cemmap

centre for microdata methods and practice

Adversarial estimation of Riesz representers

Victor Chernozhukov
Whitney Newey
Rahul Singh
Vasilis Syrgkanis

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP07/21

Economic and Social Research Council

Adversarial Estimation of Riesz Representers

Victor Chernozhukov Whitney Newey Rahul Singh Vasilis Syrgkanis

December 2020

Abstract

We provide an adversarial approach to estimating Riesz representers of linear functionals within arbitrary function spaces. We prove oracle inequalities based on the localized Rademacher complexity of the function space used to approximate the Riesz representer and the approximation error. These inequalities imply fast finite sample mean-squared-error rates for many function spaces of interest, such as high-dimensional sparse linear functions, neural networks and reproducing kernel Hilbert spaces. Our approach offers a new way of estimating Riesz representers with a plethora of recently introduced machine learning techniques. We show how our estimator can be used in the context of de-biasing structural/causal parameters in semi-parametric models, for automated orthogonalization of moment equations and for estimating the stochastic discount factor in the context of asset pricing.

Contents

1 Introduction

Many problems in econometrics, statistics, causal inference, and finance involve linear functionals of unknown functions:

$$\theta(g) = \mathbb{E}[m(Z; g)]$$

where Z denotes a random vector, and $g : \mathcal{X} \to \mathbb{R}$ is a function in some space $\mathcal{G}$. A continuous linear functional that is mean square continuous with respect to ℓ_2 norm can be written in a more benign and useful manner. Formally, for a given linear functional $\theta(\cdot)$, there exists a function a_0 such that for any $g \in \mathcal{G}$:[1]

$$\theta(g) = \mathbb{E}[a_0(X)\, g(X)]$$

This result is known as the Riesz representation theorem, and the function a_0 is the Riesz representer of the linear functional. Evaluation of a linear functional $\theta(g)$ can be achieved by simply taking the inner product between a_0 and g.

Knowing the Riesz representation of a linear functional is a critical building block in a variety of learning problems. For instance, in semi-parametric models, g_0 is an unknown regression function and $\theta(g_0)$ is a causal or structural parameter of interest. The Riesz representer a_0 of the functional $\theta(\cdot)$ can be used to debias the plug-in estimator and construct semi-parametrically efficient estimators of the parameter $\theta(g_0)$. In asset pricing applications, the Riesz representer corresponds to the stochastic discount factor, which is of primary interest when pricing financial derivatives.

Irrespective of the downstream application, the goal of this paper is to derive an estimator for the Riesz representer of any linear functional, when given access to n samples of the random vector Z and a target function space $\mathcal{A}$ that can well approximate the function a_0. We propose and analyze an estimator $\hat{a}$, with small mean-squared-error. Formally, with probability (w.p.) $1 - \zeta$:

$$\|\hat{a} - a_0\|_2 = \sqrt{\mathbb{E}\left[(\hat{a}(X) - a_0(X))^2\right]} \le \epsilon_{n,\zeta}$$

We consider estimation of the Riesz representer within some function space $\mathcal{A}$ and propose an adversarial estimator based on regularized variants of the following min-max criterion:

$$\hat{a} = \arg\min_{a\in\mathcal{A}} \max_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \left(m(Z_i; f) - a(X_i)\cdot f(X_i) - f(X_i)^2\right)$$

[1] For simplicity of exposition, throughout the paper we consider scalar-valued functions g. All our results naturally extend to vector-valued functions g, and estimate a vector valued Riesz representer that satisfies that $\theta(g) = \mathbb{E}[a(X)' g(X)]$.

We derive oracle inequalities for this estimator as a function of the localized Rademacher complexity of the function space $\mathcal{A}$ and the approximation error $\epsilon = \min_{a \in \mathcal{A}} \|a - a_0\|_2$.

We show that as long as the function class $\mathcal{F}$ contains the star-hull of differences of functions in $\mathcal{A}$, i.e. $\mathcal{F} := \{r(a - a') : a, a' \in \mathcal{A}, r \in [0,1]\}$, then the adversarial estimator satisfies w.p. $1 - \zeta$:

$$\|\hat{a} - a_0\|_2 = O\left(\epsilon + \delta_n + \sqrt{\frac{\log(1/\zeta)}{n}}\right)$$

where δ_n is the critical radius of the function classes $\mathcal{F}$ and $m \circ \mathcal{F} = \{Z \to m(Z; f) : f \in \mathcal{F}\}$. The critical radius of a function class is a widely used quantity in statistical learning theory that allows one to argue fast estimation rates that are nearly optimal. For instance, for parametric function classes, the critical radius is of order $n^{-1/2}$, leading to fast parametric rates (as compared to $n^{-1/4}$ which would be achievable via looser uniform deviation bounds).

Moreover, the critical radius has been analyzed and derived for a variety of function spaces of interest, such as neural networks, high-dimensional linear functions, reproducing kernel Hilbert spaces, and VC-subgraph classes. Thus our general theorem allows us to appeal to these characterizations and provide oracle rates for a family of Riesz representer estimators. Prior work on estimating Riesz representers only considered particular high-dimensional parametric classes and derived specialized estimators for the function space of interest. Our adversarial estimator provides a single approach that tackles generic function spaces in a uniform manner.

We also examine the computational aspect of our estimator. We provide examples of how estimation can be achieved in a computationally efficient manner for several function spaces of interest.

Finally, we show how our estimator can be used in the context of estimating causal or structural parameters in semi-parametric models. Specifically, our mean square rate for the Riesz representer is sufficiently fast to achieve semi-parametric efficiency and asymptotic normality of the causal or structural parameter.

1.1 Applications: Causal Inference and Asset Pricing

Estimation of Riesz representers arises in two important domains for economic research: causal inference and asset pricing.

Automated De-biasing of Causal Estimates. In causal inference, a variety of treatment effects and policy effects can be formulated as functionals–i.e., scalar summaries–of an underlying regression [36]. Formally, the causal parameter $\theta_0 = \theta(g_0) = \mathbb{E}[m(Z; g_0)]$ is a functional $\theta(\cdot)$ of the nuisance parameter $g_0(x) := \mathbb{E}[Y|X = x]$. In this paper, we consider a variety of treatment and policy effects including

1. Average treatment effect (ATE): $\theta_0 = \mathbb{E}[g_0(1, W) - g_0(0, W)]$, where $X = (D, W)$ consists of treatment and covariates.
2. Average policy effect: $\theta_0 = \int g_0(x) d\mu(x)$ where $\mu(x) = F_1(x) - F_0(x)$
3. Policy effect from transporting covariates: $\theta_0 = \mathbb{E}[g_0(t(X)) - g_0(X)]$

4. Cross effect: $\theta_0 = \mathbb{E}[Dg_0(0, W)]$, where $X = (D, W)$ consists of treatment and covariates.
5. Regression decomposition: $\mathbb{E}[Y|D = 1] - \mathbb{E}[Y|D = 0] = \theta_0^{response} + \theta_0^{composition}$ where

$$\begin{aligned}\theta_0^{response} &= \mathbb{E}[g_0(1, W)|D = 1] - \mathbb{E}[g_0(0, W)|D = 1] \\ \theta_0^{composition} &= \mathbb{E}[g_0(0, W)|D = 1] - \mathbb{E}[g_0(0, W)|D = 0]\end{aligned}$$

6. Average treatment on the treated (ATT): $\theta_0 = \mathbb{E}[g_0(1, W)|D = 1] - \mathbb{E}[g_0(0, W)|D = 1]$, where $X = (D, W)$ consists of treatment and covariates.
7. Local average treatment effect (LATE): $\theta_0 = \frac{\mathbb{E}[g_0(1,W)-g_0(0,W)]}{\mathbb{E}[h_0(1,W)-h_0(0,W)]}$, where $X = (V, W)$ consists of instrument and covariates and $h_0(x) := \mathbb{E}[D|X = x]$ is a second regression.

More generally, our results extend to parameters defined implicitly by $0 = \mathbb{E}[m(Z; g_0; \theta_0)]$, such as partially linear regression and partially linear instrumental variable regression.

If the regression g_0 is learned by a regularized estimator $\hat{g}$, then estimation of the causal parameter θ_0 by a plug-in estimator $\mathbb{E}_n[m(Z; \hat{g})]$ is badly biased. The solution is to use a de-biased formulation of the causal parameter instead: $\theta_0 = \mathbb{E}[m(Z; g_0) + a_0(X)\{Y - g_0(X)\}]$. Observe that a_0 arises in the bias correction term. We re-visit this example in Section 6.

Fundamental Asset Pricing Equation. In asset pricing, a variety of financial models deliver the same fundamental asset pricing equation. This equation is of both theoretical and practical interest. Theoretically, it elucidates why asset prices or returns are what they are. Practically, it can be used to identify trading opportunities when assets are mis-priced. The asset pricing equation follows from two weak assumptions: free portfolio formation, and the law of one price. In Appendix B.2, we review the derivation for a general audience.[2]

Formally, the fundamental asset pricing equation is $p_{t,i} = \mathbb{E}_t[m_{t+1}x_{t+1,i}]$ where $p_{t,i}$ is the price of asset i at time t, $x_{t+1,i}$ is payoff of asset i at time $t + 1$, and m_{t+1} is the market-wide stochastic discount factor (SDF) at time $t + 1$.[3] The expectation is conditional on information $(I_t, I_{t,i})$ known at time t: I_t are macroeconomic conditioning variables that are not asset specific, e.g. inflation rates and market return; $I_{t,i}$ are asset-specific characteristics, e.g. the size or book-to-market ratio of firm i at time t. The asset pricing equation encompasses stocks, bonds, and options. We clarify its many instantiations below, where d_{t+1} is dividend, C is the call price, S_T is the stock price at expiration, K is the strike price.

[2]The same asset pricing equation can be derived from either a model of complete markets for contingent claims, or a model of investor utility maximization. Free portfolio formation is a weaker assumption on market structure than the existence of complete markets for contingent claims. The law of one price is a weaker assumption on preference structure than investor utility maximization. We present these additional derivations in Appendix B.2.

[3]The SDF has many additional names: marginal rate of substitution, state price density, and pricing kernel. Each name corresponds to a different derivation of the asset pricing equation, starting from different first principles.

Asset	Price p_t	Payoff x_{t+1}
Stock	p_t	$p_{t+1} + d_{t+1}$
Bond	p_t	1
Option	C	$\max\{S_T - K, 0\}$
Return	1	R_{t+1}
Excess return	0	R^e_{t+1}

Table 1: Generality of asset pricing equation

The fundamental asset pricing equation can also be parametrized in terms of returns. If an investor pays one dollar for an asset i today, the gross rate of return $R_{t+1,i}$ is how many dollars the investor receives tomorrow; formally, the price is $p_{t,i} = 1$ and the payoff is $x_{t+1,i} = R_{t+1,i}$ by definition. Next consider what happens when an investor borrows a dollar today at the interest rate R^f_{t+1} and buys an asset i that gives the gross rate of return $R_{t+1,i}$ tomorrow. From the perspective of the investor, who paid nothing out-of-pocket, the price is $p_{t,i} = 0$ while the payoff is the excess rate of return $R^e_{t+1,i} := R_{t+1,i} - R^f_{t+1}$, leading to the asset pricing equation: $0 = \mathbb{E}_t[m_{t+1}R^e_{t+1,i}]$.

Following [29], we focus on the latter excess return parametrization of the asset pricing equation. Taking expectations yields the unconditional moment restriction

$$0 = \mathbb{E}[m_{t+1}R^e_{t+1,i}z(I_t, I_{t,i})] = \mathbb{E}[\mathbb{E}[m_{t+1}|R^e_{t+1,i}, I_t, I_{t,i}]R^e_{t+1,i}z(I_t, I_{t,i})], \quad \forall z(\cdot)$$

Our framework nests this final expression. Specifically,

$$\theta(g) = 0, \quad g(R^e_{t+1,i}, I_t, I_{t,i}) = R^e_{t+1,i}z(I_t, I_{t,i}), \quad a_0(R^e_{t+1,i}, I_t, I_{t,i}) = \mathbb{E}[m_{t+1}|R^e_{t+1,i}, I_t, I_{t,i}]$$

By estimating a_0, which is the projection of the SDF onto excess returns and other available information, one can pin down the price of any hypothetical asset.

1.2 Related Work

Classical Semi-parametric Statistics. Classical semi-parametric statistical theory studies the asymptotic properties of statistical quantities that are functionals of a density or a regression over a low-dimensional domain [82, 60, 65, 101, 77, 108, 128, 23, 91, 106, 129, 24, 92, 3, 93, 4, 123, 79, 5]. Any continuous linear functional has a Riesz representer. In this classical theory, the Riesz representer appears in the influence function and therefore in the asymptotic variance of semi-parametric estimators [91]. We depart from classical theory by considering the high-dimensional setting.

De-biased Machine Learning and Targeted Maximum Likelihood. Because the Riesz representer appears in the asymptotic variance of semi-parametric estimators, it can be incorporated into estimation to ensure semi-parametric efficiency. In practice, this can be achieved by introducing a de-biasing term into the estimating equation [60, 24, 133, 15, 16, 17, 18, 69, 70, 71, 124, 100, 37, 96, 103, 66, 67, 68, 27, 135, 136]. In doubly robust estimating equations for regression functionals, the de-biasing term is the product between the Riesz representer and the regression residual [107, 106, 127, 126, 84, 122]. The more general principle at play is Neyman orthogonality: the learning problem for the functional of interest becomes orthogonal to the learning problems for both the regression and the Riesz representer [97, 98, 129, 104, 134, 17, 18, 36, 14, 35, 51].

De-biased machine learning and targeted maximum likelihood combine the algorithmic insight of doubly-robust moment functions with the algorithmic insight of sample splitting [22, 113, 77, 129, 104]. In doing so, these frameworks facilitate a general analysis of residuals such that the target functional is $\sqrt{n}$-consistent under minimal assumptions on the estimators used for the regression and Riesz representer [112, 110, 111, 127, 134, 126, 44, 125, 74, 73]. In particular, any machine learning estimators are permitted that satisfy $\sqrt{n}\|\hat{g} - g_0\|_2 \cdot \|\hat{a} - a_0\|_2 \to 0$ [35, 36].

The Riesz representer may be a difficult object to estimate. Even for simple regression functionals such as policy effects, its closed form involves ratios of densities. In restricted models, where the regression is known to belong to a certain function class, there is the further difficulty of projecting the Riesz representer accordingly. A recent literature explores the possibility of directly estimating the Riesz representer, without estimating its components or even knowing its functional form [105, 95, 9, 39, 40, 62, 63, 117, 109]. A crucial insight, on which we build, is that the Riesz representer is directly identified from data.

[63] observe that to debias an average moment, it is sufficient to estimate an empirical analogue of the Riesz representer that approximately satisfies the Riesz representer moment equation on the n samples. They propose a parametric min-max criterion to estimate n parameters corresponding to the n evaluations of the empirical Riesz representer. Unlike [63], we provide a guarantee on learning the true Riesz representer, we approximate the Riesz representer within non-parametric function spaces, and our result therefore has broader application beyond causal inference. Importantly, [63] require that the same sample used to estimate the n parameters is used in final stage estimation of the causal parameter. As such, the analysis requires that the regression function g lies in a Donsker class–a restriction that precludes many machine learning estimators. By contrast, our adversarial estimator provides fast estimation rates with respect to the true Reisz representer and hence can be used in combination with cross-fitting and sample splitting to eliminate the Donsker assumption.

Adversarial Estimation. Riesz representation theorem can be viewed as a continuum of unconditional moment restrictions. The non-parametric instrumental variable problem, based on a conditional moment restriction, also implies a continuum of unconditional moment restrictions [94, 56, 25, 31, 42, 32, 33, 30]. A central insight of this work is that the min-max approach for conditional moment models may be adapted to the problem of learning the Riesz representer. In a min-max approach, the continuum of unconditional moment restrictions is enforced adversarially over a set of test functions [54, 8, 45].

The fundamental advantage of the min-max approach is its unified analysis over arbitrary function classes. In particular, via local Rademacher analysis, one can derive an abstract bound that encompasses sparse linear models, neural networks, and RKHS methods [78, 12]. As such, the min-max approach is actually a family of algorithms adaptive to a variety of data settings with a unified guarantee [90, 80, 81].

Machine Learning Approaches to Causal Inference and Asset Pricing. By pursuing a min-max approach, our work relates to previous work that incorporates a variety of machine learning methods into causal inference. Much work on de-biased machine learning focuses on sparse and approximately sparse models [39, 40, 38]. A neural network estimator with mean square rate has been successfully used to learn the nuisance regression in semiparametric estimation [34, 49] and to learn the structural function in nonparametric instrumental variable regression [59, 20, 45]. A more recent literature incorporates RKHS methods into causal inference due to their convenient closed form solutions and strong performance on smooth designs [99, 116, 89, 118, 88].

Finally, our works provides a theoretical foundation for a growing literature that incorporates machine learning into asset pricing. We follow the asset pricing literature in framing the problem of learning a stochastic discount factor as the problem of learning a Riesz representer [57]. Specifically, we propose a deep min-max approach based on free portfolio formation and the law of one price [11, 29]. This approach differs from deep learning approaches that predict asset prices via nonparametric regression [86, 50, 55, 21]. Unlike previous work, we prove mean square rates for the stochastic discount factor, and we prove $\sqrt{n}$-consistency and semiparametric efficiency for expected asset prices.

2 Adversarial Estimator

For any function space $\mathcal{G}$, let $\text{star}(\mathcal{G}) := \{r\, g : g \in \mathcal{G}, r \in [0,1]\}$, denote the star hull. Let $\partial\mathcal{G} := \{g - g' : g, g' \in \mathcal{G}\}$ denote the space of differences. We will consider estimators that estimate Riesz representers within some function space $\mathcal{A}$, equipped with some norm $\|\cdot\|_{\mathcal{A}}$. Moreover, let $\langle \cdot, \cdot \rangle_2$ be the inner product associated with the ℓ_2 norm, i.e. $\langle a, a' \rangle_2 := \mathbb{E}_X[a(X)\, a'(X)]$.[4] Given this notation, we define the class:

$$\mathcal{F} := \text{star}(\partial\mathcal{A}) := \{r\,(a - a') : a, a' \in \mathcal{A}, r \in [0,1]\}$$

and assume that the norm $\|\cdot\|_{\mathcal{A}}$ extends naturally to the larger space $\mathcal{F}$. Moreover, let $\mathbb{E}_n[\cdot]$ denote the empirical average and $\|\cdot\|_{2,n}$ the empirical ℓ_2 norm, i.e.

$$\|g\|_{2,n} := \sqrt{\mathbb{E}_n[g(X)^2]}.$$

Consider the following adversarial estimator:

$$\hat{a} = \arg\min_{a \in \mathcal{A}} \max_{f \in \mathcal{F}} \mathbb{E}_n[m(Z; f) - a(X) \cdot f(X)] - \|f\|_{2,n}^2 - \lambda \|f\|_{\mathcal{A}}^2 + \mu \|a\|_{\mathcal{A}}^2 \tag{1}$$

Remark 1 (Population limit)**.** *Consider the population limit of our criterion where $n \to \infty$ and $\lambda, \mu \to 0$. Then our criterion is:*

$$\max_{f \in \mathcal{F}} \mathbb{E}[m(Z; f) - a(X) \cdot f(X)] - \|f\|_2^2$$

By the definition of the Riesz representer we thus have:

$$\begin{aligned} \max_{f \in \mathcal{F}} \mathbb{E}\left[m(Z; f) - a(X) \cdot f(X)\right] - \|f\|_2^2 = & \max_{f \in \mathcal{F}} \mathbb{E}\left[(a_0(X) - a(X)) \cdot f(X) - f(X)^2\right] \\ = & \frac{1}{4} \mathbb{E}\left[(a_0(X) - a(X))^2\right] =: \frac{1}{4} \|\hat{a} - a_0\|_2^2 \end{aligned}$$

Thus our empirical criterion converges to the mean-squared-error criterion in the population limit, even though we don't have access to unbiased samples from $a_0(X)$.

[4] In Appendix A, we examine the relationship between $\mathcal{G}$ and the Riesz representer space $\mathcal{A}$.

Remark 2 (Norm-Based Regularization)**.** *The extra vanishing norm-based regularization can be avoided if one knows a bound on the norm of the true* a_0*. In that case, one can impose a hard norm constraint on the hypothesis space* $\mathcal{A}$ *and* $\bar{\mathcal{A}}$ *and optimize over these norm-constrained subspaces. However, regularization allows the estimator to be adaptive to the true norm of* a_0*, without knowledge of it.*

Remark 3 (Mis-specification)**.** *We in fact allow for* $a_0 \notin \mathcal{A}$*, and incur an extra bias part in our estimation error of the form of:* $\min_{a\in\mathcal{A}} \|a - a_0\|_2$*. Thus* $\mathcal{A}$ *need only be an* ℓ_2*-norm approximating sequence of function spaces.*

3 Fast Convergence Rate

We now provide fast convergence rates of our regularized minimax estimator, parameterized by the critical radii of the function classes:

$$\begin{aligned} \mathcal{F}_B &:= \{f \in \mathcal{F} : \|f\|_{\mathcal{A}}^2 \le B\} \\ m \circ \mathcal{F}_B &:= \{m(\cdot; f) : f \in \mathcal{F}_B\} \end{aligned}$$

for some appropriately defined constant B. The critical radius of a function class $\mathcal{F}$ with range in $[-1, 1]$ is defined as any solution δ_n to the inequality:

$$\mathcal{R}(\delta; \mathcal{F}) \le \delta^2 \qquad \text{with: } \mathcal{R}(\delta; \mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}: \|f\|_2 \le \delta} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right]$$

with $\epsilon_{1:n}$ are independent Rademacher random variables drawn equiprobably in $\{-1, 1\}$. For VC-subgraph function classes with constant VC dimension the critical radius is of the order of $\sqrt{\log(n)/n}$. The critical radius has been characterized by many other function classes such as reproducing kernel Hilbert spaces, neural networks and high-dimensional linear functions (c.f. [130] and Section 4).

We will also require the following norm-dominance condition:

ASSUMPTION 1 (Mean-Squared Continuity)**.** *For some constant* $M \ge 0$*, the following property holds:*

$$\forall f \in \mathcal{F} : \sqrt{\mathbb{E}\left[m(Z; f)^2\right]} \le M \|f\|_2$$

Observe that the fact that the operator $\theta(g)$ is bounded, implies that $|\mathbb{E}[m(Z; g)]| \le M \|g\|_2$. Mean-squared continuity is a stronger condition than boundedness, since: $|\mathbb{E}[m(Z; g)]| \le \mathbb{E}[|m(Z; g)|] \le \sqrt{\mathbb{E}[m(Z; g)^2]}$. In Appendix B.1, we verify this condition for a variety of popular functionals.

Example 1 (Mean-Squared Continuity for ATE)**.** *Let* $X = (D, W)$ *consist of treatment and covariates. In the case of treatment effect estimation, the above is implied by a non-parametric overlap condition, i.e.* $\Pr[D = 1 \mid w] \in (1/M, 1 - 1/M)$ *for some* $M \in (1, \infty)$*. Then observe that:*

$$\begin{aligned} \mathbb{E}[(g(1, W) - g(0, W))^2] \le{}& 2\mathbb{E}[g(1, W)^2 + g(0, W)^2] \\ \le{}& 2M\mathbb{E}\left[\Pr[D = 1 \mid W]\, g(1, W)^2 + \Pr[D = 0 \mid W]\, g(0, W)^2\right] = 2M\|g\|_2^2 \end{aligned}$$

Theorem 1. *Assume that mean-squared continuity holds for some constant $M \geq 1$ and that for some $B \geq 0$, the functions in $\mathcal{F}_B$ and $m \circ \mathcal{F}_B$ have uniformly bounded ranges in $[-1, 1]$. Let:*

$$\delta := \delta_n + \epsilon_n + c_0 \sqrt{\frac{\log(c_1/\zeta)}{n}},$$

for universal constants c_0, c_1, where δ_n upper bounds the critical radii of $\mathcal{F}_B, m \circ \mathcal{F}_B$ and ϵ_n upper bounds the bias $\min_{a \in \mathcal{A}} \|a - a_0\|_2$. Let $a_* = \arg\min_{a \in \mathcal{A}} \|a - a_0\|_2$. Then the estimator in Equation (1), with $\mu \geq 6\lambda \geq 12\delta^2/B$, satisfies w.p. $1 - \zeta$:

$$\|\hat{a} - a_0\|_2 \leq O\left(M^2 \delta + \frac{\mu}{\delta} \|a_*\|_{\mathcal{A}}^2\right)$$

For $\mu \leq C\delta^2/B$, for some constant C, the latter is: $O\left(\delta \max\left\{M^2, \frac{\|a_*\|_{\mathcal{A}}^2}{B}\right\}\right)$.

Remark 4. *Suppose we only want to approximate the Riesz representer with respect to the weaker distance metric $\|\cdot\|_{\mathcal{F}}$ defined as:*[5]

$$\|a\|_{\mathcal{F}}^2 = \sup_{f \in \mathcal{F}} \langle a, f \rangle_2 - \frac{1}{4} \|f\|_2^2 \leq \|a\|_2^2$$

Then Theorem 1 can be adapted to show that: $\|\hat{a} - a_0\|_{\mathcal{F}} \leq \delta \max\left\{M^2, \|a_*\|_{\mathcal{A}}^2/B\right\}$, where now the approximation rate is $\epsilon_n = \inf_{a \in \mathcal{A}} \|a - a_0\|_{\mathcal{F}}$. Observe that $\|\cdot\|_{\mathcal{F}}$ satisfies:

$$\|a\|_{\mathcal{F}}^2 = \inf_{f \in \mathcal{F}} \frac{1}{4}\|f\|_2^2 - \langle a, f \rangle_2 + \|a\|_2^2 - \|a\|_2^2 = \inf_{f \in \mathcal{F}} \|a - f/2\|_2^2 - \|a\|_2^2 \leq \inf_{f \in \mathcal{F}} \|a - f\|_2^2 - \|a\|_2^2$$

where in the last inequality we used the fact that $\mathcal{F}$ is star-convex. Thus it is at most the projection of a on $\mathcal{F}$. Hence, it is sufficient that $\mathcal{A}$ approximates a_0 in this weak sense that for some $a_* \in \mathcal{A}$ the projection of $a_* - a_0$ on $\mathcal{F}$ is at most ϵ_n. Thus any component of a_0 that is orthogonal to $\mathcal{F}$ can be ignored, since if we denote with $a_0 = a_0^{\perp} + a_0^{\|}$ with $\sup_{f \in \mathcal{F}} \langle a_0^{\perp}, f \rangle_2 = 0$, then $\|a_0 - a_*\|_{\mathcal{F}} = \|a_0^{\|} - a_*\|_{\mathcal{F}}$.

Our proof uses similar ideas as in the proof of Theorem 1 of [45], where an adversarial estimator was considered for the case of non-parametric instrumental variable regression. Theorem 1 of [45] provides bounds on a weaker metric than the mean-squared-error metric and requires bounds on the critical radius of more complicated function spaces.

As a corollary of Theorem 1, we can obtain a bound for the un-regularized estimator with $\lambda = \mu = 0$, where the function classes $\mathcal{F}$ and $\mathcal{G}$ are already norm constrained, e.g. $\|f\|_{\mathcal{A}} \leq U$ for all $f \in \mathcal{F}$, which also implies that $\|a\|_{\mathcal{A}} \leq U$ for all $a \in \mathcal{A}$, such that functions in $\mathcal{F}$ and $\mathcal{G}$ have uniformly bounded range. This can be achieved by using the above norm-constrained definitions of $\mathcal{F}$ and $\mathcal{G}$ and taking the limit of Theorem 3 when $B \to \infty$. In that case, $\mathcal{F}_B \to \mathcal{F}$, $m \circ \mathcal{F}_B \to m \circ \mathcal{F}$ and λ, μ are allowed to take zero value. This leads to the corollary:

[5]The metric $\|\cdot\|_{\mathcal{F}}$ satisfies the triangle inequality:

$$\|a + b\|_{\mathcal{F}} \leq \sqrt{\sup_{f \in \mathcal{F}} \langle a, f \rangle - \frac{1}{4}\|f\|_2^2 + \sup_{f \in \mathcal{F}} \langle b, f \rangle - \frac{1}{4}\|f\|_2^2} \leq \|a\|_{\mathcal{F}} + \|b\|_{\mathcal{F}}$$

and is positive definite i.e. $\|0\|_{\mathcal{F}} = 0$, but not necessarily homogeneous, i.e. $\|\lambda a\|_{\mathcal{F}} =?|\lambda|\|a\|_{\mathcal{F}}$ for $\lambda \in \mathbb{R}$.

Corollary 2. *Assume that mean-squared continuity holds for some constant $M \geq 1$ and that the functions in $\mathcal{F}$ and $m \circ \mathcal{F}$ have uniformly bounded ranges in $[-1, 1]$. Let:*

$$\delta := \delta_n + \epsilon_n + c_0 \sqrt{\frac{\log(c_1/\zeta)}{n}},$$

for universal constants c_0, c_1, where δ_n upper bounds the critical radii of $\mathcal{F}, m \circ \mathcal{F}$ and ϵ_n upper bounds the bias $\min_{a \in \mathcal{A}} \|a - a_0\|_2$. The estimator in Equation (1), with $\lambda = \mu = 0$, satisfies:

$$\|\hat{a} - a_0\|_2 \leq O\left(M^2 \delta\right)$$

3.1 Fast Rates without ℓ_2-Penalty

We will use the following notation:

$$\operatorname{span}_\kappa(\mathcal{F}) := \left\{ \sum_{i=1}^{p} w_i f_i : f_i \in \mathcal{F}, \|w\|_1 \leq \kappa, p \leq \infty \right\}$$

Theorem 3. *Consider a set of test functions $\mathcal{F} := \cup_{i=1}^{d} \mathcal{F}^i$, that is de-composable as a union of d symmetric test function spaces $\mathcal{F}^i$ and suppose that $\mathcal{A}$ is star-convex. Consider the adversarial estimator:*

$$\hat{a} = \arg\min_{a \in \mathcal{A}} \sup_{f \in \mathcal{F}} \mathbb{E}_n[m(Z; f) - a(X) \cdot f(X)] + \lambda \|a\|_{\mathcal{A}} \tag{2}$$

Let $m \circ \mathcal{F}^i = \{m(\cdot; f) : f \in \mathcal{F}^i\}$ and

$$\delta_{n,\zeta} := 2 \max_{i=1}^{d} \left(\mathcal{R}(\mathcal{F}^i) + \mathcal{R}(m \circ \mathcal{F}^i)\right) + c_0 \sqrt{\frac{\log(c_1\, d/\zeta)}{n}},$$

for some universal constants c_0, c_1 and $B_{n,\lambda,\zeta} := \left(\|a_0\|_{\mathcal{A}} + \delta_{n,\zeta}/\lambda\right)^2$. Suppose that $\lambda \geq \delta_{n,\zeta}$ and:

$$\forall a \in \mathcal{A}_{B_{n,\lambda,\zeta}} \text{ with } \|a - a_0\|_2 \geq \delta_{n,\zeta} : \tfrac{a - a_0}{\|a - a_0\|_2} \in \operatorname{span}_\kappa(\mathcal{F})$$

Then $\hat{a}$ satisfies that w.p. $1 - \zeta$:

$$\|\hat{a} - a_0\|_2 \leq \kappa \left(2 \left(\|a_0\|_{\mathcal{A}} + 1\right) \mathcal{R}(\mathcal{A}_1) + \delta_{n,\zeta} + \lambda \left(\|a_0\|_{\mathcal{A}} - \|\hat{a}\|_{\mathcal{A}}\right)\right)$$

4 Example Function Spaces

We now instantiate our two main theorems for several function classes of interest. Throughout this section we will use the following convenient characterization of the critical radius of a function class. Corollary 14.3 and Proposition 14.25 of [130] imply that the critical radius of any function class $\mathcal{F}$, uniformly bounded in $[-b, b]$, is of the same order as any solution to the inequality:

$$\frac{64}{\sqrt{n}} \int_{\frac{\delta^2}{2b}}^{\delta} \sqrt{\log\left(N_n(\epsilon; B_n(\delta; \mathcal{F})\right)} d\epsilon \leq \frac{\delta^2}{b} \tag{3}$$

where $B_n(\delta; \mathcal{F}) = \{f \in \mathcal{F} : \|f\|_{2,n} \leq \delta\}$ and $N_n(\epsilon; \mathcal{F})$ is the empirical ℓ_2-covering number at approximation level ϵ, i.e. the size of the smallest ϵ-cover of $\mathcal{F}$, with respect to the empirical ℓ_2 metric.

4.1 Sparse Linear Functions

Consider the class of s-sparse linear function classes in p dimensions, with bounded coefficients, i.e.,

$$\mathcal{A}_{\text{splin}} := \{x \to \langle \theta, x\rangle : \|\theta\|_0 \leq s, \|\theta\|_\infty \leq b\},$$

then observe that $\mathcal{F}$ is also the class of s-sparse linear functions, with bounded coefficients in $[-2b, 2b]$. Moreover, suppose that the ℓ_1-norm of the covariates x is bounded. The critical radius δ_n is of order $O\left(\sqrt{\frac{s\log(p\,n)}{n}}\right)$. It is easy to see that the ϵ-covering number of such a function class is of order $N_n(\epsilon; \mathcal{F}) = O\left(\binom{p}{s}\left(\frac{b}{\epsilon}\right)^s\right) \leq O\left(\left(\frac{p\,b}{\epsilon}\right)^s\right)$, since it suffices to choose the support of the coefficients and then place a uniform ϵ-grid on the support. Thus we get that Equation (3) is satisfied for $\delta = O\left(\sqrt{\frac{s\log(p\,b)\,\log(n)}{n}}\right)$. Moreover, observe that if $m(Z; f)$ is L-Lipschitz in f with respect to the ℓ_∞ norm, then the covering number of $m \circ \mathcal{F}$ is also of the same order. Thus we can apply Corollary 2 to get:

Corollary 4 (Sparse Linear Riesz Representer). *The estimator presented in Corollary 2, with $\mathcal{A} = \mathcal{A}_{splin}$, satisfies w.p. $1 - \zeta$:*

$$\|\hat{a} - a_0\|_2 \leq O\left(\min_{a \in A_{splin}} \|a - a_0\|_2 + \sqrt{\frac{s\log(p\,b)\,\log(n)}{n}} + \sqrt{\frac{\log(1/\zeta)}{n}}\right)$$

The latter Theorem required a hard sparsity constraint. However, our second main theorem, Theorem 3, allows us to prove a similar guarantee for the relaxed version of ℓ_1-bounded high-dimensional linear function classes. For this corollary we require a restricted eigenvalue condition which is typical for such relaxations.

Corollary 5 (Sparse Linear Riesz Representer with Restricted Eigenvalue). *Suppose that $a_0(x) = \langle \theta_0, x\rangle$ with $\|\theta_0\|_0 \leq s$ and $\|\theta_0\|_1 \leq B$ and $\|\theta_0\|_\infty \leq 1$. Moreover, suppose that the covariance matrix $V = \mathbb{E}[xx']$ satisfies the restricted eigenvalue condition:*

$$\forall \nu \in \mathbb{R}^p \;\; s.t. \;\; \|\nu_{S^c}\|_1 \leq \|\nu_S\|_1 + \delta_{n,\zeta}/\lambda : \nu^\top V \nu \geq \gamma \|\nu\|_2^2$$

Let $\mathcal{A} = \{x \to \langle \theta, x\rangle : \theta \in \mathbb{R}^p\}$, $\|\langle \theta, \cdot\rangle\|_{\mathcal{A}} = \|\theta\|_1$, and $\mathcal{F} = \{x \to \xi x_i : i \in [p], \xi \in \{-1, 1\}\}$. Then the estimator presented in Equation (2) with $\lambda \leq \frac{\gamma}{8s}$, satisfies that w.p. $1 - \zeta$:

$$\|\hat{a} - a_0\|_2 \leq O\left(\max\left\{1, \frac{1}{\lambda}\frac{\gamma}{s}\right\}\sqrt{\frac{s}{\gamma}}\left((\|\theta_0\|_1 + 1)\sqrt{\frac{\log(p)}{n}} + \sqrt{\frac{\log(p/\zeta)}{n}}\right)\right)$$

Remark 5 (Restricted Eigenvalue). *We note that if we have that the unrestricted minimum eigenvalue of V is at least γ, then the restricted eigenvalue condition always holds. Moreover, observe that we only require a condition on the population covariance matrix V and not on the empirical covariance matrix.*

4.2 Neural Networks

Suppose that the function class $\mathcal{A}$ can be represented as a RELU activation neural network with depth L and width W, denoted as $\mathcal{A}_{\text{nnet}(L,W)}$. Then observe that functions in $\mathcal{F}$ can be represented as neural networks with depth $L+1$ and width $2W$. Moreover, we assume that functions in $m \circ \mathcal{F}$ are also representable by neural networks of depth $O(L)$ and width $O(W)$. Finally, suppose that the covariates are distributed in a way that the outputs of $\mathcal{F}$ and $m \circ \mathcal{F}$ are uniformly bounded in $[-b, b]$.

Then by the L_1 covering number for VC classes of [61], the bounds of theorem 14.1 of [7] and Theorem 6 of [13], one can show that the critical radius of $\mathcal{F}$ and $m \circ \mathcal{F}$ is of the order of $\delta_n = O\left(\sqrt{\frac{L\,W\,\log(W)\,\log(b)\,\log(n)}{n}}\right)$ (c.f. Proof of Example 3 of [51] for a detailed derivation). Thus we can apply Corollary 2 to get:

Corollary 6 (Neural Network Riesz Representer)**.** *Suppose that $\mathcal{A} = \mathcal{A}_{nnet(L,W)}$, and that $m \circ \mathcal{F}$ is representable as a neural network with depth $O(L)$ and width $O(W)$. Moreover, the input covariates are such that functions in $\mathcal{F}$ and $m \circ \mathcal{F}$ are uniformly bounded in $[-b, b]$. Then the estimator presented in Corollary 2, satisfies w.p. $1-\zeta$:*

$$\|\hat{a} - a_0\|_2 \leq O\left(\min_{a \in A_{nnet(L,W)}} \|a - a_0\|_2 + \sqrt{\frac{L\,W\,\log(W)\,\log(b)\,\log(n)}{n}} + \sqrt{\frac{\log(1/\zeta)}{n}}\right)$$

If the true Riesz representer a_0 is representable as a RELU neural network, then the first term vanishes and we achieve an almost parametric rate. For non-parametric Holder function classes, one can easily combine the latter corollary with approximation results for RELU activation neural networks presented in [131, 132]. These approximation results typically require that the depth and the width of the neural network grow as some function of the approximation error ϵ, leading to errors of the form: $O\left(\epsilon + \sqrt{\frac{L(\epsilon)\,W(\epsilon)\,\log(W(\epsilon))\,\log(b)\,\log(n)}{n}} + \sqrt{\frac{\log(1/\zeta)}{n}}\right)$. Optimally balancing ϵ then typically leads to almost tight non-parametric rates, of the same order as those presented in Theorem 1 of [48].

4.3 Reproducing Kernel Hilbert Spaces

Suppose that a_0 lies in a Reproducing Kernel Hilbert Space (RKHS) with kernel K, denoted as $\mathcal{A}_{\text{rkhs}(K)}$ and with the norm $\|\cdot\|_{\mathcal{A}}$ being the RKHS norm. Then observe that $\mathcal{F}$ is the same function space. Moreover, we assume that $m \circ \mathcal{F}$ also lies in an RKHS with a potentially different kernel $\tilde{K}$. Finally, suppose that the input covariates are such that for some constant B, functions in $\mathcal{F}_B$ and $m \circ \mathcal{F}_B$ are bounded in $[-1, 1]$.

Let $\{\hat{\lambda}_j\}_{j=1}^n$ be the eigenvalues of the $n \times n$ empirical kernel matrix, with $K_{ij} = K(x_i, x_j)/n$. Similarly, let $\{\hat{\mu}_j\}_{j=1}^n$ be the eigenvalues of the empirical kernel matrix $\tilde{K}$. Then by Corollary 13.18 of [130], we can derive the following corollary of Theorem 1:

Corollary 7 (RKHS Riesz Representer)**.** *Suppose that $\mathcal{A} = \mathcal{A}_{rkhs}$, $a_0 \in \mathcal{A}_{rkhs}$, and that $m \circ \mathcal{F} \in \mathcal{A}_{rkhs(\tilde{K})}$. Let $\{\hat{\lambda}_j\}_{j=1}^n$ and $\{\hat{\mu}_j\}_{j=1}^n$ be the egienvalues of the empirical kernel matrices of K and*

$\tilde{K}$*, correspondingly. Let* δ_n *be any solution to the inequalities:*

$$B\sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^{\infty}\max\{\hat{\lambda}_j,\delta^2\}} \leq \delta^2 \qquad\qquad B\sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^{\infty}\max\{\hat{\mu}_j,\delta^2\}} \leq \delta^2$$

Moreover, the input covariates are such that functions in $\mathcal{F}_B$ *and* $m \circ \mathcal{F}_B$ *are uniformly bounded in* $[-1,1]$*. Then the estimator presented in Theorem 1, satisfies w.p.* $1-\zeta$*:*

$$\|\hat{a}-a_0\|_2 \leq O\left(\|a_0\|_{\mathcal{A}}\left(\delta_n + \sqrt{\frac{\log(1/\zeta)}{n}}\right)\right)$$

We note that the latter estimator does not need to know the RKHS norm of the true function a_0. Instead it automatically adapts to the unknown RKHS norm. Moreover, note that the bound δ_n is solely based on empirically observable quantities, as it is a function of the empirical eigenvalues. Thus these empirical quantities can be used as a data-adaptive diagnostic of the error.

Finally, we note that for particular kernels a more explicit bound can be derived as a function of the eigendecay. For instance, for the Gaussian kernel, which has an exponential eigendecay, Example 13.21 of [130] derives that the solution to the eigenvalue inequality scales as $O\left(\sqrt{\frac{\log(n)}{n}}\right)$, thus leading to almost parametric rates: $\|\hat{a}-a_0\|_2 \leq O\left(\|a_0\|_{\mathcal{A}}\sqrt{\frac{\log(n)}{n}}\right)$.

5 Computation

In this section we discuss computational aspects of the optimization problem implied by our adversarial estimator. We show how in many cases, the min-max optimization problem can be solved computationally efficiently and also discuss practical heuristics for cases where the problem is non-convex (e.g. in the case of neural networks).

5.1 Sparse Linear Function Spaces

For the case of sparse linear functions, the estimator in Theorem 3 requires solving the following optimization problem:

$$\min_{\theta\in\mathbb{R}^p:\|\theta\|_1\leq B}\ \max_{i\in[2p]} \mathbb{E}_n\left[m(Z;f_i) - f_i(X)\langle\theta,X\rangle\right] + \lambda\|\theta\|_1 \tag{4}$$

where $f_i(X) = X_i$ for $i \in \{1,\dots,p\}$ and $f_i(X) = -X_i$ for $i \in \{p+1,\dots,2p\}$. This can be solved via sub-gradient descent, which would yield an ϵ-approximate solution after $O\left(p/\epsilon^2\right)$ steps. This can be improved to $O\left(\log(p)/\epsilon\right)$ steps if one views it as a zero-sum game and uses simultaneous gradient descent, where the θ-player uses Optimistic-Follow-the-Regularized-Leader with an entropic regularizer and the f-player uses Optimistic Hedge over probability distributions on the finite set of test functions (analogous to Proposition 13 of [45]).

To present the algorithm it will be convenient to re-write the problem where the maximizing player optimizes over distributions in the $2p$-dimensional simplex, i.e.:

$$\min_{\theta\in\mathbb{R}^p:\|\theta\|_1\leq B}\ \max_{w\in\mathbb{R}^{2p}_{\geq 0}:\|w\|_1=1} \mathbb{E}_n\left[m(Z;\langle w,f\rangle) - \langle w,f\rangle(X)\,\langle\theta,X\rangle\right] + \lambda\|\theta\|_1$$

where $f = (f_1,\ldots,f_{2p})$, denote the vector of the $2p$ functions. Moreover, to avoid the non-smoothness of the ℓ_1 penalty it will be convenient to introduce the augmented vector $V = (X;-X)$ and for the minimizing player to optimize over the positive orthant of a $2p$-dimensional vector $\rho = (\rho^+;\rho^-)$, with an ℓ_1 bounded norm, such that in the end: $\theta = \rho^+ - \rho^-$. Then we can re-write the problem as:

$$\min_{\rho\in\mathbb{R}^{2p}_{\geq 0}:\|\rho\|_1\leq B}\ \max_{w\in\mathbb{R}^{2p}_{\geq 0}:\|w\|_1=1} \mathbb{E}_n\left[m(Z;\langle w,f\rangle) - \langle w,V\rangle\,\langle\rho,V\rangle\right] + \lambda\sum_{i=1}^{2p}\rho_i$$

where we also noted that $\langle w,f\rangle(X) = \langle w,V\rangle$.

Proposition 8. *Consider the algorithm that for $t = 1,\ldots,T$, sets:*

$$\begin{aligned}
\tilde{\rho}_{i,t+1} &= \tilde{\rho}_{i,t}e^{-2\frac{\eta}{B}(-\mathbb{E}_n[V_i\langle V,w_t\rangle]+\lambda)+\frac{\eta}{B}(-\mathbb{E}_n[V_i\langle V,w_{t-1}\rangle]+\lambda)} & \rho_{t+1} &= \tilde{\rho}_{t+1}\min\left\{1,\frac{B}{\|\tilde{\rho}_{t+1}\|_1}\right\}\\
\tilde{w}_{i,t+1} &= w_{i,t}e^{2\eta\,\mathbb{E}_n[m(Z;f_i)-V_i\langle V,\rho_t\rangle]-\eta\,\mathbb{E}_n[m(Z;f_i)-V_i\langle V,\rho_{t-1}\rangle]} & w_{t+1} &= \frac{\tilde{w}_{t+1}}{\|\tilde{w}_{t+1}\|_1}
\end{aligned}$$

with $\tilde{\rho}_{i,-1} = \tilde{\rho}_{i,0} = 1/e$ and $\tilde{w}_{i,-1} = \tilde{w}_{i,0} = 1/(2p)$ and returns $\bar{\rho} = \frac{1}{T}\sum_{t=1}^{T}\rho_t$. Then for $\eta = \frac{1}{4\|\mathbb{E}_n[VV^\top]\|_\infty}$,[6] *after*

$$T = 16\|\mathbb{E}_n[VV^\top]\|_\infty\frac{4B^2\log(B\vee 1)+(B+1)\log(2p)}{\epsilon}$$

iterations, the parameter $\bar{\theta} = \bar{\rho}^+ - \bar{\rho}^-$ is an ϵ-approximate solution to the minimax problem in Equation (4).

5.2 Neural Nets with Simultaneous Stochastic Gradient Descent

When the function space $\mathcal{A}$ and $\mathcal{F}$ is represented as a deep neural network then the optimization problem is highly non-convex. This is the case even if we were just solving a square loss minimization problem. On top of this we also need to deal with the non-convexity and non-smoothness introduced by the min-max structure of our estimator.

Luckily, the optimization problem that we are facing is similar to the optimization problem that is encountered in training Generative Adversarial Networks, i.e. we need to solve a non-convex, non-concave zero-sum game, where the strategy of each of the two players are the parameters of a neural net. Luckily, there has been a surge of recent work proposing iterative optimization algorithms inspired by the convex-concave zero-sum game theory (see, e.g. the Optimistic Adam algorithm of [43], also utilized in the recent work of [20, 45] in the context of solving moment

[6] For a matrix A, we denote with $\|A\|_\infty = \max_{i,j}|A_{ij}|$

equations, or the work of [64, 87] on the extra-gradient or stochastic extra-gradient algorithm). All these new algorithms for solving differentiable non-convex/non-concave zero-sum games can be deployed for our problem.

Recent work of [83] contributes to a literature on over-parameterized neural network training for square losses [6, 46, 119]. The authors show that even for min-max losses that are very similar to the loss of our estimator, neural nets that are sufficiently wide and appropriately randomly initialized essentially behave like linear functions in an appropriate reproducing kernel Hilbert space, typically referred to as the neural tangent kernel space. Given this intuition, the authors show that a simple simultaneous gradient descent/ascent algorithm and subsequent averaging of the parameters converges to the solution of the min-max problem. In this regime neural networks behave like linear functions, so one can invoke analysis similar to the analysis we invoke for sparse linear function spaces, and then carefully account for the approximation error. The intuition and results of the work of [83] can be appropriately adapted for our loss function too so as to show that the average path of the simultaneous gradient descent/ascent algorithm also converges in our setting. One caveat is that growing the width of the neural net to facilitate optimization deteriorates the statistical guarantee, since the critical radius grows as a function of the width.

5.3 Reproducing Kernel Hilbert Space

Recall the estimator is

$$\hat{a} = \arg\min_{a\in\mathcal{A}} \max_{f\in\mathcal{F}} \mathbb{E}_n[m(Z;f) - a(X)\cdot f(X)] - \|f\|_{2,n}^2 - \lambda\|f\|_{\mathcal{A}}^2 + \mu\|a\|_{\mathcal{A}}^2$$

In this section, we derive a closed form solution for $\hat{a}$ that can be computed from matrix operations.

Towards this end, we impose additional structure on the problem. If $\mathcal{G} = \mathcal{F} = \mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), then the projection $a_0^{\min}$ of any RR a_0 into $\mathcal{G}$ is clearly an element of $\mathcal{H}$ as well, so we can take $\mathcal{A} = \mathcal{H}$. Also assume that the functional m satisfies $m(z;f) = m(x;f)$. Moreover, let the functional be such that it evaluates the function in some arguments. For example, in ATE, $m(z;f) = f(1,w) - f(0,w)$ where $z = x = (d,w)$. This property holds for treatment effects and policy effects, and it ensures that $m(\cdot;f) \in \mathcal{H}$.

Denote the kernel $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, and denote the feature map $\phi : x \mapsto k(x,\cdot)$. Denote the kernel matrix K_{XX} with (i,j)-th entry $k(x_i,x_j)$. Denote the feature matrix Φ with i-th row $\phi(x_i)'$. Hence $K_{XX} = \Phi\Phi'$.

By the reproducing property $f(x) = \langle f, \phi(x)\rangle_{\mathcal{H}}$. Moreover, since m is a linear functional, we can define the linear operator $M : \mathcal{H} \to \mathcal{H},\ f(\cdot) \mapsto m(\cdot;f)$ whereby

$$m(x;f) = [Mf](x) = \langle Mf, \phi(x)\rangle_{\mathcal{H}} = \langle f, M^*\phi(x)\rangle_{\mathcal{H}}$$

where M^* is the adjoint of M. Define the matrix $\Phi^{(m)} := \Phi M$ with i-th row $\phi(x_i)'M$. Finally define Ψ as the matrix with $2n$ rows that is constructed by concatenating Φ and $\Phi^{(m)}$. We denote the induced kernel matrix by $K := \Psi\Psi'$. Formally,

$$\Psi := \begin{bmatrix} \Phi \\ \Phi^{(m)} \end{bmatrix}, \quad K := \begin{bmatrix} K^{(1)} & K^{(2)} \\ K^{(3)} & K^{(4)} \end{bmatrix} := \begin{bmatrix} \Phi\Phi' & \Phi(\Phi^{(m)})' \\ \Phi^{(m)}\Phi' & \Phi^{(m)}(\Phi^{(m)})' \end{bmatrix}$$

Note that $\{K^{(j)}\}_{j\in[4]} \in \mathbb{R}^{n\times n}$ and hence $K \in \mathbb{R}^{2n\times 2n}$ can be computed from data, though they depend on the choice of moment.

Proposition 9 (Computing kernel matrices). *For example, for ATE*

$$\begin{aligned}
[K^{(1)}]_{ij} &= k((d_i, w_i), (d_j, w_j)) \\
[K^{(2)}]_{ij} &= k((1, w_i), (d_j, w_j)) - k((0, w_i), (d_j, w_j)) \\
[K^{(3)}]_{ij} &= k((d_i, w_i), (1, w_j)) - k((d_i, w_i), (0, w_j)) \\
[K^{(4)}]_{ij} &= k((1, w_i), (1, w_j)) - k((1, w_i), (0, w_j)) - k((0, w_i), (1, w_j)) + k((0, w_i), (0, w_j))
\end{aligned}$$

We proceed in steps. First we prove the existence of a closed form for the maximizer $\hat{f} = \arg\max_{f\in\mathcal{H}} \mathbb{E}_n[m(X; f) - a(X)\cdot f(X)] - \|f\|^2_{2,n} - \lambda\|f\|^2_{\mathcal{H}}$ by extending the classic representation theorem of [76, 114].

Proposition 10 (Representation of maximizer). $\hat{f} = \Psi'\hat{\gamma}$ *for some* $\hat{\gamma} \in \mathbb{R}^{2n}$

Appealing to this abstract result, we derive the closed form expression for the maximizer in terms of kernel matrices.

Proposition 11 (Closed form of maximizer). $\hat{\gamma} = \frac{1}{2}\Delta^{-1}\left[n\Psi M'\hat{\mu} - \begin{bmatrix} K^{(1)} \\ K^{(3)} \end{bmatrix} \Phi a\right]$ *where*

$$\Delta := \begin{bmatrix} K^{(1)}K^{(1)} & K^{(1)}K^{(2)} \\ K^{(3)}K^{(1)} & K^{(3)}K^{(2)} \end{bmatrix} + n\lambda K \in \mathbb{R}^{2n\times 2n}, \quad \hat{\mu} := \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$$

Next we prove the existence of a closed form for the minimizer $\hat{a} = \arg\min_{a\in\mathcal{H}} \mathbb{E}_n[m(X; \hat{f}) - a(X)\cdot \hat{f}(X)] - \|\hat{f}\|^2_{2,n} - \lambda\|\hat{f}\|^2_{\mathcal{H}} + \mu\|a\|^2_{\mathcal{H}}$ by appealing to the classic representation theorem of [76, 114].

Proposition 12 (Representation of minimizer). $\hat{a} = \Phi'\hat{\beta}$ *for some* $\hat{\beta} \in \mathbb{R}^n$

Again, with this abstract result in hand, we derive the closed form expression for the minimizer in terms of kernel matrices.

Proposition 13 (Closed form of minimizer). $\hat{\beta} = \left\{\frac{1}{n}\Omega\Delta^{-1}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix} + 2\mu\cdot K^{(1)}\right\}^{-1} \Omega\Delta^{-1}\Psi M'\hat{\mu}$ *where*

$$\Omega := \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' - \frac{1}{2}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}\begin{bmatrix} K^{(1)}K^{(1)} & K^{(1)}K^{(2)} \\ K^{(3)}K^{(1)} & K^{(3)}K^{(2)} \end{bmatrix} - \frac{n\lambda}{2}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}K \in \mathbb{R}^{n\times 2n}$$

For practical use, we require a way to evaluate the minimizer using only kernel operations. Evaluation directly follows from the closed form expression.

Corollary 14 (Evaluation of minimizer).

$$\hat{a}(x) = K_{xX}\left\{\frac{1}{n}\Omega\Delta^{-1}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix} + 2\mu\cdot K^{(1)}\right\}^{-1}\Omega\Delta^{-1}V$$

where $V \in \mathbb{R}^{2n}$ *is defined such that*

$$v_j = \begin{cases} \frac{1}{n}\sum_{i=1}^n [K^{(2)}]_{ji} & \text{if } j \in [n] \\ \frac{1}{n}\sum_{i=1}^n [K^{(4)}]_{ji} & \text{if } j \in \{n+1, ..., 2n\} \end{cases}$$

5.4 Oracle Based Training

Consider the estimator with $\lambda = \mu = 0$:

$$\min_{a\in\mathcal{A}} \max_{f\in\mathcal{F}} \mathbb{E}_n\left[m(Z;f) - a(X)\cdot f(X) - f(X)^2\right] =: \ell(a,f) \tag{5}$$

We can solve this optimization problem by treating it as a zero-sum game, where one player controls a and the other player controls f. Observe that the game is convex a (in fact linear) and concave in f. Thus, we can solve this zero-sum game by having the f-player run a no-regret algorithm at each period $t \in \{1,\ldots,T\}$ and the a-player best responding to the current choice of the f player.

Observe that for any fixed f, the best-response of the a-player is the solution to:

$$a_t = \arg\min_{a\in\mathcal{A}} -\mathbb{E}_n[a(X)\cdot f(X)] = \arg\max_{a\in\mathcal{A}} \mathbb{E}_n[a(X)\cdot f(X)]$$

In other words, the a-player wants to match the sign of the function f. Thus the best-response of the a-player is equivalent to a weighted classification oracle, where the label is $Y_i = \mathtt{sign}(f(X_i))$ and the weight is $w_i = |f(X_i)|$.

Finally, we need to solve the no-regret problem for the f player. If the function space $\mathcal{F}$ is a convex space,[7] then we can simply run the follow the leader (FTL) algorithm, where at every period the algorithm maximizes the empirical past reward:

$$f_t = \arg\max_{f\in\mathcal{F}} \mathbb{E}_n\left[m(Z;f) - \bar{a}_{<t}(X)\cdot f(X) - f(X)^2\right] = \arg\min_{f\in\mathcal{F}} -\ell(\bar{a}_{<t}, f)$$

where $\bar{a}_{<t} = \frac{1}{t-1}\sum_{\tau<t} a_\tau$.

Proposition 15. *Suppose that the empirical operator $\mathbb{E}_n[m(Z;\cdot)]$ is bounded with operator norm upper bounded by $M_n \geq 1$ and that the function class $\mathcal{F}$ is convex. Consider the algorithm where at each period $t \in \{1,\ldots,T\}$:*

$$f_t = \arg\max_{f\in\mathcal{F}} \ell(\bar{a}_{<t}, f) \qquad\qquad a_t = \arg\min_{a\in\mathcal{A}} \ell(a, f_t)$$

Then for $T = \Theta\left(\frac{M_n \log(1/\epsilon)}{\epsilon}\right)$, the function $a_* = \frac{1}{T}\sum_{t=1}^T a_t$ is an ϵ-approximate solution to the empirical minimax problem in Equation (5)*.*

The above algorithm requires a weighted classification oracle for the a-player and an oracle for the f-player that solves the problem $\max_{f\in\mathcal{F}} \ell(a,f)$, for any a.

Example 2 (f-player oracle for ATE)**.** *For the case of ATE this problem is:*

$$f_* = \arg\min_{f\in\mathcal{F}} \mathbb{E}_n\left[f(T,X)^2 + a(T,X)f(T,X) - f(1,X) + f(0,X)\right]$$

[7] i.e. if $f, f' \in \mathcal{F}$, then $\gamma f + (1-\gamma)f' \in \mathcal{F}$ for any $\gamma \in [0,1]$

6 Debiasing Average Moment

Suppose our goal is to estimate $\theta_0 = \theta(g_0)$, where $g_0 = \mathbb{E}[Y \mid X]$. We have access to an estimate $\hat{g}$ of g_0. We consider the de-biased moment:

$$m_a(Z; g) = m(Z; g) + a(X)' (Y - g(X))$$

For simplicity of exposition, we present the remainder of the section for the case of a single-valued regression function.

6.1 Asymptotic Normality with Sample Splitting

Consider the following cross-fitted estimate:

- Partition n samples into K folds $P_1, \ldots, P_K$
- For each partition, estimate $\hat{a}_k, \hat{g}_k$ based on all out-of-fold data.
- Construct estimate:

$$\hat{\theta} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in P_k} m_{\hat{a}_k}(Z_i; \hat{g}_k)$$

Lemma 16. *Suppose that $K = \Theta(1)$ and that:*

$$\forall k \in [K] : \sqrt{n}\, \mathbb{E}[(a_0(X) - \hat{a}_k(X))\, (\hat{g}_k(X) - g_0(X))] \to_p 0 \tag{6}$$

and that for some a_* and g_* (not necessarily equal to a_0 and g_0), we have that for all $k \in [K]$: $\|\hat{a}_k - a_*\|_2 \xrightarrow{L^2} 0$ and $\|\hat{g}_k - g_*\|_2 \xrightarrow{L^2} 0$. Assume that Condition 1 is satisfied and the variables $Y, g(X), a(X)$ are bounded a.s. for all $g \in \mathcal{G}$ and $a \in \mathcal{A}$.[8] *Then if we let $\sigma_*^2 := Var(m_{a_*}(Z; g_*))$*

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to_d N\left(0, \sigma_*^2\right)$$

A sufficient condition for Condition 6 is that $\sqrt{n}\|\hat{a} - a_0\|_2 \|\hat{g} - g_0\|_2 \to_p 0$, which is a condition on the product of the two RMSE rates. However, observe that Condition (6) is much weaker as it implies that our Riesz estimate $\hat{a}$ only needs to approximately satisfy the representer moment for test functions of the form: $\hat{g} - g_0$. Thus, if we assume that $\hat{g}$ satisfies an RMSE consistency rate that $\|\hat{g} - g_0\|_2 \leq r_n$, then it suffices that it satisfies the moment for any $g \in \mathcal{G}$, with $\|g - \hat{g}\|_2 \leq r_n$, i.e. it suffices that it is a local Riesz representer around $\hat{g}$. This can potentially make the Riesz estimation task much simpler than estimating a global Riesz representer. We formalize this observation in Appendix C.

Moreover, observe that the theorem does not require consistency of both nuisance functions. Only one of the two nuisance functions needs to be consistent, while the other must simply converge to some limit function. For instance, as long as $\sqrt{n}\|\hat{a} - a_0\|_2 \to 0$ or $\sqrt{n}\|\hat{g} - g_0\|_2 \to 0$, then the result

[8] This condition can be relaxed to simply assuming bounded fourth moments of $Y, g(X), a(X)$, as long as we strengthen the requirement to assume 4-th moment convergence to a_*, g_*, i.e. that $\|\hat{a}_k - a_*\|_4, \|\hat{g}_k - g_*\|_4 \to_p 0$.

holds. Inconsistency will only impact the limit variance, which will not be equal to the efficient variance; nonetheless, confidence intervals will be asymptotically valid. The required rate for the latter scenario is implausible as it asks for faster than root-n rate for either a or g. However, we can still show that the de-biased moment satisfies a double robustness property: if one nuisance is inconsistent, as long as the other is root-n consistent, then asymptotic normality of the causal parameter still holds. This result is presented in Appendix F.2. The result is analogous to the one provided in [19], where an estimator with such a property was presented within the targeted maximum likelihood framework.

6.2 Asymptotic Normality without Sample Splitting

Consider the algorithm where no cross-fitting or sample splitting is employed:

- Estimate $\hat{a}, \hat{g}$ on all the samples
- Construct estimate:
$$\hat{\theta} = \mathbb{E}_n \left[m_{\hat{a}}(Z; \hat{g})\right]$$

Lemma 17 (Normality via Localized Complexities). *Suppose that:*

$$\forall k \in [K] : \sqrt{n}\, \mathbb{E}[(a_0(X) - \hat{a}_k(X))\, (\hat{g}_k(X) - g_0(X))] \to_p 0 \tag{7}$$

and that for some a_* and g_* (not necessarily equal to a_0 and g_0), we have that: $\|\hat{a}_k - a_*\|_2, \|\hat{g}_k - g_*\|_2 = o_p(r_n)$. Assume that Condition 1 is satisfied and the variables $Y, g(X), a(X)$ are bounded a.s. for all $g \in \mathcal{G}$ and $a \in \mathcal{A}$. Moreover, assume that with high probability $\|\hat{g}\|_{\mathcal{G}} \leq B_1$ and $\|\hat{a}\|_{\mathcal{A}} \leq B_2$. Let $\delta_{n,*} = \delta_n + c_0\sqrt{\frac{\log(c_1\, n)}{n}}$ for some appropriately defined universal constants c_0, c_1, where δ_n is a bound on the critical radius of $\mathcal{G}_{B_1}$, $m \circ \mathcal{G}_{B_1}$ and $\mathcal{A}_{B_2}$ and also at least $\sqrt{\frac{\log\log(n)}{n}}$. If

$$\sqrt{n}\left(\delta_{n,*}\, r_n + \delta_{n,*}^2\right) \to 0$$

then if we let $\sigma_*^2 := Var(m_{a_*}(Z; g_*))$

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to_d N\left(0, \sigma_*^2\right)$$

Suppose we use for both $\hat{a}$ and $\hat{g}$ an ℓ_1 constrained linear function class in p dimensions and that a_*, g_* are sparse linear functions with support size s. Moreover if $B_1 = \|a_*\|_1 + o(1)$ and $B_2 = \|g_*\|_1 + o(1)$, and the covariates satisfy a restricted eigenvalue condition, then we could show that $\delta_{n,*} = O\left(\sqrt{\frac{s\log(p)}{n}}\right)$ (a simplification by assuming $s\log(p) > \log(n)$). Then as long as $r_n \to 0$, the condition is satisfied. Moreover, for such function classes, we will typically have that $r_n = O\left(\sqrt{\frac{s\log(p)}{n}}\right)$. Therefore, the required condition is that: $\frac{s\log(p)}{\sqrt{n}} = o(1)$ or equivalently $s = o\left(\sqrt{n}/\log(p)\right)$.

Of theoretical interest, it seems that without sample splitting, the analysis essentially goes through for general function classes that are not Donsker. With sample splitting, we would require from

Condition (8) that $\frac{\sqrt{s_a s_g}\log(p)}{n} = o(n^{-1/2})$, where s_a, s_g are the sparsity bounds on a and g, respectively. Simplifying, with sample splitting we require $\sqrt{s_a\, s_g} = o\left(\sqrt{n}/\log(p)\right)$. By contrast, without sample splitting, we require this condition for both s_a and s_g. Beyond this difference, the conditions on the sparsity of the function classes seem comparable.

We also provide a proof of asymptotic normality without sample splitting for uniformly stable estimators. This proof technique handles cases beyond Donsker classes or classes with small critical radius, since stability is not only a property of the function class but also of the estimation algorithm. Thus, it could be potentially apply to large neural net classes trained via few iterations of stochastic gradient descent [58] or sub-bagged ensembles of overfitting estimators [47].

Lemma 18 (Normality via Uniform Stability)**.** *Suppose that:*

$$\forall k \in [K] : \sqrt{n}\,\mathbb{E}[(a_0(X) - \hat{a}_k(X))\,(\hat{g}_k(X) - g_0(X))] \to_p 0 \tag{8}$$

and that for some a_* and g_* (not necessarily equal to a_0 and g_0), we have that:

$$\mathbb{E}\left[\|\hat{a}_k - a_*\|_2^2\right], \mathbb{E}\left[\|\hat{g}_k - g_*\|_2^2\right] = O(r_n^2)$$

Assume that Condition 1 is satisfied and the variables $Y, g(X), a(X)$ are bounded a.s. for all $g \in \mathcal{G}$ and $a \in \mathcal{A}$. Suppose that the algorithm for estimating $\hat{h} := (\hat{a}, \hat{g})$ is symmetric across samples and satisfies β_n-mean-squared stability, i.e.:[9]

$$\mathbb{E}_Z\left[\left\|\hat{h}(Z) - \hat{h}^{-i}(Z)\right\|_\infty^2\right] \le \beta_n$$

where $\hat{h}^{-i}$ is the function that the estimation algorithm would produce if sample i was removed from the training set. If

$$r_{n-1}^2 + n\,\beta_{n-1} r_{n-2} \to 0$$

then if we let $\sigma_*^2 := Var(m_{a_*}(Z; g_*))$

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to_d N\left(0, \sigma_*^2\right)$$

Uniform stability of sub-bagged ensemble estimators. If we use sub-bagging and return as an estimate the average of a base estimator over subsamples of size $s < n$, then the sub-bagged estimate is $\beta_n := \frac{s}{n}$-uniformly stable (see e.g. [47]). If the bias of the base estimator decays as some function $\textsc{bias}(s)$, then typically sub-bagged estimators will achieve $r_n = \sqrt{\frac{s}{n}} + \textsc{bias}(s)$ (see e.g. [10, 75, 121]). Thus we need that $n\beta_n r_n = \sqrt{\frac{s^3}{n}} + s\,\textsc{bias}(s) \to 0$. As long as $s = o(n^{1/3})$ and $\textsc{bias}(s) = o(1/s)$, then the conditions of the latter theorem hold. The recent work of [121] shows that in a high-dimensional regression setting, with $p \gg n$ and only $r \ll p, n$ of the variables being μ-*strictly* relevant variables, i.e. leading to a decrease in explained variance of at least μ, (for some constant $\mu > 0$), the bias of a deep Breiman tree trained on s data points decays as $\exp(-s)$. Moreover, a deep Breiman forest where each tree is trained on $s = O\left(\frac{2^r \log(p)}{\mu}\right) = o(n^{1/3})$ samples, drawn without replacement, will achieve $r_n = O\left(\sqrt{\frac{s2^r}{n}}\right)$. Thus sub-bagged deep Breiman random forests satisfy the conditions of the theorem in the case of sparse high-dimensional non-parametric regression.

[9] The notion was originally defined in [72] and used to derive imporved bounds on k-fold cross-validation. It is weaker than the well-studied uniform stability [26]. See [47, 28, 2] for more discussion.

7 Orthogonalizing Non-Linear Moment

Suppose our goal is to estimate the solution θ_0 to a non-linear moment problem that depends on a regression function g_0, i.e.:

$$\mathbb{E}[m(Z;\theta_0,g_0)] := 0$$

One way to construct a Neyman orthogonal moment that is robust to first-stage errors of the regression is to introduce a bias correction term that involves the Riesz representer of the functional derivative of the moment with respect to g, i.e.:

$$m_a(Z;\theta,g) = m(Z;\theta,g) + a(X)'\,(Y - g(X))$$

where $a_0(X)$ is the Riesz representer of the functional derivative of m with respect to g, i.e.:

$$f(g) := \frac{\partial}{\partial\tau}\mathbb{E}\left[m(Z;\theta,g_0+\tau\,(g-g_0))\right]\Big|_{\tau=0} = \mathbb{E}[a(X)'g(X)]$$

The Riesz representer a_0 can be estimated in a first stage as follows:

- Estimate the regression function $\hat{g}$
- Estimate a preliminary $\tilde{\theta}$ using the non-orthogonal moment condition
- Calculate algebraically, or through automatic differentiation, the Gateaux derivative function:

$$\hat{f}(g) = \frac{\partial}{\partial\tau}\mathbb{E}\left[m(Z;\tilde{\theta},\hat{g}+\tau\,(g-\hat{g}))\right]\Big|_{\tau=0}$$

- Apply the adversarial Riesz representer estimator for functional $\hat{f}(g)$, to estimate a

Following similar analysis as in Section 5 of [40], one can show that the moment m_a satisfies Neyman orthogonality. Moreover, assuming that the moment function is sufficiently smooth, the estimator outlined above will achieve faster than $n^{-1/4}$ rates. These two properties are sufficient to show that the estimator for θ, based on the orthogonal moment and using cross-fitting, will be root-n asymptotically normal.

One caveat of the approach outlined above is the burden of either calculating the Gateaux derivative algebraically or auto-differentiating the moment. One can bypass this difficult, and reduce to evaluation oracles of the moment, by taking arbitrarily small approximations of the Gateaux derivative. In particular, the third step could be replaced by defining:

$$\hat{f}_\epsilon(g) = \frac{1}{\epsilon}\left(\mathbb{E}\left[m(Z;\tilde{\theta},\hat{g}+\epsilon(g-g_0)) - m(Z;\tilde{\theta},\hat{g})\right]\right)$$

For sufficiently small ϵ, the approximation error $\|\hat{f}_\epsilon - \hat{f}\|$ is negligible. Moreover, $\hat{f}_\epsilon$ only requires black-box access to evaluations of the moment function to be computed.

References

[1] http://www.vsyrgkanis.com/6853sp19/lecture4.pdf. Accessed: 2020-09-15.

[2] Karim Abou-Moustafa and Csaba Szepesvári. An exponential tail bound for lq stable learning rules. volume 98 of *Proceedings of Machine Learning Research*, pages 31–63, Chicago, Illinois, 22–24 Mar 2019. PMLR.

[3] Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

[4] Chunrong Ai and Xiaohong Chen. Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, 141(1):5–43, 2007.

[5] Chunrong Ai and Xiaohong Chen. The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, 170(2):442–457, 2012.

[6] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers. *arXiv e-prints*, page arXiv:1811.04918, November 2018.

[7] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

[8] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 214–223, 2017.

[9] Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

[10] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized Random Forests. *arXiv e-prints*, page arXiv:1610.01271, October 2016.

[11] Ravi Bansal and Salim Viswanathan. No arbitrage and arbitrage pricing: A new approach. *The Journal of Finance*, 48(4):1231–1262, 1993.

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

[13] Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *J. Mach. Learn. Res.*, 20:63–1, 2019.

[14] Alexandre Belloni, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

[15] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference for high-dimensional sparse econometric models. *arXiv:1201.0220*, 2011.

[16] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

[17] Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika*, 102(1):77–94, 2014.

[18] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788, 2014.

[19] D Benkeser, M Carone, M J Van Der Laan, and P B Gilbert. Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880, 10 2017.

[20] Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems*, pages 3559–3569, 2019.

[21] Daniele Bianchi, Matthias Büchner, and Andrea Tamoni. Bond risk premiums with machine learning. *The Review of Financial Studies*, 2020.

[22] Peter J Bickel. On adaptive estimation. *The Annals of Statistics*, pages 647–671, 1982.

[23] Peter J Bickel, Chris AJ Klaassen, Ya'acov Ritov, and Jon A Wellner. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press, 1993.

[24] Peter J Bickel and Yaacov Ritov. Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 381–393, 1988.

[25] Richard Blundell, Xiaohong Chen, and Dennis Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007.

[26] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[27] Jelena Bradic and Mladen Kolar. Uniform inference for high-dimensional quantile regression: Linear functionals and regression rank scores. *arXiv:1702.06209*, 2017.

[28] Alain Celisse and Benjamin Guedj. Stability revisited: new generalisation bounds for the Leave-one-Out. *arXiv e-prints*, page arXiv:1608.06412, August 2016.

[29] Luyang Chen, Markus Pelger, and Jason Zhu. Deep learning in asset pricing. *Available at SSRN 3350138*, 2019.

[30] Xiaohong Chen and Timothy M Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9(1):39–84, 2018.

[31] Xiaohong Chen and Demian Pouzo. Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, 152(1):46–60, 2009.

[32] Xiaohong Chen and Demian Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321, 2012.

[33] Xiaohong Chen and Demian Pouzo. Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica*, 83(3):1013–1079, 2015.

[34] Xiaohong Chen and Halbert White. Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691, 1999.

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

[36] Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*, 2016.

[37] Victor Chernozhukov, Christian Hansen, and Martin Spindler. Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, 7(1):649–688, 2015.

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

[39] Victor Chernozhukov, Whitney Newey, and Rahul Singh. Double/de-biased machine learning of global and local parameters using regularized Riesz representers. *arXiv preprint arXiv:1802.08667*, 2018.

[40] Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *arXiv preprint arXiv:1809.05224*, 8, 2018.

[41] John H Cochrane. *Asset Pricing: Revised Edition*. Princeton University Press, 2009.

[42] Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

[43] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *CoRR*, abs/1711.00141, 2017.

[44] Iván Díaz and Mark J van der Laan. Targeted data adaptive estimation of the causal dose–response curve. *Journal of Causal Inference*, 1(2):171–192, 2013.

[45] Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *arXiv preprint arXiv:2006.07201*, 2020.

[46] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[47] André Elisseeff, Massimiliano Pontil, et al. Leave-one-out error and stability of learning algorithms with applications. *NATO science series sub series iii computer and systems sciences*, 190:111–130, 2003.

[48] Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 2018.

[49] Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. *arXiv preprint arXiv:1809.09953*, 2018.

[50] Guanhao Feng, Jingyu He, and Nicholas G Polson. Deep learning for predicting asset returns. *arXiv preprint arXiv:1804.09314*, 2018.

[51] Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv:1901.09036*, 2019.

[52] Dylan J. Foster and Vasilis Syrgkanis. Orthogonal Statistical Learning. *arXiv e-prints*, page arXiv:1901.09036, January 2019.

[53] Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1):79 – 103, 1999.

[54] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27:2672–2680, 2014.

[55] Shihao Gu, Bryan Kelly, and Dacheng Xiu. Autoencoder asset pricing models. *Journal of Econometrics*, 2020.

[56] Peter Hall, Joel L Horowitz, et al. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6):2904–2929, 2005.

[57] Lars Peter Hansen and Ravi Jagannathan. Assessing specification errors in stochastic discount factor models. *The Journal of Finance*, 52(2):557–590, 1997.

[58] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1225–1234. JMLR.org, 2016.

[59] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1414–1423, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[60] Rafail Z Hasminskii and Ildar A Ibragimov. On the nonparametric estimation of functionals. In *Proceedings of the Second Prague Symposium on Asymptotic Statistics*, 1979.

[61] David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *J. Comb. Theory, Ser. A*, 69(2):217–232, 1995.

[62] David A Hirshberg and Stefan Wager. Debiased inference of average partial effects in single-index models. *arXiv:1811.02547*, 2018.

[63] David A Hirshberg and Stefan Wager. Augmented minimax linear estimation. *arXiv:1712.00038v5*, 2019.

[64] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *arXiv e-prints*, page arXiv:1908.08465, August 2019.

[65] I Ibragimov and R Has'minskii. Statistical estimation, vol. 16 of. *Applications of Mathematics*, 1981.

[66] Jana Jankova and Sara Van De Geer. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1):1205–1229, 2015.

[67] Jana Jankova and Sara Van De Geer. Confidence regions for high-dimensional generalized linear models under sparsity. *arXiv:1610.01353*, 2016.

[68] Jana Jankova and Sara Van De Geer. Semiparametric efficiency bounds for high-dimensional models. *The Annals of Statistics*, 46(5):2336–2359, 2018.

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

[70] Adel Javanmard and Andrea Montanari. Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014.

[71] Adel Javanmard and Andrea Montanari. Debiasing the lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622, 2018.

[72] Satyen Kale, Ravi Kumar, and Sergei Vassilvitskii. Cross-validation and mean-square stability. In *In Proceedings of the Second Symposium on Innovations in Computer Science (ICS2011*, pages 487–495, 2011.

[73] Edward H Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv:2004.14497*, 2020.

[74] Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, 79(4):1229, 2017.

[75] Khashayar Khosravi, Greg Lewis, and Vasilis Syrgkanis. Non-Parametric Inference Adaptive to Intrinsic Dimension. *arXiv e-prints*, page arXiv:1901.03719, January 2019.

[76] George Kimeldorf and Grace Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

[77] Chris AJ Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, pages 1548–1562, 1987.

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

[79] Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media, 2007.

[80] Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method ii: complexity dependent error rates. *The Journal of Machine Learning Research*, 18(1):5356–5403, 2017.

[81] Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method i: Sparse recovery. *Ann. Statist.*, 46(2):611–641, 04 2018.

[82] B Ya Levit. On the efficiency of a class of non-parametric estimates. *Theory of Probability & Its Applications*, 20(4):723–740, 1976.

[83] Luofeng Liao, You-Lin Chen, Zhuoran Yang, Bo Dai, Zhaoran Wang, and Mladen Kolar. Provably efficient neural estimation of structural equation model: An adversarial approach. *arXiv preprint arXiv:2007.01290*, 2020.

[84] Alexander R Luedtke and Mark J Van Der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of Statistics*, 44(2):713, 2016.

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

[86] Marcial Messmer. Deep learning and the cross-section of expected returns. *Available at SSRN 3081555*, 2017.

[87] Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting Stochastic Extragradient. *arXiv e-prints*, page arXiv:1905.11373, May 2019.

[88] Krikamol Muandet, Wittawat Jitkrittum, and Jonas Kübler. Kernel conditional moment test via maximum moment restriction. *arXiv preprint arXiv:2002.09225*, 2020.

[89] Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual iv: A single stage instrumental variable regression. *arXiv preprint arXiv:1910.12358*, 2019.

[90] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statist. Sci.*, 27(4):538–557, 11 2012.

[91] Whitney K Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, pages 1349–1382, 1994.

[92] Whitney K Newey, Fushing Hsieh, and James M Robins. Undersmoothing and bias corrected functional estimation. Technical report, MIT Department of Economics, 1998.

[93] Whitney K Newey, Fushing Hsieh, and James M Robins. Twicing kernels and a small bias property of semiparametric estimators. *Econometrica*, 72(3):947–962, 2004.

[94] Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

[95] Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv:1801.09138*, 2018.

[96] Matey Neykov, Yang Ning, Jun S Liu, and Han Liu. A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statistical Science*, 33(3):427–443, 2018.

[97] Jerzy Neyman. Optimal asymptotic tests of composite statistical hypotheses. In *Probability and Statistics*, page 416–444. Wiley, 1959.

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

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

[100] Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017.

[101] Johann Pfanzagl. Lecture notes in statistics. *Contributions to a general asymptotic statistical theory*, 13, 1982.

[102] Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3066–3074. Curran Associates, Inc., 2013.

[103] Zhao Ren, Tingni Sun, Cun-Hui Zhang, and Harrison H Zhou. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

[104] James Robins, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.

[105] James Robins, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky. Comment on "performance of double-robust estimators when inverse probability weights are highly variable". *Statistical Science*, 22(4):544–559, 2007.

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

[107] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

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

[109] Dominik Rothenhäusler and Bin Yu. Incremental causal effects. *arXiv:1907.13258*, 2019.

[110] Dan Rubin and Mark J van der Laan. A general imputation methodology for nonparametric regression with censored data. Technical report, UC Berkeley Division of Biostatistics, 2005.

[111] Daniel Rubin and Mark J van der Laan. Extending marginal structural models through local, penalized, and additive learning. Technical report, UC Berkeley Division of Biostatistics, 2006.

[112] Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[113] Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, 14(3):1139–1151, 1986.

[114] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

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

[116] Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, pages 4595–4607, 2019.

[117] Rahul Singh and Liyang Sun. De-biased machine learning in instrumental variable models for treatment effects. *arXiv preprint arXiv:1909.05244*, 2019.

[118] Rahul Singh, Liyuan Xu, and Arthur Gretton. Kernel methods for policy evaluation: Treatment effects, mediation analysis, and off-policy planning. *arXiv preprint arXiv:2010.04855*, 2020.

[119] M. Soltanolkotabi, A. Javanmard, and J. D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2019.

[120] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, pages 2989–2997, 2015.

[121] Vasilis Syrgkanis and Manolis Zampetakis. Estimation and inference with trees and forests in high dimensions. volume 125 of *Proceedings of Machine Learning Research*, pages 3453–3454. PMLR, 09–12 Jul 2020.

[122] B Toth and MJ van der Laan. TMLE for marginal structural models based on an instrument. Technical report, UC Berkeley Division of Biostatistics, 2016.

[123] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

[124] Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[125] Mark J van der Laan and Alexander R Luedtke. Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome. Technical report, UC Berkeley Division of Biostatistics, 2014.

[126] Mark J Van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media, 2011.

[127] Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

[128] Aad Van Der Vaart et al. On differentiable functionals. *The Annals of Statistics*, 19(1):178–204, 1991.

[129] Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

[130] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[131] Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

[132] Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. *arXiv preprint arXiv:1802.03620*, 2018.

[133] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[134] Wenjing Zheng and Mark J Van Der Laan. Asymptotic theory for cross-validated targeted maximum likelihood estimation. 2010.

[135] Yinchu Zhu and Jelena Bradic. Breaking the curse of dimensionality in regression. *arXiv:1708.00430.*, 2017.

[136] Yinchu Zhu and Jelena Bradic. Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, 113(524):1583–1600, 2018.

A Unrestricted and Restricted Models

In the context of semi-parametric statistics, recall that the causal parameter $\theta_0 = \theta(g_0) = \mathbb{E}[m(Z; g_0)]$ is a functional m of the underlying regression $g_0(x) := \mathbb{E}[Y|X = x]$. In an unrestricted model, we assume $g_0 \in L^2(\mathbb{P})$, the space of square integrable functions. In a restricted model, additional information about g_0 can be encoded by the restriction $g_0 \in \mathcal{G}_0 \subset L^2(\mathbb{P})$, where $\mathcal{G}_0$ is some convex function space. In this section, we give an account of Riesz representation in restricted models, following the notation and technical lemmas of [39].

Denote $\mathcal{G} := span(\mathcal{G}_0)$ and $\bar{\mathcal{G}} := closure(\mathcal{G})$. Define the modulus of continuity of $g \mapsto \theta(g)$ by

$$L := \sup_{g \in \mathcal{G} \setminus \{0\}} \frac{|\theta(g)|}{\|g\|_2}$$

Definition 1 (RR and minimal RR)**.** *A RR of the functional $\theta(g)$ is $a_0 \in L^2(\mathbb{P})$ s.t.*

$$\theta(g) = \mathbb{E}[g(X)a_0(X)], \quad \forall g \in \mathcal{G}$$

If $a_0 \in \bar{\mathcal{G}}$, then it is the minimal RR and we denote it by $a_0^{\min}$. Any RR can be reduced to the minimal RR by projecting it onto $\bar{\mathcal{G}}$.

Lemma 19 (Lemma 1 of [39])**.** *We have the following results*

1. *If $L < \infty$ then there exists a unique minimal RR $a_0^{\min}$ and $L = \|a_0^{\min}\|_2$*
2. *If there exists a RR a_0 with $\|a_0\|_2 < \infty$ then $L = \|a_0^{\min}\|_2 \leq \|a_0\|_2 < \infty$, where $a_0^{\min}$ is the unique minimal RR obtained by projecting a_0 onto $\bar{\mathcal{G}}$*

In both cases, $g \mapsto \theta(g)$ can be extended to $\bar{\mathcal{G}}$ or to all of $L^2(\mathbb{P})$ with modulus of continuity L

To interpret these results, consider a toy example of vectors in $\mathbb{R}^3$ rather than functions in $L^2(\mathbb{P})$. Suppose the functional of interest is

$$\theta : \mathbb{R}^3 \to \mathbb{R}, \quad (x, y, z) \mapsto x + 2y + 3z$$

Moreover, assume $g_0 \in \mathcal{G}$ where $\mathcal{G}$ is the (x, y)-plane, though the ambient space is $\mathbb{R}^3$. Then any vector of the form $a_0 = (1, 2, c)$ with $c \in \mathbb{R}$ is a valid RR. The unique minimal RR is $a_0^{\min} = (1, 2, 0)$. As an aside, the vector $a_0 = (1, 2, 3)$ is a universal RR; it holds for any choice of $\mathcal{G} \subset \mathbb{R}^3$, not just the (x, y)-plane. From any RR, we can obtain the minimal RR by projection onto the (x, y)-plane.

In [39, Theorem 2], we see that it is better to use $a_0^{\min}$ rather than any a_0 to attain full semi-parametric efficiency (unless of course $\mathcal{G} = L^2(\mathbb{P})$ so there is no difference). By the stated lemma, we know how to obtain $a_0^{\min}$ from any a_0: projection onto $\bar{\mathcal{G}}$.

When do these technical issues arise? In the semi-parametric literature, a popular restricted model is the additive model. It is an important setting where $\mathcal{G}$ is not dense in $L^2(\mathbb{P})$. We present a definition of the additive model, then a technical lemma about the minimal RR in an additive model.

Definition 2 (Additive model)**.** *Suppose that*

1. *the regression g_0 is additive in components $x = (x^{(1)}, x^{(2)})$: $g_0(x) = g_0^{(1)}(x^{(1)}) + g_0^{(2)}(x^{(2)})$*
2. *$g_0^{(1)} \in \mathcal{G}_0^{(1)}$, a dense subset of $L^2(\mathbb{P}^{(1)})$, where $\mathbb{P}^{(1)}$ is the distribution of $X^{(1)}$*
3. *the functional depends on only the first component: $m(z; g) = m(z; g^{(1)})$*

Lemma 20 (Lemma 6 of [39])**.** *Assume an additive model. Consider any RR $a_0 \in L^2(\mathbb{P})$. Then $\forall g \in \mathcal{G}$*

$$\theta(g) = \theta(g^{(1)}) = \int a_0^{\min}(x^{(1)}) g^{(1)}(x^{(1)}) d\mathbb{P}^{(1)}, \quad a_0^{\min}(x^{(1)}) = \mathbb{E}[a_0(X)|X^{(1)} = x^{(1)}]$$

and

$$\|a_0^{\min}\|_q \le \|a_0\|_q, \quad \forall q \in [1, \infty]$$

This preservation of order and contraction of norm is helpful in analysis.

Finally, we quote some projection geometry for sumspaces from [23, Appendix A.4]. Suppose $\mathcal{H}_1$ and $\mathcal{H}_2$ are closed subspaces of a Hilbert space $\mathcal{H}$.

Lemma 21. *If $\mathcal{H}_1 \perp \mathcal{H}_2$ then the projection onto the sumspace $\mathcal{H}_1 + \mathcal{H}_2$ is the sum of the projections onto $\mathcal{H}_1$ and $\mathcal{H}_2$*

More generally, $\mathcal{H}_1$ may not be orthogonal to $\mathcal{H}_2$. Denote by P_i the orthogonal projection onto $\mathcal{H}_i$, and denote by $Q_i := I - P_i$ the projection onto $\mathcal{H}_i^{\perp}$. Denote by Π the projection onto the closure of $\mathcal{H}_1 + \mathcal{H}_2$

Lemma 22 (Corollary 1 of [23])**.** *For any $h \in \mathcal{H}$*

$$[I - (Q_1 Q_2)^m] h \to \Pi h, \quad m \to \infty$$

Stronger versions of this result are available that provide quantitative rates of convergence and that allow for $r \ge 2$ subspaces.

B Examples

B.1 Causal Inference

Recall the definition of mean-squared continuity: $\exists M \ge 0$ s.t.

$$\forall f \in \mathcal{F} : \sqrt{\mathbb{E}\left[m(Z; f)^2\right]} \le M \|f\|_2$$

We verify mean-square continuity for several important functionals.

1. Average treatment effect (ATE): $\theta_0 = \mathbb{E}[g_0(1, W) - g_0(0, W)]$

   To lighten notation, let $\pi_0(w) := \mathbb{P}(D = 1|W = w)$ be the propensity score. Assume $\pi_0(w) \in \left(\frac{1}{M}, 1 - \frac{1}{M}\right)$ for $M \in (1, \infty)$. Then

$$\begin{aligned}
\mathbb{E}[g(1, W) - g(0, W)]^2 \le\ & 2\mathbb{E}[g(1, W)^2 + g(0, W)^2] \\
\le\ & 2M\mathbb{E}\left[\pi_0(W)\, g(1, W)^2 + [1 - \pi_0(W)]\, g(0, W)^2\right] \\
=\ & 2M\mathbb{E}[g(X)]^2
\end{aligned}$$

2. Average policy effect: $\theta_0 = \int g_0(x) d\mu(x)$ where $\mu(x) = F_1(x) - F_0(x)$

   Denote the densities corresponding to distributions (F, F_1, F_0) by (f, f_1, f_0). Assume $\frac{f_1(x)}{f(x)} \leq \sqrt{M}$ and $\frac{f_0(x)}{f(x)} \leq \sqrt{M}$ for $M \in [0, \infty)$. In this example, $m(Z; g) = m(g)$.

$$\begin{aligned}
\mathbb{E}[m(Z;g)]^2 &= \{m(g)\}^2 \\
&= \left\{ \int g(x) d\mu(x) \right\}^2 \\
&= \left\{ \mathbb{E}\left[g(X) \left\{ \frac{f_1(X)}{f(X)} - \frac{f_0(X)}{f(X)} \right\} \right] \right\}^2 \\
&\leq \left\{ 2\sqrt{M}\mathbb{E}|g(X)| \right\}^2 \\
&\leq 4M\mathbb{E}[g(X)]^2
\end{aligned}$$

3. Policy effect from transporting covariates: $\theta_0 = \mathbb{E}[g_0(t(X)) - g_0(X)]$

   Denote the density of $t(X)$ by $f_t(x)$. Assume $\frac{f_t(x)}{f(x)} \leq M$ for $M \in [0, \infty)$. Then

$$\begin{aligned}
\mathbb{E}[g(t(X)) - g(X)]^2 &\leq 2\mathbb{E}[g(t(X))^2 + g(X)^2] \\
&= 2\mathbb{E}\left[g(X)^2 \left\{ \frac{f_t(X)}{f(X)} - 1 \right\} \right] \\
&\leq 2(M+1)\mathbb{E}[g(X)]^2
\end{aligned}$$

4. Cross effect: $\theta_0 = \mathbb{E}[Dg_0(0, W)]$

   Assume $\pi_0(w) < 1 - \frac{1}{M}$ for some $M \in (1, \infty)$. Then

$$\begin{aligned}
\mathbb{E}[Dg(0,W)]^2 &\leq \mathbb{E}[g(0,W)]^2 \\
&\leq M\mathbb{E}[\{1 - \pi_0(W)\} g(0,W)^2] \\
&\leq M\mathbb{E}[g(X)]^2
\end{aligned}$$

5. Regression decomposition: $\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0] = \theta_0^{response} + \theta_0^{composition}$ where

$$\begin{aligned}
\theta_0^{response} &= \mathbb{E}[g_0(1,W)|D=1] - \mathbb{E}[g_0(0,W)|D=1] \\
\theta_0^{composition} &= \mathbb{E}[g_0(0,W)|D=1] - \mathbb{E}[g_0(0,W)|D=0]
\end{aligned}$$

   Assume $\pi_0(w) < 1 - \frac{1}{M}$ for some $M \in (1, \infty)$. Then re-write the target parameters in terms of the cross effect.

$$\begin{aligned}
\theta_0^{response} &= \frac{\mathbb{E}[DY] - \mathbb{E}[Dg_0(0,W)]}{\mathbb{E}[D]} \\
\theta_0^{composition} &= \frac{\mathbb{E}[D\gamma_0(0,W)]}{\mathbb{E}[D]} - \frac{\mathbb{E}[(1-D)Y]}{\mathbb{E}[1-D]}
\end{aligned}$$

   We implement DML for the cross effect, empirical means for the population means, then delta method.

6. Average treatment on the treated (ATT): $\theta_0 = \mathbb{E}[g_0(1,W)|D=1] - \mathbb{E}[g_0(0,W)|D=1]$

   Assume $\pi_0(w) < 1 - \frac{1}{M}$ for some $M \in (1,\infty)$. Then re-write the target parameters in terms of the cross effect.

$$\theta_0 = \frac{\mathbb{E}[DY] - \mathbb{E}[Dg_0(0,W)]}{\mathbb{E}[D]}$$

   We implement DML for the cross effect, empirical means for the population means, then delta method.

7. Local average treatment effect (LATE): $\theta_0 = \frac{\mathbb{E}[g_0(1,W)-g_0(0,W)]}{\mathbb{E}[h_0(1,W)-h_0(0,W)]}$

   The result follows from the view of LATE as a ratio of two ATEs.

B.2 Asset Pricing

We present three proofs of the existence of the stochastic discount factor. These arguments are quoted from the excellent exposition of [41].

1. Marginal rate of substitution in a consumption model.

   Consider an investor with utility function $U(c_t, c_{t+1}) = u(c_t) + \beta\mathbb{E}_t[u(c_{t+1})]$, where u is period utility, c_t is consumption at time t, and β is a subjective discount factor. Denote by e_t the original consumption level, and ξ the amount of the asset the consumer buys. The consumer solves the optimization problem

$$\max_{\xi} u(c_t) + \beta\mathbb{E}_t[u(c_{t+1})] \quad \text{s.t.} \quad c_t = e_t - p_t\xi, \quad c_{t+1} = e_{t+1} + x_{t+1}\xi$$

   Substituting constraints into the objective, the FOC yields

$$p_t = \mathbb{E}_t\left[\beta\frac{u'(c_{t+1})}{u'(c_t)}x_{t+1}\right], \quad m_{t+1} = \beta\frac{u'(c_{t+1})}{u'(c_t)}$$

   The same FOC arises in the longer-term objective $\mathbb{E}_t\left[\sum_{j=0}^{\infty}\beta^j u(c_{t+j})\right]$.

2. State price density in a contingent claim model with complete markets.

   For simplicity, consider a two-period model with S possible states of nature tomorrow. A contingent claim is a security that pays one dollar in one state s only tomorrow. $pc_t(s)$ is the price today of the contingent claim. In a *complete market*, investors can buy any contingent claims. If there are complete contingent claims, the state price density exists, and it is equal to the contingent claim price divided by probabilities. Let $x_{t+1}(s)$ denote an asset's payoff in state of nature s. The asset's price must equal the value of the contingent claims of which it is a bundle. Let $\pi_{t+1}(s)$ be the probability that state s occurs conditional on information available today. Then

$$p_t = \sum_s pc_t(s)x_{t+1}(s) = \sum_s \pi_{t+1}(s)\frac{pc_t(s)}{\pi_{t+1}(s)}x_{t+1}(s), \quad m_{t+1}(s) = \frac{pc_t(s)}{\pi_{t+1}(s)}$$

3. Pricing kernel from the law of one price.

   Let $\mathcal{X}$ be the set of all payoffs that investors can purchase (or the subset of tradeable payoffs used in a particular study). For example, if there are complete contingent claims to S states of nature then $\mathcal{X} = \mathbb{R}^S$. More generally, markets are incomplete, so $\mathcal{X} \subset \mathbb{R}^S$.

   Free portfolio formation means $x, x' \in \mathcal{X}$ implies $ax + bx' \in \mathcal{X}$ for any $a, b \in \mathbb{R}$. This assumption rules out short sales constraints, bid-ask spreads, and leverage limitations. Let $p_t(x)$ denote the price at time t of the asset that delivers payoff x at time $t+1$. The law of one price means $p_t(ax+bx') = ap_t(x)+bp_t(x')$. In other words, asset pricing is a linear functional over a vector space. This assumption says that investors cannot make instantaneous profits by repackaging portfolios. It would be satisfied in a market that has already reached equilibrium.

   Given free portfolio formation and the law of one price, there exists a unique payoff $m^*_{t+1} \in \mathcal{X}$ such that $p_t(x) = \mathbb{E}_t[m^*_{t+1}x]$ for all $x \in \mathcal{X}$. m^*_{t+1} is called the *mimicking portfolio*. Unless markets are complete, there are infinitely many SDFs that satisfy $p_t(x) = \mathbb{E}_t[m_{t+1}x]$ of the form $m_{t+1} = m^*_{t+1} + \epsilon$ where $\epsilon \in \mathcal{X}^{\perp}$. An incomplete market can be interpreted as a restricted model, and the mimicking portfolio can be interpreted as a minimal Riesz representer in the discussion of Section A.

C Local Riesz Representer Convergence Rate

Suppose that we use the constraint the test functions to lie in:

$$\mathcal{F}(r_n) = \{f \in \mathrm{star}\left(\partial(\mathcal{G} - \hat{g})\right) : \|f\|_2 \le r_n\}$$

And consider the estimator:

$$\inf_{a\in\mathcal{A}} \sup_{f\in} \Psi_n(a, f)$$

Then by a localized concentration bound we have:

$$\begin{aligned}\forall a \in \mathcal{A}, f \in \mathcal{F}(r_n) : |\Psi_n(a, f) - \Psi(a, f)| \le\ & O\left(\delta_{n,\zeta}\|m(\cdot; f) - a\, f\|_2 + \delta^2_{n,\zeta}\right) \\ \le\ & O\left((M+1)\delta_{n,\zeta}\|f\|_2 + \delta^2_{n,\zeta}\right) \\ \le\ & O\left((M+1)\delta_{n,\zeta} r_n + \delta^2_{n,\zeta}\right) =: \epsilon_n\end{aligned}$$

where $\delta_{n,\zeta} = \delta_n + c_0\sqrt{\frac{\log(c_1/\zeta)}{n}}$ and δ_n bounds the critical radius of the function class:

$$\{Z \to m(Z; f) - a(X)\, f(X) : f \in \mathcal{F}(r_n), a \in \mathcal{A}\}.$$

Thus we have that:

$$\begin{aligned}\sup_{f\in\mathcal{F}(r_n)} \Psi(\hat{a}, f) - \epsilon_n \le\ & \sup_{f\in\mathcal{F}(r_n)} \Psi_n(\hat{a}, f) \le \sup_{f\in\mathcal{F}(r_n)} \Psi_n(a_*, f) \\ \le\ & \sup_{f\in\mathcal{F}(r_n)} \Psi(a_*, f) + \epsilon_n \\ =\ & \inf_{a\in\mathcal{A}} \sup_{f\in\mathcal{F}(r_n)} \Psi(a, f) + \epsilon_n\end{aligned}$$

Concluding that:

$$\sup_{f\in\mathcal{F}(r_n)} \Psi(\hat{a}, f) \leq \inf_{a\in\mathcal{A}} \sup_{f\in\mathcal{F}(r_n)} \Psi(a, f) + 2\,\epsilon_n$$

Moreover, if a_0 is a local Riesz representer, i.e. it satisfies the Riesz equation for differences with $\hat{g}$ of any function in $\mathcal{G}$ within a ball r_n around $\hat{g}$, then:

$$\inf_{a\in\mathcal{A}} \sup_{f\in\mathcal{F}(r_n)} \Psi(a, f) = \inf_{a\in\mathcal{A}} \sup_{f\in\mathcal{F}(r_n)} \langle a_0 - a, f\rangle \leq r_n \inf_{a\in\mathcal{A}} \sup_{f\in\mathcal{F}(1)} \langle a_0 - a, f\rangle \leq r_n \inf_{a\in\mathcal{A}} \|a_0 - a\|_2$$

Thus if g_0 lies within a ball r_n of $\hat{g}$, we conclude that:

$$\mathbb{E}[(a_0(X) - \hat{a}(X))\ (\hat{g}(X) - g_0(X))] \leq O\left(M\, r_n \delta_{n,\zeta} + \delta_{n,\zeta}^2 + r_n \inf_{a\in\mathcal{A}} \|a_0 - a\|_2\right)$$

If for instance $r_n\, \delta_{n,\zeta} = o(n^{-1/2})$ and $\delta_{n,\zeta} = o(n^{-1/4})$ and $r_n \inf_{a\in\mathcal{A}} \|a_0 - a\|_2 = o(n^{-1/2})$, then we can conclude that:

$$\sqrt{n}\,\mathbb{E}[(a_0(X) - \hat{a}(X))\ (\hat{g}(X) - g_0(X))] \to_p 0$$

If $a_0 \in \mathcal{A}$ and both $\mathcal{A}$ and $\mathcal{G}$ are VC-subgraph classes with constant VC dimension, then it can be shown that $\delta_{n,\zeta} = O\left(\sqrt{\frac{\log(n/\zeta)}{n}}\right)$. Thus for the above conditions to hold, it suffices that: $r_n = o(1)$ (i.e. that $\hat{g}$ is RMSE-consistent).

Finally observe that we need $\mathcal{A}$ to have a small approximation error to a_0, not with respect to the $\|\cdot\|_2$ norm, but rather with the weaker norm:

$$\|a_0 - a\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}(1)} \langle a_0 - a, f\rangle$$

Thus a does not need to match the component of a_0 that is orthogonal to the subspace $\mathcal{F}$. If for instance, we assume that $\mathcal{F}$ lies in the space spanned by top K eigenfunctions of a reproducing kernel hilbert space, then it suffices to consider $\mathcal{A}$ the space spanned by those functions too. Then $\inf_{a\in\mathcal{A}} \|a_0 - a\|_{\mathcal{F}} = 0$. For instance, if $\mathcal{G}$ is a finite dimensional linear function space and $g_0 \in \mathcal{G}$, then it suffices to consider $\mathcal{A}$ that is also finite dimensional linear, even if the true a_0 does not lie in that sub-space. Then all the conditions of Lemma 16 will be satisfied, even if $\hat{a}$ will never be consistent with respect to a_0.

D Proofs from Section 3

For convenience, throughout this section we will use the notation:

$$\begin{aligned}
\Psi(a, f) :=&\ \mathbb{E}[m(Z; f) - a(X)\cdot f(X)] = \mathbb{E}[(a_0(X) - a(X))\, f(X)] && \text{(by Riesz definition)}\\
\Psi_n(a, f) :=&\ \frac{1}{n}\sum_{i=1}^{n} (m(Z_i; f) - a(X_i)\cdot f(X_i))
\end{aligned}$$

D.1 Proof of Theorem 1

Proof. Let:

$$\begin{aligned}\Psi_n^\lambda(a,f) =&\ \Psi_n(a,f) - \|f\|_{2,n}^2 - \lambda\|f\|_{\mathcal{A}}^2\\ \Psi^\lambda(a,f) =&\ \Psi(a,f) - \frac{1}{4}\|f\|_2^2 - \frac{\lambda}{2}\|f\|_{\mathcal{A}}^2\end{aligned}$$

Thus our estimate can be written as:

$$\hat{a} := \arg\min_{a\in\mathcal{A}} \sup_{f\in\mathcal{F}} \Psi_n^\lambda(a,f) + \mu\|a\|_{\mathcal{A}}^2$$

Relating empirical and population regularization. As a preliminary observation, we have that by Theorem 14.1 of [130], w.p. $1-\zeta$:

$$\forall f \in \mathcal{F}_B : \left|\|f\|_{n,2}^2 - \|f\|_2^2\right| \leq \frac{1}{2}\|f\|_2^2 + \delta^2$$

for our choice of $\delta := \delta_n + c_0\sqrt{\frac{\log(c_1/\zeta)}{n}}$, where δ_n upper bounds the critical radius of $\mathcal{F}_B$ and c_0, c_1 are universal constants. Moreover, for any f, with $\|f\|_{\mathcal{A}}^2 \geq B$, we can consider the function $f\sqrt{B}/\|f\|_{\mathcal{A}}$, which also belongs to $\mathcal{F}_B$, since $\mathcal{F}$ is star-convex. Thus we can apply the above lemma to this re-scaled function and multiply both sides by $\|f\|_{\mathcal{A}}^2/B$, leading to:

$$\forall f \in \mathcal{F} \text{ s.t. } \|f\|_{\mathcal{A}}^2 \geq B : \left|\|f\|_{n,2}^2 - \|f\|_2^2\right| \leq \frac{1}{2}\|f\|_2^2 + \delta^2\frac{\|f\|_{\mathcal{A}}^2}{B}$$

Thus overall, we have:

$$\forall f \in \mathcal{F} : \left|\|f\|_{n,2}^2 - \|f\|_2^2\right| \leq \frac{1}{2}\|f\|_2^2 + \delta^2 \max\left\{1, \frac{\|f\|_{\mathcal{A}}^2}{B}\right\} \tag{9}$$

Thus we have that w.p. $1-\zeta$:

$$\begin{aligned}\forall f \in \mathcal{F} : \lambda\|f\|_{\mathcal{A}}^2 + \|f\|_{2,n}^2 \geq&\ \lambda\|f\|_{\mathcal{A}}^2 + \frac{1}{2}\|f\|_2^2 - \delta^2\max\left\{1, \frac{\|f\|_{\mathcal{A}}^2}{B}\right\}\\ \geq&\ \left(\lambda - \frac{\delta^2}{B}\right)\|f\|_{\mathcal{A}}^2 + \frac{1}{2}\|f\|_2^2 - \delta^2\end{aligned}$$

Assuming that $\lambda \geq \frac{2\delta^2}{B}$, we have that, the latter is at least:

$$\forall f \in \mathcal{F} : \lambda\|f\|_{\mathcal{A}}^2 + \|f\|_{2,n}^2 \geq \frac{\lambda}{2}\|f\|_{\mathcal{A}}^2 + \frac{1}{2}\|f\|_2^2 - \delta^2$$

Upper bounding centered empirical sup-loss. We now argue that the centered empirical sup-loss:

$$\sup_{f\in\mathcal{F}}(\Psi_n(\hat{a},f) - \Psi_n(a_*,f)) = \sup_{f\in\mathcal{F}} \mathbb{E}_n[(a_*(X) - \hat{a}(X))\, f(X)]$$

is small. By the definition of $\hat{a}$:

$$\sup_{f\in\mathcal{F}} \Psi_n^\lambda(\hat{a}, f) \leq \sup_{f\in\mathcal{F}} \Psi_n^\lambda(a_*, f) + \mu\left(\|a_*\|_{\mathcal{A}}^2 - \|\hat{a}\|_{\mathcal{A}}^2\right) \tag{10}$$

By Lemma 7 of [52], the fact that $m(Z; f) - a_*(X)f(X)$ is 2-Lipschitz with respect to the vector $(m(Z; f), f(z))$ (since $a_*(X) \in [-1, 1]$) and by our choice of $\delta := \delta_n + c_0\sqrt{\frac{\log(c_1/\zeta)}{n}}$, where δ_n is an upper bound on the critical radius of $\mathcal{F}_B$ and $m \circ \mathcal{F}_B$, w.p. $1-\zeta$:

$$\forall f \in \mathcal{F}_B : |\Psi_n(a_*, f) - \Psi(a_*, f)| \leq O\left(\delta\left(\|f\|_2 + \sqrt{\mathbb{E}[m(Z; f)^2]}\right) + \delta^2\right) = O\left(\delta\, M\, \|f\|_2 + \delta^2\right)$$

where we have invoked Assumption 1. Thus, if $\|f\|_{\mathcal{A}} \geq \sqrt{B}$, we can apply the latter inequality for the function $f\sqrt{B}/\|f\|_{\mathcal{A}}$, which falls in $\mathcal{F}_B$, and then multiply both sides by $\|f\|_{\mathcal{A}}/\sqrt{B}$ (invoking the linearity of the operator $\Psi_n(a, f)$ with respect to f) to get:

$$\forall f \in \mathcal{F} : |\Psi_n(a_*, f) - \Psi(a_*, f)| \leq O\left(\delta\, M\, \|f\|_2 + \delta^2 \max\left\{1, \frac{\|f\|_{\mathcal{A}}}{\sqrt{B}}\right\}\right) \tag{11}$$

By Equations (10) and (11), we have that w.p. $1 - 2\zeta$, for some universal constant C:

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \Psi_n^\lambda(a_*, f) = & \sup_{f\in\mathcal{F}} \left(\Psi_n(a_*, f) - \|f\|_{2,n}^2 - \lambda\|f\|_{\mathcal{A}}^2\right) \\
\leq & \sup_{f\in\mathcal{F}} \left(\Psi(a_*, f) + C\delta^2 + \frac{C\delta^2}{\sqrt{B}}\|f\|_{\mathcal{A}} + CM\delta\|f\|_2 - \|f\|_{2,n}^2 - \lambda\|f\|_{\mathcal{A}}^2\right) \\
\leq & \sup_{f\in\mathcal{F}} \left(\Psi(a_*, f) + C\delta^2 + \frac{C\delta^2}{\sqrt{B}}\|f\|_{\mathcal{A}} + CM\delta\|f\|_2 - \frac{1}{2}\|f\|_2^2 - \frac{\lambda}{2}\|f\|_{\mathcal{A}}^2 + \delta^2\right) \\
\leq & \sup_{f\in\mathcal{F}} \Psi^{\lambda/2}(a_*, f) + O\left(\delta^2\right) + \sup_{f\in\mathcal{F}} \left(\frac{C\delta^2}{\sqrt{B}}\|f\|_{\mathcal{A}} - \frac{\lambda}{4}\|f\|_{\mathcal{A}}^2\right) + \sup_{f\in\mathcal{F}} \left(CM\delta\|f\|_2 - \frac{1}{4}\|f\|_2^2\right)
\end{aligned}$$

Moreover, observe that for any norm $\|\cdot\|$ and any constants $a, b > 0$:

$$\sup_{f\in\mathcal{F}} \left(a\|f\| - b\|f\|^2\right) \leq \frac{a^2}{4b}$$

Thus if we assume that $\lambda \geq 2\delta^2/B$, we have:

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \left(\frac{C\delta^2}{\sqrt{B}}\|f\|_{\mathcal{A}} - \frac{\lambda}{4}\|f\|_{\mathcal{A}}^2\right) \leq & \; \frac{C^2\delta^4}{B\lambda} \leq \frac{C^2}{2}\delta^2 \\
\sup_{f\in\mathcal{F}} \left(CM\delta\|f\|_2 - \frac{1}{4}\|f\|_2^2\right) \leq & \; C^2M^2\delta^2
\end{aligned}$$

Thus we have:

$$\sup_{f\in\mathcal{F}} \Psi_n^\lambda(a_*, f) \leq \sup_{f\in\mathcal{F}} \Psi^{\lambda/2}(a_*, f) + O\left(M^2\, \delta^2\right)$$

Moreover:

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \Psi_n^\lambda(\hat{a}, f) =& \sup_{f\in\mathcal{F}} \left(\Psi_n(\hat{a}, f) - \Psi_n(a_*, f) + \Psi_n(a_*, f) - \|f\|_{2,n}^2 - \lambda\|f\|_{\mathcal{A}}^2\right) \\
\geq& \sup_{f\in\mathcal{F}} \left(\Psi_n(\hat{a}, f) - \Psi_n(a_*, f) - 2\|f\|_{2,n}^2 - 2\,\lambda\|f\|_{\mathcal{A}}^2\right) + \inf_{f\in\mathcal{F}} \left(\Psi_n(a_*, f) + \|f\|_{2,n}^2 + \lambda\|f\|_{\mathcal{A}}^2\right)
\end{aligned}$$

Observe that since $\Psi_n(a, f)$ is a linear operator of f and $\mathcal{F}$ is a symmetric class, we have:

$$\begin{aligned}
\inf_{f\in\mathcal{F}} \left(\Psi_n(a_*, f) + \|f\|_{2,n}^2 + \lambda\|f\|_{\mathcal{A}}^2\right) =& \inf_{f\in\mathcal{F}} \left(\Psi_n(a_*, -f) + \|f\|_{2,n}^2 + \lambda\|f\|_{\mathcal{A}}^2\right) \\
=& \inf_{f\in\mathcal{F}} \left(-\Psi_n(a_*, f) + \|f\|_{2,n}^2 + \lambda\|f\|_{\mathcal{A}}^2\right) \\
=& -\sup_{f\in\mathcal{F}} \left(\Psi_n(a_*, f) - \|f\|_{2,n}^2 - \lambda\|f\|_{\mathcal{A}}^2\right) = -\sup_{f\in\mathcal{F}} \Psi_n^\lambda(a_*, f)
\end{aligned}$$

Combining this with Equation (10) yields:

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \left(\Psi_n(\hat{a}, f) - \Psi_n(a_*, f) - \|f\|_{2,n}^2 - \lambda\|f\|_{\mathcal{A}}^2\right) \leq& \; 2 \sup_{f\in\mathcal{F}} \Psi_n^\lambda(a_*, f) + \mu\left(\|a_*\|_{\mathcal{A}}^2 - \|\hat{a}\|_{\mathcal{A}}^2\right) \\
\leq& \; 2 \sup_{f\in\mathcal{F}} \Psi^{\lambda/2}(a_*, f) + \mu\left(\|a_*\|_{\mathcal{A}}^2 - \|\hat{a}\|_{\mathcal{A}}^2\right) + O\left(M^2\,\delta^2\right)
\end{aligned}$$

Lower bounding centered empirical sup-loss. First observe that:

$$\Psi_n(a, f) - \Psi_n(a_*, f) = \mathbb{E}_n[(a_*(X) - a(X))f(X)]$$

Let $\Delta = a_* - \hat{a}$. Suppose that $\|\Delta\|_2 \geq \delta$ and let $r = \frac{\delta}{2\|\Delta\|_2} \in [0, 1/2]$. Then observe that since $\Delta \in \mathcal{F}$ and $\mathcal{F}$ is star-convex, we also have that $r\Delta \in \mathcal{F}$. Thus

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \left(\Psi_n(\hat{a}, f) - \Psi_n(a_*, f) - \|f\|_{2,n}^2 - \lambda\|f\|_{\mathcal{A}}^2\right) \geq& \; \Psi_n(\hat{a}, r\Delta) - \Psi_n(a_*, r\Delta) - r^2\|\Delta\|_{2,n}^2 - \lambda r^2\|\Delta\|_{\mathcal{A}}^2 \\
=& \; r\mathbb{E}_n\left[(a_*(X) - \hat{a}(X))^2\right] - r^2\|\Delta\|_{2,n}^2 - \lambda r^2\|\Delta\|_{\mathcal{A}}^2 \\
=& \; r\|\Delta\|_{2,n}^2 - r^2\|\Delta\|_{2,n}^2 - \lambda r^2\|\Delta\|_{\mathcal{A}}^2 \\
\geq& \; r\|\Delta\|_{2,n}^2 - r^2\|\Delta\|_{2,n}^2 - \lambda\|\Delta\|_{\mathcal{A}}^2
\end{aligned}$$

Moreover, since δ_n upper bounds the critical radius of $\mathcal{F}_B$ and by Equation (9):

$$\begin{aligned}
r^2\|\Delta\|_{2,n}^2 \leq& \; r^2\left(2\|\Delta\|_2^2 + \delta^2 + \delta^2\frac{\|\Delta\|_{\mathcal{A}}^2}{B}\right) \\
\leq& \; 2\delta^2 + \delta^2\frac{\|\Delta\|_{\mathcal{A}}^2}{B} \leq 2\delta^2 + \lambda\|\Delta\|_{\mathcal{A}}^2
\end{aligned}$$

Thus we get:

$$\sup_{f\in\mathcal{F}} \left(\Psi_n(\hat{a}, f) - \Psi_n(a_*, f) - \|f\|_{2,n}^2 - \lambda\|f\|_{\mathcal{A}}^2\right) \geq \; r\|\Delta\|_{2,n}^2 - 2\delta^2 - 2\lambda\|\Delta\|_{\mathcal{A}}^2$$

Furthermore, again, since δ_n upper bounds the critical radius of $\mathcal{F}_B$ and by Equation (9):

$$\|\Delta\|_{2,n}^2 \geq \frac{1}{2}\|\Delta\|_2^2 - \frac{\delta^2}{2B}\|\Delta\|_{\mathcal{A}}^2 - \delta^2 \geq \frac{1}{2}\|\Delta\|_2^2 - \lambda\|\Delta\|_{\mathcal{A}}^2 - \delta^2$$

Thus we have:

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \left(\Psi_n(\hat{a}, f) - \Psi_n(a_*, f) - \|f\|_{2,n}^2 - \lambda\|f\|_{\mathcal{A}}^2\right) \geq\ & \frac{r}{2}\|\Delta\|_2^2 - 3\delta^2 - 3\lambda\|\Delta\|_{\mathcal{A}}^2 \\
\geq\ & \frac{\delta}{4}\|\Delta\|_2 - 3\delta^2 - 3\lambda\|\Delta\|_{\mathcal{A}}^2
\end{aligned}$$

Combining upper and lower bound. Combining the upper and lower bound on the centered population sup-loss we get that w.p. $1-3\zeta$: either $\|\Delta\|_2 \leq \delta$ or:

$$\frac{\delta}{4}\|\Delta\|_2 \leq\ O\left(M^2\,\delta^2\right) + 2\sup_{f\in\mathcal{F}} \Psi^{\lambda/2}(a_*, f) + 3\lambda\|\Delta\|_{\mathcal{A}}^2 + \mu\left(\|a_*\|_{\mathcal{A}}^2 - \|\hat{a}\|_{\mathcal{A}}^2\right)$$

We now control the last part. Since $\mu \geq 6\lambda$:

$$3\lambda\|\Delta\|_{\mathcal{A}}^2 + \mu\left(\|a_*\|_{\mathcal{A}}^2 - \|\hat{a}\|_{\mathcal{A}}^2\right) \leq\ 6\lambda\left(\|a_*\|_{\mathcal{A}}^2 + \|\hat{a}\|_{\mathcal{A}}^2\right) + \mu\left(\|a_*\|_{\mathcal{A}}^2 - \|\hat{a}\|_{\mathcal{A}}^2\right) \leq 2\mu\|a_*\|_{\mathcal{A}}^2$$

We can then conclude that:

$$\frac{\delta}{4}\|\Delta\|_2 \leq O\left(M^2\,\delta^2\right) + 2\sup_{f\in\mathcal{F}} \Psi^{\lambda/2}(a_*, f) + 2\mu\|a_*\|_{\mathcal{A}}^2$$

Dividing over by $\delta/4$, we get:

$$\|\Delta\|_2 \leq O\left(M^2\,\delta\right) + \frac{8}{\delta}\sup_{f\in\mathcal{F}} \Psi^{\lambda/2}(a_*, f) + 8\frac{\mu}{\delta}\|a_*\|_{\mathcal{A}}^2$$

Thus either $\|\Delta\|_2 \leq \delta$ or the latter inequality holds. Thus in any case the latter inequality holds.

Upper bounding population sup-loss at minimum. Observe that by the definition of the Riesz representer:

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \Psi^{\lambda/2}(a_*, f) =\ & \sup_{f\in\mathcal{F}} \mathbb{E}[(a_0(X) - a_*(X))\, f(z)] - \frac{1}{4}\|f\|_2^2 - \frac{\lambda}{4}\|f\|_{\mathcal{A}}^2 \\
\leq\ & \sup_{f\in\mathcal{F}} \mathbb{E}[(a_0(X) - a_*(X))\, f(z)] - \frac{1}{4}\|f\|_2^2 = \|a_0 - a_*\|_2^2
\end{aligned}$$

Concluding. Concluding we get that w.p. $1-3\zeta$:

$$\|\hat{a} - a_*\|_2 \leq O\left(M^2\,\delta\right) + \frac{8}{\delta}\,\|a_* - a_0\|_2^2 + 8\frac{\mu}{\delta}\|a_*\|_{\mathcal{A}}^2$$

By a trinagle inequality we get:

$$\|\hat{a} - a_0\|_2 \leq O\left(M^2\,\delta\right) + \frac{8}{\delta}\,\|a_* - a_0\|_2^2 + \|a_* - a_0\|_2 + 8\frac{\mu}{\delta}\|a_*\|_{\mathcal{A}}^2$$

Choosing $a_* = \arg\min_{a\in\mathcal{A}} \|a - a_0\|_2$ and using the fact that $\delta \ge \epsilon_n$, we get:

$$\|\hat{a} - a_0\|_2 \le O\left(M^2\delta + \|a_* - a_0\|_2 + \frac{\mu}{\delta}\|a_*\|_{\mathcal{A}}^2\right) \le O\left(M^2\delta + \frac{\mu}{\delta}\|a_*\|_{\mathcal{A}}^2\right)$$

□

D.2 Proof of Theorem 3

Proof. By the definition of $\hat{a}$:

$$0 \le \sup_f \Psi_n(\hat{a}, f) \le \sup_f \Psi_n(a_0, f) + \lambda\left(\|a_0\|_{\mathcal{A}} - \|\hat{a}\|_{\mathcal{A}}\right)$$

Let

$$\delta_{n,\zeta} = \max_i \left(\mathcal{R}(\mathcal{F}^i) + \mathcal{R}(m \circ \mathcal{F}^i)\right) + c_0\sqrt{\frac{\log(c_1/\zeta)}{n}}$$

for some universal constants c_0, c_1. By Theorem 26.5 and 26.9 of [115], and since $\mathcal{F}^i$ is a symmetric class and since $\|a_0\|_\infty \le 1$, w.p. $1-\zeta$:

$$\forall f \in \mathcal{F}^i : |\Psi_n(a_0, f) - \Psi(a_0, f)| \le \delta_{n,\zeta}$$

Since $\Psi(a_0, f) = 0$ for all $f \in \mathcal{F}$, we have that, w.p. $1-\zeta$:

$$\|\hat{a}\|_{\mathcal{A}} \le \|a_0\|_{\mathcal{A}} + \delta_{n,\zeta}/\lambda$$

Let $B_{n,\lambda,\zeta} = (\|a_0\|_{\mathcal{H}} + \delta_{n,\zeta}/\lambda)^2$, $\mathcal{A}_B \cdot \mathcal{F}^i := \{a \cdot f : a \in \mathcal{A}_B, f \in \mathcal{F}^i\}$ and

$$\epsilon_{n,\lambda,\zeta} = \max_i \left(\mathcal{R}(\mathcal{A}_{B_{n,\lambda,\zeta}} \cdot \mathcal{F}^i) + \mathcal{R}(m \circ \mathcal{F}^i)\right) + c_0\sqrt{\frac{\log(c_1/\zeta)}{n}}$$

for some universal constants c_0, c_1, then again by Theorem 26.5 and 26.9 of [115],

$$\forall a \in \mathcal{A}_{B_{n,\lambda,\zeta}}, f \in \mathcal{F}_U^i \; |\Psi_n(a, f) - \Psi(a, f)| \le \epsilon_{n,\lambda,\zeta}$$

By a union bound over the d function classes composing $\mathcal{F}$, we have that w.p. $1-2\zeta$:

$$\sup_{f\in\mathcal{F}} \Psi_n(a_0, f) \le \sup_{f\in\mathcal{F}} \Psi(a_0, f) + \delta_{n,\zeta/d} = \delta_{n,\zeta/d}$$

and

$$\sup_{f\in\mathcal{F}} \Psi_n(\hat{a}, f) \ge \sup_{f\in\mathcal{F}} \Psi(\hat{a}, f) - \epsilon_{n,\lambda,\zeta/d}$$

If $\|\hat{a} - a_0\|_2 \le \delta_{n,\zeta}$, then the theorem follows immediately. Thus we consider the case when $\|\hat{a} - a_0\|_2 \ge \delta_{n,\zeta}$. Since, by assumption, for any $a \in \mathcal{A}_B$ with $\|a - a_0\| \ge \delta_{n,\zeta}$ it holds that

$\frac{a_0-a}{\|a_0-a\|_2} \in \text{span}_\kappa(\mathcal{F})$, we have $\frac{a_0-\hat{a}}{\|a_0-\hat{a}\|_2} = \sum_{i=1}^p w_i f_i$, with $p < \infty$, $\|w\|_1 \leq \kappa$ and $f_i \in \mathcal{F}$. Thus:

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \Psi(\hat{a}, f) \geq\ & \frac{1}{\kappa}\sum_{i=1}^p w_i \Psi(\hat{a}, f_i) = \frac{1}{\kappa}\Psi\left(\hat{a}, \sum_i w_i f_i\right) \\
=\ & \frac{1}{\kappa}\frac{1}{\|\hat{a}-a_0\|_2}\Psi(\hat{a}, a_0-\hat{a}) \\
=\ & \frac{1}{\kappa}\frac{1}{\|\hat{a}-a_0\|_2}\mathbb{E}[(a_0(X)-\hat{a}(X))^2] \\
=\ & \frac{1}{\kappa}\|\hat{a}-a_0\|_2
\end{aligned}$$

Combining all the above we have, w.p. $1-2\zeta$:

$$\|\hat{a}-a_0\|_2 \leq \kappa\left(\epsilon_{n,\lambda,\zeta/d} + \delta_{n,\zeta/d} + \lambda\left(\|a_0\|_{\mathcal{A}} - \|\hat{a}\|_{\mathcal{A}}\right)\right)$$

Moreover, since functions in $\mathcal{A}$ and $\mathcal{F}$ are bounded in $[-1,1]$, we have that the function $a \cdot f$ is 1-Lipschitz with respect to the vector of functions (a, f). Thus we can apply a vector version of the contraction inequality [85] to get that:

$$\mathcal{R}(\mathcal{A}_{B_{n,\lambda,z}} \cdot \mathcal{F}^i) \leq 2\left(\mathcal{R}(\mathcal{A}_{B_{n,\lambda,z}}) + \mathcal{R}(\mathcal{F}^i)\right)$$

Finally, we have that since $\mathcal{A}$ is star-convex:

$$\mathcal{R}(\mathcal{A}_{B_{n,\lambda,z}}) \leq \sqrt{B_{n,\lambda,z}}\,\mathcal{R}(\mathcal{A}_1)$$

Leading to the final bound of:

$$\begin{aligned}
\|\hat{a}-a_0\|_2 \leq \kappa\left(2\left(\|a_0\|_{\mathcal{A}} + \delta_{n,\zeta}/\lambda\right)\mathcal{R}(\mathcal{A}_1) + 2\max_{i=1}^d\left(\mathcal{R}(\mathcal{F}^i) + \mathcal{R}(m \circ \mathcal{F}^i)\right)\right) \\
+ \kappa\left(c_0\sqrt{\frac{\log(c_1\, d/\zeta)}{n}} + \lambda\left(\|a_0\|_{\mathcal{A}} - \|\hat{a}\|_{\mathcal{A}}\right)\right)
\end{aligned}$$

Since $\lambda \geq \delta_{n,\zeta}$, we get the result. $\square$

D.3 Proof of Corollary 5

Proof. Consider any $\hat{a} = \langle\hat{\theta}, \cdot\rangle \in \mathcal{A}_{B_{n,\lambda,\zeta}}$ and let $\nu = \hat{\theta} - \theta_0$, then:

$$\delta_{n,\zeta}/\lambda + \|\theta_0\|_1 \geq \|\hat{\theta}\|_1 = \|\theta_0 + \nu\|_1 = \|\theta_0 + \nu_S\|_1 + \|\nu_{S^c}\|_1 \geq \|\theta_0\|_1 - \|\nu_S\|_1 + \|\nu_{S^c}\|_1$$

Thus:

$$\|\nu_{S^c}\|_1 \leq \|\nu_S\|_1 + \delta_{n,\zeta}/\lambda$$

and ν lies in the restricted cone for which the restricted eigenvalue of V holds. Moreover, since $|S| = s$:

$$\|\nu\|_1 \leq 2\|\nu_S\|_1 + \delta_{n,\zeta}/\lambda \leq 2\sqrt{s}\|\nu_S\|_2 + \delta_{n,\zeta}/\lambda \leq 2\sqrt{s}\|\nu\|_2 + \delta_{n,\zeta}/\lambda \leq 2\sqrt{\frac{s}{\gamma}\nu^\top V\nu} + \delta_{n,\zeta}/\lambda$$

Moreover, observe that:

$$\|\hat{a} - a_0\|_2 = \sqrt{\mathbb{E}[\langle \nu, x\rangle^2]} = \sqrt{\nu^\top V \nu}$$

Thus we have:

$$\frac{\hat{a}(x) - a_0(x)}{\|\hat{a} - a_0\|_2} = \sum_{i=1}^{p} \frac{\nu_i}{\sqrt{\nu^\top V\nu}} x_i$$

Thus for any $\hat{a} \in \mathcal{A}_{B_{n,\lambda,\zeta}}$, we can write $\frac{\hat{a}-a_0}{\|\hat{a}-a_0\|_2}$ as $\sum_{i=1}^p w_i f_i$, with $f_i \in \mathcal{F}$ and:

$$\|w\|_1 = \frac{\|\nu\|_1}{\sqrt{\nu^\top V\nu}} \le 2\sqrt{\frac{s}{\gamma}} + \frac{\delta_{n,\zeta}}{\lambda} \frac{1}{\|\hat{a} - a_0\|_2}.$$

Thus: $\frac{\hat{a}-a_0}{\|\hat{a}-a_0\|_2} \in \text{span}_\kappa(\mathcal{F})$ for $\kappa = 2\sqrt{\frac{s}{\gamma}} + \frac{\delta_{n,\zeta}}{\lambda}\frac{1}{\|\hat{a}-a_0\|_2}$.

Moreover, observe that by the triangle inequality:

$$\|a_0\|_{\mathcal{A}} - \|\hat{a}\|_{\mathcal{A}} = \|\theta_0\|_1 - \|\hat{\theta}\|_1 \le \|\theta_0 - \hat{\theta}\|_1 = \|\nu\|_1 \le 2\sqrt{\frac{s}{\gamma}\nu^\top V\nu} + \delta_{n,\zeta}/\lambda$$

Moreover, by standard results on the Rademacher complexity of linear function classes (see e.g. Lemma 26.11 of [115]), we have $\mathcal{R}(\mathcal{A}_B) \le B\sqrt{\frac{2\log(2\,p)}{n}} \max_{x\in\mathcal{X}} \|x\|_\infty$ and $\mathcal{R}(\mathcal{F}^i), \mathcal{R}(m \circ \mathcal{F}^i) \le \sqrt{\frac{2\log(2)}{n}} \max_{x\in\mathcal{X}} \|x\|_\infty$ (the latter via the fact that each $\mathcal{F}^i$; and therefore also $m \circ \mathcal{F}^i$; contains only two elements and invoking Masart's lemma). Thus invoking Theorem 3:

$$\|\hat{a} - a_0\|_2 \le \left(2\sqrt{\frac{s}{\gamma}} + \frac{\delta_{n,\zeta}}{\lambda}\frac{1}{\|\hat{a} - a_0\|_2}\right) \cdot \left(2(\|\theta_0\|_1 + 1)\sqrt{\frac{\log(2p)}{n}} + \delta_{n,\zeta} + \lambda\sqrt{\frac{s}{\gamma}}\|\hat{a} - a_0\|_2\right)$$

The right hand side is upper bounded by the sum of the following four terms:

$$\begin{aligned}
Q_1 &:= 2\sqrt{\frac{s}{\gamma}}\left(2(\|\theta_0\|_1 + 1)\sqrt{\frac{\log(2p)}{n}} + \delta_{n,\zeta}\right)\\
Q_2 &:= \left(\frac{\delta_{n,\zeta}}{\lambda}\frac{1}{\|\hat{a} - a_0\|_2}\right)\left(2(\|\theta_0\|_1 + 1)\sqrt{\frac{\log(2p)}{n}} + \delta_{n,\zeta}\right)\\
Q_3 &:= 2\lambda\frac{s}{\gamma}\|\hat{a} - a_0\|_2\\
Q_4 &:= \delta_{n,\zeta}\sqrt{\frac{s}{\gamma}}
\end{aligned}$$

If $\|\hat{a} - a_0\|_2 \ge \sqrt{\frac{s}{\gamma}}\delta_{n,\zeta}$ and setting $\lambda \le \frac{\gamma}{8s}$, yields:

$$\begin{aligned}
Q_2 &\le 8\frac{1}{\lambda}\sqrt{\frac{\gamma}{s}}\left(2(\|\theta_0\|_1 + 1)\sqrt{\frac{\log(2p)}{n}} + \delta_{n,\zeta}\right)\\
Q_3 &\le \frac{1}{4}\|\hat{a} - a_0\|_2
\end{aligned}$$

Thus bringing Q_3 on the left-hand-side and dividing by $3/4$, we have:

$$\|\hat{a} - a_0\|_2 \leq \frac{4}{3}(Q_1 + Q_2 + Q_4) = \frac{4}{3} \max\left\{\sqrt{\frac{s}{\gamma}}, \frac{1}{\lambda}\sqrt{\frac{\gamma}{s}}\right\} \left(20\left(\|\theta_0\|_1 + 1\right)\sqrt{\frac{\log(2p)}{n}} + 11\delta_{n,\zeta}\right)$$

On the other hand if $\|\hat{a} - a_0\|_2 \leq \sqrt{\frac{s}{\gamma}}\delta_{n,\zeta}$, then the latter inequality trivially holds. Thus it always holds. $\square$

E Proofs from Section 5

E.1 Proof of Proposition 8

Proposition 23. *Consider an online linear optimization algorithm over a convex strategy space S and consider the OFTRL algorithm with a 1-strongly convex regularizer with respect to some norm $\|\cdot\|$ on space S:*

$$f_t = \arg\min_{f \in S} f^\top \left(\sum_{\tau \leq t} \ell_\tau + \ell_t\right) + \frac{1}{\eta} R(f)$$

Let $\|\cdot\|_*$ denote the dual norm of $\|\cdot\|$ and $R = \sup_{f \in S} R(f) - \inf_{f \in S} R(f)$. Then for any $f^* \in S$:

$$\sum_{t=1}^{T} (f_t - f^*)^\top \ell_t \leq \frac{R}{\eta} + \eta \sum_{t=1}^{T} \|\ell_t - \ell_{t-1}\|_* - \frac{1}{4\eta} \sum_{t=1}^{T} \|f_t - f_{t-1}\|^2$$

Proof. The proof follows by observing that Proposition 7 in [120] holds verbatim for any convex strategy space S and not necessarily the simplex. $\square$

Proposition 24. *Consider a minimax objective: $\min_{\theta \in \Theta} \max_{w \in W} \ell(\theta, w)$. Suppose that Θ, W are convex sets and that $\ell(\theta, w)$ is convex in θ for every w and concave in θ for any w. Let $\|\cdot\|_\Theta$ and $\|\cdot\|_W$ be arbitrary norms in the corresponding spaces. Moreover, suppose that the following Lipschitzness properties are satisfied:*

$$\forall \theta \in \Theta, w, w' \in W : \|\nabla_\theta \ell(\theta, w) - \nabla_\theta \ell(\theta, w')\|_{\Theta,*} \leq L\|w - w'\|_W$$
$$\forall w \in W, \theta, \theta' \in \Theta : \|\nabla_w \ell(\theta, w) - \nabla_w \ell(\theta', w)\|_{W,*} \leq L\|\theta - \theta'\|_\Theta$$

where $\|\cdot\|_{\Theta,*}$ and $\|\cdot\|_{W,*}$ correspond to the dual norms of $\|\cdot\|_\Theta, \|\cdot\|_W$. Consider the algorithm where at each iteration each player updates their strategy based on:

$$\theta_{t+1} = \arg\min_{\theta \in \Theta} \theta^\top \left(\sum_{\tau \leq t} \nabla_\theta \ell(\theta_\tau, w_\tau) + \nabla_\theta \ell(\theta_t, w_t)\right) + \frac{1}{\eta} R_{\min}(\theta)$$
$$w_{t+1} = \arg\max_{w \in W} w^T \left(\sum_{\tau \leq t} \nabla_w \ell(\theta_\tau, w_\tau) + \nabla_w \ell(\theta_t, w_t)\right) - \frac{1}{\eta} R_{\max}(w)$$

such that $R_{\min}$ is 1-strongly convex in the set Θ with respect to norm $\|\cdot\|_\Theta$ and $R_{\max}$ is 1-strongly convex in the set W with respect to norm $\|\cdot\|_W$ and with any step-size $\eta \leq \frac{1}{4L}$. Then the parameters $\bar{\theta} = \frac{1}{T}\sum_{t=1}^T \theta_t$ and $\bar{w} = \frac{1}{T}\sum_{t=1}^T w_t$ correspond to an $\frac{2R_*}{\eta\cdot T}$-approximate equilibrium and hence $\bar{\theta}$ is a $\frac{4R_*}{\eta T}$-approximate solution to the minimax objective, where R is defined as:

$$R_* := \max\left\{\sup_{\theta\in\Theta} R_{\min}(\theta) - \inf_{\theta\in\Theta} R_{\min}(\theta), \sup_{w\in W} R_{\max}(w) - \inf_{w\in W} R_{\max}(w)\right\}$$

Proof. The proposition is essentially a re-statement of Theorem 25 of [120] (which in turn is an adaptation of Lemma 4 of [102]), specialized to the case of the OFTRL algorithm and to the case of a two-player convex-concave zero-sum game, which implies that the if the sum of regrets of players is at most ϵ, then the pair of average solutions corresponds to an ϵ-equilibrium (see e.g. [53] and Lemma 4 of [102]). $\square$

Proof of Proposition 8 Let $R_E(x) = \sum_{i=1}^{2p} x_i \log(x_i)$. For the space $\Theta := \{\rho \in \mathbb{R}^{2p} : \rho \geq 0, \|\rho\|_1 \leq B\}$, the entropic regularizer is $\frac{1}{B}$-strongly convex with respect to the ℓ_1 norm and hence we can set $R_{\min}(\rho) = B\, R_E(\rho)$. Similarly, for the space $W := \{w \in \mathbb{R}^{2p} : w \geq 0, \|w\|_1 = 1\}$, the entropic regularizer is 1-strongly convex with respect to the ℓ_1 norm and thus we can set $R_{\max}(w) = R_E(w)$. For this choice of regularizers, the update rules can be easily verified to have a closed form solution provided in Proposition 8, by writing the Lagrangian of each OFTRL optimization problem and invoking strong duality. Further, we can verify the lipschitzness conditions. Since the dual of the ℓ_1 norm is the ℓ_∞ norm, $\nabla_\rho \ell(\rho, w) = \mathbb{E}_n[VV^\top]w + \lambda$ and thus:

$$\begin{aligned}
\|\nabla_\rho \ell(\rho, w) - \nabla_\rho \ell(\rho, w')\|_\infty &= \|\mathbb{E}_n[VV^\top](w - w')\|_\infty \leq \|\mathbb{E}_n[VV^\top]\|_\infty \|w - w'\|_1\\
\|\nabla_w \ell(\rho, w) - \nabla_w \ell(\rho', w)\|_\infty &= \|\mathbb{E}_n[VV^\top](\rho - \rho')\|_\infty \leq \|\mathbb{E}_n[VV^\top]\|_\infty \|\rho - \rho'\|_1
\end{aligned}$$

Thus we have $L = \|\mathbb{E}_n[VV^\top]\|_\infty$. Finally, observe that:

$$\begin{aligned}
\sup_{\rho\in\Theta} B\, R_E(\rho) - \inf_{\rho\in\Theta} B\, R_E(\rho) &= B^2 \log(B \vee 1) + B \log(2p)\\
\sup_{w\in W} R_E(w) - \inf_{w\in W} R_E(w) &= \log(2p)
\end{aligned}$$

Thus we can take $R_* = B^2 \log(B \vee 1) + (B + 1)\log(2p)$. Thus if we set $\eta = \frac{1}{4\|\mathbb{E}_n[VV^\top]\|_\infty}$, then we have that after T iterations, $\bar{\theta} = \bar{\rho}^+ - \bar{\rho}^-$ is an $\epsilon(T)$-approximate solution to the minimax problem, with

$$\epsilon(T) = 16\|\mathbb{E}_n[VV^\top]\|_\infty \frac{4B^2 \log(B \vee 1) + (B + 1)\log(2p)}{T}.$$

Combining all the above with Proposition 24 yields the proof of Proposition 8.

E.2 Proof of Proposition 15

Observe that the loss function $-\ell(a, \cdot)$ is strongly convex in f with respect to the $\|\cdot\|_{2,n}$ norm, i.e.:

$$-\frac{1}{2} D_{ff}\ell(a, f)[\nu, \nu] \geq \mathbb{E}_n[\nu(X)^2]$$

and that the difference:

$$\ell(a, f) - \ell(a', f) = \mathbb{E}_n[(a(X) - a'(X)) \cdot f(X)]$$

is an $\|a - a'\|_{2,n}$-Lipschitz function with respect to the $\ell_{2,n}$ norm (via a Cauchy-Schwarz inequality). Thus we can conclude that (see Lemma 1 in [1]):

$$\|f_t - f_{t+1}\|_{2,n} \leq \|\bar{a}_{<t} - \bar{a}_{<t+1}\|_{2,n}$$

Moreover, we know that the cumulative regret of the FTL algorithm is at most (see proof of Theorem 1 in [1]):

$$R(T) \leq \sum_{t=1}^{T} |\ell(a_t, f_t) - \ell(a_t, f_{t+1})|$$

Since $\|a_t\|_\infty, \|f_t\|_\infty \leq 1$, each summand of the latter is upper bounded by:

$$|\mathbb{E}_n[m(Z; f_t - f_{t+1})]| + 3\|f_t - f_{t+1}\|_{1,n}$$

We will assume that the empirical operator $E_n[m(Z; f)]$ is also a bounded linear operator, with a bound of M_n. Thus we have:

$$|\mathbb{E}_n[m(Z; f_t - f_{t+1})]| \leq M_n \|f_t - f_{t+1}\|_{2,n}$$

Thus overall we get:

$$|\ell(a_t, f_t) - \ell(a_t, f_{t+1})| \leq (M_n + 3)\|f_t - f_{t+1}\|_{2,n} \leq (M_n + 3)\|\bar{a}_{<t} - \bar{a}_{<t+1}\|_{2,n} \leq \frac{2\,(M_n + 3)}{t}$$

where we used the fact that $|\bar{a}_{<t}(X) - \bar{a}_{<t+1}(X)| \leq \frac{2}{t}$, since $\|a\|_\infty \leq 1$. Thus we conclude that:

$$R(T) \leq 2\,(M_n + 3) \sum_{t=1}^{T} \frac{1}{t} = O(M_n\, \log(T))$$

Thus after $T = \Theta\left(\frac{M_n\, \log(1/\epsilon)}{\epsilon}\right)$ iterations, of the algorithm, the f-player has regret of at most ϵ. By standard results in solving convex-concave zero-sum games, this then implies that the average solutions: $f_* = \frac{1}{T}\sum_{t=1}^{T} f_t$ and $a_* = \frac{1}{T}\sum_{t=1}^{T} a_t$ are an ϵ-equilibrium and therefore also that a_* is an ϵ-approximate solution to the minimax problem. This concludes the proof of the proposition.

E.3 Proof of Proposition 9

Proof. For example for ATE

$$\begin{aligned}
[K^{(3)}]_{ij} &= [\Phi^{(m)}\Phi']_{ij} \\
&= \langle M^*\phi(x_i), \phi(x_j)\rangle \\
&= \langle \phi(x_i), M\phi(x_j)\rangle \\
&= \langle \phi(d_i, w_i), \phi(1, w_j) - \phi(0, w_j)\rangle \\
&= k((d_i, w_i), (1, w_j)) - k((d_i, w_i), (0, w_j))
\end{aligned}$$

Likewise

$$
\begin{aligned}
[K^{(4)}]_{ij} &= [\Phi^{(m)}(\Phi^{(m)})']_{ij} \\
&= \langle M^*\phi(x_i), M^*\phi(x_j)\rangle \\
&= \langle \phi(x_i), MM^*\phi(x_j)\rangle \\
&= \langle \phi(x_i), M^*\phi(1,w_j) - M^*\phi(0,w_j)\rangle \\
&= \langle M\phi(x_i), \phi(1,w_j) - \phi(0,w_j)\rangle \\
&= \langle \phi(1,w_i) - \phi(0,w_i), \phi(1,w_j) - \phi(0,w_j)\rangle \\
&= k((1,w_i),(1,w_j)) - k((1,w_i),(0,w_j)) - k((0,w_i),(1,w_j)) + k((0,w_i),(0,w_j))
\end{aligned}
$$

□

E.4 Proof of Proposition 10

Proof. Write the objective as

$$
\mathcal{E}_1(f) := \frac{1}{n}\sum_{i=1}^{n}\langle f, M^*\phi(x_i)\rangle_{\mathcal{H}} - a(x_i)\langle f, \phi(x_i)\rangle_{\mathcal{H}} - \langle f, \phi(x_i)\rangle_{\mathcal{H}}^2 - \lambda\|f\|_{\mathcal{H}}^2
$$

Recall that for an RKHS, evaluation is a continuous functional represented as the inner product with the feature map. Due to the ridge penalty, the stated objective is coercive and strongly convex w.r.t f. Hence it has a unique maximizer $\hat{f}$ that obtains the maximum.

Write $\hat{f} = \hat{f}_n + \hat{f}_n^{\perp}$ where $\hat{f}_n \in row(\Psi)$ and $\hat{f}_n^{\perp} \in null(\Psi)$. Substituting this decomposition of $\hat{f}$ into the objective, we see that

$$
\mathcal{E}_1(\hat{f}) = \mathcal{E}_1(\hat{f}_n) - \lambda\|\hat{f}_n^{\perp}\|_{\mathcal{H}}^2
$$

Therefore

$$
\mathcal{E}_1(\hat{f}) \leq \mathcal{E}_1(\hat{f}_n)
$$

Since $\hat{f}$ is the unique maximizer, $\hat{f} = \hat{f}_n$. □

E.5 Proof of Proposition 11

Proof. Write the objective as

$$
\begin{aligned}
\mathcal{E}_1(f) &= \frac{1}{n}\sum_{i=1}^{n}\langle Mf, \phi(x_i)\rangle - \langle a, \phi(x_i)\rangle\langle f, \phi(x_i)\rangle - \langle f, \phi(x_i)\rangle^2 - \lambda\langle f, f\rangle \\
&= f'M'\hat{\mu} - f'\hat{T}a - f'\hat{T}f - \lambda f'f
\end{aligned}
$$

where $\hat{\mu} := \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)$ and $\hat{T} := \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\otimes\phi(x_i)$. Appealing to the representer theorem

$$
\begin{aligned}
\mathcal{E}_1(\gamma) &= \gamma'\Psi M'\hat{\mu} - \gamma'\Psi\hat{T}a - \gamma'\Psi\hat{T}\Psi'\gamma - \lambda\gamma'\Psi\Psi'\gamma \\
&= \gamma'\Psi M'\hat{\mu} - \frac{1}{n}\gamma'\begin{bmatrix} K^{(1)} \\ K^{(3)} \end{bmatrix}\Phi a - \frac{1}{n}\gamma'\begin{bmatrix} K^{(1)}K^{(1)} & K^{(1)}K^{(2)} \\ K^{(3)}K^{(1)} & K^{(3)}K^{(2)} \end{bmatrix}\gamma - \lambda\gamma'K\gamma
\end{aligned}
$$

The FOC yields

$$\Psi M'\hat{\mu} - \frac{1}{n}\begin{bmatrix} K^{(1)} \\ K^{(3)} \end{bmatrix} \Phi a - \frac{2}{n}\begin{bmatrix} K^{(1)}K^{(1)} & K^{(1)}K^{(2)} \\ K^{(3)}K^{(1)} & K^{(3)}K^{(2)} \end{bmatrix} \hat{\gamma} - 2\lambda K \hat{\gamma} = 0$$

Hence

$$\begin{aligned}
\hat{\gamma} &= \frac{1}{2}\left[\frac{1}{n}\begin{bmatrix} K^{(1)}K^{(1)} & K^{(1)}K^{(2)} \\ K^{(3)}K^{(1)} & K^{(3)}K^{(2)} \end{bmatrix} + \lambda K\right]^{-1} \left[\Psi M'\hat{\mu} - \frac{1}{n}\begin{bmatrix} K^{(1)} \\ K^{(3)} \end{bmatrix} \Phi a\right] \\
&= \frac{1}{2}\left[\begin{bmatrix} K^{(1)}K^{(1)} & K^{(1)}K^{(2)} \\ K^{(3)}K^{(1)} & K^{(3)}K^{(2)} \end{bmatrix} + n\lambda K\right]^{-1} \left[n\Psi M'\hat{\mu} - \begin{bmatrix} K^{(1)} \\ K^{(3)} \end{bmatrix} \Phi a\right]
\end{aligned}$$

□

E.6 Proof of Proposition 12

Proof. Observe that

$$\begin{aligned}
\hat{f}(x) &= \langle \hat{f}, \phi(x)\rangle = \phi(x)'\Psi'\hat{\gamma} = \frac{1}{2}\phi(x)'\Psi'\Delta^{-1}\left[n\Psi M'\hat{\mu} - \begin{bmatrix} K^{(1)} \\ K^{(3)} \end{bmatrix} \Phi a\right] \\
m(x;\hat{f}) &= \langle M\hat{f}, \phi(x)\rangle = \frac{1}{2}\phi(x)'M\Psi'\Delta^{-1}\left[n\Psi M'\hat{\mu} - \begin{bmatrix} K^{(1)} \\ K^{(3)} \end{bmatrix} \Phi a\right] \\
\|\hat{f}\|^2_{\mathcal{H}} &= \hat{\gamma}'\Psi\Psi'\hat{\gamma} = \frac{1}{4}\Delta^{-1}\left[n\Psi M'\hat{\mu} - \begin{bmatrix} K^{(1)} \\ K^{(3)} \end{bmatrix} \Phi a\right]' \Delta^{-1}K\Delta^{-1}\left[n\Psi M'\hat{\mu} - \begin{bmatrix} K^{(1)} \\ K^{(3)} \end{bmatrix} \Phi a\right]
\end{aligned}$$

Write the objective as

$$\mathcal{E}_2(a) = \frac{1}{n}\sum_{i=1}^{n} m(x_i;\hat{f}) - \langle a, \phi(x_i)\rangle \hat{f}(x_i) - \hat{f}(x_i)^2 - \lambda\|\hat{f}\|^2_{\mathcal{H}} + \mu\|a\|^2_{\mathcal{H}}$$

where the various terms involving $\hat{f}$ *only* depend on a in the form Φa. Due to the ridge penalty, the stated objective is coercive and strongly convex w.r.t a. Hence it has a unique maximizer $\hat{a}$ that obtains the maximum.

Write $\hat{a} = \hat{a}_n + \hat{a}_n^{\perp}$ where $\hat{a}_n \in row(\Phi)$ and $\hat{a}_n^{\perp} \in null(\Phi)$. Substituting this decomposition of $\hat{a}$ into the objective, we see that

$$\mathcal{E}_2(\hat{a}) = \mathcal{E}_2(\hat{a}_n) + \mu\|\hat{a}_n^{\perp}\|^2_{\mathcal{H}}$$

Therefore

$$\mathcal{E}_2(\hat{a}) \geq \mathcal{E}_2(\hat{a}_n)$$

Since $\hat{a}$ is the unique minimizer, $\hat{a} = \hat{a}_n$. □

E.7 Proof of Proposition 13

Proof. Write the objective as

$$\begin{aligned}
\mathcal{E}_2(a) &= \hat{f}'M'\hat{\mu} - \hat{f}'\hat{T}a - \hat{f}'\hat{T}\hat{f} - \lambda\hat{f}'\hat{f} + \mu a'a \\
\mathcal{E}_2(\beta) &= \hat{\gamma}'\Psi M'\hat{\mu} - \hat{\gamma}'\Psi\hat{T}\Phi'\beta - \hat{\gamma}'\Psi\hat{T}\Psi'\hat{\gamma} - \lambda\hat{\gamma}'\Psi\Psi'\hat{\gamma} + \mu\beta'\Phi\Phi'\beta \\
&= \hat{\gamma}'\Psi M'\hat{\mu} - \frac{1}{n}\hat{\gamma}'\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}\beta - \frac{1}{n}\hat{\gamma}'\begin{bmatrix} K^{(1)}K^{(1)} & K^{(1)}K^{(2)} \\ K^{(3)}K^{(1)} & K^{(3)}K^{(2)} \end{bmatrix}\hat{\gamma} - \lambda\hat{\gamma}'K\hat{\gamma} + \mu\beta'K^{(1)}\beta \\
&= \sum_{j=1}^{5} E_j
\end{aligned}$$

where

$$\begin{aligned}
E_1 &= \hat{\gamma}'\Psi M'\hat{\mu} \\
E_2 &= -\frac{1}{n}\hat{\gamma}'\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}\beta \\
E_3 &= -\frac{1}{n}\hat{\gamma}'\begin{bmatrix} K^{(1)}K^{(1)} & K^{(1)}K^{(2)} \\ K^{(3)}K^{(1)} & K^{(3)}K^{(2)} \end{bmatrix}\hat{\gamma} \\
E_4 &= -\lambda\hat{\gamma}'K\hat{\gamma} \\
E_5 &= \mu\beta'K^{(1)}\beta
\end{aligned}$$

Recall that

$$\hat{\gamma} = \frac{1}{2}\Delta^{-1}\left[n\Psi M'\hat{\mu} - \begin{bmatrix} K^{(1)} \\ K^{(3)} \end{bmatrix}\Phi a\right] = \frac{1}{2}\Delta^{-1}\left[n\Psi M'\hat{\mu} - \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}\beta\right]$$

Hence

$$\hat{\gamma}' = \frac{1}{2}\left[n\hat{\mu}'M\Psi' - \beta'\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}'\right]\Delta^{-1}$$

We analyze each term

1. E_1

$$E_1 = \frac{1}{2}\left[n\hat{\mu}'M\Psi' - \beta'\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}'\right]\Delta^{-1}\Psi M'\hat{\mu}$$

   Hence

$$\frac{\partial E_1}{\partial\beta} = -\frac{1}{2}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}'\Delta^{-1}\Psi M'\hat{\mu}$$

2. E_2

$$E_2 = \frac{1}{2n}\left[\beta' \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' - n\hat{\mu}' M\Psi' \right] \Delta^{-1} \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix} \beta$$

$$= \frac{1}{2n}\beta' \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1} \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix} \beta - \frac{1}{2}\hat{\mu}' M\Psi' \Delta^{-1} \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix} \beta$$

Hence

$$\frac{\partial E_2}{\partial \beta} = \frac{1}{n} \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1} \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix} \beta - \frac{1}{2} \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}\Psi M' \hat{\mu}$$

3. E_3

$$E_3 = -\frac{1}{4n}\left[n\Psi M'\hat{\mu} - \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix} \beta \right]' \Delta^{-1} \begin{bmatrix} K^{(1)}K^{(1)} & K^{(1)}K^{(2)} \\ K^{(3)}K^{(1)} & K^{(3)}K^{(2)} \end{bmatrix} \Delta^{-1} \left[n\Psi M'\hat{\mu} - \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix} \beta \right]$$

Note that

$$\frac{\partial}{\partial s}[x - As]'W[x - As] = -2A'W(x - As)$$

Therefore

$$\frac{\partial E_3}{\partial \beta} = \frac{1}{2n} \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1} \begin{bmatrix} K^{(1)}K^{(1)} & K^{(1)}K^{(2)} \\ K^{(3)}K^{(1)} & K^{(3)}K^{(2)} \end{bmatrix} \Delta^{-1} \left[n\Psi M'\hat{\mu} - \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix} \beta \right]$$

4. E_4

$$E_4 = -\frac{\lambda}{4}\left[n\Psi M'\hat{\mu} - \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix} \beta \right]' \Delta^{-1} K \Delta^{-1} \left[n\Psi M'\hat{\mu} - \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix} \beta \right]$$

Note that

$$\frac{\partial}{\partial s}[x - As]'W[x - As] = -2A'W(x - As)$$

Therefore

$$\frac{\partial E_4}{\partial \beta} = \frac{\lambda}{2} \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1} K \Delta^{-1} \left[n\Psi M'\hat{\mu} - \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix} \beta \right]$$

5. E_5

$$\frac{\partial E_1}{\partial \beta_1} = 2\mu \cdot K^{(1)}\beta$$

Collecting these results gives the FOC

$$
\begin{aligned}
0 &= -\frac{1}{2}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}\Psi M'\hat{\mu} \\
&+ \frac{1}{n}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}\beta - \frac{1}{2}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}\Psi M'\hat{\mu} \\
&+ \frac{1}{2n}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}\begin{bmatrix} K^{(1)}K^{(1)} & K^{(1)}K^{(2)} \\ K^{(3)}K^{(1)} & K^{(3)}K^{(2)} \end{bmatrix}\Delta^{-1}\left[n\Psi M'\hat{\mu} - \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}\beta\right] \\
&+ \frac{\lambda}{2}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}K\Delta^{-1}\left[n\Psi M'\hat{\mu} - \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}\beta\right] \\
&+ 2\mu \cdot K_{XX}\hat{\beta}
\end{aligned}
$$

Grouping terms

$$
\begin{aligned}
&\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}\Psi M'\hat{\mu} \\
&- \frac{1}{2n}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}\begin{bmatrix} K^{(1)}K^{(1)} & K^{(1)}K^{(2)} \\ K^{(3)}K^{(1)} & K^{(3)}K^{(2)} \end{bmatrix}\Delta^{-1}n\Psi M'\hat{\mu} \\
&- \frac{\lambda}{2}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}K\Delta^{-1}n\Psi M'\hat{\mu} \\
&= \frac{1}{n}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}\beta \\
&- \frac{1}{2n}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}\begin{bmatrix} K^{(1)}K^{(1)} & K^{(1)}K^{(2)} \\ K^{(3)}K^{(1)} & K^{(3)}K^{(2)} \end{bmatrix}\Delta^{-1}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}\beta \\
&- \frac{\lambda}{2}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}K\Delta^{-1}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}\beta \\
&+ 2\mu \cdot K^{(1)}\beta
\end{aligned}
$$

Define

$$
\Omega := \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' - \frac{1}{2}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}\begin{bmatrix} K^{(1)}K^{(1)} & K^{(1)}K^{(2)} \\ K^{(3)}K^{(1)} & K^{(3)}K^{(2)} \end{bmatrix} - \frac{n\lambda}{2}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}K
$$

We simplify each side of the equation

1. LHS

$$
\begin{aligned}
&\left\{\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' - \frac{1}{2}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}\begin{bmatrix} K^{(1)}K^{(1)} & K^{(1)}K^{(2)} \\ K^{(3)}K^{(1)} & K^{(3)}K^{(2)} \end{bmatrix} - \frac{n\lambda}{2}\begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1}K\right\}\Delta^{-1}\Psi M'\hat{\mu} \\
&= \Omega\Delta^{-1}\Psi M'\hat{\mu}
\end{aligned}
$$

2. RHS

$$\left\{ \left(\frac{1}{n} \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' - \frac{1}{2n} \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1} \begin{bmatrix} K^{(1)}K^{(1)} & K^{(1)}K^{(2)} \\ K^{(3)}K^{(1)} & K^{(3)}K^{(2)} \end{bmatrix} - \frac{\lambda}{2} \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix}' \Delta^{-1} K \right) \Delta^{-1} \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix} \right.$$
$$\left. + 2\mu \cdot K_{XX} \right\} \hat{\beta}$$
$$= \left\{ \frac{1}{n} \Omega \Delta^{-1} \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix} + 2\mu \cdot K^{(1)} \right\} \hat{\beta}$$

□

E.8 Proof of Corollary 14

Proof.

$$\begin{aligned} \hat{a}(x) &= \langle \hat{a}, \phi(x) \rangle \\ &= \phi(x)' \Phi' \hat{\beta} \\ &= K_{xX} \left\{ \frac{1}{n} \Omega \Delta^{-1} \begin{bmatrix} K^{(1)}K^{(1)} \\ K^{(3)}K^{(1)} \end{bmatrix} + 2\mu \cdot K^{(1)} \right\}^{-1} \Omega \Delta^{-1} \Psi M' \hat{\mu} \end{aligned}$$

What remains is an account of how to evaluate $V := \Psi M' \hat{\mu} \in \mathbb{R}^{2n}$.

There are two cases

1. $j \in [n]$

   Observe that the j-th element of V is

   $$v_j = \phi(x_j)' M' \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \phi(x_j)' M' \phi(x_i)$$

   Moreover

   $$\phi(x_j)' M' \phi(x_i) = \langle \phi(x_j), M^* \phi(x_i) \rangle = [K^{(2)}]_{ji}$$

   Therefore

   $$v_j = \frac{1}{n} \sum_{i=1}^{n} [K^{(2)}]_{ji}$$

2. $j \in \{n+1, ..., 2n\}$

   Observe that the j-th element of V is

   $$v_j = \phi(x_j)' M M' \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \phi(x_j)' M M' \phi(x_i)$$

   Moreover

   $$\phi(x_j)' M M' \phi(x_i) = \langle M^* \phi(x_j), M^* \phi(x_i) \rangle = [K^{(4)}]_{ji}$$

Therefore

$$v_j = \frac{1}{n}\sum_{i=1}^{n}[K^{(4)}]_{ji}$$

□

F Proofs from Section 6

F.1 Proof of Lemma 16

Proof. Observe that $\theta_0 = \mathbb{E}[m_a(Z; g_0)]$ for all a. Moreover, we can decompose:

$$\begin{aligned}\hat{\theta} - \theta_0 = & \ \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in P_k}\left(m_{\hat{a}_k}(Z_i;\hat{g}) - \mathbb{E}_Z[m_{\hat{a}_k}(Z;\hat{g}_k)]\right) + \frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}_Z[m_{\hat{a}_k}(Z;\hat{g}_k)] - \mathbb{E}_Z[m_{\hat{a}_k}(Z;g_0)]\right)\\ = & \ \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in P_k}\left(m_{\hat{a}_k}(Z_i;\hat{g}_k) - \mathbb{E}_Z[m_{\hat{a}_k}(Z;\hat{g}_k)]\right) + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_X[(a_0(X) - \hat{a}_k(X))\,(\hat{g}_k(X) - g_0(X))]\end{aligned}$$

Thus as long as $K = \Theta(1)$ and:

$$\sqrt{n}\mathbb{E}_X[(a_0(X) - \hat{a}_k(X))\,(\hat{g}_k(X) - g_0(X))] \to_p 0$$

we have that:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = \ \underbrace{\sqrt{n}\,\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in P_k}\left(m_{\hat{a}_k}(Z_i;\hat{g}_k) - \mathbb{E}_Z[m_{\hat{a}_k}(Z;\hat{g}_k)]\right)}_{A} + o_p(1)$$

Suppose that for some a_* and g_* (not necessarily equal to a_0 and g_0), we have that: $\|\hat{a}_k - a_*\|_2 \to_p 0$ and $\|\hat{g}_k - g_*\|_2 \to_p 0$. Then we can further decompose A as:

$$A = \ \mathbb{E}_n[m_{a_*}(Z;g_*)] - \mathbb{E}_Z[m_{a_*}(Z;g_*)] + \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in P_k}\underbrace{m_{\hat{a}_k}(Z_i;\hat{g}_k) - m_{a_*}(Z_i;g_*) - \mathbb{E}_Z[m_{\hat{a}_k}(Z;\hat{g}_k) - m_{a_*}(Z;g_*)]}_{V_i}$$

Denote with:

$$B := \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in P_k} V_i =: \frac{1}{n}\sum_{k=1}^{K} B_k$$

As long as $n\,\mathbb{E}[B^2] \to 0$, then we have that $\sqrt{n}B \to_p 0$. The second moment of each B_k is:

$$\mathbb{E}\left[B_k^2\right] = \sum_{i,j\in P_k}\mathbb{E}[V_iV_j] = \sum_{i,j\in P_k}\mathbb{E}[\mathbb{E}[V_iV_j \mid \hat{g}_k]] = \sum_{i\in P_k}\mathbb{E}\left[V_i^2\right]$$

where in the last equality we used the fact that due to cross-fitting, for any $i \neq j$, V_i is independent of V_j and mean-zero, conditional on the nuisance $\hat{g}_k$ estimated on the out-of-fold data for fold k. Moreover, by Jensen's inequality with respect to $\frac{1}{K}\sum_{k=1}^{K} B_k$

$$\mathbb{E}[B^2] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^{K} B_k\right)^2\right] = \frac{K^2}{n^2}\mathbb{E}\left[\left(\frac{1}{K}\sum_{k=1}^{K} B_k\right)^2\right] \leq \frac{K}{n^2}\sum_{k=1}^{K}\mathbb{E}[B_k^2] = \frac{K}{n^2}\sum_{k=1}^{K}\sum_{i\in P_k}\mathbb{E}[V_i^2] = \frac{K}{n^2}\sum_{i=1}^{n}\mathbb{E}[V_i^2]$$

Finally, observe that $\mathbb{E}[V_i^2] \to_p 0$, by mean-squared-continuity of the moment and by boundedness of the Riesz representer function class, the function class $\mathcal{G}$ and the variable Y. More elaborately:

$$\begin{aligned}\mathbb{E}[V_i^2] \leq\; & \mathbb{E}\left[\left(m_{\hat{a}}(Z_i;\hat{g}_k) - m_{a_*}(Z_i;g_*)\right)^2\right] \\ \leq\; & 2\,\mathbb{E}\left[\left(m(Z_i;\hat{g}_k) - m(Z_i;g_*)\right)^2\right] + 2\,\mathbb{E}[(\hat{a}_k(X)\,(Y-\hat{g}_k(X)) - a_*(X)\,(Y - g_*(X)))^2]\end{aligned}$$

The latter can further be bounded as:

$$4\mathbb{E}[(a_k(X) - a_*(X))^2\,(Y - g_k(X))^2] + 4\mathbb{E}[a_*(X)^2(g_*(X) - g_k(X))^2] \leq 4C\,\left(\mathbb{E}\left[\|\hat{a}_k - a_*\|_2^2 + \|\hat{g} - g_*\|_2^2\right]\right)$$

assuming that $(Y - \hat{g}_k(X))^2 \leq C$ and $a_*(X)^2 \leq C$ a.s.. Finally, by linearity of the operator and mean-squared continuity, we have:

$$\mathbb{E}[(m(Z_i;\hat{g}_k) - m(Z_i;g_*))^2] = \mathbb{E}[(m(Z_i;\hat{g}_k - g_*))^2] \leq M\,\mathbb{E}\left[\|\hat{g}_k - g_*\|_2^2\right]$$

Thus we have:

$$\mathbb{E}[V_i^2] \leq (2M + 4C)\,\left(\mathbb{E}\left[\|\hat{a}_k - a_*\|_2^2 + \|\hat{g} - g_*\|_2^2\right]\right) \to 0$$

Thus as long as $K = \Theta(1)$, we have that:

$$n\,\mathbb{E}[B^2] = \frac{K}{n}\sum_{i=1}^{n}\mathbb{E}[V_i^2] \leq (2M+4C)\,K\,\mathbb{E}\left[\|\hat{g} - g_*\|_2^2 + \|\hat{a} - a_*\|_2^2\right] \to 0$$

and we can conclude the result that:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = \sqrt{n}\left(\mathbb{E}_n[m_{a_*}(Z;g_*)] - \mathbb{E}_Z[m_{a_*}(Z;g_*)]\right) + o_p(1)$$

where the latter term can be easily argued, invoking the Central Limit Theorem, to be asymptotically normal $N(0,\sigma_*^2)$ with $\sigma_*^2 = \mathrm{Var}(m_{a_*}(Z;g_*))$. □

F.2 Proof of Normality without Consistency

Lemma 25. *Suppose that $K = \Theta(1)$ and that for some a_* and g_* (not necessarily equal to a_0 and g_0), we have that for all $k \in [K]$: $\|\hat{a}_k - a_*\|_2 \xrightarrow{L^2} 0$ and $\|\hat{g}_k - g_*\|_2 \xrightarrow{L^2} 0$. Assume that:*

$$\forall k \in [K] : \sqrt{n}\,\mathbb{E}[(a_*(X) - \hat{a}_k(X))\,(\hat{g}_k(X) - g_*(X))] \to_p 0$$

and that $\hat{g}_k$ admits an asymptotically linear representation around the truth g_0, i.e.:

$$\sqrt{|P_k|}\,(\hat{g}_k(X) - g_0(X)) = \frac{1}{\sqrt{|P_k|}}\sum_{i\in P_k}\psi(X, Z_i; g_0) + o_p(1)$$

with $\mathbb{E}[\psi(X, Z_i; g_0) \mid X] = 0$ *and let:*

$$\sigma_*^2 := Var_{Z_i}(m_{a_*}(Z_i; g_*) + \mathbb{E}_X\left[(a_0(X) - a_*(X))\,\psi(X, Z_i; g_0)\right])$$

Assume that Condition 1 is satisfied and the variables $Y, g(X), a(X)$ *are bounded a.s. for all* $g \in \mathcal{G}$ *and* $a \in \mathcal{A}$*. Then:*

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to_d N\left(0, \sigma_*^2\right)$$

Similarly, if $\hat{a}_k$ *has an asymptotically linear representation around the truth, then the statement above holds with:*

$$\sigma_*^2 := Var_{Z_i}(m_{a_*}(Z_i; g_*) + \mathbb{E}_X\left[\psi(X, Z_i; a_0)\,(g_0(X) - g_*(X))\right])$$

Proof. Observe that $\theta_0 = \mathbb{E}[m_a(Z; g_0)]$ for all a. Moreover, we can decompose:

$$\begin{aligned}
\hat{\theta} - \theta_0 = &\ \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in P_k}\left(m_{\hat{a}_k}(Z_i; \hat{g}) - \mathbb{E}[m_{\hat{a}_k}(Z; \hat{g}_k)]\right) + \frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}[m_{\hat{a}_k}(Z; \hat{g}_k)] - \mathbb{E}[m_{\hat{a}_k}(Z; g_0)]\right)\\
= &\ \underbrace{\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in P_k}\left(m_{\hat{a}_k}(Z_i; \hat{g}_k) - \mathbb{E}[m_{\hat{a}_k}(Z; \hat{g}_k)]\right)}_{A} + \underbrace{\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[(a_0(X) - \hat{a}_k(X))\,(\hat{g}_k(X) - g_0(X))]}_{C}
\end{aligned}$$

Suppose that for some a_* and g_* (not necessarily equal to a_0 and g_0), we have that: $\|\hat{a}_k - a_*\|_2 \to_p 0$ and $\|\hat{g}_k - g_*\|_2 \to_p 0$. Then we can further decompose A as:

$$A = \ \mathbb{E}_n[m_{a_*}(Z; g_*)] - \mathbb{E}[m_{a_*}(Z; g_*)] + \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in P_k}\underbrace{m_{\hat{a}_k}(Z_i; \hat{g}_k) - m_{a_*}(Z_i; g_*) - \mathbb{E}[m_{\hat{a}_k}(Z; \hat{g}_k) - m_{a_*}(Z; g_*)]}_{V_i}$$

Denote with:

$$B := \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in P_k} V_i =: \frac{1}{n}\sum_{k=1}^{K} B_k$$

As long as $n\,\mathbb{E}[B^2] \to 0$, then we have that $\sqrt{n}B \to_p 0$. The second moment of each B_k is:

$$\mathbb{E}[B_k^2] = \sum_{i,j\in P_k}\mathbb{E}[V_i V_j] = \sum_{i,j\in P_k}\mathbb{E}[\mathbb{E}[V_i V_j \mid \hat{g}_k]] = \sum_{i\in P_k}\mathbb{E}[V_i^2]$$

where in the last equality we used the fact that due to cross-fitting, for any $i \neq j$, V_i is independent of V_j and mean-zero, conditional on the nuisance $\hat{g}_k$ estimated on the out-of-fold data for fold k. Moreover, by Jensen's inequality with respect to $\frac{1}{K}\sum_{k=1}^{K} B_k$

$$\mathbb{E}[B^2] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^{K} B_k\right)^2\right] = \frac{K^2}{n^2}\mathbb{E}\left[\left(\frac{1}{K}\sum_{k=1}^{K} B_k\right)^2\right] \leq \frac{K}{n^2}\sum_{k=1}^{K}\mathbb{E}[B_k^2] = \frac{K}{n^2}\sum_{k=1}^{K}\sum_{i\in P_k}\mathbb{E}[V_i^2] = \frac{K}{n^2}\sum_{i=1}^{n}\mathbb{E}[V_i^2]$$

Finally, observe that $\mathbb{E}[V_i^2] \to_p 0$, by mean-squared-continuity of the moment and by boundedness of the Riesz representer function class, the function class $\mathcal{G}$ and the variable Y. More elaborately:

$$\begin{aligned}
\mathbb{E}[V_i^2] \leq\ & \mathbb{E}\left[(m_{\hat{a}}(Z_i;\hat{g}_k) - m_{a_*}(Z_i;g_*))^2\right] \\
\leq\ & 2\,\mathbb{E}\left[(m(Z_i;\hat{g}_k) - m(Z_i;g_*))^2\right] + 2\,\mathbb{E}[(\hat{a}_k(X)\,(Y-\hat{g}_k(X)) - a_*(X)\,(Y-g_*(X)))^2]
\end{aligned}$$

The latter can further be bounded as:

$$4\mathbb{E}[(a_k(X)-a_*(X))^2\,(Y-g_k(X))^2] + 4\mathbb{E}[a_*(X)^2(g_*(X)-g_k(X))^2] \leq 4C\,\mathbb{E}\left[\|\hat{a}_k - a_*\|_2^2 + \|\hat{g}-g_*\|_2^2\right]$$

assuming that $(Y-\hat{g}_k(X))^2 \leq C$ and $a_*(X)^2 \leq C$ a.s.. Finally, by linearity of the operator and mean-squared continuity, we have:

$$\mathbb{E}[(m(Z_i;\hat{g}_k) - m(Z_i;g_*))^2] = \mathbb{E}[(m(Z_i;\hat{g}_k - g_*))^2] \leq M\,\mathbb{E}\left[\|\hat{g}_k - g_*\|_2^2\right]$$

Thus we have:

$$\mathbb{E}[V_i^2] \leq (2M+4C)\,\mathbb{E}\left[\|\hat{a}_k - a_*\|_2^2 + \|\hat{g}-g_*\|_2^2\right] \to 0$$

Thus as long as $K=\Theta(1)$, we have that:

$$n\,\mathbb{E}[B^2] = \frac{K}{n}\sum_{i=1}^n \mathbb{E}[V_i^2] \leq (2M+4C)\,K\,\left(\|\hat{g}-g_*\|_2^2 + \|\hat{a}-a_*\|_2^2\right) \to_p 0$$

and we can conclude the result that:

$$\sqrt{n}\,A = \sqrt{n}\,(\mathbb{E}_n[m_{a_*}(Z;g_*)] - \mathbb{E}[m_{a_*}(Z;g_*)]) + o_p(1)$$

Now we analyze term C. We will prove one of the two conditions in the "or" statement, when $\hat{g}_k$ has an asymptotically linear representation, i.e.

$$\sqrt{|P_k|}\,(\hat{g}_k(X) - g_0(X)) = \frac{1}{\sqrt{|P_k|}}\sum_{i\in P_k}\psi(X,Z_i;g_0) + o_p(1)$$

with $\mathbb{E}[\psi(X,Z_i;g_0)\mid X]=0$. The case when $\hat{a}_k$ is asymptotically linear can be proved analogously. Let:

$$C_k := \mathbb{E}[(a_0(X)-\hat{a}_k(X))\,(\hat{g}_k(X)-g_0(X))]$$

We can then write:

$$C_k = \mathbb{E}[(a_*(X)-\hat{a}_k(X))\,(\hat{g}_k(X)-g_0(X))] + \mathbb{E}[(a_0(X)-a_*(X))\,(\hat{g}_k(X)-g_0(X))]$$

Since:

$$\sqrt{|P_k|}\mathbb{E}[(a_*(X)-\hat{a}_k(X))\,(\hat{g}_k(X)-g_0(X))] \leq \sqrt{|P_k|}\|a_*-\hat{a}_k\|_2\,\|\hat{g}_k-g_0\|_2 = \|a_*-\hat{a}_k\|_2\,O_p(1) = o_p(1)$$

we have that:

$$\begin{aligned}
\sqrt{|P_k|}C_k =\ & \sqrt{|P_k|}\mathbb{E}[(a_0(X)-a_*(X))\,(\hat{g}_k(X)-g_0(X))] + o_p(1) \\
=\ & \frac{1}{\sqrt{|P_k|}}\sum_{i\in P_k}\mathbb{E}_X[(a_0(X)-a_*(X))\,\psi(X,Z_i;g_0)] + o_p(1)
\end{aligned}$$

Since $K = \Theta(1)$ and $n/|P_k| \to K$, we then also have that:

$$
\begin{aligned}
\sqrt{n}C &= \frac{\sqrt{n}}{K}\sum_{k=1}^{K} C_k = \frac{\sqrt{K}}{K}\sum_{k=1}^{K}\sqrt{|P_k|}C_k + o(1) \\
&= \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{1}{\sqrt{|P_k|}}\sum_{i\in P_k}\mathbb{E}_X[(a_0(X) - a_*(X))\,\psi(X, Z_i; g_0)] + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i\in P_k}\mathbb{E}_X[(a_0(X) - a_*(X))\,\psi(X, Z_i; g_0)] + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i\in[n]}\mathbb{E}_X[(a_0(X) - a_*(X))\,\psi(X, Z_i; g_0)] + o_p(1) \\
&= \sqrt{n}\mathbb{E}_n\left[\mathbb{E}_X[(a_0(X) - a_*(X))\,\psi(X, Z_i; g_0)]\right] + o_p(1)
\end{aligned}
$$

$$
\sqrt{n}\left(\hat{\theta} - \theta_0\right) = \sqrt{n}\left(\mathbb{E}_n\left[m_{a_*}(Z; g_*) + \mathbb{E}_X\left[(a_0(X) - a_*(X))\,\psi(X, Z_i; g_0)\right]\right] - \mathbb{E}[m_{a_*}(Z; g_*)]\right) + o_p(1)
$$

where the latter term can be easily argued, invoking the Central Limit Theorem, to be asymptotically normal $N(0, \sigma_*^2)$ with $\sigma_*^2 = \mathrm{Var}_{Z_i}(m_{a_*}(Z_i; g_*) + \mathbb{E}_X\left[(a_0(X) - a_*(X))\,\psi(X, Z_i; g_0)\right])$. □

F.3 Proof of Lemma 17

Proof. Observe that $\theta_0 = \mathbb{E}[m_a(Z; g_0)]$ for all a. Moreover, we can decompose:

$$
\begin{aligned}
\hat{\theta} - \theta_0 &= \mathbb{E}_n[m_{\hat{a}}(Z; \hat{g})] - \mathbb{E}[m_{\hat{a}}(Z; \hat{g})] + \mathbb{E}[m_{\hat{a}}(Z; \hat{g})] - \mathbb{E}[m_{\hat{a}}(Z; g_0)] \\
&= \mathbb{E}_n[m_{\hat{a}}(Z; \hat{g})] - \mathbb{E}[m_{\hat{a}}(Z; \hat{g})] + \mathbb{E}[(a_0(X) - \hat{a}(X))\,(\hat{g}(X) - g_0(X))]
\end{aligned}
$$

Thus as long as $K = \Theta(1)$ and:

$$
\sqrt{n}\mathbb{E}[(a_0(X) - \hat{a}(X))\,(\hat{g}(X) - g_0(X))] \to_p 0
$$

we have that:

$$
\sqrt{n}\left(\hat{\theta} - \theta_0\right) = \sqrt{n}\,\underbrace{\mathbb{E}_n[m_{\hat{a}}(Z; \hat{g})] - \mathbb{E}[m_{\hat{a}}(Z; \hat{g})]}_{A} + o_p(1)
$$

Suppose that for some a_* and g_* (not necessarily equal to a_0 and g_0), we have that: $\|\hat{a}_k - a_*\|_2 \to_p 0$ and $\|\hat{g}_k - g_*\|_2 \to_p 0$. Then we can further decompose A as:

$$
A = \mathbb{E}_n[m_{a_*}(Z; g_*)] - \mathbb{E}[m_{a_*}(Z; g_*)] + \mathbb{E}_n\left[m_{\hat{a}}(Z; \hat{g}) - m_{a_*}(Z; g_*)\right] - \mathbb{E}[m_{\hat{a}}(Z; \hat{g}) - m_{a_*}(Z; g_*)]
$$

Let $\delta_{n,\zeta} = \delta_n + c_0\sqrt{\frac{\log(c_1/\zeta)}{n}}$, where δ_n upper bounds the critical radius of function classes $\mathcal{G}_B$ and $m \circ \mathcal{G}_B$ and $\mathcal{A}_B$, where B is set such that these sets contain functions that are bounded in $[-1, 1]$.

By a concentration inequality, almost identical to that of Equation (11), we have that w.p. $1-\zeta$: $\forall f \in \mathcal{F}, a \in \mathcal{A}$

$$\begin{aligned}|\mathbb{E}_n\left[m_a(Z;g) - m_{a_*}(Z;g_*)\right] - \mathbb{E}[m_a(Z;a) - m_{a_*}(Z;g_*)]| \\ \leq O\left(\delta_{n,\zeta}\left(\|m \circ (g-g_*)\|_2 \|\|a\|_{\mathcal{A}} + \|a-a_*\|_2\|g\|_{\mathcal{G}} + \|g-g_*\|_2\|a\|_{\mathcal{A}}\right) + \delta_{n,\zeta}^2\|a\|_{\mathcal{A}}\|g\|_{\mathcal{G}}\right)\end{aligned}$$

Applying the latter for $\hat{g}, \hat{a}$ and invoking the MSE continuity, w.p. $1-\zeta$:

$$\begin{aligned}|\mathbb{E}_n\left[m_{\hat{a}}(Z;\hat{g}) - m_{a_*}(Z;g_*)\right] - \mathbb{E}[m_{\hat{a}}(Z;\hat{g}) - m_{a_*}(Z;g_*)]| \\ \leq O\left(\delta_{n,\zeta} M\left(\|\hat{a}-a_*\|_2\,\|\hat{g}\|_{\mathcal{G}} + \|\hat{g}-g_*\|_2\,\|\hat{a}\|_{\mathcal{A}}\right) + \delta_{n,\zeta}^2\,\|\hat{g}\|_{\mathcal{G}}\,\|\hat{a}\|_{\mathcal{A}}\right)\end{aligned}$$

If we let $\delta_{n,*} = \delta_n + c_0\sqrt{\frac{c_1 n}{n}}$, then we have that:

$$\begin{aligned}|\mathbb{E}_n\left[m_{\hat{a}}(Z;\hat{g}) - m_{a_*}(Z;g_*)\right] - \mathbb{E}[m_{\hat{a}}(Z;\hat{g}) - m_{a_*}(Z;g_*)]| \\ = O_p\left(\delta_{n,*} M\left(\|\hat{a}-a_*\|_2\,\|\hat{g}\|_{\mathcal{G}} + \|\hat{g}-g_*\|_2\,\|\hat{a}\|_{\mathcal{A}}\right) + \delta_{n,*}^2\,\|\hat{g}\|_{\mathcal{G}}\,\|\hat{a}\|_{\mathcal{A}}\right)\end{aligned}$$

If $\|\hat{a}-a_*\|_2, \|\hat{g}-g_*\|_2 = O_p(r_n)$ and $\|\hat{a}\|_{\mathcal{A}}, \|\hat{g}\|_{\mathcal{G}} = O_p(1)$, we have that:

$$|\mathbb{E}_n\left[m_{\hat{a}}(Z;\hat{g}) - m_{a_*}(Z;g_*)\right] - \mathbb{E}[m_{\hat{a}}(Z;\hat{g}) - m_{a_*}(Z;g_*)]| = O_p\left(M\,\delta_{n,*} r_n + \delta_{n,*}^2\right)$$

Thus as long as: $\sqrt{n}\left(\delta_{n,*} r_n + \delta_{n,*}^2\right) \to 0$, we have that:

$$\sqrt{n}\,|\mathbb{E}_n\left[m_{\hat{a}}(Z;\hat{g}) - m_{a_*}(Z;g_*)\right] - \mathbb{E}[m_{\hat{a}}(Z;\hat{g}) - m_{a_*}(Z;g_*)]| = o_p(1)$$

Thus we conclude that:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = \sqrt{n}\left(\mathbb{E}_n[m_{a_*}(Z;g_*)] - \mathbb{E}[m_{a_*}(Z;g_*)]\right) + o_p(1)$$

where the latter term can be easily argued, invoking the Central Limit Theorem, to be asymptotically normal $N(0, \sigma_*^2)$ with $\sigma_*^2 = \mathrm{Var}(m_{a_*}(Z;g_*))$. □

F.4 Proof of Lemma 18

Proof. Let $h = (a, g)$ and $V(Z;h) = m_a(Z;g) - m_{a_*}(Z;g_*) - \mathbb{E}[m_a(Z;g) - m_{a_*}(Z;g_*)]$. We argue that:

$$\sqrt{n}\,\mathbb{E}_n\left[V(Z;\hat{h})\right] = o_p(1)$$

The remainder of the proof is identical to the proof of Lemma 17. For the above property it suffices to show that $n\,\mathbb{E}\left[\mathbb{E}_n\left[V(Z;\hat{h})\right]^2\right] \to 0$.

First we re-write the differences $V(Z;h) - V(Z;h')$:

$$\begin{aligned}V(Z;h) - V(Z;h') = \;& m(Z; g-g') + (a(X) - a'(X))\,Y - a(X)g(X) + a'(X)g'(X) \\ & - \left(\langle a_0, g-g'\rangle_2 - \langle a, g\rangle_2 + \langle a', g'\rangle_2 + \langle a-a', g_0\rangle_2\right)\end{aligned}$$

By MSE continuity of the the moment and boundedness of the functions we have that:

$$\mathbb{E}\left[(V(Z;h) - V(Z;h'))^2\right] \le c_0 \mathbb{E}\left[\|h(X) - h'(X)\|_\infty^2\right]$$

for some constant c_0. Moreover, since, for every x, y: $x^2 \le y^2 + |x|\,|x-y| + |y|\,|x-y|$:

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}_n[V(Z;\hat{h})]^2\right] =& \frac{1}{n^2}\sum_{i,j}\mathbb{E}\left[V(Z_i;\hat{h})V(Z_j;\hat{h})\right]\\
\le& \frac{1}{n^2}\sum_{i,j}\left(\mathbb{E}\left[V(Z_i;\hat{h}^{-i,j})V(Z_j;\hat{h}^{-i,j})\right] + 2\,\mathbb{E}\left[\left|V(Z_i;\hat{h}^{-i,j})\right|\,\left|V(Z_j;\hat{h}^{-i,j}) - V(Z_j;\hat{h})\right|\right]\right)\\
\le& \frac{1}{n^2}\sum_{i,j}\left(\mathbb{E}\left[V(Z_i;\hat{h}^{-i,j})V(Z_j;\hat{h}^{-i,j})\right] + 2\sqrt{\mathbb{E}\left[V(Z_i;\hat{h}^{-i,j})^2\right]}\sqrt{\mathbb{E}\left[\left(V(Z_j;\hat{h}^{-i,j}) - V(Z_j;\hat{h})\right)^2\right]}\right)\\
\le& \frac{1}{n^2}\sum_{i,j}\left(\mathbb{E}\left[V(Z_i;\hat{h}^{-i,j})V(Z_j;\hat{h}^{-i,j})\right] + 2\,c_0\sqrt{\mathbb{E}\left[V(Z_i;\hat{h}^{-i,j})^2\right]}\sqrt{\mathbb{E}\left[\|\hat{h}^{-i,j}(X_j) - \hat{h}(X_j)\|_\infty^2\right]}\right)\\
\le& \frac{1}{n^2}\sum_{i,j}\left(\mathbb{E}\left[V(Z_i;\hat{h}^{-i,j})V(Z_j;\hat{h}^{-i,j})\right] + 8\,c_0\,\beta_{n-1}\sqrt{\mathbb{E}\left[V(Z_i;\hat{h}^{-i,j})^2\right]}\right)
\end{aligned}$$

For every $i \neq j$ we have:

$$\begin{aligned}
\mathbb{E}[V(Z_i;\hat{h}^{-i,j})V(Z_j;\hat{h}^{-i,j})] =& \mathbb{E}\left[\mathbb{E}\left[V(Z_i;\hat{h}^{-i})V(Z_j;\hat{h}^{-j}) \mid \hat{h}^{-i,j}\right]\right]\\
=& \mathbb{E}\left[\mathbb{E}\left[V(Z_i;\hat{h}^{-i,j}) \mid \hat{h}^{-i,j}\right]\mathbb{E}\left[V(Z_j;\hat{h}^{-i,j}) \mid \hat{h}^{-i,j}\right]\right] = 0
\end{aligned}$$

and

$$\begin{aligned}
\sqrt{\mathbb{E}[V(Z;\hat{h}^{-i,j})^2]} \le& O\left(\sqrt{\mathbb{E}\left[\|\hat{a}^{-i,j} - a_*\|_2^2 + \|\hat{g}^{-i,j} - g_*\|_2^2\right]}\right) = O(r_{n-2})\\
\mathbb{E}[V(Z;\hat{h}^{-i})^2] \le& O\left(\mathbb{E}\left[\|\hat{a}^{-i} - a_*\|_2^2 + \|\hat{g}^{-i} - g_*\|_2^2\right]\right) = O(r_{n-1}^2)
\end{aligned}$$

Thus we get that:

$$n\,\mathbb{E}\left[\mathbb{E}_n[V(Z;\hat{h})]^2\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[V(Z_i;\hat{h}^{-i})^2] + O\left(\beta_{n-1} r_{n-2}\right) = O\left(r_{n-1}^2 + n\,\beta_{n-1} r_{n-2}\right)$$

Thus it suffices to assume that

$$r_{n-1}^2 + n\,\beta_{n-1} r_{n-2} \to 0$$

$\square$