The Fundamental Limits of Structure-Agnostic Functional Estimation

Sivaraman Balakrishnan[†], Edward Kennedy[†] and Larry Wasserman[†]

Department of Statistics and Data Science[†]

Carnegie Mellon University,
Pittsburgh, PA 15213.

{siva,edward,larry}@stat.cmu.edu

May 9, 2023

Abstract

Many recent developments in causal inference, and functional estimation problems more generally, have been motivated by the fact that classical one-step (first-order) debiasing methods, or their more recent sample-split double machine-learning avatars, can outperform plugin estimators under surprisingly weak conditions. These first-order corrections improve on plugin estimators in a black-box fashion, and consequently are often used in conjunction with powerful off-the-shelf estimation methods. On the other hand, these first-order methods are provably suboptimal in a minimax sense for functional estimation when the nuisance functions live in Hölder-type function spaces. This suboptimality of first-order debiasing has motivated the development of "higher-order" debiasing methods [3, 5, 39, 51]. The resulting estimators are, in some cases, provably optimal over Hölder-type spaces, but in sharp contrast to first-order estimators, both the estimators which are minimax-optimal and their analyses are crucially tied to properties of the underlying function space. Along a similar vein, some work [2, 17, 49] has considered $\sqrt{n}$-consistent estimation of causal effects under weaker conditions than those required by first-order methods, once again relying on higher-order debiasing. More recent work in this area has focused on attempting to weaken the dependence of these higher-order estimators on the underlying nuisance function spaces, to make the resulting estimators and theory more robust. A central focus has been to try to make higher-order methods compatible with black-box nuisance estimators.

In this paper we investigate the fundamental limits of structure-agnostic functional estimation, where relatively weak conditions are placed on the underlying nuisance functions. We show that there is a strong sense in which *existing first-order methods are optimal*. Particularly, we show that for several canonical integral functionals of interest it is impossible to improve on first-order estimators without making further, strong structural assumptions. We achieve this goal by providing a formalization of the problem of functional estimation with black-box nuisance function estimates, and deriving minimax lower bounds for this problem. Our results highlight some clear tradeoffs in functional estimation – if we wish to remain agnostic to the underlying nuisance function spaces, impose only high-level rate conditions, and maintain compatibility with black-box nuisance estimators then first-order methods are optimal. When we have a better understanding of the structure of the underlying nuisance functions then carefully constructed higher-order estimators can outperform first-order estimators.

1 Introduction

Statistical modeling often begins by hypothesizing that the data at hand are sampled from a potentially complex, high-dimensional distribution, and the goal in a variety of applications is not to estimate the distribution itself, but rather to estimate some informative *functional* of the sampling distribution. Such functional estimation problems arise naturally in causal inference where under various identification assumptions, causal estimands are expressed as *functionals* of the observed data generating distribution. One of the main challenges in causal inference is to design statistically efficient functional estimates, while remaining as agnostic as possible to the structure of the sampling distribution (the so-called nuisance component). Beyond causal inference, functional estimation problems arise routinely in machine learning [25, 44], information theory [24, 29, 35, 53], theoretical computer science [48] and other fields.

Recent research in machine learning has led to the development of powerful prediction methods, which perform surprisingly well despite the complexity of the underlying prediction tasks as well as the high-dimensionality of the covariates [31]. Consequently, a flurry of research in causal inference [10, 12], has aimed to leverage these prediction methods to estimate causal estimands. At the heart of these works is the observation that classical one-step/first-order bias-corrected estimators of many important functionals can be constructed to leverage essentially arbitrary initial estimates of the nuisance functions. These first-order estimators interact with the nuisance function estimates in a black-box manner, improving on naïve plugin estimates by shrinking their bias, but otherwise inheriting their structure-agnostic strengths. Essentially, if we are able to construct nuisance function estimates with small error (i.e. typically solve a prediction or density estimation problem well) then the one-step estimator produces an accurate functional estimate. An important aspect of this procedure is that we don't need to be able to quantify the precise structure in the nuisance functions that allows us to solve the nuisance function estimation problem well, we simply inherit fast rates of convergence when we are able to do so. We refer to estimators of this type as *structure agnostic*. Structure agnostic functional estimates are particularly powerful because modern machine learning algorithms are in practice able to solve complex prediction tasks with high-dimensional covariates with high accuracy, but we are still far from being able to accurately quantify from a theoretical perspective the precise structures which enable this. An important question, one which we aim to formalize and answer in this paper, is: *what are the fundamental limits of structure agnostic functional estimation?*

In many cases, if we can further ensure that the nuisance estimates converge at a faster than $n^{1/4}$-rate the resulting one-step estimators achieve fast $\sqrt{n}$-rates of convergence, attain semiparametric efficiency bounds, and allow for straightforward inference [4, 10, 26]. These ideas are particularly powerful when used together with sample-splitting and cross-fitting, where the nuisance functions are estimated on one half of the data, the functional is estimated on the held-out data, and the roles are reversed and the two resulting estimates are averaged to regain efficiency.

Despite their many strengths, one-step estimators are known to be far from minimax-optimal for many non-parametric functional estimation problems over smoothness classes, even when they are based on a minimax-optimal nuisance function estimate. This basic observation dates back to at least the work of Bickel and Ritov [3] who constructed minimax-optimal estimates of the integral of the square of a density by further debiasing the one-step estimator. More generally, for estimating smooth integral functionals of a density Birgé and Massart [5] proposed a general higher-order debiasing scheme, and developed complementary lower bounds. Their scheme, in combination with ideas from the papers [28, 30, 45], yields

minimax-optimal estimates for a broad class of smooth integral functionals. For more complex functionals which arise in causal inference, the construction of higher-order estimators is more involved, and is the main contribution of a more recent line of work [33, 39, 41, 51], with complementary lower bounds appearing in the work of Robins et al. [40]. In these settings, higher-order estimators improve on the one-step estimate in (very) low-regularity settings when $\sqrt{n}$-rates are not achievable, and are also able to achieve $\sqrt{n}$-rates in a wider range of (moderately) low-regularity settings. Inspired by this latter observation, some work [2, 17, 49] has considered $\sqrt{n}$-consistent estimation of causal effects under weaker conditions than those required by first-order methods, once again relying on higher-order debiasing. It is worth noting that these higher-order estimators improve on first-order estimates, and are minimax-optimal in certain settings, but are decidedly not structure agnostic in the same way that the plugin and first-order functional estimates are[1].

Our Contributions: With this background in place we can now briefly summarize our most significant contributions:

1. In Section 3.1 we describe a formal minimax setup aimed at understanding the fundamental limits of black-box functional estimation. This minimax framework allows us to frame the discussion of structure-agnostic versus structure-aware estimators, and study their relative merits.

2. In Theorem 1, we develop consequences for estimating three canonical functionals – the quadratic functional in the Gaussian sequence model, the quadratic functional in the nonparametric density model, and a mixed bias causal functional (the expected conditional covariance). Building on relatively well-understood techniques, in Theorem 2 we give matching upper bounds. Taken together these results highlight the impossibility of improving on first-order estimators without making additional structural assumptions.

3. We conclude in Section 4 with some discussion of our results, their implications, and some important avenues for future research.

1.1 Related Work

Functional estimation problems have a rich history in many different fields and we refer the reader to the works [4, 46, 47, 50] for a broader introduction to the subject. We focus in this section on briefly reviewing some lines of work which provided most of the inspiration for our work, and which study functional estimation problems in a minimax framework. In our work we present concrete results for three canonical functional estimation problems: estimating a non-linear functional in the Gaussian sequence model, estimating a non-linear integral functional of a density, and estimating a causal functional (the expected conditional covariance).

Functional estimation in the Gaussian sequence model goes back to the work of Ibragimov and Khas' minskii [21] who initiated the study of linear functionals in this model. The work of Cai and Low [7], Donoho and Nussbaum [15], Fan [18] have considered estimating non-linear functionals in the Gaussian sequence model, over Sobolev ellipsoids, Besov bodies, ℓ_p balls, and hyperrectangles. More recent work, for instance that of Collier et al. [13], has focused on estimating linear and non-linear functionals over sparsity classes.

[1] We note that we sometimes emphasize certain differences between certain classes of estimators, but the classification of estimators and the boundaries between these classes can be blurry. Part of the motivation of our work is to ground the discussion of the relative merits of different types of estimators in a rigorous minimax framework.

The estimation of smooth integral functionals of densities was considered by Bickel and Ritov [3] who studied estimating the integral of the square of a smooth density. Bickel and Ritov [3] showed that minimax rates for this functional exhibited an "elbow effect" – when the smoothness of the density $\alpha > d/4$ it is possible to attain parametric $\sqrt{n}$-rates but for less smoothness the best achievable rate is non-parametric. Their work inspired further work on estimating other smooth integral functionals of densities and regression functions over non-parametric smoothness classes [5, 28, 30, 45] culminating in a relatively comprehensive minimax theory for these functionals. The work on estimating integral functionals of a smooth density foreshadowed many developments in causal inference: particularly identifying, the sub-optimality of plugin estimates, the improved but minimax sub-optimal performance of one-step corrected estimates, and finally minimax-optimal estimates constructed via higher-order corrections. Departing from the minimax framework, there are numerous other frameworks in which one could compare estimators. These results often provide a complementary picture. For instance, the work of Cattaneo and Jansson [9] studies estimators of the quadratic functional via the inferential lens of bootstrap consistency, highlighting other tradeoffs between some of the estimators that we study.

Functionals which arise in causal inference typically exhibit more complex structure, often depending on multiple nuisance functions. The work of Robins and Rotnitzky [37], Robins et al. [38] highlighted the so-called double robustness phenomenon, where the one-step corrected estimates exhibited (second-order) bias which depended on the product of the errors of nuisance estimates. More recent work, highlights the benefits of sample-splitting and cross-fitting when using first-order estimates [11], and attempts to characterize more precisely the set of functionals for which the first-order estimate is doubly robust [12, 42]. Moving beyond first-order estimates, the work on higher-order influence functions [39, 41] and the work on complementary minimax lower bounds [40] has aimed to more completely develop the minimax theory for various important functionals in causal inference, when the nuisance functions are Hölder smooth.

1.2 Notation

We will use the notation $\lesssim, \gtrsim$ to denote inequalities which hold up to a universal positive constant, and $\asymp$ to denote an equality which holds up to a universal positive constant.

Sobolev Ellipsoids: Some of our results will consider estimation of functionals in the Gaussian sequence model. We will discuss, for instance, the case when Θ is a Sobolev ellipsoid, i.e. for some constants $M_1, M_2 > 0$ our parameter θ^* is in the set:

$$\Theta^s(M_1, M_2) = \left\{ \theta : \sum_{j=1}^{\infty} \theta_j^2 j^{2s/d} \leq M_1, \sum_{j=1}^{\infty} \theta_j^2 \leq M_2 \right\}. \tag{1}$$

Hölder Functions: For a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, and a vector $\alpha \in \mathbb{R}_+^d$ we define,

$$D^\alpha f = \frac{\partial^{\|\alpha\|_1} f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}.$$

Then letting $\ell = \lfloor s \rfloor$ we define the Hölder function class:

$$\mathcal{H}^s(L) = \Big\{ f : f \text{ is } \ell \text{ times differentiable}, \\ \max_{\alpha} |D^\alpha f(x) - D^\alpha f(y)| \leq L\|x-y\|_2^{\ell - s},\ \forall (x,y) \in \mathbb{R}^d, \|\alpha\|_1 = s, \alpha \in \mathbb{N}^d \Big\}.$$

We sometimes refer to both Hölder functions and Sobolev ellipsoids with the terminology Hölder-type spaces.

2 Background

We begin by introducing some functional estimation problems. We then briefly introduce classical one-step estimators for this problem emphasizing their structure-agnostic nature, before discussing minimax-optimal higher-order estimators.

2.1 Functional Estimation Problems

Although our results have broader implications, we focus throughout on three important functional estimation problems, for which minimax rates are relatively well-understood. In each case, we briefly summarize some well-known results. We revisit some generalizations of our results in Section 4, and highlight some potential avenues for further investigation.

Quadratic Functionals in the Gaussian Sequence Model: The main ideas of our work are most clearly understood in the following classical infinite Gaussian sequence model. We observe,

$$y_j = \theta_j^* + \epsilon_j,$$

where $j \in \{1, 2, \ldots\}$, each ϵ_j is drawn independently with distribution $N(0, 1/n)$ and our goal is to estimate the quadratic functional:

$$Q(\theta^*) = \sum_{j=1}^{\infty} \theta_j^{*2}. \tag{2}$$

This functional is a canonical example of a smooth functional and minimax rates for estimation (over Sobolev ellipsoids) go back to a series of works [3, 15, 18, 30].

Smooth Integral Functionals: In this setting, we observe $X_1, \ldots, X_n \sim f^*$, and our goal is to estimate an integral functional,

$$T(f^*) = \int (f^*(x))^2 dx. \tag{3}$$

This can be generalized the estimation of $T_\varphi(f^*) = \int \varphi(f^*(x))dx$, for some φ which has continuous second derivative. Under suitable conditions, this general setup includes the estimation of familiar information-theoretic quantities (like the entropy), and familiar quantities which arise in non-parametric estimation (like ℓ_p^p norms). The estimation of the quadratic functional was studied by Bickel and Ritov [3] and the more general problem of estimating smooth integral functionals was considered by Birgé and Massart [5].

Causal Functionals: To illustrate our ideas we focus on the expected conditional covariance. This functional arises in biostatistics and epidemiology in the context of the estimation of the causal effect of a binary treatment [40], and is an important functional in assessing conditional independence. Concretely, we observe samples of the form $\{(Y_1, A_1, X_1), \ldots, (Y_n, A_n, X_n)\}$ drawn i.i.d from a distribution $\mathbb{P}$, where $X_i \in \mathbb{R}^d, A_i \in \{0,1\}, Y_i \in \{0,1\}$. We refer to X as the covariates, Y as the outcome, and A as the treatment. We denote the covariate density p_X, and define the regression function and propensity score:

$$\mu^*(x) = \mathbb{E}[Y|X = x],$$
$$\pi^*(x) = \mathbb{E}[A|X = x].$$

Our goal is to estimate the functional:

$$\psi^{\text{cov}} = \mathbb{E}[\text{cov}(A, Y|X)] = \mathbb{E}[AY] - \int \pi^*(x)\mu^*(x)p_X(x). \tag{4}$$

The first term is easy to estimate at fast $\sqrt{n}$-rates, and the focus is often on estimating the second term. Even in our binary setup the joint distribution over triples (X, A, Y) is not fully specified by the nuisance functions (μ^*, π^*, p_X). The joint distribution in the binary setup is fully parametrized by *quadruples* $(\mu^*, \pi^*, \eta^*, p_X)$ where we additionally define,

$$\eta^*(x) = \mathbb{E}(Y|X, A = 1) - \mathbb{E}(Y|X, A = 0).$$

Somewhat surprisingly minimax rates for estimating the expected conditional covariance are not understood in full generality when the nuisance functions are Hölder smooth. In the more restricted setting when the covariate density p_X is either known or can be estimated at a sufficiently fast rate, sharp minimax rates are better understood [39].

2.2 The Methodological Approach to Functional Estimation

The functional estimation problems introduced above are examples of semi-parametric inference problems, with non-parametric nuisance components. The first attempt to solve these problems was based on the so-called plugin principle. Using the data we estimate the non-parametric components and then plug them in to obtain estimates of the functional. For instance, in the problem of estimating the expected conditional covariance, we regress the outcome Y on the covariates X to obtain an estimate $\widehat{\mu}$, and regress the treatment A on the covariates X to obtain an estimate $\widehat{\pi}$ and construct the plugin estimate:

$$\widehat{\psi}^{\text{cov}}_{\text{pi}} = \frac{1}{n}\sum_{i=1}^{n} A_i Y_i - \frac{1}{n}\sum_{i=1}^{n} \widehat{\pi}(X_i)\widehat{\mu}(X_i). \tag{5}$$

Similar plugin estimates can be constructed for functionals in the Gaussian sequence model, and for density integral functionals. It is natural to expect that if our estimates $\widehat{\mu}$ and $\widehat{\pi}$ are accurate, then the resulting plugin estimate will also be accurate. It is worth emphasizing the *structure-agnostic nature* of the plugin estimates, and particularly its compatibility with black-box nuisance function estimates. One can use a powerful machine learning algorithm (say a random forest, or a deep neural network) to construct estimates of the propensity score and the regression function, and use these to construct accurate functional estimates.

One drawback of plugin estimates is that they inherit bias, and rates of convergence, directly from their nuisance function estimates. This in turn can complicate inference, and

has led to the development of powerful one-step correction methods which improve on plugin estimates by reducing their bias, improving their rate of convergence, and allowing for valid $\sqrt{n}$-rate inference, even when the nuisance function estimates converge at a slower than $\sqrt{n}$-rate. These one-step corrections are at the heart of semi-parametric theory, and are particularly powerful when used in conjunction with sample-splitting (and/or cross-fitting). To illustrate the main idea, suppose we consider the expected conditional covariance, and consider the following first-order estimator:

$$\widehat{\psi}^{\text{cov}}_{\text{fo}} = \frac{1}{n}\sum_{i=1}^{n}(A_i - \widehat{\pi}(X_i))(Y_i - \widehat{\mu}(X_i)), \tag{6}$$

where now we suppose that $\widehat{\pi}$ and $\widehat{\mu}$ are constructed on a separate sample. It is common to construct two estimates (reversing the roles of the two samples) and average them. This estimator can be viewed as arising from correcting the plugin estimator in (5) by adding to it an estimate of the *influence function* of the target functional. To build some intuition for the estimate $\widehat{\psi}^{\text{cov}}_{\text{fo}}$ one can observe that treating $\widehat{\mu}, \widehat{\pi}$ as fixed (or estimated on a separate sample) we can compute the bias (or conditional bias) of $\widehat{\psi}^{\text{cov}}_{\text{fo}}$ and observe that it is of second-order, i.e.

$$\left|\mathbb{E}[\widehat{\psi}^{\text{cov}}_{\text{fo}}] - \psi^{\text{cov}}\right| = \left|\int (\pi^*(x) - \widehat{\pi}(x))(\mu^*(x) - \widehat{\mu}(x))p_X(x)dx\right|.$$

The estimator $\widehat{\psi}^{\text{cov}}_{\text{fo}}$ exhibits the so called double robustness property. Roughly, one can upper bound the error of the estimate $\widehat{\psi}^{\text{cov}}_{\text{fo}}$ by a product of errors of the underlying nuisance estimates. Once again, it is worth emphasizing the structure-agnostic nature of the first-order estimate. As with the plugin estimate, the first-order estimate is agnostic to the nature of the underlying nuisance function estimates. The guarantees for the first-order estimator rely on the accuracy of the pilot estimates, but neither the estimator nor its guarantee are tailored to the structure which enabled accurate pilot estimation. This enables us to use this functional estimate along with black-box machine learning algorithms, which perform well in practice, but do so by exploiting structural properties of the underlying nuisance functions that can be difficult to describe mathematically.

2.3 Smoothness Classes and The Structural Approach to Functional Estimation

Functional estimation problems are also studied from a minimax perspective in order to understand fundamental limits and to construct optimal estimators. Absent any structural assumptions consistent functional estimation is impossible and it is classical to impose some structure on the non-parametric components in the form of smoothness assumptions.

In the Gaussian sequence model this amounts to constraining the parameter space by hypothesizing that $\theta^* \in \Theta$. Minimax rates for quadratic functional estimation are well-understood for a large variety of constraint sets Θ. In our discussion, we will primarily focus on the case when θ^* is in the Sobolev ellipsoid $\Theta^s(M_1, M_2)$ in (1). In the case of integral functionals of a density, it is common to hypothesize that the nuisance function (the sampling density f) belongs to a Hölder space, i.e. that $f \in \mathcal{H}^s(L)$, and that f specifies a valid density $f \geq 0$, $\int f(x)dx = 1$. Finally, for the expected conditional covariance a typical assumption is that $\pi \in \mathcal{H}^\alpha(L_1)$, $0 \leq \pi(x) \leq 1$, and $\mu \in \mathcal{H}^\beta(L_2)$.

Given these structural assumptions, it is then natural to wonder: given a rate-optimal (say in the ℓ_2-sense) nuisance function estimate, are the resulting plugin or first-order estimators

minimax optimal? The answer to this question is often "no" and this in turn motivates higher-order estimators.

First-order estimators can be viewed as a linear bias correction of a plugin estimate. These first-order estimators have quadratic bias and higher-order estimators are constructed, roughly, by subtracting an estimate of this quadratic bias from the first-order estimate. The estimate of the bias takes the form of a higher-order U-statistic. In the Gaussian sequence model, one second-order estimator for $Q(\theta^*)$ is a classical truncated series estimator. For a truncation threshold $T > 0$ we construct the estimator:

$$\widehat{Q}^{\theta}_{\text{ho}} = \sum_{j=1}^{T} y_j^1 y_j^2 + 2 \sum_{j=T+1}^{\infty} \left[\frac{(y_j^1 + y_j^2)\widehat{\theta}_j}{2} - \widehat{\theta}_j^2 \right], \tag{7}$$

where we assume for simplicity that we obtain two observations y^1, y^2 in the Gaussian sequence model. When this is not the case, one can use the sample-splitting device described in [36], and define:

$$y_j^1 = y_j + \Phi^{-1}(U_j)/\sqrt{n}$$
$$y_j^2 = y_j - \Phi^{-1}(U_j)/\sqrt{n},$$

where U_j are independently drawn uniform random variables. The resulting y_j^1, y_j^2 are now independent, with means θ_j^* and variance $2/n$ (i.e. their variances are inflated by a factor of 2). When $\widehat{\theta}$ is 0, the estimate (7) is an unbiased estimate of $\sum_{j=1}^{T} \theta_j^{*2}$. In this case, it is well-known [15, 18, 30], that when the truncation parameter T is chosen to scale as $n^{2d/(4s+d)}$, $\widehat{Q}^{\theta}_{\text{ho}}$ is a minimax optimal estimate of $Q(\theta^*)$ over Θ^s and is semi-parametrically efficient when $s > d/4$.

It is important to note that despite the fact the estimator was constructed as a second-order correction to the plugin estimate, its minimaxity over the Sobolev ellipsoid is no longer strongly dependent on the choice of the pilot estimate $\widehat{\theta}$, which could simply be taken to be 0. Rather, the optimality of this estimator is closely related to properties, such as decay rate of the coefficients θ_j^*, of the underlying Sobolev ellipsoid. In a similar vein, higher-order estimators have been constructed for integral functionals of a density in [5, 28, 45], and for causal functionals like the expected conditional covariance in [39]. These estimators exhibit similar properties to the estimator (7), the analysis which demonstrates their minimax-optimality (or superiority over plugin or first-order estimates) often relies on carefully exploiting properties of the underlying nuisance function space.

Comparing Higher-order Estimators and One-Step Corrections: To summarize the discussion so far, it is worth once again contrasting first-order and higher-order estimators to identify some common themes which hold across the canonical examples and more broadly. First-order estimators are black-box corrections to plugin estimates. Under very weak conditions they improve on plugin estimators, and their accuracy depends only on the (squared) errors of the nuisance function estimates. This in turn enhances their compatibility with black-box (machine learning) methods for nuisance function estimation. They are often not minimax-optimal over Hölder-type function spaces, even when used in conjunction with minimax-optimal nuisance function estimates. In contrast, minimax-optimal higher-order estimators are more carefully tailored to the underlying function space. They typically have a weaker connection to the plugin estimates on which they are based. They are analyzed via a more careful understanding of the bias-variance tradeoff in the underlying nuisance function

space. They can in many cases yield minimax-optimal functional estimates over Hölder-type spaces, even when used with trivial (zero) pilot nuisance function estimates.

3 Main Results

We begin with a description of the black-box minimax setup we will focus on. We then turn our attention to minimax lower bounds in Section 3.2, and briefly provide complementary upper bounds in Section 3.3.

3.1 Minimax Functional Estimation in the Black-Box Model

To fix ideas we first consider estimation of the quadratic functional in the Gaussian sequence model (2). As we noted previously, with no structural assumptions placed on θ^* consistent functional estimation is impossible. Rather than impose smoothness assumptions, we model the black-box setting where we construct a pilot estimate on a separate sample.

More formally, our goal is to estimate $Q(\theta^*)$ and our assumption on θ^* is that the pilot estimate $\widehat{\theta}$ is accurate in an ℓ_2 sense, i.e. that $\theta^* \in \Theta(r_n)$, where:

$$\Theta(r_n) := \left\{\theta : \|\theta - \widehat{\theta}\|_2^2 \leq r_n\right\}, \tag{8}$$

where the accuracy of the pilot estimate r_n is unknown to the statistician.

It is important to note that the assumption that the pilot estimate $\widehat{\theta}$ is r_n-accurate, imposes an (implicit) structural assumption on θ^*. The strength of this structural assumption depends on the (unknown) rate of convergence r_n. *It is precisely this structural condition that plugin and first-order estimates are tailored to leverage.* It is also worth contrasting this structural assumption with the smoothness assumptions in (1) – here the structural assumption hypothesizes that our favorite nuisance function estimator returns an accurate pilot estimate, but does not further constrain θ^* to have a particular structure. Since r_n is unknown to the statistician, estimators constructed in this model are implicitly *adaptive*, i.e. this setting bears similarities to the classical adaptive non-parametric estimation setting where functions are hypothesized to be smooth but the smoothness parameter is unknown to the statistician.

Absent any additional smoothness assumptions, our goal is to construct a minimax rate-optimal estimate, i.e. an estimate $\widehat{Q}$ such that:

$$\sup_{\theta^* \in \Theta(r_n)} \mathbb{E}(\widehat{Q} - Q(\theta^*))^2 \asymp \inf_{\widetilde{Q}} \sup_{\theta^* \in \Theta(r_n)} \mathbb{E}(\widetilde{Q} - Q(\theta^*))^2 := \mathfrak{M}_n^{\theta}(\Theta(r_n)), \tag{9}$$

and to study the minimax risk $\mathfrak{M}_n^{\theta}(\Theta(r_n))$.

In a similar vein, one can consider minimax estimation of the density functional $T(f^*)$ in (3). We assume for simplicity that the densities under consideration are uniformly upper bounded by some (large) constant $M > 0$. Our goal is to estimate $T(f^*)$ under the constraint that $f^* \in \mathcal{F}(r_n)$:

$$\mathcal{F}(r_n) := \left\{f : \int (f(x) - \widehat{f}(x))^2 dx \leq r_n, f \geq 0, \int f(x)dx = 1, \|\widehat{f}\|_\infty, \|f\|_\infty \leq M\right\}.$$

In this case we define the associated minimax risk as:

$$\mathfrak{M}_n^{\mathrm{f}}(\mathcal{F}(r_n)) := \inf_{\widehat{T}} \sup_{f^* \in \mathcal{F}(r_n)} \mathbb{E}(\widehat{T} - T(f^*))^2. \tag{10}$$

For the expected conditional covariance we are given two pilot estimates $\widehat{\mu}$ and $\widehat{\pi}$. In order to construct higher-order estimators we might further assume that we are given a third pilot estimate $\widehat{p}_X$ of the covariate density. To simplify our presentation of minimax lower bounds, we consider the case when the covariate density p_X is uniform on $[0,1]^d$. We also assume that $\widehat{\mu}$ and $\widehat{\pi}$ are bounded away from 0 and 1 on $[0,1]^d$. This latter restriction can be eliminated via a perturbation argument similar to the one used in Appendix B.2 for the integral of the squared density. Although these restrictions ease the construction of minimax lower bounds, the upper bounds in Theorem 2 hold without these restrictions.

With this setup in place our goal is to estimate ψ^{cov} in (4), under the following constraints on $(\mu^*, \pi^*, \eta^*, p_X^*) \in \mathcal{G}(r_n, s_n)$:

$$\mathcal{G}(r_n, s_n) := \Big\{(\mu, \pi, \eta, p_X) : \text{supp}(X) = [0,1]^d, p_X = \text{unif}[0,1]^d, \int (\mu(x) - \widehat{\mu}(x))^2 p_X(x) dx \leq r_n,$$
$$\int (\pi(x) - \widehat{\pi}(x))^2 p_X(x) dx \leq s_n, 0 \leq \pi(x), \mu(x) \leq 1, 1 - \varepsilon \geq \widehat{\pi}(x), \widehat{\mu}(x) \geq \varepsilon > 0, \text{for } x \in [0,1]^d \Big\}.$$

We define the associated minimax risk as:

$$\mathfrak{M}_n^{\text{cov}}(\mathcal{G}(r_n, s_n)) := \inf_{\widehat{\psi}} \sup_{(\mu^*, \pi^*) \in \mathcal{G}(r_n, s_n)} \mathbb{E}(\widehat{\psi} - \psi^{\text{cov}})^2. \tag{11}$$

With this setup in place, our goal is to understand the fundamental limits on structure-agnostic functional estimation by providing upper and lower bounds on the minimax risks in (9), (10) and (11).

3.1.1 Interpreting the Minimax Setup

The minimax problems described in this section require some care to interpret. We describe briefly some interpretations focusing again on the Gaussian sequence model:

Sample-Splitting and the Conditional Viewpoint: When analyzing black-box sample-splitting based functional estimators, it is natural to take a conditional viewpoint to remain judicious in the modeling assumptions we impose. In this viewpoint, we hypothesize that for some function $r_{n,\delta}$ our nuisance estimates are r_n-accurate with probability at least $1 - \delta$, and then proceed to analyze a functional estimate constructed on a separate sample. This perspective is for instance explicitly adopted in the work [19], and is implicit in a long series of past work [3, 4, 11, 12].

To complement this conditional viewpoint with minimax lower bounds, one can aim to understand the fundamental limitations of the second stage of this two-stage estimator construction (the first-stage corresponds to the well-studied problem of function estimation). This problem is at the heart of our proposed minimax setup.

In contrast to the traditional setting, where θ^* is fixed, and $\widehat{\theta}$ is random, in our lower bounds we treat both as fixed. To model more closely the sample-splitting based functional estimation paradigm, we might split the data into two sets $\mathcal{D}_0$ and $\mathcal{D}_1$, and construct a pilot estimate $\widehat{\theta}$ on $\mathcal{D}_0$. We would then relax the constraint set in (8) to be the random set:

$$\Theta(r_n, \delta) = \begin{cases} \Big\{\theta : \|\theta - \widehat{\theta}\|_2^2 \leq r_n\Big\} & \text{with probability } 1 - \delta \text{ independent of } \mathcal{D}_1, \\ \Big\{\theta : \sum_{j=1}^{\infty} \theta_j^2 \leq M_2\Big\} & \text{otherwise,} \end{cases}$$

where with probability δ the nuisance parameter θ^* is essentially unconstrained. Lower bounds over this random constraint set, can directly be obtained from lower bounds over the set (8) at

the cost of some additional notational burden. This is because when the nuisance parameter is (essentially) unconstrained, functional estimation is impossible.

Unstructured Local Minimax Lower Bounds: An alternative way to interpret our problem setting, is as a type of local minimax setup. The pilot estimate $\widehat{\theta}$, and the local radius r_n, define a local estimation problem via the constraints in (8). The minimax rates we study are thus quantifying the difficulty of functional estimation, locally around the pilot estimate, with an (otherwise) unconstrained nuisance parameter.

Despite the similarity in spirit, the setup and goals are quite different from classical local minimax problems. We don't make strong assumptions on the sequence r_n, the only constraints in our problem are locality constraints as opposed to locality and smoothness constraints, and we adopt a non-asymptotic view. The most significant difference is that we localize the parameter space around the pilot estimate $\widehat{\theta}$ and not the true parameter θ^* since our goal is not to elicit local minimax rates in the neighborhood of the true parameter θ^*, but rather to understand the fundamental limits of black-box functional estimation.

3.2 Lower Bounds

With the minimax setup introduced in the previous section we are now equipped to state our lower bounds on the minimax risks in (9), (10) and (11).

Theorem 1. *The minimax risks in* (9), (10) *and* (11) *are lower bounded as:*

1. **Quadratic Functional in the Gaussian Sequence Model:**

$$\mathfrak{M}_n^{\theta}(\Theta(r_n)) \gtrsim r_n^2 + \|\widehat{\theta}\|_2^2 \min\left\{r_n, \frac{1}{n}\right\}.$$

2. **Quadratic Density Integral Functional:**

$$\mathfrak{M}_n^{f}(\mathcal{F}(r_n)) \gtrsim r_n^2 + \left[\int \widehat{f}(x)^3 dx - \left(\int \widehat{f}(x)^2\right)^2 dx\right] \min\left\{r_n, \frac{1}{n}\right\}.$$

3. **Expected Conditional Covariance:**

$$\mathfrak{M}_n^{cov}(\mathcal{G}(r_n, s_n)) \gtrsim r_n \times s_n + \frac{1}{n}.$$

We prove this result in Appendix B. Our proofs are based on a well-understood recipe. We reduce the problem of lower bounding the minimax risk of functional estimation to lower bounding the risk (the sum of Type I and II errors) in an appropriate hypothesis testing problem (roughly of distinguishing if the functional is large or small). If the null and alternate are difficult to distinguish, we obtain a lower bound on the minimax risk. To lower bound the minimax hypothesis testing error we carefully construct priors on the composite null and composite alternate and show that the resulting mixture distributions are difficult to distinguish by lower bounding the error of the (optimal) likelihood ratio test. At a more technical level, we use classical ideas from Ingster and Suslina [22] for the Gaussian sequence model, from Balakrishnan and Wasserman [1] for the density integral functional, and from Robins et al. [40] for the expected conditional covariance, in order to upper bound an appropriate divergence measure between the two mixture distributions.

It is interesting to first focus on the case when $r_n, s_n \gg 1/n$, since this is the typical case. In this case, the second term in each of the lower bounds corresponds (in rate) to semi-parametric efficiency lower bounds. These lower bounds are determined by the *variance* of the estimated influence function, and decay to 0 at the standard parametric rate. A large fraction of semi-parametric theory focuses on imposing conditions on the nuisance functions under which the *bias* term (of order r_n^2 or $r_n \times s_n$) above decays to zero faster than the parametric rate, at which point semi-parametric efficient estimation and inference are possible. On the other hand, the more recent literature on higher-order estimators [2, 3, 5, 17, 39, 49, 51], has aimed at reducing the bias term to either obtain $\sqrt{n}$-rates under weaker assumptions on the nuisance functions, or in order to obtain (slower than $\sqrt{n}$) minimax-optimal rates when the nuisance functions are Hölder smooth. The main import of Theorem 1 is that under only the assumption that the pilot estimates are accurate in an ℓ_2 sense, and in the absence of further smoothness assumptions, *no further bias reduction is possible.* As we explore further in Theorem 2, first-order estimates are minimax-optimal, and achieve the limits of structure-agnostic functional estimation.

When the condition that $r_n, s_n \gg 1/n$ is violated the first-stage pilot estimates are super-accurate, i.e. are more accurate than the (fixed-dimensional) parametric rate, and the plugin functional estimate is already minimax optimal. We address this situation in more detail in Appendix D, where we construct a functional estimate which adapts between the plugin and first-order estimates paying only a small statistical price, and prove a matching lower bound which highlights the fundamental limits of adaptivity in this setup.

3.3 Upper Bounds

In this section, we develop upper bounds on the minimax risk by analyzing plugin and first-order estimators. Our analyses of these estimators are elementary, and plugin and first-order estimators have been analyzed much more generally in past work, although often from an asymptotic perspective [50]. Our simple non-asymptotic analysis enables a more direct comparison with lower bounds from Theorem 1.

To set the stage we first formally define the plugin and first-order estimates. The plugin and first-order estimates for the quadratic functionals in the Gaussian sequence model and for the density integral functional are:

$$\widehat{Q}^{\theta}_{\mathrm{pi}} = \|\widehat{\theta}\|_2^2 \qquad \widehat{T}^{\mathrm{f}}_{\mathrm{pi}} = \int (\widehat{f}(x))^2 dx$$

$$\widehat{Q}^{\theta}_{\mathrm{fo}} = 2\langle y, \widehat{\theta}\rangle - \|\widehat{\theta}\|_2^2 \qquad \widehat{T}^{\mathrm{f}}_{\mathrm{fo}} = \frac{2}{n}\sum_{i=1}^{n} \widehat{f}(X_i) - \int (\widehat{f}(x))^2 dx.$$

We also recall the definitions of the plug-in and first-order estimates of the expected conditional covariance in (5) and (6), and the higher-order estimate in the Gaussian sequence model in (7). Having introduced the plugin and first-order estimates of our three canonical functionals we have the following theorem:

Theorem 2. *The estimators described above have the following guarantees:*

1. **Quadratic Functional in the Gaussian Sequence Model:**

$$|\widehat{Q}^{\theta}_{pi} - Q(\theta^*)|^2 \lesssim r_n^2 + r_n \|\widehat{\theta}\|_2^2$$
$$\mathbb{E}(\widehat{Q}^{\theta}_{fo} - Q(\theta^*))^2 \lesssim r_n^2 + \frac{\|\widehat{\theta}\|_2^2}{n}$$
$$\mathbb{E}(\widehat{Q}^{\theta}_{ho} - Q(\theta^*))^2 \lesssim \left[\sum_{j=T+1}^{\infty} (\widehat{\theta}_j - \theta_j^*)^2\right]^2 + \frac{\|\widehat{\theta}\|_2^2}{n} + \frac{T}{n^2} \lesssim r_n^2 + \frac{\|\widehat{\theta}\|_2^2}{n} + \frac{T}{n^2}.$$

2. **Quadratic Density Integral Functional:**

$$|\widehat{T}^{f}_{pi} - T^f(f^*)|^2 \lesssim r_n^2 + r_n \int \widehat{f}(x)^2 dx$$
$$\mathbb{E}(\widehat{T}^{f}_{fo} - T^f(f^*))^2 \lesssim r_n^2 + \frac{var(\widehat{f}(X))}{n}.$$

3. **Expected Conditional Covariance:**

$$\mathbb{E}(\widehat{\psi}^{cov}_{pi} - \psi^{cov})^2 \lesssim r_n \times s_n + r_n + s_n + \frac{1}{n}$$
$$\mathbb{E}(\widehat{\psi}^{cov}_{fo} - \psi^{cov})^2 \lesssim r_n \times s_n + \frac{1}{n}.$$

Once again we initially focus our discussion on the case when $r_n, s_n \gg \frac{1}{n}$, which is the typical setting, in which case the first-order estimates outperform the plugin estimates. We design an adaptive estimate for the quadratic functional in the sequence model, which selects between the plugin and first-order estimate in a data-driven manner, and achieves close to the oracle risk, in Appendix D. For each of our functionals, focusing on terms which only depend on n and r_n, s_n, the first-order estimate achieves a maximum risk which matches the minimax lower bounds of Theorem 1, i.e. the first-order estimates are minimax optimal in our setting.

The higher-order estimator in the Gaussian sequence model depends on a truncation parameter T. The higher-order estimator can have smaller bias than the first-order estimator since it unbiasedly estimates $\sum_{j=1}^{T} \theta_j^{*2}$, and combines this with a first-order estimator of $\sum_{j=T+1}^{\infty} \theta_j^{*2}$. When T is set small relative to n, then this estimator incurs only a modest amount of additional variance. To achieve minimax-optimality over a Sobolev ellipsoid in the low-regularity regime when $s < d/4$, the truncation parameter T needs to be chosen larger than n, in a careful way, to balance the reduction in bias with the inflation in variance relative to the first-order estimator. In our structure-agnostic model, it is impossible to guarantee that the higher-order estimator has meaningfully lower bias than the first-order estimator, and consequently it is impossible to guarantee that the higher-order estimator improves on the first-order estimator. Higher-order estimators for the quadratic density functional and for the expected conditional covariance are more involved to describe [5, 39] but share the same qualitative features. This in turn highlights that in our formalization of the black-box, structure-agnostic functional estimation problem, where we are unwilling to assume more than access to a potentially accurate black-box prediction algorithm, it is impossible to improve on first-order estimators in a minimax sense. Higher-order estimates can only improve on first-order estimates when additional structural assumptions are imposed and exploited.

We note in passing that there are some slight differences between the constant factors in the upper and lower bounds for the quadratic density functional. This is due to the mismatch

between the ℓ_2^2 distance and the squared Hellinger distance. Intuitively these terms of the lower bound are determined by the modulus of continuity of the quadratic functional over a squared Hellinger neighborhood around $\widehat{f}$ [14]. On the other hand, the upper bounds are determined by the modulus of continuity over an ℓ_2^2 neighborhood. These coincide when we assume that the densities under consideration are both upper and lower bounded by universal constants, but can otherwise differ.

Finally, for the expected conditional covariance, we observe that under our conditions the plugin estimator is strictly dominated by the first-order estimator. The first-order estimator is minimax-optimal for any choice of r_n, s_n. This is because even when our pilot estimates are super accurate, i.e. for instance the true π^* and μ^* are known, we still need to estimate the term $\mathbb{E}[AY]$ to construct our functional estimate. This in turn leads to an unavoidable $\mathcal{O}(1/n)$ term in the MSE. In contrast, for the quadratic density and sequence functionals the plugin estimator is a deterministic function of the nuisance estimates and incurs no additional variance, and can dominate the first-order estimator when the pilot estimates are super-accurate.

4 Discussion and Extensions

In this work, we introduced a minimax framework for reasoning about two-stage structure-agnostic functional estimation methods. We developed consequences for estimating three canonical functionals – the quadratic functional in the Gaussian sequence model, the quadratic functional in the non-parametric density model, and a mixed bias causal functional (the expected conditional covariance).

By focusing on concrete examples, we have given results for particular estimators which are canonical plug-in and first-order estimators, but have avoided giving precise general definitions for these classes of estimators. This is by design, as we noted in Footnote 1 the distinctions between these classes of estimators can be blurry. For instance, the work of Newey and Robins [34] and Giné and Nickl [20] show that certain carefully undersmoothed plugin-type estimators can perform similarly to higher-order estimators, and inherit both their strengths and weaknesses.

There are several possible extensions of our results. Minimax theory is well-understood for more general smooth functionals in both the density model and in the Gaussian sequence model, and this theory mirrors closely results for the quadratic functionals in these models. We expect our main results will continue to hold for more general smooth functionals. For causal functionals, beyond the expected conditional covariance, minimax lower bounds are known only in a few problems [40] and we expect our results will extend to cover these functionals as well. A more ambitious extension would aim to cover larger classes of functionals for which first-order estimators are well-understood to have desirable properties [12, 42]. However, the likelihood structure in each of these statistical models is quite different, which poses some challenges to developing a unified theory of lower bounds. Minimax rates, over appropriate smoothness classes, have also been studied for certain local integral functionals which arise in causal inference and in non-parametric regression [8, 27, 43, 52], and it would be also be interesting to understand the fundamental limits of structure-agnostic estimation in these problems.

Finally, it is worth emphasizing that our lower bounds do not preclude estimators which improve in some restricted ways on first-order estimators. For instance, it is possible in some cases to construct adaptive estimators which perform nearly as well as the first-order estimator in the absence of any additional structure, but improve on the first-order estimator when the

nuisance functions have additional smoothness structure. Such estimates are developed, for instance, in a testing context in the work of Liu et al. [32]. On the other hand, our results do show that if one aims to improve on first-order estimators in a general minimax sense, this improvement is only possible at the expense of adding further assumptions, i.e. there are limits to what can be achieved by higher-order estimators without additional structural assumptions. Developing a comprehensive understanding of adaptive estimators, which adapt between smoothness classes and structure-agnosticity, and understanding their fundamental limitations, could be an interesting avenue for future research.

Acknowledgements

The authors are grateful to Jamie Robins for several helpful discussions, and for many inspiring conversations.

References

[1] S. Balakrishnan and L. Wasserman. Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *The Annals of Statistics*, 47(4):1893–1927, 2019.

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

[3] P. J. Bickel and Y. Ritov. Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā*, pages 381–393, 1988.

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

[5] L. Birgé and P. Massart. Estimation of integral functionals of a density. *The Annals of Statistics*, 23(1):11–29, 1995.

[6] L. D. Brown and M. G. Low. A constrained risk inequality with applications to nonparametric functional estimation. *The Annals of Statistics*, 24(6):2524 – 2535, 1996.

[7] T. T. Cai and M. G. Low. Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.

[8] T. T. Cai, M. Levine, and L. Wang. Variance function estimation in multivariate nonparametric regression with fixed design. *Journal of Multivariate Analysis*, 100(1):126–136, 2009.

[9] M. D. Cattaneo and M. Jansson. Average density estimators: Efficiency and bootstrap consistency. *Econometric Theory*, 38(6):1140–1174, 2022.

[10] V. Chernozhukov, M. Goldman, V. Semenova, and M. Taddy. Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. *arXiv preprint arXiv:1712.09988*, 2017.

[11] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. M. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[12] V. Chernozhukov, W. K. Newey, and R. Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022.

[13] O. Collier, L. Comminges, and A. B. Tsybakov. Minimax estimation of linear and quadratic functionals on sparsity classes. *The Annals of Statistics*, 45(3):923–958, 2017.

[14] D. L. Donoho and R. C. Liu. Geometrizing Rates of Convergence, III. *The Annals of Statistics*, 19(2):668 – 701, 1991.

[15] D. L. Donoho and M. Nussbaum. Minimax quadratic estimation of a quadratic functional. *Journal of Complexity*, 6(3):290–323, 1990.

[16] J. C. Duchi and F. Ruan. A constrained risk inequality for general losses, 2018.

[17] O. Dukes, S. Vansteelandt, and D. Whitney. On doubly robust inference for double machine learning, 2021.

[18] J. Fan. On the Estimation of Quadratic Functionals. *The Annals of Statistics*, 19(3):1273 – 1294, 1991.

[19] D. J. Foster and V. Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

[20] E. Giné and R. Nickl. A simple adaptive estimator of the integrated square of a density. *Bernoulli*, 14(1):47 – 61, 2008.

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

[22] Y. Ingster and I. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media, 2003.

[23] Y. I. Ingster. Adaptive chi-square tests. *Zapiski Nauchnykh Seminarov POMI*, 244:150–166, 1997.

[24] J. Jiao, K. Venkat, Y. Han, and T. Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

[25] K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman, and J. M. Robins. Influence functions for machine learning: Nonparametric estimators for entropies, divergences and mutual informations, 2014.

[26] E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review, 2022.

[27] E. H. Kennedy, S. Balakrishnan, J. M. Robins, and L. Wasserman. Minimax rates for heterogeneous causal effect estimation, 2022.

[28] G. Kerkyacharian and D. Picard. Estimating nonquadratic functionals of a density using haar wavelets. *The Annals of Statistics*, 24(2):485–507, 1996.

[29] L. F. Kozachenko and N. N. Leonenko. A statistical estimate for the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

[30] B. Laurent. Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 24(2):659–681, 1996.

[31] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[32] L. Liu, R. Mukherjee, and J. M. Robins. On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning. *Statistical Science*, to appear.

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

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

[35] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6): 1191–1253, 2003.

[36] J. Robins and A. van der Vaart. Adaptive nonparametric confidence sets. *The Annals of Statistics*, 34(1):229 – 253, 2006.

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

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

[39] J. M. Robins, L. Li, E. J. Tchetgen Tchetgen, and A. W. van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 335–421, 2008.

[40] J. M. Robins, E. J. Tchetgen Tchetgen, L. Li, and A. W. van der Vaart. Semiparametric minimax rates. *Electronic Journal of Statistics*, 3:1305–1321, 2009.

[41] J. M. Robins, L. Li, R. Mukherjee, E. J. Tchetgen Tchetgen, and A. W. van der Vaart. Minimax estimation of a functional on a structured high dimensional model. *The Annals of Statistics*, 45(5):1951–1987, 2017.

[42] A. Rotnitzky, E. Smucler, and J. M. Robins. Characterization of parameters with a mixed bias property. *arXiv preprint arXiv:1904.03725*, 2019.

[43] Y. Shen, C. Gao, D. Witten, and F. Han. Optimal estimation of variance in nonparametric regression with random design. *The Annals of Statistics*, 48(6):3589–3618, 2020.

[44] S. Singh. Estimating Probability Distributions and their Properties. *PhD Thesis*, 2020.

[45] E. Tchetgen, L. Li, J. Robins, and A. van der Vaart. Minimax estimation of the integral of a power of a density. *Statistics & probability letters*, 78(18):3307–3311, 2008.

[46] A. A. Tsiatis. *Semiparametric Theory and Missing Data*. New York: Springer, 2006.

[47] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. New York: Springer, 2009.

[48] G. Valiant and P. Valiant. Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new clts. *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, pages 685–694, 2011.

[49] M. J. van der Laan. Targeted estimation of nuisance parameters to obtain valid statistical inference. *The International Journal of Biostatistics*, 10(1):29–57, 2014.

[50] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 2000.

[51] A. W. van der Vaart. Higher order tangent spaces and influence functions. *Statistical Science*, 29(4):679–686, 2014.

[52] L. Wang, L. D. Brown, T. T. Cai, and M. Levine. Effect of mean on variance function estimation in nonparametric regression. *The Annals of Statistics*, 36(2):646–664, 2008.

[53] Y. Wu and P. Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

A Lower Bound Preliminaries

In this section, we collect some well-known technical facts that will aid the proof of Theorem 1. Suppose that our goal is to estimate a functional $T(P)$, given samples from $P \in \mathcal{P}$. We first recall a standard construction (see for instance Theorem 2.14 in [47]) for obtaining lower bounds in functional estimation problems.

We construct two prior distributions, π_0 and π_1 on $\mathcal{P}$, which induce two distributions Q_0 and Q_1 where for any measurable set A:

$$Q_0(A) = \int P^n(A) d\pi_0(P), \quad \text{and} \quad Q_1(A) = \int P^n(A) d\pi_1(P).$$

We further ensure that our functional takes sufficiently different values under each of the prior distributions, i.e.:

$$\pi_0(\{P : T(P) \leq c\}) = 1, \qquad \pi_1(\{P : T(P) \geq c + 2s\}) = 1.$$

Then we have the following result:

Lemma 1. *Suppose that* $\chi^2(Q_0, Q_1) \leq \alpha < \infty$, *then*

$$\inf_{\widehat{T}} \sup_{P \in \mathcal{P}} \mathbb{E}(\widehat{T} - T(P))^2 \geq s^2 \max\Big(\frac{1}{4}\exp(-\alpha), \frac{1 - \sqrt{\alpha/2}}{2}\Big).$$

If $H^2(Q_0, Q_1) \leq \alpha < 2$, *then:*

$$\inf_{\widehat{T}} \sup_{P \in \mathcal{P}} \mathbb{E}(\widehat{T} - T(P))^2 \geq s^2 \frac{1 - \sqrt{\alpha(1 - \alpha/4)}}{2}.$$

The proof of this lemma follows immediately from Theorem 2.15 in [47], and an application of Markov's inequality to obtain in-expectation bounds. The main takeaway is simply that if we can construct Q_0, Q_1 as above, ensuring that the functional is separated by at least $2s$, and ensuring that Q_0 and Q_1 have χ^2 divergence or Hellinger distance upper bounded by a sufficiently small constant, then we obtain lower bounds on the minimax risk of order s^2.

We frequently use the following well-known fact.

Fact 1. *Given two measures p, q, the squared Hellinger distance between their n-fold products can be upper bounded as:*

$$H^2(p^n, q^n) \leq nH^2(p, q).$$

Suppose we consider the Gaussian sequence model and let $Q_0 = N(\theta, 1/n)$. We then define Q_1 in the following way. We select any d coordinates, let $\mathcal{I}$ denote the selected indices. Fix an $\epsilon > 0$. Let $N := 2^d$ and let $\{u_1, \ldots, u_N\}$ denote the collection of all vectors with entries $\{+\epsilon, -\epsilon\}$. For any $v \in \mathbb{R}^d$ we denote by θ_v the vector which perturbs θ by adding v to the indices in $\mathcal{I}$. Define Q_1^ϵ to be the following mixture:

$$Q_1^\epsilon = \frac{1}{N} \sum_{i=1}^{N} N(\theta_{u_i}, 1/n).$$

The mixture distribution Q_1 is obtained by perturbing the coordinates in $\mathcal{I}$ by $\pm\epsilon$ uniformly at random. The following result is well-known:

Lemma 2. *For any θ, and $\mathcal{I}$, suppose $\epsilon \leq 1/n$, then:*

$$\chi^2(Q_0, Q_1^\epsilon) \leq \exp(dn^2\epsilon^4) - 1.$$

Furthermore, if $dn^2\epsilon^4 \leq 1$, then,

$$\chi^2(Q_0, Q_1^\epsilon) \leq 2dn^2\epsilon^4.$$

We include a short proof for completeness in Appendix A.1. We also note the following simple bound on the χ^2 distance between two Gaussians.

Lemma 3. *Suppose $P = N(\theta, I/n)$, $Q = N(\widetilde{\theta}, I/n)$, then*

$$\chi^2(P, Q) = \exp(n\|\theta - \widetilde{\theta}\|_2^2) - 1.$$

Furthermore, if $n\|\theta - \widetilde{\theta}\|_2^2 \leq 1$, then,

$$\chi^2(P, Q) \leq 2n\|\theta - \widetilde{\theta}\|_2^2.$$

We will use the following constrained risk inequality as a consequence of Theorem 1 of Brown and Low [6]. Though we only need a result for the squared loss, the form we use is most directly deduced from Corollary 2 of [16]. This inequality is a useful tool for proving lower bounds for adaptive estimators.

Lemma 4. *Fix any θ, and define $D_1 := N(\theta, I/n)$ and $D_2 := N((1 + \alpha/\|\theta\|_2)\theta, I/n)$. Let P_1 denote either D_1 or D_2, and let P_2 denote the other distribution. Suppose we have an estimator $\widehat{Q}$ for which,*

$$\mathbb{E}_{P_1}(\widehat{Q} - Q(P_1))^2 \leq \beta^2,$$

then

$$\mathbb{E}_{P_2}(\widehat{Q} - Q(P_2))^2 \geq \left[\alpha^2 + 2\alpha\|\theta\|_2 - \beta\exp(n\alpha^2/2)\right]_+^2.$$

A.1 Proof of Lemma 2

The proof follows from a direct computation. We note that,

$$\begin{aligned}
\chi^2(Q_0, Q_1^\epsilon) &= \mathbb{E}_{Q_0}\left(\frac{Q_1^\epsilon}{Q_0}\right)^2 - 1 \\
&= \mathbb{E}_{Q_0}\prod_{i=1}^d \left[\frac{\frac{1}{2}\left[\exp\left(-n\frac{(y_i-\theta_i-\epsilon)^2}{2}\right) + \exp\left(-n\frac{(y_i-\theta_i+\epsilon)^2}{2}\right)\right]}{\exp(-n(y_i-\theta_i)^2/2)}\right]^2 - 1 \\
&\overset{\text{(i)}}{=} \mathbb{E}_{Q_0}\prod_{i=1}^d \frac{1}{4}\exp(-n\epsilon^2)\left[\exp(2\sqrt{n}\epsilon Z_i) + \exp(-2\sqrt{n}\epsilon Z_i) + 1\right] - 1 \\
&\overset{\text{(ii)}}{=} \prod_{i=1}^d \frac{1}{2}\left[\exp(n\epsilon^2) + \exp(-n\epsilon^2)\right] - 1 \\
&\overset{\text{(iii)}}{\leq} \prod_{i=1}^d \exp(n^2\epsilon^4) - 1 \\
&= \exp(dn^2\epsilon^4) - 1.
\end{aligned}$$

In (i), Z_i denotes a standard Gaussian random variable, (ii) uses the fact that $\mathbb{E}(\exp(tZ_i)) = \exp(t^2/2)$, and (iii) uses the fact $\cosh(x) \leq 1 + x^2 \leq \exp(x^2)$ for $x \in [0, 1]$.

B Proof of Theorem 1

We prove each of the three lower bounds in turn.

B.1 Lower Bounds for Quadratic Functional in the Gaussian Sequence Model

Our goal is to prove minimax lower bounds for estimating the quadratic functional in the GSM. In particular, we'd like to understand lower bounds on:

$$\inf_{\widehat{Q}} \sup_{\theta:\|\theta-\widehat{\theta}\|_2^2 \leq r_n} \mathbb{E}(\widehat{Q} - Q(\theta))^2.$$

Our lower bounds will be a consequence of Lemma 1 with various choices of the priors.

LB 1: $r_n \geq \frac{1}{n}, r_n^2 \lesssim \frac{\|\widehat{\theta}\|_2^2}{n}$. In this case, we construct $Q_0 = N(\widehat{\theta}, I/n)$ and $Q_1 = N((1-\nu)\widehat{\theta}, I/n)$ where $0 \leq \nu \leq 1$. Then by Lemma 3, we have that $\chi^2(Q_1, Q_0) \leq \alpha$ if $2n\nu^2\|\widehat{\theta}\|_2^2 \leq \alpha$, so we select

$$\nu^2 = \frac{\alpha}{2n\|\widehat{\theta}\|_2^2},$$

to ensure this. Since $r_n^2 \lesssim \|\widehat{\theta}\|_2^2/n$, and $r_n \geq 1/n$, we observe that $\nu < 1$. The distance, $\|\widehat{\theta} - (1-\nu)\widehat{\theta}\|_2^2 = \nu^2\|\widehat{\theta}\|_2^2 = \alpha/2n \leq 1/n$ so the two priors are supported on the set $\|\widehat{\theta} - \theta\|_2^2 \leq r_n$ as desired.

The functional separation under the two priors is:

$$\begin{aligned} 2s &:= \|\widehat{\theta}\|_2^2 - (1-\nu)^2\|\widehat{\theta}\|_2^2 \\ &= (2\nu - \nu^2)\|\widehat{\theta}\|_2^2 \\ &\geq \nu\|\widehat{\theta}\|_2^2. \end{aligned}$$

So the minimax rate is at least $\nu^2\|\widehat{\theta}\|_2^4$, i.e. $\frac{\|\widehat{\theta}\|_2^2}{n}$ as desired.

LB 2: $r_n \leq \frac{1}{n}$. In this case, we use the pair of distributions, $Q_0 = N(\widehat{\theta}, I/n)$ and $Q_1 = N(\widehat{\theta} + \sqrt{r_n}\widehat{\theta}/\|\widehat{\theta}\|_2, I/n)$. By Lemma 3, we have that $\chi^2(Q_1, Q_0) \leq \alpha$ if $2nr_n \leq \alpha$.

The functional separation in this case:

$$\begin{aligned} 2s &:= \|\widehat{\theta}\|_2^2 \left(1 + \frac{\sqrt{r_n}}{\|\widehat{\theta}\|_2}\right)^2 - \|\widehat{\theta}\|_2^2 \\ &= r_n + 2\sqrt{r_n}\|\widehat{\theta}\|_2 \\ &\geq 2\sqrt{r_n}\|\widehat{\theta}\|_2. \end{aligned}$$

So the minimax rate is at least $r_n\|\widehat{\theta}\|_2^2$ as desired.

LB 3: Finally we show that r_n^2 is a lower bound. To see this we observe that without loss of generality we can assume that $\widehat{\theta}$ has finite norm since if it did not have finite norm we have already shown that the minimax rate is infinite. As a consequence, for any $\epsilon > 0$, and for any finite integer $d > 0$, we can find d indices, denoted as $\mathcal{I}$, such that $|\widehat{\theta}_j| \leq \epsilon/4$, for $j \in \mathcal{I}$.

We construct two mixtures Q_0 and Q_1^ϵ in the following way: we set $Q_0 = N(\widehat{\theta}, I/n)$ and Q_1^ϵ as described in the setup to Lemma 2, by perturbing each coordinate in $\mathcal{I}$ by $\pm\epsilon$ uniformly at random. Lemma 2 shows that the χ^2 distance is at most $dn^2\epsilon^4$. On the other hand the functional separation is,

$$\begin{aligned} 2s &:= \left[\|\widehat{\theta}\|_2^2 - \sum_{j\in\mathcal{I}}\widehat{\theta}_j^2 + \sum_{j\in\mathcal{I}}(\widehat{\theta}_j \pm \epsilon)^2\right] - \|\widehat{\theta}\|_2^2 \\ &= \sum_{j\in\mathcal{I}} \epsilon^2 \pm 2\widehat{\theta}_j\epsilon \\ &\geq \frac{1}{2}\sum_{j\in\mathcal{I}}\epsilon^2 = \frac{d\epsilon^2}{2}. \end{aligned}$$

On the other hand, the distribution Q_1^ϵ is supported on parameters $\widetilde{\theta}$ such that, $\|\widehat{\theta} - \widetilde{\theta}\|_2^2 = d\epsilon^2$. It remains to prescribe choices for d, ϵ to ensure that our constraints are satisfied. We choose $\epsilon = \sqrt{\alpha}\min\{1/(nr_n), 1/n\}$, and $d = r_n/\epsilon^2$. This ensures that the χ^2 distance is at most α, and as a consequence of Lemma 1 we obtain a minimax lower bound of order r_n^2 as desired.

B.2 Lower Bounds for the Integral of the Squared Density

We prove lower bounds corresponding to each term in our bound separately.

B.2.1 Lower Bound when $r_n \gtrsim 1/n$

We use Le Cam's two-point method. Define, $p_1 := \widehat{f}$ and for a sufficiently small $\varepsilon > 0$ define,

$$p_2(x) := p_1(x)\left(1 + \varepsilon p_1(x) - \varepsilon\theta\right),$$

where $\theta := \int \widehat{f}(x)^2 dx$. Since, the densities under consideration are bounded by $M > 0$, we observe that for sufficiently small $\varepsilon > 0$, p_2 is a valid density. Now, we observe that the Hellinger distance between these densities,

$$H^2(p_1, p_2) \leq \int \frac{(p_1(x) - p_2(x))^2}{p_1(x)} dx = \int \varepsilon^2 p_1(x)(p_1(x) - \theta)^2 dx = \varepsilon^2 (\int p_1(x)^3 dx - \theta^2) := \varepsilon^2 \gamma.$$

On the other hand, the functional separation is,

$$\begin{aligned} \int p_2^2(x) dx - \int p_1^2(x) dx &= 2\varepsilon (\int p_1(x)^3 dx - \theta^2) + \varepsilon^2 \int (p_1^2(x) - \theta p_1(x))^2 dx \\ &\geq 2\varepsilon (\int p_1(x)^3 dx - \theta^2) = 2\varepsilon\gamma. \end{aligned}$$

Finally, we note that the ℓ_2^2 distance is upper bounded as,

$$\int (p_1(x) - p_2(x))^2 dx = \int \varepsilon^2 p_1^2(x)(p_1(x) - \theta)^2 dx \lesssim \varepsilon^2 \gamma,$$

using the fact that $\|p_1\|_\infty \leq M$. Now, we set, ε such that, $\varepsilon^2 \gamma = \alpha/n$, for a sufficiently small constant $\alpha > 0$. We note that this construction is valid since $r_n \gtrsim 1/n$. Then the squared Hellinger distance between the n-fold product measures is at most α using Fact 1, and Lemma 1 yields a lower bound of order γ/n as claimed.

B.2.2 Lower Bound when $r_n \lesssim 1/n$

In this case, we follow the same construction as above except we choose ε such that, $\varepsilon^2 \gamma = \alpha r_n$, for a sufficiently small constant $\alpha > 0$. Then Lemma 1 yields a lower bound of order γr_n as claimed.

B.2.3 Lower Bound of Order r_n^2

To set the stage we derive a result on the chi-squared distance between a density, and a perturbed counterpart. Our result and proof are largely inspired by the proof of Lemma 4.4 in Balakrishnan and Wasserman [1], which in turn generalizes a result of Ingster [23].

For a given m, suppose that we divide S into $2m$ disjoint sets of equal volume, and pair these sets together into m (disjoint) pairs $\{(A_1, B_1), \ldots, (A_m, B_m)\}$. Now, suppose that we construct:

$$p_\lambda(x) = p_0(x) + \frac{h}{\sqrt{\mathrm{vol}(A_1)}} \sum_{j=1}^{m} \left[\lambda_j \left[\mathbb{I}(x \in A_j) - \mathbb{I}(x \in B_j)\right]\right],$$

where $0 \leq h/\sqrt{\mathrm{vol}(A_1)} \leq \inf_{x \in S} p_0(x)$, will be chosen appropriately in the sequel, $\lambda_j \in \{-1, +1\}$ will be chosen uniformly at random, and $\mathrm{vol}(A)$ denotes the Lebesgue measure of the set A. We note that by our choice of h, p_λ is a valid density, i.e. is positive everywhere and integrates to 1. We then have the following result:

Lemma 5. *Define Q to be the mixture obtained from choosing λ uniformly at random from $\{-1, +1\}^m$. If we satisfy the conditions:*

1. $h/\sqrt{vol(A_1)} \leq \inf_{x \in S} p_0(x)$

2. $nh^2/(\inf_{x\in S} p_0(x)) \leq 1$

3. $\frac{mn^2h^4}{(\inf_{x\in S} p_0(x))^2} \leq \ln(1+\alpha)$,

then we have that,

$$\chi^2(P_0^n, Q^n) \leq \alpha$$

Taking this result as given we can now complete the proof of our result. By our assumption, $\|\widehat{f}\|_\infty \leq M$, so we have that $\int \widehat{f}^2 < M$. Given an arbitrary pilot density $\widehat{f}$, and any $\epsilon > 0$ we can find a set S of volume 1 such that, $\widehat{f}(x) \leq \epsilon/4$ on S. We choose $\epsilon = \min\{\sqrt{r_n}/4\sqrt{1+M}, 1\}$.

Now, we construct an intermediate density $\widetilde{p} = (1-\epsilon)\widehat{f} + \epsilon p_U$, where p_U is the density of the uniform distribution on S. Now, we note that,

$$\int (\widetilde{p}(x) - \widehat{f}(x))^2 dx = \epsilon^2 \int (\widehat{f} - p_U)^2 dx \leq \epsilon^2 + \epsilon^2 \int \widehat{f}(x)^2 dx,$$

so for our choice of ϵ we obtain that, $\int (\widetilde{p}(x) - \widehat{f}(x))^2 dx \leq r_n/4$. We also note that $\|\widetilde{p}\|_\infty \leq M$, since $\|\widetilde{p}\|_\infty \leq \max\{\|\widehat{f}\|_\infty, 5\epsilon/4\}$. Now, on the set S, $\widetilde{p}$ is lower bounded as $\inf_x \widetilde{p}(x) \geq \epsilon$.

We construct perturbations p_λ as described above centered around the density $\widetilde{p}$, with $h = \epsilon/\sqrt{2m}$ (for an m we will choose in the sequel). We use Lemma 5 with $h = \epsilon/\sqrt{2m}$, and choose m large enough to ensure that, $n\epsilon/m \leq 1$, and $\frac{n^2\epsilon^2}{m} \leq \ln(1+\alpha)$. Each of the densities p_λ are uniformly upper bounded by a similar argument to the one for $\widetilde{p}$. Then as a consequence, we obtain that $\chi^2(\widetilde{P}^n, Q^n) \leq \alpha$. We note that,

$$\int (p_\lambda - \widehat{f})^2 dx \leq 2\int (p_\lambda - \widetilde{p})^2 dx + 2\int (\widetilde{p} - \widehat{f})^2 dx \leq 2\epsilon^2 + r_n/2 \leq r_n.$$

Finally, the functional separation is:

$$\int p_\lambda^2 - \int \widetilde{p}^2 = \int_S p_\lambda^2 - \int \widetilde{p}^2 \geq \frac{1}{2}(2\epsilon)^2 - \frac{25}{16}\epsilon^2 \geq \frac{7}{16}\epsilon^2 \gtrsim \min\{r_n/(1+M), 1\},$$

as desired.

B.2.4 Proof of Lemma 5

We first bound the expected squared likelihood ratio, which in turn implies a bound on the chi-squared distance. Let us denote $\widetilde{h} := h/\sqrt{\mathrm{vol}(A_1)}$. Suppose that we observe $\{Z_1, \ldots, Z_n\}$, then the likelihood ratio:

$$W_n(Z_1, \ldots, Z_n) = \frac{1}{2^m} \sum_{\lambda\in\{-1,+1\}^m} \prod_{i=1}^n \frac{p_\lambda(Z_i)}{p_0(Z_i)},$$

and

$$\begin{aligned} W_n^2(Z_1, \ldots, Z_n) &= \frac{1}{2^{2m}} \sum_{\lambda\in\{-1,+1\}^m} \sum_{\nu\in\{-1,+1\}^m} \prod_{i=1}^n \frac{p_\lambda(Z_i)p_\nu(Z_i)}{p_0^2(Z_i)} \\ &= \frac{1}{2^{2m}} \sum_{\lambda\in\{-1,+1\}^m} \sum_{\nu\in\{-1,+1\}^m} \prod_{i=1}^n \left(1 + \frac{\widetilde{h}\sum_{j=1}^m [\lambda_j [\mathbb{I}(Z_i \in A_j) - \mathbb{I}(Z_i \in B_j)]]}{p_0(Z_i)}\right) \times \\ &\quad \left(1 + \frac{\widetilde{h}\sum_{j=1}^m [\nu_j [\mathbb{I}(Z_i \in A_j) - \mathbb{I}(Z_i \in B_j)]]}{p_0(Z_i)}\right). \end{aligned}$$

Using the fact that the supports of the sets $\{(A_1, B_1), \ldots, (A_m, B_m)\}$ are all disjoint, taking the expected value over $Z_1, \ldots, Z_n$ and using their independence we obtain,

$$\mathbb{E}[W_n^2(Z_1, \ldots, Z_n)] = \frac{1}{2^{2m}} \sum_{\lambda \in \{-1,+1\}^m} \sum_{\nu \in \{-1,+1\}^m} \left(1 + \widetilde{h}^2 \sum_{j=1}^m \lambda_j \nu_j a_j\right)^n,$$

where

$$a_j = \int_{A_j \cup B_j} \frac{1}{p_0(x)} dx.$$

From this we can write,

$$\mathbb{E}[W_n^2(Z_1, \ldots, Z_n)] \leq \mathbb{E}_{\lambda,\nu} \exp(n\widetilde{h}^2 \sum_{j=1}^m \lambda_j \nu_j a_j) = \prod_{j=1}^m \cosh(n\widetilde{h}^2 a_j).$$

Now, using the fact that $\cosh(x) \leq 1 + x^2 \leq \exp(x^2)$ for $x \leq 1$, we obtain that,

$$\mathbb{E}[W_n^2(Z_1, \ldots, Z_n)] \leq \exp(n^2\widetilde{h}^4 \sum_{j=1}^m a_j^2) \leq \exp\left(\frac{mn^2h^4}{(\inf_{x \in S} p_0(x))^2}\right),$$

provided that, $n\widetilde{h}^2 a_j \leq 1$ for $j \in \{1, \ldots, m\}$. Finally, we note that the χ^2 distance is simply, $\mathbb{E}[W_n^2(Z_1, \ldots, Z_n)] - 1$, and so,

$$\chi^2(P_0^n, Q^n) \leq \alpha,$$

when $mn^2h^4/(\inf_{x \in S} p_0(x))^2 \leq \ln(1 + \alpha)$, as claimed.

B.3 Lower Bounds for the Expected Conditional Covariance

Now, we turn our attention to the expected conditional covariance. Our proof in this case is inspired by that of Theorem 4.1 of Robins et al. [40]. Following their work, we construct our lower bound in the case when $Y \in \{0, 1\}$. In this binary setup the joint distribution over triples (X, A, Y) can be parametrized by quadruples (μ, π, η, p_X). Recall our statistical model,

$$\mathcal{G}(r_n, s_n) := \Big\{(\mu, \pi, \eta, p_X) : \text{supp}(X) = [0,1]^d, p_X = \text{unif}[0,1]^d, \int (\mu(x) - \widehat{\mu}(x))^2 p_X(x) dx \leq r_n,$$
$$\int (\pi(x) - \widehat{\pi}(x))^2 p_X(x) dx \leq s_n, 0 \leq \pi(x), \mu(x) \leq 1, 1 - \varepsilon \geq \widehat{\pi}(x), \widehat{\mu}(x) \geq \varepsilon > 0, \text{for } x \in [0,1]^d\Big\}.$$

In our proof, we will often suppress dependence on the (universal) constant $\varepsilon > 0$.

Case 1: To begin with we show a lower bound of order $1/n$, even when the nuisance functions π, μ are known exactly, i.e. when $r_n = s_n = 0$. The likelihood in our model is given as:

$$p(X, A, Y) = p_X(X)\pi(X)^A(1 - \pi(X))^{1-A}(\mu(X) + (1 - \pi(X))\eta(X))^{AY}(\mu(X) - \pi(X)\eta(X))^{(1-A)Y} \times$$
$$(1 - \mu(X) - (1 - \pi(X))\eta(X))^{A(1-Y)}(1 - \mu(X) + \pi(X)\eta(X))^{(1-A)(1-Y)}.$$

We define a pair of distributions p_1, p_2, defined by quadruples $(\widehat{\mu}, \widehat{\pi}, 0, 1)$ and $(\widehat{\mu}, \widehat{\pi}, \zeta, 1)$, for some sufficiently small $\zeta > 0$. In particular, we will choose $\zeta \leq \varepsilon/2$. It is straighforward to verify that the functional separation between these distributions is,

$$\psi^{\text{cov}}(p_1) - \psi^{\text{cov}}(p_2) = \zeta \mathbb{E}[\widehat{\pi}^2(X)] \gtrsim \zeta.$$

On the other hand, the Hellinger distance between p_1 and p_2 can be calculated by noting:

$$p_1(X,A,Y) = \widehat{\pi}(X)^A(1-\widehat{\pi}(X))^{1-A}\widehat{\mu}(X)^Y(1-\widehat{\mu}(X))^{1-Y},$$
$$p_2(X,A,Y) = \widehat{\pi}(X)^A(1-\widehat{\pi}(X))^{1-A}(\widehat{\mu}(X)+(1-\widehat{\pi}(X))\zeta)^{AY}(\widehat{\mu}(X)-\widehat{\pi}(X)\zeta)^{(1-A)Y}\times$$
$$(1-\widehat{\mu}(X)-(1-\widehat{\pi}(X))\zeta)^{A(1-Y)}(1-\widehat{\mu}(X)+\widehat{\pi}(X)\zeta)^{(1-A)(1-Y)}.$$

For any (X,A,Y) we note that, $p_1(X,A,Y) \gtrsim \varepsilon^2$, and $|p_1(X,A,Y) - p_2(X,A,Y)| \lesssim \zeta$ from which we obtain that, the squared Hellinger distance,

$$H^2(p_1,p_2) \lesssim \zeta^2.$$

Choosing $\zeta \lesssim 1/\sqrt{n}$, using Fact 1 and applying Lemma 1 then yields a lower bound of order $1/n$ as desired.

The remainder of the analysis is devoted to proving a lower bound of order $r_n \times s_n$. We remark that in our setup this lower bound is simpler to derive that lower bounds for the expected conditional covariance under smoothness conditions [40], since we are not constrained by potentially different smoothnesses of the propensity score and outcome regression function. At a more technical level, it is not necessary in our setup to construct mixtures under both the null and alternate, and the bound on the Hellinger distance we need is essentially due to Birgé and Massart [5]. It is also not necessary to use different constructions depending on which of the propensity score or outcome regression is more difficult to estimate.

Case 2: In this case, we are aiming for a lower bound of $r_n \times s_n$, in the setting when $r_n \times s_n \gtrsim 1/n$. Consequently, we focus on lower bounds for the estimation of

$$\psi = \int \pi(x)\mu(x)p_X(x)dx,$$

since as we noted earlier the remaining term in the expected conditional covariance can be estimated at fast $\sqrt{n}$-rates.
Fix an integer $2m$, and denote by $B_1,\ldots,B_{2m}$ be $2m$ translates of the cube $(2m)^{-1/d}[0,1/2]^d$ which are disjoint, and contained in $[0,1]^d$, and let the bottom left corners of these cubes be $x_1,\ldots,x_{2m}$.

We now define,

$$\pi_\lambda(x) = \widehat{\pi}(x) + \frac{h_1}{\sqrt{\mathrm{vol}(B_1)}\widehat{\mu}(x)}\sum_{j=1}^m \lambda_j[\mathbb{I}(x\in B_{2j}(x)) - \mathbb{I}(x\in B_{2j-1}(x))],$$
$$\mu_\lambda(x) = \widehat{\mu}(x) + \frac{h_2}{\sqrt{\mathrm{vol}(B_1)}\widehat{\pi}(x)}\sum_{j=1}^m \lambda_j[\mathbb{I}(x\in B_{2j}(x)) - \mathbb{I}(x\in B_{2j-1}(x))],$$

where $\lambda_1,\ldots,\lambda_m$ will be chosen to be uniformly distributed on $\{-1,+1\}$, and h_1,h_2 will be chosen to ensure that $\varepsilon/2 \le \pi_\lambda,\mu_\lambda \le 1-\varepsilon/2$. We will set,

$$\eta_\lambda(x) = \frac{\widehat{\mu}(x)-\mu_\lambda(x)}{1-\pi_\lambda(x)}.$$

Now, we take a point null which we denote by p, to be defined by the quadruple $(\widehat{\mu},\widehat{\pi},0,1)$. The functional under the null takes the value,

$$\int \widehat{\pi}\widehat{\mu}p_X dx = \int \widehat{\pi}\widehat{\mu}.$$

Under the alternate, which we denote by q_λ we consider the mixture defined by the quadruple $(\mu_\lambda, \pi_\lambda, \eta_\lambda, 1)$, and note that the functional takes value,

$$\int \pi_\lambda \mu_\lambda p_X dx \geq \int \widehat{\pi}\widehat{\mu} + m h_1 h_2,$$

where we use the facts that the different bumps $\mathbb{I}(x \in B_j(x))$ do not overlap, and that $\int \mathbb{I}(x \in B_j(x))^2 dx = \text{vol}(B_j)$.

By construction, $\int (\widehat{\pi}(X) - \pi_\lambda(X))^2 p_X(x) dx \lesssim h_1^2/\text{vol}(B_1)$, and $\int (\widehat{\mu}(X) - \mu_\lambda(X))^2 p_X(x) dx \lesssim h_2^2 \text{vol}(B_1)$. So we can choose, $h_1 = \sqrt{\text{vol}(B_1)}\min\{\sqrt{r_n}, \varepsilon/2\}$, $h_2 = \sqrt{\text{vol}(B_1)}\min\{\sqrt{s_n}, \varepsilon/2\}$, to ensure that all the resulting nuisance functions are valid, and belong to $\mathcal{G}(r_n, s_n)$.

It remains to bound the Hellinger distance for which we will use the main result of Robins et al. [40]. In particular, we focus on upper bounding the terms a, b, d, p_j, $j \in \{1, \ldots, m\}$, in the preamble to their Theorem 2.1, which in turn yields a bound on the squared Hellinger distance. We follow their notation closely. The sample space is given by $[0,1]^d \times \{0,1\} \times \{0,1\}$ which we partition into the sets $\mathcal{X}_j$ which are $\{0,1\} \times \{0,1\} \times B_j \cup B_{j+1}$, and so the terms p_j are simply each equal to $1/m$. Under the null, we have that,

$$p(X, A, Y) = \widehat{\mu}(X)^Y (1 - \widehat{\mu}(X))^{1-Y} \widehat{\pi}(X)^A (1 - \widehat{\pi}(X))^{1-A}.$$

On the other hand, under the alternate we have that,

$$q_\lambda(X, A, Y) = \pi_\lambda(X)^A \widehat{\mu}(X)^{AY} (1 - \widehat{\mu}(X))^{A(1-Y)} (\mu_\lambda(X) - \pi_\lambda(X)\widehat{\mu}(X))^{(1-A)Y} \times$$
$$(1 - \pi_\lambda(X) - \mu_\lambda(X) + \pi_\lambda(X)\widehat{\mu}(X))^{(1-A)(1-Y)}.$$

It is easy to verify that if we denote $q = \mathbb{E}_\lambda(p_\lambda)$, then $p = q$, so the term d of Robins et al. [40] is 0. Since we do not use a mixture under the null the term a is also 0, so it only remains to bound the term b. We have that,

$$p - q_\lambda = (1 - A) \times (-1)^Y (\mu_\lambda - \widehat{\mu}) + (-1)^A \widehat{\mu}^Y (1 - \widehat{\mu})^{1-Y} (\pi_\lambda - \widehat{\pi}).$$

Since, $\varepsilon/2 \leq \pi_\lambda \leq 1 - \varepsilon/2$ we obtain that $p > 0$. A direct calculation then gives that, the term b of Robins et al. [40] is upper bounded upto constants by $m(h_1^2 + h_2^2)$. Theorem 2.1 of Robins et al. [40] then yields a bound on the Hellinger distance between the product measures:

$$H^2(p^n, \mathbb{E}_\lambda(q_\lambda^n)) \lesssim m n^2 (h_1^4 + h_2^4) \lesssim \frac{n^2}{m}(r_n^2 + s_n^2).$$

So taking m sufficiently large, we obtain that $H^2(p^n, \mathbb{E}_\lambda(q_\lambda^n)) \leq \alpha$, and via Lemma 1 we obtain a lower bound of order $m^2 h_1^2 h_2^2 \gtrsim r_n \times s_n$ as desired.

C Proof of Theorem 2

The analysis of the plugin and first-order estimators in the Gaussian sequence model follow directly from Lemma 6 of the next section. We focus first on analyzing the higher-order estimator in (7) before turning our attention to the integral of the squared density and the expected conditional covariance.

C.1 Upper Bounds for the Quadratic Functional in the Gaussian Sequence Model

We assume for simplicity that y^1 and y^2 are independent observations from the Gaussian sequence model. Observe that,

$$\begin{aligned}
\mathbb{E}|\widehat{Q}^{\theta}_{\mathrm{ho}} - Q(\theta^*)|^2 &= |\mathbb{E}\widehat{Q}^{\theta}_{\mathrm{ho}} - Q(\theta^*)|^2 + \mathbb{E}(\widehat{Q}^{\theta}_{\mathrm{ho}} - \mathbb{E}\widehat{Q}^{\theta}_{\mathrm{ho}})^2 \\
&= \left[\sum_{j=T+1}^{\infty} (\widehat{\theta}_j - \theta^*_j)^2\right]^2 + \frac{4\sum_{j=T+1}^{\infty}\theta^{*2}_j}{n} + \frac{4\sum_{j=1}^{T}\theta^{*2}_j}{n} + \frac{T}{n^2} \\
&\lesssim r_n^2 + \frac{\|\widehat{\theta}\|_2^2}{n} + \frac{T}{n^2},
\end{aligned}$$

as claimed.

C.2 Upper Bounds for the Integral of the Squared Density

We first analyze the plugin estimator.

$$\begin{aligned}
|\widehat{T}^{\mathrm{f}}_{\mathrm{pi}} - T^{\mathrm{f}}(f^*)|^2 &= \left|\int \widehat{f}^2 - \int f^{*2}\right|^2 \\
&= \left|\int (\widehat{f} - f^*)(\widehat{f} + f^*)\right|^2 \\
&\leq \left|\int (\widehat{f} - f^*)^2\right| \left|\int (2\widehat{f} + (f^* - \widehat{f}))^2\right| \\
&\lesssim \left|\int (\widehat{f} - f^*)^2\right| \left(\int \widehat{f}^2 + \int (f^* - \widehat{f}))^2\right) \\
&\leq r_n^2 + r_n \int \widehat{f}^2,
\end{aligned}$$

as claimed. For the first-order estimator we see that,

$$\begin{aligned}
\mathbb{E}|\widehat{T}^{\mathrm{f}}_{\mathrm{fo}} - T^{\mathrm{f}}(f^*)|^2 &= |\mathbb{E}\widehat{T}^{\mathrm{f}}_{\mathrm{fo}} - T^{\mathrm{f}}(f^*)|^2 + \mathbb{E}(\widehat{T}^{\mathrm{f}}_{\mathrm{fo}} - \mathbb{E}\widehat{T}^{\mathrm{f}}_{\mathrm{fo}})^2 \\
&= \left|\int (\widehat{f} - f^*)^2\right|^2 + 4\frac{\mathrm{var}(\widehat{f}(X))}{n} \\
&\lesssim r_n^2 + \frac{\mathrm{var}(\widehat{f}(X))}{n}.
\end{aligned}$$

C.3 Upper Bounds for the Expected Conditional Covariance

Once again, we first analyze the plugin estimator. We observe that,

$$\begin{aligned}
\mathbb{E}|\widehat{\psi}^{\mathrm{cov}}_{\mathrm{pi}} - \psi^{\mathrm{cov}}|^2 &= |\mathbb{E}\widehat{\psi}^{\mathrm{cov}}_{\mathrm{pi}} - \psi^{\mathrm{cov}}|^2 + \mathbb{E}(\widehat{\psi}^{\mathrm{cov}}_{\mathrm{pi}} - \mathbb{E}\widehat{\psi}^{\mathrm{cov}}_{\mathrm{pi}})^2 \\
&= \left|\int \widehat{\pi}\widehat{\mu} p_X - \int \pi^*\mu^* p_X\right|^2 + \frac{\mathrm{var}(AY - \widehat{\pi}(X)\widehat{\mu}(X))}{n}.
\end{aligned}$$

For the first term we observe that,

$$\begin{aligned}\left|\int \widehat{\pi}\widehat{\mu}p_X - \int \pi^*\mu^* p_X\right|^2 &\lesssim \left|\int (\widehat{\mu}-\mu^*)\widehat{\pi}p_X\right|^2 + \left|\int (\widehat{\pi}-\pi^*)\mu^* p_X\right|^2 \\ &\lesssim \left(\int (\widehat{\mu}-\mu^*)^2 p_X\right)\left(\int \widehat{\pi}^2 p_X\right) + \left(\int (\widehat{\pi}-\pi^*)^2 p_X\right)\left(\int \mu^{*2} p_X\right) \\ &\lesssim r_n\left(\int \widehat{\pi}^2 p_X\right) + s_n\left(\int (\widehat{\mu}^2 + (\widehat{\mu}-\mu^*)^2) p_X\right) \\ &\lesssim r_n \times s_n + r_n\left(\int \widehat{\pi}^2 p_X\right) + s_n\left(\int \widehat{\mu}^2 p_X\right),\end{aligned}$$

as desired. For the first-order estimator we have,

$$\begin{aligned}\mathbb{E}|\widehat{\psi}^{\text{cov}}_{\text{fo}} - \psi^{\text{cov}}|^2 &= |\mathbb{E}\widehat{\psi}^{\text{cov}}_{\text{fo}} - \psi^{\text{cov}}|^2 + \mathbb{E}(\widehat{\psi}^{\text{cov}}_{\text{fo}} - \mathbb{E}\widehat{\psi}^{\text{cov}}_{\text{pi}})^2 \\ &= \left|\int (\pi^* - \widehat{\pi})(\mu^* - \widehat{\mu}) p_X\right|^2 + \frac{\text{var}((A-\widehat{\pi}(X))(Y-\widehat{\mu}(X)))}{n} \\ &\leq \left(\int (\pi^* - \widehat{\pi})^2 p_X\right)\left(\int (\mu^* - \widehat{\mu})^2 p_X\right) + \frac{\text{var}((A-\widehat{\pi}(X))(Y-\widehat{\mu}(X)))}{n} \\ &\leq r_n \times s_n + \frac{\text{var}((A-\widehat{\pi}(X))(Y-\widehat{\mu}(X)))}{n},\end{aligned}$$

as claimed.

D An Adaptive Estimate of the Quadratic Functional

Throughout our paper we focused our discussion on the case when the error of the pilot estimate was larger than the variance of the first-order estimator. This situation is typical. In this section, we briefly investigate the setting where the pilot may be super-accurate.

D.1 Upper Bounds

For any $\delta > 0$, consider the following adaptive estimate:

$$\widehat{Q}^{\theta}_{\text{ad}} = \begin{cases} \widehat{Q}^{\theta}_{\text{pi}} \text{ if } |\widehat{Q}^{\theta}_{\text{pi}} - \widehat{Q}^{\theta}_{\text{fo}}| \leq 4\|\widehat{\theta}\|_2 \sqrt{\frac{4\log(2/\delta)}{n}} \\ \widehat{Q}^{\theta}_{\text{fo}} \text{ otherwise.} \end{cases}$$

The estimate is similar to a Lepski-style adaptive estimator which chooses between the plugin and first-order estimate. The following result then holds:

Theorem 3. *For any $\delta > 0$, the risk of $\widehat{Q}^{\theta}_{ad}$ is upper bounded as:*

$$\mathbb{E}[(\widehat{Q}^{\theta}_{ad} - Q(\theta^*))^2] \lesssim r_n^2 + \min\left\{ r_n\|\widehat{\theta}\|_2^2 + \frac{\delta\|\widehat{\theta}\|_2^2}{n}, \frac{\|\widehat{\theta}\|_2^2 \log(1/\delta)}{n} \right\}.$$

Remarks:

1. Suppose that ignored the terms which depend on δ, then the estimator $\widehat{Q}^{\theta}_{\text{ad}}$ is (fully) adaptive, i.e. it achieves the same performance as an oracle which always picked the estimate with lower risk.

2. More generally, there is a small price to pay to adapt between these estimates. This price is closely related to the standard Hodges superefficiency phenomenon in estimating a Gaussian mean. We develop this further in Appendix D.2.

Proof. Let us define $\|\widehat{\theta} - \theta^*\|_2^2 := R$. We first note the following error bounds on our estimates:

Lemma 6. *We have the following bounds:*

$$|\widehat{Q}^{\theta}_{pi} - Q(\theta^*)|^2 \lesssim R^2 + R\|\widehat{\theta}\|_2^2$$
$$\mathbb{E}|\widehat{Q}^{\theta}_{fo} - Q(\theta^*)|^2 \lesssim R^2 + \frac{\|\widehat{\theta}\|_2^2}{n}$$
$$\sqrt{\mathbb{E}|\widehat{Q}^{\theta}_{fo} - Q(\theta^*)|^4} \lesssim R^2 + \frac{\|\widehat{\theta}\|_2^2}{n}.$$

Taking this result as given we can complete the proof. Recall, that we observe $y = \theta^* + \epsilon$. We define the event E to be the event on which,

$$|\langle \epsilon, \widehat{\theta} \rangle| \leq \|\widehat{\theta}\|_2 \sqrt{\frac{4\log(2/\delta)}{n}},$$

which happens with probability at least $1 - \delta^2$, by applying a standard Gaussian tail bound.

Now observe that when $R \leq \frac{4\log(2/\delta)}{n}$, and on the event E:

$$\begin{aligned}|\widehat{Q}^{\theta}_{\text{pi}} - \widehat{Q}^{\theta}_{\text{fo}}| = |2(\|\widehat{\theta}\|_2^2 - \langle y, \widehat{\theta}\rangle)| &\leq 2\|\widehat{\theta}\|_2 \sqrt{\frac{4\log(2/\delta)}{n}} + 2|\langle \widehat{\theta}, \theta^* - \widehat{\theta}\rangle| \\ &\leq 2\|\widehat{\theta}\|_2 \sqrt{\frac{4\log(2/\delta)}{n}} + 2\sqrt{R}\|\widehat{\theta}\|_2 \leq 4\|\widehat{\theta}\|_2 \sqrt{\frac{4\log(2/\delta)}{n}},\end{aligned}$$

so our selection rule picks the estimate $\widehat{Q}^{\theta}_{\text{pi}}$.

Now, we are in a position to analyze our selection rule. Let us denote an index j which takes value 1 when $\widehat{Q}^{\theta}_{\text{ad}} = \widehat{Q}^{\theta}_{\text{pi}}$ and 2 otherwise. We consider two cases, when $R \leq \sqrt{\frac{4\log(2/\delta)}{n}}$ and when $R > \sqrt{\frac{4\log(2/\delta)}{n}}$. In the first case,

$$\begin{aligned}\mathbb{E}[(\widehat{Q}^{\theta}_{\text{ad}} - Q(\theta^*))^2] &= \mathbb{E}[(\widehat{Q}^{\theta}_{\text{ad}} - Q(\theta^*))^2 \mathbb{I}[j=1]] + \mathbb{E}[(\widehat{Q}^{\theta}_{\text{ad}} - Q(\theta^*))^2 \mathbb{I}[j=2]] \\ &\lesssim R^2 + R\|\widehat{\theta}\|_2^2 + \delta\sqrt{\mathbb{E}[(\widehat{Q}^{\theta}_{\text{ad}} - Q(\theta^*))^4} \\ &\lesssim R^2 + R\|\widehat{\theta}\|_2^2 + \delta\left[R^2 + \frac{\|\widehat{\theta}\|_2^2}{n}\right] \\ &\lesssim R^2 + R\|\widehat{\theta}\|_2^2 + \frac{\delta\|\widehat{\theta}\|_2^2}{n}.\end{aligned}$$

In the second case,

$$\begin{aligned}\mathbb{E}[(\widehat{Q}^{\theta}_{\text{ad}} - Q(\theta^*))^2] &= \mathbb{E}[(\widehat{Q}^{\theta}_{\text{ad}} - Q(\theta^*))^2 \mathbb{I}[j=1]] + \mathbb{E}[(\widehat{Q}^{\theta}_{\text{ad}} - Q(\theta^*))^2 \mathbb{I}[j=2]] \\ &= \mathbb{E}[(\widehat{Q}^{\theta}_{\text{pi}} - Q(\theta^*))^2 \mathbb{I}[j=1]] + \mathbb{E}[(\widehat{Q}^{\theta}_{\text{fo}} - Q(\theta^*))^2 \mathbb{I}[j=2]] \\ &\lesssim \mathbb{E}[(\widehat{Q}^{\theta}_{\text{pi}} - \widehat{Q}^{\theta}_{\text{fo}})^2 \mathbb{I}[j=1]] + \mathbb{E}[(\widehat{Q}^{\theta}_{\text{fo}} - Q(\theta^*))^2] \\ &\lesssim R^2 + \frac{\|\widehat{\theta}\|_2^2 \log(1/\delta)}{n}.\end{aligned}$$

Putting these together we obtain the desired theorem. ∎

Proof of Lemma 6: We prove each of the three claims in turn. Observe that,

$$\begin{aligned}|\widehat{Q}^{\theta}_{\text{pi}} - Q(\theta^*)|^2 &= |\|\widehat{\theta}\|_2^2 - \|\theta^*\|_2^2|^2 \\ &= |(\widehat{\theta} - \theta^*)^T(\widehat{\theta} + \theta^*)|^2 \\ &\lesssim R(\|\widehat{\theta}\|_2^2 + R),\end{aligned}$$

as desired. Now, we see that,

$$\begin{aligned}\mathbb{E}|\widehat{Q}^{\theta}_{\text{fo}} - Q(\theta^*)|^2 &= (\mathbb{E}(\widehat{Q}^{\theta}_{\text{fo}}) - Q(\theta^*))^2 + \mathbb{E}(\mathbb{E}(\widehat{Q}^{\theta}_{\text{fo}}) - \widehat{Q}^{\theta}_{\text{fo}})^2 \\ &= \|\widehat{\theta} - \theta^*\|_2^4 + \frac{4\|\widehat{\theta}\|_2^2}{n} \\ &\lesssim R^2 + \frac{\|\widehat{\theta}\|_2^2}{n},\end{aligned}$$

where the bounds on the bias and variance follow from a direct calculation. Finally,

$$\begin{aligned}\sqrt{\mathbb{E}|\widehat{Q}^{\theta}_{\text{fo}} - Q(\theta^*)|^4} &\lesssim \sqrt{(\mathbb{E}(\widehat{Q}^{\theta}_{\text{fo}}) - Q(\theta^*))^4 + \mathbb{E}(\mathbb{E}(\widehat{Q}^{\theta}_{\text{fo}}) - \widehat{Q}^{\theta}_{\text{fo}})^4} \\ &\lesssim R^2 + \frac{\|\widehat{\theta}\|_2^2}{n},\end{aligned}$$

where to bound the second term, we simply use the fact that the fourth moment of a mean zero Gaussian random variable is $3\sigma^4$.

D.2 An Adaptivity Lower Bound

In this section, we prove the following complementary lower bound which shows a sense in which our adaptive estimator is unimprovable. We denote the risk of our adaptive estimate $\widehat{Q}^{\theta}_{\text{ad}}$ as:

$$f_\delta(r) = r^2 + \min\left\{ r\|\widehat{\theta}\|_2^2 + \frac{\delta\|\widehat{\theta}\|_2^2}{n}, \frac{\log(1/\delta)\|\widehat{\theta}\|_2^2}{n} \right\},$$

where $\delta > 0$ is chosen to be sufficiently small.

Lemma 7. *Suppose that we have an estimate $\widehat{Q}$ such that for some $r_1 \geq 0$, and for a sufficiently small $\varepsilon > 0$,*

$$\sup_{\theta^* \in \Theta(r_1)} \mathbb{E}(\widehat{Q} - Q(\theta^*))^2 \lesssim \varepsilon f_\delta(r_1)$$

then there exists an r_2, such that,

$$\sup_{\theta^* \in \Theta(r_2)} \mathbb{E}(\widehat{Q} - Q(\theta^*))^2 \gtrsim \frac{\log(1/(\varepsilon\delta))}{\log(1/\delta)} f_\delta(r_2),$$

so long as $\|\widehat{\theta}\|_2^2 \geq \log^2(1/(\epsilon\delta))/(n\log(1/\delta))$.

This result captures the fact that there is a strong sense in which Theorem 3 achieves the limits of adaptation. In particular, if an estimator has a (very) small risk relative to $\widehat{Q}^{\theta}_{\text{ad}}$ for some value of the unknown radius r_1, then this must come at the expense of a worse

performance at a different value r_2. Our lower bound follows the well-trodden route of using the constrained risk inequality in Lemma 4 to argue that achieving a (very) small risk at a point in the parameter space comes at the expense of a larger risk in the neighborhood of that point.

We note the (mild) restriction that $\|\widehat{\theta}\|_2^2 \geq \log^2(1/(\epsilon\delta))/(n\log(1/\delta))$. When this assumption does not hold, the adaptation picture is different. This can be seen by observing that when $\|\widehat{\theta}\|_2 = 0$, the estimate $\widehat{Q} = 0$ is adaptively optimal for any r, and there is no price to pay for adaptation.

Proof. We consider two cases, when $r_1 \leq \log(1/\delta)/n$ and when $r_1 \geq \log(1/\delta)/n$.
Case 1: When $r_1 \leq \log(1/\delta)/n$ we can write,

$$f_\delta(r_1) \leq r_1^2 + r_1\|\widehat{\theta}\|_2^2 + \frac{\delta\|\widehat{\theta}\|_2^2}{n},$$

since $\delta > 0$ is sufficiently small to ensure that $\delta \leq \log(1/\delta)$. Now, this implies that $\widehat{Q}$ satisfies,

$$\sup_{\theta^* \in \Theta(r_1)} \mathbb{E}(\widehat{Q} - Q(\theta^*))^2 \lesssim \varepsilon \left(r_1^2 + r_1\|\widehat{\theta}\|_2^2 + \frac{\delta\|\widehat{\theta}\|_2^2}{n} \right). \tag{12}$$

The lower bounds in Appendix B.1 already show that,

$$\sup_{\theta^* \in \Theta(r_1)} \mathbb{E}(\widehat{Q} - Q(\theta^*))^2 \gtrsim r_1^2 + r_1\|\widehat{\theta}\|_2^2,$$

so the claimed improvement is only possible if the final term in (12) dominates. In this case we have that,

$$\sup_{\theta^* \in \Theta(r_1)} \mathbb{E}(\widehat{Q} - Q(\theta^*))^2 \lesssim \frac{\varepsilon\delta\|\widehat{\theta}\|_2^2}{n}.$$

We now apply Lemma 4 with the choices, $\theta := \widehat{\theta}$, $\beta := \sqrt{\varepsilon\delta}\|\widehat{\theta}\|_2/\sqrt{n}$, and $\alpha := \sqrt{\log(1/(\varepsilon\delta))/n}$, to conclude that,

$$\sup_{\theta^* \in \Theta(\alpha^2)} \mathbb{E}(\widehat{Q} - Q(\theta^*))^2 \gtrsim \alpha^4 + \frac{\log(1/(\varepsilon\delta))\|\widehat{\theta}\|_2^2}{n} \gtrsim \frac{\log(1/(\varepsilon\delta))\|\widehat{\theta}\|_2^2}{n}.$$

On the other hand, we have that the risk of $\widehat{Q}^\theta_{\text{ad}}$ is upper bounded as:

$$f_\delta(\alpha^2) \leq \alpha^4 + \frac{\log(1/\delta)\|\widehat{\theta}\|_2^2}{n} \lesssim \frac{\log(1/\delta)\|\widehat{\theta}\|_2^2}{n},$$

using our assumed lower bound on $\|\widehat{\theta}\|_2$. This in turn establishes the non-adaptivity claim.

Case 2: We now consider the case when $r_1 \geq \log(1/\delta)/n$. The lower bounds in Appendix B.1 already show that

$$\sup_{\theta^* \in \Theta(r_1)} \mathbb{E}(\widehat{Q} - Q(\theta^*))^2 \gtrsim r_1^2 + \frac{\|\widehat{\theta}\|_2^2}{n},$$

so once again the claimed improvement is only possible if the second term dominates and we have,

$$\sup_{\theta^* \in \Theta(r_1)} \mathbb{E}(\widehat{Q} - Q(\theta^*))^2 \lesssim \frac{\varepsilon \log(1/\delta) \|\widehat{\theta}\|_2^2}{n}.$$

Now, once again we apply Lemma 4 with the choices, $\theta := \widehat{\theta}$, $\beta := \sqrt{\varepsilon \log(1/\delta)} \|\widehat{\theta}\|_2 / \sqrt{n}$, and $\alpha := \sqrt{\log(1/(\varepsilon \log(1/\delta))/n}$. To obtain that,

$$\sup_{\theta^* \in \Theta(0)} \mathbb{E}(\widehat{Q} - Q(\theta^*))^2 \gtrsim \frac{\log(1/(\varepsilon \log(1/\delta))) \|\widehat{\theta}\|_2^2}{n}.$$

On the other hand, the estimator $\widehat{Q}^{\theta}_{\mathrm{ad}}$ achieves the guarantee,

$$\sup_{\theta^* \in \Theta(0)} \mathbb{E}(\widehat{Q}^{\theta}_{\mathrm{ad}} - Q(\theta^*))^2 \lesssim \frac{\delta \|\widehat{\theta}\|_2^2}{n}.$$

So we obtain that, taking $r_2 = 0$,

$$\sup_{\theta^* \in \Theta(r_2)} \mathbb{E}(\widehat{Q} - Q(\theta^*))^2 \gtrsim \frac{\log(1/(\varepsilon \log(1/\delta)))}{\delta} f_\delta(r_2) \gtrsim \frac{\log(1/(\varepsilon\delta))}{\log(1/\delta)} f_\delta(r_2).$$

These two facts taken together yield the non-adaptivity claim of the theorem.

□