda} + 2\kappa\right) \cdot \left(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2)\right)^4. \tag{11}$$

Theorem 1 shows that the for Algorithm 1, the impact of the unknown nuisance parameter on the prediction has favorable fourth-order growth: $\left(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2)\right)^4$. This means that if the desired oracle rate without nuisance parameters is of order $O(n^{-1})$, it suffices to take $\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2) = o(n^{-1/4})$.

There is one issue not addressed by Theorem 1: If the nuisance parameter g_0 were known, the rate for the target parameters would be $\mathrm{Rate}_{\mathcal{D}}(\Theta_n, \ldots; g_0)$, but the bound in (11) scales instead with $\mathrm{Rate}_{\mathcal{D}}(\Theta_n, \ldots; \widehat{g}_n)$. This is addressed in Section 5 and Section 6, where we show that for various algorithms the cost to relate these quantities grows only as $\left(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2)\right)^4$, and so can be absorbed into the second term in (10) or (11).

3.2 Beyond Strong Convexity: Slow Rates

The strong convexity assumption used by Theorem 1 in the previous section requires curvature only in the prediction space, not the parameter space. This is considerably weaker than what is assumed in prior works on double machine learning (e.g. Chernozhukov et al. (2018b)), and is a major advantage of analyzing prediction error rather than parameter recovery. Nonetheless, in some situations even assuming strong convexity on predictions may be unrealistic. A second advantage of studying prediction is that, while parameter recovery is not possible here, it is still possible to achieve low prediction error, albeit with slower rates than in the strongly convex case. In this section we give guarantees under which these (slower) oracle rates for prediction error can be obtained in the presence of nuisance variables through orthogonal learning. The key technical assumption we use to achieve such oracle rates is *universal orthogonality*, which informally states that the loss is not simply orthogonal around $\theta_n^\star$, but rather is orthogonal for all $\theta \in \Theta_n$.

Assumption 5 (Universal Orthogonality). *For all $\bar{\theta} \in \mathrm{star}(\widehat{\Theta}_n, \theta_n^\star) + \mathrm{star}(\widehat{\Theta}_n - \theta_n^\star, 0)$,*

$$D_\theta D_g L_\mathcal{D}(\bar{\theta}, g_0)[g - g_0, \theta - \theta_n^\star] = 0 \quad \forall g \in \mathcal{G}_n, \ \theta \in \Theta_n.$$

The universal orthogonality assumption is satisfied in examples we consider including treatment effect estimation (Section 8.1) and policy learning (Section 8.2), and appears to be used implicitly in previous work in these settings (Nie and Wager, 2017; Athey and Wager, 2017). Beyond orthogonality, we require a mild smoothness assumption for the nuisance class.

Assumption 6. *The derivatives $D_g^2 L_\mathcal{D}(\theta, g)$ and $D_\theta^2 D_g L_\mathcal{D}(\theta, g)$ are continuous. Furthermore, there exists a constant β such that for all $\theta \in \mathrm{star}(\widehat{\Theta}_n, \theta_n^\star)$ and $\bar{g} \in \mathrm{star}(\mathcal{G}_n, g_0)$,*

$$\left| D_g^2 L_\mathcal{D}(\theta, \bar{g})[g - g_0, g - g_0] \right| \leq \beta \cdot \|g - g_0\|_{\mathcal{G}_n}^2 \quad \forall g \in \mathcal{G}_n. \tag{12}$$

Our main theorem for slow rates is as follows.

Theorem 2. *Suppose that there is $\theta_n^\star \in \arg\min_{\theta \in \Theta_n} L_\mathcal{D}(\theta, g_0)$ such that Assumption 5 and Assumption 6 are satisfied. Then with probability at least $1 - \delta$, the target predictor $\widehat{\theta}_n$ produced by Algorithm 1 enjoys the excess risk bound*

$$L_\mathcal{D}(\widehat{\theta}_n, g_0) - L_\mathcal{D}(\theta_n^\star, g_0) \leq \mathrm{Rate}_\mathcal{D}(\Theta_n, S^{(2)}, \delta/2 \,; \widehat{g}_n) + \beta \cdot \left(\mathrm{Rate}_\mathcal{D}(\mathcal{G}_n, S^{(1)}, \delta/2) \right)^2.$$

4 Sufficient Conditions for Single Index Losses

Our setup and guarantees in the previous section were phrased at the full level of generality, with abstract assumptions on the structure of the population risk. In this section we provide conditions under which these assumptions follow from concrete structural assumptions on the risk. We give sufficient guarantees, from *first principles*, for large families of loss functions that apply immediately to the applications we consider in Section 8

4.1 Fast Rates

In this section we give a broad class of losses under which the conditions for fast rates in Subsection 3.1 are satisfied. The loss is defined as the expectation of a point-wise loss $\ell(\zeta, \gamma, z)$ acting on the predictions of the nuisance and target parameters. Specifically, we assume existence of functions Λ and Γ such that the loss has the following loss structure:

$$\ell(\theta(x), g(w), z) = \Phi(\langle \Lambda(g(w), v), \theta(x) \rangle, g(w), z), \qquad L_\mathcal{D}(\theta, g) = \mathbb{E}_{z \sim \mathcal{D}} \ell(\theta(x), g(w), z). \tag{13}$$

Here we recall from Section 2 that x, w are subsets of the data z, and let $v \subseteq z$ be an auxiliary subset of the data. We assume $\Phi(\zeta, \gamma, z)$ satisfies the condition

$$\frac{\partial}{\partial \zeta} \Phi(\zeta, \gamma, z) = \phi(\zeta) - \Gamma(\gamma, z), \tag{14}$$

where ϕ is a strictly increasing function with $T \geq \phi'(\zeta) \geq \tau$ for constants $T, \tau > 0$. A simple example is square loss regression, where $\Phi(\zeta, \gamma, z) = (\zeta - \Gamma(\gamma, z))^2$.

We now show that the following conditions are sufficient to obtain fast rates of the type in Section 3 for losses with the single index structure above.

Assumption 7 (Sufficient Fast Rate Conditions for Single Index Losses)**.** *The loss ℓ satisfies the following conditions:*

$$T \geq \phi'(t) \geq \tau.$$
$$\mathbb{E}\big[\nabla_\gamma \nabla_\zeta \ell(\theta_n^\star(x), g_0(w), z) \mid w\big] = 0. \qquad\qquad (\Rightarrow \text{Assumption 1})$$
$$\mathbb{E}\big[\nabla_\zeta \ell(\theta_n^\star(x), g_0(w), z) \cdot (\theta(x) - \theta_n^\star(x))\big] \geq 0, \qquad \forall \theta \in \text{star}(\widehat{\Theta}_n, \theta_n^\star). \qquad (\Rightarrow \text{Assumption 2})$$
$$\Lambda(\gamma, v) \text{is } L_\Lambda\text{-Lipschitz in } \gamma \text{ w.r.t } \|\cdot\|_2 \text{ a.s.} \qquad\qquad (\Rightarrow \text{Assumption 3})$$
$$\sup_{x \in \mathcal{X}, \theta \in \Theta_n} \|\theta(x)\|_2 \leq R_{\Theta_n}. \qquad\qquad (\Rightarrow \text{Assumption 3})$$
$$\frac{\mathbb{E}\langle \Lambda(g_0(w), v), \theta(x) - \theta_n^\star(x)\rangle^2}{\mathbb{E}\|\theta(x) - \theta_n^\star(x)\|_2^2} \geq \gamma, \qquad \forall \theta \in \widehat{\Theta}_n. \qquad (\Rightarrow \text{Assumption 3})$$
$$\big\|\mathbb{E}\big[\nabla_{\gamma\gamma}^2 \nabla_{\zeta_i} \ell(\theta_n^\star(x), g(w), z) \mid w\big]\big\|_\sigma \leq \mu, \qquad \forall i, \forall g \in \text{star}(\mathcal{G}_n, g_0). \qquad (\Rightarrow \text{Assumption 4 (b)})$$

As a concrete example, consider the logistic loss, where $\Phi(t, \gamma, z) = y \cdot \log(G(t)) + (1 - y) \cdot \log(1 - G(t))$, where the target class $y \in \{0, 1\}$ is a subset of the data z and $G(t) = 1/(1 + e^{-t})$ is the logistic function, so that $\frac{\partial}{\partial t} \Phi(t, \gamma, z) = G(t) - y$. Observe that the gradient of the loss with respect to the target index value can be written as

$$\nabla_\zeta \ell(\zeta, \gamma, z) = \big(\phi(\langle \Lambda(\gamma, v), \zeta\rangle) - \Gamma(\gamma, z)\big)\Lambda(\gamma, v).$$

Moreover, whenever the arguments to the loss are bounded, the Hessian can be bounded above and below via

$$T \cdot \Lambda(\gamma, v)\Lambda(\gamma, v)^\top \geq \nabla_{\zeta\zeta}^2 \ell(\zeta, \gamma, z) = \phi'(\langle \Lambda(\gamma, v), \zeta\rangle) \cdot \Lambda(\gamma, v)\Lambda(\gamma, v)^\top \geq \tau \cdot \Lambda(\gamma, v)\Lambda(\gamma, v)^\top, \tag{15}$$

and thus the ratio condition is implied by a minimum eigenvalue assumption on the conditional covariance matrix $\mathbb{E}[\Lambda(g_0(w), v)\Lambda(g_0(w), v)^\top \mid x]$.

We now show that these conditions are sufficient to satisfy the assumptions of Theorem 1, and thus guarantee higher order impact from the nuisance parameters.

Lemma 1. *If Assumption 7 holds, then Assumption 1, Assumption 2, Assumption 3 and Assumption 4 are satisfied with constants $\lambda = \frac{\tau}{4}$, $\kappa = \frac{4\tau L_\Lambda^4 R_{\Theta_n}^2}{\gamma}$, $\beta_1 = T$ and $\beta_2 = \frac{\mu\sqrt{K^{(2)}}}{\sqrt{\gamma}}$ and with respect to the norms $\|\theta\|_{\Theta_n} := \sqrt{\mathbb{E}_z\big[\langle \Lambda(g_0(w), v), \theta(x)\rangle^2\big]}$ and $\|\cdot\|_{\mathcal{G}_n} = \|\cdot\|_{L_4(\ell_2, \mathcal{D})}$.*

Combining this lemma with the guarantee from Theorem 1 directly yields the following corollary.

Corollary 1. *Suppose that there is some $\theta_n^\star \in \arg\min_{\theta \in \mathcal{G}_n} L_{\mathcal{D}}(\theta, g_0)$ such that [Assumption 7](#) is satisfied. The sample splitting meta-algorithm [Algorithm 1](#) produces a predictor $\widehat{\theta}_n$ that guarantees, with probability at least $1 - \delta$,*

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0) \le \frac{16}{\tau} \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta/2\,;\widehat{g}_n) + \frac{2\left(4\mu^2 K^{(2)} + 16\tau L_\Lambda^4 R_{\Theta_n}^2\right)}{\tau\,\gamma}\left(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2)\right)^4.$$

Furthermore, defining $\|\theta\|_{\Theta_n} = \sqrt{\mathbb{E}_z\big[\langle\Lambda(g_0(w), v), \theta(x)\rangle^2\big]}$, the following prediction error guarantee is satisfied with probability at least $1 - \delta$:

$$\left\|\widehat{\theta}_n - \theta_n^\star\right\|_{\Theta_n}^2 \le \frac{8T}{\tau} \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta/2\,;\widehat{g}_n) + \frac{2T\left(4\mu^2 K^{(2)} + 8\tau L_\Lambda^4 R_{\Theta_n}^2\right)}{\tau\,\gamma}\left(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2)\right)^4.$$

Observe that in both of the bounds in [Corollary 1](#), the only problem-dependent parameters that affect the leading term are T and τ. Importantly, this implies that if the more restrictive parameters $\gamma, L_\Lambda, \mu, R_\Theta, K^{(2)}$ and so forth are held constant as n grows, then they are negligible asymptotically so long as the nuisance parameter can be estimated quickly enough. For the square loss this is particularly desirable since $T = \tau = 1$.

Estimation for the first stage.   [Corollary 1](#) provides guarantees in terms of the L_4 estimation rate for the nuisance parameters, i.e.

$$\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta) = \|\widehat{g}_n - g_0\|_{L_4(\ell_2, \mathcal{D})} = \left(\mathbb{E}_w \|\widehat{g}_n(w) - g_0(w)\|_2^4\right)^{\frac{1}{4}}.$$

Since L_4 error rates are somewhat less common than L_2 (i.e., square loss) estimation rates, let us briefly discuss conditions under which out-of-the box algorithms can be used to give guarantees on the L_4 error.

First, for many nonparametric classes of interest, minimax L_p error rates have been characterized and can be applied directly. This includes smooth classes ([Stone, 1980, 1982](#)), Hölder classes ([Lepski, 1992](#); [Kerkyacharian et al., 2001, 2008](#)), Besov classes ([Delyon and Juditsky, 1996](#)), Sobolev classes ([Tsybakov et al., 1998](#)), and convex regression ([Guntuboyina and Sen, 2015](#))

Second, whenever the $\mathcal{G}_n$ is a linear class or more broadly a parametric class, classical statistical theory (e.g. ([Lehmann and Casella, 2006](#))) guarantees parameter recovery. Up to problem-dependent constants, this implies a bound on the L_4 error as soon as the fourth moment is bounded. This approach also extends to the high-dimesional setting ([Hastie et al., 2015](#)).

Last, if the class $\mathcal{G}_n$ has well-behaved moments in the sense that $\|g - g_0\|_{L_4(\ell_2, \mathcal{D})} \le C\|g - g_0\|_{L_2(\ell_2, \mathcal{D})}$ for all $g \in \mathcal{G}_n$, we can directly appeal to square loss regression algorithms for the first stage. This condition is related to the so-called "subgaussian class" assumption, and both have been explored in recent works ([Lecué and Mendelson, 2013](#); [Mendelson, 2014](#); [Liang et al., 2015](#)). We return to this condition in [Section 6](#).

4.2   Slow Rates

In the single index setup, assumptions much weaker than [Assumption 7](#) suffice to obtain slow rates via [Theorem 2](#). In particular, the following conditions are sufficient.

Assumption 8 (Sufficient Slow Rate Conditions for Single Index Losses)**.**

$$\mathbb{E}\big[\nabla_\zeta \nabla_\gamma \ell(\theta(x), g_0(w), z) \mid w\big] = 0. \qquad \forall \theta \in \mathrm{star}(\widehat{\Theta}_n, \theta_n^\star) + \mathrm{star}(\widehat{\Theta}_n - \theta_n^\star, 0) \qquad (\Rightarrow \text{[Assumption 5](#)})$$

$$\mathbb{E}\big[\nabla_{\gamma\gamma}^2 \ell(\theta(x), g(w), z) \mid w\big] \preceq \beta I. \qquad \forall \theta \in \mathrm{star}(\widehat{\Theta}_n, \theta_n^\star), g \in \mathrm{star}(\mathcal{G}_n, g_0) \qquad (\Rightarrow \text{[Assumption 6](#)})$$

Compared to Assumption 7, the most important difference is that since we require *universal* orthogonality, the first condition is required over all θ, not just at θ_0. Assumption 8 has the following immediate consequence.

Lemma 2. *If Assumption 8 holds, then Assumption 5 is satisfied and Assumption 6 is satisfied with constant β and with respect to the norm $\|\cdot\|_{\mathcal{G}_n} = \|\cdot\|_{L_2(\ell_2,\mathcal{D})}$.*

We have the following immediate consequence.

Corollary 2. *Suppose Assumption 8 holds. Then with probability at least $1 - \delta$, the target predictor $\widehat{\theta}_n$ produced by Algorithm 1 enjoys the excess risk bound*

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^{\star}, g_0) \leq \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta/2\,;\widehat{g}_n) + \beta \cdot \left(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2)\right)^2.$$

5 Empirical Risk Minimization with a Nuisance Component

In this section we develop algorithms and analysis for M-estimation losses, i.e. losses that take the form

$$L_{\mathcal{D}}(\theta, g) = \mathbb{E}[\ell(\theta(x), g(w), z)]. \tag{16}$$

We analyze the case where the algorithm used for the target parameter (the second stage algorithm), is one of the most natural and widely used algorithms: plug-in empirical risk minimization (ERM). Specifically, we define the empirical risk via

$$L_S(\theta, g) = \frac{1}{n}\sum_{t=1}^{n}\ell(\theta(x_t), g(w_t), z_t). \tag{17}$$

The plugin ERM algorithm returns the minimizer plug-in empirical loss obtained by plugging in the first-stage estimate of the nuisance component:

$$\widehat{\theta}_n = \arg\min_{\theta \in \Theta_n} L_S(\theta, \widehat{g}_n). \tag{18}$$

The goal of this section is to provide generalization error bounds for the plug-in ERM algorithm and variants. In particular, we will upper bound the second-stage rate $\mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta\,;\widehat{g}_n)$ as a function of standard complexity measures of the target class Θ_n. The goal of this section is to show that the impact of $\widehat{g}_n$ on the achievable rate by ERM is negligible and classical excess risk bounds carry over up to lower order terms and constant factors. One can easily combine our results on the rate $\mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta\,;\widehat{g}_n)$ from this section, with the main theorems from the previous section to obtain *oracle guarantees on the excess risk*, wherein the error due to nuisance estimation is of second order.

In the fast rate regime we offer a generalization of the local Rademacher complexity analysis of Bartlett et al. (2005) in the presence of an estimated nuisance component and show that notion of the *critical radius* of the class Θ_n still governs rate $\mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta\,;\widehat{g}_n)$ up to second order error. This result, coupled with our main theorem in the previous section, leads to several applications of our theory to particular target classes, including sparse linear models, neural networks and kernel classes; these are discussed at the end of the section.

In the slow rate regime (i.e., for generic Lipschitz losses), we show that the Rademacher complexity of the loss governs the rate, which subsequently can be upper bounded by the entropy integral of the function class. More importantly, we offer a novel moment-penalized variant of the ERM algorithm that achieves a rate whose leading term is equal to the critical radius, multiplied

by the *variance* of the population loss evaluated at the optimal target parameter. This offers an improvement over prior variance-penalized ERM approaches (Maurer and Pontil, 2009), whose leading term depends on the metric entropy of the target function class evaluated at single scale, and which typically is larger than the critical radius (the latter depending on a fixed point of the entropy integral).

Local Rademacher complexity Before moving to our main results of the section, we need to introduce standard concepts and notation from empirical processe theory and statistical learning theory. Throughout this section we will use the following shorthand for norms: for a vector valued function f, we will denote with

$$\|f\|_{p,q} = \left(\mathbb{E}\big[\|f(Z)\|_q^p\big]\right)^{1/p},\tag{19}$$

and similarly we will define an empirical analogue of the norm on n samples:

$$\|f\|_{p,q,n} = \left(\frac{1}{n}\sum_{t=1}^n \|f(z_t)\|_q^p\right)^{1/p}.\tag{20}$$

If f is real-valued we will omit q. For any $f^* \in \mathcal{F}$ let $\mathcal{F} - f^* = \{f - f^* : f \in \mathcal{F}\}$. We use the following overloaded version of the star hull notation:

$$\text{star}(\mathcal{F} - f^*) = \{r\,(f - f^*) : f \in \mathcal{F}, r \in [0,1]\}.\tag{21}$$

Moreover, let $\mathcal{F}_t = \{f_t : (f_1,\ldots,f_t,\ldots,f_d) \in \mathcal{F}\}$ denote the marginal real-valued function class that corresponds to coordinate t of the functions in class $\mathcal{F}$. Finally, for any real-valued function class $\mathcal{G}$, define the localized Rademacher complexity:

$$\mathcal{R}_n(\delta;\mathcal{G}) = \mathbb{E}_{\epsilon,z}\left[\sup_{g\in\mathcal{G}:\|g\|_2\leq\delta}\left|\frac{1}{n}\sum_{t=1}^n \epsilon_t\,g(z_t)\right|\right].\tag{22}$$

A related concept that we will also use throughout this section is that of metric entropy of a function class, which is closely related to the Rademacher complexity:

Definition 3 (Metric Entropy). *For any real-valued function class $\mathcal{G}$ and sample $z_{1:n}$, the metric entropy $\mathcal{H}_p(\epsilon,\mathcal{G},z_{1:n})$ is the logarithm of the size of the smallest function class $\mathcal{G}_\epsilon$, such that for any $g \in \mathcal{G}$ there exists $g_\epsilon \in \mathcal{G}_\epsilon$, with $\|g - g_\epsilon\|_{p,n} \leq \epsilon$. Moreover $\mathcal{H}_p(\epsilon,\mathcal{G},n)$ will denote the maximal empirical entropy over all possible samples $z_{1:n}$ of size n.*

5.1 Fast Rates via Local Rademacher Complexities

Our first contribution is an extension of the work of Bartlett et al. (2005); Koltchinskii and Panchenko (2000) on localized Rademacher complexity for bounding the excess risk. A crucial parameter in this approach is the *critical radius* δ_n of a function class $\mathcal{G}$, defined as the smallest solution to the inequality

$$\mathcal{R}_n(\delta_n;\mathcal{G}) \leq \delta_n^2.\tag{23}$$

Classical work of (see, e.g., the recent reference of Wainwright (2019)) shows that in the absence of a nuisance component, if $\ell(\theta(z),z)$ is a Lipschitz loss in its first argument and satisfies standard assumptions required for fast rates (e.g. it is strongly convex in its first argument), then empirical risk minimization achieves an excess risk bound of order δ_n^2. For the case of parametric classes, $\delta_n = \tilde{O}(n^{-1/2})$, leading to the fast $\tilde{O}(n^{-1})$ rates for strongly convex losses.

For more general classes (cf. Wainwright (2019)) the critical radius is up to constant factors equal to the solution to an inequality on the metric entropy of the function class (see Appendix E.2):

$$\int_{\frac{\delta_n^2}{8}}^{\delta_n} \sqrt{\frac{\mathcal{H}_2(\varepsilon, \mathcal{G}(\delta_n, z_{1:n}), z_{1:n})}{n}} d\varepsilon \leq \frac{\delta_n^2}{20}, \tag{24}$$

where $\mathcal{G}(\delta, z_{1:n}) = \left\{ g \in \mathcal{G} : \|g\|_{2,n} \leq \delta \right\}$. We exploit this characterization when applying the general results of this section to specific classes later on, including sparse linear models, neural networks and kernel function classes.

The theorem that follows extends this result in the presence of a nuisance component and bounds the rate of the plug-in ERM algorithm by the critical radius of the target function class Θ_n (in particular the worst-case critical radius of each coordinate of the target class, since we are dealing with vector valued function classes).

Theorem 3 (Fast Rates for Constrained ERM). *Consider a function class* $\Theta_n : \mathcal{X} \to \mathbb{R}^{K^{(2)}}$, *with* $\sup_{\theta \in \Theta_n} \|\theta\|_{\infty,2} \leq R$. *Let* $\delta_n^2 = \Omega\left(\frac{K^{(2)} \log(\log(n))}{n}\right)$ *be any solution to the inequality:*

$$\forall t \in \{1, \ldots, d\} : \mathcal{R}_n(\delta; \text{star}(\Theta_{n,t} - \theta_{n,t}^\star)) \leq \frac{\delta^2}{R}. \tag{25}$$

Suppose $\ell(\cdot, \widehat{g}_n(w), z)$ *is L-Lipschitz in its first argument with respect to the* ℓ_2 *norm and that the population risk* $L_\mathcal{D}$ *satisfies Assumption 1, Assumption 2, Assumption 3 and Assumption 4 with respect to the* $\|\cdot\|_{2,2}$ *norm. Let* $\widehat{\theta}_n$ *be the outcome of the constrained ERM algorithm. Then with probability* $1 - \delta$:

$$L_\mathcal{D}(\widehat{\theta}_n, \widehat{g}_n) - L_\mathcal{D}(\theta_n^\star, \widehat{g}_n) = O\left(\delta_n^2 + \|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4 + \frac{\log(1/\delta)}{n}\right). \tag{26}$$

Further, if $\ell(\cdot, \widehat{g}_n(w), z)$ *is also convex and twice differentiable in its first argument, with second order partial derivatives uniformly bounded by a constant and Assumption 4 is satisfied for every* $g \in \mathcal{G}_n$, *then the lower bound on* δ_n^2 *is not required and any non-negative solution to Equation 25 suffices.*

We emphasize that Theorem 3 provides an excess risk bound relative to the plugin estimate $\widehat{g}_n$, all that is required to obtain an excess risk bound at g_0 is to apply the results (Lemma 1) derived in the previous section.

5.1.1 Fast Rates for Specific Classes

We now instantiate the general ERM framework to give concrete guarantees for specific classes of interest. In all examples we use $\widetilde{O}$ to hide dependence on problem-dependent constants, $\log n$ factors, and $\log(\delta^{-1})$ factors.

Linear Classes. For our first set of examples, we focus on learning high-dimensional linear predictors. Chernozhukov et al. (2018b) gave orthogonal/debiased estimation guarantees for high-dimensional predictors using Lasso-type algorithms. Our first example shows how to recover the type of guarantee they gave, and our second example shows that we can give similar guarantees under weaker assumptions by exploiting that we work in the excess risk / statistical learning (rather than parameter estimation) framework.

Example 1 (High-Dimensional Linear Predictors with ℓ_1 Constraint). *Suppose that $\theta_n^\star$ is s-sparse with support set $T \subset [d]$ and that $\|\theta_n^\star\|_1 \leq 1$ and $\|x\|_\infty \leq 1$. Define the target class via*

$$\Theta_n = \left\{ x \mapsto \langle \theta, x \rangle \mid \theta \in \mathbb{R}^d, \ \|\theta\|_1 \leq \|\theta^\star\|_1 \right\}.$$

Given $S^{(2)} = x_{1:n}$, define the restricted eigenvalue for the target class as

$$\gamma_{\mathrm{re}} = \inf_{\Delta : \|\Delta_{T^c}\|_1 \leq \|\Delta_T\|_1} \frac{\frac{1}{n}\|X\Delta\|_2^2}{\|\Delta\|_2^2},$$

where $X \in \mathbb{R}^{n \times d}$ has $x_{1:n}$ as rows. Then under the assumptions of Theorem 3, empirical risk minimization guarantees that with probability at least $1 - \delta$,

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0) \leq \widetilde{O}\left(\frac{s \log d}{n \cdot \gamma_{\mathrm{re}}} + \left(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2) \right)^4 \right).$$

For parameter estimation it is well known that restricted eigenvalue or related conditions are required to ensure parameter consistency. For prediction however, such as assumptions are not needed if we are willing to consider inefficient algorithms. The next example shows that ERM over predictors with a hard sparsity constraint obtains the optimal high-dimensional rate for prediction in the presence of nuisance parameters with *no restricted eigenvalue assumption*.

Example 2 (High-Dimensional Linear Predictors with Hard Sparsity). *Suppose that Θ_n is a class of high-dimensional linear predictors obeying exact or "hard" sparsity:*

$$\Theta_n = \left\{ x \mapsto \langle \theta, x \rangle \mid \theta \in \mathbb{R}^d, \ \|\theta\|_0 \leq s, \ \|\theta\|_1 \leq 1 \right\},$$

and suppose $\|x\|_\infty \leq 1$. Then under the assumptions of Theorem 3, empirical risk minimization guarantees that with probability at least $1 - \delta$,

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0) \leq \widetilde{O}\left(\frac{s \log(d/s)}{n} + \left(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2) \right)^4 \right).$$

Neural Networks. We now move beyond the classical linear setting to the case to the case where the target parameters belong to a class of neural networks, a considerable more expressive class of models. Let $\sigma_{\log}(x) = (1 + e^{-x})^{-1}$ be the logistic link function and let $\sigma_{\mathrm{relu}}(x) = \max\{x, 0\}$ be the ReLU function.[4] Our first neural network example inspired by Farrell et al. (2018), who recently analyzed neural networks for nuisance parameter estimation. We depart from their approach by using neural networks to estimate *target parameters*.

Example 3. *Suppose that the target parameters are a class of neural networks $\Theta_n = \sigma_{\log} \circ \mathcal{F}$, where*

$$\mathcal{F} = \left\{ f(x) := A_L \cdot \sigma_{\mathrm{relu}}(A_{L-1} \cdot \sigma_{\mathrm{relu}}(A_{L-2} \cdot \sigma_{\mathrm{relu}}(A_1 x) \dots)) \mid A_i \in \mathbb{R}^{d_i \times d_{i-1}}, \|f\|_{L_\infty} \leq M \right\}, \quad (27)$$

and $d_0 = d$ and $d_L = 1$. Let $W = \sum_{i=1}^{L} d_i d_{i-1}$ denote the total number of weights in the network. Under the assumptions of Theorem 3, empirical risk minimization guarantees that with probability at least $1 - \delta$,

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0) \leq \widetilde{O}\left(\frac{WL \log W \log M}{n} + \left(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2) \right)^4 \right).$$

---

[4] For vector-valued inputs x we overload $\sigma_{\log}(x)$ and $\sigma_{\mathrm{relu}}(x)$ to denote element-wise application.

The target class in this example is well-suited to estimation of binary treatment effects. Note that in this example our only quantitative assumption on the network weights is that they guarantee boundedness of the output. However, the bound scales linearly with the number of parameters W, and thus may be vacuous for modern *overparameterized* neural networks. Our next example, which is based on neural networks covering bounds from Bartlett et al. (2017), shows that by making stronger assumptions on the weight matrices we can obtain weaker dependence on the number of parameters. This comes at the price of a slower rate—$n^{-\frac{1}{2}}$ vs. n^{-1}.

Example 4. *Suppose that the target parameters are a class of neural networks* $\Theta_n = \sigma_{\log} \circ \mathcal{F}$, *where*

$$\mathcal{F} = \left\{ f(x) \coloneqq A_L \cdot \sigma_{\text{relu}}(A_{L-1} \cdot \sigma_{\text{relu}}(A_{L-2} \cdot \sigma_{\text{relu}}(A_1 x) \dots)) \mid A_i \in \mathbb{R}^{d_i \times d_{i-1}}, \|A_i\|_\sigma \le s_i, \|A_i\|_{2,1} \le b_i \right\},$$

and $\|A\|_{2,1}$ *denotes the sum of row-wise* ℓ_2 *norms. Suppose* $\|x\|_2 \le 1$ *for all* $x \in \mathcal{X}$. *Under the assumptions of Theorem 3, empirical risk minimization guarantees that with probability at least* $1 - \delta$,

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0) \le \widetilde{O}\left(\frac{\prod_{i=1}^K s_i \cdot \left(\sum_{i=1}^L (b_i/s_i)^{2/3} \right)^{3/2}}{\sqrt{n}} + \left(\text{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2) \right)^4 \right).$$

Let us give a concrete example where the neural network guarantees above enable oracle rates for the target class while using a more flexible model class for the nuisance. Suppose the target parameters belong to the class in Example 3 with L_2 layers and W_2 weights, and suppose the nuisance parameters also belong to a neural network class, but with L_1 layers and W_1 weights. In the next section we will show that under certain assumptions one can guarantee $\left(\text{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2) \right)^4 = \widetilde{O}((W_1 L_1/n)^2)$ for such a class. In this case, Example 4 shows that we obtain oracle rates whenever $W_1 L_1 = o(\sqrt{W_2 L_2 n})$, meaning the number of parameters in the nuisance network can be significantly larger than for the target network. Similar guarantees can be derived for Example 3.

Deriving tight generalization bounds for neural networks is an active area of research and there are many more results that can be used as-is to give guarantees for the second stage in our general framework (Golowich et al., 2018; Allen-Zhu et al., 2018).

Kernels. We now give rates for some simple kernel classes. These examples were chosen only for concreteness, and the machinery in this section and the subsequent sections can be invoked to give guarantees for more rich and general nonparametric classes.

Example 5 (Gaussian Kernels). *Suppose that* $\Theta_n \subset ([0,1] \to \mathbb{R})$ *is unit ball in the reproducing kernel Hilbert space with the gaussian kernel* $\mathcal{K}(x, x') = e^{-\frac{1}{2}(x-x')^2}$. *Suppose* x *is drawn from the uniform distribution over* $[0, 1]$. *Under the assumptions of Theorem 3, empirical risk minimization guarantees that with probability at least* $1 - \delta$,

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0) \le \widetilde{O}\left(\frac{1}{n} + \left(\text{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2) \right)^4 \right).$$

Example 6 (Sobolev spaces). *Suppose the target class is the Sobolev space*

$$\Theta_n = \left\{ \theta : [0,1] \to \mathbb{R} \mid \theta(0) = 0, f \text{ is absolutely continuous with } \theta' \in L_2[0,1] \right\},$$

and suppose that x *is drawn from the uniform distribution on* $[0, 1]$. *Under the assumptions of Theorem 3, empirical risk minimization guarantees that with probability at least* $1 - \delta$,

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0) \le \widetilde{O}\left(\frac{1}{n^{2/3}} + \left(\text{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2) \right)^4 \right).$$

5.2 Slow Rates and Variance Penalization

We now turn to the slow rate regime, where the loss is not necessarily strongly convex in the prediction. In this setting we prove upper bounds on the generalization error of the plug-in ERM algorithm, with the main new results being slow rates that scale favorably with the variance of the loss rather than the range. Our results here are again based on local Rademacher complexities and metric entropy.

Our first contribution is a generalization bound that scales only with the worst case variance of the loss difference between any predictor $\theta \in \Theta_n$ and the optimal $\theta_n^\star$. To define our upper bound we first need to define the notion of the entropy integral of a function class:

Definition 4 (Entropy Integral). *For any real-valued function class $\mathcal{F}$ the entropy integral is defined as:*

$$\kappa(r, \mathcal{F}) = \inf_{\alpha \geq 0} \left[4\alpha + 10 \int_\alpha^r \sqrt{\frac{\mathcal{H}_2(\varepsilon/2, \mathcal{F}, n))}{n}} d\varepsilon \right], \tag{28}$$

The first theorem of this section bounds the generalization error of the plug-in ERM as a function of the normalized entropy integral of the class $\ell \circ \Theta_n$.

Theorem 4 (Slow Rates for Constrained ERM). *Consider the function class*

$$\ell \circ \Theta_n = \{\ell(\theta(\cdot), \widehat{g}_n(\cdot), \cdot) : \theta \in \Theta_n\},$$

with $\sup_{\theta \in \Theta_n} \|\ell(\theta(\cdot), \widehat{g}_n(\cdot), \cdot)\|_\infty \leq R$ *and* $r = \sup_{\theta \in \Theta_n} \|\ell(\theta(\cdot), \widehat{g}_n(\cdot), \cdot) - \ell(\theta_n^\star(\cdot), \widehat{g}_n(\cdot), \cdot)\|_2$. *Let* $\widehat{\theta}_n$ *be the outcome of the constrained ERM. Then with probability* $1 - \delta$:

$$L_{\mathcal{D}}(\widehat{\theta}_n, \widehat{g}_n) - L_{\mathcal{D}}(\theta_n^\star, \widehat{g}_n) = O\left(\mathcal{R}_n(r, \mathcal{F} - f^*) + r\sqrt{\frac{\log(1/\delta)}{n}} + R\frac{\log(1/\delta)}{n} \right) \tag{29}$$

$$= O\left(R \cdot \kappa(r/R, \ell \circ \Theta_n) + R\frac{\mathcal{H}_2(r, \ell \circ \Theta_n, n)}{n} + r\sqrt{\frac{\log(1/\delta)}{n}} + R\frac{\log(1/\delta)}{n} \right). \tag{30}$$

As a concrete example, consider the case when the class $\ell \circ \Theta_n$ is a VC-subgraph class of VC dimension d, and for simplicity assume $R = 1$. Then Theorem 2.6.7 of [Van Der Vaart and Wellner](#) (1996) shows that: $\mathcal{H}_2(\epsilon, \ell \circ \Theta_n, n) = O(d(1 + \log(1/\epsilon)))$. This implies that

$$\kappa(r, \ell \circ \Theta_n) = O\left(\int_0^r \sqrt{d(1 + \log(1/\epsilon))} d\epsilon \right) = O\left(r\sqrt{d}\sqrt{1 + \log(1/r)} \right).$$

Hence, we can conclude that:

$$L_{\mathcal{D}}(\widehat{\theta}_n, \widehat{g}_n) - L_{\mathcal{D}}(\theta_n^\star, \widehat{g}_n) = O\left(r\sqrt{1 + \log(1/r)}\sqrt{\frac{d}{n}} + \frac{d(1 + \log(1/r))}{n} + r\sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n} \right). \tag{31}$$

Variance-penalized ERM. The plug-in ERM result we just presented has two drawbacks we would like to address:

1. The generalization error depends on the worst-case variance of the loss difference r over all possible target parameters $\theta, \theta' \in \Theta_n$. This could be very large in settings like policy learning that involve importance weighting. In such a setting it is desirable to have a bound that depends only on the variance of the target policy $\theta_n^\star$.

19

2. The bound has $\log(1/r)$ factors that are potentially suboptimal

We now address the first point by providing a novel moment penalization approach that achieves a stronger guarantee. The second point is addressed for VC classes at the end of the section.

Our approach is an improvement over the variance penalization approach of Maurer and Pontil (2009) that penalizes its target parameter by its empirical variance. Existing results only show that the latter provides a generalization error whose leading term is of the form:

$$\sqrt{\frac{\mathrm{Var}_n(\ell(\theta_n^\star(\cdot), \widehat{g}_n(\cdot), \cdot))\mathcal{H}_\infty(1/n, \ell \circ \Theta_n, z_{1:n})}{n}},$$

where $\mathcal{H}_\infty$ is the metric entropy with respect to the norm $\|\cdot\|_{\infty,n}$. The drawback of such a bound is that it evaluates the metric entropy fixed approximation level, which can be suboptimal compared to the entropy integral that we derived in the previous theorem. Our new moment penalization method scales with the critical radius δ_n of the function class. The latter can be much smaller for many function classes; for instance, as we show in the next section, it is of the order of $\sqrt{d \log \tau n}$ for classes with VC dimension d and bounded Alexander capacity function τ.

Theorem 5 (Moment Penalized ERM). *Consider the function class $\mathcal{F} = \{\ell(\theta(\cdot), \widehat{g}_n(\cdot), \cdot) : \theta \in \Theta_n\}$, with $R = \sup_{f \in \mathcal{F}} \|f\|_\infty$ and $f^* = \ell(\theta_n^\star(\cdot), \widehat{g}_n(\cdot), \cdot)$. Let $\delta_n^2 \geq 0$ be any solution to the inequality*

$$\mathcal{R}_n(\delta; \mathrm{star}(\mathcal{F} - f^*)) \leq \frac{\delta^2}{R}. \tag{32}$$

Let $\widehat{\theta}_n$ be the outcome of second moment-penalized ERM:

$$\widehat{\theta}_n = \arg\min_{\theta \in \Theta_n} L_S(\theta, \widehat{g}_n) + 36\delta_n \|\ell(\theta(\cdot), \widehat{g}_n(\cdot), \cdot)\|_{2,n}. \tag{33}$$

Then with probability $1 - \delta$:

$$L_{\mathcal{D}}(\widehat{\theta}_n, \widehat{g}_n) - L_{\mathcal{D}}(\theta_n^\star, \widehat{g}_n) = O\left(\sqrt{\mathbb{E}[\ell(\theta_n^\star(\cdot), \widehat{g}_n(\cdot), \cdot)^2]}\left(\delta_n + \sqrt{\frac{\log(1/\delta)}{n}}\right) + R\left(\delta_n^2 + \frac{\log(1/\delta)}{n}\right)\right). \tag{34}$$

Note that the leading term in the new bound is multiplied by the value of the second moment of the loss *evaluated at the minimizer*. This is especially interesting in light of the fact that in the slow rate setting, under universal orthogonality, we can choose θ_n^* to be an arbitrary comparator from the target function class, i.e. it does not need to be the minimizer of the population loss. Hence, if we find comparators that have small second moments, then these comparators enjoy fast regret rates. This is achieved without any change to the algorithm, i.e. the algorithm does not need to know what θ_n^* is.

Moment vs Variance Penalization. One advantage the variance-penalized method of Maurer and Pontil (2009) enjoys relative to the new theorem is that its excess risk depends on the variance as opposed to the second moment, which is always larger. If the optimal loss $\mu^* := L_{\mathcal{D}}(\theta_n^\star, g_0)$ is zero then the two are equivalent. A general solution is as follows: if one has access to a good preliminary estimate $\hat{\mu}^*$ of the value μ^*, then using moment penalization one can always attain a bound that depends on $\|f^* - \hat{\mu}^*\|_2 = \sqrt{\mathrm{Var}(f^*)} + O(\|\hat{\mu}^* - \mu^*\|)$. The latter is achieved by simply re-defining the

function class $\mathcal{F}$ in [Theorem 5](#) to be the re-centered class of losses: $\{\ell(\theta(\cdot), \widehat{g}_n(\cdot), \cdot) - \hat{\mu}^* : \theta \in \Theta_n\}$. This leads to a slightly altered algorithm that penalizes a centered second moment:

$$\hat{\theta} = \arg \min_{\theta \in \Theta_n} L_S(\theta, \hat{g}) + 36\delta_n \|\ell(\theta(\cdot), \widehat{g}_n(\cdot), \cdot) - \hat{\mu}^*\|_{2,n}. \tag{35}$$

If the error in the preliminary estimate is vanishing, i.e. $\|\hat{\mu}^* - \mu^*\| = \epsilon_n \to 0$, then the impact of this error on the regret is only of second order, since the final regret bound will be of the form:

$$L_{\mathcal{D}}(\widehat{\theta}_n, \widehat{g}_n) - L_{\mathcal{D}}(\theta_n^\star, \widehat{g}_n) = O\left(\delta_n \sqrt{\mathrm{Var}(f^*)} + \delta_n \epsilon_n + \delta_n^2\right) = O\left(\delta_n \sqrt{\mathrm{Var}(f^*)}\right). \tag{36}$$

Such preliminary estimates can be easily attained an analysis based on the vanilla (non-localized) Rademacher complexity analysis and two-sided uniform convergence arguments over the function class $\mathcal{F}$.[5] For instance, if we have a uniform convergence rate of ϵ_n over Θ_n, i.e. $\mathbb{E}\left[\sup_{\theta \in \Theta_n} |L_S(\theta, \widehat{g}_n) - L_{\mathcal{D}}(\theta, \widehat{g}_n)|\right] \leq \epsilon_n$, then we can use $\hat{\mu}^* = \inf_{\theta \in \Theta_n} L_{S'}(\theta, \widehat{g}_n)$ on a separate sample S' of size n. Then by standard arguments we get that $\mathbb{E}[|\hat{\mu}^* - \inf_\theta L_{\mathcal{D}}(\theta, \widehat{g}_n)|] \leq \epsilon_n$. Moreover, observe that if $L_{\mathcal{D}}(\theta, g_0)$ is Lipschitz with respect to its second argument, then $|L_{\mathcal{D}}(\theta, \widehat{g}_n) - L_{\mathcal{D}}(\theta, g_0)| = O(\|g_0 - \widehat{g}_n\|)$ for any $\theta \in \Theta_n$. Hence, it must be that: $|L_{\mathcal{D}}(\theta_n^\star, \hat{g}) - \inf_\theta L_{\mathcal{D}}(\theta, \hat{g})| = O(\|g_0 - \widehat{g}_n\|)$. Thus, by consistency of the first stage estimate, the latter is also going to zero as n goes to infinity. Combining the two arguments we get that $\hat{\mu}^* \to \mu^*$ as desired. Finally, under such a Lipschitzness and consistency condition we also note that $\mathrm{Var}(\ell(\theta_n^\star(\cdot), \widehat{g}_n(\cdot), \cdot)) \to \mathrm{Var}(\ell(\theta_n^\star(\cdot), g_n^\star(\cdot), \cdot))$ and show the error from $\widehat{g}_n$ in the variance is also of lower order term as it is multiplied by δ_n. This reasoning yields the following corollary.

Corollary 3 (Centered Second Moment Penalization). *Consider the centered second moment penalized plugin empirical risk minimizer:*

$$\hat{\theta} = \arg \min_{f \in \mathcal{F}} L_S(\theta, \hat{g}) + 36\delta_n \|\ell(\theta(\cdot), \widehat{g}_n(\cdot), \cdot) - \hat{\mu}^*\|_{2,n}, \tag{37}$$

where $\hat{\mu}^* = \inf_{\theta \in \Theta_n} L_{S'}(\theta, \hat{g})$ on an auxiliary data set S' of size n. If $L_D(\theta, g)$ is Lipschitz in its second argument for all $\theta \in \Theta_n$, with respect to norm $\|\cdot\|_{\mathcal{G}_n}$ and $\widehat{g}_n$ is consistent with respect to that norm, then with probability $1 - \delta$:

$$L_{\mathcal{D}}(\widehat{\theta}_n, \widehat{g}_n) - L_{\mathcal{D}}(\theta_n^\star, \widehat{g}_n) = O\left(\sqrt{\mathrm{Var}(\ell(\theta_n^\star(\cdot), g_0(\cdot), \cdot))}\left(\delta_n + \sqrt{\frac{\log(1/\delta)}{n}}\right)\right). \tag{38}$$

5.2.1 Application to VC Classes

We now show how to instantiate the general slow rate guarantees developed in this section to give optimally efficient/variance-dependent rates for VC classes with general Lipschitz losses. Our main result shows that for VC classes the excess risk enjoyed by variance penalization grows exactly as $O(\sqrt{V^\star d/n})$ (where $V^\star$ is the variance of the loss at the pair $(\theta_n^\star, g_0)$) so long as the nuisance estimator converges at a rate of $o(n^{-1/4})$. The key to our approach is to assume boundedness of the so-called Alexander capacity function, a classical quantity that arises in the study of ratio type empirical processes ([Giné et al., 2006](#)).

To be more precise, for this example we assume that Θ_n is a class of binary predictors with VC dimension d, and let ℓ have the following policy learning structure:

$$\ell(\theta, g, z) = \Gamma(g, z) \cdot \theta(x), \qquad L_{\mathcal{D}}(\theta, g) = \mathbb{E}\, \ell(\theta, g, z),$$

---

[5]In fact we can even treat μ^* as part of the nuisance model, in which case it is easy to see that the re-centered loss satisfies universal orthogonality with respect to μ^*.

where Γ is some known function. Define the variance of the loss at $(\theta_n^\star, g_0)$ via

$$V^\star = \mathrm{Var}(\ell(\theta_n^\star(x), g_0(w), z)).$$

Our goal is to derive a bound for which the leading term only scales with $V^\star$ rather than the loss range. Our results depend on a variant of Alexander's capacity function. Letting $\Theta_n^\varepsilon = \{\theta \in \Theta_n : \mathbb{E}\,\Gamma^2(g_0, z)(\theta(x) - \theta_n^\star)^2 \le \varepsilon^2\}$, the capacity function is defined as

$$\tau^2(\varepsilon) = \frac{\mathbb{E}\sup_{\theta \in \Theta_n^\varepsilon} \Gamma^2(g_0, z)(\theta(x) - \theta_n^\star)^2}{\varepsilon^2}. \tag{39}$$

Note that when Γ is the usual unweighted classification loss this definition recovers the usual definition of the capacity function (Giné et al., 2006; Hanneke et al., 2014). Beyond boundedness of the capacity function, we make the following assumption

Assumption 9. *Assumption 8 holds with constant β. Furthermore, defining $\|g\|_{\mathcal{G}_n} = \|g\|_{L_4}$ for $g \in \mathcal{G}_n$, the following bounds hold:*
- *The loss function ℓ is bounded by R almost surely.*
- $\mathbb{E}\big[\Gamma^2(g_0, z) \mid x\big] \ge \gamma$ *almost surely.*
- $\big(\mathbb{E}(\Gamma(g, z) - \Gamma(g_0, z))^4\big)^{1/4} \le L\|g - g_0\|_{\mathcal{G}_n}$ *for all $g \in \mathcal{G}_n$.*
- *The first stage algorithm provides an estimation error bound with respect to $\|\cdot\|_{\mathcal{G}_n}$, i.e. $\|\widehat{g}_n - g_0\|_{\mathcal{G}_n} \le \mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S, \delta)$ with probability at least $1 - \delta$ over the draw of sample set S.*

Theorem 6. *Suppose that Assumption 9 holds, and define $\tau_0 := \sup_{\varepsilon \ge \sqrt{d/n}} \tau(\varepsilon)$. Then variance-penalized empirical risk minimization guarantees that with probability at least $1 - \delta$,*

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0)$$
$$\le \widetilde{O}\left(\sqrt{\frac{V^\star d \log \tau_0}{n}} + \frac{(R + L) d \log \tau_0}{n} + (\beta + (L + R)^2 \gamma^{-1/2})\big(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2)\big)^2\right).$$

Note that whenever $\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2) = o(n^{-1/4})$ the asymptotic rate depends only on the variance at $\theta_n^\star$ and g_0, not on the problem-dependent parameters $L/R/\beta\gamma$. Furthermore, whenever the capacity function is constant the asymptotic rate is exactly $O(\sqrt{V^\star d/n})$.

Variance-dependent bounds that obtain the optimal efficient $O(\sqrt{V^\star d/n})$ rate have been the subject of much recent investigation, and there is much interest in understanding when the $O(\sqrt{V^\star d \log n/n})$ rate obtained by naive approaches can be improved.

To give a brief survey, the seminal empirical variance bound due to Maurer and Pontil (2009) when applied directly gives a suboptimal to this setting gives a $O(\sqrt{V^\star d \log n/n})$ rate. An exciting recent work of Athey and Wager (2017) shows that for a specific loss and nuisance parameter setup arising in policy learning, the $\log n$ can be replaced with a certain worst-case variance parameter. Our result, Theorem 6 is complementary and shows that the $\log n$ can be replaced by the capacity function for *general losses*. We note that it appears unlikely that the $\log n$ can be removed with out at least some type of assumption; indeed the results in Rakhlin et al. (2017) imply that there are indeed VC classes for which the critical radius grows as $\sqrt{d \log n/n}$ in the worst case.

The proof of Theorem 6 can be broken into three parts: First, we apply the previous results of this section to show that the excess risk obtained by variance penalization depends on the critical radius of the class $\ell \circ \Theta_n$. Second, we show that in the absence of first-stage estimation error, the capacity function controls the critical radius. Finally, we show that the impact of nuisance estimation error on the capacity function is of second order.

6   Minimax Oracle Rates for Square Losses

We have now seen sharp orthogonal learning guarantees for a specific algorithm, ERM. We now shift our focus to understanding what rates can be achieved by *any algorithm*, and how this is determined by intrinsic properties of the target and nuisance classes.

In the absence of nuisance parameters, the minimax optimal rates for prediction with square losses (more generally, strongly convex losses) are determined by the global metric entropy of the target function class Θ_n. These minimax rates have been characterized both for the "well-specified" setting where targets are realized by some $\theta_0 \in \Theta_n$ plus zero-mean noise (Yang and Barron, 1999) and the misspecified setting where no assumptions are made on the target distribution (Rakhlin et al., 2017). Suppose we want to guarantee that this *oracle rate* (or, optimal rate without nuisance parameters) is obtained in the presence of nuisance parameters. How complex can the nuisance parameter class $\mathcal{G}_n$ be relative to the target parameter class Θ_n to ensure that the oracle rate is achieved?

In this section we answer this question, focusing on the single-index loss setup from Subsection 4.1. We characterize the relationship between the target metric entropy and nuisance metric entropy under which oracle rates are guaranteed. Given a particular growth rate for the metric entropy of the target class, we show what growth rate is required for the nuisance class to ensure the oracle rate is achieved. Our characterization extends the results of Yang and Barron (1999) and Rakhlin et al. (2017) to learning in the presence of nuisance parameters.

We focus on the Subsection 4.1 setup, where we recall the loss has the following structure:

$$\ell(\theta(x), g(w), z) = \left(\langle \Lambda(g(w), v), \theta(x) \rangle - \Gamma(g(w), z)\right)^2, \qquad L_{\mathcal{D}}(\theta, g) = \mathbb{E}_{z \sim \mathcal{D}} \, \ell(\theta, g \, ; z),$$

We distinguish between two settings: In the *well-specified* setting, we assume that there is a "ground truth" predictor $\theta_0 \in \Theta_n$ for which

$$\mathbb{E}[\Gamma(g_0(w), z) \mid w, v] = \langle \Lambda(g_0(w), v), \theta_0(x) \rangle. \tag{40}$$

In this case we may take $\theta_n^\star = \theta_0$ without loss of generality. On the other hand, in the *misspecified* setting we do not require the form of realizability in (42), but we require that $\widehat{\Theta}_n$ is convex. Both assumptions imply that the first-order condition Assumption 2 is satisfied, but interestingly the optimal oracle rate depends on whether the model is well-specified or misspecified. Consequently, the relationship between the complexity of the target class and nuisance class under which oracle rates are obtained also depends on whether the model is well-specified or misspecified.

In both cases we assume that all problem-dependent parameters are constant. Since our aim is to characterize the optimal dependence on the complexity of the classes Θ_n and $\mathcal{G}_n$, which is already quite technical, this simplifies the presentation considerably. There are no barriers to relaxing this assumption, however.

Assumption 10. *The loss satisfies Assumption 7 with constants $R_{\Theta_n}, L_\Lambda, \gamma, \mu, \tau = \Theta(1)$. The classes are bounded in the sense that for all $\theta \in \widehat{\Theta}_n$ and $g \in \mathcal{G}_n$ the following bounds hold almost surely: a) $\langle \Lambda(g(w), w), \theta(x) \rangle \in [-1, +1]$, b) $\Gamma(g(w), z) \in [-1, +1]$, c) $\Lambda(g(w), v)\Lambda(g(w), v)^\top \preceq I$, d) $\|g(w)\|_\infty \le 1$.*

Note that following Assumption 7, we work with the norms $\|\theta\|_{\Theta_n} = \sqrt{\mathbb{E}\langle \Lambda(g_0(w), v), \theta(x) \rangle^2}$ and $\|\cdot\|_{\mathcal{G}_n} = \|\cdot\|_{L_4(\ell_2, \mathcal{D})}$ throughout the section.

Our main workhorse for the results in this section is the "Aggregation of ε-Nets" or "Skeleton Aggregation" algorithm described in Yang and Barron (1999) and extended to random design in

Rakhlin et al. (2017).[6] In the square loss prediction setting with a well-specified model, Skeleton Aggregation obtains optimal rates for generic function classes in terms of metric entropy. We use Skeleton Aggregation as-is to provide rates for the first stage, and use an extension designed to work in the presence of nuisance parameters for the second stage.

6.1 Rates for First Stage

To invoke the generic framework of Section 3 to give concrete rates for specific settings of interest we must first provide bounds on the nuisance error rate $\text{Rate}_{\mathcal{D}}(\mathcal{G}_n, S, \delta)$. Here we give generic bounds on the first-stage rate based on the metric entropy of the class $\mathcal{G}_n$. We make two key assumptions: The first is that there is some subset of the variables z that identifies the true nuisance parameter g_0, and the second is that the moments of the nuisance class are well-behaved, so that we can appeal to square loss regression to provide rates; there are Assumption 11 and Assumption 12 below.

Assumption 11. *For each $i \in \left[K^{(1)}\right]$, there exists a variable $u_i \in z$ such that*

$$\mathbb{E}[u_i \mid w] = g_0(w)_i.$$

Moreover, each u_i lies in the range $[-1, +1]$.

Assumption 12 (Moment Comparison). *There is a constant $C_{2 \to 4}$ such that*

$$\frac{\|g - g_0\|_{L_4(\ell_2, \mathcal{D})}}{\|g - g_0\|_{L_2(\ell_2, \mathcal{D})}} \le C_{2 \to 4}, \qquad \forall g \in \mathcal{G}_n.$$

The assumption that g_0 is identified is standard in semiparametric literature (Chernozhukov et al., 2016, 2018c). The moment condition Assumption 12 is somewhat stronger so we spend a moment to discuss it.

The moment comparison condition has been used recently in statistics as a minimal assumption for learning without boundedness (Lecué and Mendelson, 2013; Mendelson, 2014; Liang et al., 2015). For example, suppose that each $g \in \mathcal{G}_n$ has the form $x \mapsto \langle w, x \rangle$ for $w, x \in \mathbb{R}^d$. Then $C_{2 \to 4} \le 3^{1/4}$ if x is mean-zero gaussian and $C_{2 \to 4} \le \sqrt{8}$ if x follows any distribution that is independent across all coordinates and symmetric (via the Khintchine inequality). Moment comparison is also implied by the "subgaussian class" assumption used in Mendelson et al. (2011); Lecué and Mendelson (2013).[7]

We emphasize that the moment constant $C_{2 \to 4}$ does not enter the leading term in any of our bounds—only the $\text{Rate}_{\mathcal{D}}(\mathcal{G}_n, \dots,)$ term in Theorem 1—and so it does not affect the asymptotic rates under conditions on metric entropy growth of $\mathcal{G}_n$ that we prescribe in the coming sections. We also note that, per the discussion in Section 4, this condition is not required at all for many classes of interest, where direct L_4 estimation rates are available. However, we adopt the condition here because it allows us to develop guarantees for arbitrary classes at the highest possible level of generality.

With these assumptions, we proceed by applying standard algorithms for square loss regression to each class $\mathcal{G}|_i \coloneqq \{w \mapsto g(w)_i \mid g \in \mathcal{G}_n\}$. We make the following assumption on the metric entropy growth rate for each of the nuisance classes.

Assumption 13. *One of the following conditions holds*

---

[6]We adopt the name Skeleton Aggregation from Rakhlin et al. (2017).

[7]Suppose $\mathcal{G}$ is scalar-valued and let $\|g\|_{\psi_2} = \inf\left\{c > 0 \mid \mathbb{E}\exp\left(g^2(w)/c^2\right) \le 2\right\}$. Then the subgaussian class assumption for our setting asserts that $\|g - g_0\|_{\psi_2} \le C\|g - g_0\|_{L_2(\mathcal{D})}$ for all $g \in \mathcal{G}_n$.

a) Parametric case. *There exists constants d_1 and D such that*

$$\mathcal{H}_2(\mathcal{G}|_i, \varepsilon, n) \le d_1 \log(D/\varepsilon) \quad \forall i.$$

b) Nonparametric case. *There exists constant p_1 such that.*

$$\mathcal{H}_2(\mathcal{G}|_i, \varepsilon, n) \le \varepsilon^{-p_1} \quad \forall i.$$

With this assumption, we can obtain rates by appealing to the following generic algorithms.
- *Global ERM*: For each i, select

$$(\widehat{g}_n)_i \in \underset{g \in \mathcal{G}_n|_i}{\arg\min} \sum_{t=1}^{n} ((u_i)_t - g(w_t))^2.$$

- *Skeleton Aggregation/Aggregation of ε-nets (Yang and Barron, 1999; Rakhlin et al., 2017).*
  For each i run the Skeleton Aggregation[8] algorithm with the class $\mathcal{G}_n|_i$ on the dataset of
  instance-target pairs $(w_1, (u_i)_1), \ldots, (w_n, (u_i)_n)$. Let $(\widehat{g}_n)_i$ be the result.

Proposition 1 (Rates for first stage, informal). *Suppose that Assumption 11 and Assumption 13
hold. Then Global ERM guarantees that with probability at least $1 - \delta$,*

$$\|\widehat{g}_n - g_0\|^2_{L_2(\ell_2, \mathcal{D})} \le \begin{cases} \widetilde{O}\big(K^{(1)} d_1 \log(en/d_1) \cdot n^{-1}\big), & \textit{Parametric case.} \\[3mm] \widetilde{O}\Big(K^{(1)} n^{-\frac{2}{2+p_1} \wedge \frac{1}{p_1}}\Big), & \textit{Nonparametric case.} \end{cases}$$

Skeleton Aggregation guarantees that with probability at least $1 - \delta$,

$$\|\widehat{g}_n - g_0\|^2_{L_2(\ell_2, \mathcal{D})} \le \begin{cases} \widetilde{O}\big(K^{(1)} d_1 \log(en/d_1) \cdot n^{-1}\big), & \textit{Parametric case.} \\[3mm] \widetilde{O}\Big(K^{(1)} n^{-\frac{2}{2+p_1}}\Big), & \textit{Nonparametric case.} \end{cases}$$

Here the $\widetilde{O}$ notation suppresses $\log n$, $\log(K^{(1)})$, and $\log(\delta^{-1})$ factors.

See Lemma 15 in the appendix for a precise version of Proposition 1 and detailed description of the
algorithms. Note that the minimax rate is $\Omega(n^{-\frac{2}{2+p_1}})$ (Yang and Barron, 1999), and so Skeleton
Aggregation is optimal for all values of p_1, while Global ERM is optimal only for $p_1 \le 2$. While these
are not the only algorithms in the literature for which we have generic guarantees based on metric
entropy (other choices include Star Aggregation (Liang et al., 2015) and Aggregation-of-Leaders
(Rakhlin et al., 2017)), they suffice for our goal in this section, which is to characterize the spectrum
of admissible rates.

In all applications we study, the dimension $K^{(1)}$ is constant. Nevertheless, studying procedures
that jointly learn all output dimensions of $\mathcal{G}_n$ and, in particular, deriving the correct statistical
complexity when $K^{(1)}$ is large is an interesting direction for future research and may be practically
useful.

---

[8]See Section F.2 for a formal description.

6.2 Minimax Oracle Rates

We are almost ready to state our main theorems for oracle rates. All that remains is to make an assumption on the complexity of the target class.

Assumption 14. *One of the following conditions holds*

a) Parametric case. *There exists constants d_2 and C such that*

$$\mathcal{H}_2(\Theta_n, \varepsilon, n) \le d_2 \log(C/\varepsilon).$$

b) Nonparametric case. *There exists constant p_2 such that.*

$$\mathcal{H}_2(\Theta_n, \varepsilon, n) \le \varepsilon^{-p_2}.$$

In the well-specified case, it is known that the minimax rate under Assumption 14 setting is $\Theta(n^{-\frac{2}{2+p_2}})$ (Yang and Barron, 1999). On the other hand, in the misspecified case the optimal rate is $\widetilde{\Theta}(n^{-\frac{2}{2+p_2} \wedge \frac{1}{p_2}})$, or in other words $\widetilde{\Theta}(n^{-\frac{2}{2+p_2}})$ when $p_2 \le 2$ and $\widetilde{\Theta}(n^{-\frac{2}{p_2}})$ when $p_2 > 2$ (Rakhlin et al., 2017). Our main theorems in this section show that under orthogonality, the optimal well-specified and misspecified rates can be achieved in the presence of nuisance parameters even when the nuisance class $\mathcal{G}_n$ is larger than the target class Θ_n, provided it is not *too* much larger. This generalizes the large body of results on double machine learning (Chernozhukov et al., 2018a; Mackey et al., 2017; Chernozhukov et al., 2018b), which show for various settings and assumptions that when the target class is parametric, one can obtain a $\sqrt{n}$-consistent estimator for the target as long as the nuisance estimator converges at a $n^{-\frac{1}{4}}$ rate.

We first state our main theorem for the well-specified case.

Theorem 7 (Oracle Rates, Well-Specified Case). *Suppose that we are in the well-specified setting and that Assumption 10,[9] Assumption 11, Assumption 12, Assumption 13, and Assumption 14 are satisfied for a class $\widehat{\Theta}_n$ defined below. Suppose that the following relationship holds*

$$p_1 < 2p_2 + 2. \tag{41}$$

Then for appropriate choice of sub-algorithms, the sample splitting meta-algorithm Algorithm 1 produces a predictor $\widehat{\theta}_n$ that guarantees that with probability at least $1 - \delta$,

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_0, g_0) \le \widetilde{O}\left(n^{-\frac{2}{2+p_2}}\right), \tag{42}$$

where the $\widetilde{O}$ symbol hides problem-dependent parameters and $\log(\delta^{-1})$ terms. When $p_2 \le 2$ it suffices to take $\widehat{\Theta}_n = \Theta_n$ and use global ERM for stage two, and when $p_2 > 2$ it suffices to take $\widehat{\Theta}_n = \Theta_n + \text{star}(\Theta_n - \Theta_n, 0)$ and use Skeleton Aggregation for stage two.

Theorem 7 is summarized in Figure 1. In particular, whenever Θ_n is a parametric class (i.e. $\mathcal{H}_2(\Theta_n, \varepsilon, n) \propto d_2 \log(1/\varepsilon)$), it suffices to take $p_1 < 2$, which recovers the usual setup for semiparametric inference.

We now focus on deriving oracle rates in the case where the second stage is *misspecified*. This setting has been relatively unexplored in double machine learning (Chernozhukov et al., 2018a; Mackey et al., 2017; Chernozhukov et al., 2018b); this is perhaps not surprising since for many

---

[9]When $p_1 \ge 2$, we technically require that Assumption 7 be satisfied for all $g \in \mathcal{G}_n + \text{star}(\mathcal{G}_n - \mathcal{G}_n, 0)$.
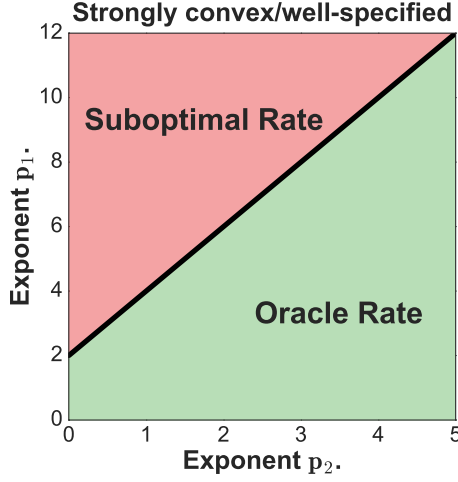
Figure 1: Relationship between first and second stage for oracle rates, well-specified case.

setups the well-specified property is critical to establish orthogonality. Interestingly, for certain settings including treatment effect estimation (Section 8.1), orthogonality can indeed hold even without model correctness. Our main theorem for the misspecified case shows that we can obtain oracle rates under both misspecification and nuisance parameters as long as the nuisance class has moderate complexity.

Theorem 8 (Oracle Rates, Misspecified Case). *Suppose that the target class Θ_n is convex. Suppose that Assumption 10,[10] Assumption 11, Assumption 12, Assumption 13, and Assumption 14. If the relationship*

$$p_1 < \max\{2p_2 + 2, 4p_2 - 2\} \tag{43}$$

holds, then for appropriate choice of sub-algorithms, the sample splitting meta-algorithm Algorithm 1 produces a predictor $\widehat{\theta}_n$ such that with probability at least $1 - \delta$,

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_0, g_0) \leq \widetilde{O}\left(n^{-\frac{2}{2+p_2} \wedge \frac{1}{p_2}}\right). \tag{44}$$

In particular, it suffices to use global ERM for the second stage.

Theorem 8 is summarized in Figure 2. Comparing to the well-specified case (Theorem 7/Figure 1), we see that in the misspecified case the condition on the nuisance parameter metric entropy is significantly more permissive when the target parameter class Θ_n is large (i.e. $p_2 > 2$). For example, if $p_2 = 5$ then we require $p_1 < 12$ for oracle rates in the well-specified case, but only require $p_1 < 18$ in the misspecified case. On the other hand, when $p_2 \leq 2$ the conditions on the nuisance metric entropy match the well-specified case. In particular, whenever Θ_n is a parametric class it again suffices to take $p_1 < 2$ (Chernozhukov et al., 2018a).

6.3 Sketch of Proof Techniques

The idea behind the second-stage rates we provide here is that the problem of obtaining a target predictor for the second stage can be solved by reducing to the classical square loss regression
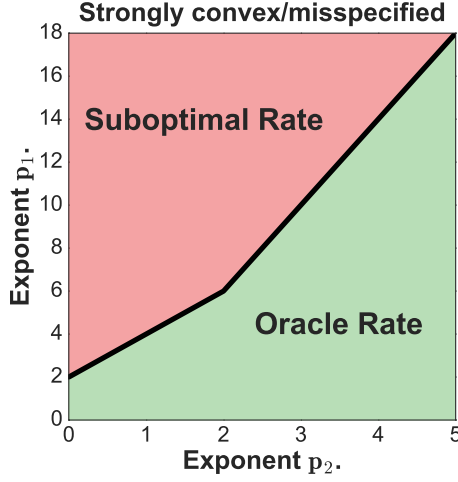
---

[10]See previous footnote.

Figure 2: Relationship between first and second stage for oracle rates, misspecified case.

setting. We map our setting onto square loss regression by defining auxiliary variables $X = w$ and $Y = \Gamma(g_0(w), w)$, and by defining auxiliary predictor classes

$$\mathcal{F}_0 = \{X \mapsto \langle \Lambda(g_0(w), w), \theta(x) \rangle \mid \theta \in \Theta_n\},$$
$$\mathcal{F} = \{X \mapsto \langle \Lambda(\widehat{g}_n(w), w), \theta(x) \rangle \mid \theta \in \Theta_n\}.$$

With these definitions, our goal to bound the excess risk in, e.g., Theorem 7 can be equivalently stated as producing a predictor $\widehat{f} \in \mathcal{F}_0$ that enjoys a bound on

$$\mathbb{E}\big(\widehat{f}(X) - Y\big)^2 - \inf_{f \in \mathcal{F}_0} \mathbb{E}(f(X) - Y)^2, \tag{45}$$

which is the standard notion of square loss excess risk used in, e.g., Liang et al. (2015); Rakhlin et al. (2017). Defining, $\widetilde{Y} = \Gamma(\widehat{g}_n(w), w)$, we can apply any standard algorithm for the class $\mathcal{F}$ to the dataset $(X_1, \widetilde{Y}_1), \ldots, (X_n, \widetilde{Y}_n)$. Note however that, due to the use of $\widehat{g}_n$ as a plug-in estimate, predictors produced via Algorithm 1 will—invoking Definition 2—give a guarantee of the form

$$\mathbb{E}\big(\widehat{f}(X) - \widetilde{Y}\big)^2 - \inf_{f \in \mathcal{F}} \mathbb{E}\big(f(X) - \widetilde{Y}\big)^2 \le \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta/2 \,; \widehat{g}_n), \tag{46}$$

where $\widehat{f}$ and the benchmark f belong to $\mathcal{F}$ instead of $\mathcal{F}_0$. The machinery developed in Section 3 relates the left-hand-side of this expression to the desired excess risk (45). Depending on the setting, more work is required to show that the right-hand-side of (46) is controlled. This challenge is only present in the well-specified setting. The difficulty is that while the original problem (45) is well-specified in this case, the presence of the plug-in estimator $\widehat{g}_n$ in (46) introduces additional "misspecification". We show for global ERM and Skeleton Aggregation the right-hand-side of the expression is controlled as well, meaning that $\mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta/2 \,; \widehat{g}_n)$ is not much larger than the rate $\mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta/2 \,; g_0)$ that would have been achieved if the true value for the nuisance predictor was plugged in. This achieved by exploiting orthogonality yet again.

In the misspecified setting, we can simply upper bound the right-hand side of (46) by the worst-case bound $\sup_{g \in \mathcal{G}_n} \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta/2 \,; g)$ and get the desired growth. Since the model is misspecified to begin with, any extra misspecification introduced by using the plugin estimate here is irrelevant. To be precise, the algorithm configuration is as follows.

28

- For stage one, use Skeleton Aggregation (Yang and Barron, 1999; Rakhlin et al., 2017). If $p_1 \leq 2$, global ERM can be used instead.
- For stage two in the misspecified setting with Θ_n convex, use global ERM.
- For stage two in the well-specified setting we use Skeleton Aggregation, with a new analysis to account for the small amount of "model misspecification" introduced by the plug-in nuisance estimate $\widehat{g}_n$. If $p_1 \leq 2$, global ERM can be used instead; this is because skeleton ERM and global ERM are both optimal for $p_1 \leq 2$, even in the presence of nuisance parameters.

Precise descriptions and analyses for these algorithms can be found in Appendix F and Appendix E.

7 Minimax Oracle Rates for Generic Lipschitz Losses

The fast oracle rates developed in the previous section rely on strong convexity, which is not satisfied for all losses used in practice. For example, linear losses used in policy learning do not satisfy this property (Athey and Wager, 2017). In this section we extend the oracle rate characterization from Section 6.2 from strongly convex losses to any loss that is Lipschitz in the target prediction $\theta(x)$. For arbitrary Lipschitz losses (in particular, for the linear loss), the optimal rate in the absence of nuisance parameters is $\widetilde{O}(n^{-\frac{1}{2} \wedge \frac{1}{p_2}})$ under the metric entropy growth assumed in Assumption 14.[11] Our main theorem for this section shows that this rate is still obtained in the presence of nuisance parameters when the nuisance metric entropy parameter p_1 is not too much larger than p_2. We make the following regularity assumption on the loss.

Assumption 15. *Assumption 8 holds with constant $\beta = O(1)$.[12] The loss ℓ has absolute value bounded by 1 and the mapping $\zeta \mapsto \ell(\zeta, \gamma, z)$ is 1-Lipschitz with respect to ℓ_2.*

Compared to the oracle rates for square losses, the assumptions required for our main Lipschitz loss theorem are relaxed significantly: We no longer require strong convexity, nor do we require any type of moment comparison for the nuisance class. On the other hand, we do require the additional universal orthogonality condition from Subsection 3.2.

Theorem 9. *Suppose that Assumption 11, Assumption 13, Assumption 14, and Assumption 15 are satisfied. If the relationship*

$$p_1 < \max\{2, 2p_2 - 2\} \tag{47}$$

holds, then for appropriate choice of sub-algorithms, the sample splitting meta-algorithm produces a predictor $\widehat{\theta}_n$ that guarantees that, with probability at least $1 - \delta$,

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_0, g_0) \leq \widetilde{O}\left(n^{-\frac{1}{2} \wedge \frac{1}{p_2}}\right). \tag{48}$$

For stage one it suffices to use global ERM for $p_1 \leq 2$ and Skeleton Aggregation for any value of p_1. For stage two it suffices to use global ERM.

Theorem 9 is summarized in Figure 3. In particular, whenever Θ_n is a parametric class (i.e. $\mathcal{H}_2(\Theta_n, \varepsilon, n) \propto d_2 \log(1/\varepsilon)$), it suffices to take $p_1 < 2$, as in the well-specified and misspecified square loss setups in Section 6.2, and as in standard semiparametric results (Chernozhukov et al., 2018a). That the condition matches is somewhat interesting given that the final rate in this case is slower than in the square loss setup. Note however that we cannot tolerate $p_1 > 2$ until $p_2 > 2$, as compared

---

[11]See, e.g., section 12.8 and section 12.9 in Rakhlin and Sridharan (2012).

[12]In line with the oracle results for fast rates, when $p_1 \geq 2$, we require that Assumption 8 is satisfied for all $g \in \mathcal{G}_n + \text{star}(\mathcal{G}_n - \mathcal{G}_n, 0)$.
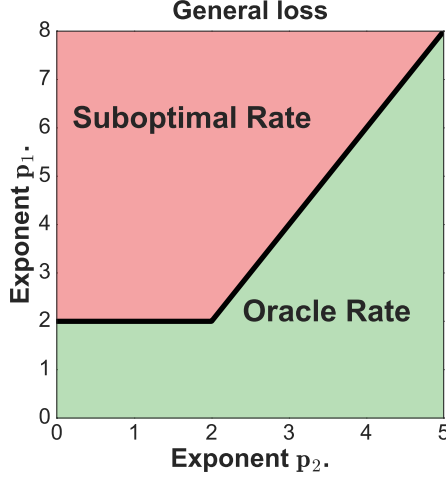
Figure 3: Relationship between first and second stage for oracle rates, general loss case.

to our results strongly convex losses, where the admissible value of p_1 is growing for all values of p_2. This result generalizes the characterization given in Athey and Wager (2017) for the specific case of policy learning, which applies only on the parametric setting and to a specific loss.

8 Applications

We now apply the techniques developed so far to four families of applications: Heterogeneous treatment effect estimation, offline policy learning/optimization, domain adaptation and sample bias correction, and learning with missing data. We how each application falls into our general orthogonal statistical learning framework, sketch some statistical consequences using the algorithmic tools developed, and show how these results generalize and extend previous work. Certain proofs from this section are deferred to Appendix G.

8.1 Treatment Effect Estimation

We first consider the problem of estimating conditional average treatment effects. Following, e.g., Robinson (1988); Nie and Wager (2017); Chernozhukov et al. (2018a), we receive examples $z = (X, W, Y, T)$ according to the following data generating process:

$$
\begin{aligned}
Y &= T \cdot \theta_0(X) + f_0(W) + \varepsilon_1, & \mathbb{E}[\varepsilon_1 \mid X, W, T] &= 0, \\
T &= e_0(W) + \varepsilon_2, & \mathbb{E}[\varepsilon_2 \mid X, W] &= 0.
\end{aligned}
\tag{49}
$$

Here $X \in \mathcal{X}$ and $W \in \mathcal{W}$, where $\mathcal{X}$ and $\mathcal{W}$ are abstract sets. $T \in \{0, 1\}$ is the treatment variable and $X \in \mathbb{R}$ is the target variable. The "ground truth" target predictor is $\theta_0 : \mathcal{X} \to \mathbb{R}$, though we do not assume $\theta_0 \in \Theta_n$. The functions $e_0 : \mathcal{W} \to [0, 1]$ and $f_0 : \mathcal{W} \to \mathbb{R}$ are unknown. The nuisance predictor class is defined based on these unknown functions as follows. We define $m_0(x, w) = \mathbb{E}[Y \mid X = x, W = w] = \theta_0(x)e_0(w) + f_0(w)$ and take $g_0 = \{m_0, e_0\}$ to be the ground truth nuisance parameter. We set $w = (X, W, T)$ and $x = (X)$, and use the loss

$$
\ell(\theta, \{m, e\}, z) = ((Y - m(X, W) - (T - e(W))\theta(X))^2.
\tag{50}
$$

30

This clearly falls into the framework of Subsection 4.1 using $\Lambda(g(w), w) = (T - e(W))$, $\Gamma(g(w), z) = (Y - m(X, W))$, and $\Phi(t, \gamma, z) = (t - \Gamma(\gamma, z))^2$. Verifying that the basic orthogonality and first-order conditions from Assumption 7 are satisfied is a simple exercise:

- The conditional expectation assumptions in Equation 49 imply that the loss has the first-order condition whenever $\theta_0 \in \Theta_n$:

$$D_\theta L_{\mathcal{D}}(\theta_0, g_0)[\theta - \theta_0] = 0 \quad \forall \theta.$$

On the other hand, even if $\theta_0 \notin \Theta_n$, the first-order condition is still satisfied as long as Θ_n is convex.

- The loss is universally orthogonal, meaning that its partial derivatives vanish not just around θ_0 but around any $\theta : \mathcal{X} \to \mathbb{R}$:

$$D_e D_\theta L_{\mathcal{D}}(\theta, \{m_0, e_0\})[\theta' - \theta, e - e_0] = 0 \quad \forall \theta, \theta', \forall e$$

and

$$D_m D_\theta L_{\mathcal{D}}(\theta, \{m_0, e_0\})[\theta' - \theta, m - m_0] = 0 \quad \forall \theta, \theta', \forall m$$

This means that the orthogonality condition (7) in Assumption 1 is satisfied for any $\theta_n^\star$, with no assumption on whether or not $\theta_0 \in \Theta_n$.

Interpreting excess risk. Let us take a moment to interpret the meaning of low prediction error in this setting. It is simple to verify that when the true nuisance parameters $g_0 = \{m_0, e_0\}$ are plugged in, and if the model is well-specified in the sense that $\theta_0 \in \Theta_n$,

$$L_{\mathcal{D}}(\theta, g_0) - L_{\mathcal{D}}(\theta_0, g_0) = \mathbb{E}((T - e_0(W)) \cdot (\theta(X) - \theta_0(X)))^2.$$

Thus, if a predictor θ has low risk, then θ must be good at predicting θ_0, so long as there is sufficient variation in the treatment T. If the model is not well-specified but Θ_n is convex, we can deduce from the excess risk that

$$L_{\mathcal{D}}(\theta, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0) \geq \mathbb{E}((T - e_0(W)) \cdot (\theta(X) - \theta_n^\star(X)))^2,$$

and so low excess risk means we predict nearly as well as the best predictor in class whenever there is sufficient variation in T.

Remarkably, our general results imply that for any class it is possible to achieve oracle rates for prediction with this loss in the presence of nuisance parameters, even when the model Θ_n is completely misspecified: If Θ_n is convex, then thanks to the universal orthogonality property, oracle rates are achievable so long as $\text{Rate}_{\mathcal{D}}(\mathcal{G}_n, n, \delta) = o\left(\text{Rate}_{\mathcal{D}}(\Theta_n, n, \delta)^{1/4}\right)$.

Fast rates, slow rates, and strong convexity. Since the treatment effects setup enjoys universal orthogonality and has a single index structure, we can appeal to the sufficient conditions in Section 4 to provide both fast and slow rates. Whether the fast or slow rate is better given finite samples depends on the assumptions on the data distribution.

Let's first consider the fast rate case, wherein we appeal to Theorem 1 via Assumption 7. For the square loss, assuming data is bounded or subgaussian, arguably the most restrictive assumption in Assumption 7 is the strong convexity-type assumption $\inf_{\theta \in \widehat{\Theta}_n} \left\{ \frac{\mathbb{E}\langle \Lambda(g_0(w), v), \theta(x) - \theta_n^\star(x) \rangle^2}{\mathbb{E}\|\theta(x) - \theta_n^\star(x)\|_2^2} \right\} \geq \gamma$, which for this setting simplifies to

$$\inf_{\theta \in \widehat{\Theta}_n} \left\{ \frac{\mathbb{E}(T - e_0(W))^2 (\theta(X) - \theta_n^\star(X))^2}{\mathbb{E}(\theta(X) - \theta_n^\star(X))^2} \right\} \geq \gamma. \tag{51}$$

31

One special case of the treatment effect setup, which was investigated in Chernozhukov et al. (2017) and Chernozhukov et al. (2018b) is where Θ_n is a class of high-dimensional predictors of the form $\theta(x) = \langle w, \phi(x) \rangle$, where $w \in \mathbb{R}^p$ and $\phi : \mathcal{X} \to \mathbb{R}^p$ is a fixed featurization. We allow the dimension p to grow with n, so in general we may have $p \gg n$. For this seting, to satisfy the conditions of Corollary 1 it suffices that $\text{Var}(\varepsilon_2 \mid X) \geq \eta$ for some $\eta > 0$ with no further assumptions on the data distribution or target parameter class. The latter condition is typically referred to as *overlap*, since for the case of a binary treatment it boils down to requiring that the treatment is not deterministic for any realization of the co-variates.

Compared to Chernozhukov et al. (2017, 2018b), our main theorems allow for misspecification of the target model. The convergence rate depends on the quantity γ in (51), which is considerably more benign than the minimum restricted eigenvalue of the matrix $\mathbb{E}[\phi(X)\phi(X)^\top]$, which was used in these works. Whenever the overlap condition is satisfied we have $\eta > \gamma$, but even when overlap is not satisfied the restricted eigenvalue assumption *does* imply (51), thereby recovering the earlier assumptions as a special case. Note that Chernozhukov et al. (2017, 2018b) investigated parameter recovery, for which restricted eigenvalue type conditions are a minimal assumption to guarantee consistency. Since we consider mean squared error, we can provide guarantees even when parameter recovery is impossible. Such is the case, for example, when the overlap condition is satisfied but the matrix $\mathbb{E}[\phi(X)\phi(X)^\top]$ has arbitrarily bad restricted eigenvalue.

Turning to slow rates, we remark that some distributions may simply not satisfy (51). In this case, we can appeal to Theorem 2 using Assumption 8. The assumption is trivially satisfied as long as the classes are bounded, and does not require any lower bounds in the vein of (51). While the dependence on first stage estimation error in Theorem 2 is worse than that in Theorem 1, orthogonality still helps out here when the target class is sufficiently large, cf. Figure 3.

Algorithms. Per the discussion above, the general guarantees for ERM developed in Section 5 can be applied to the treatment effect loss, both to obtain fast rates when sufficient conditions for strong convexity are met, and to obtain slow but variance-dependent rates otherwise. Suppose that the metric entropy of the target class Θ_n grows as $\mathcal{H}_2(\Theta_n, \varepsilon, n) \propto \varepsilon^{-p_2}$. Then the machinery of Section 6 implies that for all of the possible combinations of conditions above (fast vs. slow rate, well-specified vs. misspecified) ERM suffices to obtain optimal rates, *except* when the model is well-specified and $p_2 > 2$. In this case ERM is suboptimal (Rakhlin et al., 2017), but the Skeleton aggregation variant in Appendix F is sufficient.

Comparison to prior work on asymptotic normality in semi-parametric models. The treatment effects setting dates back to the early work of Robinson (1988) and more recently was analyzed by Chernozhukov et al. (2018a) through the lens of orthogonality. The results in these papers consider the case of a parametric class Θ_n and provide conditions for consistency and asymptotic normality of the estimated parameters. These results are mostly asymptotic in nature and require a well-specified model, i.e. θ_0 belongs to the parametric class Θ_n. Moreover, these results require strong convexity of the loss with respect to the parameters of the class, which is essentially an identifiability condition for the parameters. The rates typically depend on the degree of strong convexity in the leading term. Our results provide rates for excess risk and consistency with respect to a data dependent metric that drop both the well-specifiedness and the identifiability condition. Moreover, we do not constrain the algorithm used for fitting the second stage, while the classical results primarily consider least squares or GMM estimation for the second stage.

Comparison to prior work on penalized kernel target classes. Nie and Wager (2017) drops the well-specified model assumption and considers the case where Θ_n is a RKHS with bounded kernel norm and the second stage algorithm is penalized kernel regression. They provide results under conditions on the eigendecay of the kernel. Without making the eigendecay assumption, kernels with bounded RKHS norm are a convex class with metric entropy growth rate $p_2 = 2$ (Zhang, 2002). Hence, our results from Section 5 apply whenever ERM over the class is used. Our proof shows that one can achieve oracle rates for any such kernel class, with only second order influence from the nuisance error.

Comparison to prior work on forest-based estimation. A recent line of work (Athey et al., 2016; Oprescu et al., 2018; Friedberg et al., 2018) considers a particular type of nonparametric target class, where Θ_n is only assumed to be Lipschitz in x. These works aim for sup-norm rates and apply a specific type of local GMM estimation within a random forest. Our results provide L_2 prediction error rates as opposed to sup-norm rates. Via Section 6, this allows us to provide statistically optimal rates for *any* nonparametric class satisfying the metric entropy growth condition, not just for Lipschitz functions.[13]

Comparison to prior work on neural network estimation. Farrell et al. (2018) analyze the estimation of heterogeneous treatment effects when the nuisance estimation is conducted via deep neural network training. They provide fast statistical rates for estimation of smooth non-parametric function classes with neural nets, thereby enabling the use of these modern ML methods as nuisance estimators. In our framework, their results can be used as a black box: One can simply use the rates and algorithms they provide to instantiate the $\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S, \delta)$ quantity in our main theorems. Beyond nuisance estimation, we can also use their rates in our framework for the case when the *target parameter* is a deep neural network. In this case, our main theorems imply that we can use a larger neural net class for the nuisance parameter than for the target parameter while preserving oracle convergence rates.

In fact, we can also leverage the recent surge of work on the generalization ability of deep networks from the machine learning community (Bartlett et al., 2017; Golowich et al., 2018), which provide rates for the excess risk of deep neural networks that are almost independent of the neural network size. These results imply that one can obtain nearly size-independent rates for nuisance parameter estimation whenever the nuisance parameter can be expressed as a network with small (in appropriate norm) weights. Moreover, we can also use a deep neural network for target estimation. In this case, these results combined with our main theorems show that we can allow the norm bound for the nuisance neural net to be larger than that of the target network without affecting the leading term in the final rate.

8.2 Policy Learning

In the policy learning setting we receive examples of the form $Z = (Y, T, X)$, where $Y \in \mathbb{R}$ is an incurred loss, $T \in \mathcal{T}$ is a treatment vector and $X \in \mathcal{X}$ is a vector of covariates (also referred to as a context in the closely related contextual bandit literature). The treatment T is chosen based on some unknown, potentially randomized policy which depends on X. Specifically, we assume the following data generating process:

$$
\begin{aligned}
Y &= f_0(T, X) + \varepsilon_1, & \mathbb{E}[\varepsilon_1 \mid X, T] &= 0, \\
T &= e_0(X) + \varepsilon_2, & \mathbb{E}[\varepsilon_2 \mid X] &= 0.
\end{aligned}
\tag{52}
$$

---

[13]Lipschitz functions in d dimensions satisfy Assumption 14 with $p_2 = d$.

The learner wishes to optimize over a set of treatment policies $\Theta_n \subseteq (\mathcal{X} \to \mathcal{T})$ (i.e., policies take as input covariates X and return a treatment). Their goal is to produce a policy $\widehat{\theta}_n$ that achieves small regret with respect to the population risk:

$$\mathbb{E}\big[f_0(\widehat{\theta}_n(X), X)\big] - \min_{\theta \in \Theta_n} \mathbb{E}[f_0(\theta(X), X)]. \tag{53}$$

This formulation has been extensively studied in statistics (Qian and Murphy, 2011; Zhao et al., 2012; Zhou et al., 2017; Athey and Wager, 2017; Zhou et al., 2018) and machine learning (Beygelzimer and Langford, 2009; Dudík et al., 2011; Swaminathan and Joachims, 2015a; Kallus and Zhou, 2018) (in the latter, it is sometimes referred to as *counterfactual risk minimization*).

Note that the learner does not know the function f_0, typically referred to as the counterfactual outcome function. This function is treated as a nuisance parameter. Typically orthogonalization of this nuisance parameter is possible by utilizing the secondary treatment equation and fitting a model for the treatment policy e_0, which is also a nuisance parameter. We can then typically write the expected counterfactual reward as

$$f_0(t, X) = \mathbb{E}[\ell(t, Z; f_0, e_0) \mid X] \tag{54}$$

for some known function ℓ that utilizes the treatment model e_0. Letting $g_0 = \{f_0, e_0\}$, the learner's goal can be phrased as minimizing the population risk,

$$\mathbb{E}[f_0(\theta(X), X)] = \mathbb{E}[\mathbb{E}[\ell(\theta(X), Z; f_0, e_0) \mid X]] = \mathbb{E}[\ell(\theta(X), Z; f_0, e_0)] =: L_\mathcal{D}(\theta, g_0), \tag{55}$$

over $\theta \in \Theta_n$. This formulation clearly falls into our orthogonal statistical learning framework, where the target parameter is the policy θ and the counterfactual outcome f_0 and observed treatment policy e_0 together form the nuisance parameter $g_0 = \{f_0, e_0\} \in \mathcal{G}_n$.

Binary treatments. As a first example, consider the special case of binary treatment $T \in \{0, 1\}$, analyzed in Athey and Wager (2017). To simplify notation, define $p_0(t, x) = \mathbb{P}[T = t \mid X = x]$, and observe that $p_0(t, x) = e_0(x)$ if $t = 1$ and $1 - e_0(x)$ if $t = 0$. Then the loss function

$$\ell(t, Z; f_0, e_0) = f_0(t, X) + 1\{T = t\}\frac{Y - f_0(t, X)}{p_0(t, X)}, \tag{56}$$

has the structure in (55), i.e. it evaluates to the true risk (53) whenever the true nuisance parameter values are plugged in. This formulation leads to the well-known doubly-robust estimator for the counterfactual outcome (Cassel et al., 1976; Robins et al., 1994; Robins and Rotnitzky, 1995; Dudík et al., 2011). It is easy to verify that the resulting population risk is orthogonal with respect to both f_0 and p_0. Furthermore, note that in this setting we can obtain a new loss by subtracting the loss incurred by choosing treatment 0, it is equivalent to optimize the loss:

$$\ell(t, Z; f_0, e_0) = \underbrace{\left(f_0(1, X) - f_0(0, X) + 1\{T = 1\}\frac{Y - f_0(1, X)}{e_0(X)} + 1\{T = 0\}\frac{Y - f_0(0, X)}{1 - e_0(X)} \right)}_{\beta(f_0, e_0, Z)} \cdot t. \tag{57}$$

This formulation leads to a linear population risk:

$$L_D(\theta, f_0, e_0) = \mathbb{E}[\beta(f_0, e_0, Z) \cdot \theta(x)]. \tag{58}$$

It is straightforward to show that the this population risk satisfies universal orthogonality, and so Assumption 8 is satisfied whenever the nuisance parameters are bounded appropriately.

Multiple finite treatments. The binary setting above can easily be extended to the case of N possible treatments, analyzed in Zhou et al. (2018). Formally, let $T \in \{\vec{e}_1, \ldots, \vec{e}_N\}$, where $\vec{e}_i \in \{0,1\}^N$ is the i-th standard basis vector. We still follow the data generating process (52), but now $e_0 : \mathcal{X} \to \Delta_N$ and $f_0 : \{0,1\}^N \times \mathcal{X} \to \mathbb{R}$. To simplify notation, let $p_0(t,x) = \Pr[T = t \mid X = x]$ so $p_0(t,x) = e_0(x)_t$. Then the following loss function is an unbiased estimate of the counterfactual loss:

$$\ell(t, Z; f_0, e_0) = f_0(t, X) + 1\{T = t\} \frac{Y - f_0(t, X)}{p_0(t, X)}. \tag{59}$$

This formulation leads to the standard extension of the doubly-robust estimator to multiple outcomes Dudík et al. (2011); Zhou et al. (2018). Define an N-dimensional vector-valued function $\beta(f_0, e_0, Z)$ to have the t-th coordinate is equal to $\ell(t, Z; f_0, e_0)$. Then, as in the binary case, we can equivalently optimize a population risk that is linear in the target parameter:

$$L_\mathcal{D}(\theta, f_0, e_0) = \mathbb{E}[\langle \beta(f_0, e_0, Z), \theta(x) \rangle]. \tag{60}$$

This population risk is easily shown to satisfy universal orthogonality.

Counterfactual risk minimization and general continuous treatments. Counterfactual risk minimization (CRM) is a learning framework introduced by Swaminathan and Joachims (2015a). It is mathematically equivalent to the policy learning setup with arbitrary treatment and outcome spaces, but is motivated by a different set real-world learning scenarios and was developed in a parallel line of research in the machine learning literature. To highlight the relationship with policy learning and the applicability of our results to this setting we will present the CRM framework using the notation of policy learning.

In counterfactual risk minimization we receive data $Z = (Y, T, X)$ from the policy learning data generating process (52). The goal is to choose a hypothesis $\theta : \mathcal{X} \to \Delta(\mathcal{T})$ (i.e., the policy takes as input covariates and returns a distribution over treatments) that minimizes the population risk:

$$L_\mathcal{D}^1(\theta, f_0) = \mathbb{E}_Z \, \mathbb{E}_{t \sim \theta(X)}[f_0(t, X)]. \tag{61}$$

As in policy learning, we construct an unbiased estimate of this counterfactual loss via inverse propensity scoring. Let $p_0(t, X)$ denote the probability density of treatment t conditional on covariates X and (overloading notation) let $\theta(t, x)$ denote the density that hypothesis θ assigns to treatment t. Then we can formulate a new risk function that provides an unbiased estimate of the target risk (61):

$$L_\mathcal{D}^2(\theta, p_0) = \mathbb{E}_Z\left[Y \frac{\theta(T, X)}{p_0(T, X)}\right]. \tag{62}$$

In the CRM framework the propensity p_0 is assumed to be known (Swaminathan and Joachims, 2015a,b). When the propensity is not known, we can treat it as a nuisance parameter to be estimated from data. However, the loss (62) is not orthogonal to p_0. We can orthogonalize the population risk by also constructing an estimate of f_0 (see (52)) by regressing Y on (T, X). This leads to an analogue of the doubly robust formulation from the finite treatment setup:

$$L_\mathcal{D}(\theta, f_0, p_0) = \mathbb{E}\left[\underbrace{\left(f_0(T, X) + \frac{(Y - f_0(T, X))}{p_0(T, X)}\right) \cdot \theta(T, X)}_{\beta(f_0, p_0, Z)}\right]. \tag{63}$$

For finite treatments this formulation is mathematically equivalent to the population risk for multiple finite treatments presented in the prequel.

For continuous treatments, the empirical version of the problem (63) may be ill-posed, even if we assume that the propensity p_0 has density ower bounded by some constant (the analogue of the overlap condition). Swaminathan and Joachims (2015a) proposed to regularize the empirical risk via variance penalization. A similar variance penalization approach is also proposed in recent work of Bertsimas and McCord (2018), who consider policy learning over arbitrary treatment spaces. The variance-penalized empirical risk minimization algorithm—in the context of Algorithm 1—can be seen as a second stage algorithm that achieves a rate whose leading term scales with the variance of the optimal policy rather than some worst-case upper bound on the risk. Hence, it can be used in our framework to achieve variance-dependent excess risk bounds.

Kallus and Zhou (2018) develop alternative algorithms for policy learning with continuous treatments via a kernel smoothing approach. This approach is equivalent to adding noise to a deterministic hypothesis space Π_n, e.g. $\theta(x) = \pi(x) + \zeta$ for each $\pi \in \Pi_n$, where $\zeta \sim \mathcal{N}(0, \sigma^2)$. In our framework Θ_n is the space of density functions $\theta(t, x)$ induced by this construction. The value of a deterministic policy $\pi \in \Pi_n$ (or equivalently the value of its corresponding density $\theta \in \Theta_n$) is equal to

$$\mathbb{E}[\beta(f_0, p_0, Z) \cdot \mathcal{K}_\sigma(T - \pi(X))], \tag{64}$$

where $\mathcal{K}_\sigma$ is the pdf of a normal distribution with standard deviation σ. This is equivalent to the formulation in Kallus and Zhou (2018), since the empirical version of this risk is the kernel-weighted loss:[14]

$$L_S(\pi, f_0, p_0) = \frac{1}{n} \sum_{t=1}^{n} \beta(f_0, p_0, Z_i) \cdot \mathcal{K}_\sigma(T_i - \pi(X_i)) \tag{65}$$

This idea falls into our framework by simply defining Θ_n to be this space of randomly perturbed policy functions. The resulting analysis in our framework is slightly different than that of Kallus and Zhou (2018), where kernel weighting is invoked to show consistency of the empirical risk, and subsequently optimization of the empirical risk over *deterministic* policies is analyzed. With our framework, we directly calculate the regret with respect to *randomized* policies by applying our general theorem. This implies that we enjoy robustness to errors in estimating f_0 and p_0.

Observe that the rate for the second stage will depend on the amount of randomization σ, since the variance of the empirical risk is governed by σ. Consequently, if one is interested in regret against deterministic strategies, we can invoke Lipschitzness of the reward function f_0 to control the regret added by the extra randomness we are injecting, which would typically be of order σ. We can then choose an optimal σ as a function of the number of samples to trade-off the *bias* and *variance* of the regret. If we wish to further optimize dependence on problem-dependent parameters in the resulting rates one can use variance penalization in the kernel-based framework to achieve a regret rate whose leading term scales with the variance of the optimal policy.

8.3 Domain Adaptation and Sample Bias Correction

Domain adaptation is a widely studied topic in machine learning (Daume III and Marcu, 2006; Jiang and Zhai, 2007; Ben-David et al., 2007; Blitzer et al., 2008; Mansour et al., 2009). The goal is to choose a hypothesis that minimizes a given loss in expectation over a target data distribution, where the target distribution may be different from the distribution of data that is already collected.

We consider a particular instance of domain adaptation called covariate shift, encountered in supervised learning (Shimodaira, 2000). We assume that we have data $Z = (X, Y)$, where X

---

[14]We note that Kallus and Zhou (2018) do not restrict themselves to only the gaussian kernel.

are co-variates drawn from some distribution $\mathcal{D}_s$ with density p_s and Y are labels, drawn from some distribution $\mathcal{D}_x$ conditional on x. Our goal is to choose a hypothesis θ from some hypothesis space Θ_n, so as to minimize a loss function $\ell(\theta(x), y)$ in expectation over a different distribution of co-variates $\mathcal{D}_t$ with density p_t. Both of the densities are unknown, and we solve this issue in the orthogonal statistical learning framework by treating their ratio as a nuisance parameter for an importance-weighted loss function. Let $f_0(x) = \frac{p_t(x)}{p_s(x)}$ and $g_0 = \{f_0\}$, so that

$$L_D(\theta, g_0) := \mathbb{E}_{\mathcal{D}_s} \mathbb{E}_{y|x}[\ell(\theta(x), y) \cdot f_0(x)] = \mathbb{E}_{\mathcal{D}_t} \mathbb{E}_{y|x}[\ell(\theta(x), y)]. \tag{66}$$

We assume the hypothesis space satisfies a realizability condition, i.e. there exists $\theta_0 \in \Theta_n$ such that:

$$\mathbb{E}\big[\nabla_\zeta \ell(\theta_0(x), y) \mid x\big] = 0, \tag{67}$$

where ∇_ζ corresponds to the gradient with respect to the first input of ℓ. For instance, for the case of the square loss $\ell(\zeta, y) = (\zeta - y)^2$, then this condition has the natural interpretation that there exists $\theta_0 \in \Theta_n$ such that:

$$\mathbb{E}[y \mid x] = \theta_0(x). \tag{68}$$

Observe that when we treat the density ratio $f_0(x)$ as a nuisance function, the loss function L_D is orthogonal. Indeed,

$$
\begin{aligned}
D_f D_\theta L_D(\theta, g)[\theta - \theta_0, f - f_0] &= \mathbb{E}\big[\nabla_\zeta \ell(\theta_0(x), y) \cdot (\theta(x) - \theta_0(x)) \cdot (f(x) - f_0(x))\big] \\
&= \mathbb{E}\big[\mathbb{E}[\nabla_\zeta \ell(\theta_0(x), y) \mid x] \cdot (\theta(x) - \theta_0(x)) \cdot (f(x) - f_0(x))\big] = 0.
\end{aligned}
$$

Also, note that this setup fits in to the single index structure from Section 4 by writing $L_{\mathcal{D}}$ as the expectation of a new loss $\tilde{\ell}(\theta(x), g(w), z) := \ell(\theta(x), y) \cdot f(x)$. Focusing on the square loss $\ell(\zeta, y) = (\zeta - y)^2$ for concreteness, it is simple to show that all of the conditions of Assumption 7 are satisfied as long as we have $g(x) \geq \eta > 0$ for all $g \in \mathcal{G}_n$. Corollary 1 then implies that with probability at least $1 - \delta$, Algorithm 1 enjoys the bound

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0) \leq O\Big(\text{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta/2; \widehat{g}_n) + \text{poly}(\eta^{-1}) \cdot \big(\text{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2)\big)^4\Big).$$

Note that whenever $\text{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2) = o\big(\text{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta/2; \widehat{g}_n)^{1/4}\big)$ the dependence on η^{-1} is negligible asymptotically. Of course, it is also important to develop algorithms for which the rate of the target class does not depend on η^{-1}. As one example, we can employ the variance-penalized ERM guarantee from Theorem 6. When Θ_n has VC dimension d, and the variance of the loss at (θ_0, g_0) and the capacity function τ_0 are bounded, this gives $\text{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2) = O(\sqrt{d/n})$, with η^{-1} entering only lower-order terms. The final result is that if $\text{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2) = o(n^{-1/8})$, we get an excess risk bound for which the dominant term is $O(\sqrt{d/n})$, with no dependence on η^{-1}.

Related work. Cortes et al. (2010) gave generalization error guarantees for the important weighted loss (66) in the case where the densities p_s and p_t are known. At the other extreme, Ben-David and Urner (2012) showed strong impossibility results in the regime where the densities are unknown. Our results lie in the middle, and show that learning with unknown densities is possible in the regime where the weights belong to a nonparametric class that is not much more complex than the target predictor class Θ_n. We remark in passing that algorithms based on discrepancy minimization Ben-David et al. (2007); Mansour et al. (2009) offer another approach to domain adaptation that does not require importance weights, but these results are not directly comparable to our own.

8.4 Missing Data

As a final application, we consider the problem of regression with missing data. In this setting we receive data is generated through standard regression model, but label/target variables are sometimes "missing" or unobserved. The learner observes whether or not the target is missing for each example, and the conditional probability that the target is missing is treated as an unknown nuisance parameter. As usual, the unknown regression function is the target. This setting was addressed for low-dimensional target parameters by Graham (2011) and extended to high-dimensional target parameters by Chernozhukov et al. (2018b) using the orthogonal/debiased learning framework. Using our tools, we can provide guarantees when the target parameters belong to arbitrary, potentially nonparametric function classes.

To proceed, we formalize the setting through the following data-generating process for the observed variables $(X, W, T, \widetilde{Y})$:

$$
\begin{aligned}
\widetilde{Y} &= T \cdot Y, \\
Y &= \theta_0(X) + \varepsilon_1, \quad \mathbb{E}[\varepsilon_1 \mid X] = 0, \\
T &= e_0(W) + \varepsilon_2, \quad \mathbb{E}[\varepsilon_2 \mid W] = 0.
\end{aligned}
\tag{69}
$$

Here $T \in \{0, 1\}$ is an auxiliary variable (observed by the learner) that indicates whether the target variable is missing, and $e_0 : \mathcal{X} \to [0, 1]$ is the unknown propensity for T. The parameter $\theta_0 : \mathcal{X} \to \mathbb{R}$ is the true regression function. We make the standard *unobserved confounders* assumption that $X \subseteq W$ and $T \perp Y \mid W$. We define $h_0(w) = -2 \frac{(\mathbb{E}[Y|W=w] - \theta_0(x))}{e_0(w)}$, take $g_0 = \{h_0, e_0\}$, and use the loss

$$
\ell(\theta, \{h, e\} ; z) = \frac{T(\widetilde{Y} - \theta(X))^2}{e(W)} - \theta(X) h(W)(T - e(W)).
\tag{70}
$$

Observe that this loss has the property that

$$
L_{\mathcal{D}}(\theta, g_0) = \mathbb{E}_{X,Y}(Y - \theta(X))^2,
$$

so that the excess risk relative to θ_0 precisely corresponds to prediction accuracy whenever the true nuisance parameter is plugged in. This model satisfies Assumption 1 and Assumption 2 whenever $\theta_0 \in \Theta_n$, i.e. we have $D_\theta L_{\mathcal{D}}(\theta_0, \{h, e_0\})[\theta - \theta_0] = 0$, $D_e D_\theta L_{\mathcal{D}}(\theta_0, \{h_0, e_0\})[\theta - \theta_0, e - e_0] = 0$, and $D_h D_\theta L_{\mathcal{D}}(\theta_0, \{h_0, e_0\})[\theta - \theta_0, h - h_0] = 0$ (see Appendix G). Note that the extra nuisance parameter h is only required here because we consider the general setting in which $W \neq X$. Whenever $W = X$ this is unnecessary (and indeed $h_0 = 0$). This parameter can generally be estimated at a rate no worse than the rate for e_0 and θ_0 (absent nuisance parameters); see Chernozhukov et al. (2018b) for discussion.

As to rates and algorithms, the situation here is essentially the same as that of the domain adaptation example, so we discuss it only briefly. The setup has the single index structure from Section 4, and all of the sufficient conditions for fast rates from Assumption 7 are satisfied as long as we have $e(W) \geq \eta > 0$ for all e in the nuisance class. Thus, with probability at least $1 - \delta$, Algorithm 1 enjoys the bound

$$
L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0) \leq O\Big(\mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta/2 ; \widehat{g}_n) + \mathrm{poly}(\eta^{-1}) \cdot \Big(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2) \Big)^4 \Big).
$$

As with the previous example, the variance-penalized ERM guarantees from Section 5 can be applied here to provide bounds on $\mathrm{Rate}_{\mathcal{D}}(\Theta_n, \ldots)$ for which the dominant term in the excess risk does not scale with the inverse propensity range.

9   Discussion

This paper initiates a study of prediction error and excess risk guarantees in the presence of nuisance parameters and Neyman orthogonality. Our results highlight that orthogonality is beneficial for learning with nuisance parameters even in the presence of possible model misspecification, and even when the target parameters belong to large nonparametric classes. We also show that many of the typical assumptions used to analyze estimation in the presence of nuisance parameters can be dropped when prediction error is the target here. There are many promising future directions, including weakening assumptions, obtaining sharper guarantees for specific settings and losses of interest, and analyzing further algorithms for general function classes (along the lines of Section 5).

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018.

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 1999.

Susan Athey and Stefan Wager. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *arXiv preprint arXiv:1610.01271*, 2016.

Jean-Yves Audibert. Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems*, pages 41–48, 2008.

Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Peter L Bartlett, Nick Harvey, Chris Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Conference on Learning Theory*, 2017.

Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *International Conference on Algorithmic Learning Theory*, pages 139–153. Springer, 2012.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007.

Dimitris Bertsimas and Christopher McCord. Optimization over Continuous and Multi-dimensional Decisions with Observational Data. *arXiv preprint arXiv:1807.04183*, 2018.

Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 129–138. ACM, 2009.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Advances in neural information processing systems*, pages 129–136, 2008.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.

Claes M Cassel, Carl E Särndal, and Jan H Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620, 1976.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, and Whitney K Newey. Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*, 2016.

Victor Chernozhukov, Matt Goldman, Vira Semenova, and Matt Taddy. Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. *arXiv preprint arXiv:1712.09988*, 2017.

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):C1–C68, 2018a.

Victor Chernozhukov, Denis Nekipelov, Vira Semenova, and Vasilis Syrgkanis. Plug-in regularized estimation of high-dimensional parameters in nonlinear semiparametric models. *arXiv preprint arXiv:1806.04823*, abs/1806.04823, 2018b.

Victor Chernozhukov, Whitney Newey, and James Robins. Double/de-biased machine learning using regularized riesz representers. *arXiv preprint arXiv:1802.08667*, 2018c.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450, 2010.

Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126, 2006.

Bernard Delyon and Anatoli Juditsky. On minimax wavelet estimators. *Applied and Computational Harmonic Analysis*, 3(3):215–228, 1996.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1097–1104. Omnipress, 2011.

Max. H. Farrell, Tengyuan Liang, and Sanjog Misra. Deep Neural Networks for Estimation and Inference: Application to Causal Effects and Other Semiparametric Estimands. *ArXiv e-prints*, September 2018.

Dylan J Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. *Conference on Learning Theory*, 2018.

Rina Friedberg, Julie Tibshirani, Susan Athey, and Stefan Wager. Local Linear Forests. *ArXiv e-prints*, July 2018.

Evarist Giné, Vladimir Koltchinskii, et al. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-Independent Sample Complexity of Neural Networks. *Conference on Learning Theory*, 2018.

Bryan S Graham. Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica*, 79(2):437–452, 2011.

Adityanand Guntuboyina and Bodhisattva Sen. Global risk bounds and adaptation in univariate convex regression. *Probability Theory and Related Fields*, 163(1-2):379–411, 2015.

Steve Hanneke et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 264–271, 2007.

Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. *arXiv preprint arXiv:1802.06037*, 2018.

Gérard Kerkyacharian, Oleg Lepski, and Dominique Picard. Nonlinear estimation in anisotropic multi-index denoising. *Probability theory and related fields*, 121(2):137–170, 2001.

Gérard Kerkyacharian, Oleg Lepski, and Dominique Picard. Nonlinear estimation in anisotropic multi-index denoising. sparse case. *Theory of Probability & Its Applications*, 52(1):58–77, 2008.

V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. *High Dimensional Probability II*, 47:443–459, 2000.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, 2011.

Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer, 2008.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Meta-learners for estimating heterogeneous treatment effects using machine learning. *arXiv preprint arXiv:1706.03461*, 2017.

Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. *arXiv preprint arXiv:1305.4825*, 2013.

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

Oleg Lepski. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1992.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Proceedings of The 28th Conference on Learning Theory*, pages 1260–1285, 2015.

Lester Mackey, Vasilis Syrgkanis, and Ilias Zadik. Orthogonal machine learning: Power and limitations. *arXiv preprint arXiv:1711.00342*, 2017.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer, 2016.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *The 22nd Conference on Learning Theory (COLT*, 2009.

Shahar Mendelson. Improving the sample complexity using global data. *IEEE transactions on Information Theory*, 48(7):1977–1991, 2002.

Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

Shahar Mendelson et al. Discrepancy, chaining and subgaussian processes. *The Annals of Probability*, 39(3):985–1026, 2011.

Jerzy Neyman. Optimal asymptotic tests of composite hypotheses. *Probability and statsitics*, pages 213–234, 1959.

Jerzy Neyman. C (α) tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 1–21, 1979.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for heterogeneous treatment effect estimation. *arXiv preprint arXiv:1806.03467*, 2018.

Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180, 2011.

Alexander Rakhlin and Karthik Sridharan. Statistical learning and sequential prediction, 2012. Available at `http://stat.wharton.upenn.edu/~rakhlin/book_draft.pdf`.

Alexander Rakhlin, Karthik Sridharan, Alexandre B Tsybakov, et al. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427): 846–866, 1994.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, pages 2199–2207, 2010.

Charles J Stone. Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pages 1348–1360, 1980.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015a.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, pages 3231–3239, 2015b.

Alexandre B Tsybakov et al. Pointwise and sup-norm sharp adaptive estimation of functions on the sobolev classes. *The Annals of Statistics*, 26(6):2420–2469, 1998.

A. W. Van Der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer Series, March 1996.

Vladimir N Vapnik. The nature of statistical learning theory. 1995.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

Xin Zhou, Nicole Mayer-Hamblett, Umer Khan, and Michael R Kosorok. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517):169–187, 2017.

Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:arXiv:1810.04778*, 2018.

A Preliminaries

We invoke the following version of Taylor's theorem and its directional derivative generalization repeatedly.

Proposition 2 (Taylor expansion). *Let $a \leq b \in \mathbb{R}$ be fixed and let $f : I \to \mathbb{R}$, where $I \subseteq \mathbb{R}$ is an open interval containing a, b. If $f^{(n+1)}$ is continuous, then there exists $c \in [a, b]$ such that*

$$f(a) = f(b) + \sum_{k=1}^{n} \frac{1}{k!} f^{(k)}(b)(a-b)^k + \frac{1}{(n+1)!} f^{(n+1)}(c)(a-b)^{n+1}.$$

Let $F : \mathcal{V} \to \mathbb{R}$, where $\mathcal{V}$ is a vector space of functions. For any $g, g' \in \mathcal{V}$, if $t \mapsto F(t \cdot g + (1-t) \cdot g')$ has a continuous $(n+1)$th derivative over an open interval containing $[0, 1]$, then there exists $\bar{g} \in \mathrm{conv}(\{g, g'\})$ such that

$$F(g') = F(g) + \sum_{k=1}^{n} \frac{1}{k!} D_g^k F(g) [\underbrace{g'-g, \ldots, g'-g}_{k \; times}] + \frac{1}{(n+1)!} D_g^{n+1} F(\bar{g}) [\underbrace{g'-g, \ldots, g'-g}_{n+1 \; times}].$$

B Proofs from Section 3

Proof of Theorem 1. To simplify notation we drop the subscripts from the norms $\|\cdot\|_{\Theta_n}$ and $\|\cdot\|_{\mathcal{G}_n}$ and abbreviate $R_\theta := \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta/2; \widehat{g}_n)$ and $R_g := \mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2)$. Using a second-order Taylor expansion on the risk at $\widehat{g}_n$, there exists $\bar{\theta} \in \mathrm{star}(\widehat{\Theta}_n, \theta_n^\star)$ such that

$$\frac{1}{2} \cdot D_\theta^2 L_{\mathcal{D}}(\bar{\theta}, \widehat{g}_n)[\widehat{\theta}_n - \theta_n^\star, \widehat{\theta}_n - \theta_n^\star] = L_{\mathcal{D}}(\widehat{\theta}_n, \widehat{g}_n) - L_{\mathcal{D}}(\theta_n^\star, \widehat{g}_n) - D_\theta L_{\mathcal{D}}(\theta_n^\star, \widehat{g}_n)[\widehat{\theta}_n - \theta_n^\star].$$

Using the strong convexity assumption (Assumption 3) we get:

$$D_\theta^2 L_{\mathcal{D}}(\bar{\theta}, \widehat{g}_n)[\widehat{\theta}_n - \theta_n^\star, \widehat{\theta}_n - \theta_n^\star] \geq \lambda \cdot \left\|\widehat{\theta}_n - \theta_n^\star\right\|_{\Theta_n}^2 - \kappa \cdot \|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4 = \lambda \cdot \left\|\widehat{\theta}_n - \theta_n^\star\right\|_{\Theta_n}^2 - \kappa \cdot R_g^4.$$

Combining these statements we conclude that

$$\frac{\lambda}{2} \cdot \left\|\widehat{\theta}_n - \theta_n^\star\right\|_{\Theta_n}^2 \leq L_{\mathcal{D}}(\widehat{\theta}_n, \widehat{g}_n) - L_{\mathcal{D}}(\theta_n^\star, \widehat{g}_n) + \frac{\kappa}{2} \cdot R_g^4 - D_\theta L_{\mathcal{D}}(\theta_n^\star, \widehat{g}_n)[\widehat{\theta}_n - \theta_n^\star].$$

Using the assumed rate for $\widehat{\theta}_n$ (Definition 2), this implies the inequality

$$\frac{\lambda}{2} \cdot \left\|\widehat{\theta}_n - \theta_n^\star\right\|_{\Theta_n}^2 \leq R_\theta + \frac{\kappa}{2} \cdot R_g^4 - D_\theta L_{\mathcal{D}}(\theta_n^\star, \widehat{g}_n)[\widehat{\theta}_n - \theta_n^\star]. \tag{71}$$

We now apply a second-order Taylor expansion (using the assumed derivative continuity from Assumption 4), which implies that there exists $\bar{g} \in \mathrm{star}(\mathcal{G}_n, g_0)$ such that

$$- D_\theta L_{\mathcal{D}}(\theta_n^\star, \widehat{g}_n)[\widehat{\theta}_n - \theta_n^\star]$$

$$= -D_\theta L_{\mathcal{D}}(\theta_n^\star, g_0)[\widehat{\theta}_n - \theta_n^\star] - D_g D_\theta L_{\mathcal{D}}(\theta_n^\star, g_0)[\widehat{\theta}_n - \theta_n^\star, \widehat{g}_n - g_0] - \frac{1}{2} \cdot D_g^2 D_\theta L_{\mathcal{D}}(\theta_n^\star, \bar{g})[\widehat{\theta}_n - \theta_n^\star, \widehat{g}_n - g_0, \widehat{g}_n - g_0].$$

Using orthogonality of the loss (Assumption 1), this is equal to

$$= -D_\theta L_{\mathcal{D}}(\theta_n^\star, g_0)[\widehat{\theta}_n - \theta_n^\star] - \frac{1}{2} \cdot D_g^2 D_\theta L_{\mathcal{D}}(\theta_n^\star, \bar{g})[\widehat{\theta}_n - \theta_n^\star, \widehat{g}_n - g_0, \widehat{g}_n - g_0].$$

We use the second order smoothness assumed in Assumption 4:

$$\leq -D_\theta L_\mathcal{D}(\theta_n^\star, g_0)[\widehat{\theta}_n - \theta_n^\star] + \frac{\beta_2}{2} \cdot \|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^2 \cdot \|\widehat{\theta}_n - \theta_n^\star\|_{\Theta_n}.$$

Invoking the AM-GM inequality, we have that for any constant $\eta > 0$:

$$\leq -D_\theta L_\mathcal{D}(\theta_n^\star, g_0)[\widehat{\theta}_n - \theta_n^\star] + \frac{\beta_2}{4\eta} \cdot \|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4 + \frac{\beta_2 \eta}{4} \cdot \|\widehat{\theta}_n - \theta_n^\star\|_{\Theta_n}^2$$

Lastly, we use the assumed rate for $\widehat{g}_n$ (Definition 2):

$$\leq -D_\theta L_\mathcal{D}(\theta_n^\star, g_0)[\widehat{\theta}_n - \theta_n^\star] + \frac{\beta_2}{4\eta} \cdot R_g^4 + \frac{\beta_2 \eta}{4} \cdot \|\widehat{\theta}_n - \theta_n^\star\|_{\Theta_n}^2$$

Choosing $\eta = \frac{\lambda}{\beta_2}$ and combining this string of inequalities with (71) and rearranging, we get:

$$\|\widehat{\theta}_n - \theta_n^\star\|_{\Theta_n}^2 \leq \frac{4}{\lambda}\left(-D_\theta L_\mathcal{D}(\theta_n^\star, g_0)[\widehat{\theta}_n - \theta_n^\star] + R_\theta\right) + \left(\frac{\beta_2^2}{\lambda^2} + \frac{2\kappa}{\lambda}\right) \cdot R_g^4 \tag{72}$$

Assumption 2 implies that $D_\theta L_\mathcal{D}(\theta_n^\star, g_0)[\widehat{\theta}_n - \theta_n^\star] \geq 0$. Hence, we get the desired inequality (10).

To conclude the inequality (11), we use another Taylor expansion, which implies that there exists $\bar{\theta} \in \text{star}(\widehat{\Theta}_n, \theta_n^\star)$ such that

$$L_\mathcal{D}(\widehat{\theta}_n, g_0) - L_\mathcal{D}(\theta_n^\star, g_0) = D_\theta L_\mathcal{D}(\theta_n^\star, g_0)[\widehat{\theta}_n - \theta_n^\star] + \frac{1}{2} \cdot D_\theta^2 L_\mathcal{D}(\bar{\theta}, g_0)[\widehat{\theta}_n - \theta_n^\star, \widehat{\theta}_n - \theta_n^\star]$$

Using the smoothness bound from Assumption 4:

$$\leq D_\theta L_\mathcal{D}(\theta_n^\star, g_0)[\widehat{\theta}_n - \theta_n^\star] + \frac{\beta_1}{2} \cdot \|\widehat{\theta}_n - \theta_n^\star\|_{\Theta_n}^2 \tag{73}$$

We combine (72) with (73) to get

$$L_\mathcal{D}(\widehat{\theta}_n, g_0) - L_\mathcal{D}(\theta_n^\star, g_0) \leq \frac{2\beta_1}{\lambda} \cdot R_\theta + \frac{\beta_1}{2}\left(\frac{\beta_2^2}{\lambda^2} + \frac{2\kappa}{\lambda}\right) \cdot R_g^4 - \left(\frac{2\beta_1}{\lambda} - 1\right) \cdot D_\theta L_\mathcal{D}(\theta_n^\star, g_0)[\widehat{\theta}_n - \theta_n^\star].$$

The result follows by again using that $D_\theta L_\mathcal{D}(\theta_n^\star, g_0)[\widehat{\theta}_n - \theta_n^\star] \geq 0$, along with the fact that $\beta_1/\lambda \geq 1$, without loss of generality. $\qquad\square$

Proof of Theorem 2. To begin, we use the guarantee for the second stage from Definition 2 and perform straightforward manipulation to show

$$L_\mathcal{D}(\widehat{\theta}_n, g_0) - L_\mathcal{D}(\theta_n^\star, g_0) \leq (L_\mathcal{D}(\widehat{\theta}_n, g_0) - L_\mathcal{D}(\widehat{\theta}_n, \widehat{g}_n)) + (L_\mathcal{D}(\theta_n^\star, \widehat{g}_n) - L_\mathcal{D}(\theta_n^\star, g_0)) + \text{Rate}_\mathcal{D}(\Theta_n, S^{(2)}, \delta/2\,; \widehat{g}_n).$$

Using continuity guaranteed by Assumption 6, we perform a second-order Taylor expansion with respect to g for each pair of loss terms in the preceding expression to conclude that there exist $g, g' \in \text{star}(\mathcal{G}_n, g_0)$ such that

$$(L_\mathcal{D}(\widehat{\theta}_n, g_0) - L_\mathcal{D}(\widehat{\theta}_n, \widehat{g}_n)) + (L_\mathcal{D}(\theta_n^\star, \widehat{g}_n) - L_\mathcal{D}(\theta_n^\star, g_0))$$

$$= -D_g L_\mathcal{D}(\widehat{\theta}_n, g_0)[\widehat{g}_n - g_0] - \frac{1}{2} \cdot D_g^2 L_\mathcal{D}(\widehat{\theta}_n, g)[\widehat{g}_n - g_0, \widehat{g}_n - g_0]$$

$$+ D_g L_\mathcal{D}(\theta_n^\star, g_0)[\widehat{g}_n - g_0] + \frac{1}{2} \cdot D_g^2 L_\mathcal{D}(\theta_n^\star, g')[\widehat{g}_n - g_0, \widehat{g}_n - g_0].$$

Using the smoothness promised by Assumption 6:

$$\leq -D_g L_{\mathcal{D}}(\widehat{\theta}_n, g_0)[\widehat{g}_n - g_0] + D_g L_{\mathcal{D}}(\theta_n^\star, g_0)[\widehat{g}_n - g_0] + \beta \|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^2.$$

To relate the two derivative terms, we apply another second-order Taylor expansion (which is possible due to Assumption 6), this time with respect to the target predictor.

$$D_g L_{\mathcal{D}}(\widehat{\theta}_n, g_0)[\widehat{g}_n - g_0]$$
$$= D_g L_{\mathcal{D}}(\theta_n^\star, g_0)[\widehat{g}_n - g_0] + D_\theta D_g L_{\mathcal{D}}(\theta_n^\star, g_0)[\widehat{g}_n - g_0, \widehat{\theta}_n - \theta_n^\star] + \frac{1}{2} \cdot D_\theta^2 D_g L_{\mathcal{D}}(\bar{\theta}, g_0)[\widehat{g}_n - g_0, \widehat{\theta}_n - \theta_n^\star, \widehat{\theta}_n - \theta_n^\star],$$

where $\bar{\theta} \in \mathrm{conv}(\{\widehat{\theta}_n, \theta_n^\star\})$. Universal orthogonality immediately implies that

$$D_\theta D_g L_{\mathcal{D}}(\theta_n^\star, g_0)[\widehat{g}_n - g_0, \widehat{\theta}_n - \theta_n^\star] = 0.$$

Furthermore, observe that

$$D_\theta^2 D_g L_{\mathcal{D}}(\bar{\theta}, g_0)[\widehat{g}_n - g_0, \widehat{\theta}_n - \theta_n^\star, \widehat{\theta}_n - \theta_n^\star]$$
$$= \lim_{t \to 0} \frac{D_\theta D_g L_{\mathcal{D}}(\bar{\theta} + t(\widehat{\theta}_n - \theta_n^\star), g_0)[\widehat{g}_n - g_0, \widehat{\theta}_n - \theta_n^\star] - D_\theta D_g L_{\mathcal{D}}(\bar{\theta}, g_0)[\widehat{g}_n - g_0, \widehat{\theta}_n - \theta_n^\star]}{t}.$$

Since $\bar{\theta} + t(\widehat{\theta}_n - \theta_n^\star) \in \mathrm{star}(\widehat{\Theta}_n, \theta_n^\star) + \mathrm{star}(\widehat{\Theta}_n - \theta_n^\star, 0)$ for all $t \in [0,1]$, including $t = 0$, universal orthogonality (Assumption 5) implies that both terms in the numerator are zero, and hence $D_\theta^2 D_g L_{\mathcal{D}}(\bar{\theta}, g_0)[\widehat{g}_n - g_0, \widehat{\theta}_n - \theta_n^\star, \widehat{\theta}_n - \theta_n^\star] = 0$. We conclude that $D_g L_{\mathcal{D}}(\widehat{\theta}_n, g_0)[\widehat{g}_n - g_0] = D_g L_{\mathcal{D}}(\theta_n^\star, g_0)[\widehat{g}_n - g_0]$. Using this identity in the excess risk upper bound, we arrive at

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0)$$
$$\leq \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta/2 \,; \widehat{g}_n) + \beta \cdot \|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^2$$
$$\leq \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta/2 \,; \widehat{g}_n) + \beta \cdot \left(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta/2) \right)^2.$$

$\square$

C Proofs from Section 4

Proof of Lemma 1.

Assumption 1. By the definition of directional derivatives, the law of iterated expectations and the fact that $\mathcal{X} \subseteq \mathcal{W}$:

$$
\begin{aligned}
D_g D_\theta L_{\mathcal{D}}(\theta_n^\star, g_0)[\theta - \theta_n^\star, g - g_0] &= \mathbb{E}\big[(\theta(x) - \theta_n^\star(x))^T \cdot \nabla_\gamma \nabla_\zeta \ell(\theta_n^\star(x), g_0(w), z)(g(w) - g_0(w)) \big] \\
&= \mathbb{E}\big[(\theta(x) - \theta_n^\star(x))^T \cdot \mathbb{E}\big[\nabla_\gamma \nabla_\zeta \ell(\theta_n^\star(x), g_0(w), z) \mid w \big](g(w) - g_0(w)) \big] \\
&= 0.
\end{aligned}
$$

Assumption 2. This likewise follows immediately by expanding the directional derivative and applying the law of iterated expectation:

$$D_\theta L_{\mathcal{D}}(\theta_n^\star, g_0)[\theta - \theta_n^\star] = \mathbb{E}\big[\nabla_\zeta \ell(\theta_n^\star(x), g_0(w), z) \cdot (\theta(x) - \theta_n^\star(x)) \big] \geq 0.$$

We now argue about the remaining assumptions. We will repeatedly invoke the following expression for the second derivative of the population risk.

$$D_\theta^2 L_\mathcal{D}(\bar{\theta}, g)[\theta - \theta', \theta - \theta'] = \mathbb{E}_z\Big[\phi'\big(\langle \Lambda(g(w), v), \bar{\theta}(x)\rangle\big) \cdot \langle \Lambda(g(w), v), \theta(x) - \theta'(x)\rangle^2\Big]. \tag{74}$$

Let us introduce some additional notation. Let $g \in \mathcal{G}_n$ and $\theta \in \Theta_n$ be the free variables in the statements of Assumption 3 and Assumption 4. We define the following vector- and matrix-valued random variables:

$$W_0 = \Lambda(g_0(w), v), \qquad\qquad W_n = \Lambda(g(w), v), \tag{75}$$
$$X_0 = \theta_n^\star(x), \qquad\qquad X_n = \theta(x), \tag{76}$$
$$V_0 = g_0(w), \qquad\qquad V_n = g(w), \tag{77}$$
$$A_0 = \mathbb{E}\big[W_0 W_0^\top \mid x\big]. \tag{78}$$

To prove the lemma, it suffices to verify Assumption 3 and Assumption 4 with respect to the norm on the Θ space defined via $\|\theta - \theta'\|_{\Theta_n}^2 = \mathbb{E}\Big[\langle W_0, \theta(x) - \theta'(x)\rangle^2\Big]$ and the norm in the $\mathcal{G}_n$ space defined via $\|g - g'\|_{\mathcal{G}_n} = \|g - g'\|_{L_4(\ell_2, D)}$. One useful observation used throughout the proof is that for all $\theta \in \widehat{\Theta}_n$,

$$\|\theta - \theta_n^\star\|_{\Theta_n}^2 = \mathbb{E}\big[(\theta(x) - \theta_n^\star(x))^T A_0 (\theta(x) - \theta_n^\star(x))\big] \geq \gamma \cdot \mathbb{E}\big[\|\theta(x) - \theta_n^\star(x)\|_2^2\big] = \gamma \cdot \|\theta - \theta_n^\star\|_{L_2(\ell_2, D)}^2.$$

Assumption 3 By Equation 15 and Equation 74:

$$\begin{aligned}
D_\theta^2 L_\mathcal{D}(\bar{\theta}, g)[\theta - \theta_n^\star, \theta - \theta_n^\star] &\geq \tau \cdot \mathbb{E}_z\big[\langle W_n, X_n - X_0\rangle^2\big] \\
&\geq \frac{\tau}{2} \cdot \mathbb{E}_z\big[\langle W_0, X_n - X_0\rangle^2\big] - \tau \cdot \mathbb{E}_z\big[\langle W_0 - W_n, X_n - X_0\rangle^2\big] \\
&\geq \frac{\tau}{2} \cdot \mathbb{E}_z\big[\langle W_0, X_n - X_0\rangle^2\big] - \tau \cdot \mathbb{E}_z\big[\|W_0 - W_n\|_2^2 \|X_n - X_0\|_2^2\big] \\
&\geq \frac{\tau}{2} \cdot \|\theta - \theta_n^\star\|_{\Theta_n}^2 - \tau L_\Lambda^2 \, \mathbb{E}_z\big[\|V_n - V_0\|_2^2 \|X_n - X_0\|_2^2\big].
\end{aligned}$$

Using the AM-GM inequality we have for any $\eta > 0$:

$$\begin{aligned}
&\geq \frac{\tau}{2} \cdot \|\theta - \theta_n^\star\|_{\Theta_n}^2 - \frac{\tau L_\Lambda^4}{2\eta} \mathbb{E}_z\big[\|V_n - V_0\|_2^4\big] - \frac{\eta t}{2} \mathbb{E}\big[\|X_n - X_0\|_2^4\big] \\
&= \frac{\tau}{2} \cdot \|\theta - \theta_n^\star\|_{\Theta_n}^2 - \frac{\tau L_\Lambda^4}{2\eta} \|g - g_0\|_{\mathcal{G}_n}^4 - \frac{4\eta\tau R_{\Theta_n}^2}{2} \mathbb{E}\big[\|X_n - X_0\|_2^2\big] \\
&\geq \frac{\tau}{2} \cdot \|\theta - \theta_n^\star\|_{\Theta_n}^2 - \frac{\tau L_\Lambda^4}{2\eta} \|g - g_0\|_{\mathcal{G}_n}^4 - \frac{4\eta\tau R_{\Theta_n}^2}{2\gamma} \|\theta - \theta_n^\star\|_{\Theta_n}^2.
\end{aligned}$$

Choosing $\eta = \frac{\gamma}{8R_{\Theta_n}^2}$, yields the inequality:

$$D_\theta^2 L_\mathcal{D}(\bar{\theta}, g)[\theta - \theta_n^\star, \theta - \theta_n^\star] \geq \frac{\tau}{4} \|\theta - \theta_n^\star\|_{\Theta_n}^2 - \frac{4\tau L_\Lambda^4 R_{\Theta_n}^2}{\gamma} \|g - g_0\|_{\mathcal{G}_n}^4.$$

Thus, Assumption 3 is satisfied with $\lambda = \frac{\tau}{4}$ and $\kappa = \frac{4\tau L_\Lambda^4 R_{\Theta_n}^2}{\gamma}$.

47

Assumption 4 (a). Using Equation 15 and Equation 74 we have:

$$D_\theta^2 L_\mathcal{D}(\bar{\theta}, g_0)[\theta - \theta_n^\star, \theta - \theta_n^\star] \leq T \cdot \mathbb{E}_z\big[\langle W_0, X_n - X_0 \rangle^2\big] = T \cdot \|\theta - \theta_n^\star\|_{\Theta_n}^2$$

It follows that Assumption 4 (a) is satisfied with $\beta_1 = 2T$.

Assumption 4 (b). For simplicity of notation, define the random vectors $X_0 = \theta_n^\star(x)$, $X_n = \theta(x)$, $V_0 = g_0(w)$, $V_n = g(w)$ and $\Sigma(w) = \mathbb{E}\big[\nabla_{\gamma\gamma}^2 \nabla_{\zeta_i} \ell(\theta_n^\star(x), \bar{g}(w), z) \mid w\big]$. Then invoking the assumed structure on the loss function:

$$\big|D_g^2 D_\theta L_\mathcal{D}(\theta_n^\star, \bar{g})[\theta - \theta_n^\star, g - g_0, g - g_0]\big| = \left|\sum_{i=1}^{K^{(2)}} \mathbb{E}\big[(X_{ni} - X_{0i})(V_n - V_0)^\top \nabla_{\gamma\gamma}^2 \nabla_{\zeta_i} \ell(\theta_n^\star(x), \bar{g}(w), z)(V_n - V_0)\big]\right|.$$

Using that $x \subseteq w$:

$$= \left|\sum_{i=1}^{K^{(2)}} \mathbb{E}\big[(X_{ni} - X_{0i})(V_n - V_0)^\top \Sigma(w)(V_n - V_0)\big]\right|$$

$$\leq \sum_{i=1}^{K^{(2)}} \mathbb{E}\big[|X_{ni} - X_{0i}| \cdot \big|(V_n - V_0)^\top \Sigma(w)(V_n - V_0)\big|\big]$$

$$\leq \mu \sum_{i=1}^{K^{(2)}} \mathbb{E}\big[|X_{ni} - X_{0i}| \cdot \|V_n - V_0\|_2^2\big]$$

$$= \mu \, \mathbb{E}\big[\|X_n - X_0\|_1 \cdot \|V_n - V_0\|_2^2\big].$$

All that remains is to relate these norms to the norms appearing in the lemma definition.

$$\leq \mu\sqrt{K^{(2)}} \, \mathbb{E}\big[\|X_n - X_0\|_2 \cdot \|V_n - V_0\|_2^2\big]$$

$$\leq \mu\sqrt{K^{(2)}} \cdot \sqrt{\mathbb{E}\big[\|X_n - X_0\|_2^2\big]} \cdot \sqrt{\mathbb{E}\big[\|V_n - V_0\|_2^4\big]}$$

$$= \mu\sqrt{K^{(2)}} \cdot \|\theta - \theta_n^\star\|_{L_2(\ell_2, D)} \cdot \|g - g_0\|_{L_4(\ell_2, D)}^2$$

$$\leq \frac{\mu\sqrt{K^{(2)}}C_{2\to4}^2}{\sqrt{\gamma}} \|\theta - \theta_n^\star\|_{\Theta_n} \cdot \|g - g_0\|_{\mathcal{G}_n}^2.$$

Thus, we have

$$D_\theta^2 L_\mathcal{D}(\bar{\theta}, g_0)[\theta - \theta_n^\star, \theta - \theta_n^\star] \leq \frac{\mu\sqrt{K^{(2)}}C_{2\to4}^2}{\sqrt{\gamma}} \|\theta - \theta_n^\star\|_{\Theta_n} \cdot \|g - g_0\|_{\mathcal{G}_n}^2, \tag{79}$$

and so, Assumption 4 (b) is satisfied with $\beta_2 = \frac{\mu\sqrt{K^{(2)}}C_{2\to4}^2}{\sqrt{\gamma}}$.

$\square$

Proof of Lemma 2. Immediate. $\square$

The following lemma is used in certain subsequent results.

Lemma 3. *Suppose Assumption 7 holds. Then for any functions $\theta \in \widehat{\Theta}_n$, $\bar{\theta} \in \star(\widehat{\Theta}_n, \theta_n^\star)$, and $g \in \mathcal{G}_n$,*

$$D_\theta^2 L_\mathcal{D}(\bar{\theta}, g)[\theta - \theta_n^\star, \theta - \theta_n^\star] \leq 3T \|\theta - \theta_n^\star\|_{\Theta_n}^2 + \frac{TL_\Lambda^4 R_{\Theta_n}^2}{\gamma} \|g - g_0\|_{\mathcal{G}_n}^4. \tag{80}$$

Proof. We adopt the same notation as in the proof of Lemma 1. This proof is almost the same as the proof that Assumption 3 is satisfied in that lemma.

$$
\begin{aligned}
D_\theta^2 L_\mathcal{D}(\bar{\theta}, g)[\theta - \theta_n^\star, \theta - \theta_n^\star] \leq{} & T \cdot \mathbb{E}_z\big[\langle W_n, X_n - X_0 \rangle^2\big] \\
\leq{} & 2T \cdot \mathbb{E}_z\big[\langle W_0, X_n - X_0 \rangle^2\big] + 2\, T \cdot \mathbb{E}_z\big[\langle W_0 - W_n, X_n - X_0 \rangle^2\big] \\
\leq{} & 2T \cdot \mathbb{E}_z\big[\langle W_0, X_n - X_0 \rangle^2\big] + 2\, T \cdot \mathbb{E}_z\big[\|W_0 - W_n\|_2^2 \|X_n - X_0\|_2^2\big] \\
\leq{} & 2T \cdot \|\theta - \theta_n^\star\|_{\Theta_n}^2 + 2\, T\, L_\Lambda^2\, \mathbb{E}_z\big[\|V_n - V_0\|_2^2 \|X_n - X_0\|_2^2\big].
\end{aligned}
$$

Using the AM-GM inequality we have for any $\eta > 0$:

$$
\begin{aligned}
\leq{} & 2T \cdot \|\theta - \theta_n^\star\|_{\Theta_n}^2 + \frac{T\, L_\Lambda^4}{\eta}\, \mathbb{E}_z\big[\|V_n - V_0\|_2^4\big] + \eta t\, \mathbb{E}\big[\|X_n - X_0\|_2^4\big] \\
={} & 2T \cdot \|\theta - \theta_n^\star\|_{\Theta_n}^2 + \frac{T L_\Lambda^4}{\eta}\, \|g - g_0\|_{\mathcal{G}_n}^4 + \eta T R_{\Theta_n}^2\, \mathbb{E}\big[\|X_n - X_0\|_2^2\big] \\
\leq{} & 2T \cdot \|\theta - \theta_n^\star\|_{\Theta_n}^2 + \frac{T L_\Lambda^4}{\eta}\, \|g - g_0\|_{\mathcal{G}_n}^4 + \frac{\eta T R_{\Theta_n}^2}{\gamma}\, \|\theta - \theta_n^\star\|_{\Theta_n}^2.
\end{aligned}
$$

Choosing $\eta = \frac{\gamma}{R_{\Theta_n}^2}$, yields the inequality:

$$
D_\theta^2 L_\mathcal{D}(\bar{\theta}, g)[\theta - \theta_n^\star, \theta - \theta_n^\star] \leq\ 3T\, \|\theta - \theta_n^\star\|_{\Theta_n}^2 + \frac{T L_\Lambda^4 R_{\Theta_n}^2}{\gamma}\, \|g - g_0\|_{\mathcal{G}_n}^4.
$$

$\square$

D Preliminary Theorems for Constrained M-Estimators

Consider the case of M-estimation over a function class $\mathcal{F} : \mathcal{X} \to \mathbb{R}^d$ with loss $\ell : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$, i.e:

$$
L_\mathcal{D}(f) = \mathbb{E}[\ell(f(x), z)] \tag{81}
$$

Let $\mathcal{L}_f$ denote the random variable $\ell(f(x), z)$ and let:

$$
\mathbb{P}\mathcal{L}_f = \mathbb{E}[\ell(f(x), z)] \tag{82}
$$

$$
\mathbb{P}_n \mathcal{L}_f = \frac{1}{n} \sum_{i=1}^n \ell(f(x_i), z_i) \tag{83}
$$

denote the population risk and empirical risk over n samples. Then consider the constrained ERM algorithm:

$$
\hat{f} = \arg\min_{f \in \mathcal{F}} \mathbb{P}_n\, \mathcal{L}_f \tag{84}
$$

Throughout this section we will use the following convention on norms: for a vector valued function f, we will denote with $\|f\|_{p,q} = (\mathbb{E}[\|f(z)\|_q^p])^{1/p}$. If f is real valued, then we will omit q. For any $f^* \in \mathcal{F}$ let $\mathcal{F} - f^* = \{f - f^* : f \in \mathcal{F}\}$ and:

$$
\mathrm{star}(\mathcal{F} - f^*) = \{r\,(f - f^*) : f \in \mathcal{F}, r \in [0, 1]\} \tag{85}
$$

Finally, let:

$$
\mathcal{F}_t = \{f_t : (f_1, \ldots, f_t, \ldots, f_d) \in \mathcal{F}\} \tag{86}
$$

denote the marginal real-valued function class that corresponds to coordinate t of the functions in class $\mathcal{F}$.

For any real-valued function class $\mathcal{G}$, define the localized Rademacher complexity:

$$\mathcal{R}_n(\delta;\mathcal{G}) = \mathbb{E}_{\epsilon,z}\left[\sup_{g\in\mathcal{G}:\|g\|_2\leq\delta}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\,g(z_i)\right|\right] \tag{87}$$

The following lemmas provide a vector-valued extension of the development in Wainwright (2019). The key idea behind the extension is to invoke a vector-valued contraction theorem due to Maurer (2016). For completeness we give a include proofs for these lemmas, even though some parts are straightforward adaptations of lemmas in Wainwright (2019).

Lemma 4. *Consider a function class $\mathcal{F}: \mathcal{X} \to \mathbb{R}^d$, with $\sup_{f\in\mathcal{F}}\|f\|_{\infty,2} \leq 1$ and pick any $f^* \in \mathcal{F}$. Assume that the loss ℓ is L-Lipschitz in its first argument with respect to the ℓ_2 norm and let:*

$$Z_n(r) = \sup_{f\in\mathcal{F}:\|f-f^*\|_{2,2}\leq r}\left|\mathbb{P}_n(\mathcal{L}_f - \mathcal{L}_{f^*}) - \mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*})\right| \tag{88}$$

Then for some universal constants c_1, c_2:

$$\Pr[Z_n(r) \geq 16\,L\sum_{t=1}^d \mathcal{R}(r,\mathcal{F}_t - f_t^*) + u] \leq c_1 \exp\left\{-\frac{c_2 n u^2}{L^2 r^2 + 2Lu}\right\} \tag{89}$$

Moreover, if δ_n is any solution to the inequalities:

$$\forall t \in \{1,\ldots,d\} : \mathcal{R}(\delta; \mathrm{star}(\mathcal{F}_t - f_t^*)) \leq \delta^2 \tag{90}$$

then for each $r \geq \delta_n$:

$$\Pr[Z_n(r) \geq 16\,L\,d\,r\,\delta_n + u] \leq c_1 \exp\left\{-\frac{c_2 n u^2}{L^2 r^2 + 2Lu}\right\} \tag{91}$$

Lemma 5. *Consider a vector valued function class $\mathcal{F}: \mathcal{X} \to \mathbb{R}^d$, with $\sup_{f\in\mathcal{F}}\|f\|_{\infty,2} \leq 1$ and pick any $f^* \in \mathcal{F}$. Let $\delta_n^2 \geq \frac{4d\log(41\log(2c_2 n))}{c_2 n}$ be any solution to the inequalities:*

$$\forall t \in \{1,\ldots,d\} : \mathcal{R}(\delta; \mathrm{star}(\mathcal{F}_t - f_t^*)) \leq \delta^2 \tag{92}$$

Moreover, assume that the loss ℓ is L-Lipschitz in its first argument with respect to the ℓ_2 norm. Consider the following event:

$$\mathcal{E}_1 = \left\{\exists f \in \mathcal{F} : \|f - f^*\|_{2,2} \geq \delta_n \text{ and } \left|\mathbb{P}_n(\mathcal{L}_f - \mathcal{L}_{f^*}) - \mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*})\right| \geq 18L\,d\,\delta_n\,\|f - f^*\|_{2,2}\right\} \tag{93}$$

Then for some universal constants c_3, c_4: $\Pr[\mathcal{E}_1] \leq c_3 \exp\{-c_4 n\delta_n^2\}$.

Lemma 6. *Consider a vector valued function class $\mathcal{F}: \mathcal{X} \to \mathbb{R}^d$, with $\sup_{f\in\mathcal{F}}\|f\|_{\infty,2} \leq 1$ and pick any $f^* \in \mathcal{F}$. Let $\delta_n \geq 0$ be any solution to the inequalities:*

$$\forall t \in \{1,\ldots,d\} : \mathcal{R}(\delta; \mathrm{star}(\mathcal{F}_t - f_t^*)) \leq \delta^2 \tag{94}$$

Moreover, assume that the loss ℓ is L-Lipschitz in its first argument with respect to the ℓ_2 norm and also linear, i.e. $\mathcal{L}_{f+f'} = \mathcal{L}_f + \mathcal{L}_{f'}$ and $\mathcal{L}_{\alpha f} = \alpha\mathcal{L}_f$. Consider the following event:

$$\mathcal{E}_1 = \left\{\exists f \in \mathcal{F} : \|f - f^*\|_{2,2} \geq \delta_n \text{ and } \left|\mathbb{P}_n(\mathcal{L}_f - \mathcal{L}_{f^*}) - \mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*})\right| \geq 17L\,d\,\delta_n\,\|f - f^*\|_{2,2}\right\} \tag{95}$$

Then for some universal constants c_3, c_4: $\Pr[\mathcal{E}_1] \leq c_3 \exp\{-c_4 n\delta_n^2\}$.

Lemma 7. *Consider a function class $\mathcal{F}$, with $\sup_{f\in\mathcal{F}}\|f\|_\infty \leq 1$, and pick any $f^* \in \mathcal{F}$. Let $\delta_n^2 \geq \frac{4d\log(41\log(2c_2 n))}{c_2 n}$ be any solution to the inequalities:*

$$\forall t \in \{1,\ldots,d\} : \mathcal{R}(\delta; \mathrm{star}(\mathcal{F}_t - f_t^*)) \leq \delta^2 \tag{96}$$

Moreover, assume that the loss ℓ is L-Lipschitz in its first argument with respect to the ℓ_2 norm. Then for some universal constants c_5, c_6, with probability $1 - c_5\exp\{c_6 n\delta_n^2\}$:

$$\left|\mathbb{P}_n(\mathcal{L}_f - \mathcal{L}_{f^*}) - \mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*})\right| \leq 18 L\,d\,\delta_n\{\|f - f^*\|_{2,2} + \delta_n\} \qquad \forall f \in \mathcal{F} \tag{97}$$

Hence, the outcome $\hat{f}$ of constrained ERM satisfies that with the same probability:

$$\mathbb{P}(\mathcal{L}_{\hat{f}} - \mathcal{L}_{f^*}) \leq 18 L\,d\,\delta_n\{\|\hat{f} - f^*\|_{2,2} + \delta_n\} \tag{98}$$

If the loss $\mathcal{L}_f$ is also linear in f, i.e. $\mathcal{L}_{f+f'} = \mathcal{L}_f + \mathcal{L}_{f'}$ and $\mathcal{L}_{\alpha f} = \alpha\mathcal{L}_f$, then the lower bound on δ_n^2 is not required.

D.1   Proofs of Lemmas for Constrained M-Estimators

Proof of Lemma 4. By the Lipschitz condition on the loss and the boundedness of the functions, we have $\|\mathcal{L}_f - \mathcal{L}_{f^*}\|_\infty \leq L\|f - f^*\|_{\infty,2} \leq 2L$. Moreover:

$$\mathrm{Var}(\mathcal{L}_f - \mathcal{L}_{f^*}) \leq \mathcal{P}(\mathcal{L}_f - \mathcal{L}_{f^*})^2 \leq L^2\|f - f^*\|_{2,2}^2 \leq L^2 r^2$$

Thus by Talagrand's concentration inequality (see Theorem 3.8 of Wainwright (2019) and follow-up discussion) we have:

$$\Pr[Z_n(r) \geq 2\,\mathbb{E}[Z_n(r)] + u] \leq c_1\exp\left\{-c_2\frac{nu^2}{L^2 r^2 + 2Lu}\right\}$$

Moreover, by a standard symmetrization argument:

$$
\begin{aligned}
\mathbb{E}[Z_n(r)] &\leq 2\,\mathbb{E}\left[\sup_{\|f-f^*\|_{2,2}\leq r}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\{\ell(f(x_i), z_i) - \ell(f^*(x_i), z_i)\}\right|\right] \\
&= 2\,\mathbb{E}\left[\sup_{\|f-f^*\|_{2,2}\leq r,\,\delta\in\{-1,1\}}\delta\frac{1}{n}\sum_{i=1}^n \epsilon_i\{\ell(f(x_i), z_i) - \ell(f^*(x_i))\}\right] \\
&\leq 2\sum_{\delta\in\{-1,1\}}\mathbb{E}\left[\sup_{\|f-f^*\|_{2,2}\leq r}\delta\frac{1}{n}\sum_{i=1}^n \epsilon_i\{\ell(f(x_i), z_i) - \ell(f^*(x_i))\}\right] \\
&\leq 4\,\mathbb{E}\left[\sup_{\|f-f^*\|_{2,2}\leq r}\frac{1}{n}\sum_{i=1}^n \epsilon_i\{\ell(f(x_i), z_i) - \ell(f^*(x_i))\}\right]
\end{aligned}
$$

where the second inequality follows from the fact that each summand is non-negative, since we can always choose $f = f^*$. By invoking the multi-variate contraction inequality of Maurer (2016):

$$
\begin{aligned}
&\leq 8L\,\mathbb{E}\left[\sup_{\|f-f^*\|_{2,2}\leq r}\frac{1}{n}\sum_{i=1}^n\sum_{t=1}^d \epsilon_{i,t}\left(f_t(x_i) - f_t^*(x_i)\right)\right] \\
&\leq 8L\sum_{t=1}^d\mathbb{E}\left[\sup_{\|f_t-f_t^*\|_2\leq r}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_{i,t}\left(f_t(x_i) - f_t^*(x_i)\right)\right|\right] \\
&= 8L\sum_{t=1}^d\mathcal{R}(r, \mathcal{F}_t - f_t^*)
\end{aligned}
$$

51

where we also used the fact that for any fixed f^*, $\mathbb{E}[\epsilon_i \ell(f^*(x_i))] = \mathbb{E}[\epsilon_{i,t} f_t^*(x_i)] = 0$. This completes the proof of the first part of the lemma. For the second part, observe that: $\mathcal{R}(r, \mathcal{F}_t - f_t^*) \leq \mathcal{R}(r, \text{star}(\mathcal{F}_t - f_t^*))$. Moreover, for any star shaped function class $\mathcal{G}$, the function $r \to \frac{\mathcal{R}(r,\mathcal{G})}{r}$ is monotone non-decreasing. Thus for any $r \geq \delta_n$:

$$\frac{\mathcal{R}(r, \text{star}(\mathcal{F}_t - f_t^*))}{r} \leq \frac{\mathcal{R}(\delta_n, \text{star}(\mathcal{F}_t - f_t^*))}{\delta_n} \leq \delta_n$$

Re-arranging, yields that: $\mathcal{R}(r, \text{star}(\mathcal{F}_t - f_t^*)) \leq r \delta_n$. Hence, $\mathbb{E}[Z_n(r)] \leq 8 L d r \delta_n$. This completes the proof of the second part of the lemma. $\qquad\square$

Proof of Lemma 5. We invoke a peeling argument. Consider the events:

$$\mathcal{S}_m = \{f \in \mathcal{F} : \alpha^{m-1}\delta_n \leq \|f - f^*\|_{2,2} \leq \alpha^m \delta_n\}$$

for $\alpha = 18/17$. Since $\sup_{f \in \mathcal{F}}\|f - f^*\|_{2,2} \leq 2\sup_{f \in \mathcal{F}}\|f\|_{\infty,2} \leq 2$, it must be that any $f \in \mathcal{F}$ with $\|f - f^*\|_{2,2} \geq \delta_n$ belongs to some $\mathcal{S}_m$ for $m \in \{1, 2, \ldots, M\}$, where $M \leq \frac{\log(2/\delta_n)}{\log_\alpha(e)} \leq 41 \log(2/\delta_n)$. Thus by a union bound we have:

$$\Pr[\mathcal{E}_1] \leq \sum_{m=1}^{M} \Pr[\mathcal{E}_1 \cap \mathcal{S}_m]$$

Moreover, observe that if the event $\mathcal{E}_1 \cap \mathcal{S}_m$ occurs then there exists a $f \in \mathcal{F}$ with $\|f - f^*\|_{2,2} \leq \alpha^m \delta_n = r_m$, such that:

$$\left|\mathbb{P}_n(\mathcal{L}_f - \mathcal{L}_{f^*}) - \mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*})\right| \geq 18L d \delta_n \|f - f^*\|_2 \geq 18L d \delta_n \alpha^{m-1}\delta_n = 17L d \delta_n \alpha^m \delta_n = 17L d \delta_n r_m$$

Thus by the definition of $Z_n(r)$ we have:

$$\Pr[\mathcal{E}_1 \cap \mathcal{S}_m] \leq \Pr[Z_n(r_m) \geq 17Ld\delta_n r_m]$$

Applying Lemma 4 with $r = r_m$ and $u = L d r_m \delta_n$, yields that the latter probability is at most $c_1 \exp\{-c_2 n \frac{L^2 r_m^2 \delta_n^2}{L^2 r_m^2 + L^2 d r_m \delta_n}\} \leq c_1 \exp\{-c_2 n \frac{\delta_n^2}{2d}\}$, where we used the fact that $\delta_n \leq r_m$ in the last inequality. Subsequently, taking a union bound over the M events, we have:

$$\Pr[\mathcal{E}_1] \leq c_1 M \exp\left\{-c_2 n \frac{\delta_n^2}{2}\right\} = c_1 \exp\left\{-c_2 n \frac{\delta_n^2}{2d} + \log(M)\right\} \tag{99}$$

Since, by assumption on the lower bound on δ_n: $\log(M) \leq \log(41 \log(2/\delta_n)) \leq \log(41 \log(2c_2 n)) \leq \frac{c_2 n \delta_n^2}{4d}$, we get:

$$\Pr[\mathcal{E}_1] \leq c_1 \exp\left\{-c_2 n \frac{\delta_n^2}{4d}\right\} \tag{100}$$

$\qquad\square$

Proof of Lemma 6. For simplicity, let $\|\cdot\| = \|\cdot\|_{2,2}$. Suppose that there exists a $f \in \mathcal{F}$, with $\|f - f^*\| \geq \delta_n$, such that:

$$\left|\mathbb{P}_n\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right) - \mathbb{P}\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right)\right| \geq 17L d \delta_n \|f - f^*\|.$$

Then we will show that there exists a $f' \in \text{star}(\mathcal{F} - f^*)$, with $\|f' - f^*\| = \delta_n$, such that:

$$\left|\mathbb{P}_n\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right) - \mathbb{P}\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right)\right| \geq 17L d \delta_n^2.$$

Simply set f' to satisfy:

$$f' - f^* = \frac{\delta_n}{\|f - f^*\|} (f - f^*)$$

Since $\frac{\delta_n}{\|f-f^*\|} \leq 1$ and by the definition of the star hull, we know that $f' \in \text{star}(\mathcal{F} - f^*)$. Moreover, by definition of θ', we also have that $\|f' - f^*\|_n = \delta_n$. Moreover, by the linearity of the loss $\mathcal{L}_f^*$ with respect to f, we have:

$$
\begin{aligned}
\left| \mathbb{P}_n \left(\mathcal{L}_{f'} - \mathcal{L}_{f^*} \right) - \mathbb{P} \left(\mathcal{L}_{f'} - \mathcal{L}_{f^*} \right) \right| &= \left| \mathbb{P}_n \mathcal{L}_{f'-f^*} - \mathbb{P}\mathcal{L}_{f'-f^*} \right| \\
&= \frac{\delta_n}{\|f - f^*\|} \left| \mathbb{P}_n \mathcal{L}_{f-f^*} - \mathbb{P}\mathcal{L}_{f-f^*} \right| \\
&= \frac{\delta_n}{\|f - f^*\|} \left| \mathbb{P}_n \left(\mathcal{L}_f - \mathcal{L}_{f^*} \right) - \mathbb{P} \left(\mathcal{L}_f - \mathcal{L}_{f^*} \right) \right| \\
&\geq \frac{\delta_n}{\|f - f^*\|} 17 L \, d \, \delta_n \|f - f^*\| = 17 L \, d \, \delta_n^2.
\end{aligned}
$$

Thus we have that the probability of event $\mathcal{E}_1$ is upper bounded by the probability of the event:

$$\mathcal{E}_1' = \left\{ \sup_{f \in \text{star}(\mathcal{F}-f^*):\|f-f^*\|\leq\delta_n} \left| \mathbb{P}_n \left(\mathcal{L}_f - \mathcal{L}_{f^*} \right) - \mathbb{P} \left(\mathcal{L}_f - \mathcal{L}_{f^*} \right) \right| \geq 17 L \, d \, \delta_n^2 \right\} = \left\{ Z_n(\delta_n) \geq 17 L \, d \, \delta_n^2 \right\}$$

Invoking Lemma 4, with $r = \delta_n$ and $u = L \, d \, \delta_n^2$, we get that the probability of the second event is also at most $\mu_1' \exp\{-\mu_2' n \delta_n^2\}$, for some universal constants μ_1', μ_2'. □

Proof of Lemma 7. Consider the events:

$$
\begin{aligned}
\mathcal{E}_0 &= \left\{ Z_n(\delta_n) \geq 17 L \, d \, \delta_n^2 \right\} \\
\mathcal{E}_1 &= \left\{ \exists f \in \mathcal{F} : \|f - f^*\|_{2,2} \geq \delta_n \text{ and } \left| \mathbb{P}_n(\mathcal{L}_f - \mathcal{L}_{f^*}) - \mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*}) \right| \geq 18 L \, d \, \delta_n \|f - f^*\|_{2,2} \right\}
\end{aligned}
$$

with $Z_n(r)$ as defined in Lemma 4. Observe that if Equation 97 is violated, then one of these events must occur. Applying Lemma 4 with $r = \delta_n$ and $u = L \, d \, \delta_n^2$ yields, that event $\mathcal{E}_0$ happens with probability at most $c_1 \exp\{c_2' n \delta_n^2\}$, where $c_2' = c_2/(L^2 + L \, d)$. Moreover, applying Lemma 5 we get that $\Pr[\mathcal{E}_1] \leq c_3 \exp\{-c_4 n \delta_n^2\}$. Thus by a union bound with probability $1 - c_5 \exp\{c_6 n \delta_n^2\}$, neither events occur. If the loss $\mathcal{L}_f$ is linear then we apply Lemma 6 instead of Lemma 5, which does not require a lower bound on δ_n^2. □

E Proofs from Section 5

For notational convenience, we will use throughout the proofs of this section an empirical process theory notation. In particular, let $\mathcal{L}_{\theta,g}$ denote the random variable $\ell(\theta(x), g(w), z)$ and let:

$$\mathbb{P}\mathcal{L}_{\theta,g} = \mathbb{E}[\ell(\theta(x), g(w), z)] \tag{101}$$

$$\mathbb{P}_n\mathcal{L}_{\theta,g} = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta(x_i), g(w_i), z_i) \tag{102}$$

denote, correspondingly, the population risk and empirical risk over n samples. Then consider the constrained the two-stage plugin ERM algorithm:

$$\widehat{\theta}_n = \arg \min_{\theta \in \Theta_n} \mathbb{P}_n \mathcal{L}_{\theta,\widehat{g}_n} \tag{103}$$

E.1 Proof of Theorem 3

We split the theorem into two parts. In the first part we prove the result that requires a lower bound on the cricial radius δ_n^2 of order $O(\log\log(n)/n)$. In the second part we prove the improved result that does not require the lower bound, albeit under the extra assumption that the loss ℓ is Lipschitz and twice differentiable. We prove both results for the case when $R = 1$. The general R case follows easily by a standard re-scaling of the loss argument.

Lemma 8 (Fast Rates for Constrained ERM). *Consider a function class $\Theta_n : \mathcal{X} \to \mathbb{R}^d$, with $\sup_{\theta \in \Theta_n} \|\theta\|_{\infty,2} \le 1$. Let $\delta_n^2 \ge \frac{4\,d\,\log(41\log(2c_2 n))}{c_2 n}$ be any solution to the inequality:*

$$\forall t \in \{1,\ldots,d\} : \mathcal{R}(\delta; \mathrm{star}(\Theta_{n,t} - \theta_{n,t}^\star)) \le \delta^2. \tag{104}$$

Suppose $\ell(\cdot, \widehat{g}_n(w), z)$ is L-Lipschitz in its first argument with respect to the ℓ_2 norm and that the population risk L_D satisfies Assumption 1, Assumption 2, Assumption 3 and Assumption 4 with respect to the $\|\cdot\|_{2,2}$ norm. Let $\widehat{\theta}_n$ be the outcome of the constrained ERM algorithm. Then with probability $1 - c_7 \exp\{-c_8 n \delta_n^2\}$:

$$\mathbb{P}(\mathcal{L}_{\widehat{\theta}_n,\widehat{g}_n} - \mathcal{L}_{\theta_n^\star,\widehat{g}_n}) \le O(\delta_n^2 + \|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4) \tag{105}$$

Proof of Lemma 8. Since $\widehat{\theta}_n$ is the outcome of the constrained ERM and since $\theta_n^\star \in \Theta_n$, we have:

$$\mathbb{P}_n\left(\mathcal{L}_{\widehat{\theta}_n,\widehat{g}_n} - \mathcal{L}_{\theta_n^\star,\widehat{g}_n}\right) \le 0$$

Applying Lemma 5, with $\mathcal{F} = \Theta_n$, $f^\star = \theta_n^\star$ and $\mathcal{L}_\cdot = \mathcal{L}_{\cdot,\widehat{g}_n}$, we know that the probability of the event:

$$\mathcal{E}_1 = \left\{ \exists \theta \in \Theta_n : \|\theta - \theta_n^\star\|_{2,2} \ge \delta_n \text{ and } \left|\mathbb{P}_n(\mathcal{L}_{\theta,\widehat{g}_n} - \mathcal{L}_{\theta_n^\star,\widehat{g}_n}) - \mathbb{P}(\mathcal{L}_{\theta,\widehat{g}_n} - \mathcal{L}_{\theta_n^\star,\widehat{g}_n})\right| \ge 18 L\, d\, \delta_n \|\theta - \theta_n^\star\|_{2,2}\right\}$$

is at most $\zeta = c_3 \exp\{-c_4 n \delta_n^2\}$. Thus with probability $1 - \zeta$, either $\|\widehat{\theta}_n - \theta_n^\star\|_{2,2} \le \delta_n$ or

$$\left|\mathbb{P}_n(\mathcal{L}_{\widehat{\theta}_n,\widehat{g}_n} - \mathcal{L}_{\theta_n^\star,\widehat{g}_n}) - \mathbb{P}(\mathcal{L}_{\widehat{\theta}_n,\widehat{g}_n} - \mathcal{L}_{\theta_n^\star,\widehat{g}_n})\right| \le 18 L\, d\, \delta_n \|\theta - \theta_n^\star\|_{2,2}.$$

Invoking the first inequality, the latter also implies that:

$$\mathbb{P}(\mathcal{L}_{\widehat{\theta}_n,\widehat{g}_n} - \mathcal{L}_{\theta_n^\star,\widehat{g}_n}) \le 18 L\, d\, \delta_n \|\widehat{\theta}_n - \theta_n^\star\|_{2,2}$$

Invoking the strong convexity Assumption 3, we have:

$$\mathbb{P}(\mathcal{L}_{\widehat{\theta}_n,\widehat{g}_n} - \mathcal{L}_{\theta_n^\star,\widehat{g}_n}) \ge D_\theta L_D(\theta_n^\star,\widehat{g}_n)[\widehat{\theta}_n - \theta_n^\star] + \lambda\|\widehat{\theta}_n - \theta_n^\star\|_{2,2}^2 - \kappa\|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4$$

Invoking Assumption 1, Assumption 2 and Assumption 4 and following the same step of inequalities as in the proof of Theorem 1, we have that for any $\eta > 0$:

$$D_\theta L_D(\theta_n^\star,\widehat{g}_n)[\widehat{\theta}_n - \theta_n^\star] \ge D_\theta L_D(\theta_n^\star,g_0)[\widehat{\theta}_n - \theta_n^\star] - \frac{\beta_2}{4\eta}\|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4 - \frac{\beta_2\eta}{4}\|\theta - \theta_n^\star\|_{2,2}^2$$

$$\ge -\frac{\beta_2}{4\eta}\|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4 - \frac{\beta_2\eta}{4}\|\widehat{\theta}_n - \theta_n^\star\|_{2,2}^2$$

Choosing $\eta = \frac{2\lambda}{\beta_2}$ and combining the last two inequalities, yields:

$$\mathbb{P}(\mathcal{L}_{\widehat{\theta}_n,\widehat{g}_n} - \mathcal{L}_{\theta_n^\star,\widehat{g}_n}) \ge \frac{\lambda}{2}\|\widehat{\theta}_n - \theta_n^\star\|_{2,2}^2 - \left(\kappa + \frac{\beta_2^2}{8\lambda}\right)\|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4$$

Combining with the upper bound on the left hand side and using the AM-GM inequality, we have for any $\eta > 0$:

$$\frac{\lambda}{2}\|\widehat{\theta}_n - \theta_n^\star\|_{2,2}^2 - \left(\kappa + \frac{\beta_2^2}{8\lambda}\right)\|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4 \le 18L\,d\,\delta_n\|\widehat{\theta}_n - \theta_n^\star\|_{2,2} \le \frac{9L^2\,d^2}{\eta}\delta_n^2 + \frac{\eta}{2}\|\widehat{\theta}_n - \theta_n^\star\|_{2,2}^2$$

Choosing $\eta = \frac{\lambda}{2}$ and re-arranging, yields:

$$\|\widehat{\theta}_n - \theta_n^\star\|_{2,2}^2 \le \frac{80L^2d^2}{\lambda^2}\delta_n^2 + \frac{4}{\lambda}\left(\kappa + \frac{\beta_2^2}{8\lambda}\right)\|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4$$

Thus we conclude that in any case:

$$\|\widehat{\theta}_n - \theta_n^\star\|_{2,2} \le O\left(\delta_n + \|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^2\right)$$

Applying Lemma 7 and invoking Equation 97 together with the fact that $\mathbb{P}_n\left(\mathcal{L}_{\widehat{\theta}_n,\widehat{g}_n} - \mathcal{L}_{\theta_n^\star,\widehat{g}_n}\right) \le 0$, by the definition of constrained ERM, yields that:

$$\mathbb{P}(\mathcal{L}_{\widehat{\theta}_n,\widehat{g}_n} - \mathcal{L}_{\theta_n^\star,\widehat{g}_n}) \le O\left(\delta_n^2 + \delta_n\|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^2\right) \le O\left(\delta_n^2 + \|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4\right)$$

where we used the AM-GM inequality in the final inequality. $\qquad\square$

Lemma 9 (Improved Fast Rates for Constrained ERM). *Consider a function class $\Theta_n : \mathcal{X} \to \mathbb{R}^d$, with $\sup_{\theta\in\Theta_n}\|\theta\|_{\infty,2} \le 1$. Let $\delta_n \ge 0$ be any solution to the inequality:*

$$\forall t \in \{1,\ldots,d\} : \mathcal{R}(\delta; \text{star}(\Theta_{n,t} - \theta_{n,t}^*)) \le \delta^2 \tag{106}$$

Suppose $\ell(\cdot,\widehat{g}_n(w),z)$ is L-Lipschitz with respect to the ℓ_2 norm, convex and twice differentiable in its first argument, with second order partial derivatives all uniformly bounded by a constant. Moreover, the population risk L_D satisfies Assumption 1, Assumption 2, Assumption 3 and Assumption 4, with respect to the $\|\cdot\|_{2,2}$ norm and Assumption 4 (a) is satisfied at every $g \in \mathcal{G}_n$, rather than only at g_0. Let $\widehat{\theta}_n$ be the outcome of the constrained ERM algorithm. Then with probability $1 - c_7\exp\{-c_8 n\delta_n^2\}$:

$$\mathbb{P}(\mathcal{L}_{\widehat{\theta}_n,\widehat{g}_n} - \mathcal{L}_{\theta_n^\star,\widehat{g}_n}) \le O(\delta_n^2 + \|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4) \tag{107}$$

Proof of Lemma 9. For shorthand notation, let $\|\theta\| = \|\theta\|_{2,2} = \sqrt{\mathbb{E}\left[\|\theta(x)\|_2^2\right]}$ and $\|\theta\|_n = \|\theta\|_{n,2,2} = \sqrt{\frac{1}{n}\sum_{i=1}^n \|\theta(x_i)\|_2^2}$. Moreover, throughout the section we use the fact (see Lemma 14.1 of Wainwright (2019)) that for δ_n that satisfies the conditions in the theorem, with probability $\rho \ge 1 - \mu_1\exp\{-\mu_2 n\delta_n^2\}$ (for some universal constants μ_1, μ_2): for all $\theta \in \Theta_n$

$$\left|\|\theta - \theta_n^\star\|_n^2 - \|\theta - \theta_n^\star\|^2\right| \le \frac{1}{2}\|\theta - \theta_n^\star\|^2 + d\frac{\delta_n^2}{2} \tag{108}$$

Since $\widehat{\theta}_n$ is the outcome of the constrained ERM and since $\theta_n^\star \in \Theta_n$, we have:

$$\mathbb{P}_n\left(\mathcal{L}_{\widehat{\theta}_n,\widehat{g}_n} - \mathcal{L}_{\theta_n^\star,\widehat{g}_n}\right) \le 0$$

Moreover, we also show that the strong convexity Assumption 3 on the population risk, also implies up to $O(\delta_n^2)$ terms a strong convexity condition of the empirical risk:

Lemma 10. *Let $\delta_n \geq 0$ be any solution to the inequality:*

$$\forall t \in \{1, \ldots, d\} : \mathcal{R}(\delta; \mathrm{star}(\Theta_{n,t} - \theta^*_{n,t})) \leq \delta^2 \tag{109}$$

Suppose that the loss function ℓ is convex, twice differentiable in its first argument, with second order partial derivatives all uniformly bounded by a constant. Moreover, suppose that the population risk satisfies [Assumption 3](#). Then for some universal constants μ_3, μ_4, with probability $1 - \mu_3 \exp\{-\mu_4 n \delta_n^2\}$:

$$\mathbb{P}_n\left(\mathcal{L}_{\widehat{\theta}_n, \widehat{g}_n} - \mathcal{L}_{\theta^*_n, \widehat{g}_n}\right) \geq \mathbb{P}_n\langle \nabla_\zeta \ell(\theta^*_n(x), \widehat{g}_n(w), z), \widehat{\theta}_n(z) - \theta^*_n(z)\rangle + \frac{\lambda}{4}\|\widehat{\theta}_n - \theta^*_n\|^2 - \kappa\|\widehat{g}_n - g_0\|^4_{\mathcal{G}_n} - O(\delta_n^2)$$

We defer the proof of this lemma and continue with the remainder of the proof. Let $\mathcal{L}^*_{\theta,g} = \langle \nabla_\zeta \ell(\theta^*_n(x), g(w), z), \theta(z)\rangle$, denote the linearized loss around θ^*_n. Combining the last two inequalities we derive:

$$\frac{\lambda}{2}\|\widehat{\theta}_n - \theta^*_n\|^2 \leq -\mathbb{P}_n\left(\mathcal{L}^*_{\widehat{\theta}_n, \widehat{g}_n} - \mathcal{L}^*_{\theta^*_n, \widehat{g}_n}\right) + \kappa\|\widehat{g}_n - g_0\|^4_{\mathcal{G}_n} + O(\delta_n^2) \tag{110}$$

Consider the following events:

$$\mathcal{E}_1 = \left\{ \sup_{\|\theta - \theta^*_n\| \leq \delta_n} \left|\mathbb{P}_n\left(\mathcal{L}_{\theta, \widehat{g}_n} - \mathcal{L}_{\theta^*_n, \widehat{g}_n}\right) - \mathbb{P}\left(\mathcal{L}_{\theta, \widehat{g}_n} - \mathcal{L}_{\theta^*_n, \widehat{g}_n}\right)\right| > 17\, L\, d\, \delta_n^2\right\}$$

$$\mathcal{E}_2 = \left\{\exists \theta \in \Theta_n : \|\theta - \theta^*_n\| \geq \delta_n \text{ and } \left|\mathbb{P}_n\left(\mathcal{L}^*_{\theta, \widehat{g}_n} - \mathcal{L}^*_{\theta^*_n, \widehat{g}_n}\right) - \mathbb{P}\left(\mathcal{L}^*_{\theta, \widehat{g}_n} - \mathcal{L}^*_{\theta^*_n, \widehat{g}_n}\right)\right| \geq 17 L\, d\, \delta_n \|\theta - \theta^*_n\|\right\}$$

Invoking [Lemma 4](#), with $\mathcal{F} = \Theta_n$, $f^* = \theta^*_n$ and $\mathcal{L}. = \mathcal{L}_{., \widehat{g}_n}$, $r = \delta_n$ and $u = L\, d\, \delta_n^2$, we get that the probability of event $\mathcal{E}_1$ is at most $\mu'_1 \exp\{-\mu'_2 n \delta_n^2\}$, for some universal constants μ'_1, μ'_2. Noting that the loss $\mathcal{L}^*_{\theta, \widehat{g}_n}$ is linear in θ and invoking [Lemma 6](#), with $\mathcal{F} = \Theta_n$, $f^* = \theta^*_n$ and $\mathcal{L}. = \mathcal{L}^*_{., \widehat{g}_n}$, we get that the probability of $\mathcal{E}_2$ is also at most $\mu'_1 \exp\{-\mu'_2 n \delta_n^2\}$. Thus we conclude that by a union bound none of the events occur with probability at least $1 - \zeta$, for $\zeta = \mu''_1 \exp\{-\mu''_2 n \delta_n^2\}$, for some universal constants μ''_1, μ''_2. Thus with probability $1 - \zeta$, $\mathcal{E}_1$ does not hold and either $\|\widehat{\theta}_n - \theta^*_n\| \leq \delta_n$ or

$$\left|\mathbb{P}_n\left(\mathcal{L}^*_{\widehat{\theta}_n, \widehat{g}_n} - \mathcal{L}^*_{\theta^*_n, \widehat{g}_n}\right) - \mathbb{P}\left(\mathcal{L}^*_{\widehat{\theta}_n, \widehat{g}_n} - \mathcal{L}^*_{\theta^*_n, \widehat{g}_n}\right)\right| \leq 17 L\, d\, \delta_n \|\widehat{\theta}_n - \theta^*_n\|.$$

In the former case, together with the fact that $\mathcal{E}_1$ does not hold:

$$\left|\mathbb{P}_n\left(\mathcal{L}_{\widehat{\theta}_n, \widehat{g}_n} - \mathcal{L}_{\theta^*_n, \widehat{g}_n}\right) - \mathbb{P}\left(\mathcal{L}_{\widehat{\theta}_n, \widehat{g}_n} - \mathcal{L}_{\theta^*_n, \widehat{g}_n}\right)\right| \leq \sup_{\|\theta - \theta^*_n\| \leq \delta_n} \left|\mathbb{P}_n\left(\mathcal{L}_{\theta, \widehat{g}_n} - \mathcal{L}_{\theta^*_n, \widehat{g}_n}\right) - \mathbb{P}\left(\mathcal{L}_{\theta, \widehat{g}_n} - \mathcal{L}_{\theta^*_n, \widehat{g}_n}\right)\right| \leq 17\, L\, d\, \delta_n^2$$

Combining with the fact that $\mathbb{P}_n\left(\mathcal{L}_{\widehat{\theta}_n, \widehat{g}_n} - \mathcal{L}_{\theta^*_n, \widehat{g}_n}\right) \leq 0$, by the definition of ERM, yields:

$$\mathbb{P}\left(\mathcal{L}_{\widehat{\theta}_n, \widehat{g}_n} - \mathcal{L}_{\theta^*_n, \widehat{g}_n}\right) \leq 17\, L\, d\, \delta_n^2 = O\left(\delta_n^2 + \|\widehat{g}_n - g_0\|^4_{\mathcal{G}_n}\right)$$

as desired by the theorem.

In the latter case, we further derive, by invoking [Equation 110](#), that:

$$\frac{\lambda}{2}\|\widehat{\theta}_n - \theta^*_n\|^2 \leq 17 L\, d\, \delta_n\|\widehat{\theta}_n - \theta^*_n\| - \mathbb{P}\left(\mathcal{L}^*_{\widehat{\theta}_n, \widehat{g}_n} - \mathcal{L}^*_{\theta^*_n, \widehat{g}_n}\right) + \kappa\|\widehat{g}_n - g_0\|^4_{\mathcal{G}_n} + O(\delta_n^2)$$

$$= 17 L\, d\, \delta_n\|\widehat{\theta}_n - \theta^*_n\| - D_\theta L_D(\theta^*_n, \widehat{g}_n)[\widehat{\theta}_n - \theta^*_n] + \kappa\|\widehat{g}_n - g_0\|^4_{\mathcal{G}_n} + O(\delta_n^2)$$

Applying the AM-GM inequality, for any $\eta > 0$:

$$\frac{\lambda}{2}\|\widehat{\theta}_n - \theta^*_n\|^2 \leq \frac{17^2 L^2\, d^2}{2\eta}\delta_n^2 + \frac{\eta}{2}\|\widehat{\theta}_n - \theta^*_n\|^2 - D_\theta L_D(\theta^*_n, \widehat{g}_n)[\widehat{\theta}_n - \theta^*_n] + \kappa\|\widehat{g}_n - g_0\|^4_{\mathcal{G}_n} + O(\delta_n^2)$$

Choosing $\eta = \frac{\lambda}{2}$, re-arranging yields:

$$\|\widehat{\theta}_n - \theta_n^\star\|^2 \leq O(\delta_n^2) - \frac{4}{\lambda} D_\theta L_D(\theta_n^\star, \widehat{g}_n)[\widehat{\theta}_n - \theta_n^\star] + \frac{4\kappa}{\lambda}\|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4$$

Performing a second order Taylor expansion of the population risk around $\theta_n^\star$ and invoking second order smoothness of the population risk:

$$\mathbb{P}(\mathcal{L}_{\widehat{\theta}_n, \widehat{g}_n} - \mathcal{L}_{\theta_n^\star, \widehat{g}_n}) \leq D_\theta L_{\mathcal{D}}(\theta_n^\star, \widehat{g}_n)[\widehat{\theta}_n - \theta_n^\star] + \frac{\beta_1}{2} \cdot \|\widehat{\theta}_n - \theta_n^\star\|^2$$

Combining the latter two inequalities yields:

$$\mathbb{P}(\mathcal{L}_{\widehat{\theta}_n, \widehat{g}_n} - \mathcal{L}_{\theta_n^\star, \widehat{g}_n}) \leq O\left(\delta_n^2 + \|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4\right) - \left(\frac{4\beta_1}{\lambda} - 1\right) D_\theta L_D(\theta_n^\star, \widehat{g}_n)[\widehat{\theta}_n - \theta_n^\star] \tag{111}$$

Since $\beta_1/\lambda \geq 1$ (by standard properties of smoothness and strong convexity of a function) the coefficient $\left(\frac{4\beta_1}{\lambda} - 1\right)$ in front of the last term is non-negative. Invoking [Assumption 1], [Assumption 2] and [Assumption 4] and following the same step of inequalities as in the proof of [Theorem 1]/[Lemma 1], we have that for any $\eta > 0$:

$$
\begin{aligned}
-D_\theta L_D(\theta_n^\star, \widehat{g}_n)[\theta - \theta_n^\star] &\leq -D_\theta L_D(\theta_n^\star, g_0)[\widehat{\theta}_n - \theta_n^\star] + \frac{\beta_2}{4\eta}\|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4 + \frac{\beta_2\eta}{4}\|\theta - \theta_n^\star\|^2 \\
&\leq \frac{\beta_2}{4\eta}\|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4 + \frac{\beta_2\eta}{4}\|\theta - \theta_n^\star\|^2 \\
&\leq \frac{\beta_2}{4\eta}\|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4 + \frac{\beta_2\eta}{4}\left(O\left(\delta_n^2 + \|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4\right) - \frac{4}{\lambda} D_\theta L_D(\theta_n^\star, \widehat{g}_n)[\widehat{\theta}_n - \theta_n^\star]\right)
\end{aligned}
$$

Choosing $\eta = \frac{\lambda}{2\beta_2}$ and re-arranging:

$$-D_\theta L_D(\theta_n^\star, \widehat{g}_n)[\theta - \theta_n^\star] \leq O\left(\delta_n^2 + \|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4\right)$$

Combining with [Equation 111], yields:

$$\mathbb{P}(\mathcal{L}_{\widehat{\theta}_n, \widehat{g}_n} - \mathcal{L}_{\theta_n^\star, \widehat{g}_n}) \leq O\left(\delta_n^2 + \|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4\right) \tag{112}$$

as desired by the theorem. $\qquad\square$

Proof of [Lemma 10]. Consider a second order Taylor expansion of $\mathbb{P}_n\left(\mathcal{L}_{\widehat{\theta}_n, \widehat{g}_n} - \mathcal{L}_{\theta_n^\star, \widehat{g}_n}\right)$:

$$\mathbb{P}_n\left\langle \nabla_\zeta \ell(\theta_n^\star(x), \widehat{g}_n(w), z), \widehat{\theta}_n(z) - \theta_n^\star(z)\right\rangle + \frac{1}{2}\mathbb{P}_n\left(\widehat{\theta}_n(x) - \theta_n^\star(x)\right)^\top \nabla_{\zeta\zeta}\ell(\bar{\theta}(x), \widehat{g}_n(w), z)\left(\widehat{\theta}_n(x) - \theta_n^\star(x)\right)$$

for some $\bar{\theta}$ that is determined by $\widehat{\theta}_n$ by the mean value theorem. Since ℓ is a convex function with respect to its first argument, we know that the Hessian is positive semidefinite. Hence, we can decompose it as:

$$\nabla_{\zeta\zeta}\ell(\bar{\theta}(x), \widehat{g}_n(w), z) = \sum_{t=1}^d v_t(\bar{\theta}(x), \widehat{g}_n(w), z)\, v_t(\bar{\theta}(x), \widehat{g}_n(w), z)^\top$$

For simplicity of notation, since $\widehat{g}_n$ is fixed and the rest of the variables are random, we will denote with $V_{t,\widehat{\theta}_n}$ the random vector that corresponds to vector $v_t(\bar{\theta}(x), \widehat{g}_n(w), z)$, where we remind that $\bar{\theta}$ is determined by $\widehat{\theta}_n$. Thus we can write the second order part of the Taylor expansion as:

$$\sum_{t=1}^{d} \mathbb{P}_n \left(\widehat{\theta}_n(x) - \theta_n^{\star}(x) \right)^{\top} V_{t,\widehat{\theta}_n} V_{t,\widehat{\theta}_n}^{\top} \left(\widehat{\theta}_n(x) - \theta_n^{\star}(x) \right) = \sum_{t=1}^{d} \mathbb{P}_n \langle V_{t,\widehat{\theta}_n}, \widehat{\theta}_n(x) - \theta_n^{\star}(x) \rangle^2$$

Moreover, we note that since $\|\nabla_{\zeta\zeta} \ell(\theta(x), \widehat{g}_n(w), z)\|_{\infty} \leq U$, it must be that $\|V_{t,\theta}\|_2^2 \leq U$, for all $\theta \in \Theta_n$, which implies that $\|V_{t,\widehat{\theta}_n}\|_2 \leq \sqrt{U}$. We now argue that for any δ_n that satisfies the conditions of the theorem, for some universal constants μ_1, μ_2, with probability $1 - \mu_1 \exp\{-\mu_2 n \delta_n^2\}$:

$$\mathbb{P}_n \langle V_{t,\widehat{\theta}_n}, \widehat{\theta}_n(x) - \theta_n^{\star}(x) \rangle^2 \geq \frac{1}{2} \mathbb{P} \langle V_{t,\widehat{\theta}_n}, \widehat{\theta}_n(x) - \theta_n^{\star}(x) \rangle^2 - O(\delta_n^2).$$

Consider the function class: $\mathcal{G} = \{Z \to \langle V_{t,\theta}, \theta(X) - \theta_n^{\star}(X) \rangle : \theta \in \Theta_n\}$. Let $\|g\|_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} g(Z_i)^2}$ and $\|g\| = \sqrt{\mathbb{E}[g(Z)^2]}$. Invoking Theorem 4.1 of Wainwright (2019), we have that for any $\hat{\delta}_n$ that satisfies the inequality:

$$\mathcal{R}(\hat{\delta}_n, \text{star}(\mathcal{G}, 0)) \leq \hat{\delta}_n^2$$

with probability $1 - \mu_3 \exp\{-\mu_4 n \hat{\delta}_n^2\}$:

$$\|g\|_n^2 \geq \frac{1}{2} \|g\|^2 - \frac{\hat{\delta}_n^2}{2} \qquad\qquad \forall g \in \mathcal{G}$$

The latter immediately translates to:

$$\mathbb{P}_n \langle V_{t,\theta}, \theta(x) - \theta_n^{\star}(x) \rangle^2 \geq \frac{1}{2} \mathbb{P} \langle V_{t,\theta}, \theta(x) - \theta_n^{\star}(x) \rangle^2 - \frac{1}{2} \hat{\delta}_n^2 \qquad\qquad \forall \theta \in \Theta_n$$

Thus the latter also holds for $\theta = \widehat{\theta}_n$. We now relate $\hat{\delta}_n$, with the δ_n from our theorem. Observe that because $\|V_{t,\theta}\| \leq \sqrt{U}$, we have that the function $Z \to \langle V_{t,\theta}, \theta(X) - \theta_n^{\star}(X) \rangle$ is $\sqrt{U}$-Lipschitz in $\theta(X) - \theta_n^{\star}(X)$. Thus by the vector-valued contraction lemma of Maurer (2016), we have (see also proof of Lemma 4), for any $r > 0$:

$$\mathcal{R}(r, \text{star}(\mathcal{G})) \leq 4\sqrt{U} \sum_{t=1}^{d} \mathcal{R}(r, \text{star}(\Theta_{n,t} - \theta_{n,t}^{\star}))$$

Now for any $r \geq \delta_n$, we have:

$$\mathcal{R}(r, \text{star}(\mathcal{G})) \leq 4\sqrt{U} \sum_{t=1}^{d} \mathcal{R}(r, \text{star}(\Theta_{n,t} - \theta_{n,t}^{\star})) \leq 4\sqrt{U} d \delta_n r$$

Thus if we choose $\hat{\delta}_n = 4\sqrt{U} d \delta_n$, then:

$$\mathcal{R}(\hat{\delta}_n, \text{star}(\mathcal{G})) \leq 4\sqrt{U} d \delta_n \hat{\delta}_n = \hat{\delta}_n^2$$

which is what is required for $\hat{\delta}_n$. Thus we get that with probability $1 - \mu_3 \exp\{-16 U d^2 \mu_4 n \delta_n^2\}$:

$$\mathbb{P}_n \langle V_{t,\widehat{\theta}_n}, \widehat{\theta}_n(x) - \theta_n^{\star}(x) \rangle^2 \geq \frac{1}{2} \mathbb{P} \langle V_{t,\widehat{\theta}_n}, \widehat{\theta}_n(x) - \theta_n^{\star}(x) \rangle^2 - 8 U d^2 \delta_n^2 \qquad\qquad \forall \theta \in \Theta_n$$

as desired.

Now taking a union bound over all t, and since d is a constant, we get that the latter holds uniformly for all t. Hence, we conclude that with a probability of $1 - \mu_3' \exp\{-\mu_4' n \delta_n^2\}$:

$$\sum_{t=1}^{d} \mathbb{P}_n \langle V_{t,\widehat{\theta}_n}, \widehat{\theta}_n(x) - \theta_n^\star(x) \rangle^2 \geq \frac{1}{2} \sum_{t=1}^{d} \mathbb{P} \langle V_{t,\widehat{\theta}_n}, \widehat{\theta}_n(x) - \theta_n^\star(x) \rangle^2 - O(\delta_n^2)$$

Subsequently the latter implies that:

$$\mathbb{P}_n \left(\widehat{\theta}_n(x) - \theta_n^\star(x) \right)^\top \nabla_{\zeta\zeta} \ell(\bar{\theta}(x), \widehat{g}_n(w), z) \left(\widehat{\theta}_n(x) - \theta_n^\star(x) \right)$$
$$\geq \frac{1}{2} \mathbb{P} \left(\widehat{\theta}_n(x) - \theta_n^\star(x) \right)^\top \nabla_{\zeta\zeta} \ell(\bar{\theta}(x), \widehat{g}_n(w), z) \left(\widehat{\theta}_n(x) - \theta_n^\star(x) \right) - O(\delta_n^2)$$

The first term on the right hand side is equal to $D_\theta^2 L_{\mathcal{D}}(\bar{\theta}, g)[\widehat{\theta}_n - \theta_n^\star, \widehat{\theta}_n - \theta_n^\star]$. Hence, by Assumption 3, we get that the latter is lower bounded by:

$$\frac{\lambda}{2} \|\widehat{\theta}_n - \theta_n^\star\|^2 - \frac{\kappa}{2} \|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4 - O(\delta_n^2)$$

$\square$

E.2 Proofs for Specific Function Class Guarantees

Lemma 11 (Mendelson (2002), Lemma 4.5)**.** *For any real-valued function class $\mathcal{F}$ with $\|f\|_n \leq 1$ for all $f \in \mathcal{F}$ and any $f^\star$ with $\|f^\star\|_n \leq 1$,*

$$\mathcal{H}(\text{star}(\mathcal{F}, f^\star), \varepsilon, z_{1:n}) \leq \mathcal{H}(\mathcal{F}, \varepsilon/2, z_{1:n}) + \log(4/\varepsilon). \tag{113}$$

Lemma 12 (Wainwright (2019), Proposition 14.1)**.** *Let δ_n be the minimal solution to*

$$\mathfrak{R}_n(\delta\,;\,\mathcal{F}) \leq \delta^2,$$

where $\mathcal{F} \subseteq (\mathcal{Z} \to \mathbb{R})$ is a star-shaped set with $\sup_{f \in \mathcal{F}} \sup_{z \in \mathcal{Z}} |f(z)| \leq 1$. Then with probability at least $1 - e^{-cn\delta_n^2}$ over the draw of data $z_{1:n}$,

$$\delta_n \leq 34 \widehat{\delta}_n,$$

where $\widehat{\delta}_n$ is the minimal solution to

$$\mathfrak{R}_n(\delta\,;\,\mathcal{F}, z_{1:n}) \coloneqq \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}: \|f\|_n \leq \delta} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(z_t) \right| \right] \leq \delta^2. \tag{114}$$

Proposition 3. *Define*

$$\mathcal{F}(\delta, z_{1:n}) = \{ f \in \mathcal{F} : \|f\|_n \leq \delta \}. \tag{115}$$

Then any minimal solution to

$$\int_{\frac{\delta^2}{8}}^{\delta} \sqrt{\frac{\mathcal{H}(\varepsilon, \mathcal{F}(\delta, z_{1:n}), z_{1:n})}{n}} d\varepsilon \leq \frac{\delta^2}{20}. \tag{116}$$

is an upper bound for the fixed point $\widehat{\delta}_n$ in (114).

Proof. Immediately follows from Lemma 13. $\square$

Proof for Example 1. Let data $x_{1:n}$ be fixed. Let $\mathcal{G}_b = \{x \mapsto \langle \theta, x \rangle \mid \|\theta\|_1 \le b\}$. Under our assumption that $\|x_t\|_\infty \le 1$, Zhang (2002), Theorem 3, implies that

$$\mathcal{H}(\varepsilon, \mathcal{G}_b, x_{1:n}) \le O\left(\frac{b^2 \log(d)}{\varepsilon^2}\right).$$

We will now establish that $\Theta_n(\delta, x_{1:n}) \coloneqq \{x \mapsto \langle \theta - \theta_n^\star, x \rangle \mid \theta \in \Theta_n, \|\theta - \theta_n^\star\|_n \le \delta\} \subseteq \mathcal{G}_b$ for an appropriate choice of b using the restricted eigenvalue bound. Let $\mathcal{C} = \{\Delta \in \mathbb{R}^d \mid \|\Delta_{T^c}\|_1 \le \|\Delta_T\|_1\}$. We first claim $\Theta_n - \theta_n^\star \subset \mathcal{C}$. Indeed, fix $\theta \in \Theta_n$ and let $\Delta = \theta - \theta_n^\star$. Then we have

$$\|\theta_n^\star\|_1 \ge \|\theta\|_1 = \|\theta_n^\star + \Delta\|_1 = \|\theta_n^\star + \Delta_T\|_1 + \|\Delta_{T^c}\|_1 \ge \|\theta_n^\star\|_1 - \|\Delta_T\|_1 + \|\Delta_{T^c}\|_1.$$

Rearranging, we get $\|\Delta_{T^c}\|_1 \le \|\Delta_T\|_1$ as desired. Now observe that for any $\Delta \in \mathcal{C}$, we have

$$\|\Delta\|_1 \le \|\Delta_T\|_1 + \|\Delta_{T^c}\|_1 \le 2\|\Delta_T\|_1 \le 2\sqrt{s}\|\Delta\|_2 \le \frac{2\sqrt{s}}{\sqrt{\gamma_{\mathrm{re}}}} \cdot \frac{1}{\sqrt{n}} \|X\Delta\|_2.$$

This implies that $\Theta_n(\delta, x_{1:n}) \subseteq \mathcal{G}_b$ for $b = \frac{2\sqrt{s}}{\sqrt{\gamma_{\mathrm{re}}}}\delta$, and as a consequence

$$\mathcal{H}(\varepsilon, \Theta_n(\delta, x_{1:n}), x_{1:n}) \le O\left(\frac{s \log(d)}{\gamma_{\mathrm{re}} \cdot \varepsilon^2}\delta^2\right).$$

We plug this bound into (116) and derive an upper bound of

$$\int_{\frac{\delta^2}{8}}^{\delta} \sqrt{\frac{\mathcal{H}(\varepsilon, \mathrm{star}(\Theta_n - \theta_n^\star, 0), x_{1:n})}{n}} d\varepsilon \le O\left(\delta \cdot \sqrt{\frac{s \log(d)}{\gamma_{\mathrm{re}} n}} \cdot \int_{\frac{\delta^2}{8}}^{\delta} \frac{1}{\varepsilon} d\varepsilon\right) \le O\left(\delta \log(1/\delta) \cdot \sqrt{\frac{s \log(d)}{\gamma_{\mathrm{re}} n}}\right).$$

where we have used that $\mathrm{star}(\Theta_n - \theta_n^\star, 0) = \Theta_n - \theta_n^\star$. Using Proposition 3, we may now take $\delta_n \le O\left(\sqrt{\frac{s \log(d/s)}{n}} \log n\right)$ in Equation 25, then combine with Theorem 3 and Lemma 1 to get the result.

$\square$

Proof for Example 2. Since $\|\theta\|_1 \le 1$ and $\|x\|_\infty \le 1$, the standard covering number bound for linear classes states that the covering number at scale ε for any fixed sparsity pattern is at most $C \cdot s \log(1/\varepsilon)$. We take the union over all such covers for all $\binom{d}{s} \le \left(\frac{ed}{s}\right)^s$ sparsity patterns, which implies $\mathcal{H}(\varepsilon, \Theta_n - \theta_n^\star, x_{1:n}) \propto s(\log(d/s) + \log(1/\varepsilon))$. Lemma 11 further implies that

$$\mathcal{H}(\varepsilon, \mathrm{star}(\Theta_n - \theta_n^\star, 0), x_{1:n}) \propto s(\log(d/s) + \log(1/\varepsilon)).$$

It is now a standard calculation to show that

$$\int_{\frac{\delta^2}{8}}^{\delta} \sqrt{\frac{\mathcal{H}(\varepsilon, \mathrm{star}(\Theta_n - \theta_n^\star, 0), x_{1:n})}{n}} d\varepsilon \le O\left(\delta\sqrt{\log(1/\delta)} \cdot \sqrt{\frac{s \log(d/s)}{n}}\right).$$

Thus, via Lemma 12 and Proposition 3, we may take $\delta_n \le O\left(\sqrt{\frac{s \log(d/s) \log n}{n}}\right)$ in Equation 25. The final result follows by combining Theorem 3 and Lemma 1.

$\square$

Proof for Example 3. Since our the target class is bounded, Lemma 11 implies

$$\mathcal{H}(2\varepsilon, \text{star}(\Theta_n - \theta_n^\star), x_{1:n}) \le \mathcal{H}(\varepsilon, \Theta_n - \theta_n^\star, x_{1:n}) + \log(2/\varepsilon) = \mathcal{H}(\varepsilon, \Theta_n, x_{1:n}) + \log(2/\varepsilon).$$

Recall

$$\mathcal{F} = \left\{ f(x) \coloneqq A_L \cdot \sigma_{\text{relu}}(A_{L-1} \cdot \sigma_{\text{relu}}(A_{L-2} \cdot \sigma_{\text{relu}}(A_1 x) \dots)) \mid A_i \in \mathbb{R}^{d_i \times d_{i-1}}, \|f\|_{L_\infty} \le M \right\}.$$

Then, since σ_{relu} is 1-Lipschitz and positive-homogeneous, we have $\mathcal{H}(\varepsilon, \Theta_n, x_{1:n}) \le \mathcal{H}(\varepsilon, \mathcal{F}, x_{1:n})$. Recall that for $[0, M]$-valued classes of regressors we can relate the L_2 metric to the L_1 metric for a closely related VC class as follows. Let $Y \sim \text{unif}([0, M])$, let $f, f' \in \mathcal{G}$, and write

$$\mathbb{P}_n(f(X) - f'(X))^2 = M^2 \mathbb{P}_n(\mathbb{P}_Y(Y \le f(X)) - \mathbb{P}_Y(Y \le f'(X)))^2$$
$$\le M^2(\mathbb{P}_n \times \mathbb{P}_Y)(\mathbb{1}\{Y \le f(X)\} - \mathbb{1}\{Y \le f'(X)\}).$$

Consequently, we see that the L_2 covering number for $\mathcal{F}$ on the distribution $\mathbb{P}_n$ at scale ε, is at most the size of the L_1 cover of the class $\mathcal{F}' = \{(x, y) \mapsto \{y \le f(x)\} \mid f \in \mathcal{F}\}$ on distribution $\mathbb{P}_n \times \mathbb{P}_Y$ at scale ε^2/M. Thus, invoking Haussler's L_1 covering number bound for VC classes (Haussler, 1995), we have

$$\mathcal{H}(\varepsilon, \Theta_n, x_{1:n}) \le 2 \cdot \text{vc}(\mathcal{F}') \log\left(\frac{CM}{\varepsilon}\right) = 2 \cdot \text{pdim}(\mathcal{F}) \log\left(\frac{CM}{\varepsilon}\right),$$

where $\text{vc}(\cdot)$ denotes the VC dimension and $\text{pdim}(\cdot)$ denotes the pseudodimension. Using Theorem 14.1 from Anthony and Bartlett (1999) and Theorem 6 from Bartlett et al. (2017), we have

$$\text{pdim}(\mathcal{F}) \le O(LW \log(W)).$$

With this bound on the metric entropy we have

$$\int_{\frac{\delta^2}{8}}^{\delta} \sqrt{\frac{\mathcal{H}(\varepsilon, \text{star}(\Theta_n - \theta_n^\star, 0), x_{1:n})}{n}} d\varepsilon \le O\left(\sqrt{\frac{LW \log W \log M}{n}}\right) \cdot \int_{\frac{\delta^2}{8}}^{\delta} \sqrt{\log(1/\varepsilon)} d\varepsilon$$
$$\le O\left(\sqrt{\frac{LW \log W \log M}{n}} \cdot \delta \sqrt{\log(1/\delta)}\right).$$

Thus, it suffices to take $\delta_n \le O\left(\sqrt{\frac{LW \log W \log M \log n}{n}}\right)$ in Equation 25 and appeal to Theorem 3 and Lemma 1. $\qquad\square$

Proof for Example 4. As in Example 3, we have $\mathcal{H}(\varepsilon, \text{star}(\Theta_n - \theta_n^\star), x_{1:n}) \le \mathcal{H}(\varepsilon, \mathcal{F}, x_{1:n})$. Theorem 3.3 of Bartlett et al. (2017) implies that under our assumptions,

$$\mathcal{H}(\varepsilon, \mathcal{F}, x_{1:n}) \le O\left(\frac{n \log W}{\varepsilon^2} \prod_{i=1}^{K} s_i^2 \cdot \left(\sum_{i=1}^{L} (b_i/s_i)^{2/3}\right)^3\right).$$

The result follows by plugging this bound into Proposition 3 and proceeding exactly as in the previous examples. $\qquad\square$

Proof for Example 5 and Example 6. Note that each target class Θ_n has range bounded by 1. By examples 14.4 and 14.3 in Wainwright (2019), we may take $\delta_n = c\sqrt{\log n/n}$ and $\delta_n = cn^{-1/3}$ in Equation 25 for the gaussian and Sobolev classes respectively. We combine with this with Theorem 3 and Lemma 1. $\qquad\square$

E.3 Proof of Theorem 4

We prove the theorem for the case of $R = 1$. The general R version then easily follows by a standard re-scaling argument. Applying the first part of Lemma 4 with $\mathcal{F} = \ell \circ \Theta_n$, $f^* = \ell(\theta_n^\star(\cdot), \widehat{g}_n(\cdot), \cdot)$, $\mathcal{L}_f = f$, and $r = \sup_{f \in \mathcal{F}} \|f - f^*\|_2$, we have that with probability $1 - \delta$:

$$\sup_{f \in \mathcal{F}} \left| \mathbb{P}_n(\mathcal{L}_f - \mathcal{L}_{f^*}) - \mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*}) \right| \leq 16\mathcal{R}(r, \mathcal{F} - f^*) + O\left(r\sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n} \right)$$

Hence, the latter also holds for $\hat{f} = \ell(\widehat{\theta}_n(\cdot), \widehat{g}_n(\cdot), \cdot)$. Moreover by the definition of plug-in ERM, $\mathbb{P}_n(\mathcal{L}_{\hat{f}} - \mathcal{L}_{f^*}) \leq 0$. Thus we get:

$$\mathbb{P}(\mathcal{L}_{\hat{f}} - \mathcal{L}_{f^*}) \leq 16\mathcal{R}(r, \mathcal{F} - f^*) + O\left(r\sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n} \right)$$

This proves the first part of the Theorem. We now analyze the quantity $\mathcal{R}(r, \mathcal{F} - f^*)$ so as to provide the second part.

Let $\mathfrak{R}_n(\delta, \mathcal{F}, z_{1:n}) = \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(z_i) \right| \right]$ denote the Rademacher complexity conditional on the dataset $z_{1:n}$. Thus $\mathcal{R}(r, \mathcal{F} - f^*) = \mathbb{E}_{z_{1:n}}[\mathfrak{R}_n(r, \mathcal{F} - f^*, z_{1:n})]$. Let $r_n = \sup_{f \in \mathcal{F}} \|f - f^*\|_{2,n}$, which is a random variable that depends on the dataset $z_{1:n}$. Invoking Lemma 13, we have

$$\mathfrak{R}_n(r, \mathcal{F} - f^\star, z_{1:n}) \leq \inf_{\alpha \geq 0}\left[4\alpha + 10 \underbrace{\int_\alpha^{r_n} \sqrt{\frac{\mathcal{H}_2(\varepsilon, \mathcal{F} - f^\star, n)}{n}} d\varepsilon}_{K(\alpha, r_n)} \right].$$

Thus we have that:

$$\mathcal{R}(r, \mathcal{F} - f^*) \leq \inf_{\alpha \geq 0}[4\alpha + 10\,\mathbb{E}_{z_{1:n}}[K(\alpha, r_n)]]$$

Since metric entropy is a non-negative, non-increasing function of ϵ, we have that $K(\alpha, \cdot)$ is a non-decreasing, concave function. Thus by Jensen's inequality:

$$\mathbb{E}_{z_{1:n}}[K(\alpha, r_n)] \leq K(\alpha, \mathbb{E}_{z_{1:n}}[r_n]) \leq K(\alpha, \sqrt{\mathbb{E}_{z_{1:n}}[r_n^2]})$$

Moreover, by the definition of r_n^2 and invoking a standard symmetrization argument for uniform convergence and Lipschitz contraction (since elements of $\mathcal{F}$ are bounded by 1), we have

$$\mathbb{E}[r_n^2 - r^2] \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \|f - f^*\|_{n,2}^2 - \|f - f^*\|_2^2 \right| \right] \leq 4\mathcal{R}(r, \mathcal{F} - f^\star).$$

Moreover, by concavity of $K(\alpha, \cdot)$ and an application of the AM-GM inequality, we also have:

$$\begin{aligned}
K(\alpha, \sqrt{\mathbb{E}_{z_{1:n}}[r_n^2]}) &\leq K(\alpha, r + 2\sqrt{\mathcal{R}(r, \mathcal{F} - f^\star)}) && \text{(non-decreasing } K(\alpha, \cdot)) \\
&\leq K(\alpha, r) + \frac{\partial K(\alpha, r)}{\partial r} 2\sqrt{\mathcal{R}(r, \mathcal{F} - f^\star)} && \text{(concavity of } K(\alpha, \cdot)) \\
&= K(\alpha, r) + \sqrt{\frac{\mathcal{H}_2(r, \mathcal{F} - f^\star, n)}{n}} 2\sqrt{\mathcal{R}(r, \mathcal{F} - f^\star)} && \text{(definition of derivative)} \\
&= K(\alpha, r) + 2\frac{\mathcal{H}_2(r, \mathcal{F} - f^\star, n)}{n} + \frac{1}{2}\mathcal{R}(r, \mathcal{F} - f^\star) && \text{(AM-GM)}
\end{aligned}$$

Thus we conclude that:

$$\mathcal{R}(r, \mathcal{F} - f^*) \leq \inf_{\alpha \geq 0} [4\alpha + 10K(\alpha, r)] + 2\frac{\mathcal{H}_2(r, \mathcal{F} - f^\star, n)}{n} + \frac{1}{2}\mathcal{R}(r, \mathcal{F} - f^\star)$$

Re-arranging yields:

$$\mathcal{R}(r, \mathcal{F} - f^*) \leq \inf_{\alpha \geq 0} [8\alpha + 20K(\alpha, r)] + 4\frac{\mathcal{H}_2(r, \mathcal{F} - f^\star, n)}{n}$$

Moreover, observe that if $\mathcal{F}_\epsilon$ is an ϵ-cover of $\mathcal{F}$, the $\mathcal{F}_\epsilon - f^*$ is an ϵ-cover of $\mathcal{F} - f^*$. Thus: $\mathcal{H}_2(\varepsilon, \mathcal{F} - f^\star, n) \leq \mathcal{H}_2(\varepsilon, \mathcal{F}, n)$. This completes the proof of the theorem.

E.4 Proof of Theorem 5

Proof of Theorem 5. Applying Lemma 7, the fact that $\|f - f^*\|_2 \leq 2\|f - f^*\|_{n,2} + \delta_n$ with probability $1 - c_7 \exp\{c_8 n \delta_n^2\}$ (via Theorem 4.1 of Wainwright (2019)) and the fact that $\mathcal{L}_f$ is linear (it is the identity function), we get that for any $\delta_n \geq 0$ that satisfies the conditions of the theorem, with probability $1 - c_9 \exp\{c_{10} n \delta_n^2\}$:

$$\left| \mathbb{P}_n(\mathcal{L}_f - \mathcal{L}_{f^*}) - \mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*}) \right| \leq 18\delta_n\{\|f - f^*\|_2 + \delta_n\} \leq 36\delta_n\{\|f - f^*\|_{n,2} + \delta_n\} \quad \forall f \in \mathcal{F} \quad (117)$$

Then we know that with probability $1 - c_9 \exp\{c_{10} n \delta_n^2\}$:

$$\begin{aligned}
\mathbb{P}(\mathcal{L}_{\hat{f}} - \mathcal{L}_{f^*}) &\leq \mathbb{P}_n(\mathcal{L}_{\hat{f}} - \mathcal{L}_{f^*}) + 36\delta_n\|\hat{f} - f^*\|_{n,2} + 36\delta_n^2 \\
&\leq \mathbb{P}_n(\mathcal{L}_{\hat{f}} - \mathcal{L}_{f^*}) + 36\delta_n\|\hat{f}\| + 36\delta_n\|f^*\|_{n,2} + 36\delta_n^2 \\
&\leq 72\delta_n\|f^*\|_{n,2} + 36\delta_n^2
\end{aligned}$$

where the second inequality follows by the definition of the moment penalized algorithm, since:

$$\mathbb{P}_n\mathcal{L}_f + 36\delta_n\|\hat{f}\|_{n,2} \leq \mathbb{P}_n\mathcal{L}_{f^*} + 36\delta_n\|f^*\|_{n,2}$$

$\square$

E.5 Proof of Theorem 6

Proof of Theorem 6.
Part 1: Overview. Define the function class $\mathcal{F} = \{z \mapsto \ell(\theta(x), \widehat{g}_n(w), z) : \theta \in \Theta_n\}$. We assume for now that $\|f\|_\infty \leq 1$ for all $f \in \mathcal{F}$, that Γ is 1-Lipschitz (i.e. $L = R = 1$ in the theorem statement) and that $\mathbb{E}[\Gamma^2(g_0, z) \mid x] \geq \gamma$; the general case will be handled by rescaling at the end of the proof..

Our starting point is to appeal to Theorem 5 . In particular, let $\delta_n \geq 0$ be any solution to the inequality:

$$\mathcal{R}_n(\delta, \text{star}(\mathcal{F} - f^*)) \leq \delta^2, \quad (118)$$

where $f^* = z \mapsto \ell(\theta_n^\star(x), \widehat{g}_n(w), z)$. Then if $\widehat{\theta}_n$ is the outcome of variance-penalized ERM— per (36) and the discussion around it—with probability at least $1 - \delta$,

$$L_\mathcal{D}(\widehat{\theta}_n, \widehat{g}_n) - L_\mathcal{D}(\theta_n^\star, \widehat{g}_n) = O\left(\sqrt{V^\star}\left(\delta_n + \sqrt{\frac{\log(1/\delta)}{n}} \right) + \delta_n^2 + \frac{\log(1/\delta)}{n} + \left(\text{Rate}_\mathcal{D}(\mathcal{G}_n, S^{(1)}, \delta/2) \right)^2 \right).$$

Furthermore, Corollary 1, orthogonality implies that

$$L_\mathcal{D}(\widehat{\theta}_n, g_0) - L_\mathcal{D}(\theta_n^\star, g_0) = O\left(\sqrt{V^\star}\left(\delta_n + \sqrt{\frac{\log(1/\delta)}{n}} \right) + \delta_n^2 + \frac{\log(1/\delta)}{n} + \beta\left(\text{Rate}_\mathcal{D}(\mathcal{G}_n, S^{(1)}, \delta/2) \right)^2 \right).$$

Part 2: Moving to capacity function at g_0. Per the discussion in the previous section, we focus on bounding the critical radius in the case where $\mathcal{F}$ is bounded by 1 and Γ is 1-Lipschitz. We wish to make use of the capacity function, which is defined at g_0, but the local Rademacher complexity we need to bound is that of $\mathcal{F}$, which evaluates the weight function Γ at $\widehat{g}_n$. To make progress, we show how to use the capacity function defined in the theorem statement to bound the an alternative capacity function

$$\widetilde{\tau}^2(\varepsilon) = \frac{\mathbb{E}\sup_{\theta:\mathbb{E}[\Gamma(\widehat{g}_n,z)^2(\theta(x)-\theta_n^\star(x))^2]\leq\varepsilon^2}\Gamma^2(\widehat{g}_n,z)(\theta(x)-\theta_n^\star(x))^2}{\varepsilon^2}.$$

We first show how to relate the L_2 norm at $\widehat{g}_n$ to the L_2 norm at g_0. Define $\|\theta\|_{\Theta_n} = \sqrt{\mathbb{E}\,\Gamma(\widehat{g}_n,z)^2(\theta(x)-\theta_n^\star(x))^2}$. Then for any $\theta \in \Theta_n$ we have

$$\mathbb{E}\big[\Gamma(\widehat{g}_n,z)^2(\theta(x)-\theta_n^\star(x))^2\big]$$
$$\geq \frac{1}{2}\,\mathbb{E}\,\Gamma(g_0,z)^2(\theta(x)-\theta_n^\star(x))^2 - \mathbb{E}(\Gamma(\widehat{g}_n,z)-\Gamma(g_0,z))^2(\theta(x)-\theta_n^\star(x))^2$$
$$= \frac{1}{2}\|\theta-\theta_n^\star\|_{\Theta_n}^2 - \mathbb{E}(\Gamma(\widehat{g}_n,z)-\Gamma(g_0,z))^2(\theta(x)-\theta_n^\star(x))^2.$$

Using AM-GM and boundedness of θ, for any $\eta > 0$ this is lower bounded by

$$\geq \frac{1}{2}\|\theta-\theta_n^\star\|_{\Theta_n}^2 - \frac{1}{2\eta}\,\mathbb{E}(\Gamma(\widehat{g}_n,z)-\Gamma(g_0,z))^4 - \frac{\eta}{2}\,\mathbb{E}(\theta(x)-\theta_n^\star(x))^2$$

Using the Lipschitz assumption and conditional lower bound on Γ, we further lower bound by

$$\geq \frac{1}{2}\|\theta-\theta_n^\star\|_{\Theta_n}^2 - \frac{1}{2\eta}\|\widehat{g}_n-g_0\|_{\mathcal{G}_n}^4 - \frac{\eta}{2\gamma}\|\theta-\theta_n^\star\|_{\Theta_n}^2.$$

Hence, by choosing $\eta = \gamma/2$ and rearranging, we get

$$\|\theta-\theta_n^\star\|_{\Theta_n}^2 \leq 4\,\mathbb{E}\big[\Gamma(\widehat{g}_n,z)^2(\theta(x)-\theta_n^\star(x))^2\big] + \frac{4}{\gamma}\|\widehat{g}_n-g_0\|_{\mathcal{G}_n}^4. \tag{119}$$

We now proceed to bound the capacity function $\widetilde{\tau}$. Let $\varepsilon_0 = \frac{2}{\gamma^{1/2}}\|\widehat{g}_n-g_0\|_{\mathcal{G}_n}^2$. Let $\varepsilon \geq \varepsilon_0$ be fixed and let $\theta \in \Theta_n$ be any policy with $\mathbb{E}\big[\Gamma(\widehat{g}_n,z)^2(\theta(x)-\theta_n^\star(x))^2\big] \leq \varepsilon^2$. Then equation (119) implies that that $\|\theta-\theta_n^\star\|_{\Theta_n}^2 \leq 5\varepsilon^2$, and so

$$\widetilde{\tau}^2(\varepsilon) \leq \frac{\mathbb{E}\sup_{\theta:\|\theta-\theta_n^\star\|_{\Theta_n}^2\leq5\varepsilon^2}\Gamma^2(\widehat{g}_n,z)(\theta(x)-\theta_n^\star(x))^2}{\varepsilon^2}.$$

To handle the term in the numerator we proceed similar to the proof of (119). We have

$$\mathbb{E}\sup_{\theta:\|\theta-\theta_n^\star\|_{\Theta_n}^2\leq5\varepsilon^2}\Gamma^2(\widehat{g}_n,z)(\theta(x)-\theta_n^\star(x))^2$$
$$\leq 2\,\mathbb{E}\sup_{\theta:\|\theta-\theta_n^\star\|_{\Theta_n}^2\leq5\varepsilon^2}\Gamma^2(g_0,z)(\theta(x)-\theta_n^\star(x))^2 + 2\,\mathbb{E}\sup_{\theta:\|\theta-\theta_n^\star\|_{\Theta_n}^2\leq5\varepsilon^2}(\Gamma(\widehat{g}_n,z)-\Gamma(g_0,z))^2(\theta(x)-\theta_n^\star(x))^2.$$

Fix any $\eta > 0$. We use AM-GM and boundedness of policies to upper bound the second term as

$$\mathbb{E} \sup_{\theta : \|\theta - \theta_n^\star\|_{\Theta_n}^2 \le 5\varepsilon^2} (\Gamma(\widehat{g}_n, z) - \Gamma(g_0, z))^2 (\theta(x) - \theta_n^\star(x))^2$$

$$\le \frac{1}{\eta} \mathbb{E}(\Gamma(\widehat{g}_n, z) - \Gamma(g_0, z))^4 + \eta \, \mathbb{E}_x \sup_{\theta : \|\theta - \theta_n^\star\|_{\Theta_n}^2 \le 5\varepsilon^2} (\theta(x) - \theta_n^\star(x))^2$$

$$\le \frac{1}{\eta} \|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4 + \frac{\eta}{\gamma} \mathbb{E}_x \sup_{\theta : \|\theta - \theta_n^\star\|_{\Theta_n}^2 \le 5\varepsilon^2} \mathbb{E}\big[\Gamma^2(g_0, z) \mid x\big](\theta(x) - \theta_n^\star(x))^2$$

$$\le \frac{1}{\eta} \|\widehat{g}_n - g_0\|_{\mathcal{G}_n}^4 + \frac{\eta}{\gamma} \mathbb{E} \sup_{\theta : \|\theta - \theta_n^\star\|_{\Theta_n}^2 \le 5\varepsilon^2} \Gamma^2(g_0, z)(\theta(x) - \theta_n^\star(x))^2$$

We choose $\eta = \gamma$ and recall the definition of ε_0, which gives

$$\le \varepsilon_0^2/4 + \mathbb{E} \sup_{\theta : \|\theta - \theta_n^\star\|_{\Theta_n}^2 \le 5\varepsilon^2} \Gamma^2(g_0, z)(\theta(x) - \theta_n^\star(x))^2.$$

Putting everything together, we get

$$\widetilde{\tau}^2(\varepsilon) \le \frac{4 \cdot \mathbb{E} \sup_{\theta : \|\theta - \theta_n^\star\|_{\Theta_n}^2 \le 5\varepsilon^2} \Gamma^2(g_0, z)(\theta(x) - \theta_n^\star(x))^2 + \varepsilon_0^2/2}{\varepsilon^2}.$$

Thus, for all $\varepsilon \ge \varepsilon_0$ we have

$$\widetilde{\tau}^2(\varepsilon) \le 20 \cdot \tau^2(5\varepsilon) + 3. \tag{120}$$

Part 3: Bounding the critical radius. For any class $\mathcal{G}$ we define $\mathcal{G}_\delta = \{g \in \mathcal{G} : \|g\|_2 \le \delta\}$. We work with the following localized empirical Rademacher complexity

$$\mathfrak{R}_n(\delta, \mathcal{G}, z_{1:n}) = \mathbb{E}_\epsilon \left[\sup_{g \in \mathcal{G}_\delta} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i g(z_i) \right| \right], \tag{121}$$

which has $\mathcal{R}_n(\delta, \mathcal{G}) = \mathbb{E}_{z_{1:n}} [\mathfrak{R}_n(\delta, \mathcal{G}, z_{1:n})]$.

For the remainder of the proof we define $\mathcal{G} = \mathrm{star}(\mathcal{F} - f^\star)$. Let the draw of $z_{1:n}$ be fixed. Invoking Lemma 13, we have

$$\mathfrak{R}_n(\delta, \mathrm{star}(\mathcal{F} - f^\star), z_{1:n}) \le \inf_{\alpha \ge 0} \left[4\alpha + 10 \int_\alpha^{\sup_{g \in \mathcal{G}_\delta} \|g\|_n} \sqrt{\frac{\mathcal{H}_2(\varepsilon, \mathcal{G}_\delta, z_{1:n})}{n}} d\varepsilon \right].$$

We also define $\mathcal{F}(\delta) = \{f \in \mathcal{F} : \|f - f^\star\|_2 \le \delta\}$. Using that any $g \in \mathcal{G}$ can be written as $r \cdot (f - f^\star)$, where $\|f - f^\star\|_2 \le \delta$ and $r \in [0, 1]$, a simple discretization argument (see the proof of Lemma 11) shows that

$$\mathcal{H}_2(\varepsilon, \mathcal{G}_\delta, z_{1:n}) \le \mathcal{H}_2(\varepsilon/2, (\mathcal{F} - f^\star)_\delta, z_{1:n}) + \log \left(2 \sup_{f \in \mathcal{F}(\delta)} \|f - f^\star\|_n / \varepsilon \right).$$

Let us adopt the shorthand $v_n = \sup_{f \in \mathcal{F}(\delta)} \|f - f^\star\|_n$. It follows from the usual symmetrization argument that $\mathbb{E} v_n^2 \le \delta^2 + 2\mathcal{R}_n(\delta, \mathcal{F} - f^\star)$. Letting $\alpha = 0$ be fixed, we can summarize our argument so far as

$$\mathfrak{R}_n(\delta, \mathrm{star}(\mathcal{F} - f^\star), z_{1:n}) \le 10 \int_0^{v_n} \sqrt{\frac{\mathcal{H}_2(\varepsilon/2, (\mathcal{F} - f^\star)_\delta, z_{1:n})}{n}} d\varepsilon + 10 \int_0^{v_n} \sqrt{\log(2v_n/\varepsilon)/n} d\varepsilon.$$

65

Furthermore, using a change of variables we have

$$\int_0^{v_n} \sqrt{\log(2v_n/\varepsilon)/n}\,d\varepsilon \le v_n \int_0^1 \sqrt{\log(2/\varepsilon)/n}\,d\varepsilon \le C \cdot \frac{v_n}{\sqrt{n}}.$$

We now handle the covering integral for $(\mathcal{F}-f^\star)_\delta$. We first need some additional notation. Let $g = (f - f^\star)$ and $g' = (f'-f^\star)$ be fixed elements of $(\mathcal{F}-f^\star)_\delta$. Let $\Theta_n(\delta) = \{\theta \in \Theta_n \mid \mathbb{E}\,\Gamma(\widehat{g}_n, z)(\theta(x) - \theta_n^\star(x))^2\delta^2\}$, and let $\theta, \theta' \in \Theta_n(\delta)$ be such that $f = \Gamma(\widehat{g}_n, z)(\theta(x) - \theta_n^\star(x))$ and likewise for f' and θ'. The following quantity will feature prominently in the analysis: $u_n^2 = \frac{1}{n}\sum_{t=1}^n \sup_{\theta\in\Theta_n(\delta)} \Gamma_t^2(\theta(x_t) - \theta_n^\star(x_t))^2$. Note in particular that $u_n^2 \ge v_n^2$ by definition.

Our approach is to upper bound the empirical L_2 covering number for $(\mathcal{F} - f^\star)$ by the covering number of the class Θ_n with respect to Hamming error. In particular, we claim that there is a dataset $S' = x_1', \ldots, x_M'$ such that for all $g, g' \in (\mathcal{F} - f^\star)_\delta$, the empirical L_2 error on S is upper bounded by the empirical Hamming error of the associated policies θ, θ' on S'.

Let the Hamming error on S' be defined via

$$d_{H,S'}(\theta, \theta') = \frac{1}{M}\sum_{t=1}^M \mathbb{1}\{\theta(x_t') \ne \theta'(x_t')\}.$$

Define $\Gamma_t = \Gamma(\widehat{g}_n, z_t)$, and let S' consist of $m_t := \left\lceil \frac{\sup_{\theta\in\Theta_n(\delta)}\Gamma_t^2(\theta(x_t)-\theta^\star(x_t))^2}{\delta^2}\right\rceil$ copies of example x_t. With this definition we have

$$\begin{aligned}
d_{H,S'}(\theta, \theta') &= \frac{1}{M}\sum_{t=1}^M \mathbb{1}\{\theta(x_t') \ne \theta'(x_t')\} \\
&= \frac{1}{M}\sum_{t=1}^n \left\lceil \frac{\sup_{\widetilde{\theta}\in\Theta_n(\delta)}\Gamma_t^2(\widetilde{\theta}(x_t)-\theta^\star(x_t))^2}{\delta^2}\right\rceil \mathbb{1}\{\theta(x_t) \ne \theta'(x_t)\} \\
&\ge \frac{1}{2M}\sum_{t=1}^n \frac{\Gamma_t^2(\theta(x_t)-\theta^\star(x_t))^2 + \Gamma_t^2(\theta'(x_t)-\theta^\star(x_t))^2}{\delta^2}\mathbb{1}\{\theta(x_t) \ne \theta'(x_t)\} \\
&\ge \frac{1}{4M}\sum_{t=1}^n \frac{\Gamma_t^2(\theta(x_t)-\theta'(x_t))^2}{\delta^2}\mathbb{1}\{\theta(x_t) \ne \theta'(x_t)\} \\
&= \frac{1}{4M}\sum_{t=1}^n \frac{\Gamma_t^2(\theta(x_t)-\theta'(x_t))^2}{\delta^2} \\
&= \frac{n}{4M\delta^2}d_{2,S}^2(g, g').
\end{aligned}$$

Thus, if we let $\varepsilon' = \frac{n}{4M\delta^2}\varepsilon^2$, then any ε'-cover in Hamming error is an ε-cover in L_2. We now use the following facts:

- $M \le n + \sum_{t=1}^n \frac{\sup_{\theta\in\Theta_n(\delta)}\Gamma_t^2(\theta(x_t)-\theta^\star(x_t))^2}{\delta^2} = n(1 + u_n^2/\delta^2)$.
- Haussler's bound (Haussler, 1995) implies that any class with VC dimension d admits a ε-Hamming error cover of size $e(d+1)\left(\frac{2e}{\varepsilon}\right)^d$.

Putting everything together, we get that

$$\int_0^{v_n} \sqrt{\frac{\mathcal{H}_2(\varepsilon/2, (\mathcal{F}-f^\star)_\delta, z_{1:n})}{n}}\,d\varepsilon \le \int_0^{v_n} \sqrt{\frac{d\log(2e(\delta^2+u_n^2)/\varepsilon^2)}{n}}\,d\varepsilon + C \cdot v_n\sqrt{\log d/n}.$$

It follows from the usual symmetrization argument that $\mathbb{E}\,v_n^2 \le \delta^2 + 2\mathcal{R}_n(\delta, \mathcal{F} - f^\star)$. Furthermore, using the Talagrand-type concentration bound in (89),[15] there exists a constant $C \ge 1$ such that for

---

[15]We use the assumed boundedness of elements of $\mathcal{F}$ to simplify (89) to the form that appears on this page.

any $s > 0$, with probability at least $1 - e^{-s}$ over the draw of $z_{1:n}$,

$$v_n^2 \le C\left(\delta^2 + \mathcal{R}_n(\delta, (\mathcal{F} - f^\star)) + \frac{s}{n}\right) =: \tilde{\delta}^2.$$

Thus, conditioning on this event, we have

$$\int_0^{v_n} \sqrt{\frac{d \log(2e(\delta^2 + u_n^2)/\varepsilon^2)}{n}} \, d\varepsilon$$

$$\le \int_0^{\tilde{\delta}} \sqrt{\frac{d \log(2e(\delta^2 + u_n^2)/\varepsilon^2)}{n}} \, d\varepsilon$$

$$\le \tilde{\delta} \int_0^1 \sqrt{\frac{d \log(2e(1 + u_n^2/\delta^2)/\varepsilon^2)}{n}} \, d\varepsilon$$

$$\le C \cdot \tilde{\delta} \sqrt{\frac{d \log(2e(1 + u_n^2/\delta^2))}{n}},$$

where the second inequality uses a change of variables and that $\delta \le \tilde{\delta}$.

Now, to summarize our developments so far, we have shown that with probability at least $1 - e^{-s}$,

$$\mathfrak{R}_n(\delta, \text{star}(\mathcal{F} - f^\star), z_{1:n})$$

$$\le C\left(\tilde{\delta}\sqrt{\frac{d \log(2e(1 + u_n^2/\delta^2))}{n}} + v_n\sqrt{\log d/n}\right)$$

$$\le C\left(\tilde{\delta}\sqrt{\frac{d \log(2e(1 + u_n^2/\delta^2))}{n}} + (\delta + \sqrt{\mathcal{R}_n(\delta, \mathcal{F} - f^\star)})\sqrt{\log d/n}\right)$$

$$\le C\left((\delta + \sqrt{\mathcal{R}_n(\delta, \mathcal{F} - f^\star)} + \sqrt{s/n})\sqrt{\frac{d \log(2e(1 + u_n^2/\delta^2))}{n}} + (\delta + \sqrt{\mathcal{R}_n(\delta, \mathcal{F} - f^\star)})\sqrt{\log d/n}\right).$$

Using Markov's inequality, we also have that with probability at least $1 - e^{-s}$, $u_n^2 \le \mathbb{E}\, u_n^2 \cdot e^s$. Thus, by union bound, we have that with probability at least $1 - e^{-2s}$,

$$\mathfrak{R}_n(\delta, \text{star}(\mathcal{F} - f^\star), z_{1:n})$$

$$\le C\left(\sqrt{s}(\delta + \sqrt{\mathcal{R}_n(\delta, \mathcal{F} - f^\star)} + \sqrt{s/n})\sqrt{\frac{d \log(2e^2(1 + \mathbb{E}_{z_{1:n}} u_n^2/\delta^2))}{n}} + (\delta + \sqrt{\mathcal{R}_n(\delta, \mathcal{F} - f^\star)})\sqrt{\log d/n}\right).$$

Integrating out this tail bound, we get that

$$\mathcal{R}_n(\delta, \text{star}(\mathcal{F} - f^\star))$$

$$\le C\left((\delta + \sqrt{\mathcal{R}_n(\delta, \mathcal{F} - f^\star)} + \sqrt{1/n})\sqrt{\frac{d \log(2e^2(1 + \mathbb{E}_{z_{1:n}} u_n^2/\delta^2))}{n}} + (\delta + \sqrt{\mathcal{R}_n(\delta, \mathcal{F} - f^\star)})\sqrt{\log d/n}\right).$$

Using AM-GM and that $\mathcal{R}_n(\delta, \mathcal{F} - f^\star) \le \mathcal{R}_n(\delta, \text{star}(\mathcal{F} - f^\star))$ and rearranging, this implies

$$\mathcal{R}_n(\delta, \text{star}(\mathcal{F} - f^\star)) \le C\left(\delta\sqrt{\frac{d \log(2e^2(1 + \mathbb{E}_{z_{1:n}} u_n^2/\delta^2))}{n}} + \frac{d \log(2e(1 + \mathbb{E}_{z_{1:n}} u_n^2/\delta^2))}{n}\right).$$

67

We now bound the ratio $\mathbb{E}_{z_{1:n}} u_n^2/\delta^2$. We have

$$\mathbb{E}_{z_{1:n}} u_n^2 = \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{z_t}\sup_{\theta\in\Theta_n(\delta)}\Gamma^2(\widehat{g}_n,z_t)(\theta(x_t)-\theta^\star(x_t))^2 = \mathbb{E}_z\sup_{\theta\in\Theta_n(\delta)}\Gamma^2(\widehat{g}_n,z_t)(\theta(x_t)-\theta^\star(x_t))^2 \le \widetilde{\tau}^2(\delta)\delta^2.$$

Thus, using the relationship between τ and $\widetilde{\tau}$ established in the previous section of the proof, if $\delta > \varepsilon_0$ we have

$$\mathbb{E}_{z_{1:n}} u_n^2/\delta^2 \le 20\cdot\tau^2(5\delta)+3,$$

and so for all $\delta > \varepsilon_0$,

$$\mathcal{R}_n(\delta,\text{star}(\mathcal{F}-f^\star)) \le C\left(\delta\sqrt{\frac{d\log\tau(5\delta)}{n}}+\frac{d\log\tau(5\delta)}{n}\right).$$

In particular, we can see from this expression that taking

$$\delta_n \propto \sqrt{\frac{d\log\tau_0}{n}}+\varepsilon_0$$

yields a valid upper bound on the critical radius.

Part 4: Final bound. Putting together the excess risk bound and the critical radius bound, we get

$$\begin{aligned}
&L_\mathcal{D}(\widehat{\theta}_n,g_0) - L_\mathcal{D}(\theta_n^\star,g_0)\\
&= O\left(\sqrt{V^\star}\left(\delta_n+\sqrt{\frac{\log(1/\delta)}{n}}\right)+\delta_n^2+\frac{\log(1/\delta)}{n}+(1+\beta)\left(\text{Rate}_\mathcal{D}(\mathcal{G}_n,S^{(1)},\delta/2)\right)^2\right)\\
&\le O\left(\sqrt{\frac{V^\star d\log(\tau_0/\delta)}{n}}+\frac{d\log(\tau_0/\delta)}{n}+(\gamma^{-1/2}+\beta)\left(\text{Rate}_\mathcal{D}(\mathcal{G}_n,S^{(1)},\delta/2)\right)^2\right).
\end{aligned}$$

To deduce the final bound in the general R-bounded L-Lipschitz case we divide the class by $(L+R)$, then rescale the final bound (observing that β, γ, and $V^\star$ all change under the rescaling). $\qquad\square$

F Proofs from Section 6 and 7

F.1 Preliminaries

Definition 5 (Empirical Rademacher Complexity). *For a real-valued function class $\mathcal{F}$ and sample set $S = z_1,\ldots,z_n$, the Rademacher complexity is defined via*

$$\mathfrak{R}_n(\mathcal{F},S) = \mathbb{E}_\epsilon\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(z_t)\right], \tag{122}$$

where $\epsilon = \epsilon_1,\ldots,\epsilon_n$ are i.i.d. Rademacher random variables.

We require the following technical lemmas. First is the Dudley entropy integral bound; we use the following form from Srebro et al. (2010) (Lemma A.3). In all results that follow we will use $C > 0$ to denote an absolute constant whose value may change from line to line.

Lemma 13. *For any real-valued function class $\mathcal{F} \subseteq (\mathcal{Z} \to \mathbb{R})$, we have*

$$\mathfrak{R}_n(\mathcal{F}, S) \leq \inf_{\alpha > 0} \left\{ 4\alpha + 10 \int_\alpha^{\sup_{f \in \mathcal{F}} \sqrt{\mathbb{E}_{z \sim S} f^2(z)}} \sqrt{\frac{\mathcal{H}_2(\mathcal{F}, \varepsilon, S)}{n}} d\varepsilon \right\}. \tag{123}$$

As a consequence, whenever $\mathcal{F}$ takes values in $[-1, +1]$ the following bounds hold:
- *If $\mathcal{H}_2(\mathcal{F}, \varepsilon, S) \leq \varepsilon^{-p}$, then $\mathfrak{R}_n(\mathcal{F}, S) \leq \cdot r_{n,p}$, where $r_{n,p}$ satisfies*

$$r_{n,p} \leq C_p \cdot \begin{cases} n^{-\frac{1}{2}}, & p < 2. \\ n^{-\frac{1}{2}} \cdot \log n, & p = 2. \\ n^{-\frac{1}{p}}, & p > 2. \end{cases}$$

- *If $\mathcal{H}_2(\mathcal{F}, \varepsilon, S) \leq d \log(D/\varepsilon)$, then $\mathfrak{R}_n(\mathcal{F}, S) \leq C \cdot \sqrt{d/n}$.*

We also require the following lemma, which controls the rate at which the empirical L_2 metric converges to the population L_2 metric in terms of metric entropy behavior.

Lemma 14. *Let $\mathcal{F} \subseteq (\mathcal{Z} \to [-1, +1])$, and let $S = z_{1:n}$ be a collection of samples in $\mathcal{Z}$ drawn i.i.d. from $\mathcal{D}$.*
- *If $\mathcal{H}_2(\mathcal{F}, \varepsilon, n) \propto \varepsilon^{-p}$ for some p, then with probability at least $1 - \delta$, for all $f, f' \in \mathcal{F}$,*

$$\|f - f'\|_{L_2(\mathcal{D})}^2 \leq 2d_S^2(f, f') + R_{n,p} + C \frac{\log(\log n/\delta)}{n},$$

  where

$$R_{n,p} \leq C_p \cdot \begin{cases} n^{-1} \log^3 n, & p < 2. \\ n^{-1} \log^4 n, & p = 2. \\ n^{-\frac{2}{p}} \log^3 n, & p > 2. \end{cases}$$

- *If $\mathcal{H}_2(\mathcal{F}, \varepsilon, n) \leq d \log(D/\varepsilon)$, then with probability at least $1 - \delta$, for all $f, f' \in \mathcal{F}$,*

$$\|f - f'\|_{L_2(\mathcal{D})}^2 \leq 2d_S^2(f, f') + C \frac{d \log(en/d) + \log(\log n/\delta)}{n}.$$

Proof of Lemma 14. Using Lemma 8 and Lemma 9 from Rakhlin et al. (2017), it also holds that with probability at least $1 - 4\delta$, for all $f, f' \in \mathcal{F}$, we have

$$\|f - f'\|_{L_2(\mathcal{D})}^2 \leq 2d_S^2(f, f') + C(r^\star + \beta),$$

where $\beta = (\log(1/\delta) + \log \log n)/n$, and where $r^\star \leq d_2 \log(en/d_2)$ in the parametric case, and $r^\star \leq \mathfrak{R}_n^2(\mathcal{F}) \log^3(n)$ in the general case. The final result follows by applying the Rademacher complexity bounds from Lemma 13.

$\square$

Remark 4. *Technically, the result in Rakhlin et al. (2017) we appeal to in the proof above is stated for $[0, 1]$-valued classes, but it may be applied to our $[-1, +1]$-valued setting by shifting and rescaling the class $\mathcal{F}$ (i.e., invoking with $\mathcal{F}' := (\mathcal{F} + 1)/2$). We appeal to the same reasoning throughout this section, shifting regression targets in the same fashion when necessary.*

F.2 Skeleton Aggregation

Here we briefly describe the Skeleton Aggregation meta-algorithm for real-valued regression (Yang and Barron, 1999; Rakhlin et al., 2017). The setting is as follows: we receive n examples $S = (X_1, Y_1), \ldots, (X_n, Y_n) \in (\mathcal{X} \times \mathbb{R})^n$ i.i.d. from a distribution $\mathcal{D}$. For a function class $\mathcal{F} \subseteq (\mathcal{X} \to \mathbb{R})$, we define $L_{\mathcal{D}}(f) = \mathbb{E}_{\mathcal{D}}(f(X) - Y)^2$. Our goal is to produce a predictor $\widehat{f}_S$ for which the excess risk $L_{\mathcal{D}}(\widehat{f}) - \inf_{f \in \mathcal{F}} L_{\mathcal{D}}(f)$ is small.

We call a *sharp model selection aggregate* any algorithm that, given a finite collection of M functions $f_1, \ldots, f_M$ and n i.i.d. samples, returns a convex combination $\widehat{f} = \sum_{i=1}^{M} \nu_i f_i$ for which

$$L(\widehat{f}) \leq \min_{i \in [M]} L(f_i) + C \frac{\log(M/\delta)}{n} \tag{124}$$

with probability at least $1 - \delta$. One such model selection aggregate is the *star aggregation* strategy of Audibert (2008), which produces a 2-sparse convex combination $\widehat{f}$ with the property (124) whenever $|Y| \leq 1$ almost surely and the functions in $f_1, \ldots, f_M$ take values in $[-1, +1]$.

We use the following variant of skeleton aggregation, following Rakhlin et al. (2017). Given a dataset $S = (X_1, Y_1), \ldots, (X_n, Y_n)$, we split it into two equal-sized parts S' and S''.

- Fix a scale $\varepsilon > 0$, and let $N = \mathcal{H}(\mathcal{F}, \varepsilon, S')$. Let $\{\widehat{f}_i\}_{i \in [N]}$ be a collection of functions that realize the cover, and assume the cover is proper without loss of generality.[16]
- Let $\widehat{f}$ be the output of the star aggregation algorithm run with the collection $\{\widehat{f}_i\}_{i \in [N]}$ on the dataset S''.

For a simple analysis of this strategy, see Section 6 of Rakhlin et al. (2017). In general, the strategy is optimal only in the well-specified setting in which $\mathbb{E}[Y \mid X] = f^{\star}(X)$ for some $f^{\star} \in \mathcal{F}$. We give a more refined analysis in the presence of nuisance parameters in the sequel. As final remark, note that since we use a proper cover and the star aggregate is 2-sparse, the final predictor $\widehat{f}$ lies in the class $\mathcal{F}' \coloneqq \mathcal{F} + \mathrm{star}(\mathcal{F} - \mathcal{F}, 0)$

F.3 Rates for Specific Algorithms

Given an example $z \in \mathcal{Z}$, we define an auxiliary example $(\widetilde{X}, \widetilde{Y})$ via $\widetilde{X}(z) = w$, $\widetilde{Y}(z) = \Gamma(\widehat{g}_n(w), z)$. For the remainder of this section we make use of the auxiliary second-stage dataset $\widetilde{S}$ defined via

$$\widetilde{S} = \left\{ (\widetilde{X}(z), \widetilde{Y}(z)) \right\}_{z \in S^{(2)}}. \tag{125}$$

We make use of the following auxiliary predictor classes:

$$\mathcal{F}_0 = \left\{ \widetilde{X} \mapsto \langle \Lambda(g_0(w), w), \theta(x) \rangle \mid \theta \in \Theta_n \right\}, \tag{126}$$

$$\mathcal{F} = \left\{ \widetilde{X} \mapsto \langle \Lambda(\widehat{g}_n(w), w), \theta(x) \rangle \mid \theta \in \Theta_n \right\}. \tag{127}$$

Finally, we define $\widetilde{\ell}(\widehat{y}, y) = (\widehat{y} - y)^2$ and $\widetilde{L}(f) = \mathbb{E}_{\widetilde{X}, \widetilde{Y}} \widetilde{\ell}(f(\widetilde{X}), \widetilde{Y})$, where $(\widetilde{X}, \widetilde{Y})$ are sampled from the distribution introduced by drawing $z \sim \mathcal{D}$, and taking $(\widetilde{X}(z), \widetilde{Y}(z))$. With these definitions, observe that for any $f \in \widehat{F}$ of the form $\widetilde{X} \mapsto \langle \Lambda(\widehat{g}_n(w), w), \theta(x) \rangle$, we have

$$\widetilde{L}(f) = L_{\mathcal{D}}(\theta, \widehat{g}_n) \quad \text{and} \quad \inf_{f \in \mathcal{F}} \widetilde{L}(f) \leq L_{\mathcal{D}}(\theta_n^{\star}, \widehat{g}_n). \tag{128}$$

We relate the metric entropy of the auxiliary class $\widehat{\mathcal{F}}$ to that of Θ_n as follows.

---

[16] Any improper ε-cover can be made into a proper 2ε-cover.

70

Proposition 4. *Under Assumption 10, it holds that*

$$\mathcal{H}_2(\widehat{\mathcal{F}}, \varepsilon, n) \le \mathcal{H}_2(\widehat{\Theta}_n, \varepsilon, n). \tag{129}$$

Lemma 15 (Rates for first stage). *Suppose that Assumption 11 and Assumption 13 hold. Then Global ERM guarantees that with probability at least $1 - \delta$,*

$$\|\widehat{g}_n - g_0\|_{L_2(\ell_2, \mathcal{D})}^2 \le K^{(1)} \cdot \begin{cases} C \cdot \big(d_1 \log(en/d_1) + \log(\delta^{-1})\big) n^{-1}, & \textit{Parametric case.} \\ C_{p_1} \cdot n^{-\frac{2}{2+p_1}} + \log(\delta^{-1}) n^{-1}, & \textit{Nonparametric case, } p_1 < 2 \\ C\sqrt{(\log n + \log(\delta^{-1}))/n}, & \textit{Nonparametric case, } p_1 = 2 \\ C_{p_1} \cdot n^{-\frac{1}{p_1}} + \sqrt{\log(\delta^{-1})/n}, & \textit{Nonparametric case, } p_1 > 2 \end{cases}$$

Skeleton Aggregation guarantees that with probability at least $1 - \delta$,

$$\|\widehat{g}_n - g_0\|_{L_2(\ell_2, \mathcal{D})}^2 \le K^{(1)} \cdot \begin{cases} C \cdot \big(d_1 \log(en/d_1) + \log(\delta^{-1})\big) n^{-1}, & \textit{Parametric case.} \\ C_{p_1} \cdot n^{-\frac{2}{2+p_1}} + \log(\delta^{-1}) n^{-1}, & \textit{Nonparametric case.} \end{cases}$$

C_{p_1} is some constant that depends only on p_1.

Proof of Lemma 15. In what follows we analyze the algorithms under consideration for the class $\mathcal{G}_n|_i$ for a fixed coordinate i. The final result follows by union bounding over coordinates and summing the coordinate-wise error bounds we establish.

Global ERM. When we are either in the parametric case or the nonparametric case with $p_1 < 2$, the result is given by Theorem 5.2 of Koltchinskii (2011). See Example 3 and Example 4 that follow the theorem for precise calculations under these assumptions. See also Remark 4.

On the other hand, when $p_1 \ge 2$ we apply to the standard Rademacher complexity bound for ERM (e.g. Shalev-Shwartz and Ben-David (2014)), which states that with probability at least $1 - \delta$,

$$\mathbb{E}(\widehat{g}_n)_i(w) - g_0(w))^2 \le 2 \cdot \mathfrak{R}_{n/2}(\ell \circ \mathcal{G}_n|_i, S) + \sqrt{\frac{\log(1/\delta)}{n}}.$$

The result follows by applying Lipschitz contraction to the Rademacher complexity (using that the class is bounded) and appealing to the Rademacher complexity bounds from Lemma 13.

Skeleton Aggregation. We appeal to Section 6 of Rakhlin et al. (2017). See Remark 4. $\qquad\square$

Lemma 16. *Consider the plug-in global ERM strategy for the setting in Section 6.2, i.e.*

$$\widehat{\theta}_n \in \underset{\theta \in \Theta_n}{\arg\min} \sum_{z \in S^{(2)}} \ell(\theta, \widehat{g}_n ; z).$$

Under the assumptions of Theorem 7 and Theorem 8, global ERM guarantees

$$L_{\mathcal{D}}(\widehat{\theta}_n, \widehat{g}_n) - L_{\mathcal{D}}(\theta_n^\star, \widehat{g}_n) \le C \cdot \begin{cases} \big(d_2 \log(en/d_2) + \log(\delta^{-1})\big) n^{-1}, & \textit{Parametric case.} \\ C_{p_2} \cdot n^{-\frac{2}{2+p_2}} + \log(\delta^{-1}) n^{-1}, & \textit{Nonparametric case, } p_2 < 2. \\ \sqrt{(\log n + \log(\delta^{-1}))/n}, & \textit{Nonparametric case, } p_2 = 2. \\ C_{p_2} \cdot n^{-\frac{1}{p_2}} + \sqrt{\log(\delta^{-1})/n}, & \textit{Nonparametric case, } p_2 > 2. \end{cases}$$

Proof of Lemma 16. Let $\widehat{f} = \widetilde{X} \mapsto \langle \Lambda(\widehat{g}_n(w), w), \widehat{\theta}_n(x) \rangle$. Observe that we can write $\widehat{f}$ as the global ERM for the auxiliary dataset $\widetilde{S}$:

$$\widetilde{f} \in \arg\min_{f \in \widehat{F}} \sum_{(\widetilde{X}, \widetilde{Y}) \in \widetilde{S}} \widetilde{\ell}(f(\widetilde{X}), \widetilde{Y}).$$

Case $p_2 < 2$. In the misspecified case we appeal to Theorem 5.1 in Koltchinskii (2011), using that $\widehat{f}$ is the global ERM for the class $\widehat{\mathcal{F}}$. To invoke the theorem, we verify that a) $\widehat{\mathcal{F}}$ takes values in $[-1, +1]$ under Assumption 10,[17] b) $\widehat{\mathcal{F}}$ inherits convexity from Θ_n, and c) $\mathcal{H}_2(\widehat{\mathcal{F}}, \varepsilon, n) \leq \mathcal{H}_2(\Theta_n, \varepsilon, n)$, following Proposition 4. The theorem (see also the following discussion in Example 3 and Example 4) therefore guarantees that with probability at least $1 - \delta$,

$$\widetilde{L}(\widehat{f}) - \inf_{f \in \widehat{\mathcal{F}}} \widetilde{L}(f) \leq C \cdot \begin{cases} n^{-1}\big(d_2 \log(en/d_2) + \log(\delta^{-1})\big), & \text{Parametric case.} \\ C_{p_2} \cdot n^{-\frac{2}{2+p_2}} + n^{-1}\log(\delta^{-1}), & \text{Nonparametric case, } p_1 < 2. \end{cases}$$

The result now follows from (128), in particular that $\widetilde{L}(\widehat{f}) = L_{\mathcal{D}}(\widehat{\theta}_n, \widehat{g}_n)$.

Case $p_2 \geq 2$. We apply the standard Rademacher complexity bound for ERM (e.g. Shalev-Shwartz and Ben-David (2014)), which states that with probability at least $1 - \delta$,

$$\widetilde{L}(\widehat{f}) - \inf_{f \in \widehat{\mathcal{F}}} \widetilde{L}(f) \leq 2 \cdot \mathfrak{R}_{n/2}(\widetilde{\ell} \circ \widehat{\mathcal{F}}) + C\sqrt{\frac{\log(1/\delta)}{n}}$$

$$\leq C' \cdot \mathfrak{R}_{n/2}(\widehat{\mathcal{F}}) + C\sqrt{\frac{\log(1/\delta)}{n}},$$

where we have applied Lipschitz contraction to the Rademacher complexity (using that the class is bounded). To complete the result, we use that $\mathcal{H}_2(\widehat{\mathcal{F}}, \varepsilon, n) \leq \mathcal{H}_2(\Theta_n, \varepsilon, n)$ and appeal to the Rademacher complexity bound from Lemma 13. $\square$

Lemma 17. *Consider the following Skeleton Aggregation variant:*[18]
- *Split $S^{(2)}$ into equal-sized subsets S' and S''.*
- *Fix a scale $\varepsilon > 0$, and let $N = \mathcal{H}_2(\Theta_n, \varepsilon, S')$. Let $\{\theta_i\}_{i \in [N]}$ be a collection of functions that realize the cover, and assume the cover is proper without loss of generality. Define $f_i = \widetilde{X} \mapsto \langle \Lambda(\widehat{g}_n(w), w), \theta_i(x) \rangle$ for each $i \in [N]$.*
- *Let $\widehat{\theta}_n \in \Theta_n + \mathrm{star}(\Theta_n - \Theta_n, 0) =: \widehat{\Theta}_n$ realize the output of the star aggregation algorithm using the function class $\{\widehat{f}_i\}_{i \in [N]}$ on the dataset $\{(\widetilde{X}(z), \widetilde{Y}(z))\}_{z \in S''}$.*

Under the assumptions of Theorem 7, when the model is well-specified, Skeleton Aggregation guarantees that with probability at least $1 - \delta$,

$$L_{\mathcal{D}}(\widehat{\theta}_n, \widehat{g}_n) - L_{\mathcal{D}}(\theta_n^\star, \widehat{g}_n)$$

$$\leq C \cdot \|\widehat{g}_n - g_0\|_{L_4(\ell_2, \mathcal{D})}^4 + C' \cdot \begin{cases} \big(d_2 \log(en/d_2) + K^{(2)} \log(K^{(2)} \log n/\delta)\big)n^{-1}, & \text{Parametric case,} \\ C_{p_2} \cdot n^{-\frac{2}{2+p_2}} + K^{(2)} \log(K^{(2)} \log n/\delta)n^{-1}, & \text{Nonparametric case,} \end{cases}$$

so long as $K^{(2)} = o\big(n^{\frac{p_2}{2+p_2} \wedge \frac{4}{p_2(2+p_2)}}\big)$.

---

[17]See Remark 4.

[18]See Section F.2 for background.

Proof of Lemma 17. Let $\widehat{\mathcal{F}} = \{f_i\}_{i \in [N]}$. The Skeleton Aggregation algorithm as described outputs a predictor $\widehat{f} \in \widehat{\mathcal{F}} + \mathrm{star}(\widehat{\mathcal{F}} - \widehat{\mathcal{F}}, 0)$ (see Section F.2) such that

$$\widetilde{L}(\widehat{f}) \leq \min_{i \in [N]} \widetilde{L}(\widehat{f}_i) + C\left(\frac{\log N}{n} + \frac{\log(1/\delta)}{n}\right) \leq \min_{i \in [N]} \widetilde{L}(\widehat{f}_i) + C\left(\frac{\mathcal{H}_2(\Theta_n, \varepsilon, S)}{n} + \frac{\log(1/\delta)}{n}\right).$$

Translating back into the language of the lemma statement, recall that we can express each f_i via $f_i = X \mapsto \langle \Lambda(\widehat{g}_n(w), w), \theta_i(x) \rangle$, with $\{\theta_i\}_{i \in [N]} \subset \Theta_n$ since we have assumed a proper cover. Since this parameterization is linear in θ, there must be some $\widehat{\theta}_n \in \Theta_n + \mathrm{star}(\Theta_n - \Theta_n, 0)$ that realizes $\widehat{f}$. Using the expression for the risk in (128), this implies

$$L_{\mathcal{D}}(\widehat{\theta}_n, \widehat{g}_n) \leq \min_{i \in [N]} L_{\mathcal{D}}(\theta_i, \widehat{g}_n) + C\left(\frac{\mathcal{H}_2(\Theta_n, \varepsilon, S)}{n} + \frac{\log(1/\delta)}{n}\right).$$

Adding and subtracting from both sides, we rewrite the inequality as

$$L_{\mathcal{D}}(\widehat{\theta}_n, \widehat{g}_n) - L_{\mathcal{D}}(\theta_0, \widehat{g}_n) \leq \min_{i \in [N]} L_{\mathcal{D}}(\theta_i, \widehat{g}_n) - L_{\mathcal{D}}(\theta_0, \widehat{g}_n) + C\left(\frac{\mathcal{H}_2(\Theta_n, \varepsilon, S)}{n} + \frac{\log(1/\delta)}{n}\right).$$

The idea going forward is to use orthogonality to bound the approximation error on the right-hand-side. Let i be fixed. Using a second-order Taylor expansion there exists $\bar{\theta} \in \mathrm{star}(\Theta_n, \theta_0)$ such that

$$L_{\mathcal{D}}(\theta_i, \widehat{g}_n) - L_{\mathcal{D}}(\theta_0, \widehat{g}_n) = D_\theta L_{\mathcal{D}}(\theta_0, \widehat{g}_n)[\theta_i - \theta_0] + \frac{1}{2} D_\theta^2 L_{\mathcal{D}}(\bar{\theta}, \widehat{g}_n)[\theta_i - \theta_0, \theta_i - \theta_0].$$

Using another second-order Taylor expansion, there exists $\bar{g} \in \mathrm{star}(\mathcal{G}_n, g_0)$ for which

$$D_\theta L_{\mathcal{D}}(\theta_0, \widehat{g}_n)[\theta_i - \theta_0]$$
$$= D_\theta L_{\mathcal{D}}(\theta_0, g_0)[\theta_i - \theta_0] + D_g D_\theta L_{\mathcal{D}}(\theta_0, g_0)[\theta_i - \theta_0, \widehat{g}_n - g_0] + D_g^2 D_\theta L_{\mathcal{D}}(\theta_0, \bar{g})[\theta_i - \theta_0, \widehat{g}_n - g_0, \widehat{g}_n - g_0].$$

The model is well-specified and we have assumed orthogonality, so the first two terms on the right-hand-side vanish. Let $\|\theta - \theta'\|_{\Theta_n}^2 = \mathbb{E}\left[\langle \Lambda(g_0(w), w), \theta(x) - \theta'(x) \rangle^2\right]$. Using Lemma 3, we have

$$D_\theta^2 L_{\mathcal{D}}(\bar{\theta}, \widehat{g}_n)[\theta_i - \theta_0, \theta_i - \theta_0] = C \cdot \|\theta_i - \theta_0\|_{\Theta_n}^2 + C' \cdot \|\widehat{g}_n - g_0\|_{L_4(\ell_2, \mathcal{D})}^4.$$

Next, using (79), we have

$$D_g^2 D_\theta L_{\mathcal{D}}(\theta_0, \bar{g})[\theta_i - \theta_0, \widehat{g}_n - g_0, \widehat{g}_n - g_0] \leq C \cdot \|\theta_i - \theta_0\|_{\Theta_n} \cdot \|\widehat{g}_n - g_0\|_{L_4(\ell_2, \mathcal{D})}^2 \leq C \cdot \left(\|\theta_i - \theta_0\|_{\Theta_n}^2 + \|\widehat{g}_n - g_0\|_{L_4(\ell_2, \mathcal{D})}^4\right).$$

Furthermore, Assumption 10 implies that $\|\theta_i - \theta_0\|_{\Theta_n} \leq \|\theta_i - \theta_0\|_{L_2(\ell, \mathcal{D})}$. Plugging these bounds back into the excess risk guarantee, we have that there are constants C, C', C'' such that

$$L_{\mathcal{D}}(\widehat{\theta}_n, \widehat{g}_n) - L_{\mathcal{D}}(\theta_0, \widehat{g}_n) \leq C \cdot \min_{i \in [N]} \|\theta_i - \theta_0\|_{L_2(\ell_2, \mathcal{D})}^2 + C'\left(\frac{\mathcal{H}_2(\mathcal{F}, \varepsilon, S)}{n} + \frac{\log(1/\delta)}{n}\right) + C'' \cdot \|\widehat{g}_n - g_0\|_{L_4(\ell_2, \mathcal{D})}^4.$$

We now invoke Lemma 14 for each of the $K^{(2)}$ output coordinates of the target space separately and union bound, which implies that with probability at least $1 - \delta$ over the draw of S',

$$\|\theta_i - \theta_0\|_{L_2(\ell_2, \mathcal{D})}^2 \leq 2d_{2, S'}^2(\theta_i, \theta_0) + U,$$

73

where $U \le K^{(2)} R_{n,p_2} + C \frac{K^{(2)} \log(K^{(2)} \log n/\delta)}{n}$ in the nonparametric case, and $U \le C K^{(2)} \frac{d_2 \log(en/d_2)}{n} + C' \frac{K^{(2)} \log(K^{(2)} \log n/\delta)}{n}$ in the parametric case. Returning to the excess risk, this implies

$$L_{\mathcal{D}}(\widehat{\theta}_n, \widehat{g}_n) - L_{\mathcal{D}}(\theta_0, \widehat{g}_n) \le C \cdot \min_{i \in [N]} d_{S'}^2(\theta_i, \theta_0) + C' \left(\frac{\mathcal{H}_2(\mathcal{F}, \varepsilon, S)}{n} + \frac{\log(1/\delta)}{n} \right) + U + C'' \cdot \|\widehat{g}_n - g_0\|_{L_4(\ell_2, \mathcal{D})}^4.$$

The cover property implies that $\min_{i \in [N]} d_{2,S'}^2(\theta_i, \theta_0) \le \varepsilon^2$. So we are left with

$$L_{\mathcal{D}}(\widehat{\theta}_n, \widehat{g}_n) - L_{\mathcal{D}}(\theta_0, \widehat{g}_n) \le C\varepsilon^2 + C' \left(\frac{\mathcal{H}_2(\mathcal{F}, \varepsilon, S)}{n} + \frac{\log(1/\delta)}{n} \right) + U + C'' \cdot \|\widehat{g}_n - g_0\|_{L_4(\ell_2, \mathcal{D})}^4.$$

Under Assumption 14, solving for the balance $\varepsilon^2 \asymp \frac{\mathcal{H}_2(\mathcal{F}, \varepsilon, S)}{n}$, leads the first two terms to be of order $d_2 \log(en/d_2)$ in the parametric case and $n^{-\frac{2}{2+p_2}}$ in the nonparametric case. Thus, in the parametric case, the term U dominates and the final bound is $C K^{(2)} \frac{d_2 \log(en/d_2)}{n} + C' \frac{K^{(2)} \log(K^{(2)} \log n/\delta)}{n}$. In the nonparametric case, our assumption on the growth of $K^{(2)}$ implies that U is of lower order.

$\qquad\qquad\square$

F.4 Proofs for Oracle Rates

Proof of Theorem 7. First we invoke Lemma 15, which along with Assumption 12 implies that depending on whether $p_1 > 2$, one of either global ERM or skeleton aggregation guarantees that with probability at least $1 - \delta$,

$$\|\widehat{g}_n - g_0\|_{L_4(\ell_2, \mathcal{D})}^2 \le \left(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta) \right)^2 \le \widetilde{O}\left(C_{2 \to 4}^2 \cdot n^{-\frac{2}{2+p_1}} \right).$$

Observe that the assumption that the model is well-specified at (g_0, θ_0) implies that Assumption 2 is satisfied. We invoke Assumption 7 and Theorem 1, which guarantees

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0) \le C \cdot \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta \,;\, \widehat{g}_n) + C'' \left(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta) \right)^4.$$

We now invoke Lemma 17 using that the model is assumed to be well-specified, which implies that with probability at least $1 - \delta$, Skeleton Aggregation enjoys

$$\mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta \,;\, \widehat{g}_n) \le \widetilde{O}\left(n^{-\frac{2}{2+p_2}} \right) + C' \left(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta) \right)^4.$$

Combining these results, we get

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0) \le \widetilde{O}\left(n^{-\frac{2}{2+p_2}} + C_{2 \to 4}^4 \cdot n^{-\frac{4}{2+p_1}} \right).$$

The final result follows by setting p_1 to guarantee that the first term dominates this expression.

We mention in passing that to show that global ERM achieves the desired rate for stage two when $p_2 \le 2$, one can appeal to the rates in Appendix E. $\qquad\square$

Proof of Theorem 8. As in the proof of Theorem 7, we invoke Lemma 15, which implies that Skeleton Aggregation[19] guarantees that with probability at least $1 - \delta$,

$$\|\widehat{g}_n - g_0\|_{L_2(\ell_2, \mathcal{D})}^2 \le C_{2 \to 4}^2 \left(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta) \right)^2 \le \widetilde{O}\left(C_{2 \to 4}^2 \cdot n^{-\frac{2}{2+p_1}} \right).$$

---

[19]Alternatively, global ERM can be applied when $p_1 \le 2$.

Observe that since Θ_n is convex, Assumption 2 is satisfied, and so we can invoke Assumption 7 and Theorem 1 to get

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0) \leq C \cdot \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta\,;\widehat{g}_n) + C''\Big(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta)\Big)^4.$$

We use global ERM for the second stage. Lemma 16 implies that since the class Θ_n is convex, with probability at least $1 - \delta$, global ERM guarantees

$$\mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta\,;\widehat{g}_n) \leq \widetilde{O}\Big(n^{-\frac{2}{2+p_2} \wedge \frac{1}{p_2}}\Big).$$

This leads to a final guarantee of

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0) \leq \widetilde{O}\Big(n^{-\frac{2}{2+p_2} \wedge \frac{1}{p_2}} + C_{2\to4}^4 \cdot n^{-\frac{4}{2+p_1}}\Big),$$

and the stated result follows by setting p_1 to guarantee that the first term dominates this expression. $\qquad\square$

Proof of Theorem 9. Lemma 15 implies that either Skeleton Aggregation or global ERM (for $p_1 \leq 2$) guarantees that with probability at least $1 - \delta$,

$$\|\widehat{g}_n - g_0\|_{L_2(\ell_2, \mathcal{D})}^2 = \Big(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta)\Big)^2 \leq \widetilde{O}\Big(n^{-\frac{2}{2+p_1}}\Big).$$

Since Assumption 8 is satisfied, we appeal to Lemma 2, which guarantees

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0) \leq C \cdot \mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta\,;\widehat{g}_n) + C''\Big(\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}_n, S^{(1)}, \delta)\Big)^2.$$

We use global ERM for the second stage. The standard Rademacher complexity bound for ERM (e.g. Shalev-Shwartz and Ben-David (2014)), states that with probability at least $1 - \delta$, the excess risk is bounded by the Rademacher complexity of the target class Θ_n composed with the loss class as follows:

$$L_{\mathcal{D}}(\widehat{\theta}_n, \widehat{g}_n) - L_{\mathcal{D}}(\theta_n^\star, \widehat{g}_n) \leq 2 \cdot \underbrace{\mathbb{E}_\epsilon \sup_{\theta \in \Theta_n} \frac{2}{n} \sum_{t=1}^{n/2} \epsilon_t \ell(\theta(x_t), \widehat{g}_n(w_t), z_t)}_{=:\,\mathfrak{R}_{n/2}(\ell \circ \Theta_n, S^{(2)})} + C\sqrt{\frac{\log(1/\delta)}{n}}.$$

Using Lemma 13 and boundedness of the loss, we have

$$\mathfrak{R}_{n/2}(\ell \circ \Theta_n, S^{(2)}) \leq C \cdot \inf_{\alpha > 0}\left\{\alpha + \int_\alpha^1 \sqrt{\frac{\mathcal{H}_2(\ell \circ \Theta_n, \varepsilon, S^{(2)})}{n}}d\varepsilon\right\}.$$

Since the loss is 1-Lipschitz with respect to ℓ_2, we have $\mathcal{H}_2(\ell \circ \Theta_n, \varepsilon, S^{(2)}) \leq \mathcal{H}_2(\Theta_n, \varepsilon, S^{(2)})$. Under the assumed growth of $\mathcal{H}_2(\Theta_n, \varepsilon, S^{(2)}) \propto \varepsilon^{-p_2}$, this gives

$$\mathrm{Rate}_{\mathcal{D}}(\Theta_n, S^{(2)}, \delta\,;\widehat{g}_n) \leq \widetilde{O}\Big(n^{-\frac{1}{2} \wedge \frac{1}{p_2}}\Big).$$

This leads to a final guarantee of

$$L_{\mathcal{D}}(\widehat{\theta}_n, g_0) - L_{\mathcal{D}}(\theta_n^\star, g_0) \leq \widetilde{O}\Big(n^{-\frac{1}{2} \wedge \frac{1}{p_2}} + n^{-\frac{2}{2+p_1}} \cdot\Big),$$

The theorem statement follows by setting p_1 so that the first term dominates.

$\qquad\square$

G Proofs from Section 8

Orthogonality for Treatment Effects Example (Section 8.1). We first show that Assumption 2 holds whenever the second stage is well-specified (i.e. $\theta_0 \in \Theta_n$). We have

$$D_\theta L_\mathcal{D}(\theta_0, g_0)[\theta - \theta_0] = -2 \cdot \mathbb{E}[((Y - m_0(X, W)) - (T - e_0(W))\theta_0(X))(T - e_0(W))(\theta(X) - \theta_0(X))].$$

In particular, for any x we have

$$\mathbb{E}[(Y - m_0(X, W)) - (T - e_0(W))\theta_0(X))(T - e_0(W)) \mid X = x]$$
$$= \mathbb{E}[(T\theta_0(W) + f_0(W) + \varepsilon_1 - e_0(X)\theta_0(W) - f_0(W)) - (T - e_0(W))\theta_0(X))(T - e_0(W)) \mid X = x]$$
$$= \mathbb{E}[\varepsilon_1 \cdot \varepsilon_2 \mid X = x] = \mathbb{E}[\mathbb{E}[\varepsilon_1 \mid X = x, \varepsilon_2] \cdot \varepsilon_2 \mid X = x] = 0,$$

and so $D_\theta L_\mathcal{D}(\theta_0, g_0)[\theta - \theta_0] = 0$. To establish orthogonality for the propensities e, let θ, θ' be fixed. Then we have

$$D_e D_\theta L_\mathcal{D}(\theta, \{m_0, e_0\})[\theta' - \theta, e - e_0]$$
$$= 2 \cdot \mathbb{E}[((Y - m_0(X, W)) - (T - e_0(W))\theta(X))(\theta'(X) - \theta(X))(e(W) - e_0(W))]$$
$$- 2 \cdot \mathbb{E}[(\theta(X)(T - e_0(W))(\theta'(X) - \theta(X))(e(W) - e_0(W)))].$$

To handle the first term, we use that for any x, w,

$$\mathbb{E}[((Y - m_0(X, W)) - (T - e_0(W))\theta(X)) \mid X = x, W = w]$$
$$= \mathbb{E}[\varepsilon_1 + \varepsilon_2 \cdot (\theta_0(X) - \theta(X)) \mid X = x, W = w] = 0.$$

Similarly, the second term is handled by using that

$$\mathbb{E}[\theta(X)(T - e_0(W)) \mid X = x, W = w] = \mathbb{E}[\theta(X) \cdot \varepsilon_2 \mid X = x, W = w] = 0.$$

To establish orthogonality for the expected value parameter m, for any θ, θ' we have

$$D_m D_\theta L_\mathcal{D}(\theta, \{m_0, e_0\})[\theta' - \theta, m - m_0]$$
$$= 2 \cdot \mathbb{E}[(T - e_0(W))(\theta'(X) - \theta(X))(m(X, W) - m_0(X, W))]$$
$$= 2 \cdot \mathbb{E}[\varepsilon_2 \cdot (\theta'(X) - \theta(X))(m(X, W) - m_0(X, W))]$$
$$= 0,$$

which follows from the assumption $\mathbb{E}[\varepsilon_2 \mid X, W] = 0$. Note that both of these orthogonality proofs held for any choice of θ, not just θ_0, and hence universal orthogonality holds.

$\qquad\square$

Orthogonality for Missing Data Example (Section 8.4). We first show that the gradient vanishes in the sense of Assumption 2 when evaluated at θ_0. In particular, for any choice of h we have

$$D_\theta L_\mathcal{D}(\theta_0, \{h, e_0\})[\theta - \theta_0]$$
$$= \mathbb{E}\left[\left(-2T\frac{(\widetilde{Y} - \theta_0(X))}{e_0(W)} - h(W)(T - e_0(W))\right)(\theta(X) - \theta_0(X))\right].$$

Using that $\mathbb{E}[T \mid W] = e_0(W)$ and $X \subseteq W$:

$$= \mathbb{E}\left[\left(-2T\frac{(\widetilde{Y} - \theta_0(X))}{e_0(W)}\right)(\theta(X) - \theta_0(X))\right].$$

Using that $T \in \{0, 1\}$:

$$= \mathbb{E}\left[\left(-2T\frac{(Y - \theta_0(X))}{e_0(W)}\right)(\theta(X) - \theta_0(X))\right].$$

Using that $T \perp Y \mid W$ and that $X \subseteq W$:

$$= \mathbb{E}_W\left[\left(-2\,\mathbb{E}[T \mid W]\frac{\mathbb{E}[(Y - \theta_0(X)) \mid W]}{e_0(W)}\right)(\theta(X) - \theta_0(X))\right]$$
$$= -2\,\mathbb{E}[(Y - \theta_0(X))(\theta(X) - \theta_0(X))].$$

Using that $\mathbb{E}[Y \mid X] = \theta_0(X)$:

$$= 0.$$

To establish orthogonality with respect to e, we have

$$D_e D_\theta L_\mathcal{D}(\theta_0, \{h_0, e_0\})[\theta - \theta_0, e - e_0]$$
$$= \mathbb{E}\left[\left(2T\frac{(Y - \theta_0(X))}{e_0^2(W)} + h_0(W)\right)(\theta(X) - \theta_0(X))(e(W) - e_0(W))\right].$$

Using that $X \subseteq W$:

$$= \mathbb{E}_W\left[\left(2\frac{(\mathbb{E}[Y \mid W] - \theta_0(X))}{e_0(W)} + h_0(W)\right)(\theta(X) - \theta_0(X))(e(W) - e_0(W))\right].$$

The result follows immediately, using that $h_0(w) = -2\frac{(\mathbb{E}[Y \mid W = w] - \theta_0(x))}{e_0(w)}$.

$$= 0.$$

To establish orthogonality with respect to h, we have

$$D_h D_\theta L_\mathcal{D}(\theta_0, \{h_0, e_0\})[\theta - \theta_0, h - h_0]$$
$$= \mathbb{E}[(T - e_0(W))(\theta(X) - \theta_0(X))(h(W) - h_0(W))]$$
$$= \mathbb{E}[\varepsilon_2 \cdot (\theta(X) - \theta_0(X))(h(W) - h_0(W))].$$

Using that $\mathbb{E}[\varepsilon_2 \mid W] = 0$ and $X \subseteq W$:

$$= 0.$$

$\square$