Minimax-inspired Semiparametric Estimation and Causal Inference

David A. Hirshberg

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

Minimax-inspired Semiparametric Estimation and Causal Inference

David A. Hirshberg

This thesis focuses on estimation and inference for a large class of semiparametric estimands: the class of continuous functionals of regression functions. This class includes a number of estimands derived from causal inference problems, among then the average treatment effect for a binary treatment when treatment assignment is unconfounded and many of its generalizations for non-binary treatments and individualized treatment policies.

Chapter 2, based on work with Stefan Wager, introduces the augmented minimax linear estimator (AMLE), a general approach to the problem of estimating a continuous linear functional of a regression function. In this approach, we estimate the regression function, then subtract from a simple plug-in estimator of the functional a weighted combination of the estimated regression function's residuals. For this, we use weights chosen to minimize the maximum of the mean squared error of the resulting estimator over regression functions in a chosen neighborhood of our estimated regression function. These weights are shown to be a universally consistent estimator our linear functional's Riesz representer, the use of which would result in an exact bias correction for our plug-in estimator. While this convergence can be slow, especially when the Riesz representer is highly nonsmooth, the action of these weights on functions in the aforementioned neighborhood imitates that of the Riesz representer accurately even when they are slow to converge in other respects. As a result, we show that under no regularity conditions on the Riesz representer and minimal regularity conditions on the regression function, the proposed estimator is semiparametrically efficient. In simulation, it is shown to perform very well in the context of estimating the average partial effect in the conditional linear model, a simultaneous generalization of the average treatment effect to address continuous-valued treatments and of the partial linear model to address treatment effect heterogeneity.

Chapter 3, based on work with Arian Maleki and José Zubizarreta, studies the minimax linear estimator, a simplified version of the AMLE in which the estimated regression function is taken to be zero, for a class of estimands generalizing the mean with outcomes missing at random. We show semiparametric efficiency under conditions that are only slightly stronger than those required for the AMLE. In addition, we bound the deviation of our estimator's error from the averaged efficient influence function, characterizing the degree to which the first order asymptotic characterization of semiparametric efficiency is meaningful in finite samples. In simulation, this estimator is shown to perform well relative to alternatives in high-noise, small-sample settings with limited overlap

between the covariate distribution of missing and nonmissing units, a setting that is challenging for approaches reliant on accurate estimation of either or both of the regression function and the propensity score.

Chapter 4 discusses an approach to rounding linear estimators for the targeted average treatment effect into matching estimators. The targeted average treatment effect is a generalization of the average treatment effect and the average treatment effect on the treated units.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank José Zubizarreta for introducing me to Causal Inference as a field and as a community; for his guidance, collaboration, and friendship during my time in graduate school; and for his generosity with his time and ideas. He never gave me his second-best project. Had he not invited me as an inexperienced student to join him in a promising collaboration with Arian Maleki on minimax interpretations of treatment effect estimators, I could not have written this thesis.

I would also like to thank Arian, from whom I learned the basics in my first year as a graduate student and have continued to learn as we've worked together since. His advice has been invaluable throughout my time at Columbia, both in the context of our work together and more broadly as I've tried to develop an understanding of a diverse and at times bewildering literature.

I am grateful to Stefan Wager, whose intuition and knack for explaining problems simply has been immensely helpful on the projects we've worked together on over the last few years and in my attempts to think about new problems. I am also grateful to David Madigan and Michael Sobel for thoughtful comments on my work and on what needs doing in Causal Inference over the years, and to Whitney Newey, from whom I've learned much in recent discussions.

I feel lucky to have been a part of the Statistics Department at Columbia and am grateful to everyone there who has taught me, helped me, or worked with me over the last five years, especially Bodhi Sen, Zhiliang Ying, Ming Yuan, Dave Blei, Andrew Gelman, Tian Zheng, Victor de la Peña, Dood Kalicharan, Anthony Cruz, Rohit Patra and Susanna Makela. I also had the good fortune to spend a semester at the Department of Health Care Policy at Harvard Medical School, and I am grateful to José for inviting me and supporting me during my time there, to Cynthia Hobbs and others there who made that possible, and to students and faculty members I met at Methods Happy Hour and elsewhere at HCP for making me feel welcome.

I would like to thank Michael Black, who taught me how to write and sell a paper while employing me as a computer vision researcher in his group before my time at Columbia, first at Brown and later at MPI Tübingen. Also Alex Weiss, Matt Loper, Eric Rachlin, Aggeliki Tsoli, and other folks from Michael's group who I shared authorship, code, and lunch with during those years.

Finally, I'd like to thank the friends and roommates I spent much of my time with while I was in New York, Sonia and Tanya Saraiya, Leah Shabshelowitz, Ilana Rothkopf, Tessa Conrad, and Tommy Wu; the Rhode Island and NYC Climbing Communities; my girlfriend Anne Jonas; and my parents Debbe Fate and Larry Hirshberg, my sister Shira Hirshberg, and my brother Dave Stuebe.

To my parents and my sister.

Chapter One

Introduction

1.1 Observational Studies and Causality

Randomized experiments are the gold standard for comparing treatments and pervasive in many disciplines: medicine, public health, economics, and psychology among them (Hernán and Robins, 2015; Imbens and Rubin, 2015; Rosenbaum, 2002). But often we are interested in comparing treatments we cannot randomize because it is either infeasible or unethical, for example when the comparison is between home and hospital birth (Daysal et al., 2016) or when we are interested in adverse effects of a medication that has already become available (Bernardo et al., 2011). Considering the former, we could compare the rate of infant mortality for babies born at home to that of babies born in hospitals, but the result will be difficult to interpret, as association is not causation. Taking a difference of these rates would give us a mixture of the effect of the difference in birth setting and the baseline difference in infant mortality between those who elect home birth and those who elect hospital birth. These groups are systematically different, so it is reasonable to expect a baseline difference, possibly due to age, diet, exposure to toxins, etc. Because of this, the observed difference in infant mortality rates is not necessarily predictive of how the aggregate infant mortality rate would change if home birth became more (or less) popular. If we are able to disentangle the effect of the difference in birth setting from these baseline differences, we have a stronger basis for individual decisionmaking, advocacy, and policy.

However, with rare exceptions, we need to make strong untestable assumptions to disentangle causal effects like this one from baseline differences in observational studies. To discuss these, we need language we can use to define our assumptions. Working with such language has allowed scholars to more precisely define and compare estimands and evaluate estimators (Sobel, 2000). In this dissertation, I will use the framework of *potential outcomes*, in which we imagine that for each study participant i and treatment of interest $w \in \mathcal{W}$, there is an outcome $Y_i^{(w)}$ that would have happened if, possibly contrary to fact, participant i had received treatment w (Neyman, 1923; Rubin, 1974). The vector all potential outcomes will be written $Y_i^{\mathcal{W}}$ and the index set $\mathcal{W}$ can be arbitrary.

Categorical indices w may represent totally distinct treatments or discrete levels of a treatment dose, continuous indices w may represent doses, mixed categorical/continuous indices may represent doses of distinct treatments, etc. We observe only W_i, indicating which treatment occurred; $Y_i = Y_i^{(W_i)}$, the outcome under that treatment; and X_i, a vector of *covariates* describing the participant. I will also assume that our participants are independently drawn from an infinite superpopulation, i.e. the complete data vectors $(X_i, W_i, Y_i^{\mathcal{W}})$ are independent and identically distributed random variables from some distribution P. In these terms, we define the comparison we want to make, or *causal estimand*, as a function of this distribution. When the treatment w is binary, two of the most common are the average treatment effect (ATE), $\tau := \mathbb{E}[Y_i^{(1)}] - \mathbb{E}[Y_i^{(0)}]$, and the average treatment effect on the treated (ATT), $\tau_T := \mathbb{E}[Y_i^{(1)} \mid W_i = 1] - \mathbb{E}[Y_i^{(0)} \mid W_i = 1]$. Each of these estimands is a difference in potential outcomes under the two treatments averaged; the ATT is an average over the distribution of all the units in our study whereas the ATT is an average over that of the units that receive treatment.

The assumptions we feel most comfortable making dictate our *identification strategy*, the means by which we convert a causal estimand defined in terms of unobserved potential outcomes, into a *statistical estimand*, defined in terms of the distribution of the observed data. I will focus on the assumption that no unobserved factor influences both the selection of treatment and the vector of potential outcomes, often called *unconfounded treatment assignment*, *strong ignorability*, *selection on observables*, or *exogenous noise*. This assumption justifies estimation of causal effects by adjusting for the observed covariates X. It is rare that any identifying assumptions are satisfied in practice; for that reason once we've estimated the statistical estimand, we study how it can deviate from the causal estimand as a function of the degree to which our identifying assumptions are violated. This last step, called *sensitivity analysis*, allows in some cases a quantitative defense against arguments that 'an association does not imply causation.' This was used convincingly by Cornfield in 1959 when prominent statisticians questioned the causal relationship between cigarettes and lung cancer on that basis (Cornfield et al., 1959).

1.2 Causal Estimands and Identification

In this section, we will consider a simple causal estimand, the ATT τ_T. First note that $Y_i^{(1)} = Y_i$ when $W_i = 1$, so the first term in τ_T is simply $\mathbb{E}[Y_i \mid W_i = 1]$ and can be estimated by a sample mean. The latter term, $\mathbb{E}[Y_i^{(0)} \mid W_i = 1]$, is what requires our attention. We make the simplest nontrivial assumption that allows us to estimate it, the unconfoundedness assumption that $\mathbb{E}[Y_i^{(0)} \mid$

$X_i, W_i = 1] = \mathbb{E}[Y_i^{(0)} \mid X_i, W_i = 0]$. Then,

$$\begin{aligned}
\mathbb{E}\left[Y_i^{(0)} \mid W_i = 1\right] &= \mathbb{E}\left[\mathbb{E}\left[Y_i^{(0)} \mid X_i, W_i = 1\right] \mid W_i = 1\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i^{(0)} \mid X_i, W_i = 0\right] \mid W_i = 1\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i \mid X_i, W_i = 0\right] \mid W_i = 1\right].
\end{aligned} \tag{1.1}$$

In order to ensure these expectations are defined, we make a so-called positivity assumption, $P\{W_i = 1 \mid X_i\} > 0$. We call the last expression in (1.1), which is defined in terms of the distribution of the observed data, the statistical estimand. It is a scalar-valued function, henceforth *functional*, of the conditional mean function $m(x, w) = \mathbb{E}[Y_i \mid X_i = x, W_i = w]$. In particular, we have $\mathbb{E}[Y_i^{(0)} \mid W_i = 1] = \psi(m)$ where $\psi(f) = \mathbb{E}[f(X_i, 0) \mid W_i = 1]$ is a linear functional. Many of the simpler causal estimands, including all of those that we will discussed in this dissertation, can be identified as linear functionals of conditional mean functions using essentially this argument. The appropriate unconfoundedness assumption varies slightly depending on the estimand, but the core notion at play is that given covariates, the outcome $Y_i^{(w)}$ that would occur under some fixed treatment w does not depend on the treatment W_i that is actually observed.

1.3 Estimation

Having identified our causal estimand as some kind of function or functional of the observed data, we can now forget about causality until it comes time to do sensitivity analysis. This is slightly cavalier, as some sensitivity analysis methods are available only for estimators of specific forms, but very general and even completely estimator-agnostic approaches are available (see Zhao et al. (2017a) and Ding and VanderWeele (2016) respectively). I say this to emphasize that the estimation problems in causal inference need not be the specific domain of the causal inference community — they are ordinary nonparametric or semiparametric estimation problems.

1.3.1 Estimation of Linear Functionals and Riesz Representation

In this dissertation, I focus on semiparametric problems, and in particular on the estimation of continuous linear functionals ψ of conditional means $m(x, w)$.[1] One of the core objects that we will be discussing is the Riesz representer γ_ψ, the unique function that satisfies

$$\mathbb{E}\,\gamma_\psi(X_i, W_i) f(X_i, W_i) = \psi(f) \quad \text{for all square integrable functions } f(x, w). \tag{1.2}$$

[1] The method considered in the first chapter can be applied to linearizations of differentiable functionals around an initial estimator $\hat{m}(x, w)$, but this will not be discussed.

The Riesz representation theorem guarantees that such a function exists and is unique for any continuous functional ψ (see e.g. Peypouquet, 2015, Theorem 1.4.1). The relevant notion of continuity is continuity in mean square, i.e. $\psi(f) \leq C\|f\|_{L_2(P)}$ for all f and some constant C where $\|f\|_{L_2(P)} = \sqrt{\mathbb{E}\, f(X_i, W_i)^2}$. If we know γ_ψ, we have a good estimator: $n^{-1}\sum_{i=1}^n \gamma_\psi(X_i, W_i)Y_i \to \mathbb{E}[\gamma_\psi(X_i, W_i)\,\mathbb{E}[Y_i \mid X_i, W_i]] = \psi(m)$ at $n^{-1/2}$ rate by the central limit theorem.

While this sort of Riesz representer is not frequently discussed as a general phenomenon in this context, instances are discussed pervasively. For example, the Riesz representer for the functional $\psi(f) = \mathbb{E}[f(X_i, 0) \mid W_i = 1]$ that we identified in the context of the ATT is the inverse propensity weighting function[2]

$$\gamma_\psi(w, x) = \frac{1_{\{w=0\}}(1 - e(x))}{p_1 e(x)} \quad \text{where} \quad e(x) = P\{W_i = 0 \mid X_i = x\},\ p_1 = P\{W_i = 1\}.$$

This function $e(X)$ is called the propensity score. The 'overlap' condition $0 < e(x)$ that we need to identify the ATT implies continuity of the functional ψ that we identify as the ATT. Furthermore, the 'strong overlap' condition $0 < \eta \leq e(x)$ that is typically assumed is equivalent to boundedness of the Riesz representer γ_ψ. This boundedness property is almost invariably assumed in the semiparametric estimation literature, and it has in fact been shown by Khan and Tamer (2010) that it is required for $n^{-1/2}$-rate estimation of the ATE and other estimands. We will assume it throughout.

Estimation of the Riesz representer γ_ψ is straightforward in general, as the defining property (1.2) of the Riesz representer is a set of moment conditions. These moment conditions have been used for some time in the causal inference literature for checking the adequacy of inverse propensity score estimators (see e.g Rubin, 2004) and for estimation of weights that act like γ_ψ (Graham et al., 2012; Hainmueller, 2012; Imai and Ratkovic, 2014; Robins et al., 2007; Zubizarreta, 2012). In that tradition, the approximate satisfaction of the equations (1.2) is called 'balance,' and arises from minimax considerations. More recently, Chernozhukov et al. (2016, 2018); Newey and Robins (2018) have begun speaking more explicitly about estimating Riesz representers using the moment conditions (1.2), and have done so in fairly general settings. Both influences appear in what follows.

In the next chapter, I will discuss a general approach to the estimation of linear functionals ψ of conditional mean function m. The essential approach is to bias-correct simple plugin estimator $\psi(\hat{m})$ by subtracting a weighted sum of the regression residuals $Y_i - \hat{m}(X_i, W_i)$. The weights we choose solve a sort of minimax problem — minimizing the maximum, over conditional mean functions m in

[2]It is common to write this without the factor of p_1 in the denominator use it in a weighted average $n_1^{-1}\sum_{i=1}^n 1_{\{W_i=0\}}(1 - e(X_i))/e(X_i)Y_i$ where $n_1 = \sum_{i=1}^n 1_{\{W_i=1\}}$, which behaves like $n^{-1}\sum_{i=1}^n \gamma_\psi(X_i, W_i)Y_i$ because $n_1 \approx p_1 n$.

a neighborhood of our estimate $\hat{m}$, of the design-conditional mean squared error of the bias-corrected estimator. This minimax criterion demands that our weights act like and ultimately converge to the Riesz representer γ_ψ of our functional, and as a consequence of this behavior we are able to establish semiparametric efficiency under weak regularity conditions for a large class of estimands. This class includes the ATE for a categorical treatment and various generalizations for continuous-valued treatments and individualized treatment assignment policies.

In the chapter following, I focus on the estimation of average treatment effects using minimax linear estimators, a degenerate case of the approach discussed in the first chapter in which we use a trivial estimate $\hat{m} = 0$ of the conditional mean. Using a sharper characterization of this estimator's design-conditional bias than the one offered in the first chapter, I establish semiparametric efficiency for these estimators under similarly weak conditions and characterize the degree to which this first-order asymptotic characterization justifies inference by bounding higher order terms.

Chapter Two

Augmented Minimax Linear Estimation

In this chapter, we address problems in which we observe n independent and identically distributed samples $(Z_i, Y_i) \sim P$ with support in $\mathcal{Z} \times \mathbb{R}$, and we want to estimate a continuous linear functional of the form

$$\psi(m) = \mathbb{E}\left[h(Z_i,\, m)\right] \quad \text{at} \quad m(z) = \mathbb{E}\left[Y_i \,\middle|\, Z_i = z\right]. \tag{2.1}$$

Our main result establishes that we can build efficient estimators for a wide variety of such problems simply by subtracting from a plugin estimator $\psi(\hat{m})$ a minimax linear estimate of its error $\psi(\hat{m}) - \psi(m)$.

The following estimands from the literature on causal inference and missing data are of this type and can be estimated efficiently by our approach.

Example 2.1 (Mean with Outcomes Missing at Random). Suppose we observe covariates X_i and some but not all of the corresponding outcomes $Y_i^\star$. Then for an indicator W_i that the outcome $Y_i^\star$ was observed, we have observed $Z_i = (X_i, W_i)$ and $Y_i = W_i Y_i^\star$, and we may estimate the linear functional $\psi(m) = \mathbb{E}\left[m(X_i,\, 1)\right]$ at $m(x, w) = \mathbb{E}\left[Y_i \,\middle|\, X_i = x, W_i = w\right]$. This will be equal to the mean $\mathbb{E}\left[Y_i^\star\right]$ if, conditional on covariates X_i, each outcome $Y_i^\star$ is independent of its nonmissingness W_i (Rosenbaum and Rubin, 1983).

Example 2.2 (Average Partial Effect). Letting $Z_i = (X_i,\, W_i) \in \mathcal{X} \times \mathbb{R}$, we estimate the average of the derivative of the response surface $m(x, w)$ with respect to w, $\psi(m) = \mathbb{E}\left[\frac{d}{dw}\left\{m(X_i,\, w)\right\}_{w=W_i}\right]$. This estimand—and weighted generalizations of it—present a natural quantification of the average effect of a continuous treatment W_i under exogeneity (Powell, Stock, and Stoker, 1989).

Example 2.3 (Average Partial Effect in the Conditionally Linear Model). Considering the estimand discussed in the previous example, we make the additional assumption that the regression function m is conditionally linear in w, $m(x, w) = \mu(x) + w\tau(x)$. Then the average partial effect is $\psi(m) = \mathbb{E}\left[\tau(X_i)\right]$ (Robinson, 1988).

Example 2.4 (Distribution Shift). We estimate the effect of a shift in the distribution of the conditioning variable Z from one known distribution, P_0, to another, P_1. $\psi(m) = \int m(z)(dP_1(z) - dP_0(z))$

for $m(z) = \mathbb{E}[Y_i \mid Z_i = z]$. Under exogeneity assumptions, this estimand can be used to compare policies for assigning personalized treatments, and estimators for it form a key building block in methods for estimation of optimal treatment policies (Athey and Wager, 2017).

In this section, we will discuss our estimator in the simple case that our functional of interest $\psi(\cdot)$ is *known*, in the sense that given a function f, we are able to evaluate $\psi(f)$. This is the case in Example 2.4. Our problem formulation (2.1) is more general, allowing $\psi(\cdot)$ to depend on the unknown distribution P in a limited way, as $\mathbb{E}[h(Z, \cdot)]$ depends on the marginal distribution of Z. We address this sort of dependence later in Section 2.1 by working with sample average approximations to $\psi(\cdot)$.

2.0.1 Estimation of Known Linear Functionals

Consider the estimation of $\psi(m)$ where $\psi(\cdot)$ is a known mean-square-continuous linear functional. As discussed in our introductory remarks, the estimator we propose is a plugin estimator $\psi(\hat{m})$ with a estimate of its error $\psi(\hat{m}) - \psi(m) = \psi(\hat{m} - m)$ subtracted,

$$\hat{\psi} = \psi(\hat{m}) - \frac{1}{n} \sum_{i=1}^{n} \hat{\gamma}_i \left(\hat{m}(Z_i) - Y_i\right). \tag{2.2}$$

Our focus will be on this error estimate $n^{-1} \sum_{i=1}^{n} \hat{\gamma}_i \left(\hat{m}(Z_i) - Y_i\right)$. The existence of a good estimate of this form follows from the Riesz representation theorem, which implies that any continuous linear functional $\psi(\cdot)$ on the square integrable functions from $\mathcal{Z}$ to $\mathbb{R}$ has a Riesz representer $\gamma_\psi(\cdot)$, i.e. a function satisfying $\mathbb{E}[\gamma_\psi(Z_i) f(Z_i)] = \psi(f)$ for all square-integrable functions f (see e.g. Peypouquet, 2015, Theorem 1.4.1).

Chernozhukov, Escanciano, Ichimura, and Newey (2016) show that using this function γ_ψ, it is possible to define an oracle estimator of the proposed form. To do this, consider the function $f = \hat{m} - m$, approximate this expectation by a sample average $n^{-1} \sum_{i=1}^{n} \gamma_i (\hat{m}(Z_i) - m(Z_i))$ with $\gamma_i = \gamma_\psi(Z_i)$, and substitute for the unknown quantity $m(Z_i)$ the unbiased estimator Y_i:

$$\begin{aligned} \psi(\hat{m} - m) &= \mathbb{E}[\gamma_\psi(Z)(\hat{m} - m)(Z)] \\ &\approx \frac{1}{n} \sum_{i=1}^{n} \gamma_i (\hat{m}(Z_i) - m(Z_i)) \\ &= \frac{1}{n} \sum_{i=1}^{n} \gamma_i (\hat{m}(Z_i) - Y_i) + \frac{1}{n} \sum_{i=1}^{n} \gamma_i \left(Y_i - m(Z_i)\right). \end{aligned} \tag{2.3}$$

As a result, the error of the estimator (2.2) with the oracle weights $\hat{\gamma}_i = \gamma_\psi(Z_i)$ will be roughly equal to a weighted sum of mean-zero noise $n^{-1} \sum_{i=1}^{n} \gamma_i \varepsilon_i$ where $\varepsilon_i = Y_i - m(Z_i)$. This behavior is known to be asymptotically optimal with a great deal of generality (see e.g. Newey, 1994, Proposition 4).

Our goal will be to imitate the behavior of this oracle estimator. One possible approach is to determine the form of the Riesz representer $\gamma_\psi(\cdot)$ by solving the set of equations that define it,

$$\mathbb{E}\left[\gamma_\psi(Z) f(Z)\right] = \psi(f) \quad \text{for all } f \text{ satisfying } \quad \mathbb{E}\left[f(Z)^2\right] < \infty, \tag{2.4}$$

then estimate it and plug the resulting weights $\hat{\gamma}_i = \hat{\gamma}_\psi(Z_i)$ into (2.2). In the context of our first example, the estimation of a mean with outcomes missing, the Riesz representer is the inverse probablility weight $\gamma_\psi(w, x) = w/e(x)$ where $e(x) = P[W_i = 1 \mid X_i = x]$, and this approach results in the well-known Augmented Inverse Probability Weighting (AIPW) estimator of Robins and Rotnitzky (1995).

We take another approach. Considering our regression estimator $\hat{m}$ and the design $Z_1 \dots Z_n$ to be fixed[1], we simply choose the weights $\hat{\gamma} \in \mathbb{R}^n$ that make our correction term $n^{-1} \sum_{i=1}^n \hat{\gamma}_i \left(\hat{m}(Z_i) - Y_i\right)$ a minimax linear estimator of what it is intended to correct for, $\psi(\hat{m} - m)$. To be precise, we choose the weights that perform best in terms of mean squared error in the worst case over regression functions m in a neighborhood $\hat{m} - \mathcal{F}$ of our regression estimator $\hat{m}$ and over conditional variance functions $\text{Var}\left[Y_i \mid Z_i = z\right]$ bounded by σ^2, having chosen $\mathcal{F}$ to be an absolutely convex set of functions which, given our beliefs about the regression function m and the properties of our estimator $\hat{m}$, should contain the regression error $\hat{m} - m$. This specifies the weights $\hat{\gamma}$ as the solution to a convex optimization problem,

$$\hat{\gamma} = \underset{\gamma \in \mathbb{R}^n}{\text{argmin}}\, I^2_{\psi,\mathcal{F}}(\gamma) + \frac{\sigma^2}{n^2}\|\gamma\|^2, \quad I_{\psi,\mathcal{F}} = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \gamma_i f(Z_i) - \psi(f). \tag{2.5}$$

The good properties of minimax linear estimators like this one are well known. Donoho (1994) and related papers (Armstrong and Kolesár, 2018; Cai and Low, 2003; Donoho and Liu, 1991; Ibragimov and Khas'minskii, 1985; Johnstone, 2015; Juditsky and Nemirovski, 2009) show that when a regression function m is in a convex set $\mathcal{F}$ and $Y_i \,\big|\, Z_i \sim N(0, \sigma_i^2)$, a minimax linear estimator of a linear functional $\psi(m)$ will come within a factor 1.25 of the minimax risk over all estimators. In addition to strong conceptual support, estimators of the type have been found to perform well in practice across several application areas (Armstrong and Kolesár, 2018; Imbens and Wager, 2017; Kallus, 2016; Zubizarreta, 2015). Because we 'augment' the minimax linear estimator by applying it after regression adjustment in the same way that the AIPW estimator augments the inverse probability weighting estimator, we refer to this approach as the Augmented Minimax Linear (AML) estimator.

[1] If we estimate $\hat{m}$ on an auxilliary sample, this is the case when we condition on both that sample and on $Z_1 \dots Z_n$. While it is not necessary to estimate $\hat{m}$ on an auxilliary sample when estimating linear functionals, we do this in the nonlinear case discussed in a forthcoming commentary (Hirshberg and Wager, 2018).

These weights $\hat{\gamma}$ can be interpreted as a penalized least-squares solution to a set of estimating equations suggested by the definition (2.4) of the Riesz representer γ_ψ,

$$\frac{1}{n}\sum_{i=1}^{n}\gamma_i f(Z_i) \approx \psi(f) \quad \text{for all} \quad f \in \mathcal{F} \tag{2.6}$$

Note that the restriction of f to a strict subset $\mathcal{F}$ of the square-integrable functions is necessary, as there are infinitely many square-integrable functions f that agree our sample $Z_1 \dots Z_n$ and they need not even approximately agree in terms of $\psi(f)$. Our choice of this subset $\mathcal{F}$, a set that characterizes our uncertainty about the regression error function $\hat{m} - m$, focuses our estimated weights $\hat{\gamma}$ on the role they play in our correction term's derivation (2.3) — the role of ensuring that (2.6) is satisfied for this function $f = \hat{m} - m$. The size of this subset $\mathcal{F}$, measured e.g. by its Rademacher Complexity, determines the accuracy with which these equations (2.6) can be simultaneously satisfied. So that we do not 'waste' accuracy at $f = \hat{m} - m$ by working with too large a set $\mathcal{F}$, it is helpful to encode the complexity-limiting assumptions that we believe are satisfied by $\hat{m} - m$ in our choice. For example, we may take $\mathcal{F}$ to be a set of smooth functions, functions that are approximately sparse in some basis, functions of bounded variation, etc.

That our weights $\hat{\gamma}_i$ approximately solve these estimating equations (2.6) does not imply that they estimate the Riesz representer $\gamma_\psi(\cdot)$ well in the mean-square sense. However, to whatever degree the oracle weights $\gamma_i = \gamma_\psi(Z_i)$ also approximately solve (2.6), it will imply that $\hat{\gamma}$ and $\gamma_\psi(\cdot)$ are close in the sense that

$$\frac{1}{n}\sum_{i=1}^{n}[\hat{\gamma}_i - \gamma_\psi(Z_i)]f(Z_i) \approx 0 \quad \text{for all } f \in \mathcal{F}. \tag{2.7}$$

This property will hold if the vector with elements $\hat{\gamma}_i - \gamma_\psi(Z_i)$ is small *or* if it is approximately orthogonal to the vector with elements $f(Z_i)$ for all functions $f \in \mathcal{F}$, and so long as $\hat{m} - m$ is in $\mathcal{F}$ or a scaled version of it, this will imply that our estimator with weights $\hat{\gamma}_i$ and our oracle estimator with weights $\gamma_i = \gamma_\psi(Z_i)$ will be close as well — the difference between them is $n^{-1}\sum_{i=1}^{n}[\hat{\gamma}_i - \gamma_\psi(Z_i)][\hat{m}(Z_i) - m(Z_i) - \varepsilon_i]$.

We state below a simple version of our main result. In essence, if an estimator $\hat{m}$ converges to m in mean square and our regression error $\hat{m} - m$ is in a uniformly bounded Donsker class $\mathcal{F}$ or more generally satisfies $(\hat{m} - m)/\mathcal{O}_p(1) \in \mathcal{F}$, then our approach can be used to define an asymptotically efficient estimator of a known continuous linear functional $\psi(m)$ at $m(z) = \mathbb{E}[Y_i \mid Z_i = z]$.

2.0.2 Definitions

As a measure of the scale of a function f relative to an absolutely convex set $\mathcal{F}$, we define the *gauge*[2] $\|f\|_{\mathcal{F}} := \inf\{\alpha \geq 0 : f \in \alpha\mathcal{F}\}$. We will write $L_2(P)$ to refer to $\{f : \mathbb{E}\left[f(Z)^2 \leq 1\right]\}$ and $L_2(P_n)$ for

$\{f : n^{-1}\sum_{i=1}^n f(Z_i)^2 \leq 1\}$, so that the gauges $\|\cdot\|_{L_2(P)}$ and $\|\cdot\|_{L_2(P_n)}$ have their typical meanings as the root mean squared error and empirical root mean squared error. We will write $\overline{\mathcal{M}}$ to denote the closure of a subspace $\overline{\mathcal{M}}$ of the square-integrable functions and will also also write $\overline{\operatorname{span}}\,\mathcal{F}$ to denote the closure of $\operatorname{span}\mathcal{F}$. We say a class $\mathcal{F}$ is pointwise separable if it has a countable subset $\mathcal{F}_0$ such that for every function $f \in \mathcal{F}$, there is a sequence $f_m \in \mathcal{F}_0$ converging to g pointwise and in $\|\cdot\|_{L_2(P)}$ (see, e.g., van der Vaart and Wellner, 1996, section 2.3.3).

2.0.3 Setting

We observe $(Y_1, Z_1) \ldots (Y_n, Z_n) \overset{iid}{\sim} P$ with $Y_i \in \mathbb{R}$, $Z_i \in \mathcal{Z}$ for a complete separable metric space $\mathcal{Z}$. We assume that $m(z) = \mathbb{E}_P[Y_i \mid Z_i = z]$ is in a subspace $\mathcal{M}$ of the square integrable functions and that $v(z) = \operatorname{Var}[Y_i \mid Z_i = z]$ is bounded. And we let $\mathcal{F}$ be absolutely convex set of square integrable functions $\mathcal{F}$ that believed to contain, at least up to scale, the regression error $\hat{m} - m$.

Our estimand is $\psi(m)$ for a known and continuous linear functional $\psi(\cdot)$ on a subspace $\mathcal{M} \cup \operatorname{span}\mathcal{F}$ of the square integrable functions. The Riesz representation theorem guarantees the existence and uniqueness of a function $\gamma_\psi \in \overline{\operatorname{span}}\,\mathcal{F}$ satisfying the set of equations $\{\mathbb{E}_P\,\gamma_\psi(Z)f(Z) = \psi(f) : f \in \overline{\operatorname{span}}\,\mathcal{F}\}$.[3] We call this function the Riesz representer of ψ on the *tangent space* $\overline{\operatorname{span}}\,\mathcal{F}$ and observe that when $\overline{\operatorname{span}}\,\mathcal{F}$ is the space of square integrable functions, this agrees with our prior definition (2.4).

We will assume that $\psi(\cdot)$ satisfies the following continuity property, which ensures that our Riesz representer γ_ψ is bounded.[4]

$$\|\psi\|_{L_1^\star(P)} < \infty \text{ for } \|\psi\|_{L_1^\star(P)} := \sup_{\substack{f \in \operatorname{span}\mathcal{F} \\ \|f\|_{L_1(P)} \leq 1}} \psi(f). \tag{2.8}$$

Theorem 2.1. *In the setting above, consider the estimator*

$$\begin{aligned} \hat{\psi}_{AML} &= \psi(\hat{m}) - \frac{1}{n}\sum_{i=1}^n \hat{\gamma}_i\left(\hat{m}(Z_i) - Y_i\right), \\ \hat{\gamma} &= \operatorname*{argmin}_{\gamma \in \mathbb{R}^n} I^2_{\psi, \tilde{\mathcal{F}}_n}(\gamma) + \frac{\sigma^2}{n^2}\|\gamma\|^2, \quad I_{\psi,\mathcal{F}} = \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n \gamma_i f(Z_i) - \psi(f). \end{aligned} \tag{2.9}$$

[2]We write the gauge $\|\cdot\|_{\mathcal{F}}$ because for the sets $\mathcal{F}$ we will be working with, the gauge is a norm. While in general, the gauge of an absolutely convex set is a pseudonorm, we will be working with sets for which point evaluation is gauge-continuous, i.e. $f(x) \leq c(x)\|f\|_{\mathcal{F}}$ for $c(x) < \infty$, and which therefore satisfy $\|f\|_{\mathcal{F}} = 0 \implies f(x) = 0$ for all x.

[3]In this statement we implicitly work with the unique extension of the continuous functional $\psi(\cdot)$ defined on $\operatorname{span}\mathcal{F}$ to a functional defined on its closure $\overline{\operatorname{span}}\,\mathcal{F}$ (Lang, 1993, Theorem IV.3.1).

[4] Boundedness of the Riesz representer γ_ψ follows from the Hahn-Banach extension theorem (Lang, 1993, Theorem IV.1.1), which guarantees that the equivalent linear functionals $f \to \psi(f)$ and $f \to \mathbb{E}\left[\gamma_\psi(Z)f(Z)\right]$ defined on the tangent space have an extension to the space $L_1(P)$ of all integrable functions which satisfies the same $\|\cdot\|_{L_1^\star(P)}$ bound and therefore satisfies $\|\gamma_\psi\|_\infty = \sup_{f \in L_1(P)} \mathbb{E}\left[\gamma_\psi(Z)f(Z)\right] = \sup_{f \in L_1(P)} \psi(f) < \infty$.

for $\tilde{\mathcal{F}}_n = \mathcal{F} \cap \rho_n L_2(P_n)$, $\rho_n \in \mathbb{R}_+ \cup \{\infty\}$ *satisfying* $n^{1/2}\rho_n \to \infty$, *and any finite* $\sigma > 0$. *If* $\mathcal{F}$ *is a pointwise separable uniformly bounded Donsker class, then the weights converge to the Riesz representer of* ψ *on the tangent space* $\overline{\operatorname{span}}\,\mathcal{F}$ *in the sense*

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\gamma}_i - \gamma_\psi(Z_i))^2 \to_P 0. \tag{2.10}$$

If, in addition, $\hat{m}$ *has the tightness and consistency properties*

a. $\|\hat{m} - m\|_{\mathcal{F}} \in \mathcal{O}_P(1)$ *and* $\|\hat{m} - m\|_{L_2(P_n)} \in \mathcal{O}_P(\rho_n)$ *if* $\rho_n \to 0$,

b. $\|\hat{m} - m\|_{\mathcal{F}} \in o_P(1)$ *otherwise,*

then our estimator $\hat{\psi}_{AML}$ *has the asymptotic linear characterization*

$$\begin{aligned} &\hat{\psi}_{AML} - \psi(m) = \frac{1}{n}\sum_{i=1}^{n}\iota(Y_i,\, Z_i) + o_P(n^{-1/2}) \quad \textit{where} \\ &\iota(y,\, z) = \gamma_\psi(z)(y - m(z)) \end{aligned} \tag{2.11}$$

and therefore $\sqrt{n}(\hat{\psi}_{AML} - \psi(m))/V^{1/2} \Rightarrow \mathcal{N}(0,\, 1)$ *with* $V = \mathbb{E}\left[\iota(Y,\, Z)^2\right]$. *When this happens,* $\hat{\psi}_{AML}$ *is regular if* $\mathcal{M} \subseteq \overline{\operatorname{span}}\,\mathcal{F}$ *and is semiparametrically efficient iff it is regular and* $v(\cdot)\gamma_\psi(\cdot) \in \overline{\mathcal{M}}$.

This theorem is a straightforward consequence of a more general asymptotic result, Theorem 2.4, discussed in Section A.1. It is proven in Appendix A.1. We end this section with a few remarks.

Remark 2.1. Our assumptions boil down to continuity of the functional $\psi(\cdot)$ and the tightness and consistency properties $\|\hat{m} - m\|_{\mathcal{F}} \in \mathcal{O}_P(1)$ and $\|\hat{m} - m\|_{L_2(P_n)} \in \mathcal{O}_P(\rho_n)$ that we require of our estimator. While we can do nothing about the continuity of the functional $\psi(\cdot)$, there is a general recipe for ensuring these tightness and consistency properties. If we can choose $\mathcal{F}$ to be an absolutely convex Donsker class such that $\|m\|_{\mathcal{F}} < \infty$, then the estimator $\hat{m}$ minimizing the penalized empirical risk $n^{-1}\sum_{i=1}^{n}(Y_i - m(Z_i))^2 + \lambda\|m\|_{\mathcal{F}}^{\nu}$ for appropriately chosen λ, ν will typically have these properties with $\rho_n = n^{-1/4}$ (see e.g. Lecué et al. (2018, Theorem 3.2) and van de Geer (2000, Theorem 10.2)).

Remark 2.2. Our estimator does not require knowledge of the form of the Riesz representer $\gamma_\psi(\cdot)$. This spares us the trouble of determining it for each estimand we consider. And while our efficiency condition $v(\cdot)\gamma_\psi(\cdot) \in \overline{\mathcal{M}}$ is phrased in terms of γ_ψ, we can often think in terms of the sufficient condition $\{v(\cdot)f(\cdot) : f \in \overline{\operatorname{span}}\,\mathcal{F}\} \subseteq \overline{\mathcal{M}}$.

Remark 2.3. We note two particular ways to define our weights in this theorem. A simple approach is to just take $\rho_n = \infty$, which results in weights which control our error uniformly over functions

in a fixed class $\mathcal{F}$. This takes advantage of the decay of the regression error $\hat{m} - m$ as measured by the gauge $\|\cdot\|_{\mathcal{F}}$, a very strong type of convergence, but not its decay in any weaker norm like $\|\cdot\|_{L_2(P_n)}$. In this case, our theorem applies if $\hat{m}$ is $\|\cdot\|_{\mathcal{F}}$-consistent for m. However, we can also exploit a known rate of convergence ρ_n for $\hat{m} - m$ in $\|\cdot\|_{L_2(P_n)}$ to work uniformly over a smaller class $\tilde{\mathcal{F}}_n = \mathcal{F} \cap \rho_n L_2(P_n)$ appropriate to our sample size; in this case, it is sufficient to have tightness of $\hat{m} - m$ in $\|\cdot\|_{\mathcal{F}}$ rather than consistency.

Remark 2.4. This theorem is valid in the general case that $\psi(m) = \mathbb{E}\left[h(Z_i, m)\right]$ if we substitute $\tilde{\psi}(\cdot) = n^{-1}\sum_{i=1}^{n} h(Z_i, \cdot)$ for $\psi(\cdot)$ where it appears in (2.9), change the influence function to $\iota(y, z) = h(z, m) - \psi(m) + \gamma_\psi(y - m(z))$, and make the additional assumptions that (i) $\{h(z, f) : f \in \mathcal{F}\}$ is a pointwise separable uniformly bounded Donsker class and that (ii) $h(Z, f)$ is uniformly continuous at zero in the sense that $\sup_{f \in \mathcal{F} \cap r L_2(P)} \operatorname{Var}\left[h(Z, f)\right]^{1/2} \to 0$ as $r \to 0$. This is proven in Appendix A.1.

Remark 2.5. Our estimator $\hat{\psi}_{AML}$ is defined in terms of an estimator $\hat{m}$ of our regression function and the class $\mathcal{F}$ of possible regression errors $\hat{m} - m$ that we correct for. The choices we make for $\hat{m}$ and $\mathcal{F}$ correspond to assumptions about the regression function m. In addition to complexity-limiting assumptions like smoothness, we may in some cases choose to make parametric or semiparametric assumptions about the form of the model. Such an assumption distinguishes Examples 2.2 and 2.3, which consider the Average Partial Effect for arbitrary functions $m(w, x)$ and for functions of the form $m(w, x) = \mu(x) + w\tau(x)$ respectively.

In the latter case, which we discuss in detail in Section 2.2, it is natural to use an estimator $\hat{m}$ of this form and to take $\mathcal{F}$ to be a class of functions having this form. As a result, the tangent space $\overline{\operatorname{span}}\,\mathcal{F}$ is smaller than the space of all square integrable functions, and the Riesz representer $\gamma_{\mathcal{F}}$ for $\psi(\cdot)$ will be the orthogonal projection of the Riesz representer γ_{L_2} for $\psi(\cdot)$ on the tangent space of all square-integrable functions onto $\overline{\operatorname{span}}\,\mathcal{F}$. An important consequence is that the optimal asymptotic variance in Example 2.3 is strictly lower than that in Example 2.2 so long as our stated conditions for efficiency are satisfied.[5] This reflects the ease of estimating the APE in the Conditionally Linear Model relative to the general case.

We pay for this reduction in asymptotic variance with a corresponding reduction in robustness. When these parametric or semiparametric assumptions are violated and $\hat{m} - m \notin \operatorname{span}\mathcal{F}$, the theorem above says nothing about the performance of our estimator. Characterization of the behavior of our estimator in settings in which these assumptions tend to be violated in practice, as in Example 2.2, is important but beyond the scope of this paper.

[5] The the difference in asymptotic variance between estimators using weights converging to γ_{L_2} (Example 2.2) and weights converging to $\gamma_{\mathcal{F}}$ (Example 2.3) is $\mathbb{E}\, v(Z)[\gamma_{L_2}^2(Z) - \gamma_{\mathcal{F}}^2(Z)] = \mathbb{E}\, v(Z)[\gamma_{L_2}(Z) - \gamma_{\mathcal{F}}(Z)]^2 +$

Remark 2.6. Although we assume no regularity conditions on the Riesz representer $\gamma_\psi(\cdot)$ beyond boundedness, our weights $\hat{\gamma}_i$ still estimate it consistently. This is a universal consistency result, in line with well known results about k-nearest neighbors regression and related estimators (Lugosi and Zeger, 1995; Stone, 1977). Heuristically, the reason for this phenomenon is that the Riesz representer γ_ψ is the unique[6] weighting function that sets a population-analogue of $I_{\psi,\mathcal{F}}$ to 0; because $\hat{\gamma}$ comes close to doing the same, it must also approximate γ_ψ. This universal consistency property is not what controls the bias of our estimator $\hat{\psi}$ (in fact the rate of convergence of $\hat{\gamma}_i$ to $\gamma_\psi(X_i)$ is in general too slow for standard arguments for plugin estimators to apply); however, it plays a key role in understanding why we get efficiency under heteroskedasticity even though we choose our weights by solving an optimization problem (2.5) that is not calibrated to the conditional variance structure of Y_i.

To understand this phenomenon, observe that under the conditions of Theorem 2.1, the conditional bias term $n^{-1}\sum_{i=1}^n \hat{\gamma}_i(\hat{m}(Z_i) - m(Z_i))$ in our error is $o_P(n^{-1/2})$. It is therefore unnecessary to make an optimal bias-variance tradeoff by this sort of calibration to get efficiency under heteroskedasticity and heteroskedasticity-robust confidence intervals; the asymptotic behavior of our estimator is determined by the asymptotic behavior of our noise term $n^{-1}\sum_{i=1}^n \hat{\gamma}_i\varepsilon_i$ and therefore by the limiting weights $\gamma_\psi(Z_i)$.

For the same reason, it is not necessary to know the error scale $\|\hat{m} - m\|_{\mathcal{F}}$ to form asymptotically valid confidence intervals. We stress that this is an asymptotic statement; in finite samples, there are strong impossibility results for uniform inference that is adaptive to the scale of an unknown signal(Armstrong and Kolesár, 2018). Furthermore, tuning approaches that estimate and incorporate individual variances σ_i into the minimax weighting problem (2.5) like those discussed in Armstrong and Kolesár (2018) may offer some finite-sample improvement.

2.0.4 Comparison with Double-Robust Estimation

Perhaps the most popular existing paradigm for building semiparametrically efficient estimators in this setting is via constructions that first compute stand-alone estimates $\hat{m}(\cdot)$ and $\hat{\gamma}_\psi(\cdot)$ for the regression function and the Riesz representer, and then plug them into (Chernozhukov et al., 2016;

$2\,\mathbb{E}\, v(Z)\gamma_{\mathcal{F}}(Z)[\gamma_{L_2}(Z) - \gamma_{\mathcal{F}}(Z)]$. The first term in this decomposition is positive and the second term is zero *if* $v(\cdot)\gamma_{\mathcal{F}}(\cdot) \in \overline{\operatorname{span}}\,\mathcal{F}$, as in this case therefore $\mathbb{E}\,\gamma_{L_2}(Z)[v(Z)\gamma_{\mathcal{F}}(Z)] = \psi(v(Z)\gamma_{\mathcal{F}}(Z)) = \mathbb{E}\,\gamma_{\mathcal{F}}[v(Z)\gamma_{\mathcal{F}}]$. This condition is satisfied under our efficiency conditions.

[6]This uniqueness is violated when the tangent space $\operatorname{span}\mathcal{F}$ that ψ acts on is not dense in the space of square integrable functions. However, the dual characterization Lemma 2.5 shows that our weights must converge to a function in the closure of this tangent space, and it follows that they converge to the unique Riesz representer γ_ψ on this tangent space.

Newey, 1994; Robins and Rotnitzky, 1995)

$$\hat{\psi}_{DR} = \gamma(\hat{m}) - \frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}_\psi(Z_i)\left(\hat{m}(Z_i) - Y_i\right) \tag{2.12}$$

or an asymptotically equivalent expression (see e.g. van der Laan and Rubin, 2006). This estimator has a long history in the context of many specific estimands, e.g. the aforementioned AIPW estimator for the estimation of a mean with outcomes missing at random (Cassel, Särndal, and Wretman, 1976; Robins, Rotnitzky, and Zhao, 1994). In recent work, Chernozhukov, Newey, and Robins (2018) describe a general approach of this type, making use of a novel estimator for the Riesz representer of a functional γ_ψ in high dimensions motivated by the Dantzig selector of Candès and Tao (2007).

In considerable generality, this estimator $\hat{\psi}_{DR}$ is efficient when we use sample splitting[7] to construct $\hat{m}$ and $\hat{\gamma}_\psi$ and these estimators satisfy

$$\frac{1}{n}\sum_{i=1}^{n}[\hat{\gamma}_\psi(Z_i) - \gamma_\psi(Z_i)][\hat{m}(Z_i) - m(Z_i)] \in o_P(n^{-1/2}) \tag{2.13}$$

(Chernozhukov et al., 2017; Zheng and van der Laan, 2011). Taking the Cauchy-Schwartz bound on this bilinear form results in the well known sufficient condition on the product of the errors, $\|\hat{\gamma}_\psi - \gamma_\psi\|_{L_2(P_n)}\|\hat{m} - m\|_{L_2(P_n)} \in o_P(n^{-1/2})$. This phenomenon, that we can trade off accuracy in how well the two nuisance functions m and γ_ψ are estimated, is called *double-robustness*.

While the estimator $\hat{\psi}_{AML}$ defined in (2.9) shares the form of $\hat{\psi}_{DR}$, it is in no reasonable sense doubly robust. This is by design. The weights $\hat{\gamma}$ used in $\hat{\psi}_{AML}$ are optimized for the task of correcting the error of the plugin estimator $\psi(\hat{m})$ when our assumptions on the regression error function $\hat{m} - m$ are correct. When this is the case and the class $\mathcal{F}$ characterizing our uncertainty about this function is sufficiently small (e.g. Donsker), this allows us to be completely robust to the difficulty of estimating the Riesz representer γ_ψ. Our estimator will be efficient essentially because the error $\hat{\gamma} - \gamma_\psi$ will be sufficiently orthogonal to all functions $f \in \mathcal{F}$ that (2.13) will be satisfied uniformly over the class of possible regression error functions $\hat{m} - m \in \mathcal{F}$. As the existence of an estimator $\hat{m}$ whose error $\hat{m} - m$ is tight in the gauge of some Donsker class $\mathcal{F}$ is essentially equivalent to the existence of an $o_P(n^{-1/4})$-consistent regression estimator of m, one way to interpret this is that our use of minimax linear weights $\hat{\gamma}_i$ rather than plug-in estimates of $\gamma_\psi(Z_i)$ has let us completely eliminate the regularity requirements on the Riesz representer γ_ψ while requiring the same level of regularity on the regression function $m(\cdot)$.

On the other hand, we sacrifice robustness to the difficulty of estimating the regression function m. In terms of the regularity assumptions necessary for asymptotic efficiency, $\hat{\psi}_{DR}$ is preferable to

[7] In particular, this result holds if we use the cross-fitting construction of Schick (1986), where separate data folds are used to estimate the nuisance components $\hat{m}(\cdot)$ and $\hat{\gamma}_\psi(\cdot)$ and to compute the expression (2.12) given those estimates.

$\hat{\psi}_{AML}$ whenever estimates of γ_ψ with faster than $\mathcal{O}_P(n^{-1/4})$ convergence are available (and vice-versa). Furthermore, for some specific choices of estimators $\hat{\gamma}_\psi(\cdot)$ and $\hat{m}(\cdot)$, it has been shown that the errors in estimating the nuisance parameters are sufficiently orthogonal that the rate-product bound can be relaxed (Newey and Robins, 2018). Thus, our aim is by no means to suggest that the AMLE dominates existing doubly-robust methods, but rather only to show that the approach can achieve efficiency under surprisingly general conditions.

In addition, we typically sacrifice robustness to any semiparametric or parametric assumptions we make on the form our regression function m. For example, when estimating a mean with outcomes missing at random in a high-dimensional linear model $m(w, z) = wx^T\beta$, it is natural to control error over a set $\mathcal{F}$ of similar linear models. In this case, the Riesz representer for $\psi(\cdot)$ on the tangent space $\overline{\text{span}}\,\mathcal{F}$ will be not the inverse propensity weight $w/e(x)$ but its best linear approximation. This can result in greater efficiency of estimation than using the true or estimated inverse propensity weights but it does not correct for misspecification of the linear model as the use of inverse propensity weights would. This phenomenon is not unique to our approach, as some other methods can estimate something like a Riesz representer on a tangent space of their choosing. See e.g. Remark 2.5 of Chernozhukov et al. (2017) or Section 3 of Robins et al. (2007).

Thus, while our estimator (2.9) can potentially be seen as an instance of (2.12) because our weights $\hat{\gamma}_i$ do converge to $\gamma_\psi(Z_i)$, the way the two estimators work is very different. Convergence of our weights to the Riesz representer is slow and plays only a second-order role in our analysis. The reason our weights succeed in debiasing $\psi(\hat{m})$ is the form of the optimization problem (2.5), not our universal consistency result. Thus, we often find it more helpful to think of our method in the context of minimax linear estimation rather than that of doubly robust methods.

However, these two approaches are not really discrete alternatives. The following section's Theorem 2.2 shows that our weights $\hat{\gamma}$ will, if our tuning parameter σ in (2.9) is allowed to grow with sample size at the correct rate, typically give a rate-optimal estimate of the Riesz representer $\hat{\gamma}_\psi$. Thus, by varying this parameter σ in our estimator (2.9), we trace out a family of estimators including the AMLE and a doubly-robust estimator using a very reasonable estimate of $\hat{\gamma}_\psi$. This is discussed briefly in Appendix A.1. In this paper, we will focus on the AMLE case, deferring the exploration of this continuum and strategies for choosing this tuning parameter σ to later work.

2.0.5 Related Work

As discussed above, our approach is primarily motivated as a refinement of minimax linear estimators as developed and studied by a large community over the past decades (Armstrong and Kolesár, 2018;

Cai and Low, 2003; Donoho, 1994; Donoho and Liu, 1991; Ibragimov and Khas'minskii, 1985; Imbens and Wager, 2017; Johnstone, 2015; Juditsky and Nemirovski, 2009; Kallus, 2016; Zubizarreta, 2015); meanwhile, our main efficiency result is most closely comparable to results from the literature on semiparametrically efficient inference, including results on doubly robust methods (Belloni et al., 2017; Bickel et al., 1998; Chen et al., 2008; Chernozhukov et al., 2017, 2018; Farrell, 2015; Hahn, 1998; Hirano et al., 2003; Mukherjee et al., 2017; Newey, 1994; Newey and Robins, 2018; Scharfstein et al., 1999; Robins and Rotnitzky, 1995; Robins et al., 2017; van der Laan and Robins, 2003; van der Laan and Rose, 2011; van der Vaart, 1991).

We are aware of two estimators that can be understood as special cases of our augmented minimax linear estimator (2.2). In the case of parameter estimation in high-dimensional linear models, Javanmard and Montanari (2014) propose a type of debiased lasso that combines a lasso regression adjustment with weights that debias the L_1-ball (i.e., a convex class known to capture the error of the lasso); meanwhile, Athey, Imbens, and Wager (2016) develop a related idea for average treatment effect estimation with high-dimensional confounding. The contribution of our paper relative to this line of work lies in the generality of our results, and also in characterizing the asymptotic variance of the estimator under heteroskedasticity and proving efficiency in the fixed-dimensional nonparametric setting. Given heteroskedasticity, Athey, Imbens, and Wager (2016) and Javanmard and Montanari (2014) only prove $\sqrt{n}$-consistency but do not characterize the the asymptotic variance directly in terms of the distribution of the data; rather, they have an expression for the variance that depends explicitly on the solution to an optimization problem analogous to (2.5).

In the special case of mean estimation with data missing at random, the optimization problem (2.5) takes on a particularly intuitive form, and

$$I_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left(1 - W_i \hat{\gamma}_i\right) f(X_i,\, 1) \tag{2.14}$$

measures how well the $\hat{\gamma}$-weighted average of f over the observed samples matches its average over everyone. In other words, the minimax linear weights enforce "balance", which has been emphasized as fundamental to this problem by several authors including Rosenbaum and Rubin (1983) and Hirano, Imbens, and Ridder (2003). More recently, there has been considerable interest in practical methodologies that emphasize balance when paired with AIPW methodology (Athey et al., 2016; Chan et al., 2015; Graham et al., 2012, 2016; Hainmueller, 2012; Hirano et al., 2001, 2003; Imai and Ratkovic, 2014; Kallus, 2016; Wang and Zubizarreta, 2017; Zhao, 2016; Zubizarreta, 2015). In addition to generalizing beyond the missing-at-random problem, our Theorem 2.4 also provides the

sharpest results we are aware of for balancing-type estimators in this specific problem.

2.1 Estimating Linear Functionals

In this section, we will address the problem of estimating continuous linear functionals of the form $\psi(m) = \mathbb{E}\left[h(Z, m)\right]$ at $m = \mathbb{E}\left[Y_i \mid Z_i = z\right]$. We will be working with a generalization of the estimator described in the previous section that substitutes sample averages of $h(Z_i, \cdot)$ for the possibly unknown functional $\psi(\cdot)$,

$$\begin{aligned} \hat{\psi}_{AML} &= \frac{1}{n}\sum_{i=1}^{n} h(Z_i, \hat{m}) - \frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}_i \left(\hat{m}(Z_i) - Y_i\right), \\ \hat{\gamma} &= \operatorname*{argmin}_{\gamma \in \mathbb{R}^n} I^2_{h,\tilde{\mathcal{F}}}(\gamma) + \frac{\sigma^2}{n^2}\|\gamma\|^2, \quad I_{h,\mathcal{F}} = \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} [\gamma_i f(Z_i) - h(Z_i, f)]. \end{aligned} \tag{2.15}$$

Note that in the case that $\psi(\cdot)$ is known, $h(Z_i, \cdot) = \psi(\cdot)$ for all Z_i, and this reduces to our estimator from Theorem 2.1 when we take $\tilde{\mathcal{F}} = \mathcal{F} \cap \rho_n L_2(P_n)$. Here we allow $\tilde{\mathcal{F}}$ to be an arbitrary set defined in terms of $Z_1 \dots Z_n$ and we will characterize our estimator primarily in terms of a pair of nonrandom 'bounds' $\mathcal{F}_L$ and $\mathcal{F}$ satisfying $\mathcal{F}_L \subseteq \tilde{\mathcal{F}} \subseteq \mathcal{F}$ with high probability.

To better understand the behavior of our estimator, we decompose its error into a bias-like term and a noise-like term. We will consider estimation of a sample-average version of our estimand, $\tilde{\psi}(m) := n^{-1}\sum_{i=1}^{n} h(Z_i, m)$, as the behavior of the latter term in the error decomposition $\hat{\psi}_{AML} - \psi(m) = (\hat{\psi} - \tilde{\psi}(m)) + (\tilde{\psi}(m) - \psi(m))$ is entirely out of our hands. We write

$$\begin{aligned} \hat{\psi}_{AML} - \tilde{\psi}(m) &= \frac{1}{n}\sum_{i=1}^{n} \left(h(Z_i, \hat{m}) - h(Z_i, m)\right) - \hat{\gamma}_i \left(\hat{m}(Z_i) - Y_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \underbrace{h(Z_i, \hat{m} - m) - \hat{\gamma}_i(\hat{m} - m)(Z_i)}_{\text{bias}} + \underbrace{\hat{\gamma}_i \left(Y_i - m(Z_i)\right)}_{\text{noise}}. \end{aligned} \tag{2.16}$$

We will establish finite sample bounds on the bias term and the difference between the noise term and the noise term of the oracle estimator with weights $\gamma_\psi(Z_i)$. If both of these quantities are $o_p(n^{-1/2})$, our estimator will be asymptotically linear with influence function $\iota(y, z) = h(z, m) - \psi(m) + \gamma_\psi(z)(y - m(z))$, which implies asymptotic efficiency under a few conditions stated in Proposition 2.3.

We establish these bounds in essentially three steps.

1. Establish a bound on $n^{-1}\sum_{i=1}^{n}(\hat{\gamma}_i - \gamma_i^\star)^2$ for $\gamma_i^\star = \gamma_\psi(Z_i)$.

2. Our bias term can be bounded by $\|\hat{m} - m\|_{\tilde{\mathcal{F}}} I_{h,\tilde{\mathcal{F}}}(\hat{\gamma})$. Observe that as a consequence of the definition of our weights $\hat{\gamma}$ in (2.15), they satisfy

$$I_{h,\tilde{\mathcal{F}}}(\hat{\gamma})^2 \leq I_{h,\tilde{\mathcal{F}}}(\gamma^\star)^2 + \frac{\sigma^2}{n^2}\sum_{i=1}^{n} \gamma_i^{\star 2} - \hat{\gamma}_i^2. \tag{2.17}$$

Empirical process techniques can be used to characterize the first term in this bound, as the weights $\gamma^\star$ have the property that $I_{h,\mathcal{F}}(\gamma^\star)$ is the supremum of the empirical process $n^{-1}\sum_{i=1}^n \delta_{X_i}$ indexed by the class of mean-zero functions $\mathcal{H} = \{z \to h(z,f) - \gamma_\psi(z)f(z) : f \in \mathcal{F}\}$, while the second term can be bounded using the previous step and some simple arithmetic. This bound, in combination with a bound on $\|\hat{m} - m\|_{\tilde{\mathcal{F}}}$, will imply a bound on our bias term.

3. Bound the difference between our noise term and that of the oracle estimator, $n^{-1}\sum_{i=1}^n(\hat{\gamma}_i - \gamma_i^\star)(Y_i - m(Z_i))$, using our the first step.

The first step represents the core technical contribution of our paper. Following a few definitions, we will state and prove these bounds.

2.1.1 Definitions

As it will be useful to discuss the behavior of $h(Z_i, f)$ for $f \in \mathcal{F} - \gamma_\psi$, if $\gamma_\psi \notin \operatorname{span}\mathcal{F}$ we will we will work implicitly with the extension of the z-indexed family of linear functionals $h(z,\cdot)$ to the space spanned by this set that satisfies $h(z,\gamma_\psi) = \gamma_\psi(z)^2$ for all z. Note that when working on this larger space, γ_ψ is still a Riesz representer, as $\psi(f) = \mathbb{E}\left[h(Z_i,f)\right] = \mathbb{E}\left[\gamma_\psi(Z_i)f(Z_i)\right]$ for all f in it. It will often be convenient to work on a slight enlargement of this set, $\mathcal{F} - [0,1]\gamma_\psi$, which is star-shaped around zero.

To characterize the size of a set $\mathcal{G}$, we will use its *Rademacher complexity*, defined $R_n(\mathcal{G}) := \mathbb{E}\sup_{g\in\mathcal{G}}|n^{-1}\sum_{i=1}^n \epsilon_i g(Z_i)|$ where $\epsilon_i = \pm 1$ each with probability $1/2$ independently and independently of the sequence $Z_1 \dots Z_n$. A useful type of fixed point of the Rademacher complexity of a parameterized family of classes $\mathcal{G}(r)$ will be written $R_n^\star(c, \mathcal{G}(r)) := \inf\{r > 0 : R_n(\mathcal{G}(r)) \leq cr^2\}$. In this context, we will take $\mathcal{G}(r) = \mathcal{F} \cap rL_2(P)$ or a related class, and we call $R_n(\mathcal{G}(r))$ a *local Rademacher Complexity* (see, e.g., Bartlett et al., 2005; Koltchinskii, 2006). We will also use its maximal supremum norm $M_\mathcal{G} := \sup_{g\in\mathcal{G}}\|g\|_\infty$.

We will be interested in the Rademacher complexity and local Rademacher complexities of the classes $\mathcal{F}(r) = \mathcal{F}\cap rL_2(P)$, $\mathcal{H}(r) = \{h(z,f) - \gamma_\psi(z)f(z) : f \in \mathcal{F}(r)\}$, $\mathcal{F}^\star(r) = (\mathcal{F} - [0,1]\gamma_\psi)\cap rL_2(P)$, $\mathcal{H}^\star(r) = \{h(z,f) - \gamma_\psi(z)f(z) : f \in \mathcal{F}^\star(r)\}$, and as a shorthand will write $\mathcal{H} = \mathcal{H}(\infty), \mathcal{F}^\star = \mathcal{F}^\star(\infty), \mathcal{H}^\star = \mathcal{H}^\star(\infty)$ for the non-localized versions. Specifically, the primary factors determining our bound will be a measure r_Q of the local complexity of $\mathcal{F}^\star$, measures $u(\mathcal{H})$ and r_C of the complexity and local complexity of the classes $\mathcal{H}$ and $\mathcal{H}^\star$, and a measure κ of the degree of $\|\cdot\|_{\mathcal{F}_L}$-size necessary to approximate γ_ψ well. We define these measures, which are similar to those in Lecué and Mendelson (2017), below.

$$
\begin{aligned}
r_Q(\eta_Q) &= \frac{\sqrt{24(1+\eta_Q)}}{1-\eta_Q} R_n^\star\left(\frac{1}{2M_{\mathcal{F}^\star}}, \mathcal{F}^\star(\cdot)\right); \\
r_C(\eta_C, \delta) &= \inf\left\{r > 0 : u(\mathcal{H}^\star(r), \delta) \leq \eta_C r^2\right\} \text{ where} \\
u(\mathcal{H}, \delta) &> \sup_{h \in \mathcal{H}} n^{-1} \sum_{i=1}^n h(Z_i) \text{ with probability } 1-\delta; \text{[8]} \\
\kappa^2(\sigma, \delta) &= \inf_{\tilde{\gamma}} \left\{ \|\tilde{\gamma} - \gamma_\psi\|_{L_2(P)}^2 + \frac{\delta \sigma^2 \|\tilde{\gamma}\|_{\mathcal{F}_L}^2}{2n} \right\};
\end{aligned}
\tag{2.18}
$$

It may be helpful to have a sense of the behavior of these quantities before we state our main result. If $\tilde{\mathcal{F}}$ has an upper bound $\mathcal{F}$ that is a Donsker class, typically the local complexity fixed points $r_Q(\eta_Q)$ and $r_C(\eta_C, \delta)$ will be $o(n^{-1/4})$ and $u(\mathcal{H}_n, \delta)$ will be $O(n^{-1/2})$ — typically the latter will be $o(n^{-1/2})$ when we exploit the consistency of the regression $\hat{m}$ by choosing $\tilde{\mathcal{F}}_n$ satisfying with high probability $\sup_{f \in \tilde{\mathcal{F}}_n} \|f\|_{L_2(P)} \to 0$.[9] And for fixed $\sigma > 0$, we will have $\kappa(\sigma, \delta) \to 0$ essentially without assumptions. Roughly speaking, these properties will be sufficient to establish asymptotic results analogous to Theorem 2.1.

2.1.2 Main Results

Theorem 2.2. *Suppose that we observe iid* $(Y_1, Z_1) \dots (Y_n, Z_n)$ *with* $Y_i \in \mathbb{R}$, Z_i *in an arbitrary set* $\mathcal{Z}$, *and* $v(z) = \mathrm{Var}[Y_i \mid Z_i = z]$ *bounded. Let* $\{h(z, \cdot) : z \in \mathcal{Z}\}$ *be a family of linear functionals and the linear functional* $\psi(\cdot) = \mathbb{E}[h(Z_i, \cdot)]$ *be continuous. Consider the estimator* $\hat{\psi}_{AML}$ *defined in* (2.15) *in terms of* $\sigma > 0$ *and an absolutely convex set* $\tilde{\mathcal{F}}$ *defined in terms of* $Z_1 \dots Z_n$. *Let there exist nonrandom sets* $\mathcal{F}_L$ *and* $\mathcal{F}$ *satisfying* $\mathcal{F}_L \subseteq \tilde{\mathcal{F}} \subseteq \mathcal{F}$ *with probability* $1 - \delta_{\tilde{\mathcal{F}}}$ *with* $\mathcal{F}$ *pointwise separable, absolutely convex, and either reflexive or totally bounded in* $\|\cdot\|_\infty$. *If* $\{h(z, f) : f \in \mathcal{F}\}$ *is pointwise separable and* $h(Z_1, \cdot) \dots h(Z_n, \cdot)$ *are continuous on the normed vector space* $(\operatorname{span} \mathcal{F}, \|\cdot\|_{\mathcal{F}})$ *and on* $(\operatorname{span} \mathcal{F}, \|\cdot\|_\infty)$ *as well if the former space is not reflexive, then on an event of probability at least* $1 - \exp\{-c_1(\eta_Q) n r_Q(\eta_Q)^2 / M_{\mathcal{F}^\star}^2\} - 5\delta - 2\delta_{\tilde{\mathcal{F}}}$,

1. *The weights* $\hat{\gamma}$ *defined in* (2.15) *satisfy* $n^{-1} \sum_{i=1}^n (\hat{\gamma}_i - \hat{\gamma}_\psi(Z_i))^2 \leq a \wedge b$ *where*

[8] We may use the bound $u(\mathcal{H}, \delta) = 2\delta^{-1} R_n(\mathcal{H})$, which arises from Markov's inequality and symmetrization (see e.g. van der Vaart and Wellner, 1996, Lemma 2.3.1). Bounds based on Talagrand-type concentration inequalities (see e.g. Bartlett et al., 2005, Theorem 2.1) offer much weaker dependence on δ when $\sup_{h \in \mathcal{H}} \|h\|_{L_2(P)}$ is comparable to $R_n(\mathcal{H})$, which will typically be the case.

[9] This is tantamount to saying that the deviation of an empirical process from its mean is $o_p(n^{-1/2})$ when the class indexing it decays to zero in $\|\cdot\|_{L_2(P)}$, a phenomenon typically referred to as the asymptotic equicontinuity of the empirical process.

$$
\begin{aligned}
a &= \alpha u(\mathcal{H}^\star, \delta) + \bar{R};\\
b &= 2\alpha^2 r^2 \vee 2\frac{\bar{R} + \sigma^2/n}{\eta_Q - 2\alpha^{-1}\eta_C} \vee \frac{44 M_{\mathcal{F}^\star}^2 \alpha^2 \log(\delta^{-1})}{n};\\
r &= r_Q(\eta_Q) \vee r_C(\eta_C) \vee \sigma \eta_Q^{-1/2} n^{-1/2};\\
\alpha &= 1 \vee \left[2\eta_C \sigma^{-2} n r^2 + \sigma^{-1} n^{1/2} \bar{R}^{1/2}\right]\\
\bar{R} &= 2\delta_{\tilde{\mathcal{F}}}^{-1}\left[\kappa^2 + 2\sigma^{-1}\kappa\overline{\sigma}(\mathcal{H}^\star(\kappa))\right] + 4\delta_{\tilde{\mathcal{F}}}^{-1/2} n^{-1/2}\overline{\sigma}(\mathcal{H}^\star(\kappa)),\ \ \kappa = \kappa(\sigma, \delta_{\tilde{\mathcal{F}}}).
\end{aligned}
\tag{2.19}
$$

2. *The uniform version of our bias term satisfies the bound*

$$
I_{h,\tilde{\mathcal{F}}} \leq u(\mathcal{H}, \delta) + 2^{1/2}\|\gamma_\psi\|_{L_2(P_n)}^{1/2} \sigma n^{-1/2}(a \wedge b)^{1/4}. \tag{2.20}
$$

3. *The difference between our noise term and that of the oracle estimator satisfies*

$$
\left| n^{-1}\sum_{i=1}^n (\hat{\gamma}_i - \gamma_\psi)(Y - m(Z_i)) \right| \leq \delta^{-1/2}\|v\|_\infty n^{-1/2}(a \wedge b)^{1/2}. \tag{2.21}
$$

Here $\eta_Q \in (0, .47)$ and $\eta_C > 0$ are arbitrary and the function c_2 is defined in Lemma 2.7.

These bounds yield straightforward conditions under which our estimator is asymptotically linear, i.e.

$$
\begin{aligned}
\hat{\psi}_{AML} - \tilde{\psi}(m) &= n^{-1}\sum_{i=1}^n \tilde{\iota}(Y_i, Z_i) + o_P(n^{-1/2}),\ \ \tilde{\iota}(y,z) = \gamma_\psi(z)(y - m(z)); \quad \text{therefore}\\
\hat{\psi}_{AML} - \psi(m) &= n^{-1}\sum_{i=1}^n \iota(Y_i, Z_i) + o_P(n^{-1/2}),\ \ \iota(y,z) = h(z,m) - \psi(m) + \tilde{\iota}(y,z).
\end{aligned}
\tag{2.22}
$$

Typically, such estimators are asymptotically efficient. The following proposition, proven in Appendix A.1.4, generalizes the conditions for efficiency stated in Theorem 2.1

Proposition 2.3. *Suppose we observe an iid sample $(Z_i, Y_i)_{i \leq n}$ from P where $Y_i \in \mathbb{R}$ and $Z_i \in \mathcal{Z}$, a complete separable metric space, and that the set of possible regression functions $m(z) = \mathbb{E}[Y_i \mid Z_i = z]$ is a linear space $\mathcal{M}$. An estimator for a continuous linear functional of the form $\psi(m) = \mathbb{E}[h(Z_i, m)]$ at $m(z) = \mathbb{E}[Y_i \mid Z_i = z]$ is regular if (2.22) holds where γ_ψ is the Riesz representer for the functional $\psi(\cdot)$ on a space containing the closure of $\mathcal{M}$. It is semiparametrically efficient if, in addition, the function $z \to \gamma_\psi(z)\operatorname{Var}[Y_i \mid Z_i = z]$ is in the closure of $\mathcal{M}$.*

Now consider the expansion of our estimator around this characterization.

$$
\begin{aligned}
&\left| \hat{\psi}_{AML} - \tilde{\psi}(m) - \frac{1}{n}\sum_{i=1}^n \tilde{\iota}(Y_i, Z_i) \right|\\
&\leq \left| \frac{1}{n}\sum_{i=1}^n [h(Z_i, \hat{m} - m) - \hat{\gamma}_i(\hat{m} - m)(Z_i)] + \frac{1}{n}\sum_{i=1}^n (\hat{\gamma}_i - \gamma_\psi)(Y - m(Z_i)) \right|\\
&\leq \|\hat{m} - m\|_{\tilde{\mathcal{F}}} I_{h,\tilde{\mathcal{F}}}(\hat{\gamma}) + \frac{1}{n}\sum_{i=1}^n (\hat{\gamma}_i - \gamma_\psi)(Y_i - m(Z_i)).
\end{aligned}
\tag{2.23}
$$

This difference will be negligible if both the product of $\|\hat{m} - m\|_{\tilde{\mathcal{F}}}$ and our bound (2.20) and our bound (2.21) are $o_P(n^{-1/2})$. An inspection of these bounds, which we carry out in Appendix A.1, shows that this happens under conditions generalizing those of Theorem 2.1. This yields the following asymptotic result.

Theorem 2.4. *Let $(Z_{i,n}, Y_{i,n})_{i \leq n}$ be an iid sample from P^n with $Y_{i,n} \in \mathbb{R}$, $Z_{i,n}$ in an arbitrary set $\mathcal{Z}_n$, and $v_n(z) = \mathrm{Var}\left[Y_{i,n} \mid Z_{i,n} = z\right]$ bounded uniformly in n, and define $m_n(z) = \mathbb{E}\left[Y_{i,n} \mid Z_{i,n} = z\right]$. In terms of a family of linear functionals $\{h_n(z, \cdot) : z \in \mathcal{Z}_n\}$, define the continuous linear functional $\psi_n(\cdot) = \mathbb{E}\left[h(Z_{i,n}, \cdot)\right]$. Choose $\tilde{\mathcal{F}}_n$ to be an absolutely convex set, defined in terms of $Z_1 \ldots Z_n$, of square integrable functions on $\mathcal{Z}_n$. In terms of that set, an estimator $\hat{m}$ for m_n, and tuning parameters $\sigma_n = O(1)$, define the estimator*

$$\begin{aligned} \hat{\psi} &= \frac{1}{n}\sum_{i=1}^{n} h_n(Z_{i,n}, \hat{m}) - \frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}_i \left(\hat{m}(Z_{i,n}) - Y_{i,n}\right), \\ \hat{\gamma} &= \operatorname*{argmin}_{\gamma \in \mathbb{R}^n} I^2_{h_n, \tilde{\mathcal{F}}_n}(\gamma) + \frac{\sigma_n^2}{n^2}\|\gamma\|^2, \quad I_{h,\mathcal{F}} = \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} [\gamma_i f(Z_{i,n}) - h(Z_{i,n}, f)]. \end{aligned} \tag{2.24}$$

Let there exist nonrandom sets $\mathcal{F}_{L,n}$ and $\mathcal{F}_n$ such that $P^n\{\mathcal{F}_{L,n} \subseteq \tilde{\mathcal{F}}_n \subseteq \mathcal{F}_n\} \to 1$ with $\mathcal{F}_n$ pointwise separable and either reflexive or totally bounded in $\|\cdot\|_\infty$; let γ_{ψ_n} be the Riesz representer of ψ_n on the tangent space $\overline{\mathrm{span}}\, \mathcal{F}_n$; and define $\mathcal{H}_n(r)$, $\mathcal{F}_n^\star(r)$, $\mathcal{H}_n^\star(r)$ as is Section 2.1.1 in terms of $\mathcal{F}_n$, h_n, and γ_{ψ_n}. Then if

i. for each $Z_{i,n}$, the functional $h(Z_{i,n}, \cdot)$ is continuous on $(\mathrm{span}\, \mathcal{F}_n, \|\cdot\|_{\mathcal{F}_n})$ and if this space is not reflexive, on $(\mathrm{span}\, \mathcal{F}_n, \|\cdot\|_\infty)$ as well.

ii. our functional $\psi_n(\cdot)$ satisfies the condition $\sup\{|\psi_n(f)| : f \in \mathcal{F}_n, \|f\|_{L_1(P^n)} \leq 1\} = O(1)$, which is equivalent to uniform boundedness of its Riesz representer;

iii. its Riesz representer is approximable in the sense that there exist functions $\tilde{\gamma}_n$ satisfying $\|\tilde{\gamma}_n - \gamma_{\psi_n}\|_{L_2(P^n)} \to 0$ and $\|\tilde{\gamma}_n\|_{\mathcal{F}_{L,n}} = o(n^{1/2})$;

iv. $\mathcal{F}_n$ is uniformly bounded in the sense that $M_{\mathcal{F}_n} = O(1)$;

v. $R_n^\star(1, \mathcal{F}_n^\star(\cdot)), R_n^\star(1, \mathcal{H}_n^\star(\cdot)) = o(n^{-1/4})$;

vi. $\|\hat{m} - m\|_{\tilde{\mathcal{F}}_n} = O_{P^n}(1)$, $R_n(\mathcal{H}_n) = O_{P^n}(n^{-1/2})$, $\|\hat{m} - m\|_{\tilde{\mathcal{F}}_n} R_n(\mathcal{H}_n) = o_{P^n}(n^{-1/2})$;

our weights $\hat{\gamma}$ converge to the Riesz representer and our estimator $\hat{\psi}$ is asymptotically linear, i.e.

$$n^{-1}\sum_{i=1}^{n}(\hat{\gamma}_i - \hat{\gamma}_{\psi_n}(Z_{i,n})) \to_{P^n} 0; \tag{2.25}$$

$$\hat{\psi} - \psi_n(m_n) = n^{-1}\sum_{i=1}^{n}\iota_n(Y_{i,n}, Z_{i,n}) + o_{P^n}(n^{-1/2}) \quad \textit{with} \tag{2.26}$$

$$\iota_n(y,z) = h_n(z, m_n) - \psi_n(m_n) + \gamma_{\psi_n}(z)(y - m(z))$$

Here our assumptions (i,ii,iv) are triangular-array equivalents of assumptions stated in Theorem 2.1; (v,vi) generalize the Donskerity assumption and assumptions on the tightness and consistency of $\hat{m}$ in Theorem 2.1 for the estimation of a non-known functional and to the triangular-array setting; and (iii) is a new assumption that is essentially vacuous in the non-triangular asymptotic setting ($P^n = P$). This is the case for (iii) because any fixed function in $\overline{\text{span}}\,\mathcal{F}$ including γ_ψ can be approximated by a sequence $\tilde{\gamma}_n$ with $\|\tilde{\gamma}_n\|_{\mathcal{F}} \to \infty$. We need to include this condition in the triangular-array asymptotics because γ_{ψ_n} is not a fixed function. It may, for example, be be a function of increasing dimension.

When our estimator has the asymptotic characterization (2.26), $\sqrt{n}(\hat{\psi} - \psi_n(m_n))$ is asymptotically normal with variance $V_n = \mathbb{E}\left[\iota_n(Y_i, Z_i)^2\right]$. We can then form confidence intervals $\hat{\psi} \pm z_{\alpha/2} n^{-1/2}\widehat{V}^{1/2}$ of asymptotic size $1-\alpha$ using a consistent variance estimate $\widehat{V}$. A simple choice is

$$\widehat{V} = \frac{1}{n}\sum_{i=1}^{n}\left(h_n(Z_{i,n}, \hat{m}) - \hat{\psi}\right)^2 + \hat{\gamma}_i^2\left(Y_{i,n} - \hat{m}\left(Z_{i,n}\right)\right)^2. \tag{2.27}$$

2.1.3 Proof of Finite Sample Results

We will now prove Theorem 2.2. In our proof, we will write $P_n f$ and Pf for averages of the function f over the empirical and population distributions of Z respectively in accordance with convention in the empirical process literature (see e.g. van der Vaart and Wellner, 1996). As a slight abuse of notation, we also write P_n to indicate an empirical sum in other expressions.

2.1.3.1 Consistency of the Minimax Linear Weights

To show that our weights converge to the $\hat{\gamma}$, we will first characterize them as $\hat{\gamma}_i = \hat{g}(X_i)$ for a least squares estimator $\hat{g}$ of the Riesz representer γ. This least squares problem is the dual of the problem (2.15) solved by our weights $\hat{\gamma}$.

2.1.3.2 Dual Characterization as a Least Squares Problem

Lemma 2.5. *Let $\mathcal{G}$ be an absolutely convex set and the space $(\text{span}\,\mathcal{G}, \|\cdot\|_{\mathcal{G}})$ be a reflexive vector space. Let a linear functional $L(f)$ and the point evaluation functionals $\delta_z(f) := f(z)$ for all $z \in$*

$Z_1 \dots Z_n$ be continuous in $\|\cdot\|_{\mathcal{G}}$. Then,

$$\inf_{\gamma \in \mathbb{R}^n} \ell_{n,\mathcal{G}}(\gamma) = \sup_{g \in \operatorname{span} \mathcal{G}} \mathbb{M}_{n,\mathcal{G}}(g) \qquad \textit{where}$$

$$\ell_{n,\mathcal{G}}(\gamma) = P_n \gamma_i^2 + \sup_{f \in \mathcal{G}} \left[L(f) - P_n \gamma_i f(Z_i) \right]^2 \qquad \textit{will be called the primal and}$$

$$\mathbb{M}_{n,\mathcal{G}}(g) = -\|g\|_{\mathcal{G}}^2 - P_n g(Z_i)^2 + 2L(g) \qquad \textit{will be called the dual.}$$

Furthermore, the primal has a unique minimum at $\hat{\gamma}$ irrespective of the reflexiveness of our space, the dual has a potentially non-unique maximum at $\hat{g}$, and for any $\hat{g}$ at which the dual maximum is attained, $\hat{\gamma}_i = \hat{g}(Z_i)$.

This result is proven in the Section A.2 of the appendix by working with a constrained optimization problem equivalent to the primal. After introducing a Lagrange multiplier for the constraint, the resulting saddle point problem is reduced to maximization of $\mathbb{M}_{n,\mathcal{G}}$ by explicitly solving for γ and our Lagrange multiplier as functions of $\hat{g}$.

In our estimator (2.15), we use the weights $\hat{\gamma}$ that minimize $(\sigma^2/n)\ell_{n,\mathcal{G}}$ where $L(f) = P_n h(Z_i, f)$ and $\mathcal{G} = \sigma^{-1} n^{1/2} \tilde{\mathcal{F}}$, so we may characterize our weights via the function $\hat{g}$ that maximizes $\mathbb{M}_{n,\lambda\tilde{\mathcal{F}}}$ for $\lambda = \sigma^{-1} n^{1/2}$. This characterization will be valid at least on the high-probability event that $\tilde{\mathcal{F}} \subseteq \mathcal{F}$, as on this event $\|\cdot\|_{\mathcal{F}} \leq \|\cdot\|_{\tilde{\mathcal{F}}}$ and therefore the functionals $\delta_{Z_1} \dots \delta_{Z_n}$ and L will be continuous in $\|\cdot\|_{\tilde{\mathcal{F}}}$ and therefore in $\|\cdot\|_{\mathcal{G}}$. There is one remaining assumption that we've made in Lemma 2.5 but not in Theorem 2.2: the assumption that the space $(\operatorname{span} \tilde{\mathcal{F}}, \|\cdot\|_{\tilde{\mathcal{F}}})$ is reflexive. We will assume this holds for now, as it lets us simplify exposition but does not materially affect the final result. Later, we will derive a bound without this assumption by application of this Lemma to a sequence finite-dimensional and therefore reflexive approximations to $\tilde{\mathcal{F}}$.

It is perhaps not immediately obvious that maximizing $\mathbb{M}_{n,\lambda\tilde{\mathcal{F}}}$ is a penalized least squares problem for estimation of γ_ψ. To show this, we will consider the excess loss $\mathcal{L}_{\tilde{\gamma}}(g) := -\mathbb{M}_{n,\lambda\tilde{\mathcal{F}}}(g) + \mathbb{M}_{n,\lambda\tilde{\mathcal{F}}}(\tilde{\gamma})$ relative to an approximation $\tilde{\gamma}$ of the Riesz representer γ_ψ. This excess loss is minimized and no larger than zero at $\hat{g}$. We work with an approximation $\tilde{\gamma}$ because we are not assuming that γ_ψ is in the span of $\mathcal{F}$, so $\|\gamma_\psi\|_{\lambda\tilde{\mathcal{F}}}$ may be infinite and therefore the excess loss relative to γ_ψ itself uninformative. We then write[10]

[10] This expression can be checked via simple algebra as follows, $\mathcal{L}_{\tilde{\gamma}}(g) = P_n(g^2 - \tilde{\gamma}^2) - 2P_n\left[h(Z,g) - h(Z,\tilde{\gamma})\right] + (\|g\|_{\tilde{\mathcal{F}}}^2 - \|\tilde{\gamma}\|_{\tilde{\mathcal{F}}}^2)/\lambda^2 = P_n\left[(g-\gamma_\psi)^2 - (\tilde{\gamma}-\gamma_\psi)^2 + 2\gamma_\psi(g-\tilde{\gamma})\right] - 2P_n\left[h(Z,g) - h(Z,\tilde{\gamma})\right] + (\|g\|_{\tilde{\mathcal{F}}}^2 - \|\tilde{\gamma}\|_{\tilde{\mathcal{F}}}^2)/\lambda^2 = P_n(g-\gamma_\psi)^2 - 2P_n\left[h(Z, g-\gamma_\psi) - \gamma_\psi(g-\gamma_\psi)\right] + \|g\|_{\tilde{\mathcal{F}}}^2/\lambda^2 - \{P_n(\tilde{\gamma}-\gamma_\psi)^2 - 2P_n\left[h(Z,\tilde{\gamma}-\gamma_\psi) - \gamma_\psi(\tilde{\gamma}-\gamma_\psi)\right] - \|\tilde{\gamma}\|_{\tilde{\mathcal{F}}}^2/\lambda^2\}$.

$$\begin{aligned}
&\mathcal{L}_{\tilde{\gamma}}(g) = P_n(g-\gamma_\psi)^2 - 2P_n\check{h}(Z,\, g-\gamma_\psi) + \|g\|_{\tilde{\mathcal{F}}}^2/\lambda^2 - R_{n,\lambda\tilde{\mathcal{F}}}(\tilde{\gamma})^2, \quad \text{where} \\
&\check{h}(Z,\, g) = h(Z,\, g) - \gamma_\psi(Z)g(Z) \quad \text{and} \\
&R_{n,\lambda\mathcal{F}}(\tilde{\gamma}) = P_n(\tilde{\gamma}-\gamma_\psi)^2 - 2P_n\check{h}(Z,\, \tilde{\gamma}-\gamma_\psi) + \|\tilde{\gamma}\|_{\mathcal{F}}^2/\lambda^2
\end{aligned} \tag{2.28}$$

Here $\check{h}$ is, in a sense, a centered version of our linear functional h, as our Riesz representer γ_ψ satisfies $P\gamma_\psi(Z)g(Z) = Ph(Z,g)$ for all $g \in \text{span}(\mathcal{F} \cup \{\gamma_\psi\})$. Consequently, we have the typical form of the excess loss for a penalized least squares estimator: it is a sum of the empirical MSE, a centered empirical process, and a difference in penalties $\|g\|_{\tilde{\mathcal{F}}}^2/\lambda^2 - R_{n,\lambda\tilde{\mathcal{F}}}(\tilde{\gamma})^2$. Note that in the case that we take $\tilde{\gamma} = \gamma_\psi$, this difference in penalties is the more familiar $\|g\|_{\tilde{\mathcal{F}}}^2/\lambda^2 - \|\gamma_\psi\|_{\tilde{\mathcal{F}}}^2/\lambda^2$. We work with the noisy measurement $R_{n,\lambda\tilde{\mathcal{F}}}(\tilde{\gamma})$ of the regularity of γ_ψ indirected through $\tilde{\gamma}$ to establish useful bounds even when $\|\gamma_\psi\|_{\mathcal{F}} = \infty$.

2.1.3.3 Consistency of the Dual Solution

We will use this dual characterization to prove a high-probability finite-sample bound on $\|\hat{g} - \gamma_\psi\|_{L_2(P_n)}$. To do this, we will show that on a high-probability event, $\mathcal{L}_{\tilde{\gamma}}(g) > 0$ for all g such that $\|g-\gamma_\psi\|_{L_2(P_n)} > r$ for some radius r. Our main workhorse is the following inequality for $\mathcal{L}_{\tilde{\gamma}}(g)$: for any $\bar{R}$ and $\mathcal{F}$ such that $\bar{R} > R_{n,\lambda\tilde{\mathcal{F}}}(\tilde{\gamma})$ and $\mathcal{F} \supseteq \tilde{\mathcal{F}}$,

$$\begin{aligned}
&\mathcal{L}_{\tilde{\gamma}}(g) \geq \check{\mathcal{L}}(g-\gamma_\psi) - 1\left(\|g\|_{\mathcal{F}} < 1\right)\|g-\gamma_\psi\|_{\mathcal{F}^\star}^2/\lambda^2 \quad \text{for} \\
&\check{\mathcal{L}}(\check{g}) := P_n\check{g}^2 - 2\big|P_n\check{h}(Z,\check{g})\big| + \|\check{g}\|_{\mathcal{F}^\star}^2/\lambda^2 - \bar{R},
\end{aligned} \tag{2.29}$$

where $\mathcal{F}^\star := \mathcal{F} - [0,1]\gamma_\psi$ and $\check{g}$ should be interpreted as short-hand for $g-\gamma_\psi$. In our argument, we will choose $\bar{R}$ and $\mathcal{F}$ to be deterministic, and then verify that the required conditions $\bar{R} > R_{n,\lambda\tilde{\mathcal{F}}}(\tilde{\gamma})$ and $\mathcal{F} \supseteq \tilde{\mathcal{F}}$ hold with high probability. The lower bound (2.29) follows directly from the definition (2.28) once we verify that

$$1\left(\|g\|_{\mathcal{F}} \geq 1\right)\|g-\gamma_\psi\|_{\mathcal{F}^\star}^2 \leq \|g\|_{\tilde{\mathcal{F}}}^2.$$

To do so, first observe that the containment $\mathcal{F} \supseteq \tilde{\mathcal{F}}$ implies that $\|g\|_{\tilde{\mathcal{F}}} \geq \|g\|_{\mathcal{F}}$. Then observe that if $g \in \alpha\mathcal{F}$, $g - \gamma_\psi \in \alpha(\mathcal{F} - \alpha^{-1}\gamma_\psi) \subseteq \alpha\mathcal{F}^\star$ as long as $\alpha^{-1} \in [0,1]$. This implies that $\|g\|_{\tilde{\mathcal{F}}} \geq \|g\|_{\mathcal{F}} \geq \|g-\gamma_\psi\|_{\mathcal{F}^\star}$ whenever $\|g\|_{\mathcal{F}} \geq 1$, which is equivalent to what we wanted to check.

From this point, our argument will be fairly standard, and we will base our presentation on that in Lecué and Mendelson (2017). We will first establish a sort of tightness result, in which we show that for $\check{g}$ outside a $\|\cdot\|_{\mathcal{F}^\star}$-ball, we will have $\check{\mathcal{L}}(\check{g}) > 0$. And with it, we will get a $\|\cdot\|_{L_2(P)}$ bound, although we will express it strangely for a reason that will become clear later when we prove

Lemma 2.8. Our core approach will be to lower bound the difference $P_n\check{g}^2 - 2\left|P_n\check{h}(Z,\check{g})\right|$ between our empirical MSE and our empirical process term as a proportion of the population MSE $P\check{g}^2$. We will first state a purely deterministic result in terms of two uniform-over-$\mathcal{F}^\star$ bounds: a lower bound on the ratio of the empirical and population MSE and an upper bound on our empirical process term. We prove this lemma at the end of this section.

Lemma 2.6. *Let $\mathcal{F}^\star$ be a class of functions mapping support$(P) \to \mathbb{R}$ that is star-shaped around zero and $\{h(z,\cdot) : z \in \text{support}(P)\}$ be a set of linear functionals on the span of $\mathcal{F}^\star$ and and define $\check{\mathcal{L}}(\check{g})$ as in (2.29). Suppose r_Q, η_Q, r_C, and η_C satisfy*

$$\inf_{\check{g}\in\mathcal{F}^\star : P\check{g}^2 \geq r_Q^2} \frac{P_n\check{g}^2}{P\check{g}^2} \geq \eta_C \tag{2.30}$$

$$\sup_{\check{g}\in\mathcal{F}^\star \cap r_C L_2(P)} \left|P_n\check{h}(Z,\check{g})\right| \leq \eta_C r_C^2. \tag{2.31}$$

Then for $r = r_Q \vee r_C \vee \lambda^{-1}\eta_Q^{-1/2}$ and $\alpha = 2\lambda^2\eta_C r^2 + \lambda\bar{R}^{1/2}$, $\check{\mathcal{L}}(\check{g}) > 0$ for all $\check{g}$ satisfying $\|\check{g}\|_{\mathcal{F}^\star} \geq \alpha$. Furthermore, $\check{\mathcal{L}}(\check{g}) > t$ for all $\check{g}$ satisfying $\|\check{g}\|_{\mathcal{F}^\star} \leq \alpha$ and $\|\check{g}\|_{L_2(P)}^2 > \alpha^2 r^2 \vee [\bar{R}+t]/[\eta_Q - 2\alpha^{-1}\eta_C]$.

The given value of α is determined by the behavior of bounds like (2.30) and (2.31) over a scale of classes $s\mathcal{F}^\star$ for $s \in \mathbb{R}_+$.

The condition (2.31) holds with probability $1-\delta$ for $r_C = r_C(\eta_C, \delta)$ as defined in (2.18). To establish (2.30) with high probability, we use the following conveniently rewritten form of Bartlett et al. (2005, Theorem 3.3). It is proven in Appendix A.2.

Lemma 2.7. *Let $\mathcal{F}$ be pointwise separable, star-shaped around zero, and uniformly bounded in sup-norm. For any $\eta_Q \in (0,1)$,*

$$\inf_{f\in\mathcal{F} : Pf^2 \geq r_Q^2} \frac{P_nf^2}{Pf^2} \geq \eta_Q \quad \text{with probability } 1 - \exp\left\{-\frac{c_1(\eta_Q) n r_Q^2}{M_{\mathcal{F}}^2}\right\}$$

with

$$r_Q = c_0(\eta_Q)\inf\left\{r > 0 : R_n\left(\mathcal{F} \cap rL_2(P)\right) \leq \frac{r^2}{2M_{\mathcal{F}}}\right\}$$

and $c_0(\eta_Q) = \frac{\sqrt{24(1+\eta_Q)}}{(1-\eta_Q)}$, $c_1(\eta_Q) = \frac{(1-\eta_Q)^2}{2(1+\eta_Q)(21-11\eta_Q)}$.

Having established conditions under which the assumptions of Lemma 2.6 hold, it will now be straightforward to prove a bound of the form $\|\hat{g} - \gamma_\psi\|_{L_2(P_n)} < a \wedge b$ like the one in Theorem 2.2.

Lemma 2.8. *Suppose that we observe $Z_1 \dots Z_n \overset{iid}{\sim} P$ and that for each $z \in$ support(P), we have a real linear functional $h(z,\cdot)$ acting on the real-valued functions $f(z)$ on support(P). Let $\tilde{\mathcal{F}}$ be an absolutely convex set that may depend on the sample $Z_1 \dots Z_n$ and define $\mathbb{M}_{n,\lambda\tilde{\mathcal{F}}}(g) = -\|g\|_{\tilde{\mathcal{F}}}^2/\lambda^2 - P_ng(Z_i)^2 + 2P_nh(Z_i,g)$.*

Let $\mathcal{F}$ be a nonrandom set of real-valued functions on support(P) that is pointwise measurable and absolutely convex; $\{h(z,f) : f \in \mathcal{F}\}$ also be pointwise measurable; $\psi(\cdot) = \mathbb{E}[h(Z,\cdot)]$ be a continuous linear functional on the space $(\operatorname{span}\mathcal{F}, \|\cdot\|_{L_2(P)})$ and $\gamma_\psi \in \overline{\operatorname{span}}\,\mathcal{F}$ be its Riesz representer; and define $R_{n,\lambda\mathcal{F}}$ as in (2.28), $\mathcal{F}^\star(r) = (\mathcal{F} - [0,1]\gamma_\psi) \cap r L_2(P)$, and $\mathcal{H}^\star(r) = \{z \to h(z,f) - \gamma_\psi(z) f(z) : f \in \mathcal{F}^\star(r)\}$.

Let $\hat{g}$ and $\tilde{\gamma}$ be two random functions on support(P). If $\tilde{\mathcal{F}} \subseteq \mathcal{F}$, $R_{n,\lambda\tilde{\mathcal{F}}}(\tilde{\gamma}) < \bar{R}$, and $\mathbb{M}_{n,\lambda\tilde{\mathcal{F}}}(\hat{g}) \geq \mathbb{M}_{n,\lambda\tilde{\mathcal{F}}}(\tilde{\gamma})$ on an event of probability $1 - 2\delta'$ for some nonrandom $\bar{R} > 0$, then on an event $\mathcal{A}$ of probability $1 - \exp\{-c_2(\eta_Q) n r_Q(\eta_Q)^2 / M_{\mathcal{F}^\star}^2\} - 3\delta - 2\delta'$, $P_n(\hat{g} - \gamma_\psi)^2 \leq a \wedge b$ where

$$
\begin{aligned}
a &= \alpha u(\mathcal{H}^\star, \delta) + \bar{R};\\
b &= 2\alpha^2 r^2 \vee 2\frac{\bar{R} + \lambda^{-2}}{\eta_Q - 2\alpha^{-1}\eta_C} \vee \frac{44 M_{\mathcal{F}^\star}^2 \alpha^2 \log(\delta^{-1})}{n};\\
\alpha &= 1 \vee \left[2\lambda^2 \eta_C r^2 + \lambda \bar{R}^{1/2}\right];\\
r &= r_Q(\eta_Q) \vee r_C(\eta_C, \delta) \vee \lambda^{-1} \eta_Q^{-1/2};
\end{aligned}
\tag{2.32}
$$

for $\eta_Q \in (0, .47)$ and $\eta_C > 0$.

We prove this lemma shortly, using different arguments to establish our bounds a and b. Our bound a will follow from a simple consistency-given-tightness argument: we show that when the empirical MSE is greater than a, it will exceed the centered empirical process term $P_n h(Z, \check{g})$ uniformly over $\check{g} \in \alpha\mathcal{F}^\star$ and therefore imply that the excess loss is positive. Our bound b will follow from the $\|\cdot\|_{L_2(P)}$ from Lemma 2.6.

This gets us nearly to our goal. But this shows convergence of the solution $\hat{g}$ to our dual problem to the Riesz representer γ_ψ, whereas we want convergence of the weights $\hat{\gamma}$ minimizing $\ell_{n,\lambda\tilde{\mathcal{F}}}$ to γ_ψ. By Lemma 2.5, this is equivalent when $\tilde{\mathcal{F}}$ is reflexive. The following lemma, proven in Appendix A.2, uses a finite dimensional approximation argument to show that reflexiveness is not necessary.

Lemma 2.9. *Under the assumptions of Lemma 2.5 excepting reflexiveness, the assumptions of Lemma 2.8 with the condition $\mathbb{M}_{n,\lambda\tilde{\mathcal{F}}}(\hat{g}) \geq \mathbb{M}_{n,\lambda\tilde{\mathcal{F}}}(\tilde{\gamma})$ involving $\hat{g}$ dropped, and the additional assumption that $\mathcal{F}$ is totally bounded in $\|\cdot\|_\infty$, the weights $\hat{\gamma}$ minimizing the primal $\ell_{n,\lambda\tilde{\mathcal{F}}}$ satisfy $P_n(\hat{\gamma}_i - \gamma_\psi(Z_i))^2 \leq a \wedge b$ on $\mathcal{A}$ with those quantities defined as in Lemma 2.8.*

We conclude our proof of our theorem's first claim by establishing a specific value of $\bar{R}$ to use in this bound. To do this, we make use of our theorem's assumption that $\tilde{\mathcal{F}}$ satisfies $\mathcal{F}_L \subseteq \tilde{\mathcal{F}} \subseteq \mathcal{F}$ on an event of probability $1 - \delta'$. On the event, $R_{n,\lambda\tilde{\mathcal{F}}} \leq R_{n,\lambda\mathcal{F}_L}$. Therefore given $\bar{R}$ such that for some $\tilde{\gamma}$, $R_{n,\lambda\mathcal{F}_L}(\tilde{\gamma}) \leq \bar{R}$ on an event of probability $1 - \delta'$, the conditions $\tilde{\mathcal{F}} \subseteq \mathcal{F}$ and $R_{n,\lambda\mathcal{F}}(\tilde{\gamma}) \leq \bar{R}$ will be satisfied on the interestion of these events which has probability $1 - 2\delta'$ as required. To choose

$\bar{R}$ satisfying this condition for a deterministic function $\tilde{\gamma}$, we use the following bound, proven in Appendix A.2.

Lemma 2.10. *Under the assumptions of Theorem 2.2, with probability* $1-\delta$,

$$\begin{aligned} &R_{n,\,\lambda\mathcal{F}_L}(\tilde{\gamma}) < 2\delta^{-1}[\kappa^2 + 2\lambda n^{-1/2}\kappa\overline{\sigma}(\mathcal{H}^\star(\kappa))] + 2^{3/2}\delta^{-1/2}n^{-1/2}\overline{\sigma}(\mathcal{H}^\star(\kappa)), \\ &\kappa^2(\tilde{\gamma}) = \|\tilde{\gamma} - \gamma_\psi\|_{L_2(P)}^2 + \delta\|\tilde{\gamma}\|_{\mathcal{F}_L}^2/(2\lambda^2). \end{aligned} \tag{2.33}$$

Letting $\kappa_\star^2$ be the infimum of $\kappa_\star^2$, for any $\epsilon > 0$ we may take $\bar{R}$ to be the value of our bound at a point $\tilde{\gamma}$ with $\kappa(\tilde{\gamma}) = \kappa_\star + \epsilon$. And then, as the linearity of $h(Z,\cdot)$ implies the continuity of $\kappa \to \overline{\sigma}(\mathcal{H}^\star(\kappa))$, the effect of this ϵ on our bound $\bar{R}$ is infinitesimal. To state a cleaner result, we increase our factor of $2^{3/2}$ to 4 and drop this ϵ in our statement of Theorem 2.2.

We close the section with proofs of our core lemmas, Lemma 2.6 and Lemma 2.8.

Proof of Lemma 2.6. We will first prove the following claim. Suppose we have the bounds

$$\inf_{\check{g}\in\alpha\mathcal{F}^\star : P\check{g}^2 \geq r_{Q,\alpha}^2} \frac{P_n\check{g}^2}{P\check{g}^2} \geq \eta_{Q,\alpha} \tag{2.34}$$

$$\sup_{\check{g}\in\alpha\mathcal{F}^\star} \frac{|P_n\check{h}(Z,\check{g})|}{P\check{g}^2 \vee r_{C,\alpha}^2} \leq \eta_{C,\alpha}. \tag{2.35}$$

Then if $\eta_{Q,\alpha} > 2\eta_{C,\alpha}$ and $(r_{Q,\alpha}\vee r_{C,\alpha})^2 < \frac{\alpha^2/\lambda^2 - \bar{R}}{2\eta_{C,\alpha}}$, $\check{\mathcal{L}}(\check{g}) > 0$ for all $\check{g}$ satisfying $\|\check{g}\|_{\mathcal{F}^\star} \geq \alpha$ and furthermore $\check{\mathcal{L}}(\check{g}) > t$ for all $\check{g}$ satisfying $\|\check{g}\|_{\mathcal{F}^\star} \leq \alpha$ and $\|\check{g}\|_{L_2(P)}^2 > (r_{Q,\alpha}\vee r_{C,\alpha})^2 \vee [\bar{R}+t]/[\eta_{Q,\alpha} - 2\eta_{C,\alpha}]$.

To prove this claim, we begin by showing that $\check{\mathcal{L}}(\check{g}) > 0$ for all $\check{g}$ in the sphere $\alpha\mathcal{S} := \{\check{g} : \|g\|_{\mathcal{F}^\star} = \alpha\}$. If $\check{g} \in \alpha\mathcal{S}$ satisfies $P\check{g}^2 \geq (r_{Q,\alpha}\vee r_{C,\alpha})^2$, then $\check{\mathcal{L}}(\check{g}) \geq [\eta_{Q,\alpha} - 2\eta_{C,\alpha}]\,P\check{g}^2 + \left[\alpha^2/\lambda^2 - \bar{R}\right] > 0$. All other $\check{g} \in \alpha\mathcal{S}$ satisfy $P\check{g}^2 \leq (r_{Q,\alpha}\vee r_{C,\alpha})^2$, in which case $\check{\mathcal{L}}(\check{g}) \geq -2\eta_{C,\alpha}(r_{Q,\alpha}\vee r_{C,\alpha})^2 + \alpha^2/\lambda^2 - \bar{R} > 0$ under our assumption $(r_{Q,\alpha}\vee r_{C,\alpha})^2 < \frac{\alpha^2/\lambda^2-\bar{R}}{2\eta_{C,\alpha}}$.

We will now extend this result to show that $\check{\mathcal{L}} > 0$ outside the sphere $\alpha\mathcal{S}$ as well, on the set $\{\check{g}' : \|\check{g}'\|_{\mathcal{F}^\star} > \alpha\}$. Because $\mathcal{F}^\star$ is star-shaped around zero, any point $\check{g}'$ with $\|\check{g}'\|_{\mathcal{F}^\star} < \infty$ can be written in the form $\check{g}' = R\check{g}$ for $\check{g} \in \alpha\mathcal{S}$, and the aforementioned points outside the sphere may be written in this form for $R > 1$. Consider such a point.

$$\begin{aligned} \check{\mathcal{L}}(R\check{g}) &= R^2 P_n\check{g}^2 - 2R|P_n\check{h}(Z,\check{g})| + R^2\|\check{g}\|_{\mathcal{F}^\star}^2/\lambda^2 - \bar{R} \\ &\geq R^2\left[P_n\check{g}^2 - 2|P_n\check{h}(Z,\check{g})| + \|\check{g}\|_{\mathcal{F}^\star}^2/\lambda^2 - \bar{R}\right] \\ &= R^2\check{\mathcal{L}}(\check{g}) > 0. \end{aligned}$$

Consequently, under the stated conditions $\check{\mathcal{L}}(\check{g}) > 0$ if $\|\check{g}\|_{\mathcal{F}^\star} \geq \alpha$ as claimed.

We will complete our proof of this initial claim by checking that $\check{\mathcal{L}}(\check{g}) > t$ when $\|\check{g}\|_{\mathcal{F}^\star} \leq \alpha$ and $\|\check{g}\|^2_{L_2(P)} > (r_{Q,\alpha} \vee r_{C,\alpha})^2 \vee [\bar{R} + t]/[\eta_{Q,\alpha} - 2\eta_{C,\alpha}]$. For such $\check{g}$, $\check{\mathcal{L}}(\check{g}) \geq [\eta_{Q,\alpha} - 2\eta_{C,\alpha}]\, P\check{g}^2 - \bar{R}$, and this exceeds t because $P\check{g}^2 > [\bar{R} + t]/[\eta_{Q,\alpha} - 2\eta_{C,\alpha}]$.

Our initial claim proven, we will now establish that its assumptions hold under the assumptions of our Lemma. First, observe that (2.35) is implied by the bound

$$\sup_{\check{g} \in \alpha\mathcal{F}^\star \cap r_C L_2(P)} \left| P_n \check{h}(Z, \check{g}) \right| \leq \eta_C r_C^2. \tag{2.36}$$

This follows from an argument used in the proof of Mendelson (2014, Theorem 3.1), which we restate for convenience. For $\|\check{g}\|_{L_2(P)} \leq r_C$, the bound above directly implies $|P_n \check{h}(Z, \check{g})| \leq \eta_C r_C^2$. For $\|\check{g}\|_{L_2(P)} \geq r_C$, we may apply (2.36) to $\check{g}' = (r_C / \|\check{g}\|_{L_2(P)})\check{g}$, which satisfies the condition $\|\check{g}'\|_{L_2(P)} \leq r_C$ by construction and is in $\mathcal{F}^\star$ because it is a scaled-down version of $\check{g}$ and $\mathcal{F}^\star$ is star-shaped around zero. Therefore

$$\left| P_n \check{h}(Z, \check{g}) \right| = \left| P_n \check{h}(Z, \check{g}') \right| \frac{\|\check{g}\|_{L_2(P)}}{r_C} \leq \eta_C r_C^2 \frac{\|\check{g}\|_{L_2(P)}}{r_C} \leq \eta_C \|\check{g}\|^2_{L_2(P)}.$$

Taking the maximum of the upper bounds for the two cases $\|\check{g}\|_{L_2(P)} \leq r_C$ and $\|\check{g}\|_{L_2(P)} \geq r_C$ gives a bound $|P_n \check{h}(Z, \check{g})| \leq \eta_C (r_C^2 \vee P\check{g}^2)$ valid for all $\check{g} \in \alpha\mathcal{F}^\star$ and therefore our claimed bound (2.35).

Because the ratio $P_n\check{g}^2 / P\check{g}^2$ is invariant to scale,

$$\inf_{\check{g} \in \mathcal{F}^\star : P\check{g}^2 \geq r_Q^2} \frac{P_n \check{g}^2}{P\check{g}^2} \geq \eta_Q \iff \inf_{\check{g} \in \alpha\mathcal{F}^\star : P\check{g}^2 \geq (\alpha r_Q)^2} \frac{P_n \check{g}^2}{P\check{g}^2} \geq \eta_Q.$$

Similarly, scaling (2.36) by α gives

$$\sup_{\check{g} \in \mathcal{F}^\star \cap r_C L_2(P)} \left| P_n \check{h}(Z, \check{g}) \right| \leq \eta_C r_C^2 \iff \sup_{\check{g} \in \alpha\mathcal{F}^\star \cap \alpha r_C L_2(P)} \left| P_n \check{h}(Z, \check{g}) \right| \leq (\eta_C / \alpha)(\alpha r_C)^2.$$

Therefore under the assumptions of our Lemma, the conditions (2.34) and (2.35) for our claim are satisfied with parameters $\eta_{Q,\alpha} = \eta_Q$, $\eta_{C,\alpha} = \eta_C / \alpha$, $r_{Q,\alpha} = \alpha r_Q$, $r_{C,\alpha} = \alpha r_C$.

For those parameters, the additional condition $(r_{Q,\alpha} \vee r_{C,\alpha})^2 < \frac{\alpha^2/\lambda^2 - \bar{R}}{2\eta_{C,\alpha}}$ can be equivalently written as the quadratic inequality $\alpha^2/\lambda^2 - 2\eta_C r^2 \alpha - \bar{R} > 0$ for $r = r_Q \vee r_C$. This convex quadratic function of α has one positive and one negative root, so it will be positive for $\alpha > 0$ iff α exceeds its positive root

$$\frac{2\eta_C r^2 + \sqrt{4\eta_C^2 r^4 + 4\bar{R}/\lambda^2}}{2/\lambda^2} = \lambda^2 \left[\eta_C r^2 + \sqrt{\eta_C^2 r^4 + \bar{R}/\lambda^2} \right].$$

Because $\sqrt{a+b} < \sqrt{a} + \sqrt{b}$ for $a, b > 0$, the condition $\alpha \geq 2\lambda^2 \eta_C r^2 + \lambda\sqrt{\bar{R}}$ is sufficient.

The final condition for our initial claim is $\eta_{Q,\alpha} > 2\eta_{C,\alpha}$, i.e., $\alpha\eta_Q > 2\eta_C$. For $\alpha \geq 2\lambda^2\eta_C r^2 + \lambda\sqrt{\bar{R}}$, it suffices to take η_C satisfying $(2\lambda^2\eta_C r^2 + \lambda\sqrt{\bar{R}})\eta_Q > 2\eta_C$ or equivalently $2(\lambda^2 r^2 \eta_Q - 1)\eta_C + \lambda\sqrt{\bar{R}}\eta_Q > 0$, which is satisfied for all η_C when $r \geq \lambda^{-1}\eta_Q^{-1/2}$. □

Proof of Lemma 2.8. To simplify our proof, we will assume that $u(\mathcal{H}^\star(r_C(\eta_C,\delta)),\delta) \leq \eta_C r_C(\eta_C,\delta)^2$, i.e. that the infimum defining $r_C(\eta_C,\delta)$ is attained. We will work on an event $\mathcal{A}$ on which $\tilde{\mathcal{F}} \subseteq \mathcal{F}$, $R_{n,\lambda\tilde{\mathcal{F}}}(\tilde{\gamma}) \leq \bar{R}$, and $\mathbb{M}_{n,\lambda\tilde{\mathcal{F}}}(\hat{g}) \geq \mathbb{M}_{n,\lambda\tilde{\mathcal{F}}}(\tilde{\gamma})$; the conditions (2.30) and (2.31) for Lemma 2.6 are satisfied; and we have

$$\sup_{h\in\mathcal{H}^\star} |P_n h| < u(\mathcal{H}^\star,\delta); \tag{2.37}$$

$$\mathcal{F}^\star \cap r_E L_2(P) \subseteq \mathcal{F}^\star \cap \sqrt{2} r_E L_2(P_n) \quad \text{for} \quad r_E = \sqrt{\frac{22M_{\mathcal{F}^\star}^2 \log(\delta^{-1})}{n}} \vee R_n^\star\left(\frac{1}{20M_{\mathcal{F}^\star}}, \mathcal{F}^\star(\cdot)\right). \tag{2.38}$$

Our first set of three conditions is satisfied w.p. $1-2\delta'$ by assumption; the conditions (2.30) and (2.31) hold w.p. $1-\exp\{-c_2(\eta_Q)nr_Q(\eta_Q)^2/M_{\mathcal{F}^\star}^2\}$ and $1-\delta$ respectively by Lemmas 2.7 and our definition of $u(\cdot,\delta)$; (2.37) holds with probability $1-\delta$ again by our definition of $u(\cdot,\delta)$; and (2.38) holds with probability $1-\delta$ by Bartlett et al. (2005, Corollary 2.2). Consequently, by the union bound this event $\mathcal{A}$ has probability $1-\exp\{-c_2(\eta_Q)nr_Q(\eta_Q)^2/M_{\mathcal{F}^\star}^2\} - 3\delta - 2\delta'$.

We have set up our problem so that $\hat{g}$ satisfies $\mathcal{L}_{\tilde{\gamma}}(\hat{g}) \leq 0$, so we will derive bounds on $\hat{g}$ from conditions on g that rule out the possibility that $\mathcal{L}_{\tilde{\gamma}}(g) \leq 0$. We will work with the lower bound

$$\mathcal{L}_{\tilde{\gamma}}(g) \geq \begin{cases} \check{\mathcal{L}}(g-\gamma_\psi) & \text{if } \|g-\gamma_\psi\| > 1 \\ \check{\mathcal{L}}(g-\gamma_\psi) - \lambda^{-2} & \text{if } \|g-\gamma_\psi\| \leq 1 \end{cases}.$$

This follows from (2.29), as because $g \in \mathcal{F} \implies g-\gamma_\psi \in \mathcal{F}^\star$, we have $1\left(\|g\|_{\mathcal{F}} < 1\right)\|g-\gamma_\psi\|_{\mathcal{F}^\star}^2 = 0$ if $\|g-\gamma_\psi\|_{\mathcal{F}^\star} > 1$ and $1\left(\|g\|_{\mathcal{F}} < 1\right)\|g-\gamma_\psi\|_{\mathcal{F}^\star}^2 \leq 1$ otherwise.

First, consider the case that $\|g-\gamma_\psi\|_{\mathcal{F}^\star} > \alpha$. Then as $\alpha \geq 1$, $\mathcal{L}_{\tilde{\gamma}}(g) \geq \check{\mathcal{L}}(g-\gamma_\psi)$, and by Lemma 2.6, $\check{\mathcal{L}}(g-\gamma_\psi) > 0$. It follows that $\hat{g}$ must satisfy $\|\hat{g}-\gamma_\psi\|_{\mathcal{F}^\star} \leq \alpha$.

Now consider the case that $\|g-\gamma_\psi\|_{\mathcal{F}^\star} \leq \alpha$. Substituting into our definition (2.28) of $\mathcal{L}_{\tilde{\gamma}}(g)$ our bounds (2.37) and $R_{n,\lambda\tilde{\mathcal{F}}}(\tilde{\gamma}) \leq \bar{R}$, we have $\mathcal{L}_{\tilde{\gamma}}(g) > \|g-\gamma_\psi\|_{L_2(P_n)}^2 - 2\alpha u(\mathcal{H}^\star,\delta) - \bar{R}$. Thus, we will have $\mathcal{L}_{\tilde{\gamma}}(g) > 0$ if $\|g-\gamma_\psi\|_{L_2(P_n)}^2 > 2\alpha u(\mathcal{H}^\star,\delta) + \bar{R}$. This implies our bound a on $\|\hat{g}-\gamma_\psi\|_{L_2(P_n)}^2$.

Finally, consider again the case that $\|g-\gamma_\psi\|_{\mathcal{F}^\star} \leq \alpha$. $\mathcal{L}_{\tilde{\gamma}}(g) \geq \check{\mathcal{L}}(g-\gamma_\psi) - \lambda^{-2}$ will be strictly positive if $\check{\mathcal{L}}(g-\gamma_\psi) > \lambda^{-2}$. By Lemma 2.6, this will happen if $\|g-\gamma_\psi\|_{L_2(P)}^2 > \alpha^2 r^2 \vee [\bar{R}+\lambda^{-2}]/[\eta_Q - 2\alpha^{-1}\eta_C]$. And by (2.38), this will happen if $\|g-\gamma_\psi\|_{L_2(P_n)}^2 > 2\alpha^2(r\vee r_E)^2 \vee 2[\bar{R}+\lambda^{-2}]/[\eta_Q - 2\alpha^{-1}\eta_C]$. This implies that $\|\hat{g}-\gamma_\psi\|_{L_2(P_n)}^2$ is no larger than the right side above.

To derive our bound b, we upper bound the right side by something without this new constant r_E. To do this, first separate out the two components of r_E, writing this quantity as

$$2\alpha^2(r\vee s_E)^2 \vee \frac{22\alpha^2 M_{\mathcal{F}^\star}^2 \log(\delta^{-1})}{n} \vee 2\frac{\bar{R}+\lambda^{-2}}{\eta_Q - 2\alpha^{-1}\eta_C} \quad \text{for} \quad s_E = R_n^\star\left(\frac{1}{20M_{\mathcal{F}^\star}}, \mathcal{F}^\star(\cdot)\right).$$

Then we will bound s_E in terms of r_Q, which we will write $r_Q = cs_Q$ where $c = \frac{\sqrt{14(1+\eta_Q)}}{1-\eta_Q}$ and $s_Q = R_n^\star\left(\frac{1}{2M_{\mathcal{F}^\star}}, \mathcal{F}^\star(\cdot)\right)$. Recall the definition $R_n^\star(\eta, \mathcal{F}^\star(\cdot)) = \inf\{s > 0 : \tau(s) \leq \eta s^2\}$ for $\tau(s) = R_n(\mathcal{F}^\star \cap sL_2(P))$. Here the ratio $\tau(s)/s$ is decreasing (Bartlett et al., 2005, Lemma 3.4) and these infima are attained with equality, i.e. $\tau(s_Q) = s_Q^2/(2M_{\mathcal{F}^\star})$ and $\tau(s_E) = s_E^2/(20M_{\mathcal{F}^\star})$ (Bartlett et al., 2005, Lemma 3.2). Because s_E satisfies $\tau(s) \leq s^2/(2M_{\mathcal{F}^\star})$ and s_Q is the minimal point with this property, we have $s_E \geq s_Q$, and therefore $s_E/(20M_{\mathcal{F}^\star}) = \tau(s_E)/s_E \leq \tau(s_Q)/s_Q = s_Q/(2M_{\mathcal{F}^\star})$ and therefore $s_E \leq 10s_Q = (10/c)r_Q$. This constant $(10/c)$ will be less than one if $\eta_Q \leq .47$, so we simply add this restriction and drop s_E from the bound above. □

2.1.3.4 Bounding the bias term

In this section, we will use our bound $P_n(\hat{\gamma}_i - \gamma_\psi)^2 \leq a \wedge b$ to bound the quantity $I_{h,\tilde{\mathcal{F}}}(\hat{\gamma})$. We will work on the intersection $\mathcal{A}'$ of the event $\mathcal{A}$ from Lemma 2.8 and an event on which $\sup_{h\in\mathcal{H}} P_n h < u(\mathcal{H}, \delta)$. As this new condition holds with probability $1 - \delta$, our new event $\mathcal{A}'$ has probability $1 - \exp\{-c_2(\eta_Q)nr_Q(\eta_Q)^2/M_{\mathcal{F}^\star}^2\} - 4\delta - 2\delta'$ by the union bound.

Recall from our sketch that

$$I_{h,\tilde{\mathcal{F}}} \leq I_{h,\tilde{\mathcal{F}}}(\gamma^\star)^2 + \frac{\sigma^2}{n^2}\sum_{i=1}^n \gamma_\psi(Z_i)^2 - \hat{\gamma}_i^2 \quad \text{where} \quad \gamma_i^\star = \gamma_\psi(Z_i).$$

To bound the first term of the right side, observe that because $\tilde{\mathcal{F}} \subseteq \mathcal{F}$, $I_{h,\tilde{\mathcal{F}}}(\gamma^\star) \leq I_{h,\mathcal{F}}(\gamma^\star) = \sup_{h\in\mathcal{H}} P_n h \leq u(\mathcal{H}, \delta)$. To bound the second term, we use the elementary identity $a^2 - b^2 = 2a(a-b) - (a-b)^2$. Using this and Cauchy-Schwartz,

$$\frac{1}{n}\sum_{i=1}^n \gamma_\psi(Z_i)^2 - \hat{\gamma}_i^2 \leq 2\|\gamma_\psi\|_{L_2(P_n)}\|\gamma_\psi - \hat{\gamma}\|_{L_2(P_n)} \leq 2\|\gamma_\psi\|_{L_2(P_n)}(a \wedge b)^{1/2}.$$

Thus, using the elementary inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$,

$$I_{h,\tilde{\mathcal{F}}} \leq u(\mathcal{H}^\star, \delta) + 2^{1/2}\sigma\|\gamma_\psi\|_{L_2(P_n)}^{1/2} n^{-1/2}(a \wedge b)^{1/4}.$$

2.1.3.5 Convergence of the noise term

In this section, we will use our bound $P_n(\hat{\gamma}_i - \gamma_\psi)^2 \leq a \wedge b$ to bound the difference between our noise term and the iid sum $P_n\gamma_\psi(Z_i)\varepsilon_i$, $\varepsilon_i = Y_i - m(Z_i)$. Because $\hat{\gamma}$ is a function of $Z_1 \ldots Z_n$, we can apply Chebyshev's inequality conditionally on $Z_1 \ldots Z_n$ to the difference between our noise term and this sum. With conditional probability $1 - \delta$,

$$|P_n(\hat{\gamma}_i - \gamma_\psi(Z_i))\varepsilon_i| \leq \delta^{-1/2}n^{-1/2}\sqrt{P_n[\hat{\gamma} - \gamma_\psi(Z_i)]^2 v(Z_i)}.$$

If we instead do this with an indicator for our event $\mathcal{A}$, on which $\|\hat{\gamma} - \gamma_\psi\|_{L_2(P_n)} \leq (a \wedge b)^{1/2}$, and apply Cauchy-Schwarz to the inner product appearing in the right side above, we get the bound

$$1_{\mathcal{A}} |P_n(\hat{\gamma}_i - \gamma_\psi(Z_i))\varepsilon_i| \leq 1_{\mathcal{A}} \delta^{-1/2} n^{-1/2} \|\hat{\gamma} - \gamma_\psi\|_{L_2(P_n)} \|v\|_{L_2(P_n)} \leq \delta^{-1/2} n^{-1/2} (a \wedge b)^{1/2} \|v\|_\infty.$$

This last bound does not depend on $Z_1 \dots Z_n$ and therefore holds unconditionally. Thus, on the intersection of our event $\mathcal{A}'$ from the previous section and our probability $1 - \delta$ event here and therefore with probability $1 - \exp\{-c_2(\eta_Q) n r_Q(\eta_Q)^2 / M_{\mathcal{F}^\star}^2\} - 5\delta - 2\delta'$, all of our theorem's claims hold.

2.2 Example: Estimating Average Partial Effects

As a concrete instantiation of our augmented minimax linear approach, we consider the problem of average partial effect estimation in the conditionally linear treatment effect model: A statistician observes features $X \in \mathcal{X}$, a treatment assignment $W \in \mathbb{R}$ and an outcome $Y \in \mathbb{R}$ related by a functional form restriction as below and wants to estimate ψ, where

$$\mathbb{E}\left[Y \,\middle|\, X = x, W = w\right] = \mu(x) + w\,\tau(x), \quad \psi = \mathbb{E}\left[\tau(X)\right]. \tag{2.39}$$

By Proposition 2.3, our AML estimator will be efficient for ψ under regularity conditions when $\mathrm{Var}\left[Y_i \,\middle|\, X_i, W_i\right] = \sigma^2(X_i)$ is only a function of X_i.

In the classical case of an unconfounded binary treatment, the model (2.39) is general and the estimand ψ corresponds to the average treatment effect (Rosenbaum and Rubin, 1983; Imbens and Rubin, 2015). At the other extreme, if W is real valued but $\tau(x) = \tau$ is constrained not to depend on x, then (2.39) reduces to the partially linear model as studied by Robinson (1988). The specific model (2.39) has recently been studied by Athey, Tibshirani, and Wager (2018) and Zhao, Small, and Ertefaie (2017b). We consider the motivation for (2.39) further in Section 2.3 in the context a real-world application; here, we focus on estimating ψ for this model.

Both $\mu(\cdot)$ and $\tau(\cdot)$ in the model (2.39) are assumed to have finite gauge with respect to an absolutely convex class $\mathcal{M}$, and we define

$$\mathcal{F}_{\mathcal{M}} = \left\{ m(\cdot) : m(x, w) = \mu(x) + w\tau(x),\ \|\mu\|_{\mathcal{M}}^2 + \|\tau\|_{\mathcal{M}}^2 \leq 1 \right\}. \tag{2.40}$$

Then we can define a minimax linear estimator conditional on X and W,

$\hat{\psi}_{MLIN} = \frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}_i Y_i$ with

$$\hat{\gamma} = \text{argmin}\left\{ \frac{\|\gamma\|^2}{n^2} + \sup_{\mu\in\mathcal{M}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}_i \mu(X_i) \right\}^2 + \sup_{\tau\in\mathcal{M}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left(W_i\hat{\gamma}_i - 1\right)\tau(X_i) \right\}^2 \right\}. \tag{2.41}$$

Given any estimators $\hat{\mu}(\cdot)$ and $\hat{\tau}(\cdot)$, we can define an augmented minimax linear estimator

$$\hat{\psi}_{AML} = \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\tau}(X_i) - \hat{\gamma}_i \left(\hat{\mu}(X_i) + W_i\hat{\tau}(X_i) - Y_i\right)\right). \tag{2.42}$$

And as the Riesz representer can be shown to have the form $\gamma_\psi(x, w) = (w - e(x))/v(x)$ with $e(x) = \mathbb{E}\left[W \,\middle|\, X = x\right]$ and $v(x) = \text{Var}\left[W \,\middle|\, X = x\right]$, we also consider a natural doubly robust estimator based on plug-in estimates of these quantities,[11]

$$\hat{\psi}_{DR} = \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\tau}(X_i) - \left(\frac{W_i - \hat{e}(X_i)}{\hat{v}(X_i)}\right)\left(\hat{\mu}(X_i) + W_i\hat{\tau}(X_i) - Y_i\right)\right). \tag{2.43}$$

The goal of our simulation study is to compare the relative merits of minimax linear, augmented minimax linear, and plug-in doubly robust estimation of the average partial effect.[12]

All experiments can be replicated using the `R` package **`amlinear`**. We computed minimax linear weights via the cone solver `ECOS` (Domahidi, Chu, and Boyd, 2013), available in `R` via the package `CVXR` (Fu et al., 2017). When needed, we run penalized regression using the `R` package **`glmnet`** (Friedman, Hastie, and Tibshirani, 2010).

2.2.1 Simulation Design

In all our simulations, we start by generating data $(X_i,\, Y_i,\, W_i)$, such that the expectation of Y_i and W_i has a non-linear dependence on a low-dimensional set of covariates X_i. We then fit our signal of interest using a sparse linear combination of transformations Ψ_i of the original features X_i. We considered data-generating distributions of the form

$$X_i \sim \mathcal{N}\left(0,\, I_{d\times d}\right), \quad W_i \,\middle|\, X_i \sim \mathcal{L}_{X_i}, \quad Y_i \,\middle|\, X_i,\, W_i = \mathcal{N}\left(b(X_i) + W_i\tau(X_i),\, 1\right), \tag{2.44}$$

for different choices of dimension d, treatment assignment distribution $\mathcal{L}_{X_i}$, baseline main effect $b(\cdot)$ and treatment effect function $\tau(\cdot)$. We considered the following 4 setups, each of which depends on a sparsity level k that controls the complexity of the signal.

[11] For example, a random forest version of this estimator is available in the **`grf`** package of Athey, Tibshirani, and Wager (2018).

[12] In the binary treatment assignment case $W_i \in \{0, 1\}$, we know that $v(x) = e(x)(1 - e(x))$; and if we set $\hat{v}(x) = \hat{e}(x)(1 - \hat{e}(x))$, then the estimator in (2.43) is equivalent to the augmented inverse-propensity weighted estimator of Robins, Rotnitzky, and Zhao (1994). For more general W_i, however, $v(x)$ is not necessarily determined by $e(x)$ and so we need to estimate it separately.

1. Beta-distributed treatment, $W_i \,|\, X_i \sim B(\alpha(X_i),\, \beta(X_i))$, with $\zeta(x) = \sum_{j=1}^{k} x_j/\sqrt{k}$, $\psi(x) = \text{sign}(\zeta(x))\zeta^2(x)$, $\alpha(x) = \max\{0.05,\, \min\{0.95,\, 1/(1+\exp[-\psi(x)])\}\}$, $\beta(x) = 1-\alpha(x)$, $b(x) = \psi(x) + 0.2(\alpha(x)-0.5)$, and $\tau(x) = -0.2$.

2. Scaled Gaussian treatment, $W_i \,|\, X_i \sim \mathcal{N}\left(\lambda(X_i),\, \lambda^2(X_i)\right)$, with $\psi(x) = 2^{k-1}\prod_{j=1}^{k} x_j$, $b(x) = \text{sign}(\psi(x))\sqrt{|\psi(x)|}$, $\lambda(x) = 0.1\,\text{sign}(b(x)) + b(x)$, and $\tau(x) = \max\{x_1 + x_2,\, 0\}/2$.

3. Poisson treatment, $W_i \,|\, X_i \sim \text{Poisson}(\lambda(X_i))$, with $\tau(x) = k^{-1}\sum_{j=1}^{k}\cos(\pi x_j/3)$, $\lambda(x) = 0.2 + \tau^2(x)$, and $b(x) = 4d^{-1}\sum_{j=1}^{d} x_j + 2\lambda(x)$.

4. Log-normal treatment, $\log(W_i) \,|\, X_i \sim \mathcal{N}\left(\lambda(X_i),\, 1/3^2\right)$, with $\zeta(x) = \sum_{j=1}^{k} x_j/\sqrt{k}$, $b(x) = \max\{0,\, 2\zeta(x)\}$, $\lambda(x) = 1/(1+\exp[-\text{sign}(\zeta(x))\zeta^2(x)])$, and $\tau(x) = \sin(2\pi x_1)$.

2.2.2 Methods under Comparison

We first consider two variants of the **minimax linear** estimator. The simpler option is minimax over the class $\mathcal{F}_{\mathcal{M}}$ described in (2.40) where $\mathcal{M}$ is defined in terms of a basis expansion Ψ of our covariates,

$$\mathcal{M} = \left\{ f(x) : f(x) = \sum_{j=1}^{\infty} \beta_j \psi_j(x),\ \ \sum_{j=1}^{\infty} |\beta_j| \le 1 \right\}. \tag{2.45}$$

Throughout, we use a basis sequence $\psi_j = a_j \psi'_j$, where ψ'_j are d-dimensional interactions of standardized Hermite polynomials that are orthonormal with respect to the standard Gaussian distribution. The sequence of weights $\{a_j\}$ varies with order k of the polynomial ψ_j; $a_j = 1/(k\sqrt{n_{k,d}})$ where $n_{k,d}$ is the number of terms of order k. Observe that $\sum_{j=1}^{\infty} a_j^2 = 1$ and therefore, for standard normal X, $\sum_{j=1}^{\infty} \mathbb{E}\,\psi_j(X)^2 = 1$. It follows that if the density of X with respect to Gaussian measure is bounded, $\sum_{j=1}^{\infty} \mathbb{E}\,\psi_j(X)^2 < \infty$, and so $\mathcal{M}$ is Donsker. When W_i is bounded, this implies that $\mathcal{F}_{\mathcal{M}}$ is also Donsker; see, e.g., van der Vaart and Wellner (1996, Section 2.13.2 and Section 2.10).

Then, motivated by popular idea of propensity-stratified estimation in the causal inference literature (Rosenbaum and Rubin, 1984), we consider minimax linear estimation over the expanded class $\mathcal{F}_{\mathcal{M}_+}$ where $\mathcal{M}_+$ extends $\mathcal{M}$ by adding to our basis expansion Ψ the following random basis functions:

- Multi-scale strata of the estimated average treatment intensity $\hat{e}(X_i)$ (we balanced over histogram bins of length 0.05, 0.1, and 0.2),

- Basis elements obtained by depth-3 recursive dyadic partitioning (i.e., pick a feature, split along its median, and recurse), and

- Leaves generated by a regression tree on the W_i (Breiman et al., 1984).

The idea behind using this expanded class is that we may be able to improve the practical performance of the method by opportunistically adding a small number of basis functions that help mitigate bias in case of misspecification (i.e., when μ and τ do not have finite gauge $\|\cdot\|_{\mathcal{M}}$). The motivation for focusing on transformations of $\hat{e}(X_i)$ is that accurately stratifying on $e(X_i)$ would suffice to eliminate all confounding in the model (2.39).[13] We emphasize that this estimator is a heuristic method motivated by popular ideas in the applied literature, and is not covered by the formal results developed in this paper.

The remaining methods we consider all combine a regression adjustment $(\hat{\mu}(x),\, \hat{\tau}(x))$ with various weighting schemes. To get such regression adjustments, we first fit the conditional marginal response functions $\mathbb{E}\left[Y_i \,\middle|\, X_i = x\right]$ and $e(x)$ via a cross-validated lasso (Tibshirani, 1996) on the design Ψ. We then fit the $\tau(x)$ function via the R-lasso method proposed by Nie and Wager (2017), again on Ψ, and finally set $\hat{\mu}(x) = \widehat{\mathbb{E}}\left[Y_i \,\middle|\, X_i = x\right] - \hat{\tau}(x)\hat{e}(x)$. As discussed in Nie and Wager (2017), this method is appropriate when the treatment effect function $\tau(x)$ is simpler than $\mathbb{E}\left[Y_i \,\middle|\, X_i = x\right]$ and $e(x)$, and allows for faster rates of convergence on $\tau(x)$ than the other regression components whenever the nuisance components can be estimated at $o_p(n^{-1/4})$ rates in root-mean squared error. We use the same regression adjustment for all 4 methods listed below. Note that we only use the basis Ψ for this regression; we do not use the random basis functions that we used to define $\mathcal{M}_+$.

We consider an **augmented minimax linear** estimator that combines this regression adjustment with minimax linear weights as in (2.42), as well as **augmented minimax linear estimation over an extended class** that uses the same functional form but with the minimax linear weights for $\mathcal{F}_{\mathcal{M}_+}$ instead of $\mathcal{F}_{\mathcal{M}}$. We also consider the **plug-in doubly robust** estimator defined in (2.43), where $\hat{v}(\cdot)$ is estimated via a separate lasso on Ψ as above, as well as an **oracle doubly robust** estimator that uses the same functional form (2.43) but with oracle values of $e(X_i)$ and $v(X_i)$.

2.2.3 Results

We first compare the two minimax linear estimators with the corresponding augmented minimax linear estimators. Figure 2.1 compares the resulting mean-squared errors for ψ across several variants of the simulation designs considered in Section 2.2.1 (the exact parameters used are the same as

[13] In the case of binary treatments W_i, this corresponds to the classical result of Rosenbaum and Rubin (1983), who showed that the propensity score is a balancing score. With non-binary treatments, $\mathbb{E}\left[W_i \,\middle|\, X_i\right]$ is not in general a balancing score (Imbens, 2000); however, it is a balancing score for our specific model (2.39).

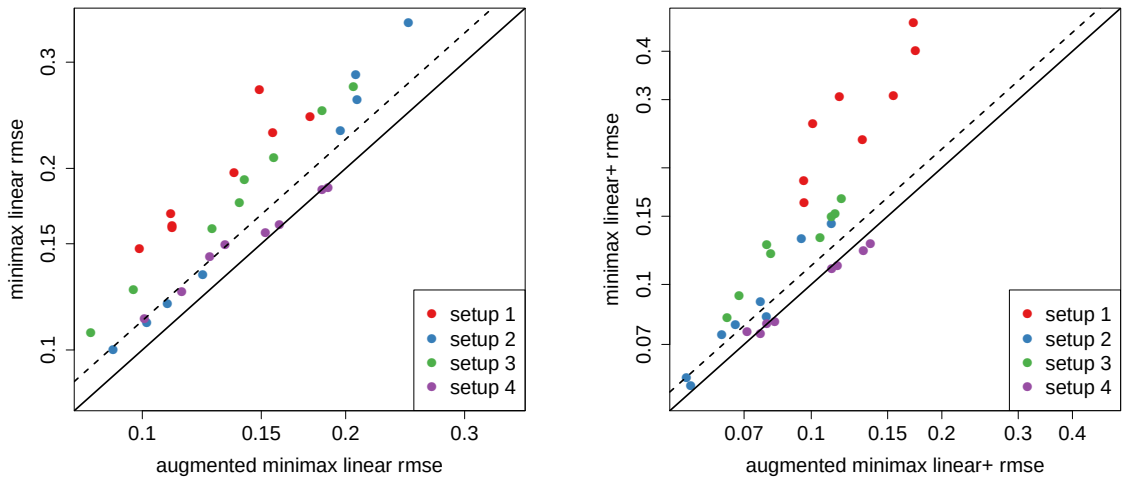

Figure 2.1: Comparing augmented minimax linear estimation with minimax linear estimation. The solid line $y = x$ indicates equivalent performance, while the dashed line $y = 1.25x$ corresponds to the best possible improvement over the minimax linear estimation in the setup of Donoho (1994), i.e., where $\mathcal{M}$ is known and convex.

those used in Table 2.1). The left panel shows results where the weights are minimax over $\mathcal{M}$, while the right panel has minimax weights over $\mathcal{M}_+$.

Overall, we see that the augmented minimax linear estimator is sometimes comparable to the minimax linear one, and sometimes substantially better. As discussed earlier, the improvements due to augmenting the minimax linear estimator can come from several different sources. First, even when $m \in \mathcal{F}$, the minimax linear estimator is only guaranteed to be within a factor of 1.25 of minimax in terms of mean-squared error (Donoho, 1994), meaning that there is room for small improvements even in this well specified setting. Second, perhaps more importantly, our method is less sensitive to the unknown signal-to-noise ratio because the bias-like term tends to decay faster than that of the minimax linear estimator; and finally, our approach only requires that $\|\hat{m} - m\|_{\mathcal{F}} \in \mathcal{O}_P(1)$ instead of $\|m\|_{\mathcal{F}} \in \mathcal{O}_P(1)$, meaning that we can accommodate signals in non-convex model classes, e.g. sparsity classes, as long as the residual error $\hat{m} - m$ is captured by the convex class $\mathcal{F}$. In Figure 2.1, we see that augmenting the minimax linear estimator often improves mean-squared error by substantially more than a factor 1.25, meaning that this second and third factors play a role in at least some examples.

Second, in Table 2.1, we compare augmented minimax linear estimation with doubly robust estimators, both using an estimated and an oracle Riesz representer. In terms of mean-squared error, our simple AML estimator already performs well relative to the main baseline (i.e., plug-in

doubly robust estimation), and the heuristically improved AML+ estimator does better yet. Perhaps more surprisingly, our methods sometimes also beat the doubly robust oracle, suggesting that the AML approach has good second order properties that manifest themselves in finite samples. In terms of coverage, some of our simulation designs are extremely difficult and all non-oracle estimators have substantial relative bias. However, settings 1 and 4, the asymptotics appear to be kicking in and our estimators get close to nominal coverage.

	method			double rob. plugin			augm. minimax			augm. minimax+			double rob. oracle		
	n	p	κ	rmse	bias	covg	rmse	bias	covg	rmse	bias	covg	rmse	bias	covg
setup 1	600	6	3	**0**.13	0.03	0.98	0.14	0.03	0.98	**0**.13	0.00	0.98	0.18	-0.01	0.96
	600	6	4	0.16	0.06	0.92	0.16	0.04	0.94	**0**.15	0.03	0.93	0.21	0.00	0.92
	600	12	3	0.22	0.09	0.78	0.18	-0.00	0.87	**0**.17	0.05	0.90	0.27	-0.04	0.90
	600	12	4	0.21	0.14	0.78	**0**.15	0.01	0.94	0.17	0.09	0.90	0.23	-0.03	0.93
	1200	6	3	**0**.10	0.03	0.94	0.11	0.06	0.92	**0**.10	0.02	0.96	0.12	0.00	0.98
	1200	6	4	0.11	0.03	0.94	0.11	0.05	0.92	**0**.10	0.02	0.96	0.13	0.00	0.94
	1200	12	3	0.11	0.02	0.90	**0**.10	0.01	0.95	**0**.10	0.02	0.94	0.14	0.00	0.94
	1200	12	4	0.15	0.06	0.86	**0**.11	0.00	0.92	0.12	0.04	0.90	0.16	-0.00	0.94
setup 2	600	6	1	0.15	0.12	0.52	0.11	0.09	0.74	**0**.08	0.02	0.94	0.09	0.00	0.92
	600	6	2	0.23	0.22	0.08	0.21	0.20	0.04	**0**.09	0.07	0.85	0.10	0.00	0.94
	600	12	1	0.16	0.14	0.44	0.12	0.11	0.62	**0**.08	0.03	0.93	0.08	0.00	0.98
	600	12	2	0.27	0.26	0.02	0.25	0.24	0.00	**0**.11	0.09	0.76	0.10	0.01	0.95
	1200	6	1	0.12	0.11	0.30	0.09	0.08	0.52	**0**.05	0.01	0.95	0.06	-0.00	0.96
	1200	6	2	0.20	0.20	0.00	0.20	0.19	0.00	**0**.06	0.04	0.90	0.06	-0.00	0.96
	1200	12	1	0.12	0.11	0.31	0.10	0.09	0.48	**0**.05	0.01	0.96	0.06	-0.00	0.98
	1200	12	2	0.22	0.22	0.00	0.21	0.20	0.00	**0**.07	0.04	0.86	0.07	0.00	0.94
setup 3	600	6	3	0.23	0.23	0.04	0.14	0.13	0.44	**0**.11	0.09	0.72	0.08	-0.00	0.96
	600	6	4	0.20	0.20	0.12	0.13	0.11	0.54	**0**.10	0.09	0.72	0.07	-0.00	0.96
	600	12	3	0.25	0.24	0.03	0.21	0.20	0.10	**0**.12	0.10	0.70	0.08	-0.01	0.95
	600	12	4	0.21	0.20	0.09	0.18	0.17	0.16	**0**.11	0.10	0.72	0.08	-0.01	0.94
	1200	6	3	0.20	0.19	0.01	0.10	0.09	0.55	**0**.07	0.05	0.78	0.05	-0.01	0.97
	1200	6	4	0.18	0.18	0.01	0.08	0.07	0.68	**0**.06	0.05	0.85	0.05	-0.01	0.96
	1200	12	3	0.23	0.22	0.00	0.16	0.15	0.02	**0**.08	0.07	0.76	0.05	-0.00	0.96
	1200	12	4	0.19	0.19	0.00	0.14	0.14	0.13	**0**.08	0.07	0.70	0.05	0.00	0.94
setup 4	600	6	4	0.22	0.16	0.84	0.16	-0.03	0.94	**0**.11	-0.02	1.00	0.16	0.03	0.94
	600	6	5	0.20	0.14	0.88	0.15	-0.05	0.93	**0**.11	-0.02	1.00	0.15	0.00	0.93
	600	12	4	0.23	0.15	0.86	0.18	-0.09	0.88	**0**.14	-0.04	0.96	0.17	-0.01	0.91
	600	12	5	0.24	0.17	0.82	0.19	-0.09	0.89	**0**.13	-0.05	0.97	0.17	-0.01	0.94
	1200	6	4	0.13	0.09	0.90	0.10	-0.03	0.94	**0**.07	-0.01	1.00	0.10	0.00	0.96
	1200	6	5	0.14	0.08	0.91	0.11	-0.05	0.94	**0**.08	-0.01	1.00	0.11	0.00	0.94
	1200	12	4	0.14	0.08	0.88	0.13	-0.07	0.88	**0**.08	-0.02	0.98	0.11	-0.00	0.94
	1200	12	5	0.14	0.09	0.87	0.13	-0.07	0.90	**0**.08	-0.02	1.00	0.11	-0.00	0.96

Table 2.1: Performance of 4 methods described in Section 2.2.2 on the simulation designs from Section 2.2.1. We report root-mean squared error, bias, and coverage of 95% confidence intervals averaged over 200 simulation replications.

2.3 Application: The Effect of Lottery Winnings on Earnings

To test the behavior of our method in practice, we revisit a study of Imbens, Rubin, and Sacerdote (2001) on the effect of lottery winnings on long-term earnings. It is of considerably policy interest to understand how people react to reliable sources of unearned income; such questions come up, for example, in discussing how universal basic income would affect employment. In an attempt to get some insight about this effect, Imbens, Rubin, and Sacerdote (2001) study a sample of people who won a major lottery whose prize is paid out in installments over 20 years. The authors then ask how \$1 in yearly lottery income affects the earnings of the winner.

To do so, the authors consider $n = 194$ people who all won the lottery, but got prizes of different sizes (\$1,000–\$100,000 per year).[14] They effectively use a causal model

$$\mathbb{E}[Y_i^{(w)} \mid X_i = x] = m(x) + \tau w. \tag{2.46}$$

for observations $Y_i = Y_i^{(W_i)}$ of the average yearly earnings in the 6 years following the win, W_i of the yearly lottery payoff, and X_i of a set of $p = 12$ pre-win covariates (year won, number of tickets bought, age at win, gender, education, whether employed at time of win, earnings in 6 years prior to win). The authors also consider several other specifications in their paper.

As discussed at length by Imbens, Rubin, and Sacerdote (2001), although the lottery winnings were presumably randomly assigned, we cannot assume exogeneity of the form $W_i \perp\!\!\!\perp Y_i^{(\mathcal{W})}$ because of survey non-response. The data was collected by mailing out surveys to lottery winners asking about their earnings, etc., so there may have been material selection effects in who responded to the survey. A response rate of 42% was observed, and older people with big winnings appear to have been relatively more likely to respond than young people with big winnings. For this reason, the authors only assume exogeneity conditionally on the covariates, i.e., $W_i \perp\!\!\!\perp Y_i^{(\mathcal{W})} \,\big|\, X_i$, which suffices to establish that $\mathbb{E}[Y_i \mid X_i = x, W_i = w] = m(x) + \tau w$ for $m(x), \tau$ in our causal model (2.46).

Here, we examine the robustness of the conclusions of Imbens, Rubin, and Sacerdote (2001) to potential effect heterogeneity. Instead of assuming a fixed τ parameter as in (2.46), we let $\tau(x)$ vary with x and seek to estimate $\psi = \mathbb{E}\left[\tau(X)\right]$; this corresponds exactly to an average partial effect in the conditionally linear model, as studied in Section 2.2. In our comparison, we consider 3 estimators that implicitly assume the partially linear specification (2.46) and estimate τ, and 6 that allow $\tau(x)$ to vary and estimate $\mathbb{E}\left[\tau(X)\right]$.

[14] The paper also considers some people who won very large prizes (more than \$100k per year) and some who won smaller prizes (not paid in installments); however, we restrict our analysis to the smaller sample of people who won prizes paid out in installments worth \$1k–\$100k per year.

estimand	estimator	estimate	std. err
partial effect	OLS without controls	-0.176	0.039
partial effect	OLS with controls	-0.106	0.032
partial effect	residual-on-residual OLS	-0.110	0.032
avg. partial effect	plugin Riesz weighting	-0.175	—
avg. partial effect	doubly robust plugin	-0.108	0.042
avg. partial effect	minimax linear weighting	-0.074	—
avg. partial effect	augm. minimax linear	-0.091	0.044
avg. partial effect	minimax linear+ weighting	-0.083	—
avg. partial effect	augm. minimax linear+	-0.097	0.045

Table 2.2: Various estimates, estimands and estimators for the effect of unearned income on earnings, using the dataset of Imbens, Rubin, and Sacerdote (2001). The first 3 methods are justified under the assumption of no heterogeneity in $\tau(x)$ (i.e., $\tau(x) = \tau$), in which case the methods estimate τ, while the latter 6 allow for heterogeneity and estimate $\mathbb{E}[\tau(X)]$. We do not report standard errors for the 3 weighting-based estimators, as these may be asymptotically biased and so valid confidence intervals would also need to explicitly account for possible bias.

Among methods that use (2.46), the first runs ordinary least squares for Y_i on W_i, ignoring potential confounding due to non-response. The second, which most closely resembles the method used by Imbens, Rubin, and Sacerdote (2001), controls for the X_i ordinary least squares, i.e., it regresses Y_i on (X_i, W_i) and considers the coefficient on W_i. The third uses the method of Robinson (1988) with cross-fitting as in Chernozhukov et al. (2017): it first estimates the marginal effect of X_i on W_i and Y_i via a non-parametric adjustment and then regresses residuals $Y_i - \widehat{\mathbb{E}}\left[Y_i \,\middle|\, X_i\right]$ on $W_i - \widehat{\mathbb{E}}\left[W_i \,\middle|\, X_i\right]$. In each case, we report robust standard errors obtained via the `R`-package `sandwich` (Zeileis, 2004).

The 6 methods that allow for treatment effect heterogeneity correspond to the 5 methods discussed in Section 2.2, along with a pure weighting estimator using the estimated Riesz representer, $\hat{\psi} = n^{-1} \sum_{i=1}^{n} \hat{g}(X_i) Y_i$, with the same choice of $\hat{g}(\cdot)$ as used in (2.43). For all non-parametric regression adjustments, we run penalized regression as in Section 2.2, on a basis obtained by taking order-3 Hermite interactions of the 10 continuous features, and then creating full interactions with the two binary variables (gender and employment), resulting in a total of 1140 basis elements. For AML+, we augment the balancing class with multi-scale propensity strata (at scales 0.05, 0.1, and 0.2).

Table 2.2 reports results using the 9 estimators described above, along with standard error estimates. We do not report standard errors for the 3 pure weighting methods, as these may not be asymptotically unbiased and so confidence intervals should also account for bias. The reported estimates are unitless; in other words, the majority of the estimators suggest that survey respondents on average respond to a $1 increase in unearned yearly income by reducing their yearly earnings by

roughly $0.10.

Substantively, it appears reassuring that most point estimates are consistent with each other, whether or not they allow for heterogeneity in $\tau(x)$. The only two divergent estimators are the one that doesn't control for confounding at all, and the one that uses pure plug-in weighting (which may simply be unstable here). From a methodological perspective, it is encouraging that our method (and here, also the plug-in doubly robust method) can rigorously account for potential heterogeneity in $\tau(x)$ without excessively inflating uncertainty.

Chapter Three

Minimax Linear Estimation

In this chapter, we focus on the estimation of a small class of linear functionals that arises frequently in causal inference. We consider an observational study in which we observe for each unit a covariate vector X_i, a categorical treatment status $W_i \in 0 \dots C$, and an outcome $Y_i = Y_i^{(W_i)} \in \mathbb{R}$, and assume that as a function of (X_i, W_i), we can calculate indicators $T_i = T(X_i, W_i)$ that mark units as members of a *target group* of scientific interest. Our goal will be to estimate the average, over this target group, of the potential outcome $Y_i^{(0)}$ that they would have been experienced had they received the treatment of interest $W_i = 0$.

The well-known problem of estimating a mean outcome when some outcomes are missing by a strongly ignorable mechanism (Rosenbaum and Rubin, 1983) is such a problem. In that case, we 'observe the outcome of interest' if we observe an outcome at all our target group is the entire population we sample from, i.e. we have $W_i = 0$ iff we actually observe the outcome Y_i and $T_i = 1$ for all i. However, the flexibility afforded us in this framework to define our target group can be very valuable. For example, if we are wondering whether to recommend a change to treatment $W_i = 0$ for those who are above a given age and currently taking another treatment $W_i = 1$, it is natural to estimate of the average outcome we'd expect to see for that specific group if that recommendation were followed, which we can do by defining T_i in terms of both X_i and W_i.

Assumptions that identify a quantity like this as a functional of the distribution of (X_i, W_i, Y_i) are discussed in Dahabreh et al. (2017). We restate them, adapted for our purposes, below.

Assumption 3.1 (Mean Exchangeability). Conditional on the covariates, the potential outcome mean does not depend on treatment assignment,

$$\mathbb{E}[Y_i^{(0)} \mid X_i, W_i = 0] = \mathbb{E}[Y_i^{(0)} \mid X_i].$$

To ensure the existence of the conditional mean above, we assume a positivity condition, $P\{W_i = 0 \mid X_i\} > 0$.

Assumption 3.2 (Mean Transportability). Conditional on covariates, the potential outcome mean

does not depend on membership in the target group,

$$\mathbb{E}[Y_i^{(0)} \mid X_i, T_i = 1] = \mathbb{E}[Y_i^{(0)} \mid X_i].$$

These assumptions identify our causal estimand $\mathbb{E}[Y_i^{(0)} \mid T_i = 1]$ as the linear functional

$$\psi'(m) = \mathbb{E}\left[m(X_i, 0) \mid T_i = 1\right] \quad \text{at} \quad m(x, w) = \mathbb{E}\left[Y_i \mid X_i = x, W_i = w\right]. \tag{3.1}$$

While the problem of estimating $\psi'(m)$ can be solved by the methods described in the previous chapter, in this chapter we will consider a simple linear estimator of the form $\hat{\psi}' = n_T^{-1} \sum_{i=1}^n 1_{\{W_i=0\}} \hat{\gamma}_i Y_i$ for $n_T = \sum_{i=1}^n T_i$. In many applications, estimators of this form are desirable for their ease of interpretation: we observe our potential outcome of interest $Y_i^{(0)}$ on the subsample receiving the treatment $W_i = 0$, so we weight that subsample so that it looks like the target subsample and calculate the weighted average outcome.

It will be helpful to reframe this problem essentially as a special case of the problem we considered in the previous chapter, the estimation of a functional of the form $\psi(m) = \mathbb{E}\left[h(X_i, W_i, m)\right]$. $\psi'(\cdot)$ does not have this form itself, but it may be expressed as a ratio $\psi'(m) = \mathbb{E}\left[T_i m(X_i, 0)\right] / \mathbb{E}\left[T_i\right]$ in which the numerator has this form,

$$\psi(m) = \mathbb{E}\left[h(X_i, W_i, m)\right] \quad \text{with} \quad h(X_i, W_i, m) = T(X_i, W_i) m(X_i, 0). \tag{3.2}$$

We will focus on an estimator for the numerator of the form $\hat{\psi} = n^{-1} \sum_{i=1}^n 1_{\{W_i=0\}} \hat{\gamma}_i Y_i$, as dividing by $n^{-1} \sum_{i=1} T_i$ will yield an estimator $\hat{\psi}'$ of the desired form.

The previous chapter's logic suggests a natural oracle estimator of this form defined in terms of the Riesz representer for this functional,

$$\gamma_\psi(x, w) = 1_{\{w=0\}} g_\psi(x) \quad \text{for } g_\psi(x) = \frac{P\{T_i = 1 \mid X_i = x\}}{P\{W_i = 0 \mid X_i = x\}}, \tag{3.3}$$

the unique square integrable function satisfying $\mathbb{E}\left[\gamma_\psi(X_i, W_i) f(X_i, W_i)\right] = \psi(f)$ for all square integrable functions $f(x, w)$. This property, which holds under the 'overlap' condition $g_\psi > 0$ a.s., ensures that this oracle estimator $\psi^\star = n^{-1} \sum_{i=1}^n \gamma_\psi(X_i, W_i) Y_i$ will be unbiased, as $\mathbb{E}\left[\psi^\star\right] = \mathbb{E}\left[\gamma_\psi(X_i, W_i) m(X_i, W_i)\right]$, and as it is a sum of independent terms, it will converge to its mean $\psi(m)$ at $n^{-1/2}$ rate by the central limit theorem. This motivates the use of estimators of the form $\hat{\psi}_{IPW} = n^{-1} \sum_{i=1}^n \hat{\gamma}_\psi(X_i, W_i) Y_i$ where $\hat{\gamma}_\psi(\cdot)$ is an estimate of the Riesz representer $\gamma_\psi(\cdot)$. It is conventional to call weights like $\gamma_\psi(X_i, W_i)$ 'inverse probability weights', as they essentially invert the probabilistic mechanism that assigns units to our treatment and target groups to ensure unbiasedness, and we call estimators of the form $\hat{\psi}_{IPW}$ Inverse Probability Weighting estimators.

But as in the previous chapter, we will take a different approach here. Our estimator $\hat{\psi}_{ML}$ will be a minimax linear estimator of a sample-average version of our estimand $\tilde{\psi}(m) = n^{-1}\sum_{i=1}^{n} T_i m(X_i, 0)$ conditional on the study design $(X_i, W_i)_{i \leq n}$. Specifically, we choose the weights that result in the best estimate of $\tilde{\psi}(m)$ of the form $n^{-1}\sum_{i=1}^{n}\gamma_i Y_i$ in the worst case over regression functions $m(\cdot, 0)$ in an absolutely convex class $\mathcal{F}$ and over conditional variance functions $\text{Var}\left[Y_i \mid X_i = x, W_i = w\right]$ bounded by a constant σ^2. This defines our weights $\hat{\gamma} \in \mathbb{R}^n$ as the solution to an instance of the convex optimization problem (2.15) discussed in the previous chapter, which reduces in this case to the choice of weights satisfying $\hat{\gamma}_i = 0$ for $W_i \neq 0$ and minimizing

$$I_{\mathcal{F}}^2(\gamma) + \frac{\sigma^2}{n^2}\|\gamma\|^2 \quad \text{where} \quad I_{\mathcal{F}}(\gamma) = \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}[1_{\{W_i=0\}}\gamma_i - T_i]f(X_i). \tag{3.4}$$

Much has already been said in favor of such estimators. In regression problems with fixed design $Z_1 \ldots Z_n$, if we observe $Y_i = m(Z_i) + \varepsilon_i$ with independent Gaussian noise $\varepsilon_i \sim N(0, \sigma_i)^2$, Donoho (1994) and related papers (Armstrong and Kolesár, 2018; Cai and Low, 2003; Donoho and Liu, 1991; Ibragimov and Khas'minskii, 1985; Johnstone, 2015; Juditsky and Nemirovski, 2009) have established a number of desirable properties, among them that if the regression function $m(\cdot)$ is in a convex set $\mathcal{F}$, a minimax linear estimator of a linear functional $\psi(m)$ will come within a factor 1.25 of the minimax risk over all estimators. And these estimators have been found to perform well in practice in a variety of applications including the missing outcomes problem discussed above (Armstrong and Kolesár, 2018; Imbens and Wager, 2017; Kallus, 2016; Wang and Zubizarreta, 2017; Zubizarreta, 2015).

However, there is a mismatch between this fixed-design approach and the typical way of thinking about these problems, in which the random variation of treatment status W_i plays an essential role. In particular, while our oracle estimator $\psi^\star$ is widely considered to be the gold standard among linear estimators, and inverse probability weighting estimators are very popular, it is difficult to make sense of them in the fixed-design terms. For that reason, we will focus here on characterizing the random-design behavior of our estimator $\hat{\psi}_{ML}$. In essence, what we will show here is that $\hat{\psi}_{ML}$ will be asymptotically efficient under assumptions similar to those under which we showed the efficiency of the more complicated estimator $\hat{\psi}_{AML}$ in the previous chapter. This is in part a consequence of the convergence of the weights $\hat{\gamma}_i$ to the evaluated Riesz representer $\gamma_\psi(X_i, W_i)$, a result established in greater generality in the previous chapter and also familiar from Chan et al. (2015) and Wang and Zubizarreta (2017). This happens because (3.4) requires that our weights satisfy a set of estimating equations $I_{\mathcal{F}}(\hat{\gamma}) \approx 0$ derived from the condition $\psi(f) = \mathbb{E}\,\gamma_\psi(X_i, W_i)f(X_i, W_i)$ that defines γ_ψ. While this good random-design asymptotic behavior does seem to line up well with the fixed-design

near-optimality results discussed in the previous paragraph, these assumptions are substantially weaker than those that have been used to show efficiency in most previous work. As asymptotics do not tell the whole story, we include finite sample bounds on the error of our estimator and a small simulation study as well as an examination of the estimator's performance on the well-known LaLonde study.

3.1 Understanding the Estimator

To better understand the behavior of our estimator, we decompose its error into a bias-like term and a noise-like term. We will consider estimation of the sample-average version of our estimand, $\tilde{\psi}(m) := n^{-1}\sum_{i=1}^n T_i m(X_i, 0)$, as the behavior of the difference $\tilde{\psi}(m) - \psi(m)$ is out of our hands. We write

$$\begin{aligned}\hat{\psi}_{ML} - \tilde{\psi}(m) &= \frac{1}{n}\sum_{i=1}^n 1_{\{W_i=0\}}\hat{\gamma}_i Y_i - T_i m(X_i, 0) \\ &= \frac{1}{n}\sum_{i=1}^n \underbrace{\left[1_{\{W_i=0\}}\hat{\gamma}_i - T_i\right] m(X_i, 0)}_{\text{bias}} + \underbrace{1_{\{W_i=0\}}\hat{\gamma}_i \varepsilon_i}_{\text{noise}},\ \varepsilon_i = Y_i - m(X_i, W_i).\end{aligned} \tag{3.5}$$

It is clear from this expression that what we are minimizing in (3.4) to define our weights is, in fact, the mean squared error conditional on $(X_i, W_i)_{i\leq n}$. Our bias-like term is $\mathbb{E}\left[\hat{\psi}_{ML} \mid X, W\right] - \tilde{\psi}(m)$ and our noise-like term is $\hat{\psi}_{ML} - \mathbb{E}\left[\hat{\psi}_{ML} \mid X, W\right]$.

Supposing that $m(\cdot, 0)$ is really in the class $\mathcal{F}$ that our estimator is minimax over, our bias term is bounded by $I_{\mathcal{F}}$. This allows us to bound our bias term using a simple argument. We use the property that our maximal risk (3.4) at its minimizer $\hat{\gamma}$ is smaller than it is at the oracle weights $\gamma_i^\star = \gamma_\psi(X_i, W_i)$. Rearranging this condition yields the bound

$$I_{\mathcal{F}}^2(\hat{\gamma}) \leq I_{\mathcal{F}}^2(\gamma^\star) + \frac{\sigma^2}{n^2}(\|\gamma^\star\|^2 - \|\hat{\gamma}\|^2). \tag{3.6}$$

As a result, we have $I_{\mathcal{F}}(\hat{\gamma}) \leq I_{\mathcal{F}}(\gamma^\star) + \sigma\|\gamma_\psi\|_\infty n^{-1/2}$. Furthermore, $I_{\mathcal{F}}(\gamma^\star)$ is the supremum of the empirical process $n^{-1}\sum_{i=1}^n \delta_{X_i, W_i}$ indexed by the class of functions $\mathcal{H} = \{[\gamma_\psi(x, w) - T(x, w)]f(x) : f \in \mathcal{F}\}$, and for the same reason that weighting by the Riesz representer γ_ψ results in unbiased estimation, each function in this class has mean zero. Using well known tools from Empirical Process Theory, this supremum can be shown to be concentrate at $n^{-1/2}$-rate on a quantity comparable to the Rademacher complexity $R_n(\mathcal{F})$ of the set of outcome models $\mathcal{F}$.[1] Consequently, this argument shows that our estimator will be consistent at $n^{-1/2}$ rate when the class $\mathcal{F}$ is small enough that $R_n(\mathcal{F}) = O(n^{-1/2})$.

[1]The relevant tools are the symmetrization technique, the Ledoux-Talagrand Contraction Lemma, and the Bounded Difference Inequality (see e.g. Giné and Nickl, 2015, Theorems 3.1.21, 3.2.1, and 3.3.14 respectively).

However, this is essentially the limit of this argument's power in this context. While we used a variant of this argument to show that the bias term of the regression-adjusted weighting estimator $\hat{\psi}_{AML}$ was $o_p(n^{-1/2})$, it cannot be used for the same purpose without regression adjustment. The 'bias term' of the linear estimator with the oracle weights $\gamma^\star$, $n^{-1}\sum_{i=1}^n[\gamma_\psi(X_i,W_i)-T_i]m(X_i,0)$, has mean zero but standard deviation on the order of $n^{-1/2}$. As a result, arguments relying solely on the characterization that our estimator performs as well as this oracle estimator cannot be used to show that our estimator's bias term is negligible.

Using more refined arguments, many methods like ours have been shown to control the bias term at $o_p(n^{-1/2})$ rate and as a consequence achieve semiparametric efficiency. These methods include Empirical Balancing Calibration Weighting (Chan et al., 2015), the Covariate Balancing Propensity Score (Fan et al., 2016), and the Minimal Approximately Balancing Weights (Wang and Zubizarreta, 2017). Each of them optimizes for some desirable property of the weights subject to bounds on the maximal conditional bias $I_{\mathcal{F}_n}$ over some finite-dimensional class $\mathcal{F}_n$. All of these arguments rely on the phenomenon that there are weights that achieve better control on $I_{\mathcal{F}_n}$ than the inverse probability weights γ_ψ do for sufficiently small classes $\mathcal{F}_n$. Clearly this is the case for classes $\mathcal{F}_n$ of dimension no larger than n, as in that case the condition $I_{\mathcal{F}_n}(\gamma)=0$ is a solvable set of linear equations. This phenomenon is quite robust. Under sufficient regularity assumptions, even methods that estimate inverse propensity weights $\hat{\gamma}_\psi$ by maximum likelihood within some appropriate sequence of finite-dimensional model classes $\mathcal{G}_n$ have been shown to achieve better control on the bias term and therefore semiparametric efficiency (Hirano et al., 2003).

But these approaches do not line up well with the minimax framework we've discussed. After all, the regression function $m(\cdot,0)$ is not smoother because we have a small sample size, although from the perspective of its behavior on the sample $X_1\dots X_n$, it may admit approximation by a smoother function. What we show here is that even when we solve the optimization problem (3.4) for a nontrivially large infinite-dimensional class of possible regression functions $\mathcal{F}$, we see this phenomenon. That is, the use of these finite-dimensional sieves $\mathcal{F}_n$ is not necessary to control the bias better than the inverse propensity weights γ_ψ. Instead, we will take the class $\mathcal{F}$ to be the unit ball of a Reproducing Kernel Hilbert Space (RKHS), e.g. the Sobolev space H^k of square-integrable functions with square integrable weak partial derivatives of order up to $k > d/2$. Kallus (2016) worked with this class of estimators as well, using the aforementioned argument based on (3.6) to show consistency at $n^{-1/2}$ rate. Our primary contribution is a sharper characterization of the estimator's behavior.

At the heart of our argument will be a characterization of our estimator $\hat{\psi}_{ML}$ as the average, over

our target subsample, of a kernel-ridge-regression estimate $\hat{m}(\cdot)$ of the regression function $m(\cdot, 0)$ based on the subsample $\{i : W_i = 0\}$ of units receiving the treatment of interest. Specifically,

Lemma 3.1. *If $\mathcal{F}$ is the unit ball of an RKHS with norm $\|\cdot\|$, then*

$$\frac{1}{n}\sum_{i=1}^{n} 1_{\{W_i=0\}}\hat{\gamma}_i Y_i = \frac{1}{n}\sum_{i=1}^{n} T_i \hat{m}(X_i) \quad \textit{where} \tag{3.7}$$

$$\hat{\gamma} = \operatorname*{argmin}_{\gamma} I^2_{\mathcal{F}}(\gamma) + \frac{\sigma^2}{n^2}\|\gamma\|^2, \;\; I_{\mathcal{F}}(\gamma) = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} [1_{\{W_i=0\}}\gamma_i - T_i] f(X_i); \tag{3.8}$$

$$\hat{m} = \operatorname*{argmin}_{m} \frac{1}{n_Z} \sum_{i:W_i=0} (Y_i - m(X_i))^2 + \frac{\sigma^2}{n_Z}\|m\|^2 \quad \textit{where} \;\; n_Z = |\{i : W_i = 0\}|. \tag{3.9}$$

We use this result, proven in Appendix B.4, to characterize the bias term of our estimator as the bias of this ridge regression estimator, conditional on the design $(X_i, W_i)_{i\leq n}$, averaged over the target subsample. As the weight σ^2/n_Z of the penalty term in our ridge regression is small, our estimator will be fairly unbiased, but it is sufficient to allow generalization to our target sample so long as our oracle weights γ_ψ are bounded. Using this argument to characterize our estimator's bias term and a variant of the previous chapter's Theorem 2.2 to characterize our noise term, we establish finite sample bounds on the error of our estimator $\hat{\psi}_{ML}$.

As a consequence, we obtain a simple characterization of the set of regression functions $m(\cdot, 0)$ for which our estimator will be semiparametrically efficient: it suffices for $m(\cdot, 0)$ to be in a certain sense smoother than the least smooth functions in our RKHS. Smoothness in excess of this level improves the higher order terms in our bound. We will state this result formally after introducing the necessary definitions in the following section.

3.1.1 A Review of Reproducing Kernel Hilbert Spaces[2]

Let $\mathcal{X}$ be a compact metric space. A Reproducing Kernel Hilbert Space $\mathcal{H}_K$ of functions on $\mathcal{X}$ is a complete normed vector space with its norm $\|\cdot\|$ induced by an inner product $\langle\cdot,\cdot\rangle$ in the sense that $\|f\|^2 = \langle f, f\rangle$ and on which point evaluation is continuous in the sense that for all $x \in \mathcal{X}$ there is a constant C_x such that $f(x) \leq C_x\|f\|$. By the Riesz representation theorem (see e.g. Peypouquet, 2015, Theorem 1.4.1), this implies that each $x \in \mathcal{X}$ corresponds to a unique element K_x in the RKHS such that $f(x) = \langle K_x, f\rangle$. We call the function $K(x, y) = \langle K_x, K_y\rangle$ the *kernel* associated $\mathcal{H}_K$. If the kernel is continuous, it is bounded as a consequence of the the compactness of $\mathcal{X}\times\mathcal{X}$, and furthermore $\|\cdot\|_\infty \leq M_K\|\cdot\|$ for the finite constant $M_K = \sqrt{\sup_x K(x, x)}$, as by Cauchy-Schwartz $\|f\|_\infty = \sup_x\langle K_x, f\rangle \leq \sqrt{\sup_x\langle K_x, K_x\rangle\langle f, f\rangle}$.

[2]This review is based largely on Chapters 2 and 4 of Cucker and Zhou (2007), although what is taken as definitional and what is considered a derived property differ somewhat between this account and that one. Results from those chapters will be stated without individual citation.

Given any finite measure ν with support equal to $\mathcal{X}$, we can completely characterize an RKHS $\mathcal{H}_K$ with a continuous kernel K in terms of the spectral decomposition of a compact positive integral operator

$$(L_{K,\nu}f)(x) = \int K(x,x')f(x')d\nu(x') \tag{3.10}$$

mapping the space of square integrable functions $L_2(\nu)$ to itself. Its eigenfunctions $(\phi_j)_{j\in\mathbb{N}}$ form an orthonormal basis for $L_2(\nu)$ and its scaled eigenfunctions $\sqrt{\lambda_j}\phi_j$, where λ_j is the eigenvalue corresponding to ϕ_j, form an orthonormal basis for $\mathcal{H}_K$. One useful consequence is that the square root of our integral operator, the operator $L_{K,\nu}^{1/2}$ mapping $\sum_j f_j\phi_j$ to $\sum_j f_j\sqrt{\lambda_j}\phi_j$, maps an orthonormal basis of $L_2(\nu)$ to an orthonormal basis of our $\mathcal{H}_K$, so (i) $\mathcal{H}_K$ is the image of the square integrable functions $L_2(\nu)$ under $L_{K,\nu}^{1/2}$ and (ii) $\|f\|_{L_2(\nu)} = \|L_{K,\nu}^{1/2}f\|$.

Generalizing (i), we can think of the space of square integrable functions $L_2(\nu)$ and our RKHS $\mathcal{H}_K$ as elements of a continuum of spaces, the images $L_{K,\nu}^{\kappa}(L_2(\nu))$ of $L_2(\nu)$ under powers of $L_{K,\nu}$. Corresponding to these spaces, we define the family of norms $\|f\|_{L_{K,\nu}^{\kappa}} = \|L_{K,\nu}^{-\kappa}f\|_{L_2(\nu)}$, with $\|\cdot\| = \|\cdot\|_{L_{K,\nu}^{1/2}}$. This exponent κ will be a useful quantitative notion of smoothness.

One familiar scale of spaces like this is the scale of Sobolev spaces H^s of s-times weakly differentiable periodic functions on the unit cube endowed with Lebesgue measure μ (see e.g. Kühn et al., 2014). These spaces have the characterization $H^s = \{\sum_{k\in\mathbb{Z}^d} f_k(1+\|k\|_2^2)^{-s/2}e^{2\pi ik} : \sum_{k\in\mathbb{Z}^d} f_k^2 \leq 1\}$, with the fourier basis functions as eigenfunctions irrespective of s. It is clear from this that if K_s is the kernel of H^s, then for all s', $H^{s'}$ is the image of $L_2(\mu)$ under $L_{K_s,\mu}^{(s'/s)/2}$ or equivalently the image of H^s under $L_{K_s,\mu}^{(s'/s-1)/2}$.

While our space $\mathcal{H}_K$ itself is not defined with reference to any particular measure ν, many of the the objects discussed above are. One useful relation between operators $L_{K,\nu}$ and $L_{k,\nu'}$ defined in terms of different measures is that for all ϕ,

$$\langle\phi, L_{K,\nu'}\phi\rangle_{L_2(\nu')} = \int K(x,y)\phi(x)\phi(y)\frac{d\nu'}{d\nu}(x)\frac{d\nu'}{d\nu}(y)d\nu(x)d\nu(y) \leq \left\|\frac{d\nu'}{d\nu}\right\|_\infty^2 \langle\phi, L_{K,\nu}\phi\rangle.$$

As mentioned in Bach (2017), this identity and the extremal characterization of the eigenvalues $\lambda_{j,\nu}$ and $\lambda_{j,\nu'}$ offered by the Courant-Fischer minimax theorem (see e.g. Horn et al., 1990) imply that $\lambda_{j,\nu'} \leq \|d\nu'/d\nu\|_\infty^2\lambda_{j,\nu}$ for all j. As our finite-sample bounds will depend on the eigenvalues of integral operators defined in terms of the unknown distribution of our data, this phenomenon makes our bounds much less opaque than they otherwise would be. In particular, under weak assumptions the relevant eigenvalues decay at the same rate as those of the operator $L_{K,\mu}$ for Lebesgue measure μ, which can often be calculated straightforwardly. This approach will be used in the proof of Lemma 3.5.

3.1.2 Main Results

Having reviewed these properties, we are prepared to state and prove our results. We'll start with the asymptotic results.

Setting We observe $(X_i, W_i, Y_i)_{i \le n}$ iid from a distribution P with $m(x,w) = \mathbb{E}[Y_i \mid X_i = x, W_i = w]$ and $v(x,w) = \mathrm{Var}[Y_i \mid X_i = x, W_i = w]$. For some binary function $T(x,w)$, we define $T_i = T(W_i, X_i)$, and we will assume that T is chosen so that the target and treatment groups overlap in the sense that $g_\psi(x) < \infty$ $P-a.s.$ for g_ψ defined in (3.3). We consider the estimands $\psi'(m)$ and $\psi(m)$ defined in (3.1) and (3.2) in terms of these observations as well as the sample variant of $\psi(m)$, $\tilde{\psi}(m) = n^{-1}\sum_{i=1}^n T(X_i, W_i) m(X_i, 0)$. We write P_Z for the distribution of X_i conditional on $W_i = 0$ and assume that its support is a compact metric space $\mathcal{X}$, working with an RKHS $\mathcal{H}_K$ of functions on $\mathcal{X}$ with kernel K and norm $\|\cdot\|$.

Smoothness assumptions For spaces of functions on subsets of $\mathbb{R}^d$, we measure smoothness by one of two standards, (i) the maximal Sobolev norm $\sup_{f:\|f\|\le 1}\|f\|_{H^s}$ of an element of the unit ball of $\mathcal{H}_K$ or (ii) the Hölder norm $\|K\|_{C^{2s}}$ of the kernel K. We define the aforementioned norms in Appendix B.2.

Assumption 3.3. The unit ball of our RKHS is contained in a ball in the Sobolev space H^s, i.e. $\sup_{\|f\|_{\mathcal{H}_K}\le 1}\|f\|_{H^s} < \infty$, for $s > d$.

Assumption 3.4. The kernel K of our space satisfies the Hölder-type smoothness condition $\|K\|_{C^{2s}} < \infty$ for noninteger $s > d/2$.

This latter assumption implies containment of the unit ball of $\mathcal{H}_K$ in a ball in a Hölder space, i.e. $\sup_{\|f\|_{\mathcal{H}_K}\le 1}\|f\|_{C^s} < \infty$ (Cucker and Zhou, 2007, Theorem 5.5.).

Theorem 3.2. *In the setting described above, let $\mathcal{H}_K$ be dense in $L_2(P_Z)$ and satisfy Assumption 3.3 or 3.4, and $\|m(\cdot,0)\|_{L^\kappa_{K,P_Z}} < \infty$ for $\kappa > 1/2$. Then for any constant $\sigma > 0$, the estimator $\hat{\psi}_{ML} = n^{-1}\sum_{i=1}^n 1_{\{W_i=0\}}\hat{\gamma}_i Y_i$ with weights $\hat{\gamma}$ defined in (3.8) has the asymptotic characterization*

$$\begin{aligned} \hat{\psi}_{ML} - \psi(m) &= \frac{1}{n}\sum_{i=1}^n \iota(X_i, W_i, Y_i) + o_p(n^{-1/2}) \quad \textit{where} \\ \iota(x,w,y) &= T(x,w)m(x,0) - \psi(m) + \gamma_\psi(x,w)(y - m(x,0)). \end{aligned} \tag{3.11}$$

As a consequence of this characterization, $\sqrt{n}(\hat{\psi}_{ML} - \psi(m))$ is asymptotically normal with variance $V = \mathbb{E}\,\iota(X_i, W_i, Z_i)^2$. Given a consistent estimate $\widehat{V}$ of this variance, $\hat{\psi}_{ML} \pm z_{\alpha/2}\widehat{V}^{1/2}/n^{1/2}$ is an asymptotically valid confidence interval of level $1-\alpha$.

An analogous result applies for our original estimand $\psi'(m)$, justifying analogous normality-based inference for this quantity. This follows from the theorem above using the convergence of $n_T/n \to p_T = P\{T_i = 1\}$.

Corollary 3.3. *Under the assumptions of Theorem 3.2,* $\hat{\psi}'_{ML} = (n_T/n)^{-1}\hat{\psi}_{ML}$ *has the asymptotic characterization*

$$\begin{aligned} &\hat{\psi}'_{ML} - \psi'(m) = \frac{1}{n}\sum_{i=1}^{n} \iota'(X_i, W_i, Y_i) + o_p(n^{-1/2}) \quad \textit{where} \\ &\iota'(x, w, y) = p_T^{-1}T(x, w)m(x, 0) - \psi'(m) + p_T^{-1}\gamma_\psi(x, w)(y - m(x, 0)). \end{aligned} \tag{3.12}$$

Proposition 2.3 establishes that these estimators are semiparametrically efficient[3], meaning that no other estimator has better first-order asymptotic behavior uniformly over a neighborhood of the true data generating process (see e.g. van der Vaart, 2000, Theorem 25.21). But it is not clear that in any finite sample the $o_p(n^{-1/2})$ 'remainder term' in these characterizations will not invalidate inference based on these first order asymptotic characterizations (3.11) and (3.12). To inform about the magnitude of this remainder, we will now state a nonasymptotic characterization of our estimator's error. As this result is fairly complex, we will discuss the rate at which this remainder converges to zero in a remark below.

Theorem 3.4. *In the setting described above, consider the estimator* $\hat{\psi}_{ML} = n^{-1}\sum_{i=1}^{n} 1_{\{W_i=0\}}\hat{\gamma}_i Y_i$ *with weights* $\hat{\gamma}$ *defined in (3.8). Let the decreasing sequences of eigenvalues* $\lambda_{j,T}$ *and* $\lambda_{j,Z}$ *of* L_{K,P_T} *and* L_{K,P_Z} *respectively satisfy the bounds* $\lambda_{j,T} \leq C_{\lambda,T}j^{-\alpha}, \lambda_{j,Z} \leq C_{\lambda,Z}j^{-\alpha}$ *and the eigenfunctions* ϕ_j *of* L_{K,P_Z} *satisfy the bound* $\|\phi_j\|_\infty \leq C_\phi \lambda_j^{-\beta/2}$ *with* $\alpha > 1$ *and* $\alpha(1-\beta) > 1$. *Define* $\lambda = \sigma^2/n$, $p_Z = P\{W_i = 0\}$, $p_T = P\{T_i = 1\}$.

For any $\eta > 0, \delta \in (0, 1)$ *and any* $\kappa_m > 1/2, \kappa_\gamma > 0$ *such that* $\|m(\cdot, 0)\|_{L^{\kappa_m}_{K,P_Z}}, \|g_\psi\|_{L^{\kappa_\gamma}_{K,P_Z}} < \infty$,

[3]Specifically, Proposition 2.3 establishes efficiency and therefore regularity of the asymptotically linear estimator $\hat{\psi}$; it is clear that we get another regular asymptotically linear estimator when we divide $\hat{\psi}$ by n_T/n to yield $\hat{\psi}'$; and all regular asymptotically linear estimators are efficient in problems like this one, in which the space of models we allow is nonparametric (see e.g. Newey, 1994, Theorem 2.1).

with probability $1-3\delta$,

$$\left|\mathbb{E}\left[\hat{\psi}_{ML} \mid X, W\right] - \tilde{\psi}(m)\right| \leq sp_Z \|g_\psi\|_{L_2(P_Z)} + n^{-1/2} s (r/s)^{1/\alpha}(1+\eta)C$$
$$+ n^{-1/2} s \left(2p_Z \|g_\psi\|_\infty \log(2\delta^{-1})\right) + n^{-1} r \left(2M_K \left(1/3 + 1/\eta\right) \log(2\delta^{-1})\right); \tag{3.13}$$

$$s = \begin{cases} \zeta^{-1}\lambda^{\kappa_m} \|m(\cdot,0)\|_{L^{\kappa_m}_{K,P_Z}} & \kappa_m \in [1/2, 1); \\ \zeta^{-1}\lambda\lambda_{1,Z}^{\kappa_m - 1} \|m(\cdot,0)\|_{L^{\kappa_m}_{K,P_Z}} & \kappa_m \geq 1; \end{cases}$$

$$r = \begin{cases} \zeta^{-1}\lambda^{\kappa_m - 1/2} \|m(\cdot,0)\|_{L^{\kappa_m}_{K,P_Z}} & \kappa_m \in [1/2, 3/2); \\ \zeta^{-1}\lambda\lambda_{1,Z}^{\kappa_m - 3/2} \|m(\cdot,0)\|_{L^{\kappa_m}_{K,P_Z}} & \kappa_m \geq 3/2; \end{cases}$$

$$\zeta = \max\left\{0,\ 1 - \frac{8C_\zeta \lambda^{-(1/\alpha+\beta)} \log(4\delta^{-1}n^2)}{np_Z - \sqrt{2np_Z \log(\delta^{-1})}} - \sqrt{\frac{16C_\zeta \lambda^{-(1/\alpha+\beta)} \log(4\delta^{-1}n^2)}{np_Z - \sqrt{2np_Z \log(\delta^{-1})}}}\right\};$$

and with probability $1 - \exp\{-c_1(\eta_Q) n r^2 / M^2_{\mathcal{F}^\star}\} - 4\delta$,

$$\left|\hat{\psi}_{ML} - \mathbb{E}[\hat{\psi}_{ML} \mid X, W] - n^{-1} \sum_{i=1}^n \gamma_\psi(X_i, W_i)(Y_i - m(X_i, 0))\right| \leq n^{-1/2} (a \wedge b)^{1/2} \|v\|_\infty \delta^{-1/2}; \tag{3.14}$$

$$a = \alpha\left(C_{u,1} n^{-1/2} + C_{u,2} n^{-1}\right) + \bar{R} + \lambda;$$

$$b = 2\alpha^2 r^2 \vee 2\frac{\bar{R} + \lambda}{\eta_Q - 2\alpha^{-1}\eta_C} \vee \frac{44 M^2_{\mathcal{F}^\star} \alpha^2 \log(\delta^{-1})}{n};$$

$$\alpha = 1 \vee \left[2\eta_C \lambda^{-1} r^2 + \lambda^{-1/2} \bar{R}^{1/2}\right];$$

$$r = 7C_Q n^{-\frac{1}{2(1+1/\alpha)}} \vee \lambda^{1/2} \eta_Q^{-1/2};$$

$$\bar{R} = \begin{cases} C_{1,R} \lambda^{\frac{4\kappa_\gamma}{1+2\kappa_\gamma}} + C_{2,R} n^{-1/2} \lambda^{\frac{2\kappa_\gamma}{1+2\kappa_\gamma}} & \kappa_\gamma \in (0, 1/2); \\ \lambda \|g_\psi\|^2_{\mathcal{H}_K} & \kappa_\gamma \geq 1/2. \end{cases}$$

in terms of 'constants', which may be functions of δ *but not of* n *or* λ*, defined in Appendix B.1.*

Via the triangle inequality, the sum of these two bounds is a bound on the magnitude of the remainder, i.e. the deviation of our estimator from our idealized asymptotic characterization $\psi(m) + n^{-1/2}\sum_{i=1}^n \iota(X_i, W_i, Y_i)$ in (3.11), as

$$\hat{\psi}_{ML} - \tilde{\psi}(m) - n^{-1}\sum_{i=1}^n \gamma_\psi(X_i, W_i)(Y_i - m(X_i, 0)) = \hat{\psi}_{ML} - \psi(m) - n^{-1}\sum_{i=1}^n \iota(X_i, W_i, Y_i).$$

Thus, the claim (3.11) made by Theorem 3.2 holds if the two bounds (3.13) and (3.14) are $o(n^{-1/2})$ for all $\delta > 0$. To prove Theorem 3.2, it suffices to establish bounds on the eigenvalues of L_{K,P_Z} and L_{K,P_T} and the supremum norm of the eigenfunctions of the former.

The behavior of these eigenvalues and eigenfunctions can be characterized in terms of the measures of the smoothness of the space $\mathcal{H}_K$ that we discussed above. We prove the following lemma in Appendix B.2 and using it prove Theorem 3.2 in Appendix B.3.

Lemma 3.5. *Let $\mathcal{H}_K$ be an RKHS of functions on a compact set $\mathcal{X} \subseteq \mathbb{R}^d$ and ν be a measure on $\mathcal{X}$ that is strongly equivalent to Lebesgue measure μ in the sense that $\eta \leq d\nu/d\mu \leq \eta^{-1}$ for some $\eta > 0$. Then the decreasing sequence of eigenvalues λ_j and the corresponding eigenfunctions ϕ_j of $L_{K,\nu}$ satisfy $\lambda_j = O(j^{-\alpha})$ and $\|\phi_j\|_\infty = O(\lambda_j^{-\beta/2})$ with (i) $\alpha = 2s/d$ and $\beta = d/(2s)$ if $\sup_{\|f\|_{\mathcal{H}_K} \leq 1} \|f\|_{H^s} < \infty$ and (ii) $\alpha = (2s+d)/d$ and $\beta = d/(2s+d)$ if $\|K\|_{C^{2s}} < \infty$ and s is not an integer.*

We close the section with a few remarks.

Remark 3.1. Some estimators of $\psi'(m)$ are translation invariant in the sense that estimates based on observations Y_i and translated versions $Y_i' = Y_i + t$ differ by exactly t. The estimator $\hat{\psi}'_{ML}$ that we discuss here is not. This is a consequence of the regularization implicit in our choice of weights. In the averaged ridge regression interpretation of our estimator (3.7), the penalty $\|\cdot\|^2_{\mathcal{H}_K}$ that we use when we estimate $m(\cdot, 0)$ penalizes deviation of our estimator from the constant function $f(x) = 0$, even if that deviation takes the form of a constant translation. As penalties on translations are light for most reasonable RKHS norms, this is not generally a problem if $\mathbb{E}[Y_i \mid W_i = 0]$ is not too large. However, modifying our estimator so that it is translation invariant makes it somewhat more foolproof. A simple way to do this is to use the estimator $\hat{\psi}_{ML_t} = (n_T/n)\bar{Y}_0 + n^{-1}\sum_{i:W_i=0} \hat{\gamma}_i(Y_i - \bar{Y}_0)$ where $\bar{Y}_0 = n_Z^{-1}\sum_{i:W_i=0} Y_i$. This is a very simple augmented minimax linear estimator incorporating a constant estimate $\bar{Y}_0$ of $m(\cdot, 0)$. See Kallus (2016, Section 4.5) for an alternative approach to translation invariance and its generalizations that substitutes a translation invariant seminorm for the norm $\|\cdot\|_{\mathcal{H}_K}$.

Remark 3.2. Our first-order asymptotic result, Theorem 3.2, requires no assumptions on the Riesz representer γ_ψ beyond its existence and boundedness. Our assumption that $\gamma_\psi(x)$ is bounded is a strict overlap assumption in the sense of D'Amour et al. (2017), which ensures that our target population and the population that receives our treatment of interest are sufficiently similar that the rate at which $\psi(m)$ can be estimated is not impacted by identification issues (see e.g. Khan and Tamer, 2010).

Theorem 3.2 does require smoothness of the regression function $m(\cdot, 0)$. In particular, it requires that the RKHS $\mathcal{H}_K$ that we work with satisfies Assumption 3.3 or 3.4 and that $m(\cdot, 0)$ is smoother than the least smooth function in $\mathcal{H}_K$ in the sense that $\|m(\cdot, 0)\|_{L^{\kappa_m}_{K,P_Z}} < \infty$ for $\kappa_m > 1/2$.

Under Assumption 3.3, this implies Sobolev-type smoothness of $m(\cdot, 0)$ of order $s > d$. This seems to be twice as strong as should be necessary. Efficient estimation of $\psi(m)$ is possible so long as $m(\cdot, 0)$ is Hölder-smooth of order $s > d/2$ (Robins et al., 2009), and the linear 'plug-in' estimator

of Newey and Robins (2018) is efficient in this case. Furthermore, Sobolev-type smoothness of order $s > d/2$ is sufficient to show that that our estimator is consistent at $O_p(n^{-1/2})$ rate by the simple argument based on (3.6) discussed in Section 3.1. Thus, it would be strange if twice this level of smoothness were required for efficiency of our esitimator. This seemingly excessive level of smoothness is needed only to ensure an adequately slow rate of growth for the eigenfunctions of L_{K,P_Z} in supremum norm, a significant challenge in the characterization of the performance of RKHS methods (see discussion in Zhou, 2002). In some cases, it is clear that we do not need this degree of smoothness. In particular if $\mathcal{H}_K$ is the Sobolev space of periodic functions on the unit cube in $\mathbb{R}^d$ and P_Z is uniform measure on this cube, the eigenfunctions of L_{K,P_Z} will be the fourier basis functions, which are bounded in supremum norm.

The implication of Assumption 3.4 that $\|m(\cdot,0)\|_{C^s} < \infty$ for $s > d/2$ is closer to what we expect. This is the aforementioned minimal level of Hölder-type smoothness required for efficient estimation of $\psi(m)$. However, insofar as the finiteness of $\|m(\cdot,0)\|_{C^s}$ is implied by and not equivalent to our assumptions, this should not be taken as a claim that our results are comparable to those of Newey and Robins (2018).

Remark 3.3. While the first order asymptotic behavior of our estimator is not impacted by the smoothness of this Riesz representer γ_ψ, the higher order 'remainder' terms are strongly affected. In particular, our bound (3.14) decays no faster than $n^{-1/2}\lambda^{\min\{\frac{1}{2},\frac{2\kappa_\gamma}{1+2\kappa_\gamma}\}}$ where κ_γ is the largest κ such that $\|g_\psi\|_{L^\kappa_{K,P_Z}} < \infty$.

Remark 3.4. In Theorem 3.2, we take $\lambda = \sigma^2/n$ for constant σ. This choice is the natural one in our minimax approach, as $\sigma_n \to \infty$ or $\sigma_n \to 0$ would yield minimax estimators in settings in which the noise level was either increasing or decreasing with sample size. In addition, it is a robust choice, as it results in first-order asymptotic efficiency under no smoothness assumptions on γ_ψ and a smoothness assumption $\kappa_m > 1/2$ on $m(\cdot,0)$ that we cannot weaken by tuning λ differently.

However, other perspectives on our estimator motivate the use of λ asymptotically larger than $1/n$. In what follows, we will use the notation $a_n \ll b_n$ meaning $a_n/b_n \to 0$, $a_n \lesssim b_n$ meaning $\sup_n a_n/b_n < \infty$, and $a_n \sim b_n$ meaning $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Interpreting our estimator as an averaged ridge regression estimator (3.9) or as an inverse probability weighting estimator with inverse probability weights estimated by least squares (2.28), the choice $\lambda \sim 1/n$ results in unusually weak regularization. As discussed in Appendix A.1.3, by taking $\lambda \gg 1/n$, it is possible to get faster convergence of $\hat{\gamma}$ to the Riesz representer γ_ψ, and this phenomenon holds for convergence of our ridge regression estimator $\hat{m}$ to the regression function $m(\cdot,0)$ as well.

While this tuning approach requires greater smoothness of γ_ψ and $m(\cdot,0)$ for first-order asymp-

totic efficiency, it can lead to faster decay of the remainder. If the Riesz representer is smooth enough that the conditions of our Lemma are satisfied with $\kappa_\gamma \geq 1/2$, characterization of the optimal λ is straightforward. Our remainder rate is bounded by the sum of the dominant term $\lambda^{\tilde{\kappa}_m}$ from (3.13) and the dominant term $n^{-1/2}a^{1/2}$ from (3.14), which is on the order of $\max\{(n^{-3/4}r)^{\tilde{\kappa}_m/(\tilde{\kappa}_m+1/2)},\ n^{-\tilde{\kappa}_m}\}$ for $\tilde{\kappa}_m = \min\{\kappa_m, 1\}$ at the optimal choice $\lambda \sim \max\{(n^{-3/4}r)^{1/(\tilde{\kappa}_m+1/2)},\ n^{-1}\}$. As $r \ll n^{-1/4}$ whenever Theorem 3.4 applies, this rate is faster than $n^{-\tilde{\kappa}_m/(\tilde{\kappa}_m+1/2)}$, which ranges from $n^{-(1/2+\epsilon)}$ to $n^{-2/3}$ as the smoothness parameter $\tilde{\kappa}_m$ describing $m(\cdot, 0)$ increases from $1/2+\epsilon'$ to 1. Characterization of an optimal tuning parameter in terms of κ_m and $\kappa_\gamma \in (0, 1/2)$ is more complicated.

The immediate utility of this knowledge is limited, as this choice of the tuning parameter λ depends on the unknown parameters κ_m, κ_γ. However, it does provide potentially useful intuition: the tuning parameter λ that results in the smallest remainder is typically between our robust choice $\lambda \sim 1/n$ and the choice $\lambda \sim r^2$ that optimizes the rate of convergence of $\hat{\gamma}$ to γ_ψ (see Appendix A.1.3). Thus, we should not necessarily expect optimal performance either from tuning approaches that assume $\sigma = n\lambda$ should be roughly constant as a function of sample size or from approaches that tune λ for estimation of γ_ψ by cross-validation.

3.2 Proving the finite sample bounds

3.2.1 Proof of the bias term bound (3.13)

In this section, we prove bound (3.13) from Theorem 3.4. We will begin by proving the lemma below, then show that it implies the bound (3.13).

Lemma 3.6. *In the setting described in Section 3.1.2, consider the estimator* $\hat{\psi}_{ML} = n^{-1}\sum_{i=1}^{n} 1_{\{W_i=0\}}\hat{\gamma}_i Y_i$ *with weights* $\hat{\gamma}$ *defined in* (3.8). *Define* $\lambda = \sigma^2/n$, $L = L_{K,P_Z}$, $L_\lambda = L + \lambda I$, $p_Z = P\{W_i = 0\}$, *and* $p_T = P\{T_i = 1\}$.

For any $\eta, \delta \in (0,1)$, on an event of probability $1 - 3\delta$,

$$\begin{aligned}
\left|\mathbb{E}\left[\hat{\psi}_{ML} \mid X, W\right] - \tilde{\psi}(m)\right| &\leq s p_Z \|g_\psi\|_{L_2(P_Z)} \\
&\quad + 2(1+\eta) R_n\{T(x,w)b(x) : \|b\| \leq r, \|b\|_{L_2(P_Z)} \leq s\} \\
&\quad + s n^{-1/2} \sqrt{2 p_Z \|g_\psi\|_\infty \log(2\delta^{-1})} + 2 r M_K \left(\frac{1}{3} + \frac{1}{\eta}\right) \log(2\delta^{-1}) \qquad (3.15)
\end{aligned}$$

$$(r, s) = \zeta^{-1}\lambda \cdot \left(\|L_\lambda^{-1} m(\cdot, 0)\|,\ \|L_\lambda^{-1/2} m(\cdot, 0)\|\right);$$

$$\zeta = \max\left\{0,\ 1 - \frac{8U^2 \log(4\delta^{-1} n_\delta^2)}{n_\delta} - \sqrt{\frac{16U^2 \log(4\delta^{-1} n_\delta^2)}{n_\delta}}\right\}; \qquad (3.16)$$

$$U = \operatorname*{ess\,sup}_{X \sim P_Z} \left\| L_\lambda^{-1/2} K_X \right\|; \qquad (3.17)$$

$$n_\delta = n p_Z - \sqrt{2 n p_Z \log(\delta^{-1})}. \qquad (3.18)$$

Our approach works with the averaged ridge regression interpretation our our estimator. In (3.5) above, we decomposed the error $\hat{\psi}_{ML} - \psi(m)$ of our estimator, written in weighting form (3.8), into its design-conditional bias and its variation around it. Consider the same decomposition of our estimator expressed in averaged ridge regression form (3.9).

$$\hat{\psi}_{ML} - \tilde{\psi}(m) = \underbrace{\frac{1}{n}\sum_{i=1}^n T_i[\mathbb{E}\left[\hat{m}(X_i) \mid X, W\right] - m(X_i, 0)]}_{\text{bias}} + \underbrace{\frac{1}{n}\sum_{i=1}^n T_i[\hat{m}(X_i) - \mathbb{E}\left[\hat{m}(X_i) \mid X, W\right]]}_{\text{noise}}. \qquad (3.19)$$

The quantity $\mathbb{E}[\hat{\psi}_{ML} \mid X, W] - \tilde{\psi}(m)$ that we are bounding is an average $n^{-1}\sum_{i=1}^n T_i b(X_i)$ of the conditional bias function $b = \mathbb{E}\left[\hat{m}(X_i) \mid X, W\right] - m(X_i, 0)$ of our regression estimator.

To bound this quantity, we proceed in two steps. In the first step, we will show that on a high-probability event $\mathcal{A}_1$, this function b is in a set $\mathcal{B} := \{b' : \|b'\| \leq r, \|b'\|_{L_2(P_Z)} \leq s\}$ for certain r, s. To do this, we work conditionally on W, considering $\{X_i : i \leq n, W_i = 0\}$ to be an iid sequence of length $n_Z = \sum_{i=1}^n 1_{\{W_i = 0\}}$ from the conditional distribution P_Z of X_i given $W_i = 0$. As this will determine bounds r', s' in terms of the random variable n_Z, we will show that on a high probability event $\mathcal{A}_2$, n_Z/n is nearly as large as its mean $P\{Z_i = 1\}$, and as a consequence define nonrandom bounds r, s holding on $\mathcal{A}_1 \cap \mathcal{A}_2$. In the second step, we will bound $\sup_{b' \in \mathcal{B}} |n^{-1}\sum_{i=1}^n T_i b'(X_i)|$ on another high probability event $\mathcal{A}_3$. As a consequence, we will have a bound on $|n^{-1}\sum_{i=1}^n T_i b(X_i)|$ on the event $\cap_{k=1}^3 \mathcal{A}_k$, which holds with probability $1 - \sum_{k=1}^3 P\{\mathcal{A}_k\}$ by the union bound.

3.2.1.1 Characterizing the conditional bias function b

The optimization problem (3.9) defininging $\hat{m}$ has an explicit solution (see e.g. Hsu et al., 2012)

$$\hat{m} = \left[\frac{1}{n_Z}\sum_{i:W_i=0} K_{X_i} \otimes K_{X_i} + \lambda I\right]^{-1} \frac{1}{n_Z}\sum_{i:W_i=0} K_{X_i} Y_i,$$

written in terms of $\lambda = \sigma^2/n_Z$ and the rank-one operator $K_x \otimes K_x$ defined by $[K_x \otimes K_x]f = K_x\langle f, K_x\rangle$. Then because for the $W_i = 0$ units $\mathbb{E}[Y_i \mid X, W] = m(X_i, 0) = \langle K_{X_i}, m(\cdot, 0)\rangle$, we have $K_{X_i}\mathbb{E}[Y_i \mid X, W] = K_{X_i}\langle K_{X_i}, m(\cdot, 0)\rangle = [K_{X_i} \otimes K_{X_i}]m(\cdot, 0)$. In terms of the operator $\widehat{L} = n_Z^{-1}\sum_{i:W_i=0} K_{X_i} \otimes K_{X_i}$ we may write

$$\mathbb{E}[\hat{m} \mid X, W] = \left[\widehat{L} + \lambda I\right]^{-1} \widehat{L}\, m(\cdot, 0) = \left(I - \lambda\left[\widehat{L} + \lambda I\right]^{-1}\right) m(\cdot, 0)$$

and therefore

$$b = -\lambda\left[\widehat{L} + \lambda I\right]^{-1} m(\cdot, 0).$$

Note that the operator $\widehat{L}$ is an empirical version of our integral operator $L := L_{K,P_Z}$. Our characterization of b will rely on the convergence of $\widehat{L}$ to its mean L in the operator norm $\|A\| = \sup_{\|f\|\le 1}\|Af\|$. Using the shorthand $\widehat{L}_\lambda := \widehat{L} + \lambda I$ and $L_\lambda := L + \lambda I$ for the regularized versions of these operators, we may write

$$b = -\lambda[\widehat{L}_\lambda^{-1} L_\lambda][L_\lambda^{-1} m(\cdot, 0)].$$

From this decomposition, we get the bound

$$\|b\| \le \lambda\|\widehat{L}_\lambda^{-1} L_\lambda\|\|L_\lambda^{-1} m(\cdot, 0)\|. \tag{3.20}$$

Furthermore, because $\|b\|_{L_2(P_Z)} \le \|L_\lambda^{1/2} b\|$, analogously we have the bound

$$\|b\|_{L_2(P_Z)} \le \lambda\left\|L_\lambda^{1/2}\widehat{L}_\lambda^{-1} L_\lambda^{1/2}\right\|\left\|L_\lambda^{-1/2} m(\cdot, 0)\right\|. \tag{3.21}$$

These two bounds form the basis for our characterization of b as an element of the set $\mathcal{B}$. The operator norm factors in the two expressions above are the same, as $\|AB\| = \|B^{1/2}AB^{1/2}\|$ for any operator A and positive operator B.[4] We bound this quantity using an argument of Hsu, Kakade, and Zhang (2012, Lemmas 25 and 26), observing first that

$$\left\|L_\lambda^{1/2}\widehat{L}_\lambda^{-1} L_\lambda^{1/2}\right\| \le (1 - \|\Delta_\lambda\|)^{-1} \quad \text{where} \quad \Delta_\lambda = L_\lambda^{-1/2}(\widehat{L}_\lambda - L_\lambda)L_\lambda^{-1/2}.$$

Here Δ_λ is a centered mean of iid rank-one operators,

$$\Delta_\lambda = n_Z^{-1}\sum_{i:W_i=0} \tilde{X}_i \otimes \tilde{X}_i - \mathbb{E}_{X\sim P_Z}\tilde{X}_i \otimes \tilde{X}_i \text{ where } \tilde{X}_i = L_\lambda^{-1/2} K_{X_i},$$

[4] Check that when ϕ is an eigenvector of AB, $B\phi$ is an eigenvector of BA with the same eigenvalue.

and it satisfies the condition $\|\mathbb{E}_{X\sim P_Z} \tilde{X}_i \otimes \tilde{X}_i\| = \|L_\lambda^{-1/2} L L_\lambda^{-1/2}\| \leq 1$. We bound $\|\Delta_\lambda\|$ using a concentration inequality for such averages (Oliveira et al., 2010, Lemma 1).[5] If $\|\tilde{X}_i\| \leq U$ almost surely, with probability $1-\delta$,

$$\left\| n_Z^{-1} \sum_{i:W_i=0} \tilde{X}_i \otimes \tilde{X}_i - \mathbb{E}_{X\sim P_Z} \tilde{X}_i \otimes \tilde{X}_i \right\| < \frac{8U^2 \log(4\delta^{-1} n_Z^2)}{n_Z} + \sqrt{\frac{16U^2 \log(4\delta^{-1} n_Z^2)}{n_Z}}$$

Consequently, with probability $1-\delta$,

$$\left\| L_\lambda^{1/2} \widehat{L}_\lambda^{-1} L_\lambda^{1/2} \right\| \leq \max\left\{ 0, 1 - \frac{8U^2 \log(4\delta^{-1} n_Z^2)}{n_Z} - \sqrt{\frac{16U^2 \log(4\delta^{-1} n_Z^2)}{n_Z}} \right\}^{-1}.$$

As long as U is small relative to $\sqrt{n_Z}$, this operator norm will be essentially one, and our bounds on $\|b\|$ and $\|b\|_{L_2(P_Z)}$ will be roughly $\lambda \|L_\lambda^{-1} m(\cdot, 0)\|$ and $\lambda \|L_\lambda^{-1/2} m(\cdot, 0)\|$ respectively. We state our results in terms of the sharp upper bound $U = \operatorname{ess\,sup}_{X\sim P_T} \|L_\lambda^{-1/2} K_X\|$.

To eliminate this bound's dependence on n_Z, observe that $\log(4\delta^{-1}x)/x$ is an increasing function, so our bound will remain valid if we substitute an upper bound on n_Z. Furthermore, in terms of $p_Z = P\{W_i = 0\}$, $n_Z \geq n(1-\epsilon)p_Z$ with probability $1 - \exp\{-n\epsilon^2 p_Z/2\}$ by the lower tail of the multiplicative Chernoff bound (see e.g. Mitzenmacher and Upfal, 2005, Theorem 4.5). For $\epsilon = [2\log(\delta^{-1})/(np_Z)]^{1/2}$, this is probability $1-\delta$, and we have $(1-\epsilon)n = np_Z - \sqrt{2np_Z \log(\delta^{-1})}$. Therefore by the union bound, with probability $1-2\delta$,

$$\left\| L_\lambda^{1/2} \widehat{L}_\lambda^{-1} L_\lambda^{1/2} \right\| \leq \zeta^{-1} \quad \text{for} \quad \zeta = \max\left\{ 0, 1 - \frac{8U^2 \log(4\delta^{-1} n_\delta^2)}{n_\delta} - \sqrt{\frac{16U^2 \log(4\delta^{-1} n_\delta^2)}{n_\delta}} \right\}; \tag{3.22}$$
$$n_\delta = np_Z - \sqrt{2np_Z \log(\delta^{-1})}.$$

We take r, s in the definition of the set $\mathcal{B}$ to be the values of the bounds (3.20) and (3.21) with this bound substituted.

3.2.1.2 Bounding empirical averages over $\mathcal{B}$

First, note that because $\mathcal{B}$ is symmetric, we can drop the absolute value. We then rewrite our empirical average $n^{-1}\sum_{i=1}^n T_i b'(X_i)$ as the sum of its mean and its deviation around it, $\mathbb{E}\, T_i b'(X_i) + n^{-1}\sum_{i=1}^n [T_i b'(X_i) - \mathbb{E}\, T_i b'(X_i)]$. We will take the supremum of each term over $\mathcal{B}$ separately.

[5] Hsu, Kakade, and Zhang (2012) complete this argument by invoking a similar inequality. Theirs involves a log factor involving a parameter of L_λ, whereas the one we use here involves a factor of $\log(n)$ in its place.

We bound the mean via a change of measure and Cauchy-Schwartz.

$$\begin{aligned}
\mathbb{E}\, T_i b'(X_i) &= p_T \mathbb{E}\left[b'(X_i) \mid T_i = 1\right] \quad \text{for } p_T = P\{T_i = 1\} \\
&= p_Z\, \mathbb{E}_{X \sim P_Z}\left[b'(X) g_\psi(X)\right] \quad \text{as} \quad g_\psi(x) = \frac{P\{T_i = 1 \mid X = x\}}{P\{W_i = 0 \mid X = x\}} = \frac{p_T dP_T}{p_Z dP_Z}(x) \\
&\leq p_Z \|b'\|_{L_2(P_Z)} \|g_\psi\|_{L_2(P_Z)} \\
&\leq s p_Z \|g_\psi\|_{L_2(P_Z)}.
\end{aligned} \tag{3.23}$$

Furthermore, by the same line of reasoning we have $\mathbb{E}(T_i b'(X_i))^2 \leq s^2 p_Z \|g_\psi\|_\infty$, as via Hölder's inequality,

$$\mathbb{E}\, T_i b'(X_i)^2 = p_Z\, \mathbb{E}_{X \sim P_Z}\left[b'(X)^2 g_\psi(X)\right] \leq p_Z \|b'\|^2_{L_2(P_Z)} \|g_\psi\|_\infty.$$

In addition, we have $\|b'\|_\infty \leq r M_K$ where $M_K = \sup_{\|f\| \leq 1} \|f\|_\infty$. We use these in our bound on the deviation term, for which we use a form of Talagrand's inequality (Bartlett et al., 2005, Theorem 2.1). As the class of functions $\{T(x,w)b'(x) : b \in \mathcal{B}\}$ satisfies the bounds $\|T(x,w)b'(x)\|_{L_2(P)} \leq s p_Z^{1/2} \|g_\psi\|_\infty^{1/2}$ and $\|T(x,w)b'(x)\|_\infty \leq r M_K$, with probability $1 - \delta$,

$$\begin{aligned}
&\sup_{b' \in \mathcal{B}} \left| n^{-1} \sum_{i=1}^n \left(T_i b'(X_i) - \mathbb{E}\, T_i b'(X_i)\right) \right| \leq t_\eta \quad \text{for all} \quad \eta > 0; \\
&t_\eta = 2(1+\eta) R_n\{T(x,w)b'(x) : b' \in \mathcal{B}\} + s\sqrt{\frac{2 p_Z \|g_\psi\|_\infty \log(2\delta^{-1})}{n}} + 2 r M_K \left(\frac{1}{3} + \frac{1}{\eta}\right) \log(2\delta^{-1}).
\end{aligned} \tag{3.24}$$

By the union bound, the intersection of this event and the event on which (3.22) holds has probability at least $1 - 3\delta$. On this intersection, our mean and deviation bounds above apply to b. Adding them yields the bound (3.15) that we set out to prove. This completes our proof of Lemma 3.6.

3.2.1.3 Proving (3.13) from Lemma 3.6

To prove (3.13) from Lemma 3.6, we substitute upper bounds for a few quantities in (3.15). To establish these bounds, we use the lemmas stated below, which are proven in Appendix B.4. Lemma 3.7 implies that that our expression for ζ in terms of α, β, n in (3.13) bounds the corresponding quantity ζ in (3.15). Lemma 3.8 implies that our expressions for s and r as multiples of ζ^{-1} in (3.13) bound those in (3.15). As in these cases our lemmas give exactly the quantities that appear in (3.13), we will not discuss those terms further.

To bound the second term in (3.15), we will use Lemma 3.9, a generalization of Mendelson's bound on the local Rademacher complexity of the unit ball in an RKHS (Mendelson, 2002). This term is $2(1+\eta)r$ times the the Rademacher complexity of the set $\{T_i b'(x) : \|b'\| \leq 1, \|b'\|_{L_2(P_Z)} \leq s/r\}$, and as established above, this $\|\cdot\|_{L_2(P_Z)}$ bound on b' implies that $\|T(x,w)b'(x)\|_{L_2(P)} \leq t$ for

$t = (s/r)(p_Z \|g_\psi\|_\infty)^{1/2}$. Thus, it suffices to bound the Rademacher complexity R of the set $\{T_i b(x) : \|b\| \le 1, \mathbb{E}(T_i b(X_i))^2 \le t^2\}$, and we apply Lemma 3.9 with $g = 0$, $Z_i = T_i$, and an iid Rademacher sequence $\sigma_1 \dots \sigma_n$ independent of $(X_i, W_i)_{i \le n}$. This yields the bound $R^2 \le (2/n) \sum_{j=1}^\infty \lambda_j \wedge t^2$ in terms of the eigenvalues λ_j of the integral operator $L_{K,\nu}$ associated with the measure $\nu = p_T \cdot P_T$, a scaled version of the distribution of the covariate X_i on the target population. Thus, $\lambda_j = p_T \lambda_{j,T}$ for eigenvalues $\lambda_{j,T}$ of L_{K,P_T}, and our bound may be rewritten in the form $R^2 \le (2/n) \sum_{j=1}^\infty (p_T \lambda_{j,T}) \wedge t^2$ and bounded using Lemma 3.10 to complete our proof.

Lemma 3.7. *Let $\mathcal{H}_K$ be an RKHS of functions on a compact set $\mathcal{X}$, Let ν be a finite measure with support equal to $\mathcal{X}$, define $[L_{K,\nu} f](x) = \int K(x,t) f(t) d\nu(t)$, and let $(\lambda_j, \phi_j)_{i \in \mathbb{N}}$ be its eigenvalues and eigenfunctions scaled so that $\|\phi_j\|_{L_2(\nu)} = 1$, and assume that $\lambda_j \le C_\lambda j^{-\alpha}$ and that $\|\phi_j\|_{L_\infty(\nu)} \le C_\phi \lambda_j^{-\beta/2}$ with $\alpha(1-\beta) > 2$. Then,*

$$\operatorname*{ess\,sup}_{X \sim \nu} \left\| [L_{K,\nu} + \lambda I]^{-1/2} K_x \right\| \le C \lambda^{-(1/\alpha + \beta)/2} \quad \text{where}$$

$$C = C_\phi C_\lambda^{1/(2\alpha)} \left[\left(\frac{\beta}{1-\beta} \right)^{1/\alpha + \beta} + \frac{\alpha}{(1+\alpha\beta)(\alpha - (1+\alpha\beta))} \right]^{1/2}.$$

Lemma 3.8. *Let $\mathcal{H}_K$ be an RKHS of functions on a compact set $\mathcal{X}$, let ν be a finite measure with support equal to $\mathcal{X}$, and let λ_1 be the largest eigenvalue of $[L_{K,\nu} f](x) = \int K(x,t) f(t) d\nu(t)$. Then,*

$$\begin{aligned} \|[L_{K,\nu} + \lambda I]^{-1/2} f\| &\le \begin{cases} \lambda^{\kappa - 1} \|f\|_{L^\kappa_{K,\nu}} & \kappa \in [1/2, 1) \\ \lambda_1^{\kappa - 1} \|f\|_{L^\kappa_{K,\nu}} & \kappa \ge 1 \end{cases} \\ \|[L_{K,\nu} + \lambda I]^{-1} f\| &\le \begin{cases} \lambda^{\kappa - 3/2} \|f\|_{L^\kappa_{K,\nu}} & \kappa \in [1/2, 3/2) \\ \lambda_1^{\kappa - 3/2} \|f\|_{L^\kappa_{K,\nu}} & \kappa \ge 3/2. \end{cases} \end{aligned} . \tag{3.25}$$

Lemma 3.9. *Let $\mathcal{H}_K$ be an RKHS of functions on a compact set $\mathcal{X}$, let $(X_1, Z_1) \dots (X_n, Z_n) \overset{iid}{\sim} \nu_{x,z}$ where the marginal ν_x on X_i has support equal to $\mathcal{X}$, and let $s_z(x) = \mathbb{E}\left[Z_i^2 \mid X_i = z\right]$ satisfy $s_z(x) > 0$ a.e.$-\nu_x$. Define the measure ν by $d\nu = s_z d\nu_x$ and let $\{\lambda_j : j \in 1 \dots \infty\}$ be the eigenvalues of $[L_{K,\nu} f](x) = \int K(x,t) f(t) d\nu(t)$ in decreasing order. For a ν-square-integrable function g, define the set $\mathcal{B}^\star = \{f - sg : \|f\|_{\mathcal{H}_K} \le 1, s \in [0,1]\}$. In terms of an identically distributed sequence $\sigma_1 \dots \sigma_n$ satisfying $\mathbb{E}\left[\sigma_i \sigma_j \mid X_1, Z_1 \dots X_n, Z_n\right] = 0$ for $i \ne j$, the local multiplier complexity*

$$M_n\{z f(x) : f \in \mathcal{B}^\star,\ \mathbb{E}(Z_i f(X_i))^2 \le t^2\} := \sup_{\substack{f \in \mathcal{B}^\star \\ \mathbb{E}(Z_i f(X_i))^2 \le t^2}} \left| n^{-1} \sum_{i=1}^n \sigma_i Z_i f(X_i) \right|$$

is bounded by

$$3^{1/2} \left\| \mathbb{E}\left[\sigma_i^2 \mid X_i, Z_i\right] \right\|_{L_\infty(\nu_{x,z})} n^{-1/2} \sqrt{\sum_{j=0}^\infty \lambda_j \wedge t^2} \quad \text{where} \quad \lambda_0 = (1 + \sqrt{\lambda_1})^2 \|g\|^2_{L_2(\nu)}.$$

If $g = 0$, we may take $\lambda_0 = 0$ and the leading constant to be $2^{1/2}$. For $t = \infty$, we have the tighter bound

$$2^{1/2} \left\| \mathbb{E}\left[\sigma_i^2 \mid X_i, Z_i\right] \right\|_{L_\infty(\nu_{x,z})} n^{-1/2} \left(\|g\|_{L_2(\nu)} + \sqrt{\sum_{j=1}^{\infty} \lambda_i} \right).$$

If $g = 0$, we may take the leading constant to be 1.

Lemma 3.10. *If $\lambda_j \leq Cn^{-\alpha}$ for $\alpha > 1$, $\sum_{j=1}^{\infty} \lambda_j \wedge t^2 \leq C^{1/\alpha}(1 - 1/\alpha)^{-1} t^{2(1-1/\alpha)}$.*

3.2.2 Proof of the noise term bound (3.14)

In this section, we prove the bound (3.14) from Theorem 3.4. This is a slight variation on the bound (2.21) from the previous chapter. We will work with a characterization of the noise term from (3.5),

$$\hat{\psi}_{ML} - \mathbb{E}[\hat{\psi}_{ML} \mid X, W] = 1_{\{W_i = 0\}} \hat{\gamma}_i \varepsilon_i \quad \text{where} \quad \varepsilon_i = Y_i - m(X_i, W_i).$$

We will show convergence of this quantity to the iid sum $n^{-1} \sum_{i=1}^{n} \gamma_\psi(X_i, W_i)\varepsilon_i$ by showing convergence of $1_{\{W_i=0\}}\hat{\gamma}_i$ to $\gamma_\psi(X_i, W_i)$. To do this, we use the previous chapter's Lemma 2.8. This suffices, as in Section 2.1.3.5 we've shown that if $\gamma_1 \dots \gamma_n$ satisfy the bound $n^{-1} \sum_{i=1}^{n} (\gamma_i - \hat{\gamma}_\psi(X_i, W_i))^2 \leq a \wedge b$ with probability $1 - \delta'$, then the bound (3.14) we aim to prove holds with probability $1 - \delta - \delta'$.

In order to apply Lemma 2.8 in this setting, we must establish that the weights $\hat{\gamma}$ that we discuss here are an instance of the weights $\hat{\gamma}$ that we discuss in the previous chapter. We use the following characterization of the solution to optimization problem (2.15), which specializes the previous chapter's Lemma 2.5 to our setting. We prove this proposition in Appendix B.4.

Proposition 3.11. *Let $h(x, w, f) = T(x, w)f(x, 0)$, let $\mathcal{B}$ be the unit ball of a reflexive space of functions on a set $\mathcal{X}$, and let $\mathcal{B}^C$ be the unit ball for the cartesian product of $C + 1$ copies of this space considered as functions $f(x, w)$ on $(\mathcal{X}, \{0 \dots C\})$. Then the primal problem*

$$\ell_{n,\mathcal{B}^C}(\gamma) = I^2_{h,\mathcal{B}^C}(\gamma) + \frac{\sigma^2}{n^2}\|\gamma\|^2, \; I_{h,\mathcal{F}} = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left[h(X_i, W_i, f) - \gamma_i f(X_i, W_i) \right], \tag{3.26}$$

has a unique minimum at $\hat{\gamma}$ satisfying $\hat{\gamma}_i = 0$ for $W_i \neq 0$. Furthermore, the dual

$$M_{n,\mathcal{B}^C}(g) = -\frac{\sigma^2 \|g\|^2_{\mathcal{B}^C}}{n} - \frac{1}{n} \sum_{i=1}^{n} g(X_i, W_i)^2 + \frac{2}{n} \sum_{i=1}^{n} h(X_i, W_i, g), \tag{3.27}$$

has a possibly nonunique maximum, and for any $\hat{g}$ at which its maximum is attained, $\hat{\gamma}_i = \hat{g}(X_i, W_i)$ and $\hat{g}(\cdot, w) = 0$ for $w \neq 0$.

Here we take $\mathcal{B}$ to be the unit ball of our RKHS $\mathcal{H}_K$. Subject to the constraint that $\hat{\gamma}_i = 0$ if $W_i \neq 0$, a constraint that is satisfied by the solution of (3.26), this problem reduces to the problem

(3.8) that defines our weights. Thus, our weights solve it. Having established this, we may now show convergence of $\hat{\gamma}$ to γ_ψ by using Lemma 2.8 to establish convergence of $\hat{g}$ to γ_ψ.

As we know that both $\hat{g}$ and γ_ψ satisfy the property $g(\cdot, w) = 0$ for $w \neq 0$, we apply Lemma 2.8 with $\tilde{\gamma}$ satisfying this property and with $\tilde{\mathcal{F}} = \mathcal{F} = \{f : \|f(\cdot,0)\|_{\mathcal{H}_K} \leq 1, f(\cdot, w) = 0 \text{ for } w \neq 0\}$. The resulting bound will be stated in terms of a few properties of the sets $\mathcal{F}^\star(t) = \{f - s\gamma_\psi : f \in \mathcal{F}, s \in [0,1], \ \|f - \gamma_\psi\|_{L_2(P)} \leq t\}$ and $\mathcal{H}^\star(t) = \{T(x,w)f(x,0) - \gamma_\psi(x,w)f(x,w) : f \in \mathcal{F}^\star(t)\}$. The relevant properties are, in terms of a convenient choice of constant $\eta_Q = (61 - 8\sqrt{39})/49 \approx .23$ and arbitrary $\eta_C > 0$,

$$
\begin{aligned}
&r_Q(\eta_Q) = 7 \inf\{r > 0 : R_n(\mathcal{F}^\star(r)) \leq r^2/(2M_{\mathcal{F}^\star})\} \text{ and} \\
&r_C(\eta_C, \delta) = \inf\{r > 0 : u(\mathcal{H}^\star(r), \delta) \leq \eta_C r^2\} \text{ where} \\
&u(\mathcal{H}, \delta) = \min_{\eta > 0} 2(1+\eta)R_n(\mathcal{H}) + \overline{\sigma}(\mathcal{H})\sqrt{\frac{2\log(2\delta^{-1})}{n}} + 2M_{\mathcal{H}}\left(\frac{1}{3} + \frac{1}{\eta}\right)\frac{\log(2\delta^{-1})}{n};^{6} \\
&M_{\mathcal{G}} = \sup_{g \in \mathcal{G}} \|g\|_\infty; \ \overline{\sigma}(\mathcal{G}) = \sup_{g \in \mathcal{G}} \|g\|_{L_2(P)}.
\end{aligned}
$$

It will also be stated in terms of a measure of the approximability of the Riesz representer γ_ψ by a function $\tilde{\gamma}$ with $\tilde{\gamma}(\cdot, 0) \in \mathcal{H}_K$, specifically a bound $\bar{R}$ satisfying with probability $1 - \delta$

$$
\bar{R} \geq \frac{1}{n}\sum_{i=1}^n [(\tilde{\gamma} - \gamma_\psi)(X_i, W_i)]^2 - \frac{2}{n}\sum_{i=1}^n (T_i - \gamma_\psi(X_i, W_i))(\tilde{\gamma} - \gamma_\psi)(X_i, W_i) + \frac{\sigma^2 \|\tilde{\gamma}(\cdot, 0)\|^2_{\mathcal{H}_K}}{n}. \quad (3.28)
$$

In terms of these quantities, Lemma 2.8 yields the bound $n^{-1}\sum_{i=1}^n (\hat{g}(X_i, W_i) - \gamma_\psi(X_i, W_i))^2 \leq a \wedge b$ with probability $1 - \exp\{-c_1(\eta_Q) n r_Q(\eta_Q)^2 / M^2_{\mathcal{F}^\star}\} - 4\delta$ where

$$
\begin{aligned}
a &= \alpha u(\mathcal{H}^\star, \delta) + \bar{R}; \\
b &= 2\alpha^2 r^2 \vee 2\frac{\bar{R} + \sigma^2/n}{\eta_Q - 2\alpha^{-1}\eta_C} \vee \frac{44 M^2_{\mathcal{F}^\star}\alpha^2 \log(\delta^{-1})}{n}; \\
\alpha &= 1 \vee \left[2\eta_C \sigma^{-2} n r^2 + \sigma^{-1} n^{1/2} \bar{R}^{1/2}\right]; \\
r &= r_Q(\eta_Q) \vee r_C(\eta_C, \delta) \vee \sigma n^{-1/2} \eta_Q^{-1/2}; \\
c_1(\eta_Q) &= \frac{(1 - \eta_Q)^2}{2(1 + \eta_Q)(21 - 11\eta_Q)} \approx .02
\end{aligned} \quad (3.29)
$$

To complete our proof, it suffices to bound these quantities. We do this in Appendix B.5, bounding $u(\mathcal{H}^\star, \delta)$, $r_Q(\eta_Q)$, and $r_C(\eta_C, \delta)$ using Lemmas 3.9 and 3.10. The lemma stated below, which is proven in Appendix B.4, characterizes $\bar{R}$ when we take $\lambda = \sigma^2/n$.

Lemma 3.12. *Suppose that we observe $X_1, W_1 \ldots X_n, W_n$ iid and let P_Z be the conditional distribution of X_i given $W_i = 0$ and have support equal to a compact set $\mathcal{X}$. Let $\mathcal{H}_K$ be an RKHS of*

[6] Here rather than the general definition (2.18) of $u(\cdot, \delta)$ used in the previous chapter, we use a specific instance based on a convenient form of Talagrand's inequality (Bartlett et al., 2005, Theorem 2.1).

functions of $\mathcal{X}$ *and* $\gamma_\psi(x,w) = 1_{\{w=0\}} g_\psi(x)$ *be the Riesz representer for the functional* $f(x,w) \to \mathbb{E}\, T(x,w) f(x,0)$ *and for an approximation* $\tilde{\gamma}(x,w) = 1_{\{w=0\}}\tilde{g}(x)$*, define*

$$\bar{R}_{\lambda,\tilde{\gamma}} = \frac{1}{n}\sum_{i=1}^{n}(\tilde{\gamma}(X_i,W_i) - \gamma_\psi(X_i,W_i))^2 - \frac{2}{n}\sum_{i=1}^{n}(T_i - \gamma_\psi(X_i,W_i))(\tilde{\gamma}(X_i,W_i) - \gamma_\psi(X_i,W_i)) + \lambda\|\tilde{g}\|^2_{\mathcal{H}_K}.$$

1. *If* $\|g_\psi\|_{L_2(P_Z)} < \infty$*,* $\mathcal{H}_K$ *is dense in* $L_2(P_Z)$*, and* $\lambda_n \to 0$*, then* γ_ψ *has a sequence of approximations* $\tilde{\gamma}_n(x,w) = 1_{\{w=0\}}\tilde{g}_n(x)$ *such that* $\bar{R}_{\lambda_n,\tilde{\gamma}_n} = o_p(1)$*.*

2. *Furthermore, if* $\|g_\psi\|_{L^\kappa_{K,P_Z}} < \infty$ *for* $\kappa \in (0,1/2)$*,* γ_ψ *has an approximation* $\tilde{\gamma}(x,w) = 1_{\{w=0\}}\tilde{g}(x)$ *such that with probability* $1-\delta$*,*

$$\begin{aligned}\bar{R}_{\lambda,\tilde{\gamma}} \le\ & 2\left(\delta^{-1}p_Z\right)^{\frac{1-2\kappa}{1+2\kappa}} \|g_\psi\|^{\frac{4}{1+2\kappa}}_{L_2(P_Z)} \left(\theta^{-\frac{\theta}{1+\theta}} + \theta^{\frac{1}{1+\theta}}\right)\lambda^{\frac{4\kappa}{1+2\kappa}} \\ & + 2^{\frac{4(1-\kappa)}{1-2\kappa}}\left(\delta^{-1}p_Z\right)^{\frac{1}{1+2\kappa}} \|g_\psi\|_{L_\infty(P_Z)}\|g_\psi\|^{\frac{4}{1-4\kappa^2}}_{L_2(P_Z)}\theta^{\frac{\theta}{2(\theta+1)}} n^{-1/2}\lambda^{\frac{2\kappa}{1+2\kappa}}.\end{aligned}$$

where $p_Z = P\{W_i = 0\}$ *and* $\theta = 4\kappa/(1-2\kappa)$*.*

3.3 Empirical Performance

We evaluate the performance of our estimator on the famous example of Kang and Schafer (2007). In this example, we estimate a mean outcome when some outcomes are missing by a strongly ignorable mechanism (Rosenbaum and Rubin, 1983), an instance of the estimand $\psi(m)$ that we've been discussing in which we take $W_i \in \{0,1\}$ and $T_i = 1$ for all i.

The estimators under comparison are (i) an averaged regression estimator $n^{-1}\sum_{i=1}^n \hat{m}(X_i)$ where $\hat{m}$ is an estimate of $m(x,0)$ by ordinary least squares (OLS) on the treated units; (ii) an inverse propensity weighting (IPW) estimator $n^{-1}\sum_{i=1}^n 1_{\{W_i=0\}}\hat{e}(X_i)^{-1}Y_i$ where $\hat{e}(x)$ is a logistic regression estimator of $P\{Z_i = 0 \mid X_i = x\}$; (iii) an augmented inverse probability weighting (AIPW) estimator $n^{-1}\sum_{i=1}^n \hat{m}(X_i) + 1_{\{W_i=0\}}\hat{e}(X_i)^{-1}(Y_i - \hat{m}(X_i))$ incorporating the aforementioned estimators $\hat{m}$ and $\hat{e}$; and (iv) the minimax linear estimator $\hat{\psi}_{ML}$ (ML) and (v) the translation invariant variant $\hat{\psi}_{ML_t}$ (MLt) discussed in Remark 3.1. The latter estimators use the Matérn Kernel, $K(x,y) = k_\nu(\|x-y\|)$ for $k_\nu(x) = \frac{(\sqrt{2\nu}x)^\nu}{2^{\nu-1}\Gamma(\nu)}BK_\nu(\sqrt{2\nu}x))$ where BK_ν is a modified Bessel function of the second kind. The RKHS associated with this kernel is the Sobolev space H^s for $s = d/2+\nu$ (Schaback, 2011). We take ν to be $3/2$ and the primary level of the parameter σ in (3.8) to be 0.1, although we will display some additional results for $\sigma = 1$ and $\sigma = 10$. Calculation of the estimators is straightforward, amounting to the solution of a symmetric $n \times n$ linear system, as discussed in the Proof of Lemma 3.1 in Appendix B.4.

We will look at, in addition to root mean squared error and bias, the width and coverage of 95% confidence intervals of the form $\hat{\psi} \pm z_{.025}\widehat{V}^{1/2}/n^{1/2}$, where

$$\widehat{V} = n^{-1} \sum_{i:T_i=1} \left(\hat{m}(X_i) - \hat{\psi}\right)^2 + n^{-1} \sum_{i:W_i=0} \hat{\gamma}_i^2 \left(Y_i - \hat{m}(X_i)\right)^2. \tag{3.30}$$

Here $\hat{\gamma}_i$ are the weights used in the given estimator[7] and $\hat{m}$ is an OLS estimate of $m(x, 0)$ based on the sample receiving treatment $W_i = 0$. $\widehat{V}$ is based on a variant of the asymptotic characterization (3.11) with the limit of $\hat{\gamma}_i$ substituted for $\gamma_\psi(X_i, W_i)$, which will hold for these estimators so long as the conditional bias $\mathbb{E}[\hat{\psi} \mid X, W] - \psi(m) = o_p(n^{-1/2})$.

The Kang and Schafer example was designed to illustrate that methods using estimated inverse propensity weights can be unstable. Our observations $X_i \in \mathbb{R}^4, W_i \in \{0, 1\}, Y_i \in \mathbb{R}$ are defined in terms of a latent vector of standard normal random variables $Z_i \in \mathbb{R}^4$: we have $X_{i1} = \exp(Z_{i1}/2), X_{i2} = Z_{i2}/(1+\exp(Z_{i1})+10), X_{i3} = (Z_{i1}Z_{i3}/25+.06)^3, X_{i4} = (Z_{i2}+Z_{i4}+20)^2$; $P\{W_i = 0 \mid Z_i\} = logit^{-1}(-Z_{i1}+0.5Z_{i2}-0.25Z_{i3}-0.1Z_{i4})$; and $Y_i = 210+27.4Z_{i1}+13.7(Z_{12}+Z_{i3}+Z_{i4})+\sigma_\varepsilon\varepsilon_i$ for standard normal ε_i when $W_i = 0$. In this example, the instability of the IPW and AIPW estimators persists even into large sample sizes, while the OLS estimator performs extremely well even in small samples. These phenomena are explained in detail by a comment on Kang and Schafer (2007) by Robins et al. (2007). In summary, there are regions of poor overlap between the distributions of the covariate X_i between the treated and untreated subpopulations, which results in large inverse probability weights and therefore instability, but $m(x, 0)$ is sufficiently linear throughout the support of X_i that an estimator fit on the treated units extrapolates well into these regions of poor overlap. We show here that our estimator $\hat{\psi}_{ML_t}$, while not reliant on the linearity of $m(x, 0)$, also performs very well in all sample sizes. Furthermore, when the sample size is small and noise level σ_ε is large, the inclusion of regularization in our estimator's implict estimate of $\hat{m}$ helps us — in these settings, $\hat{\psi}_{ML}$ and $\hat{\psi}_{ML_t}$ with larger values of the tuning parameter σ outperform OLS.

[7] While the OLS estimator is not typically considered a weighting estimator, it is linear in Y, and can therefore be expressed in that form. Lemma 3.1 shows that it is, in fact, a limiting ($\sigma \to 0$) case of our estimator $\hat{\psi}_{ML}$ in which we work with the RKHS of linear functions $f(x) = f^T x$ with the Euclidean inner product $\langle f(x), g(x) \rangle = f^T g$.

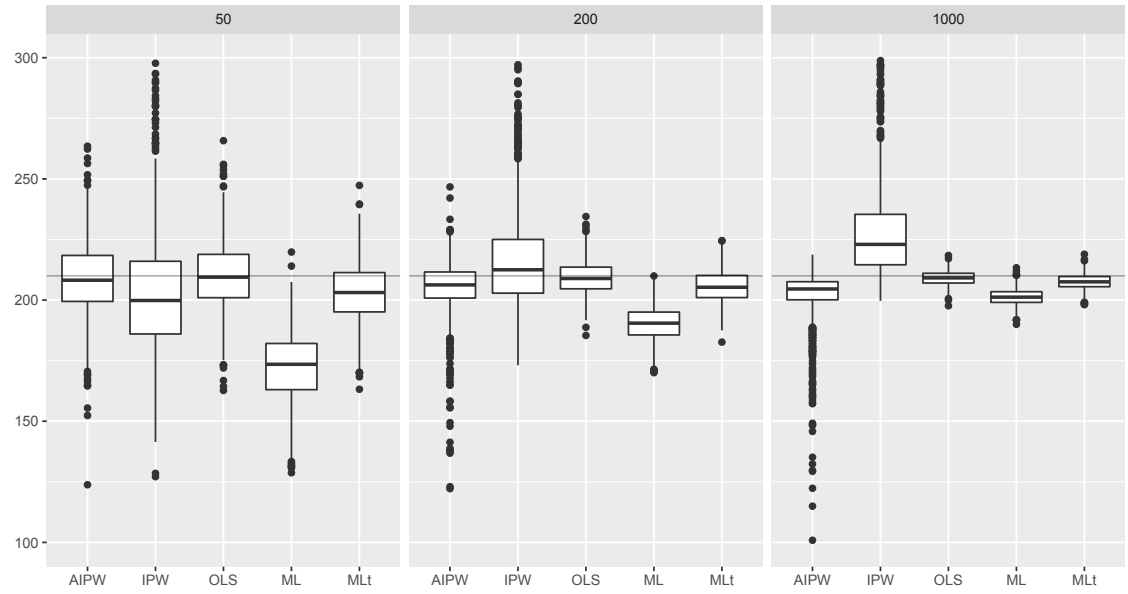

Figure 3.1: Boxplots of our estimates over 1000 replications with $\sigma_\varepsilon = 50$ at sample sizes 50, 200, 1000. The grey horizontal line indicates the value of estimand. As the IPW and AIPW estimators were very variable, some larger estimates are cut off to allow some detail to be visible in the plot.

	n	50	200	1000	4000	50	200	1000	4000
		rmse bias				half-width coverage			
$\sigma_\varepsilon = 1$	IPW	40.8 -3.1	79.3 19	126.5 33.7	169.6 55	17.6 0.53	17.5 0.51	26.7 0.25	51.5 0.08
	AIPW	8.1 -1.4	14.9 -5.4	51.4 -11.9	92 -23.7	15.8 0.96	17 0.98	26.6 0.95	51.5 0.66
	OLS	6.8 0.1	3.3 -0.5	1.7 -0.7	1.2 -0.9	14.8 0.97	8 0.98	3.7 0.98	1.9 0.91
	ML	38.2 -36.7	20 -19.5	8.8 -8.7	4.3 -4.3	16.5 0.01	7.5 0	3.6 0	1.9 0
	MLt	8.9 -6.6	5.2 -4.4	2.6 -2.3	1.5 -1.3	12.9 0.85	7.1 0.83	3.5 0.83	1.9 0.82
	MLt 10σ	9.8 -7.7	6.4 -5.7	3.6 -3.4	2.2 -2.1	12 0.76	6.4 0.6	3.1 0.44	1.7 0.28
	MLt 100σ	11.9 -9.7	10.2 -9.6	8.7 -8.6	7.1 -7.1	11 0.57	5.4 0.1	2.5 0	1.3 0
$\sigma_\varepsilon = 50$	IPW	40.5 -3.3	308 30.8	479.7 59.4	842.3 78.7	33.9 0.75	35.9 0.74	54 0.5	86.7 0.15
	AIPW	15.4 -1.2	69.8 -8.5	143.4 -20.2	513.4 -35.5	32.6 0.95	35.2 0.98	53.8 0.98	86.7 0.9
	OLS	14.1 0.2	6.8 -0.8	3.1 -0.9	1.7 -0.9	34.4 0.97	17.6 0.99	8 0.99	4 0.98
	ML	40.1 -37.5	20.8 -19.6	9.4 -8.8	4.5 -4.2	29.9 0.27	17.2 0.37	8.4 0.47	4.4 0.53
	MLt	14.2 -6.9	8 -4.5	3.9 -2.4	2.1 -1.3	27.9 0.94	17 0.96	8.4 0.97	4.4 0.98
	MLt 10σ	14.3 -8.1	8.4 -5.8	4.4 -3.5	2.5 -2	22.9 0.88	14.1 0.91	7.2 0.9	3.8 0.89
	MLt 100σ	15.8 -10.1	11.5 -9.7	9.1 -8.7	7.1 -7	12.8 0.53	5.9 0.25	3.7 0.03	2.6 0
$\sigma_\varepsilon = 200$	IPW	68.3 -1.3	631.3 39	5869.4 236	442.3 58.7	115.8 0.94	87.3 0.94	159.2 0.86	92.7 0.66
	AIPW	55.6 0.1	55.5 -3.9	687 -41.8	345.4 -27.3	115.2 0.95	85.7 0.98	158 1	92.7 1
	OLS	51.1 1.3	23.3 -0.4	10.4 -0.9	5.1 -0.8	124.5 0.97	62.9 0.99	28.3 1	14.2 1
	ML	53.6 -36.5	29.9 -19.4	14.3 -8.8	6.9 -4.2	101.1 0.93	61.1 0.96	30.8 0.98	16.1 0.99
	MLt	44.7 -6.6	24.2 -4.2	11.7 -2.5	5.7 -1.3	100.5 0.96	61 0.98	30.8 0.99	16.1 0.99
	MLt 10σ	42.8 -7.5	22.2 -5.6	10.5 -3.5	5.4 -2.1	79.4 0.92	49.9 0.96	25.9 0.99	13.8 0.99
	MLt 100σ	42.3 -9.9	22.5 -9.5	12.7 -8.8	8.4 -7.1	29.1 0.49	11.4 0.4	11.2 0.59	9 0.66
$\sigma_\varepsilon = 1000$	IPW	305.6 -0.6	458.6 30.4	627.1 50.9	24863.4 928.5	585.9 0.96	436.6 0.97	352.7 0.97	912.8 0.97
	AIPW	287 -1.5	281.9 2.3	355 -19.9	5577.7 -142.7	584.8 0.94	436 0.99	352.5 1	912.2 1
	OLS	257.4 -3.3	115.8 3.4	50.7 1.5	26.2 -0.4	619.7 0.98	312.9 0.99	140.3 1	70.3 0.99
	ML	199.7 -39.3	111.8 -18.8	56.4 -7.3	29.5 -3.7	497.1 0.99	305.7 0.99	152.6 1	79.9 0.99
	MLt	224.9 -8.4	116.7 -3.3	57.1 -0.8	29.6 -0.8	496.9 0.97	305.6 0.99	152.6 1	79.9 1
	MLt 10σ	211.5 -8.5	105.1 -2	50.1 -1.8	26 -1.4	392.3 0.93	248.9 0.98	128.5 0.99	68.8 0.99
	MLt 100σ	205.3 -9.6	99.3 -5.6	46.4 -6.8	24.6 -7.1	137.8 0.46	51.5 0.37	54.7 0.75	44.5 0.93

Figure 3.2: Root mean squared error (rmse), bias, and confidence interval half-width and coverage over 1000 replications. Here we take the tuning parameter σ to be 0.1 in the estimators ML and MLt. The notation MLt 10σ and 100σ indicates the substitution of 1 and 10 respectively.

3.4 Application: the LaLonde Study

We apply our method to estimate the impact of the National Supported Work (NSW) Demonstration, a labor training program, on post-intervention income levels. In this study, participants were randomly selected for admission to the program, so experimental estimates of a treatment effect are available. As a result, it has been used to test methods for estimation of treatment effects in observational studies. Attempts have been made to use larger nonexperimental control groups to replicate the experimental estimate, but this has proven challenging for many of the methods considered. This problem was famously discussed in LaLonde (1986) and later in Dehejia and Wahba (1999).

We follow Dehejia and Wahba (1999) in working with a subset of the male participants in the experimental sample in which pre-intervention income history is available for at least two years. The latter restriction allows us to adjust for 4 continuous covariates and 4 binary ones: two years of pre-intervention income, age (in years), education (in years of schooling), and indicators for attainment of a high-school diploma, marriage status (married/unmarried), identification as black, and identification as hispanic. The former is in recognition of both substantially different eligibility criteria and realization of the intervention for men and women (see LaLonde, 1986). In this subset, the experimental treatment and control subsamples have 185 and 260 units respectively.

In this context, the primary causal estimand that has been discussed is the average treatment effect on the treated, $\tau_T = \mathbb{E}[Y_i^{(1)} \mid W_i = 1] - \mathbb{E}[Y_i^{(0)} \mid W_i = 1]$. In the experimental sample, randomization ensures that $\mathbb{E}[Y_i^{(0)} \mid W_i = 1] = \mathbb{E}[Y_i^{(0)} \mid W_i = 0] = \mathbb{E}[Y_i \mid W_i = 0]$, and a simple simple difference-in-means estimate $n_T^{-1} \sum_{i:W_i=1} Y_i - n_Z^{-1} \sum_{i:W_i=0} Y_i$ for $n_T = \sum_{i=1}^n 1_{\{W_i=1\}}$ and $n_Z = \sum_{i=1}^n 1_{\{W_i=0\}}$ yields a 95% confidence interval of $\$1794 \pm 1315$. In our attempt to replicate this estimate this using a nonexperimental control group, we observe that under our identification assumptions, $\mathbb{E}[Y_i^{(0)} \mid W_i = 1] = \psi'(m)$ where ψ' is defined as in (3.1) for $T_i = 1_{\{W_i=1\}}$. For the treatment effect τ_T, we use the point estimator $\hat{\tau}_T = n_T^{-1} \sum_{i:W_i=1} Y_i - \hat{\psi}'_{ML}$, taking the parameter σ in (3.8) to be 0.1 and using the Matérn kernel with $\nu = 3/2$ when calculating $\hat{\psi}'_{ML} = (n/n_T)\hat{\psi}_{ML}$. Around it, we give 95% confidence intervals $\hat{\tau}_T \pm z_{.025} \widehat{V}^{1/2} / n_T^{1/2}$ based on the variance estimator

$$
\begin{aligned}
&\widehat{V} = \widehat{V}_1 + \widehat{V}_2 \quad \text{for} \\
&\widehat{V}_1 = n_T^{-1} \sum_{i:W_i=1} Y_i^2 - \left(n_T^{-1} \sum_{i:W_i=1} Y_i \right)^2 ; \\
&\widehat{V}_2 = n_T^{-1} \sum_{i:W_i=1} \left(\hat{m}(X_i) - \hat{\psi}_{ML} \right)^2 + n_T^{-1} \sum_{i:W_i=0} \hat{\gamma}_i^2 \left(Y_i - \hat{m}(X_i) \right)^2 ;
\end{aligned}
$$

in which we use an auxilliary ordinary least squares estimator $\hat{m}$ of $m(X_i, 0)$.

We consider the use of non-experimental control samples constructed by LaLonde from the Population Survey of Income Dynamics (PSID-1) and the Current Population Survey (CPS-1) and well as a small subset of the latter called CPS-3 chosen to have characteristics like the experimental sample. This data is made available with and summarized in (Dehejia and Wahba, 1999). Our point estimates vary substantially depending on the control group used. We estimate 95% confidence intervals of 525 ± 2684, 1233 ± 2733, 1783 ± 1652, and 770 ± 1785 using the additional control units from the CPS-3 sample, the PSID-1 sample, the CPS-1 sample, and the PSID-1 and CPS-1 samples combined. This may be suggestive of a problem, perhaps caused by adjusting for a fairly limited set of covariates, but the standard error of our estimators is sufficiently large that differences between these estimates could simply be explained by random variation. Thus, the experimental data provide little evidence that can be used to validate or invalidate our approach. The same qualitative behavior is observed in the results of Dehejia and Wahba (1999).

Chapter Four

Matching by Rounding

In observational studies, it is common to compare outcomes on *matched* subsamples of our study sample which received different treatments. The role of matching is to select subsamples which are comparable in terms of measured pre-treatment covariates. Insofar as we are able to do this, and these measured covariates include the ones salient to both selection of treatment and outcome under treatment, we can attribute observed differences in outcome to differences in treatment. If, in addition, our matched subsamples are representative of the *target population* on which we hope to compare treatments, we may with some reservation act as if we've observed exactly what we'd like best: a randomized experiment conducted on a sample from our target population.

In this chapter, we focus on matching methods for estimation of the targeted average treatment effect (TATE) for categorical treatments: the average, over our target population, of the difference $Y^{(w)} - Y^{(w')}$ between the outcomes that would have occured under treatments w and w'. As in the previous chapter, we consider an observational study in which we observe for each unit a covariate vector X_i, a categorical treatment status $W_i \in 0 \dots C$, and an outcome $Y_i = Y_i^{(W_i)} \in \mathbb{R}$, and assume that as a function of (X_i, W_i), we can calculate indicators $T_i = T(X_i, W_i)$ that mark units as members of a target subsample. Under the previous chapter's identification assumptions, the TATE is identified as $\mathbb{E}[Y^{(w)} \mid T_i = 1] - \mathbb{E}[Y^{(w')} \mid T_i = 1]$, a difference between two quantities like the estimand we focused on in the previous chapter.

The majority of matching methods in the literature are for estimation of two special cases of the TATE: the average treatment effect (ATE) and the average treatment effect on the treated units (ATT). In the former, the target population is the population from which our study sample is drawn; in the latter, it is the population from which the subsample of treated units is drawn. Additional specialized methods exist for the pairwise comparison of three or more nominal treatments (Lopez and Gutman, 2017), which focus on estimation of the TATE for various subpopulations, defined in terms of received treatment, of the population from which the study sample was drawn. Matching methods are often categorized as 'without replacement', in which individuals are either included in the matched subsample or not, or 'with replacement', in which an individual can appear multiple

times times in the matched subsample. While methods of both types appear in the literature for the ATT and similarly-defined estimands, extant methods for the ATE and its ilk are all, to our knowledge, with replacement. In this paper, we will focus on matching without replacement.

Matching, while favored in many scientific communities for its transparency and its familiarity for those used to randomized experiments, is not known for its statistical efficiency. For some methods, this is merited. Nearest neighbor matching methods, in particular, have been shown to suffer badly from the curse of dimensionality (Abadie and Imbens, 2006). However, approaches have been shown to be $\sqrt{n}$-consistent, namely matching on an estimated propensity score[1] (Abadie and Imbens, 2016) and integer programming methods which optimize for distributional similarity between matched groups (Zubizarreta, 2012; Zubizarreta et al., 2014; Kallus, 2016). And while the former has been established only under parametric assumptions on the propensity score, the latter approaches can be shown to achieve $n^{-1/2}$ rates under fairly weak nonparametric assumptions (Kallus, 2016, Theorem 9).

Randomized rounding offers a simple approach to proving rates for these integer programming methods. In this argument, we think of matching estimators as a subclass of weighting estimators, which may fractionally include individuals in the matched subsamples. More concretely, if we let $A_{i,w}$ and $A_{i,w'}$ be indicators for membership in the matched groups of equal size receiving treatments w and w' respectively, a matched difference in means estimator $n^{-1}\sum_{i:W_i=w} A_{i,w}Y_i - n^{-1}\sum_{i:W_i=w'} A_{i,w'}Y_i$ is simply a weighted difference in means estimator with binary weights satisfying the constraint $\sum_{i:W_i=w} A_{i,w} = \sum_{i:W_i=w'} A_{i,w'}$. To bound the minimum of an objective function over binary weights, we first find a bound on its minimum v over the larger set of non-binary weights, and then exhibit a randomized algorithm that rounds non-binary weights to binary ones in such a way that with nonzero probability, the value v' of the objective function at the rounded weights is close to its value with the weights we round, i.e. $v' \leq v + \epsilon$. As this implies the existence of a binary solution with value no larger than $v+\epsilon$, it follows that $v+\epsilon$ bounds the minimum over binary weights. And insofar as the value of this optimization problem can be used to bound the risk of our estimator, this results in a risk bound.

For example, consider the approach of Kallus (2016) to estimation of the ATT using a matched difference in means $n_1^{-1}\sum_{i:W_i=1} Y_i - n_1^{-1}\sum_{i:W_i=0} \hat{A}_i Y_i$. Kallus uses matched group membership indicators $\hat{A}_i$ solving a constrained variant of the minimax problem (3.4) we considered in the

[1] This result was established for matching with replacement.

previous chapter[2],

$$\min_{\substack{A\in\{0,1\}^n \\ A_i=0 \text{ if } W_i=1 \\ \sum_{i:W_i=0} A_i = n_1}} I_{\mathcal{F}}^2(A) + \frac{\sigma^2}{n_1^2}\|A\|^2 \quad \text{where} \quad I_{\mathcal{F}}(A) = \sup_{f\in\mathcal{F}} \frac{1}{n_1}\sum_{i=1}^{n}[A_i - 1_{\{W_i=1\}}]f(X_i). \tag{4.1}$$

Rounding scaled inverse propensity weights $\alpha\gamma_\psi(X_i)$ into binary weights $\tilde{A}_i$, Kallus bounds the deviation of $I_{\mathcal{F}}(\tilde{A})$ from zero and therefore also that of $I_{\mathcal{F}}^2(\tilde{A}) + \frac{\sigma^2}{n_1^2}\|\tilde{A}\|^2$. As $I_{\mathcal{F}}^2(\hat{A}) + \frac{\sigma^2}{n_1^2}\|\hat{A}\|^2$ will be no larger than this, it follows that the maximal risk of our estimator satisfies the same bound (Kallus, 2016, Proof of Theorem 9).

Here we will use an even simpler rounding argument. Starting with a weighting estimator using non-binary weights, we derive a matching estimator by rounding those weights directly and bound its error using the triangle inequality, as the sum of the error of the weighting estimator and the deviation of the rounded weighting estimator from the weighting estimator it is based on. Considering again the estimation of the ATT using a matched difference in means, given a weighting estimator of the form $n_1^{-1}\sum_{i:W_i=1} Y_i - n_1^{-1}\sum_{i:W_i=0}\hat{\gamma}_i Y_i$, we can derive a matching estimator by substituting for $\hat{\gamma}_i$ matching indicators $\hat{A}_i$ with the property that $n_1^{-1}\sum_{i:W_i=0}\hat{A}_i Y_i \approx n_1^{-1}\sum_{i:W_i=0}\hat{\gamma}_i Y_i$. We propose the use of a rounding method that guarantees that for any possible vector of outcomes y, $n_1^{-1}\sum_{i:W_i=0}\hat{A}_i y_i \approx n_1^{-1}\sum_{i:W_i=0}\hat{\gamma}_i y_i$. A well-known randomized algorithm of Srinivasan (2001) rounds weights $\hat{\gamma}_i \in [0,1]$ into binary weights $\hat{A}_i$ which satisfy the constraints in (4.1) deterministically; are unbiased in the sense that $\mathbb{E}[\hat{A}_i \mid W, X, Y] = \hat{\gamma}_i$; and has a negative dependence property that guarantees that for any bounded vector y, $n_1^{-1}\sum_{i:W_i=0}\hat{A}_i y_i = n_1^{-1}\sum_{i:W_i=0}\hat{\gamma}_i y_i + Z$ where Z is $O_p(n_1^{-1/2})$ and subgaussian conditional on the observed data W, X, Y (Brändén and Jonasson, 2012; Pemantle and Peres, 2014). Thus, when outcomes Y_i are bounded, given any $\sqrt{n}$-consistent weighting estimator with weights $\hat{\gamma}_i \in [0,1]$, our rounding approach results in a $\sqrt{n}$-consistent matching estimator. Furthermore, as the rounding algorithm runs in $O(n)$ time, it does so with essentially no additional computational cost over the weighting method on which it is based.

This approach to estimating the ATT generalizes straightforwardly to estimation of the TATE $\tau_{w,w'} = \mathbb{E}[Y^{(w)} \mid T_i = 1] - \mathbb{E}[Y^{(w')} \mid T_i = 1]$. Suppose we have a weighted difference in means estimator $\hat{\tau}_{w,w'} = n_T^{-1}\sum_{i:W_i=w}\hat{\gamma}_i Y_i - n_T^{-1}\sum_{i:W_i=w'}\hat{\gamma}_i Y_i$ that uses nonnegative weights $\hat{\gamma}_i$ with the normalization property $\sum_{i:W_i=w}\hat{\gamma}_i = \sum_{i:W_i=w'}\hat{\gamma}_i = n_T$ that makes them translation-invariant in the sense of Remark 3.1. Letting $\tilde{\gamma}_i = \hat{\gamma}_i(n_A/n_T)$ for any $n_A \le \lfloor n_T / \max_{\{i:W_i\in w,w'\}}\hat{\gamma}_i \rfloor$, we have $\hat{\tau}_{w,w'} = n_A^{-1}\sum_{i:W_i=w}\tilde{\gamma}_i Y_i - n_A^{-1}\sum_{i:W_i=w'}\tilde{\gamma}_i Y_i$ where each weight $\tilde{\gamma}_i$ is in $[0,1]$. Separately

[2]Here the latter term $(\sigma^2/n_1^2)\|A\|^2$ is equal to the constant σ^2 under our constraint $\sum_{i=1}^n A_i = n_1$, so this reduces to the minimization of $|I_{\mathcal{F}}(A)|$. This can be expressed as a integer linear program with a possibly infinite set of constraints.

rounding the vectors $[\tilde{\gamma}_i : W_i = w]$ and $[\tilde{\gamma}_i : W_i = w']$ using the randomized algorithm of Srinivasan (2001) yields indicators A_i such that $\hat{\tau}^A_{w,w'} = n_A^{-1} \sum_{i:W_i=w} A_i Y_i - n_A^{-1} \sum_{i:W_i=w'} A_i Y_i$ is a matched difference in means ($\sum_{i:W_i=w} A_i = \sum_{i:W_i=w'} A_i = n_A$) and each of its two terms is unbiased for the corresponding term of $\hat{\tau}_{w,w'}$ conditional of X, W, Y and concentrates around it at $O_p(n_A^{-1/2}\|Y\|_\infty)$ rate. Specifically, using a concentration inequality of Pemantle and Peres (2014, Theorem 1), we get the following result.

Theorem 4.1. *For $\hat{\tau}_{w,w'}$ and $\hat{\tau}^A_{w,w'}$ above, $\mathbb{E}[\hat{\tau}^A_{w,w'} \mid X, Y, W] = \hat{\tau}_{w,w'}$ and*

$$\left|\hat{\tau}^A_{w,w'} - \hat{\tau}_{w,w'}\right| \leq \frac{t\|Y\|_\infty}{\sqrt{n_A}} \quad \textit{with probability} \quad 1 - 2e^{-t^2/4} \quad \textit{conditional on the observed data } X, Y, W.$$

We end our discussion with a few remarks.

Remark 4.1. As the difference $\hat{\tau}_{w,w'} - \hat{\tau}^A_{w,w'}$ between our matching estimator and the rounding estimator on which it is based has mean zero conditional on the observed data, by using the former we are essentially just adding noise the the latter. Thus, from the perspective of estimation error only, we should prefer the weighting estimator. The primary advantage of the matching estimator $\hat{\tau}^A_{w,w'}$ is its interpretability — it allows us to interpret our estimator as a comparison between two subsamples of the treatment and control groups that are chosen to be both comparable and representative of our target population. One option is to use weighting to estimate the treatment effect and the rounded version as an aid to the interpretation of that estimator. Using the randomized algorithm of Srinivasan (2001) to sample multiple matched difference-in-means estimators, we get an explicit representation of our weighting estimator as an average of matching estimators, each of which should be reasonable on its own.

Remark 4.2. If we are using a matched difference in means as an estimator and not as an interpretation tool for a weighting estimator, integer programming approaches (Zubizarreta, 2012; Zubizarreta et al., 2014; Kallus, 2016) may be expected to perform better in terms of mean squared error. After all, those estimators choose the 'rounding error' in such a way that a proxy for the estimator's maximal risk is minimized. However, these integer programming approaches also have a few disadvantages.

First, the randomized rounding argument of Kallus described above establishes a rate r_n but does not separately characterize bias and variance, and thus does not justify inference based on concentration of the scaled estimation error $r_n(\hat{\tau} - \tau)$ around zero. In contrast, the bias of our matching estimator $\hat{\tau}^A_{w,w'}$ is equal to that of the weighting estimator $\hat{\tau}_{w,w'}$ on which it is based, and we can estimate its variance by adding to an estimate of $\mathrm{Var}\left[\hat{\tau}_{w,w'}\right]$ a simple sample-based

estimate of the variance of Var $\left[\hat{\tau}^A_{w,w'} - \hat{\tau}_{w,w'}\right]$ which exploits our ability to sample efficiently from the distribution of $\hat{\tau}^A_{w,w'}$ conditional on $\hat{\tau}_{w,w'}$.

Second, in cases in which the regression functions $m^{(w)}(x) = \mathbb{E}[Y \mid X = x, W = w]$ are less smooth than the inverse propensity weights $g^{(w)}_\psi(x) = P\{T_i = 1 \mid X_i = x\}/P\{W_i = w \mid X_i = x\}$, approaches based on estimated inverse propensity weights can perform better than estimators based on design-conditional minimax criteria like the integer programming approaches we've discussed. Our approach allows us to derive matching estimators from estimated inverse propensity weighting estimators as well as from minimax-type weighting estimators.

Third, in large sample settings, solving the integer programs in Zubizarreta (2012); Zubizarreta et al. (2014); Kallus (2016) becomes computationally intractable. Furthermore, expressing nonparametric minimax-type problems like (4.1) exactly requires an infinite number of constraints, and insofar as it is necessary to use a large number of constraints to get a good approximation of the intended problem, solving these problems even in small sample settings can be computationally demanding. As the computational cost of our matching method is essentially that of the weighting method on which it is based, and optimization over continuous-valued weights is often substantially less computationally demanding than integer programming, $\sqrt{n}$-consistent matching estimators in extremely large samples sizes. Our reduction to weighting is particularly computationally advantageous in nonparametric minimax-type problems, where strong duality (see e.g. Lemma 2.5 and Lemma 3.1) can dramatically simplify the computation of the weights.

Remark 4.3. We do not require any particular upper bound on $\hat{\gamma}_i$ for estimation of the TATE, but when we discussed estimation of the ATT above, we required the weights $\hat{\gamma}_i$ to be in $[0, 1]$. This bound was necessary to ensure that we could take the size n_A of our matched group of control units ($W_i = 0$) to be n_1, and thus correspond in size to the subsample of units that receive treatment ($W_i = 1$). This is typical for matching estimators of the ATT. This requires that our control subsample be substantially larger than the treatment sample, and thus tends to hold primarily in case-control studies.

Bibliography

A. Abadie and G. W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

A. Abadie and G. W. Imbens. Matching on the estimated propensity score. *Econometrica*, 84(2): 781–807, 2016.

T. B. Armstrong and M. Kolesár. Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683, 2018.

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

S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*, 2016.

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *The Annals of Statistics*, forthcoming, 2018.

F. Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

A. Belloni, V. Chernozhukov, I. Fernández-Val, and C. Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West. Bayesian methods in pharmacovigilance. *Oxford University Press*, 23:29, 2011.

P. Bickel, C. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, 1998.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

P. Brändén and J. Jonasson. Negative dependence in sampling. *Scandinavian Journal of Statistics*, 39(4):830–838, 2012.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. CRC press, 1984.

T. T. Cai and M. G. Low. A note on nonparametric estimation of linear functionals. *Annals of Statistics*, pages 1140–1153, 2003.

E. Candès and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

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

K. C. G. Chan, S. C. P. Yam, and Z. Zhang. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2015.

J. T. Chang and D. Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317, 1997.

X. Chen, H. Hong, and A. Tarozzi. Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics*, pages 808–843, 2008.

V. Chernozhukov, J. C. Escanciano, H. Ichimura, and W. K. Newey. Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*, 2016.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2017.

V. Chernozhukov, W. Newey, and J. Robins. Double/de-biased machine learning using regularized riesz representers. *arXiv preprint arXiv:1802.08667*, 2018.

J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Nat. Cancer Inst*, 22:173–203, 1959.

F. Cucker and D. X. Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

I. Dahabreh, S. Robertson, E. Stuart, and M. Hernan. Extending inferences from randomized participants to all eligible individuals using trials nested within cohort studies. *arXiv preprint arXiv:1709.04589*, 2017.

A. D'Amour, P. Ding, A. Feller, L. Lei, and J. Sekhon. Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*, 2017.

N. Daysal, M. Trandafir, and R. van Ewijk. Re: A recent study by economists on the impact of home births on infant outcomes confuses the debate on home birth. *BJOG: International Journal of Obstetrics and Gynaecology*, 123:17131714, 2016. doi: 10.1111/1471-0528.14250.

R. H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

P. Ding and T. J. VanderWeele. Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.)*, 27(3):368, 2016.

A. Domahidi, E. Chu, and S. Boyd. ECOS: An SOCP solver for embedded systems. In *European Control Conference (ECC)*, pages 3071–3076, 2013.

D. L. Donoho. Statistical estimation and optimal recovery. *The Annals of Statistics*, pages 238–270, 1994.

D. L. Donoho and R. C. Liu. Geometrizing rates of convergence, III. *The Annals of Statistics*, pages 668–701, 1991.

J. Fan, K. Imai, H. Liu, Y. Ning, and X. Yang. Improving covariate balancing propensity score: A doubly robust and efficient approach. Technical report, Tech. rep., Princeton University, 2016.

M. H. Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

A. Fu, B. Narasimhan, S. Diamond, and J. Miller. *CVXR: Disciplined Convex Optimization*, 2017. URL https://CRAN.R-project.org/package=CVXR. R package version 0.94-4.

E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press, 2015.

B. Graham, C. Pinto, and D. Egel. Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies*, pages 1053–1079, 2012.

B. Graham, C. Pinto, and D. Egel. Efficient estimation of data combination models by the method of auxiliary-to-study tilting (ast). *Journal of Business and Economic Statistics*, pages –, 2016.

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

J. Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

H. Hajaiej, L. Molinet, T. Ozawa, and B. Wang. Sufficient and necessary conditions for the fractional gagliardo-nirenberg inequalities and applications to navier-stokes and generalized boson equations. *arXiv preprint arXiv:1004.4287*, 2010.

M. Hernán and J. Robins. Causal inference book, 2015.

K. Hirano, G. W. Imbens, G. Ridder, and D. B. Rubin. Combining panel data sets with attrition and refreshment samples. *Econometrica*, 69(6):1645–1659, 2001.

K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

D. Hirshberg and S. Wager. Debiased inference of average partial effects in single-index models: A comment on wooldridge and zhu. *preprint*, 2018.

R. A. Horn, R. A. Horn, and C. R. Johnson. *Matrix analysis*. Cambridge university press, 1990.

D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. In *Conference on Learning Theory*, pages 9–1, 2012.

I. A. Ibragimov and R. Z. Khas'minskii. On nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory of Probability & Its Applications*, 29(1):18–32, 1985.

K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

G. Imbens and S. Wager. Optimized regression discontinuity designs. *arXiv preprint arXiv:1705.01677*, 2017.

G. W. Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000.

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

G. W. Imbens, D. B. Rubin, and B. I. Sacerdote. Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *American Economic Review*, 91(4):778–794, 2001.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

I. M. Johnstone. Gaussian estimation: Sequence and wavelet models. *Manuscript*, 2015.

A. B. Juditsky and A. S. Nemirovski. Nonparametric estimation by convex programming. *The Annals of Statistics*, 37(5A):2278–2300, 2009.

N. Kallus. Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*, 2016.

J. D. Kang and J. L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, pages 523–539, 2007.

S. Khan and E. Tamer. Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78(6):2021–2042, 2010.

V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

T. Kühn. Eigenvalues of integral operators with smooth positive definite kernels. *Archiv der Mathematik*, 49(6):525–534, 1987.

T. Kühn, W. Sickel, and T. Ullrich. Approximation numbers of sobolev embeddingssharp constants and tractability. *Journal of Complexity*, 30(2):95–116, 2014.

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pages 604–620, 1986.

S. Lang. *Real and functional analysis*. Springer-Verlag, New York, 1993.

G. Lecué and S. Mendelson. Regularization and the small-ball method ii: complexity dependent error rates. *Journal of Machine Learning Research*, 18(146):1–48, 2017.

G. Lecué, S. Mendelson, et al. Regularization and the small-ball method i: sparse recovery. *The Annals of Statistics*, 46(2):611–641, 2018.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

M. J. Lopez and R. Gutman. Estimation of causal effects with multiple treatments: a review and new ideas. *arXiv preprint arXiv:1701.05132*, 2017.

G. Lugosi and K. Zeger. Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, 41(3):677–687, 1995.

P. Massart. Some applications of concentration inequalities to statistics. In *Annales-Faculte des Sciences Toulouse Mathematiques*, volume 9, pages 245–303. Université Paul Sabatier, 2000.

S. Mendelson. Geometric parameters of kernel machines. In *International Conference on Computational Learning Theory*, pages 29–43. Springer, 2002.

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

M. Mitzenmacher and E. Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge university press, 2005.

R. Mukherjee, W. K. Newey, and J. M. Robins. Semiparametric efficient empirical higher order influence function estimators. *arXiv preprint arXiv:1705.07577*, 2017.

W. K. Newey. Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2):99–135, 1990.

W. K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382, 1994.

W. K. Newey and J. R. Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

J. Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1923. translated into English and edited by D.M. Dabrowska and T.P. Speed (1990).

X. Nie and S. Wager. Learning objectives for treatment effect estimation. *arXiv preprint arXiv:1712.04912*, 2017.

R. I. Oliveira et al. Sums of random hermitian matrices and an inequality by rudelson. *Electron. Commun. Probab*, 15(203-212):26, 2010.

R. Pemantle and Y. Peres. Concentration of lipschitz functionals of determinantal and other strong rayleigh measures. *Combinatorics, Probability and Computing*, 23(1):140–160, 2014.

J. Peypouquet. *Convex Optimization in Normed Spaces: Theory, Methods and Examples*. Springer, 2015.

J. L. Powell, J. H. Stock, and T. M. Stoker. Semiparametric estimation of index coefficients. *Econometrica*, pages 1403–1430, 1989.

J. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(1):122–129, 1995.

J. Robins, M. Sued, Q. Lei-Gomez, and A. Rotnitzky. Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, 22(4):544–559, 2007.

J. Robins, E. T. Tchetgen, L. Li, and A. van der Vaart. Semiparametric minimax rates. *Electronic journal of statistics*, 3:1305, 2009.

J. Robins, L. Li, R. Mukherjee, E. Tchetgen Tchetgen, and A. van der Vaart. Minimax estimation of a functional on a structured high dimensional model. *Annals of Statistics, forthcoming*, 2017.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

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

P. R. Rosenbaum. *Observational Studies*. Springer, 2002.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

P. R. Rosenbaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, 1984.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

D. B. Rubin. On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and drug safety*, 13(12):855–857, 2004.

R. Schaback. The missing wendland functions. *Advances in Computational Mathematics*, 34(1): 67–81, 2011.

D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448): 1096–1120, 1999.

A. Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, pages 1139–1151, 1986.

M. E. Sobel. Causal inference in the social sciences. *Journal of the American Statistical Association*, 95(450):647–651, 2000.

A. Srinivasan. Distributions on level-sets with applications to approximation algorithms. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 588–597. IEEE, 2001.

C. J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, pages 595–620, 1977.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 267–288, 1996.

V. Tikhomirov. ε-entropy and ε-capacity of sets in functional spaces. In *Selected works of AN Kolmogorov*, pages 86–170. Springer, 1993.

A. B. Tsybakov. Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats, 2009.

S. van de Geer. *Empirical Processes in M-Estimation (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press Cambridge, 2000.

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Science & Business Media, 2003.

M. J. van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):1–40, 2006.

A. van der Vaart. On differentiable functionals. *The Annals of Statistics*, pages 178–204, 1991.

A. van der Vaart. Bracketing smooth functions. *Stochastic Processes and their Applications*, 52(1): 93–105, 1994.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 1998.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Univ Pr, 2000.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

Y. Wang and J. R. Zubizarreta. Approximate balancing weights: Characterizations from a shrinkage estimation perspective. *arXiv preprint arXiv:1705.00998*, 2017.

A. Zeileis. Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, 11(10):1–17, 2004. URL http://www.jstatsoft.org/v11/i10/.

Q. Zhao. Covariate balancing propensity score by tailored loss functions. *arXiv preprint arXiv:1601.05890*, 2016.

Q. Zhao, D. S. Small, and B. B. Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *arXiv preprint arXiv:1711.11286*, 2017a.

Q. Zhao, D. S. Small, and A. Ertefaie. Selective inference for effect modification via the lasso. *arXiv preprint arXiv:1705.08020*, 2017b.

W. Zheng and M. J. van der Laan. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pages 459–474. Springer, 2011.

D.-X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

J. R. Zubizarreta. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500):1360–1371, 2012.

J. R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

J. R. Zubizarreta, R. D. Paredes, P. R. Rosenbaum, et al. Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. *The Annals of Applied Statistics*, 8(1):204–231, 2014.

Appendix A

Additional Proofs for Chapter 2

A.1 Asymptotics

In this section, we will examine the asymptotic consequences of Thereom 2.2. Our primary aim will be to prove Theorem 2.4 and Theorem 2.1, but we will discuss the behavior of our estimator in other asymptotic regimes (e.g. $\sigma_n \to \infty$) as well.

A.1.1 Proof of Theorem 2.4

To show that our estimator $\hat{\psi}$ is asymptotically linear (2.26), by our characterization (2.23) it suffices to show that

$$\|\hat{m} - m_n\|_{\tilde{\mathcal{F}}_n} I_{\tilde{\mathcal{F}}_n}(\hat{\gamma}) \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}(\hat{\gamma}_i - \gamma_{\psi_n})(Y_{i,n} - m_n(Z_{i,n}))$$

are $o_{P^n}(n^{-1/2})$.

A.1.1.1 Reduction to Consistency of $\hat{\gamma}$

Under the assumptions of Theorem 2.2, (2.20) and (2.21) imply that these are bounded respectively by

$$\|\hat{m} - m_n\|_{\tilde{\mathcal{F}}_n}\left[u(\mathcal{H}_n, \delta) + 2^{1/2}\|\gamma_{\psi_n}\|^{1/2}_{L_2(P_n)}\sigma_n n^{-1/2}(a \wedge b)^{1/4}\right] \quad \text{and}$$
$$\delta^{-1/2}\|v_n\|_\infty n^{-1/2}(a \wedge b)^{1/2}.$$

In this proof, we will use the simple bound $u(\mathcal{H}_n, \delta) = 2\delta^{-1}R_n(\mathcal{H}_n)$ discussed in a footnote to its definition in (2.18). Thus, when v_n and $\|\gamma_{\psi_n}\|_\infty$ are bounded and $\|\hat{m} - m_n\|_{\tilde{\mathcal{F}}_n} = O_{P^n}(1)$ as assumed, what we have to do is show that $\|\hat{m} - m_n\|_{\tilde{\mathcal{F}}_n} R_n(\mathcal{H}_n)$ is $o_{P^n}(n^{-1/2})$ and that $\sigma_n(a \wedge b)^{1/4}$ and $(a \wedge b)^{1/2}$ are $o(1)$. The first of these rates is guaranteed by assumption (vi), and as we've assumed $\sigma_n = O(1)$, the second follows from the consistency property $a \wedge b \to 0$.

A.1.1.2 Establishing Consistency of $\hat{\gamma}$

Our claim has been reduced to the claim that $a \wedge b \to 0$, i.e. consistency of $\hat{\gamma}$. We will focus on the sufficient condition $a \to 0$ because b tends to be a useful bound only when $\sigma_n \to \infty$. Generally speaking, a is the bound we use to show consist estimation of $\hat{\gamma}_\psi$ without tuning for that purpose, and b is the bound we use to establish rates when we do. a has two relevant terms, $\alpha u(\mathcal{H}^\star, \delta)$ and $\bar{R}$.

Consider $\bar{R}$. Clearly it goes to zero as $\kappa = \kappa(\sigma_n, \delta)$ does. And the approximation condition (iii) of Theorem 2.2 is exactly what is necessary to establish that $\kappa(\sigma_n, \delta) \to 0$ for $\sigma_n = O(1)$. This property has a simple interpretation in terms of the dual problem, discussed in Section 2.1.3.3. We study the dual to establish the convergence to γ_ψ of the function $\hat{g}(\cdot)$ that determines $\hat{\gamma}_i$ in the sense that $\hat{\gamma}_i = \hat{g}(Z_i)$. It is is a penalized least squares problem, and this condition is what is necessary to ensure that in the 'noiseless case' the penalty term term $(\sigma^2/n)\|g(\cdot)\|^2_{\mathcal{F}_L}$ is small enough that it does not prevent convergence to γ_ψ.

We will now address the term $\alpha u(\mathcal{H}^\star, \delta) = 2\delta^{-1}\alpha R_n(\mathcal{H}^\star)$. Our first step will be to show that $R_n(\mathcal{H}_n^\star)$ is $O(n^{-1/2})$. By assumption $R_n(\mathcal{H}_n) = O(n^{-1/2})$; the Rademacher complexity of $\mathcal{H}_n^\star = \mathcal{H}_n - [0,1]\gamma_{\psi_n}$ is bounded by $R_n(\mathcal{H}_n) + R_n([0,1]\gamma_{\psi_n})$ and the latter is equal to the Rademacher complexity of its extreme points $R_n(\{0, \gamma_{\psi_n}\})$, which bounded by $\sqrt{2\log(2)}\,\mathbb{E}\|\gamma_{\psi_n}\|_{L_2(P_n)}$ by Massart's finite class lemma (Massart, 2000, Lemma 5.2). As our $L_1(P^n)$-continuity assumption on ψ_n guarantees $\|\gamma_{\psi_n}\|_\infty = O(1)$, this implies that $R_n(\mathcal{H}_n^\star) = O(n^{-1/2})$. Here we've used well-known properties of Rademacher complexity (see e.g. Bartlett and Mendelson, 2002, Theorem 12). What is left is to show that $\alpha n^{-1/2} = o(1)$ or equivalently that $\alpha = o(n^{1/2})$.

Recalling our dual problem in which we are optimizing over functions g that determine weights $\gamma_i = g(Z_i)$, the role α plays in our proof is the radius of a $\|\cdot\|_{\mathcal{F}_n^\star}$ ball. Outside this ball, we can reject the possibility that a recentered function $\breve{g}' = g' - \gamma_{\psi_n}$ is our recentered estimator $\breve{g} = \hat{g} - \gamma_{\psi_n}$. Insofar as the role of our penalty $(\sigma^2/n)\|\breve{g}'(\cdot)\|^2_{\mathcal{F}_L}$ is to reduce our problem to minimization over this ball, we are not requiring it to have done that well. Optimal tuning typically ensures that this radius α is $O(1)$.

We conclude our proof by bounding the value of α that we actually get. It will be on the order of $n(r_Q \vee r_C)^2 + n^{1/2}\bar{R}$. We've previously shown that $\bar{R} \to 0$, so the latter term is $o(n^{1/2})$ as desired. Furthermore, r_Q and r_Q are proportional to fixed points of localized Rademacher complexity $R_n^\star(1, \mathcal{F}_n^\star(\cdot))$ and $R_n^\star(1, \mathcal{H}_n^\star(\cdot))$. These fixed points are $o(n^{-1/4})$ by our assumption (v), so the former term will also be $o(n^{1/2})$.

A.1.2 Proof of Theorem 2.1

In this section, we prove the generalization of Theorem 2.1 mentioned in Remark 2.4. We prove this theorem by showing that its assumptions imply those of our more general asymptotic theorem, Theorem 2.4. As for efficiency, we make the same assumptions in Theorem 2.1 that we do in our efficiency characterization Proposition 2.3.

Notice that for a Donsker class $\mathcal{F}$, $\mathcal{F}(r) = \mathcal{F} \cap rL_2(P)$ satisfies satisfies $R_n^\star(1, \mathcal{F}(\cdot)) = o(n^{-1/4})$. This follows from the following simple lemma,

Lemma A.1. *Let $\tau_n(r)$ be a sequence of positive functions, each increasing in r, and satisfying $\tau_n(s_n) = o(n^{-1/2})$ for all positive sequences $s_n \to 0$. For any η, there exists a positive sequence r_n satisfying $r_n = o(n^{-1/4})$ and $\tau_n(r_n) \leq \eta r_n^2$.*

Proof. Let $r_n = \sqrt{\tau_n(n^{-1/4})/\eta}$. Then $r_n = o(n^{-1/4})$ and $\tau(r_n) \leq \eta r_n^2 = \tau(n^{-1/4})$ for n sufficiently large that $r_n \leq n^{-1/4}$. If necessary, increase finitely many elements of r_n to ensure that this condition is satisfied for all n. □

Its assumption that $\tau_n(r) = R_n(\mathcal{F}(r))$ satisfies $\tau_n(s_n) = o(n^{-1/2})$ for $s_n \to 0$ is, in this case, the asymptotic equicontinuity of the Rademacher process indexed by a Donsker class (see e.g. Ledoux and Talagrand, 2013, Theorem 14.6).

First we'll choose bounds $\mathcal{F}_n$ and $\mathcal{F}_{L,n}$ in the Theorem 2.4 sense. Theorem 2.1 defines $\tilde{\mathcal{F}}_n = \mathcal{F} \cap \rho_n L_2(P_n)$ for a Donsker class $\mathcal{F}$. For $\mathcal{F}_n$, we take $\mathcal{F}(\rho_n')$ with $\rho_n' = 2^{1/2}(\rho_n \vee n^{-1/4})$, which will contain $\tilde{\mathcal{F}}_n$ with probability going to one (Bartlett et al., 2005, Lemma 3.6). The role of $n^{-1/4}$ here is to ensure that $r = \rho_n'$ is large enough that $\mathcal{F} \cap rL_2(P_n) \subseteq \mathcal{F} \cap 2^{1/2} rL_2(P)$ w.h.p.; for r smaller than some multiple of $R_n^\star(1, \mathcal{F}(\cdot))$ this will not necessarily be the case. Furthermore, for such r we also have $\mathcal{F} \cap rL_2(P_n) \supseteq \mathcal{F} \cap 2^{-1/2} rL_2(P)$ (Bartlett et al., 2005, Corollary 2.2), and thus $\mathcal{F}_{L,n} = (2^{-1/2}\rho_n/\rho_n')\mathcal{F}(\rho_n')$ is a lower bound on $\tilde{\mathcal{F}}_n$. This set $\mathcal{F}_{L,n}$ has the form $r_n \mathcal{F} \cap 2^{-1/2}\rho_n L_2(P)$ where by assumption $\rho_n \gg n^{-1/2}$ and as a consequence $r_n = 2^{-1/2}\rho_n/\rho_n' = 2^{-1/2}(1 \wedge n^{1/4}\rho_n) \gg n^{-1/4}$. Thus there exists a sequence $s_n \ll n^{1/2}$ such that $s_n r_n \to \infty$ and $s_n \rho_n \to \infty$ and therefore $\cup_{n=1}^\infty s_n \mathcal{F}_{L,n} = \operatorname{span} \mathcal{F}$, implying our approximation condition (iii) from Theorem 2.4.

Conditions (i,ii) are satisfied directly by assumption and (iv) follows from the uniform boundedness of $\{h(z, f) : f \in \mathcal{F}\}$ and the boundedness of γ_ψ. To verify (v), it suffices to show that the Donskerity of $\mathcal{F}$ and $\{h(z, \mathcal{F}) : f \in \mathcal{F}\}$ implies the Donskerity of the classes $\mathcal{F}^\star$ and $\mathcal{H}^\star$. $\mathcal{F}^\star$ is contained in the convex hull of the union of two Donsker classes, $\mathcal{F}$ and $-[0,1]\gamma_\psi$; $\mathcal{H}^\star$ is contained in the convex hull of the union of two Donsker classes, $\{h(z, f) : f \in \mathcal{F}\}$, $-[0,1]h(z, \gamma_\psi)$, and the

product of a bounded function γ_ψ and a uniformly bounded Donsker class class $\mathcal{F}^\star$; all of these operations preserve Donskerity each of those operations preserves Donskerity (see e.g. van der Vaart and Wellner, 1996, Chapter 2.10).

Considering (vi), the property $\|\hat{m} - m\| = O_P(1)$ is assumed; the property $R_n(\mathcal{H}_n) = O_P(n^{-1/2})$ follows from Donskerity of the class $\mathcal{H}$, which we established for the superset $\mathcal{H}^\star$ in the previous step; and the property $\|\hat{m} - m\|_{\tilde{\mathcal{F}}_n} R_n(\mathcal{H}_n)$ follows from the tightness and consistency conditions Theorem 2.1. To see this last property, consider separately the cases $\rho_n \to 0$ and $\rho_n \not\to 0$. Consider first the case $\rho_n \to 0$. $\mathcal{H}_n$ lies in a $\|\cdot\|_{L_2(P)}$ ball dictated by the decreasing radius ρ_n' and the modulus of continuity of the functional $f \to h(z,f) - \gamma_\psi(z)f(z) = h(z,f) - \psi(f) + \psi(f) - \gamma_\psi(z)f(z)$. This radius will converge to zero because $h(Z,\cdot) - \psi$ is uniformly continuous by assumption and $f \to \psi(z) - \gamma_\psi(z)f(z)$ is by boundedness of γ_ψ and of the functional $\psi(\cdot)$. Note that we lack the uniform continuity assumption in the original Theorem 2.1, but that $h(Z,\cdot) - \psi = 0$ in that case. And as a consequence of the asymptotic equicontinuity of the Rademacher process indexed by a Donsker class, this implies that $R_n(\mathcal{H}_n) = o_P(n^{-1/2})$. In the case that $\rho_n \not\to 0$, we have $\|\hat{m} - m\|_{\tilde{\mathcal{F}}_n} = o_P(1)$ and its product with $R_n(\mathcal{H}_n) = O_P(n^{-1/2})$ will be $o_P(n^{-1/2})$.

This completes our proof of Theorem 2.1 and the generalization mentioned in Remark 2.4.

A.1.3 Improved Rates: Taking $\sigma_n \to \infty$

By increasing σ_n with sample size, we can improve the rate at which our weights $\hat{\gamma}$ converge to $\hat{\gamma}_\psi$ in $\|\cdot\|_{L_2(P_n)}$. If we are working with the bound (2.20) that we use to control bias in our proof of Theorem 2.4, this is not helpful. In particular, σ_n and our rate of convergence to γ_ψ enter into that bound in the same term, which is on the order of $\sigma n^{-1/2}(a \wedge b)^{1/4}$. And at best, when the bound b is the relevant one and it is dominated by $(\alpha r)^2 \approx (\sigma^{-2} n r^3)^2$, after taking this fourth root our factors of σ cancel. In short, when we do this, we'll want to use a different argument to characterize our estimator. The typical one is the standard argument for doubly robust estimators discussed in Section 2.0.4: by attaining the best rate of convergence to γ_ψ, we make the rate-product bound $\|\hat{m} - m\|_{L_2(P_n)} \|\hat{\gamma} - \gamma_\psi\|_{L_2(P_n)}$ as small as possible.

If this is the approach we want to take, and we are willing to commit to the idea that $\|\gamma_\psi\|_{\mathcal{F}} = O(1)$ for some class $\mathcal{F}$, then the optimal tuning strategy is straightforward. So long as this assumption is valid, if we take $\sigma = n^{1/2} r$ for $r = r_Q(\eta_Q) \vee r_C(\eta_C)$ our bound will be on the order of r. To see this in (2.19), observe that with this tuning, α is constant order and and we use the b bound with the two branches $\bar{R} \approx \sigma^2/n$ and $(\alpha r)^2 \approx r^2$ comparable.

While the general problem of estimating a Riesz representer is somewhat nonstandard, one point

of reference is Example 2.1, the estimation of a mean with outcomes missing at random. In this case, the Riesz representer is the inverse propensity weight $W_i/e(X_i)$. Here $e(x) = \mathbb{E}[W_i \mid X_i = x]$ is the mean of the non-missingness indicator conditional the covariates. And in this example, our functional $h(x, w, m) = m(x, 1)$ is simple enough that $\mathcal{F}$ and $\mathcal{H}$ have comparable local Rademacher complexity, so we can take r to be roughly $R_n^\star(1, \mathcal{F}(\cdot))$. If, for example, $\mathcal{F}$ is a class with empirical metric entropy $\log N(\mathcal{F}; L_2(P_n); \epsilon) = O(\epsilon^{-2\rho})$, then it can be shown that our rate $r = O(n^{-\frac{1}{2(1+\rho)}})$ using a bound of Giné and Koltchinskii (Koltchinskii, 2006, Equation 2.4). In the case of a Hölder smoothness class $C^s(R^d)$, we have $\rho = d/(2s)$ (Tikhomirov, 1993; van der Vaart, 1994) and we recover the well-known minimax rate $r = O(n^{-\frac{s}{2s+d}})$ (Tsybakov, 2009).

A.1.4 Regularity and Efficiency

Our first step is to characterize the tangent space $\mathcal{T}$ to our probability measure P. We show that it is $\{s(y, z) = a(z) + b(y, z) : \mathbb{E}[a(Z)] = 0, \mathbb{E}[b(Y, Z) \mid Z] = 0, \mathbb{E}[Yb(Y, Z) \mid Z] \in \mathcal{M}\}$. Consider a one-dimensional parametric submodel $P_t, t \in [0, \epsilon)$ with score s. We will first show that $s \in \mathcal{T}$.

First we will deal with the technical details necessary to write our submodel in terms of factored densities $p_t(y \mid z)p(z)$ with respect to a common σ-finite measure λ. We will use disintegrations as described in Chang and Pollard (1997), using their notation $p_{t,z}$ for conditional densities rather $p_t(\cdot \mid z)$. It suffices to consider rational t, as the limit defining the score for the submodel converges only if it converges on the rationals. This set of rational-indexed submodels is countable and therefore has a σ-finite dominating measure λ. Under topological assumptions stated in Chang and Pollard (1997, Theorem 1), λ has a disintegration $\{\lambda_z : z \in \mathcal{Z}\}$ and each P_t has a disintegration $\{P_{t,z} : z \in \mathcal{Z}\}$ with $P_{t,z}$ is dominated by λ_z. This allows us to define conditional probability densities, denoted $p_{t,z}$, for almost all z (Chang and Pollard, 1997, Theorem 5 i,v). Doing this for all rational t gives a set of probability densities $p_{t,z}$ with respect to λ_z simultaneously at all rational t for almost all z. It follows that $p_{t,z}(y, z)p_t(z)$ is a density with respect to λ, where $p_t(z)$ is the density of the marginal of P_t on Z with respect to the marginal of λ on Z.

Now the score s will be the derivative at $t = 0$ of $\log p_{t,z}p_t = \log p_{t,z} + \log p_t$ with respect to t. We will call the derivative of the first term s_y and the second s_z. Our submodel must satisfy $\mathbb{E}[Y \mid Z] = m_t(Z)$ for $m_t \in \mathcal{M}$, which we may write $\int y p_{t,z} d\lambda_z = m_t(z)$. Differentiating with respect to t at $t = 0$, we have $\int y \frac{\partial}{\partial t}|_{t=0} p_{t,z} d\lambda_z = \mathbb{E}[Y s_y(Y, Z) \mid Z] = \lim_{t\to 0} t^{-1}(m_t - m) \in \mathcal{M}$. We make no assumptions on the marginal on z, so we have no conditions on s_z other than that it, like all scores, has mean zero. Consequently, our tangent space $\mathcal{T}$ is contained in the proposed set. To show that $\mathcal{T}$ is equal to the proposed set, we exhibit a submodel with every score in the set. As in

Van der Vaart (1998, Example 25.16), we take densities $p_t(y,z) = c(t)k(ts(z))p_0(y,z)$ for scores s in the proposed set where where k is a bounded nonnegative function satisfying $k(0) = k'(0) = 1$, for example $k(x) = 2(1+e^{-2x})^{-1}$, and $c(t)$ is a normalizing constant. Note that because $m' - m \in \mathcal{M}$ for all $m, m' \in \mathcal{M}$, each of these is a valid parametric submodel.

A.1.4.1 The Pathwise Derivative of χ

We will calculate the derivative of our functional $\chi(P)$ with respect to the tangent space discussed above. As before, we will work with paths with factored densities $p_t = p_{t,z}p_t$ with respect to the measure λ. Along a path $p_t(y,z) \in \mathcal{P}$, our derivative may be written

$$\frac{\partial}{\partial t}|_{t=0} \int h(z, m_t) p_t(y,z) d\lambda = \int h(z, m_0) \frac{\partial}{\partial t}|_{t=0} p_t(y,z) d\lambda + \int h(z, \frac{\partial}{\partial t}|_{t=0} m_t) d\lambda$$

The first term is just $\mathbb{E}\, h(Z,m)s(Y,Z) = \mathbb{E}(h(Z,m) - \mathbb{E}\, h(Z,m))s(Y,Z)$; this equality follows from the condition $\mathbb{E}\,(s(Y,Z) \mid Z) = 0$. If $g(Z)$ is a Riesz representer for the functional $\mathbb{E}\, h(Z,\cdot)$ on $\mathcal{M}$, we can write our second term as $\mathbb{E}\, g(Z) \frac{\partial}{\partial t}|_{t=0} m_t$, where

$$\begin{aligned}
\frac{\partial}{\partial t}|_{t=0} m_t &= \frac{\partial}{\partial t}|_{t=0} \frac{\int y p_t(y,z) d\lambda_z}{\int p_t(y,z) d\lambda_z} \\
&= \frac{[\int y \frac{\partial}{\partial t}|_{t=0} p_t(y,z) d\lambda_z][\int p_t(y,z) d\lambda_z] - [\int y p_t(y,z) d\lambda_z][\int \frac{\partial}{\partial t}|_{t=0} p_t(y,z) d\lambda_z]}{[\int p_t(y,z) d\lambda_z]^2} \\
&= E[Y s(Y,Z) \mid Z] - E[Y|Z]E[s(Y,Z)|Z] = E[(Y - m(Z))s(Y,Z) \mid Z].
\end{aligned}$$

That is, we can write our derivative in the form $E[h(Z,m) - \mathbb{E}\, h(Z,m) + g(Z)(Y - m(Z))]s(Y,Z)$.

A.1.4.2 Regularity

Paraphrasing Newey (1990, Theorem 2.2), an asymptotically linear estimator of a functional $\chi(P)$ at P_0 is regular iff its influence function is a Riesz representer for the derivative of that functional χ at P_0 *on a space containing the tangent space.* From our characterization of this derivative above, this happens if the influence function has the form $\iota(y,z) = h(z,m) - \mathbb{E}\, h(Z,m) + \gamma_\psi(z)(y - m(z))$ and γ_ψ is a Riesz representer on a space containing the space $\mathcal{M}$.

A.1.4.3 Efficiency

The projection of the bracketed term onto the closure of the tangent space $\mathcal{T}$ gives the efficient influence function. It follows that the bracketed term is the efficient influence function iff it is in this closure, i.e. if $\mathbb{E}\,[Y g(Z)(Y - m(Z)) \mid Z]$ is in the closure of $\mathcal{M}$. As this conditional expectation is equal to $\mathbb{E}\left[g(Z)(Y - m(Z))^2 \mid Z\right] = g(Z) \operatorname{Var}[Y \mid Z]$, the bracketed term is the efficient influence function if $g(Z) \operatorname{Var}[Y \mid Z] = m(Z)$ for m in the closure of $\mathcal{M}$.

A.2 Additional proofs for lemmas used in Section 2.1

Proof of Lemma 2.5. Because $\ell_{n,\mathcal{F}}$ and $-\mathbb{M}_{n,\mathcal{F}}$ are proper, convex, coercive, continuous functions on reflexive spaces they have minima $\hat{\gamma}$ and $\hat{g}$ respectively. Because in $\ell_{n,\mathcal{F}}$ is strictly convex, its minimum is unique (Peypouquet, 2015, Theorem 2.19, Corollary 2.20).

We transform our primal into an equivalent constrained problem and then, by introducing a Lagrange multiplier, a saddle point problem.

$$\inf_{\gamma\in\mathbb{R}^n} \ell_{n,\mathcal{F}}(\gamma) = \inf\{P_n\gamma_i^2 + t^2 : (\gamma,t)\in\mathbb{R}^n\times\mathbb{R},\ \sup_{f\in\mathcal{F}}(L(f) - P_n\gamma_i f(Z_i)) \le t\} \tag{A.1}$$

$$= \inf_{(\gamma,t)\in\mathbb{R}^n\times\mathbb{R}} \sup_{\lambda\ge 0}\sup_{f\in\mathcal{F}} P_n\gamma_i^2 + t^2 + 2\lambda\left(L(f) - P_n\gamma_i f(Z_i) - t\right). \tag{A.2}$$

Assume we can reorder the the infimum over (γ,t) and the suprema over λ and f in (A.2), so (A.2) is equal to

$$\sup_{f\in\mathcal{F}}\sup_{\lambda\ge 0}\inf_{(\gamma,t)\in\mathbb{R}_n\times\mathbb{R}} P_n\gamma_i^2 + t^2 + 2\lambda\left(L(f) - P_n\gamma_i f(Z_i) - t\right). \tag{A.3}$$

We will simplify this expression. Our first step is to explicitly minimize

$$P_n\gamma_i^2 + t^2 + 2\lambda\left(L(f) - P_n\gamma_i f(Z_i) - t\right)$$

with respect to (γ,t) for fixed (λ,f). The expression is convex and differentiable in (γ,t) and attains its infimum at $\gamma_i = \lambda f(Z_i)$ and $t=\lambda$, which can be seen from the first order optimality conditions

$$0 = \frac{\partial}{\partial\gamma_i} P_n\gamma_i^2 + t^2 + 2\lambda\left(L(f) - P_n\gamma_i f(Z_i) - t\right) = \frac{2}{n}\gamma_i - \frac{2}{n}\lambda f(Z_i),$$
$$0 = \frac{\partial}{\partial t} P_n\gamma_i^2 + t^2 + 2\lambda\left(L(f) - P_n\gamma_i f(Z_i) - t\right) = 2t - 2\lambda$$

Substituting these values shows that (A.3) is equal to

$$\sup_{f\in\mathcal{F}}\sup_{\lambda\ge 0} P_n(\lambda f(Z_i))^2 + \lambda^2 + 2L(\lambda f) - 2P_n(\lambda f(Z_i))^2 - 2\lambda^2$$
$$= \sup_{f\in\mathcal{F}}\sup_{\lambda\ge 0} -\lambda^2 - P_n g(Z_i)^2 + 2L(g) \quad \text{where } g = \lambda f.$$

Reparameterizing in terms of g, the constraint $f\in\mathcal{F}$ is equivalent to $g\in\lambda\mathcal{F}$, and the supremum of the expression above over λ is attained at $\lambda = \inf\{\lambda : g\in\lambda\mathcal{F}\} = \|g\|_{\mathcal{F}}$. Substituting this value of λ results in the expression $\sup_g \mathbb{M}_{n,\mathcal{F}}(g)$, and we've established that this supremum is attained at $\hat{g}$. Retracing our steps, (A.3) is equal to $\mathbb{M}_{n,\mathcal{F}}(\hat{g})$.

We conclude by establishing the equality of (A.2) and (A.3). We begin with the constrained problem (A.1) equivalent to (A.2). This is a finite dimensional convex optimization problem, and the Slater condition holds, i.e., the constraint $\sup_{f\in\mathcal{F}}(L(f) - P_n\gamma_i f(Z_i)) \le t$ is satisfiable with strict

inequality by taking t sufficiently large, so we have strong Lagrange duality (Boyd and Vandenberghe, 2004, Section 5.2.3). That is, the Lagrange multiplier problem (A.2) is equal to its dual

$$\sup_{\lambda \geq 0} \inf_{(\gamma,t) \in \mathbb{R}^n \times \mathbb{R}} \sup_{f \in \mathcal{F}} P_n \gamma_i^2 + t^2 + 2\lambda \left(L(f) - P_n \gamma_i f(Z_i) - t \right)$$

and furthermore there exists $\lambda^\star$ such that is equal to

$$\inf_{(\gamma,t) \in \mathbb{R}^n \times \mathbb{R}} \sup_{f \in \mathcal{F}} P_n \gamma_i^2 + t^2 + 2\lambda^\star \left(L(f) - P_n \gamma_i f(Z_i) - t \right).$$

This saddle point problem is convex and continuous in (γ, t) and concave in f, so the Kneser-Kuhn minimax theorem (Johnstone, 2015, Theorem A.1). implies that if we restrict our infimum to a compact convex set $\mathcal{C}$, reordering the infimum and supremum does not change the value, i.e.

$$\begin{aligned} &\inf_{(\gamma,t) \in \mathcal{C}} \sup_{f \in \mathcal{F}} P_n \gamma_i^2 + t^2 + 2\lambda^\star \left(L(f) - P_n \gamma_i f(Z_i) - t \right) \\ &= \sup_{f \in \mathcal{F}} \inf_{(\gamma,t) \in \mathcal{C}} P_n \gamma_i^2 + t^2 + 2\lambda^\star \left(L(f) - P_n \gamma_i f(Z_i) - t \right). \end{aligned}$$

Our final step in showing equality of (A.2) and (A.3) is to show that the restriction to $\mathcal{C}$ can be dropped on each side of this equality without changing the value, i.e.

$$\begin{aligned} &\inf_{(\gamma,t) \in \mathbb{R}^n \times \mathbb{R}} \sup_{f \in \mathcal{F}} P_n \gamma_i^2 + t^2 + 2\lambda^\star \left(L(f) - P_n \gamma_i f(Z_i) - t \right) \\ &= \inf_{(\gamma,t) \in \mathcal{C}} \sup_{f \in \mathcal{F}} P_n \gamma_i^2 + t^2 + 2\lambda^\star \left(L(f) - P_n \gamma_i f(Z_i) - t \right) \end{aligned} \tag{A.4}$$

and

$$\begin{aligned} &\sup_{f \in \mathcal{F}} \inf_{(\gamma,t) \in \mathcal{C}} P_n \gamma_i^2 + t^2 + 2\lambda^\star \left(L(f) - P_n \gamma_i f(Z_i) - t \right) \\ &= \sup_{f \in \mathcal{F}} \inf_{(\gamma,t) \in \mathbb{R}^n \times \mathbb{R}} P_n \gamma_i^2 + t^2 + 2\lambda^\star \left(L(f) - P_n \gamma_i f(Z_i) - t \right). \end{aligned} \tag{A.5}$$

The first equality (A.4) follows because the function of (γ, t) which takes the value

$$\sup_{f \in \mathcal{F}} P_n \gamma_i^2 + t^2 + 2\lambda^\star \left(L(f) - P_n \gamma_i f(Z_i) - t \right)$$

is proper and coercive, so its infimum must occur on some bounded set $\mathcal{C}'$. The second equality (A.5) follows because taking the unconstrained minimum results in the previously discussed problem (A.3), and we've shown that this problem has a solution $(\gamma^\star, t^\star)$ with $\gamma_i^\star = \hat{g}(Z_i)$, $t^\star = \|\hat{g}\|_{\mathcal{F}}$. Therefore, for any compact convex superset $\mathcal{C}$ of $\mathcal{C}' \cup \{(\gamma^\star, t^\star)\}$, both equalities (A.4) and (A.5) are satisfied. This completes our proof. □

Proof of Lemma 2.7. This is a straightforward calculation based on Bartlett et al. (2005, Theorem 3.3, Part 2). We apply it to the class $\mathcal{F}^2 = \{f^2 : f \in \mathcal{F}\}$ with $T(f^2) = Pf^4 \leq M_{\mathcal{F}}^2 Pf^2$ and

$\psi(r) = 2M_{\mathcal{F}}^3 R_n\{f \in \mathcal{F} : Pf^2 \leq r/M_{\mathcal{F}}^2\}$. This gives the following bound with probability $1 - e^{-x}$ and any $K > 1$.

$$\forall\ f \in \mathcal{F} \quad Pf^2 \leq \frac{K}{K-1} P_n f^2 + \frac{6K}{M_{\mathcal{F}}^2} r^\star + \frac{x(11+5K)M_{\mathcal{F}}^2}{n}$$

where $r^\star$ is a unique fixed point of $\psi(r)$. For this, $\psi(r)$ must be a sub-root function satisfying $\psi(r) \geq M_{\mathcal{F}}^2 R_n\{f^2 \in \mathcal{F}^2 : Pf^4 \leq r\}$. Our choice is sub-root by Bartlett et al. (2005, Lemma 3.4). To see that it is a bound, observe that $\{f^2 \in \mathcal{F}^2 : Pf^4 \leq r\} \subseteq \{f^2 \in \mathcal{F}^2 : Pf^2 \leq r/M_{\mathcal{F}}^2\}$ and $R_n\{f^2 \in \mathcal{F}^2 : Pf^2 \leq r/M_{\mathcal{F}}^2\} \leq 2M_{\mathcal{F}} R_n\{f \in \mathcal{F} : Pf^2 \leq r/M_{\mathcal{F}}^2\}$ by the contraction principle for Rademacher processes (see e.g Bartlett et al., 2005, Theorem A.6), as $\phi(f) = f^2$ is $2M_{\mathcal{F}}$-Lipschitz for $f \in [-M_{\mathcal{F}}, M_{\mathcal{F}}]$. Define $r' = \sqrt{r^\star/M_{\mathcal{F}}^2}$, so the condition $r^\star = \psi(r^\star) = 2M_{\mathcal{F}}^3 R_n\{f \in \mathcal{F} : Pf^2 \leq r^\star/M_{\mathcal{F}}^2\}$ may be written $r'^2/(2M_{\mathcal{F}}) = R_n(\mathcal{F} \cap r' L_2(P))$. In these terms, we may restate our bound in the form

$$\forall\ f \in \mathcal{F} \quad Pf^2 \leq \frac{K}{K-1} P_n f^2 + 6Kr'^2 + \frac{x(11+5K)M_{\mathcal{F}}^2}{n}.$$

Take $x = sKnr'^2/[(11+5K)M_{\mathcal{F}}^2]$ so the last two terms sum to $(6+s)Kr'^2$. We may rearrange our bound as follows.

$$\forall\ f \in \mathcal{F} \quad \frac{P_n f^2}{Pf^2} \geq \frac{K-1}{K} - \frac{(6+s)(K-1)r'^2}{Pf^2}.$$

For $Pf^2 \geq (6+s)K(K-1)r'^2$, this second term is no larger than $1/K$, so we have

$$\forall\ f \in \mathcal{F} \ \text{ with } \ Pf^2 \geq (6+s)K(K-1)r'^2, \quad \frac{P_n f^2}{Pf^2} \geq \frac{K-2}{K}.$$

Letting our lower bound $(K-2)/K = \eta_Q$, we have $K = 2/(1-\eta_Q)$. Therefore,

$$\forall\ f \in \mathcal{F} \ \text{ with } \ Pf^2 \geq b_1(\eta_Q)r'^2, \quad \frac{P_n f^2}{Pf^2} \geq \eta_Q \ \text{ with probability } \ 1 - \exp\Big\{b_2(\eta_Q) n r'^2/M_{\mathcal{F}}^2\Big\}$$

where

$$b_1(\eta_Q) = (6+s)\frac{2}{1-\eta_Q}\left(\frac{2}{1-\eta_Q} - 1\right) = 2(6+s)\frac{1+\eta_Q}{(1-\eta_Q)^2}$$
$$b_2(\eta_Q) = sK/(11+5K) = \frac{\frac{2s}{1-\eta_Q}}{11 + \frac{10}{1-\eta_Q}} = \frac{2s}{21 - 11\eta_Q}.$$

Reparameterizing in terms of $r_Q^2 = b_1(\eta_Q)r'^2$ yields the bound

$$\forall\ f \in \mathcal{F} \ \text{ with } \ Pf^2 \geq r_Q^2, \quad \frac{P_n f^2}{Pf^2} \geq \eta_Q \ \text{ with probability } \ 1 - \exp\left\{\frac{s}{s+6} c_1(\eta_Q) n r_Q^2/M_{\mathcal{F}}^2\right\}$$

where

$$c_1(\eta_Q) = \frac{b_2(\eta_Q)/s}{b_1(\eta_Q)/(s+6)} = \frac{(1-\eta_Q)^2}{(21-11\eta_Q)(1+\eta_Q)}.$$

Taking $s = 6$ gives the claimed bound.

□

Proof of Lemma 2.9. As we care only about the behavior of $\hat{\gamma}$ on an event on which $\tilde{\mathcal{F}} \subseteq \mathcal{F}$, we will give a construction specific to that event. In particular, we will use the implications that $\tilde{\mathcal{F}}$ inherits from $\mathcal{F}$ the properties that it is totally bounded in $\|\cdot\|_\infty$ and that $h(Z_1, \cdot) \dots h(Z_n, \cdot)$ are continuous on the space $(\operatorname{span} \tilde{\mathcal{F}}, \|\cdot\|_{\tilde{\mathcal{F}}})$.

Let $\tilde{\mathcal{F}}_\tau$ be the absolutely convex hull of the centers of a finite internal τ-cover of $\tilde{\mathcal{F}}$ in $\|\cdot\|_\infty$. The space normed by $\|\cdot\|_{\tilde{\mathcal{F}}}$ is finite-dimensional and therefore reflexive (see e.g. Peypouquet, 2015, Theorem 1.24), so we can apply Lemma 2.5 and Lemma 2.8. Letting $\hat{\gamma}_\tau$ be the weights minimizing $\ell_{n,\lambda\mathcal{F}_\tau}(\gamma)$ and $\hat{g}_\tau$ be the corresponding maximizer of $M_{n,\lambda\tilde{\mathcal{F}}_\tau}$, we have $\hat{\gamma}_{i,\tau} = \hat{g}(Z_i)$. We will compare this solution to an approximate maximizer of $M_{n,\lambda\tilde{\mathcal{F}}}$.

Let $\tilde{\mathcal{F}}$, $\hat{g}$, and $\tilde{\gamma}$ satisfy the conditions of Lemma 2.8 for given $\mathcal{F}, \bar{R}$: on an event $\mathcal{A}$ of the stated probability, $\tilde{\mathcal{F}} \subseteq \mathcal{F}$, $R_{n,\lambda\tilde{\mathcal{F}}}(\tilde{\gamma}) < \bar{R}$, and $\mathbb{M}_{n,\lambda\tilde{\mathcal{F}}}(\hat{g}) \geq \mathbb{M}_{n,\lambda\tilde{\mathcal{F}}}(\tilde{\gamma})$. We will show shortly that there exists $\tilde{\gamma}_\tau$ such that $R_{n,\lambda\tilde{\mathcal{F}}_\tau}(\tilde{\gamma}_\tau) < \bar{R}$ on this event. Because in addition $\tilde{\mathcal{F}}_\tau \subseteq \tilde{\mathcal{F}}$, and $\hat{g}_\tau$ minimizes $\mathbb{M}_{n,\lambda\tilde{\mathcal{F}}_\tau}$, Lemma 2.8 applies with the same $\mathcal{F}, \bar{R}$ and therefore the same bound $a \wedge b$ applies to both $P_n(\hat{g} - \gamma_\psi)^2$ and $P_n(\hat{g}_\tau - \gamma_\psi)^2 = P_n(\hat{\gamma}_{i,\tau} - \gamma_\psi)^2$. We will complete our proof by showing that the minimizer $\hat{\gamma}$ of $\ell_{n,\tilde{\mathcal{F}}}$ is arbitrarily close to $\hat{\gamma}_\tau$, so that our bound $a \wedge b$ applies to $P_n(\hat{\gamma}_i - \gamma_\psi)^2$ as claimed.

Before we do that, we will construct $\tilde{\gamma}_\tau$ such that $R_{n,\lambda\tilde{\mathcal{F}}_\tau}(\tilde{\gamma}_\tau) < \bar{R}$ as promised. Recall that $R_{n,\lambda\mathcal{F}}(\tilde{\gamma}) = P_n(\tilde{\gamma} - \gamma_\psi)^2 - 2P_n\check{h}(Z, \tilde{\gamma} - \gamma_\psi) + \|\tilde{\gamma}\|_{\mathcal{F}}^2/\lambda^2$, so

$$\begin{aligned}
&\left|R_{n,\lambda\tilde{\mathcal{F}}_\tau}(\tilde{\gamma}_\tau) - R_{n,\lambda\mathcal{F}}(\tilde{\gamma})\right| \\
&\leq \left|P_n(\tilde{\gamma}_\tau - \gamma_\psi)^2 - P_n(\tilde{\gamma} - \gamma_\psi)^2\right| + 2\left|P_n\check{h}(Z, \tilde{\gamma}_\tau - \tilde{\gamma})\right| + \lambda^{-2}\left(\|\tilde{\gamma}_\tau\|_{\tilde{\mathcal{F}}_\tau}^2 - \|\tilde{\gamma}\|_{\tilde{\mathcal{F}}}\right).
\end{aligned}$$

Letting $\tilde{\gamma}_\tau$ be the center of the ball in a $\|\tilde{\gamma}\|_{\tilde{\mathcal{F}}}$-scaled version of our τ-cover that contains $\|\tilde{\gamma}\|_{\tilde{\mathcal{F}}}$, we have the properties $\|\tilde{\gamma}_\tau\|_{\tilde{\mathcal{F}}_\tau} \leq \|\tilde{\gamma}\|_{\tilde{\mathcal{F}}_\tau}$ and $\|\tilde{\gamma}_\tau - \tilde{\gamma}\|_\infty \leq \tau\|\tilde{\gamma}\|_{\tilde{\mathcal{F}}}$. The first property ensures that the last term in the difference above is zero or negative. The second implies the deterministic bound $\|\tilde{\gamma}_\tau - \tilde{\gamma}\|_\infty \leq \tau\|\tilde{\gamma}\|_{\mathcal{F}}$ on the event $\mathcal{A}$, so we can choose τ such that on this event these functions are arbitrarily close in $\|\cdot\|_\infty$. As the first and second terms of our difference are zero at $\tilde{\gamma}_\tau = \tilde{\gamma}$ and $\|\cdot\|_\infty$ continuous, they go to zero as τ does. Consequently, for sufficiently small τ our strict bound $\bar{R}$ on $R_{n,\lambda\mathcal{F}}(\tilde{\gamma})$ applies to $R_{n,\lambda\tilde{\mathcal{F}}_\tau}(\tilde{\gamma}_\tau)$ as desired.

We'll now complete our proof by showing that the minimizer $\hat{\gamma}$ of $\ell_{n,\tilde{\mathcal{F}}}$ is arbitrarily close to $\hat{\gamma}_\tau$. To do this, we use the $2/n$-strong convexity of $\ell_{n,\lambda\tilde{\mathcal{F}}}$, $P_n(\hat{\gamma}_{\tau,i} - \hat{\gamma}_i)^2 \leq \ell_{n,\lambda\tilde{\mathcal{F}}}(\hat{\gamma}_\tau) - \ell_{n,\lambda\tilde{\mathcal{F}}}(\hat{\gamma})$. In order to get a useful upper bound on the right side in the expression above, we exploit the similarity of

$\ell_{n,\lambda\tilde{\mathcal{F}}}$ and $\ell_{n,\lambda\tilde{\mathcal{F}}_\tau}$, $\ell_{n,\lambda\tilde{\mathcal{F}}_\tau}(\hat{\gamma}_\tau) \leq \ell_{n,\lambda\tilde{\mathcal{F}}}(\hat{\gamma}) \leq \ell_{n,\lambda\tilde{\mathcal{F}}}(\hat{\gamma}_\tau)$ where

$$\ell_{n,\lambda\tilde{\mathcal{F}}}(\hat{\gamma}_\tau) - \ell_{n,\lambda\tilde{\mathcal{F}}_\tau}(\hat{\gamma}_\tau) = \lambda \sup_{f\in\tilde{\mathcal{F}}} \left[P_n h(Z_i, f) - \hat{\gamma}_{\tau,i} f(Z_i)\right]^2 - \lambda \sup_{f'\in\tilde{\mathcal{F}}_\tau} \left[P_n h(Z_i, f') - \hat{\gamma}_{\tau,i} f'(Z_i)\right]^2.$$

Given any sequence f_n in $\tilde{\mathcal{F}}$ along which the first term converges to its supremum, there is a corresponding sequence $f_{n,\tau} \in \tilde{\mathcal{F}}_\tau$ such that the value of $P_n h(Z_i, f) - \hat{\gamma}_{\tau,i} f(Z_i)$ at $f = f_n$ and $f = f_{n,\tau}$ can be made arbitrarily close by choice of τ, and consequently this difference shinks to zero with τ. This completes our proof. □

Proof of Lemma 2.10. We will be bounding

$$R_{n,\,\lambda\mathcal{F}_L}(\tilde{\gamma}) = P_n(\tilde{\gamma} - \gamma)^2 - 2P_n\check{h}(Z, \tilde{\gamma} - \gamma_\psi) + \|\tilde{\gamma}\|^2_{\mathcal{F}_L}/\lambda^2.$$

Consider first the middle term. By Chebyshev's inequality, with probability greater than $1-\delta/2$,

$$\left|P_n\check{h}(Z, \tilde{\gamma} - \gamma)\right| < 2^{1/2}\delta^{-1/2}n^{-1/2} \operatorname{Var}\left[\check{h}(Z, \tilde{\gamma} - \gamma_\psi)\right]^{1/2}.$$

We can bound $\operatorname{Var}\left[\check{h}(Z, \tilde{\gamma} - \gamma_\psi)\right]^{1/2}$ by invoking some uniformity,

$$\operatorname{Var}\left[\check{h}(Z, \tilde{\gamma} - \gamma_\psi)\right]^{1/2} \leq \|\tilde{\gamma} - \gamma_\psi\|_{\mathcal{F}^\star(r)} \sup_{h\in\mathcal{H}^\star(r)} \operatorname{Var}\left[h(Z)\right]^{1/2} \quad \text{where } r = \|\tilde{\gamma} - \gamma_\psi\|_{L_2(P)}.$$

Therefore, with probability at least $1 - \delta/2$, our middle term is less than

$$2^{3/2}\delta^{-1/2}n^{-1/2}\|\tilde{\gamma} - \gamma_\psi\|_{\mathcal{F}^\star(r)}\overline{\sigma}(\mathcal{H}^\star(r))$$

Now consider the first term. By Markov's inequality, with probability greater than $1 - \delta/2$,

$$P_n(\tilde{\gamma} - \gamma_\psi)^2 < 2\delta^{-1}P(\tilde{\gamma} - \gamma_\psi)^2.$$

Then by the union bound, with probability $1 - \delta$, we have

$$R_{n,\,\lambda\mathcal{F}_L}(\tilde{\gamma}) \leq 2\delta^{-1}[P(\tilde{\gamma} - \gamma_\psi)^2 + \delta\|\tilde{\gamma}\|^2_{\mathcal{F}_L}/(2\lambda^2)] + 2^{3/2}\delta^{-1/2}n^{-1/2}\|\tilde{\gamma} - \gamma_\psi\|_{\mathcal{F}^\star(r)}\overline{\sigma}(\mathcal{H}^\star(r)). \quad \text{(A.6)}$$

We will call the bracketed term in this bound κ^2 and bound the remaining term in terms of it, using the obvious properties that $r = \|\tilde{\gamma} - \gamma_\psi\|_{L_2(P)} \leq \kappa$ and $\|\tilde{\gamma}\|_{\mathcal{F}_L} \leq 2^{1/2}\delta^{-1/2}\lambda\kappa$. Recalling our discussion of the relationship of $\|\tilde{\gamma} - \gamma_\psi\|_{\mathcal{F}^\star}$ to $\|\tilde{\gamma}\|_{\mathcal{F}}$ in Section 2.1.3.3, we have $\|\tilde{\gamma} - \gamma_\psi\|_{\mathcal{F}^\star} \leq \|\tilde{\gamma}\|_{\mathcal{F}} \vee 1 \leq \|\tilde{\gamma}\|_{\mathcal{F}_L} + 1$, and it follows that $\|\tilde{\gamma} - \gamma_\psi\|_{\mathcal{F}^\star(r)} \leq 2^{1/2}\delta^{-1/2}\lambda\kappa + 1$.

Substituting this into our bound (A.6), we see that with probability $1 - \delta$,

$$\begin{aligned} R_{n,\,\lambda\mathcal{F}_L}(\tilde{\gamma}) &\leq 2\delta^{-1}\kappa^2 + 2^{3/2}\delta^{-1/2}n^{-1/2}[2^{1/2}\delta^{-1/2}\lambda\kappa + 1]\overline{\sigma}(\mathcal{H}^\star(\kappa)) \\ &= 2\delta^{-1}[\kappa^2 + 2\lambda n^{-1/2}\kappa\overline{\sigma}(\mathcal{H}^\star(\kappa))] + 2^{3/2}\delta^{-1/2}n^{-1/2}\overline{\sigma}(\mathcal{H}^\star(\kappa)). \end{aligned}$$

□

Appendix B

Additional Proofs for Chapter 3

B.1 Constants used in the statement of Theorem 3.4

B.1.0.1 Constants used in Equation 3.13

$$C = \sqrt{8\|g_\psi\|_\infty^{1-1/\alpha}(p_T C_{\lambda,T})^{1/\alpha}\Big/(1-1/\alpha)}\,;$$
$$C_\zeta = C_\phi^2 C_{\lambda,Z}^{1/\alpha}\left[\left(\frac{\beta}{1-\beta}\right)^{1/\alpha+\beta} + \frac{\alpha}{(1+\alpha\beta)(\alpha-(1+\alpha\beta))}\right].$$

B.1.0.2 Constants used in Equation 3.14

$$c_1(\eta_Q) = \frac{(1-\eta_Q)^2}{2(1+\eta_Q)(21-11\eta_Q)} \approx .02;$$
$$\eta_Q = (61-8\sqrt{39})/49 \approx .23$$
$$\eta_C = \left(C_{1,C} + C_{2,C}n^{-\frac{1}{2(\alpha+1)}} + C_{3,C}n^{-\frac{1}{\alpha+1}}\right)\Big/(7C_Q)^{1+1/\alpha}\,;$$
$$M_{\mathcal{F}^\star} = M_K + \|g_\psi\|_\infty;$$
$$M_{\mathcal{H}^\star} = \|g_\psi\|_\infty(M_K + \|g_\psi\|_\infty);$$
$$C_{u,1} = (1+\eta)2^{3/2}p_Z^{1/2}\max\{1,\|g_\psi\|_\infty\}\left(\|g_\psi\|_{L_2(P_Z)} + C_{\lambda,Z}^{1/2}(1-\alpha)^{-1/2}\right);$$
$$+ 2^{1/2}p_Z^{1/2}\max\{1,\|g_\psi\|_\infty\}\left(\lambda_{1,Z} + \|g_\psi\|_{L_2(P_Z)}\right)\sqrt{\log(2\delta^{-1})};$$
$$C_{u,2} = 2M_{\mathcal{H}^\star}(1/3+1/\eta)\log(2\delta^{-1});$$
$$C_Q = \{12p_Z^2M_{\mathcal{F}^\star}^2[1+(p_Z C_{\lambda,Z})^{1/\alpha}(1-1/\alpha)^{-1}])\}^{1/[2(1+1/\alpha)]};$$
$$C_{1,C} = 2(1+\eta)\|g_\psi\|_\infty\{3[1+(p_Z C_{\lambda,Z})^{1/\alpha}(1-1/\alpha)^{-1}]\}^{1/2};$$
$$C_{2,C} = \max\{1,\|g_\psi\|_\infty\}\sqrt{2\log(2\delta^{-1})};$$
$$C_{3,C} = 2M_{\mathcal{H}^\star}\left(\frac{1}{3}+\frac{1}{\eta}\right)\log(2\delta^{-1});$$
$$C_{1,R} = 2\left(\delta^{-1}p_Z\right)^{\frac{1-2\kappa_\gamma}{1+2\kappa_\gamma}}\|g_\psi\|_{L_2(P_Z)}^{\frac{4}{1+2\kappa_\gamma}};\left(\theta^{-\frac{\theta}{1+\theta}} + \theta^{\frac{1}{1+\theta}}\right);$$
$$C_{2,R} = 2^{\frac{4(1-\kappa_\gamma)}{1-2\kappa_\gamma}}\left(\delta^{-1}p_Z\right)^{\frac{1}{1+2\kappa_\gamma}}\|g_\psi\|_\infty\|g_\psi\|_{L_2(P_Z)}^{\frac{4}{1-4\kappa_\gamma^2}}\theta^{\frac{\theta}{2(\theta+1)}} \quad \text{for} \quad \theta = 4\kappa_\gamma/(1-2\kappa_\gamma).$$

B.2 Smoothness, Eigenvalues, and Eigenfunctions

The conditions of Lemma 3.5 are defined in terms of the Hölder norm $\|\cdot\|_{C^{2s}}$ and the Sobolev norm $\|\cdot\|_{H^s}$. We define these norms, then prove the lemma.

$$\|f\|_{C^s} = \sum_{\beta\in\mathbb{N}^d:\|\beta\|_1\le\lfloor s\rfloor} \|D^\beta f\|_\infty + \sum_{\beta\in\mathbb{N}^d:\|\beta\|_1=\lfloor s\rfloor} \sup_{x,x'\in\mathbb{R}^d} \frac{|D^\beta f(x) - D^\beta f(x')|}{\|x-x'\|_{\ell_2}^{s-\lfloor s\rfloor}};$$

$$\|f\|_{H^s} = \sum_{\beta\in\mathbb{N}^d:\|\beta\|_1\le\lfloor s\rfloor} \|D^\beta f\|_{L_2(\mu)} + \sum_{\beta\in\mathbb{N}^d:\|\beta\|_1=\lfloor s\rfloor} \left[\int \frac{|D^\beta f(x) - D^\beta f(x')|^2}{\|x-x'\|_{\ell_2}^{2(s-\lfloor s\rfloor)+d}} d\mu(x)d\mu(x')\right]^{1/2};$$

$$D^\beta f = \frac{\partial^{\beta_1}}{\partial x_1}\cdots\frac{\partial^{\beta_d}}{\partial x_d} f.$$

Here μ is Lebesgue measure.

Proof of Lemma 3.5. In this proof, we will use the Gagliardo Nirenberg inequality (Hajaiej et al., 2010, Theorem 1.2), in particular the bounds

$$\|f\|_\infty \lesssim \|f\|_{C^s}^{1-\theta}\|f\|_{L_2(\mu)}^{\theta}, \qquad \theta = 1 - d/(2s+d);$$
$$\|f\|_\infty \lesssim \|f\|_{H^s_\mu}^{1-\theta}\|f\|_{L_2(\mu)}^{\theta}, \qquad \theta = 1 - d/(2s).$$

In case (ii), Kühn (1987, Theorem 4) established the claimed eigenvalue bound with $\alpha = 2s/d + 1$ under a condition $\sup_{x\in\mathcal{X}}\|K(x,\cdot)\|_{C^{2s}} < \infty$ weaker than our assumption on K. In addition, Cucker and Zhou (2007, Theorem 5.5) established the bound $\sup_{f:\|f\|_{\mathcal{H}_K}\le 1}\|f\|_{C^s} < \infty$. This, in combination with the Gagliardo-Nirenberg inequality, imply that $\|f\|_\infty \lesssim \|f\|_{\mathcal{H}_K}^{d/(2s+d)}\|f\|_{L_2(\mu)}^{1-d/(2s+d)}$. As $\|\phi_j\|_{\mathcal{H}_K} = \lambda_j^{-1/2}$ and $\|\phi\|_{L_2(\mu)} \le \eta^{-1}\|\phi\|_{L_2(\nu)} = \eta^{-1}$, we have $\|\phi_j\|_\infty \lesssim \lambda_j^{-d/[2(2s+d)]}$ as claimed.

In case (i), we use the bound $a_j(\mathcal{B}_{H^s_\mu}) \lesssim j^{-s/d}$ (see e.g. Kühn et al., 2014) where

$$a_j(\mathcal{F}) = \inf_{\{\text{rank } A<j\}} \sup_{\{f\in\mathcal{F}\}} \|f - Af\|_{L_2(\mu)} \quad \text{and} \quad \mathcal{B}_{H^s_\mu} = \{f : \|f\|_{\mathcal{H}^s_\mu} \le 1\}.$$

Observe that a_j has the homogeneity property $a_j(s\mathcal{F}) = sa_j(\mathcal{F})$ and the increasingess property $A \subseteq B \implies a_j(A) \le a_j(B)$. As our assumption $\sup_{\|f\|\le 1}\|f\|_{H^s_\mu} < \infty$ implies that the unit ball $\mathcal{B}_K$ of our RKHS is in $s\mathcal{B}_{H^s_\mu}$ for some s, we have $a_j(\mathcal{B}_K) \lesssim j^{-s/d}$ as well. This is helpful because $a_j(\mathcal{B}_K) = \lambda_j^{1/2}$ where λ_j is the jth eigenfunction of the integral operator $L_{K,\mu}$. To see this, observe that if the range of A does not contain the span of the first $j-1$ eigenfunctions $\phi_1\ldots\phi_{j-1}$, there is a function $f = \sum_{k=1}^{j-1} f_k\lambda_k^{1/2}\phi_j$ in $\mathcal{B}_K$ with $\|f - Af\|_{L_2(\mu)} = \|f\|_{L_2(\mu)} = [\sum_{k=1}^{n-1} f_k^2\lambda_k]^{1/2} \ge \lambda_j^{1/2}$, whereas if it is the identity restricted to that span, we have $\|f - Af\|_{L_2(\nu)} \le \lambda_j^{1/2}$ whenever $f \in \mathcal{B}_K$. Thus, our saddle point is attained with A equal to this restricted identity and $f = \phi_j$. This implies

that $\lambda_j \lesssim j^{-2s/d}$, and as discussed in our review of RKHSes, strong equivalence of μ and ν implies the same rate for the eigenvalues $\lambda_{j,\nu}$ of $L_{K,\nu}$.

To bound the eigenfunctions, recall that $\|f\|_{H^s_\mu} \lesssim \|f\|_{\mathcal{H}_K}$. This and the Gagliardo-Nirenberg inequality imply the bound $\|f\|_\infty \lesssim \|f\|_{\mathcal{H}_K}^{d/(2s)} \|f\|_{L_2(\mu)}^{1-d/(2s)}$. As $\|\phi_j\|_{\mathcal{H}_K} = \lambda_j^{-1/2}$ and $\|\phi\|_{L_2(\mu)} \leq \eta^{-1}\|\phi\|_{L_2(\nu)} = \eta^{-1}$, we have $\|\phi_j\|_\infty \lesssim \lambda_j^{-d/(4s)}$ as claimed. □

B.3 Asymptotics

Proof of Theorem 3.2. Lemma 3.5 implies that the eigenvalue and eigenfunction bounds assumed in Theorem 3.4 are satisfied with $\alpha = 2s/d > 1$ and $\alpha(1-\beta) = 2s/d - 1 > 1$ under Assumption 3.3 or $\alpha = (2s+d)/d > 1$ and $\alpha(1-\beta) = (2s+d)/d - 1 > 1$ under Assumption 3.4.

Consider the bound (3.13) on the bias term and assume for a moment that ζ^{-1} is bounded. We will characterize the order of the leading terms T_1 and T_2 in our bound for a range of κ_m, as it is clear that third and fourth terms are irrelevant.

$$T_1 + T_2 \sim \begin{cases} \lambda^{\kappa_m} + n^{-1/2}\lambda^{\kappa_m - 1/(2\alpha)} & \kappa_m \in [1/2, 1) \\ \lambda + n^{-1/2}\lambda^{1+(\kappa_m - 3/2)/\alpha} & \kappa_m \in [1, 3/2) \\ \lambda + n^{-1/2}\lambda & \kappa_m \in [3/2, \infty). \end{cases}$$

$T_2 = o(n^{-1/2})$ irrespective of κ_m as long as $\lambda \ll 1$, whereas the dominant term T_1 will be $o(n^{-1/2})$ iff $\lambda \ll n^{-1/(2\min\{\kappa_m, 1\})}$. This will be the rate at which the bias term goes to zero if ζ^{-1} is bounded, which occurs if $\lambda^{-(1/\alpha+\beta)}\log(n)/n \to 0$ and equivalently if $[\log(n)/n]^{1/(1/\alpha+\beta)} \ll \lambda$. In summary, our bias term is $o_p(n^{-1/2})$ iff $[\log(n)/n]^{1/(1/\alpha+\beta)} \ll \lambda \ll n^{-1/(2\min\{\kappa_m,1\})}$.

Now consider the bound (3.14) on the deviation of our noise term from our desired asymptotic characterization. This bound will be negligible if the factor a goes to zero. This will happen if (i) $\bar{R} \to 0$, (ii) $\alpha \ll \sqrt{n}$ and therefore the first term in a goes to zero. Referring to the first claim of Lemma 3.12, we have (i) given $\lambda \to 0$ and our assumption that $\mathcal{H}_K$ is dense in $L_2(P_Z)$. Unpacking (ii), we are assured that the second term in α is $o(\sqrt{n})$ given (i) if $\lambda \gtrsim n^{-1}$ and the first term in α is $o(\sqrt{n})$ if $\lambda^{-1}n^{-1/(1+1/\alpha)}$ is or equivalently $n^{-(1/2+1/(1+1/\alpha))} \ll \lambda$.

Collecting all of our conditions, we have efficiency if $\mathcal{H}_K$ is dense, $n^{-1} \lesssim \lambda$, and

$$\max\{[\log(n)/n]^{1/(1/\alpha+\beta)},\ n^{-(1/2+1/(1+1/\alpha))}\} \ll \lambda \ll n^{-1/(2\min\{\kappa_m,1\})}.$$

The condition $n^{-1} \lesssim \lambda$ implies all of our lower bounds, as our condition $\alpha(1-\beta) > 1$ is equivalent to $1/\alpha + \beta < 1$ and therefore $[\log(n)/n]^{1/(1/\alpha+\beta)} \ll n^{-1}$ and our assumption $\alpha > 1$ implies that

$n^{-(1/2+1/(1+1/\alpha))} \ll n^{-1}$. Thus, if $\mathcal{H}_K$ is dense, we have efficiency for λ satisfying $n^{-1} \lesssim \lambda \ll n^{-1/(2\min\{\kappa_m,1\})}$ and, in particular, for $\lambda = \sigma^2/n$. □

B.4 Proofs for lemmas used in Section 3.2

Here we collect proofs for all the lemmas and propositions stated in the section.

Proof of Lemma 3.1. To simplify our notation, we'll use Z_i as a shorthand for $1_{\{W_i=0\}}$. Our weighting problem (3.8) is

$$\begin{aligned}
&\frac{\sigma^2}{n^2}\sum_{i:Z_i=1}\gamma_i^2 + \sup_{f:\|f\|\le 1}\left[\frac{1}{n}\sum_i (T_i - Z_i\gamma_i)\langle K_{X_i}, f\rangle\right]^2 \\
&= \frac{\sigma^2}{n^2}\sum_{i:Z_i=1}\gamma_i^2 + \left\langle \frac{1}{n}\sum_i (T_i - Z_i\gamma_i)K_{X_i}, \frac{1}{n}\sum_j (T_j - Z_j\gamma_j)K_{X_j}\right\rangle \\
=&\frac{\sigma^2}{n^2}\sum_{i:Z_i=1}\gamma_i^2 + \frac{1}{n^2}\sum_{i,j}(T_i - Z_i\gamma_i)(T_j - Z_j\gamma_j)K(X_i, X_j) \\
=&\frac{1}{n^2}\left[\sigma^2\gamma^T\gamma + 1^T K_{T,T}1 - 2\gamma^T K_{Z,T}1 + \gamma^T K_{Z,Z}\gamma\right] \\
=&\frac{1}{n^2}\left[1^T K_{T,T}1 - 2\gamma^T K_{Z,T}1 + \gamma^T\left(K_{Z,Z} + \sigma^2 I\right)\gamma\right]
\end{aligned}$$

where K is the Gram matrix ($K_{i,j} = K(X_i, X_j)$), 1 is a vector of $|\{i : T_i = 1\}|$ ones, and subscripting by Z or T takes the rows of columns corresponding to units in those groups. At the minimum over γ, the derivative with respect to γ will be zero, so our weights solve $(K_{Z,Z} + \sigma^2 I)\gamma = K_{Z,T}1$, and the weighted average of treatment outcomes is

$$n^{-1}\sum_{i=1}^n Z_i\hat{\gamma}_i Y_i = n^{-1}Y_Z^T\hat{\gamma} = n^{-1}Y_Z^T(K_{Z,Z} + \sigma^2 I)^{-1}K_{Z,T}1. \tag{B.1}$$

Now consider ridge regression on the treated units. We estimate $\hat{m}$ solving

$$\min_f \sum_{i:Z_i=1}(Y_i - \langle K_{X_i}, f\rangle)^2 + \sigma^2\|f\|^2.$$

We can write it equivalently in constrained form,

$$\min_{r,f}\sum_{i:Z_i=1} r_i^2 + \sigma^2\|f\|^2 \quad \text{where} \quad r_i = \langle K_{X_i}, f\rangle - Y_i.$$

This problem is solved by a saddle point of the Lagrangian (Peypouquet, 2015, Theorem 3.6.8),

$$L((r,f),\lambda) = \sum_{i:Z_i=1} r_i^2 + \sigma^2\|f\|^2 + 2\sum_{i:Z_i=1}\lambda_i\left(\langle K_{X_i}, f\rangle - Y_i - r_i\right).$$

For given λ, we can minimize over (r, f) explicitly, solving the conditions $r_i - \lambda_i = 0$ and $\sigma^2 f + \sum_{i:Z_i=1} \lambda_i K_{X_i} = 0$ that arise from setting the derivatives with respect to r_i and f to zero. Substituting the optimal values $\hat{r}_i = \lambda_i$ and $\hat{f} = -\sigma^{-2} \sum_{i:Z_i=1} \lambda_i K_{X_i}$,

$$
\begin{aligned}
L((\hat{r}, \hat{f}), \lambda)) &= \sum_{i \in Z} \lambda_i^2 + \sigma^{-2} \left\langle \sum_{i:Z_i=1} \lambda_i K X_i, \sum_{j:Z_j=1} \lambda_j K_{X_j} \right\rangle + 2 \sum_{i:Z_i=1} \lambda_i \left(-\sigma^{-2} \left\langle \sum_{j:Z_j=1} \lambda_j K_{X_j}, K_{X_i} \right\rangle - Y_i - \lambda_i \right) \\
&= \sum_{i:Z_i=1} \lambda_i^2 + \sigma^{-2} \sum_{i,j:Z_i=Z_j=1} \lambda_i \lambda_j K(X_i, X_j) - 2\sigma^{-2} \sum_{i,j:Z_i=Z_j=1} \lambda_i \lambda_j K(X_j, X_i) - 2 \sum_{i:Z_i=1} \left(\lambda_i Y_i + \lambda_i^2 \right) \\
&= -\lambda^T \lambda - 2\lambda^T Y_Z - \sigma^{-2} \lambda^T K_{Z,Z} \lambda \\
&= -2\lambda^T Y_Z - \lambda^T \left(\sigma^{-2} K_{Z,Z} + I \right) \lambda.
\end{aligned}
$$

This is maximized at $\hat{\lambda} = -\left(\sigma^{-2} K_{Z,Z} + I \right)^{-1} Y_Z = -\sigma^2 \left(K_{Z,Z} + \sigma^2 I \right)^{-1} Y_Z$. Thus, we have a saddle at $((\hat{r}, \hat{f}), \hat{\lambda})$ and the function $\hat{m}$ solving our problem is $\hat{f}$. Substituting in $\hat{\lambda}$ into our expression for $\hat{f}$ above,

$$
\begin{aligned}
\langle K_x, \hat{f} \rangle &= \left\langle -\sigma^{-2} \sum_{i:Z_i=1} \left[-\sigma^2 \left(K_{Z,Z} + \sigma^2 I \right)^{-1}_{i,Z} Y_Z \right] K_{X_i}, K_x \right\rangle \\
&= \sum_{i:Z_i=1} Y_Z^T \left(K_{Z,Z} + \sigma^2 I \right)^{-1}_{Z,i} K(X_i, x).
\end{aligned}
$$

Therefore our ridge regression prediction $\hat{f}$, averaged over our target sample, is

$$
\begin{aligned}
\left\langle n^{-1} \sum_{j:T_j=1} K_{X_j}, \hat{f} \right\rangle &= n^{-1} \sum_{j:T_j=1} \sum_{i:Z_i=1} Y_Z^T \left(K_{Z,Z} + \sigma^2 I \right)^{-1}_{Z,i} K(X_i, X_j). \\
&= n^{-1} Y_Z^T \left(K_{Z,Z} + \sigma^2 I \right)^{-1} K_{Z,T} 1.
\end{aligned}
$$

This is the weighted average of treatment outcomes using our minimax weights, completing our proof. □

Proof of Lemma 3.7. Expanding K_x in the orthonormal basis $(\lambda_j^{1/2} \phi_j)_{j \in \mathbb{N}}$ of $\mathcal{H}_K$, we have

$$
K_x = \sum_j \langle K_x, \lambda_j^{1/2} \phi_j \rangle \lambda_j^{1/2} \phi_j = \sum_j \lambda_j^{1/2} \phi_j(x) \lambda_j^{1/2} \phi_j
$$

and consequently

$$
\left\| [L_{K,\nu} + \lambda I]^{-1/2} K_x \right\|^2 = \left\| \sum_j \frac{\lambda_j^{1/2} \phi_j(x)}{(\lambda_j + \lambda)^{1/2}} \lambda_j \phi_j \right\|^2 = \sum_j \frac{\lambda_j \phi_j(x)^2}{\lambda_j + \lambda} \leq C_\phi^2 \sum_j \frac{\lambda_j^{1-\beta}}{\lambda_j + \lambda}, \tag{B.2}
$$

Here the last step holds ν-almost-everywhere by our assumption $\|\phi_j\|_{L_\infty(\nu)} \leq C_\phi \lambda_j^{-\beta/2}$.

The function $t^{1-\beta}/(t+\lambda)$ is increasing for $0 \le t < \lambda(1-\beta)/\beta$ — the sign of its derivative is that of $(1-\beta)t^{-\beta}(t+\lambda) - t^{1-\beta} = t^{-\beta}[(1-\beta)\lambda - \beta t]$. Thus, we may substitute our bound $C_\lambda j^{-\alpha}$ for eigenvalues λ_j smaller than this threshold. Ordering the eigenvalues λ_j so they are decreasing and taking $J = \max\{j \in \mathbb{N} : \lambda_j \ge \lambda(1-\beta)/\beta\}$, we bound the sum (B.2) above by

$$C_\phi^2 \sum_{j \le J} \frac{\lambda_j^{1-\beta}}{\lambda_j + \lambda} + C_\phi^2 \sum_{j > J} \frac{(C_\lambda j^{-\alpha})^{1-\beta}}{C_\lambda j^{-\alpha} + \lambda}. \tag{B.3}$$

Furthermore, as (i) for all J terms in the first sum here, $\lambda_j^{1-\beta}/(\lambda_j + \lambda) \le \lambda_j^{-\beta} \le [\lambda(1-\beta)/\beta]^{-\beta}$ and (ii) $\lambda(1-\beta)/\beta \le \lambda_J \le C_\lambda J^{-\alpha}$ and therefore $J \le [C_\lambda \beta/(\lambda(1-\beta))]^{1/\alpha}$, we may bound it by

$$C_\phi^2 \left[\frac{C_\lambda \beta}{\lambda(1-\beta)}\right]^{1/\alpha} \left[\frac{\lambda(1-\beta)}{\beta}\right]^{-\beta} = C_\phi^2 C_\lambda^{1/\alpha} \left[\frac{\beta}{1-\beta}\right]^{1/\alpha+\beta} \lambda^{-(1/\alpha+\beta)}.$$

In addition, we may bound the second sum here by an integral,

$$C_\phi^2 \int_J^\infty \frac{(C_\lambda^{-1/\alpha} t)^{-\alpha(1-\beta)}}{(C_\lambda^{-1/\alpha} t)^{-\alpha} + \lambda} dt = C_\phi^2 C_\lambda^{1/\alpha} \int_{C_\lambda^{-1/\alpha} J}^\infty \frac{s^{-\alpha(1-\beta)}}{s^{-\alpha} + \lambda} ds,$$

decompose that integral into two pieces,

$$\int_{C_\lambda^{-1/\alpha} J}^{\lambda^{-1/\alpha}} \frac{s^{-\alpha(1-\beta)}}{s^{-\alpha} + \lambda} ds + \int_{\lambda^{-1/\alpha}}^\infty \frac{s^{-\alpha(1-\beta)}}{s^{-\alpha} + \lambda} ds,$$

and bound each piece as follows:

$$\int_{C_\lambda^{-1/\alpha} J}^{\lambda^{-1/\alpha}} \frac{s^{-\alpha(1-\beta)}}{s^{-\alpha} + \lambda} ds \le \int_{C_\lambda^{-1/\alpha} J}^{\lambda^{-1/\alpha}} s^{\alpha\beta} \le \frac{1}{1+\alpha\beta} \lambda^{-(1/\alpha+\beta)}$$

$$\int_{\lambda^{-1/\alpha}}^\infty \frac{s^{-\alpha(1-\beta)}}{s^{-\alpha} + \lambda} ds \le \lambda^{-1} \int_{\lambda^{-1/\alpha}}^\infty s^{-\alpha(1-\beta)} ds = \frac{\lambda^{-1}}{1-\alpha(1-\beta)} \left[0 - \left(\lambda^{-1/\alpha}\right)^{1-\alpha(1-\beta)}\right] = \frac{\lambda^{-(1/\alpha+\beta)}}{\alpha(1-\beta)-1}.$$

To guarantee that the latter integral converges, we use our assumption $\alpha(1-\beta) > 1$.

Putting everything together, (B.3) and therefore (B.2) is bounded by

$$C_\phi^2 C_\lambda^{1/\alpha} \left[\left(\frac{\beta}{1-\beta}\right)^{1/\alpha+\beta} + \frac{1}{1+\alpha\beta} + \frac{1}{\alpha(1-\beta)-1}\right] \lambda^{-(1/\alpha+\beta)}$$
$$= C_\phi^2 C_\lambda^{1/\alpha} \left[\left(\frac{\beta}{1-\beta}\right)^{1/\alpha+\beta} + \frac{\alpha}{(1+\alpha\beta)(\alpha-(1+\alpha\beta))}\right] \lambda^{-(1/\alpha+\beta)}.$$

Thus, the square root of this quantity bounds $\left\|[L_{K,\nu} + \lambda I]^{-1/2} K_x\right\|$ ν-a.e. as claimed. □

Proof of Lemma 3.8. We may write $f = \sum_j f_j \lambda_j^\kappa \psi_j$ with $\|f\|_{L_{K,\nu}^\kappa}^2 = \sum_j f_j^2$. In terms of this decom-

position,

$$\begin{aligned}\left\|L_\lambda^{-1/2} f\right\|^2 &= \left\| \frac{f_j \lambda_j^{\kappa-1/2}}{(\lambda_j+\lambda)^{1/2}} \lambda_j^{1/2} \phi_j \right\|^2 \\ &= \sum_j f_j^2 \frac{\lambda_j^{2\kappa-1}}{\lambda_j+\lambda} \\ &\le \left(\sum_j f_j^2\right) \sup_j \frac{\lambda_j^{2\kappa-1}}{\lambda_j+\lambda} \\ &\le \left(\sum_j f_j^2\right) \sup_j \lambda_j^{2\kappa-2} \wedge \lambda^{-1}\lambda_j^{2\kappa-1}. \end{aligned} \tag{B.4}$$

The last two steps above are, respectively, via Hölder's inequality and the substitution of the lower bound $\lambda_j \vee \lambda$ for the denominator.

Ordering the eigenvalues λ_j so that they are decreasing, note that for $\kappa \in [1/2, 1)$, $\lambda_j^{2\kappa-2}$ is increasing and $\lambda^{-1}\lambda_j^{2\kappa-1}$ decreasing in λ_j. For $\lambda_j \le \lambda$, the minimum of the two expressions is equal to the first, and because that expression is increasing in λ, it is bounded by $\lambda^{2\kappa-2}$. For $\lambda_j \ge \lambda$, the minimum of the two expressions is equal to the second, and because that expression is decreasing in λ, it too is bounded by $\lambda^{2\kappa-2}$. Thus, this quantity bounds the supremum (B.4) that we are interested in. For $\kappa \ge 1$, both of these expressions are increasing in λ_j, and therefore decreasing in j. It follows that the supremum (over j) of this minimum is the minimum of the first term in the left branch and the first term in the right, and we may take either as an upper bound. We choose the first, $\lambda_1^{2\kappa-2}$. This establishes our first claim.

The second claim follows by the same argument working with the analogous bound

$$\left\|L_\lambda^{-1} f\right\|^2 \le \left(\sum_j f_j^2\right) \sup_j \lambda_j^{2\kappa-3} \wedge \lambda^{-1}\lambda_j^{2\kappa-2}.$$

□

Proof of Lemma 3.9. The unit ball of $\mathcal{H}_K$ can be characterized as $\{\sum_{j=1}^\infty f_j \lambda_j^{1/2}\phi_j : \sum_{j=1}^\infty f_j^2 \le 1\}$ where ϕ_j are eigenfunctions of $L_{K,\nu}$ that form an orthonormal basis of $L_2(\nu)$. Let $\phi_0 = g/\|g\|_{L_2(\nu)}$ and $\phi_j^\perp = \phi_j - \langle \phi_j, \phi_0\rangle_{L_2(\nu)}\phi_0$ for $j \ge 1$. Any function in our set $\mathcal{B}^\star$ can be written in the form

$$f - sg = \sum_{j=1}^\infty f_j \lambda_j^{1/2}\phi_j^\perp + \left[\sum_{j=1}^\infty f_j \lambda_j^{1/2}\langle \phi_j, \phi_0\rangle_{L_2(\nu)} - s\|g\|_{L_2(\nu)}\right]\phi_0 \text{ for } \sum_{j=1}^\infty f_j^2 \le 1, s \in [0,1].$$

By Cauchy-Schwartz, the bracketed term is bounded by $\lambda_0^{1/2} = \sqrt{\sum_{j=1}^\infty \lambda_j \langle \phi_j, \phi_0\rangle_{L_2(\nu)}^2} + \|g\|_{L_2(\nu)}$. Thus, $\mathcal{B}^\star$ is contained in the set $\mathcal{B}'$ of functions of the form

$$f = f_0 \lambda_0^{1/2}\phi_0 + \sum_{j=1}^\infty f_j \lambda_j^{1/2}\phi_j^\perp \text{ for } \sum_{j=1}^\infty f_j^2 \le 1, f_0^2 \le 1.$$

Define the rescaled basis functions $\tilde{\phi}_j = \phi_j^\perp / \|\phi_j^\perp\|_{L_2(\nu)}$ and $\tilde{\lambda}_j = \lambda_j \|\phi_j^\perp\|^2_{L_2(\nu)} = \lambda_j(1 - \langle \phi_j, \phi_0 \rangle^2_{L_2(\nu)})$ for $j \geq 1$ and let $\tilde{\phi}_0 = \phi_0$ and $\tilde{\lambda}_0 = \lambda_0$. Equivalently, we may say that $\mathcal{B}'$ is the set of functions

$$f = \sum_{j=0}^{\infty} f_j \tilde{\lambda}_j^{1/2} \tilde{\phi}_j \quad \text{for} \quad \sum_{j=1}^{\infty} f_j^2 \leq 1, f_0^2 \leq 1.$$

From this point we imitate the proof of Mendelson (2002, Theorem 41). If f is a function of the form above, $Z_i f(X_i)$ satisfies

$$\mathbb{E}(Z_i f(X_i))^2 = \int f(x)^2 s_z d\nu_x = \int f(x)^2 d\nu(x) = \sum_{j=1}^{\infty} f_j^2 \tilde{\lambda}_j.$$

Therefore if $f(x)$ is in the set $\mathcal{B}'_t = \{f \in \mathcal{B}' : \mathbb{E}(Z_i f(X_i))^2 \leq t^2\}$, it has coefficients that satisfy

$$f_0^2 \leq 1, \sum_{j=1}^{\infty} f_j^2 \leq 1, \sum_{j=0}^{\infty} f_j^2 \tilde{\lambda}_j / t^2 \leq 1.$$

Now consider the set $\mathcal{E}_t$ of functions f with coefficients satisfying $\sum_{j=0}^{\infty} f_j^2 (1 \vee \tilde{\lambda}_j / t^2) \leq 1$. As $\sum_{j=0}^{\infty} f_j^2 (1 \vee \tilde{\lambda}_j / t^2) \leq f_0^2 + \sum_{j=1}^{\infty} f_j^2 + \sum_{j=0}^{\infty} f_j^2 \tilde{\lambda}_j / t^2 \leq 3$ for all functions in $\mathcal{B}'_t$, $\sqrt{3}\mathcal{E}_t$ contains $\mathcal{B}'_t$ and therefore also $\mathcal{B}^\star_t = \{f \in \mathcal{B}^\star : \mathbb{E}(Z_i f(X_i))^2 \leq t^2\}$. Thus, we will use $\sqrt{3} M_n\{zf(x) : f \in \mathcal{E}_t\}$ to bound $M_n\{zf(x) : f \in \mathcal{B}^\star_t\}$.

Note that in the case that $t = \infty$, we can improve this constant $\sqrt{3}$ to $\sqrt{2}$. $\mathcal{E}_\infty$ is the set of functions f with coefficients satisfying $\sum_{j=0}^{\infty} f_j^2 \leq 1$, and as $f_0^2 + \sum_{j=1}^{\infty} f_j^2 \leq 2$ for $f \in \mathcal{B}'$, $\sqrt{2}\mathcal{E}_\infty \supseteq \mathcal{B}' \supseteq \mathcal{B}^\star$.

We will bound $M_{n,2}\{zf(x) : f \in \mathcal{E}_t\} = [\mathbb{E} \sup_{f \in \mathcal{E}_t} |n^{-1} \sum_{i=1}^n \sigma_i Z_i f(X_i)|]^{1/2}$, as by Jensen's inequality this quantity bounds $M_n\{zf(x) : f \in \mathcal{E}_t\}$ itself. Writing $e_0, e_1, \ldots$ for the standard basis

for ℓ_2, we have

$$
\begin{aligned}
M_{n,2}\{zf(x) : f \in \mathcal{E}_t\}^2 &= \mathbb{E} \sup_{f \in \mathcal{E}_t} \left\langle \sum_{j=0}^{\infty} f_j e_j, n^{-1} \sum_{i=1}^{n} \sum_{j=0}^{\infty} \sigma_i Z_i \tilde{\lambda}_j^{1/2} \tilde{\phi}_j(X_i) e_j \right\rangle_{\ell_2}^2 \\
&= \mathbb{E} \sup_{f \in \mathcal{E}_t} \left\langle \sum_{j=0}^{\infty} f_j \sqrt{1 \vee \tilde{\lambda}_j / t^2}\ e_j, n^{-1} \sum_{i=1}^{n} \sum_{j=0}^{\infty} \sigma_i Z_i \sqrt{\frac{\tilde{\lambda}_j}{1 \vee \tilde{\lambda}_j / t^2}} \tilde{\phi}_j(X_i) e_j \right\rangle_{\ell_2}^2 \\
&\leq \mathbb{E} \left\| n^{-1} \sum_{i=1}^{n} \sum_{j=0}^{\infty} \sigma_i Z_i \sqrt{\tilde{\lambda}_j \wedge t^2} \tilde{\phi}_j(X_i) e_j \right\|_{\ell_2}^2 \\
&= n^{-2} \sum_{i=1}^{n} \sum_{j=0}^{\infty} \left(\tilde{\lambda}_j \wedge t^2 \right) \mathbb{E} \left[\sigma_i^2 Z_i^2 \tilde{\phi}_j(X_i)^2 \right] \\
&= n^{-1} \sum_{j=0}^{\infty} \left(\tilde{\lambda}_j \wedge t^2 \right) \mathbb{E} \left[\mathbb{E} \left[\sigma_i^2 \mid X_i, Z_i \right] \mathbb{E} \left[Z_i^2 \tilde{\phi}_j(X_i)^2 \right] \right] \\
&\leq n^{-1} \sum_{j=0}^{\infty} \left(\tilde{\lambda}_j \wedge t^2 \right) \left\| \mathbb{E} \left[\sigma_i^2 \mid X_i, Z_i \right] \right\|_{L_\infty(\nu_{x,z})} \left\| Z_i^2 \tilde{\phi}_j(X_i)^2 \right\|_{L_1(\nu_{x,z})} \\
&= \left(n^{-1} \sum_{j=0}^{\infty} \tilde{\lambda}_j \wedge t^2 \right) \left\| \mathbb{E} \left[\sigma_i^2 \mid X_i, Z_i \right] \right\|_{L_\infty(\nu_{x,z})}.
\end{aligned}
$$

The inequalities above are via Cauchy-Schwartz and Hölder's inequality respectively.

All that remains now is to simplify this bound so we need not discuss $\tilde{\lambda}_j$. Recall that $\tilde{\lambda}_0 = \left(\sqrt{\sum_{j=1}^{\infty} \lambda_j \langle \phi_j, \phi_0 \rangle_{L_2(\nu)}^2} + \|g\|_{L_2(\nu)} \right)^2$ and that $\tilde{\lambda}_j = \lambda_j (1 - \langle \phi_j, \phi_0 \rangle_{L_2(\nu)}^2)$ for $j \geq 1$. Typically when we consider the local Rademacher complexity, i.e. the case $t < \infty$, we take t small, so we will have t in our sum rather than the large $\tilde{\lambda}_0$ term. Thus, in that case we will not be able to cancel terms appearing in $\tilde{\lambda}_0$ and the rest of the sum. To bound $\tilde{\lambda}_0$, we apply Hölder's inequality to the term inside the square root in our expression for $\tilde{\lambda}_0$, yielding $\tilde{\lambda_0} \leq \lambda_0' = [(\lambda_1^{1/2} + 1)\|g\|_{L_2(\nu)}]^2$. Note that typically $\tilde{\lambda}_0 > t$, so the looseness of this bound will be irrelevant. For the other terms, we use the simple bound $\tilde{\lambda}_j \leq \lambda_j$ for $j \geq 1$, Thus, letting $\lambda_j' = \lambda_j$, we have our claimed bound $M_n\{zf(x) : f \in \mathcal{B}^\star,\ \mathbb{E}(Z_i f(X_i))^2 \leq t^2\} \leq \sqrt{3} M_{n,2}\{zf(x) : f \in \mathcal{E}_t\} \leq \|\mathbb{E}\left[\sigma_i^2 \mid X_i, Z_i\right]\|_{L_\infty(\nu_{x,z})}^{1/2} \cdot 3^{1/2} n^{-1/2} \sum_{j=0}^{\infty} \lambda_j' \wedge t$.

In the case $t = \infty$, terms from $\tilde{\lambda}_0$ and the other terms in our sum will cancel, giving us a better

bound.

$$\begin{aligned}
\sum_{j=0}^{\infty}\tilde{\lambda}_j &= \sum_{j=1}^{\infty}\lambda_j\langle\phi_j,\phi_0\rangle^2_{L_2(\nu)} + 2\|g\|_{L_2(\nu)}\sqrt{\sum_{j=1}^{\infty}\lambda_j\langle\phi_j,\phi_0\rangle^2_{L_2(\nu)}} + \|g\|^2_{L_2(\nu)} + \sum_{j=1}^{\infty}\lambda_j\left(1-\langle\phi_j,\phi_0\rangle^2_{L_2(\nu)}\right)\\
&= 2\|g\|_{L_2(\nu)}\sqrt{\sum_{j=1}^{\infty}\lambda_j\langle\phi_j,\phi_0\rangle^2_{L_2(\nu)}} + \|g\|^2_{L_2(\nu)} + \sum_{j=1}^{\infty}\lambda_j\\
&\le 2\|g\|_{L_2(\nu)}\sqrt{\sum_{j=1}^{\infty}\lambda_j} + \|g\|^2_{L_2(\nu)} + \sum_{j=1}^{\infty}\lambda_j\\
&= \left(\|g\|_{L_2(\nu)} + \sqrt{\sum_{j=1}^{\infty}\lambda_i}\right)^2.
\end{aligned}$$

Therefore $M_n\{zf(x) : x \in \mathcal{B}^\star\} \le 2^{1/2}M_{n,2}(\mathcal{E}_\infty) \le 2^{1/2}\|\mathbb{E}\left[\sigma_i^2 \mid X_i, Z_i\right]\|^{1/2}_{L_\infty(\nu_{x,z})}n^{-1/2}(\|g\|_{L_2(\nu)} + \sqrt{\sum_{j=1}^{\infty}\lambda_i})$ as claimed.

□

Proof of Lemma 3.10. Let $J = \max\{j : \lambda_j \ge t^2\}$. J satisfies $t^2 \le \lambda_J \le CJ^{-\alpha}$ and therefore $J \le (C/t^2)^{1/\alpha}$. Therefore,

$$\begin{aligned}
\sum_{j=1}^{\infty}\lambda_j \wedge t^2 &= Jt^2 + \sum_{j=J+1}^{\infty}\lambda_j\\
&\le (C/t^2)^{1/\alpha}t^2 + \int_{(C/t^2)^{1/\alpha}}^{\infty} Cs^{-\alpha}ds\\
&= C^{1/\alpha}t^{2(1-1/\alpha)} + [C/(\alpha-1)][(C/t^2)^{1/\alpha}]^{1-\alpha}\\
&= C^{1/\alpha}[\alpha/(\alpha-1)]t^{2(1-1/\alpha)}.
\end{aligned}$$

□

Proof of Proposition 3.11. Consider the decomposition of the expression maximized in $I_{h,\mathcal{B}^C}$

$$\frac{1}{n}\sum_{i=1}^{n}[h(X_i,W_i,f) - \gamma_i f(X_i,W_i)] = \frac{1}{n}\sum_{i=1}^{n}[T(X_i,W_i) - 1_{\{W_i=0\}}\gamma_i]f(x,0) + \frac{1}{n}\sum_{i:W_i\neq 0}\gamma_i f(x,w)$$

and note that if there were nonzero weights γ_i in the second sum, functions $f(\cdot,1)\dots f(\cdot,C)$ could be chosen from the (symmetric) unit ball $\mathcal{B}$ that make the second term match the first in sign. It follows that the weights $\gamma_i' = 1_{\{W_i=0\}}\gamma_i$ satisfy $I^2_{h,\mathcal{B}^C}(\gamma') \le I^2_{h,\mathcal{B}^C}(\gamma)$ and, unless $\gamma_i' = \gamma_i$, $\|\gamma'\|^2 < \|\gamma'\|^2$. Therefore it suffices to optimize over weights of the form $1_{\{W_i=1\}}\gamma_i$.

The space normed by $\mathcal{B}^C$ that we consider is reflexive, as cartesian products of reflexive spaces are reflexive. Then Lemma 2.5 establishes that $M_{n,\mathcal{B}^C}$ has a maximum at some possibly nonunique function $\hat{g}$ and that the weights satisfy $\hat{\gamma}_i = \hat{g}(X_i,W_i)$ for all such functions. To establish our last

claimed property, observe that taking $g(\cdot, w) \neq 0$ for $w \neq 0$ decreases the second term of $M_{n,\mathcal{B}^C}$ without increasing any other term. $\square$

Proof of Lemma 3.12. Let $Z_i = 1_{\{W_i=0\}}$. As $\tilde{\gamma}(w,x)$ takes the form $1_{\{w=0\}}g'(x)$ and $\gamma_\psi(w,x)$ the form $1_{\{w=0\}}g(x)$, we can rewrite the quantity we are bounding as

$$\frac{1}{n}\sum_{i=1}^n Z_i(g'(X_i) - g(X_i))^2 - \frac{2}{n}\sum_{i=1}^n (T_i - Z_i g(X_i))Z_i(g'(X_i) - g(X_i)) + \lambda\|g'\|_{\mathcal{H}_K}^2.$$

The middle term is centered, so we bound it using Chebyshev's inequality. With probability greater than $1 - \delta/2$,

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n (T_i - Z_i g(X_i))Z_i(g'(X_i) - g(X_i))\right| &< (n\delta/2)^{-1/2}\mathbb{E}\left[(T_i - Z_i g(X_i))^2 Z_i(g'(X_i) - g(X_i))^2\right]^{1/2} \\ &\leq (n\delta/2)^{-1/2}\|g\|_{L_\infty(P_Z)}\mathbb{E}\left[Z_i(g'(X_i) - g(X_i))^2\right]^{1/2} \\ &= (n\delta/2)^{-1/2}\|g\|_{L_\infty(P_Z)}\mathbb{E}\left[Z_i\right]^{1/2}\|g' - g\|_{L_2(P_Z)}.\end{aligned}$$

We use Markov's inequality for the first term. With probability greater than $1 - \delta/2$,

$$\frac{1}{n}\sum_{i=1}^n Z_i(g'(X_i) - g(X_i))^2 \leq (\delta/2)^{-1}\mathbb{E}\left[Z_i(g'(X_i) - g(X_i))^2\right] = (\delta/2)^{-1}\mathbb{E}\left[Z_i\right]\|g' - g\|_{L_2(P_Z)}^2.$$

Thus, with probability $1 - \delta$, we have the bound

$$2\delta^{-1}\mathbb{E}\left[Z_i\right]\|g' - g\|_{L_2(P_Z)}^2 + 2^{3/2}(\delta^{-1}\mathbb{E}\left[Z_i\right])^{1/2}n^{-1/2}\|g\|_{L_\infty(P_Z)}\|g' - g\|_{L_2(P_Z)} + \lambda\|g'\|_{\mathcal{H}_K}^2. \tag{B.5}$$

Our first claim follows by observing that because $\mathcal{H}_K$ is dense in $L_2(P_Z)$, there exists a sequence g_j' satisfying $\|g_j' - g\|_{L_2(P_T)} \to 0$ as $j \to \infty$ and $\|g_j'\|_{\mathcal{H}_K} < \infty$, as so long as $\lambda_n \to 0$, we can take a subsequence $g_n' = g_{j_n}$ such that $\lambda_n\|g_n\|_{\mathcal{H}_K} \to 0$.

Our second claim is a consequence of Cucker and Zhou (2007, Theorem 4.1), which establishes that if $\|g\|_{L_{k,\nu}^\kappa} < \infty$,

$$\inf_{g':\|g'\|_{\mathcal{H}_K} \leq r}\|g' - g\|_{L_2(\nu)} \leq \left(2\|g\|_{L_2(\nu)}\right)^{\frac{2}{1-2\kappa}} r^{-\frac{4\kappa}{1-2\kappa}}. \tag{B.6}$$

Thus, (B.5) is bounded by $ar^{-2\theta} + br^{-\theta} + \lambda r^2$ for $\theta = 4\kappa/(1-2\kappa)$, $a = 2\delta^{-1}\mathbb{E}\left[Z_i\right](2\|g\|_{L_2(P_Z)})^{\frac{4}{1-2\kappa}}$, and $b = 2^{3/2}(\delta^{-1/2}\mathbb{E}\left[Z_i\right])^{1/2}n^{-1/2}\|g\|_{L_\infty(P_Z)}(2\|g\|_{L_2(P_Z)})^{\frac{2}{1-2\kappa}}$. As this expression will be difficult to minimize analytically and b is very small relative to a, we simply evaluate this expression at the minimizer $r_\star = (a\theta/\lambda)^{1/[2(\theta+1)]}$ of $ar^{-2\theta} + \lambda r^2$. This yields the bound

$$\begin{aligned}&a^{\frac{1}{\theta+1}}\theta^{-\frac{\theta}{\theta+1}}\lambda^{\frac{\theta}{\theta+1}} + a^{\frac{\theta}{2(\theta+1)}}b\,\theta^{\frac{\theta}{2(\theta+1)}}\lambda^{\frac{\theta}{2(\theta+1)}} + a^{\frac{1}{\theta+1}}\theta^{\frac{1}{\theta+1}}\lambda^{\frac{\theta}{\theta+1}} \\ &= a^{\frac{1}{\theta+1}}\left(\theta^{-\frac{\theta}{\theta+1}} + \theta^{\frac{1}{\theta+1}}\right)\lambda^{\frac{\theta}{\theta+1}} + a^{\frac{\theta}{2(\theta+1)}}b\,\theta^{\frac{\theta}{2(\theta+1)}}\lambda^{\frac{\theta}{2(\theta+1)}}\end{aligned}$$

and substituting the above definitions of θ, a, b, we have the claimed bound

$$2\left(\delta^{-1}\mathbb{E}\left[Z_i\right]\right)^{\frac{1-2\kappa}{1+2\kappa}}\|g\|_{L_2(P_Z)}^{\frac{4}{1+2\kappa}}\left(\theta^{-\frac{\theta}{1+\theta}}+\theta^{\frac{1}{1+\theta}}\right)\lambda^{\frac{4\kappa}{1+2\kappa}}$$
$$+2^{\frac{4(1-\kappa)}{1-2\kappa}}\left(\delta^{-1}\mathbb{E}\left[Z_i\right]\right)^{\frac{1}{1+2\kappa}}\|g\|_{L_\infty(P_Z)}\|g\|_{L_2(P_Z)}^{\frac{4}{1-4\kappa^2}}\theta^{\frac{\theta}{2(\theta+1)}}n^{-1/2}\lambda^{\frac{2\kappa}{1+2\kappa}}.$$

□

B.5 Calculations used in the proof of (3.14)

In this section, we establish bounds on the quantities appearing in the bound (3.29). Note that because we have $f(\cdot, w) = 0$ when $w \neq 0$ for all $f \in \mathcal{F}$, we have

$$\begin{aligned}
\mathcal{F}^\star(t) &= \{1_{\{w=0\}}f'(x) : f' \in \mathcal{B}^\star, \|1_{\{w=0\}}f'\|_{L_2(P)} \leq t\};\\
\mathcal{H}^\star(t) &= \{[T(x,0) - g_\psi(x)]1_{\{w=0\}}f'(x) : 1_{\{w=0\}}f'(x) : f' \in \mathcal{B}^\star, \|1_{\{w=0\}}f'\|_{L_2(P)} \leq t\};\\
\mathcal{B}^\star &= \{f - sg_\psi(x) : \|f\| \leq 1, s \in [0,1]\}.
\end{aligned} \tag{B.7}$$

Via the triangle inequality, $M_{\mathcal{F}^\star} \leq M_K + \|g_\psi\|_\infty$ and $M_{\mathcal{H}^\star} \leq \|g_\psi\|_\infty(M_K + \|g_\psi\|_\infty)$.

B.5.1 Bounding non-aggregate terms

In this section, we bound the terms that appear in our bound: $\bar{R}$, $r_Q(\eta_C)$, $r_C(\eta_C)$, and $u(\mathcal{H}^\star, \delta)$.

B.5.1.1 Bounding $\bar{R}$.

In the event that $\|g_\psi\|_{\mathcal{H}_K} < \infty$, we may take $\tilde{\gamma} = \gamma_\psi$, in which case the condition (3.28) is satisfied deterministically with $\bar{R} = (\sigma^2/n)\|g_\psi\|_{\mathcal{H}_K}^2$. If instead we have $\|g_\psi\|_{L_{K,P_Z}^{\kappa_\gamma}} < \infty$ for $\kappa_\gamma \in (0,1)$, we use the second claim of Lemma 3.12.

B.5.1.2 Bounding $r_Q(\eta_Q)$.

$r_Q(\eta_Q)$ is a fixed point of the local Rademacher complexity of the class $\mathcal{F}^\star$. In the terms of Lemma 3.9, $R_n(\mathcal{F}^\star(t)) = M_n\{zf(x) : f \in \mathcal{B}^\star, \mathbb{E}(Z_i f(X_i))^2 \leq t^2\}$ for $g(x) = g_\psi(x)$, $Z_i = 1_{\{W_i=0\}}$, $\nu_{x,z} = P$, an iid Rademacher sequence $\sigma_1 \ldots \sigma_n$ independent of $(X_i, W_i)_{i\leq n}$. As $s_z = P\{W_i = 0 \mid X_i\}$, and $d\nu = p_Z dP_Z$ for $p_Z = P\{W_i = 0\}$. Thus, we have the bound $R_n(\mathcal{F}^\star(t)) \leq [(3/n)\sum_{j=0}^\infty (p_Z\lambda_{j,Z}) \wedge t^2]^{1/2}$ where $\lambda_{0,Z} = p_Z(1 + \sqrt{p_Z\lambda_{1,Z}})^2\|g_\psi\|_{L_2(P_Z)}^2$. Via Lemma 3.10, the assumptions of Theorem 3.4 guarantee that $\sum_{j=1}^n \lambda_{j,Z} \wedge t^2 \leq (p_Z C_{\lambda,Z})^{1/\alpha}(1-1/\alpha)^{-1}t^{2(1-1/\alpha)}$, and so long as $t^2 \leq \min\{1, \lambda_{0,Z}\}$ adding the first term to this sum increases it by no more than t^2 and therefore by no more than $t^{2(1-1/\alpha)}$. Consequently, $R(\mathcal{F}^\star(t)) \leq C' n^{-1/2} t^{1-1/\alpha}$ where $C' = \{3[1 + (p_Z C_{\lambda,Z})^{1/\alpha}(1-1/\alpha)^{-1}]\}^{1/2}$. To bound $r_Q(\eta_Q)$, we take 7 times the solution to fixed point

equation $C'n^{-1/2}t^{1-1/\alpha} = t^2/(2M_{\mathcal{F}^\star})$, which is $t = (2M_{\mathcal{F}^\star}C'n^{-1/2})^{1/(1+1/\alpha)} = C_Q n^{-1/[2(1+1/\alpha)]}$ for $C_Q = \{12p_Z^2 M_{\mathcal{F}^\star}^2[1 + (p_Z C_{\lambda,Z})^{1/\alpha}(1-1/\alpha)^{-1}])\}^{1/[2(1+1/\alpha)]}$. Recalling our recent assumption that $t^2 \leq \min\{1, \lambda_{0,Z}\}$, this means that we have the bound

$$r_Q(\eta_Q) \leq 7C_Q n^{-1/[2(1+1/\alpha)]} \tag{B.8}$$

if it is no larger than $7\min\{1, \lambda_{0,Z}\}^{1/2}$ and therefore if it is no larger than $7p_Z^{1/2}\|g_\psi\|_{L_2(P_Z)}$.

B.5.1.3 Bounding $r_C(\eta_C)$

Our approach will be to find a simple function $u'(\cdot)$ for which we can solve the fixed point equation $u'(t) = \eta_C t^2$ and which, at that fixed point t, we have $\eta_C t^2 = u'(t) \geq u(\mathcal{H}^\star(t), \delta)$ and therefore $t \geq r_C(\eta_C)$. Recall that

$$u(\mathcal{H}^\star(t), \delta) = \min_{\eta > 0} u_\eta, u_\eta = 2(1+\eta)R_n(\mathcal{H}^\star(t)) + \overline{\sigma}(\mathcal{H}^\star(t))\sqrt{\frac{2\log(2\delta^{-1})}{n}} + 2M_{\mathcal{H}^\star}\left(\frac{1}{3} + \frac{1}{\eta}\right)\frac{\log(2\delta^{-1})}{n}.$$

Our first step will be to establish that for all $\eta > 0$, when $t \leq p_Z^{1/2}\|g_\psi\|_{L_2(P_Z)}$,

$$\begin{aligned}
&u(\mathcal{H}^\star(t), \delta) \leq C_1 n^{-1/2}t^{1-1/\alpha} + C_2 n^{-1/2}t + +C_3 n^{-1} \\
&C_1 = 2(1+\eta)\|g_\psi\|_\infty\{3[1 + (p_Z C_{\lambda,Z})^{1/\alpha}(1-1/\alpha)^{-1}]\}^{1/2}; \\
&C_2 = \max\{1, \|\gamma_\psi\|_\infty\}\sqrt{2\log(2\delta^{-1})}; \\
&C_3 = 2M_{\mathcal{H}^\star}\left(\frac{1}{3} + \frac{1}{\eta}\right)\log(2\delta^{-1}).
\end{aligned} \tag{B.9}$$

Here we include the third term in $u(\mathcal{H}^\star(t), \delta)$ as-is and include a bound on the second using $\overline{\sigma}(\mathcal{H}^\star(t)) \leq \|T(x,0) - g_\psi(x)\|_\infty t \leq \max\{1, \|g_\psi(x)\|_\infty\}t$, so what remains to do is show that $C_1 n^{-1/2}t^{1-1/\alpha}$ bounds $R_n(\mathcal{H}^\star(t))$. To do this, we apply Lemma 3.9 with $g = g_\psi$, $Z_i = 1_{\{W_i=0\}}$ and $\sigma_i = \sigma_i'(T(X_i, W_i) - \gamma_\psi(X_i, W_i))$ for an iid Rademacher sequence $\sigma_i' \ldots \sigma_n'$ independent of $(X_i, W_i)_{i \leq n}$. Then, noting that $\|\sigma_i^2\|_\infty = \|(T(x,w) - \gamma_\psi)^2\|_\infty \leq \|\gamma_\psi\|_\infty^2$ and that as in the previous case $\nu = p_Z \cdot P_Z$ and therefore $\lambda_j = p_Z\lambda_{j,Z}$, we have the bound $R_n(\mathcal{H}^\star(t)) \leq \|\gamma_\psi\|_\infty[(3/n)\sum_{j=0}^\infty (p_Z\lambda_{j,Z}) \wedge t^2]^{1/2}$ where $\lambda_{0,Z} = p_Z(1 + \sqrt{p_Z\lambda_{1,Z}})\|g_\psi\|_{L_2(P_Z)}$. Using Lemma 3.10 to bound this, we have $R_n(\mathcal{H}^\star(t)) \leq C_1 n^{-1/2}t^{1-1/\alpha}$ when $t \leq p_Z^{1/2}\|g_\psi\|_{L_2(P_Z)}$.

Having established the validity of our bound (B.9), we now define something that will act as a bound on it: $u_a(t) = an^{-1/2}t^{1-1/\alpha}$, a multiple of its asymptotically dominant term. We will solve the fixed point equation $u_a'(t) = \eta_C t^2$ and then select a so that $u_a(t_a)$ upper bounds the right side above at $t = t_a$, ensuring that $t_a \geq r_C(\eta_C)$ as desired. The solution to this fixed point equation is $t_a = (an^{-1/2}/\eta_C)^{1/(1+1/\alpha)}$. When our condition $t_a \leq p_Z^{1/2}\|g_\psi\|_{L_2(P_Z)}$ for the validity of our bound on $R_n(\mathcal{F}^\star(t))$ is satisfied, clearly t_a at $a = C_1$ bounds the first term in (B.9). To incorporate the

other terms as well, we will bound their ratios with $u_a(t_a)$, too. For all $a \geq 1$, the ratio of the second and third terms in (B.9) and $u_a(t_a)$ are respectively

$$C_2 a^{-1} t_a^{1/\alpha} = C_2 \eta_C^{-1/(\alpha+1)} a^{-1/(1+1/\alpha)} n^{-1/[2(\alpha+1)]} \leq C_2 \eta_C^{-1/(\alpha+1)} n^{-1/[2(\alpha+1)]};$$
$$C_3 n^{-1/2} a^{-1} t_a^{1/\alpha - 1} = C_3 n^{-1/2} a^{-1} (a n^{-1/2}/\eta_C)^{(1-\alpha)/(1+\alpha)} \leq C_3 \eta_C^{(\alpha-1)/(\alpha+1)} n^{-1/(\alpha+1)}.$$

It follows that $t_a \geq r_C(\eta_C)$ if (i) a is no smaller than the sum of our three ratio bounds and also no smaller than one, the latter being required for the validity of our second and third ratio bounds and (ii) t_a is no larger than $p_Z^{1/2} \|g_\psi\|_{L_2(P_Z)}$, required for the validity of our local Rademacher complexity bound. Thus, in terms of C_1, C_2, C_3 defined in (B.9), we have

$$r_C(\eta_C) \leq \max\left\{\eta_C^{-1}, C_1\eta_C^{-1} + C_2\eta_C^{-\left(1+\frac{1}{\alpha+1}\right)} n^{-\frac{1}{2(\alpha+1)}} + C_3\eta_C^{-\left(1-\frac{\alpha-1}{\alpha+1}\right)} n^{-\frac{1}{\alpha+1}}\right\}^{\frac{1}{1+1/\alpha}} n^{-\frac{1}{2(1+1/\alpha)}} \tag{B.10}$$

so long as the entire bound is no larger than $p_Z^{1/2}\|g_\psi\|_{L_2(P_Z)}$. Here the expression within the maximum plays the role of $a\eta_C^{-1}$, and we take this maximum to ensure that our condition $a \geq 1$ holds.

We will use a variant of this bound with simpler dependence on η_C,

$$r_C(\eta_C) \leq \left[\eta_C^{-1}\left(C_1 + C_2 n^{-\frac{1}{2(\alpha+1)}} + C_3 n^{-\frac{1}{\alpha+1}}\right)\right]^{\frac{1}{1+1/\alpha}} n^{-\frac{1}{2(1+1/\alpha)}}, \tag{B.11}$$

valid when $2^{\alpha+1} \leq \eta_C \leq \left[[C_2/(2C_3)]n^{1/2}\right]^{\frac{\alpha+1}{\alpha-1}}$, the parenthesized term exceeds 1, and the entire bound is no larger than $p_Z^{1/2}\|g_\psi\|_{L_2(P_Z)}$. To show that this is a valid upper bound, we will show that the bracketed expression in (B.11) exceeds the right branch of the maximum in (B.10). The difference between these two expressions is

$$C_2 n^{-\frac{1}{2(\alpha+1)}}\left(\eta_C^{-1} - \eta_C^{-\left(1+\frac{1}{\alpha+1}\right)}\right) + C_3 n^{-\frac{1}{\alpha+1}}\left(\eta_C^{-1} - \eta_C^{-\left(1-\frac{\alpha-1}{\alpha+1}\right)}\right),$$

which is positive when the ratio

$$C_2 n^{-\frac{1}{2(\alpha+1)}}\left(\eta_C^{-1} - \eta_C^{-\left(1+\frac{1}{\alpha+1}\right)}\right) \Big/ C_3 n^{-\frac{1}{\alpha+1}}\left(\eta_C^{-\left(1-\frac{\alpha-1}{\alpha+1}\right)} - \eta_C^{-1}\right)$$

exceeds one. This ratio is bounded above

$$(C_2/C_3) n^{\frac{1}{2(\alpha+1)}} \eta_C^{-\frac{\alpha-1}{\alpha+1}}\left(1 - \eta_C^{-\frac{1}{\alpha+1}}\right),$$

To complete our argument, observe that our lower bound $2^{\alpha+1}$ on η_C implies that the parenthesized factor is at least $1/2$, and consequently that our upper bound on η_C implies that the quantity above is at least one as required.

B.5.1.4 Bounding $u(\mathcal{H}^\star, \delta)$

To bound $u(\mathcal{H}^\star, \delta)$, we bound $R_n(\mathcal{H}^\star)$ and $\overline{\sigma}(\mathcal{H}^\star)$. We bound the latter using Hölder's inequality and the triangle inequality,

$$\begin{aligned}
\overline{\sigma}(\mathcal{H}^\star) &\leq \max\{1, \|g_\psi\|_\infty\} \sqrt{\sup_{\|f\|\leq 1} \mathbb{E}\, 1_{\{W_i=1\}} (f(X_i) + g_\psi(X_i))^2} \\
&= \max\{1, \|g_\psi\|_\infty\} \sqrt{\sup_{\|f\|\leq 1} p_Z \mathbb{E}\left[(f(X_i) + g_\psi(X_i))^2 \mid W_i = 1\right]} \\
&\leq p_Z^{1/2} \max\{1, \|g_\psi\|_\infty\} \left(\sup_{\|f\|\leq 1} \|f\|_{L_2(P_Z)} + \|g_\psi\|_{L_2(P_Z)} \right) \\
&= p_Z^{1/2} \max\{1, \|g_\psi\|_\infty\} \left(\lambda_{1,Z}^{1/2} + \|g_\psi\|_{L_2(P_Z)} \right)
\end{aligned}$$

The identity $\sup_{\|f\|\leq 1} \|f\|_{L_2(P_Z)} = \lambda_{1,Z}$ used in the last step follows from the representation of this unit ball as the set $\{\sum_{j=1}^\infty f_j \lambda_{j,Z}^{1/2} \psi_j(x) : \sum_{j=1}^\infty f_j^2 \leq 1\}$ in terms of $L_2(P_Z)$-orthonormal eigenfunctions ψ_j.

We bound the Rademacher complexity using Lemma 3.9 with $t = \infty$, $g = g_\psi$, $Z_i = 1_{\{W_i=0\}}$ and $\sigma_i = \sigma_i'(T(X_i, W_i) - \gamma_\psi(X_i, W_i))$ for an iid Rademacher sequence $\sigma_i' \ldots \sigma_n'$ independent of $(X_i, W_i)_{i\leq n}$. Then, noting that $\|\sigma_i^2\|_\infty = \|(T(x,w) - \gamma_\psi)^2\|_\infty \leq \|\gamma_\psi\|_\infty^2$ and that as in the previous case $\nu = p_Z \cdot P_Z$ and therefore $\lambda_j = p_Z \lambda_{j,Z}$, we have the bound

$$\begin{aligned}
R_n(\mathcal{H}^\star) &\leq 2^{1/2} \|g_\psi\|_\infty n^{-1/2} \left(p_Z^{1/2} \|g_\psi\|_{L_2(P_Z)} + \sqrt{\sum_{j=1}^\infty p_Z \lambda_{j,Z}} \right) \\
&\leq 2^{1/2} p_Z^{1/2} \|g_\psi\|_\infty n^{-1/2} \left(\|g_\psi\|_{L_2(P_Z)} + \sqrt{C_{\lambda,Z} \int_{s=1}^\infty s^{-\alpha}} \right) \\
&= 2^{1/2} p_Z^{1/2} \|\|g_\psi\|\|_\infty \left(\|g_\psi\|_{L_2(P_Z)} + C_{\lambda,Z}^{1/2} (\alpha - 1)^{-1/2} \right) n^{-1/2}.
\end{aligned}$$

and therefore

$$\begin{aligned}
u(\mathcal{H}^\star, \delta) &\leq \min_{\eta > 0} C_{u,1,\eta} n^{-1/2} + C_{u,2,\eta} n^{-1}; \qquad \text{(B.12)} \\
C_{u,1,\eta} &= (1+\eta) 2^{3/2} p_Z^{1/2} \max\{1, \|g_\psi\|_\infty\} \left(\|g_\psi\|_{L_2(P_Z)} + C_{\lambda,Z}^{1/2} (1-\alpha)^{-1/2} \right) \\
&\quad + 2^{1/2} p_Z^{1/2} \max\{1, \|g_\psi\|_\infty\} \left(\lambda_{1,Z} + \|g_\psi\|_{L_2(P_Z)} \right) \sqrt{\log(2\delta^{-1})}; \\
C_{u,2,\eta} &= 2 M_{\mathcal{H}^\star} (1/3 + 1/\eta) \log(2\delta^{-1}).
\end{aligned}$$

B.5.2 Aggregating terms

In order to simplify our statement of (3.14) as much as possible, we equate our our bounds (B.8) and (B.11) on $r_Q(\eta_Q)$ and $r_C(\eta_C)$ by setting

$$\eta_C = \left(C_1 + C_2 n^{-\frac{1}{2(\alpha+1)}} + C_3 n^{-\frac{1}{\alpha+1}} \right) \Big/ (7 C_Q)^{1+1/\alpha}. \qquad \text{(B.13)}$$

Having chosen this value of η_C, $r = r_Q(\eta_Q) \vee r_C(\eta_C) \vee n^{-1/2}\sigma^{-1}\eta_Q^{-1/2}$ satisfies the bound $r \leq 7C_Q n^{-\frac{1}{2(1+1/\alpha)}} \vee n^{-1/2}\sigma^{-1}\eta_Q^{-1/2}$. This is usually optimal. Among choices of η_C, this one results in the sharpest bound (3.29) except in the case that $b < a$ and b is equal to the second of the three expressions of which it is the maximum. This completes our proof. All other bounds above appear in the bound (3.13) as-is.