Submitted to the Annals of Statistics

NOISY RECOVERY FROM RANDOM LINEAR OBSERVATIONS: SHARP MINIMAX RATES UNDER ELLIPTICAL CONSTRAINTS

BY REESE PATHAK[1,a], MARTIN J. WAINWRIGHT[2,b] AND LIN XIAO[3,c]

[1]*EECS, University of California, Berkeley,* [a]*pathakr@berkeley.edu*

[2]*EECS, Mathematics, Statistics and Data Science Center, Massachusetts Institute of Technology,* [b]*wainwrigwork@gmail.com*

[3]*FAIR, Meta Inc.,* [c]*linx@meta.com*

Estimation problems with constrained parameter spaces arise in various settings. In many of these problems, the observations available to the statistician can be modelled as arising from the noisy realization of the image of a random linear operator; an important special case is random design regression. We derive sharp rates of estimation for arbitrary compact elliptical parameter sets and demonstrate how they depend on the distribution of the random linear operator. Our main result is a functional that characterizes the minimax rate of estimation in terms of the noise level, the law of the random operator, and elliptical norms that define the error metric and the parameter space. This nonasymptotic result is sharp up to an explicit universal constant, and it becomes asymptotically exact as the radius of the parameter space is allowed to grow. We demonstrate the generality of the result by applying it to both parametric and nonparametric regression problems.

1. Introduction.

In this paper, we study the problem of estimating an unknown vector $\theta^\star$ on the basis of random linear observations corrupted by noise. More concretely, suppose that we observe a random operator T_ξ and a random vector y, which are linked via the equation

$$y = T_\xi(\theta^\star) + w. \tag{1}$$

This observation model involves two forms of randomness: the unobserved vector w, which is a form of additive observation noise, and the observed operator T_ξ, which is random, as indicated by its dependence on an underlying random variable ξ, and is linear in the argument $\theta^\star$.

While relatively simple in appearance, the observation model (1) captures a broad range of statistical estimation problems.

EXAMPLE 1 (Linear regression). We begin with a simple but widely used model: linear regression. The goal is to estimate the coefficients $\theta^\star \in \mathbf{R}^d$ that define the best linear predictor $x \mapsto \langle x, \theta^\star \rangle$ of some real-valued response variable $Y \in \mathbf{R}$. In order to do so, we observe a collection of (x_i, y_i) pairs linked via the noisy observation model

$$y_i = \langle x_i, \theta^\star \rangle + w_i \qquad \text{for } i = 1, \ldots, n.$$

If we define the concatenated vector $y = (y_1, \ldots, y_n)$, with an analogous definition for w, this is a special case of our general setup with the random linear operator $T_\xi : \mathbf{R}^d \to \mathbf{R}^n$ given by

$$[T_\xi(\theta)]_i = \langle x_i, \theta \rangle \quad \text{for } i = 1, \ldots, n. \tag{2}$$

Here, the random index corresponds to the covariate vectors so that $\xi = (x_1, \ldots, x_n)$; note that we have imposed no assumptions on the dependence structure of these covariate vectors. In the classical setting, these covariates are assumed to be drawn in an i.i.d. manner; however, our general set-up is by no means limited to this classical setting. In the sequel, we consider various examples with interesting dependence structure, and our theory gives some very precise insights into the effects of such dependence.

EXAMPLE 2 (Nonparametric regression). In the preceding example, we discussed the problem of predicting a response variable $Y \in \mathbf{R}$ in a linear manner. Let us consider the nonparametric generalization: here our goal is to estimate the regression function $f^\star(x) := \mathbf{E}[Y \mid X = x]$, which need not be linear as a function of x. Given observations $\{(x_i, y_i)\}_{i=1}^n$, we can write them in the form

$$y_i = f^\star(x_i) + w_i, \qquad \text{for } i = 1, \ldots, n,$$

where $w_i = y_i - \mathbf{E}[Y \mid X = x_i]$ are zero-mean noise variables.

Now let us suppose that $f^\star$ belongs to some function class $\mathcal{F}$ contained with $L^2(\mathcal{X})$, and show how this observation model can be understand as a special case of our setup with $\theta^\star \in \ell^2(\mathbf{N})$. Take some orthonormal basis $\{\phi_j\}_{j \geqslant 1}$ of $L^2(\mathcal{X})$. Any function in $\mathcal{F}$ can then be expanded as $f = \sum_{j \geqslant 1} \theta_j \phi_j$ for some sequence $\theta \in \ell^2(\mathbf{N})$. Letting $\xi = (x_1, \ldots, x_n)$, we can define the operator $T_\xi : \ell^2(\mathbf{N}) \to \mathbf{R}^n$ via

$$\theta \mapsto [T_\xi(\theta)]_i := \sum_{j=1}^{\infty} \theta_j \phi_j(x_i) \quad \text{for } i = 1, \ldots, n,$$

so that this problem can be written in the form of our general model (1). Observe that the randomness in the observation operator T_ξ arises via the randomness in sampling the covariates $\{x_i\}_{i=1}^n$.

EXAMPLE 3 (Tomographic reconstruction). The problem of tomographic reconstruction refers to the problem of recovering an image, modeled as a real-valued function $f^\star$ on some compact domain $\mathcal{X} \subset \mathbf{R}^2$, based on noisy integral measurements. Formally, we observe responses of the form

$$y_i = \int_{\mathcal{X}} h(x_i, u) f^\star(u)\, \mathrm{d}u + w_i \qquad \text{for } i = 1, \ldots, n,$$

where $h : \mathbf{R}^2 \times \mathbf{R}^2 \to \mathbf{R}$ is a known window function. If we again view $f^\star$ as belonging to some function class $\mathcal{F}$ within $L^2(\mathcal{X})$, then we can write this model in our general form with

$$[T_\xi(v)]_i = \sum_{j \geqslant 1} v_j \Big[\int_{\mathcal{X}} h(x_i, u) \phi_j(u)\, \mathrm{d}u \Big], \quad \text{and } \xi = (x_1, \ldots, x_n).$$

Here we have followed the same conversion as in Example 2, in particular re-expressing $f^\star$ in terms of its generalized Fourier coefficients with respect to an orthonormal family $\{\phi_j\}_{j \geqslant 1}$.

EXAMPLE 4 (Error-in-variables). Consider the Berkson variant [6, 14] of the error-in-variables problem in nonparametric regression. In this problem, an observed covariate x—instead of being associated with a noisy observation of $f^\star(x)$—is associated with a noisy observation of the "jittered" evaluation $f^\star(x + u)$, where $u \in \mathbf{R}$ is the random jitter. Formally, we observe n pairs (x_i, y_i) of the form

$$y_i = f^\star(x_i + u_i) + \varepsilon_i \qquad \text{for } i = 1, \ldots, n,$$

where the unobserved random jitter u_i is drawn independently of the pair (x_i, ε_i). We can re-write these observations as a special case of our general model with $\xi = (x_1, \ldots, x_n)$, and

$$[T_\xi(f)]_i := \mathbf{E}_u \big[f(x_i + u) \big], \quad \text{and} \quad w_i := \varepsilon_i + \Big\{ f(x_i + u_i) - \mathbf{E}_u \big[f(x_i + u) \big] \Big\} \quad \text{for } i = 1, \ldots, n.$$

Note that the new noise variables w_i are again zero-mean, and our assumption that T_ξ is observed means that the distribution of the jitter u is known.

These examples (and others, as discussed below in Section 1.2) motivate our study of the operator model (1). As we discuss in further detail later, a key advantage of writing the observation model in this form is that it will allow us to separate three key components of the difficulty of the problem: (i) the distribution of the random operator T_ξ, as expressed via the distribution of ξ, (ii) the distribution of the noise variable $w \coloneqq y - T_\xi \theta^\star$, and (iii) the constraints on the unknown parameter $\theta^\star$.

1.1. *Problem formulation, notation, and assumptions.* With these motivating examples in mind, we now turn to a more precise mathematical formulation of the estimation problem introduced above.

1.1.1. *Assumptions on the random variables* (ξ, w). Let us start by discussing properties of the random operator T_ξ. In the examples previously introduced, the domain of the observation operator T_ξ was either a subset of $\mathbf{R}^d$, or more generally, a subset of the sequence space $\ell^2(\mathbf{N})$. The bulk of our analysis focuses on the finite-dimensional setting —i.e., with domain $\mathbf{R}^d$—so that T_ξ can be identified with a random matrix $\mathbf{R}^{n \times d}$, for some pair (n, d) of positive but finite integers. However, as we highlight in Section 3.2, simple approximation arguments can be used to leverage our finite-dimensional results to determine minimax rates of convergence for estimating an element $\theta^\star$ of the infinite-dimensional sequence space $\ell^2(\mathbf{N})$.

In terms of the probabilistic structure of T_ξ, we assume the random element ξ lies in the measurable space $(\Xi, \mathcal{E})$, and is drawn from a probability measure $\mathbb{P}$ on the same space. Throughout we take $\mathcal{E}$ to be large enough such that linear functionals of T_ξ are measurable.

As for the noise vector $w \in \mathbf{R}^n$, we assume it is drawn—conditionally on ξ—from a noise distribution with conditional mean zero, and bounded conditional covariance. Formally, we assume that $w \sim \nu(\cdot \mid \xi)$ where ν is a Borel regular conditional probability on $\mathbf{R}^n$ that satisfies the following two conditions:

(N1) For $\mathbb{P}$-almost every $\xi \in \Xi$, we have $\int w \, \nu(\mathrm{d}w \mid \xi) = 0$; and
(N2) For $\mathbb{P}$-almost every $\xi \in \Xi$, we have

$$\int (u^\mathsf{T} w)^2 \, \nu(\mathrm{d}w \mid \xi) \leqslant u^\mathsf{T} \Sigma_w u, \qquad \text{for any fixed } u \in \mathbf{R}^n.$$

We write that the measure ν lies in the set $\mathcal{P}(\Sigma_w)$ when these two conditions are satisfied.

In words, Assumption (N1) requires that w is conditionally centered, and Assumption (N2) assumes that the conditional covariance of w is almost surely upper bounded in the semidefinite ordering by Σ_w. Let $\mathbb{P} \times \nu$ denote the distribution of the tuple (ξ, w); in explicit terms, writing $(\xi, w) \sim \mathbb{P} \times \nu$ means that $\xi \sim \mathbb{P}$ and $w \mid \xi \sim \nu(\cdot \mid \xi)$. Having specified the joint law of (ξ, w), the random variable y then satisfies the stated observation model (1).

1.1.2. *Decision-theoretic formulation.* In this paper, our goal to estimate $\theta^\star$ to the best possible accuracy as measured by a fixed quadratic form. To make this rigorous, we introduce two symmetric positive definite matrices K_e and K_c, which induce (respectively) the squared norms

$$\|\theta\|_{K_e}^2 \coloneqq \langle \theta, K_e \theta \rangle \quad \text{and} \quad \|\theta\|_{K_c^{-1}}^2 \coloneqq \langle \theta, {K_c}^{-1} \theta \rangle,$$

defined for any $\theta \in \mathbf{R}^d$. We seek estimates $\widehat{\theta}$ of $\theta^\star$ that have low squared *estimation error* $\|\widehat{\theta} - \theta^\star\|_{K_e}^2$, as defined by the matrix K_e. In parallel, we assume that underlying parameter is bounded in the *constraint norm*, so that it lies in the ellipse

$$\Theta(\varrho, K_c) \coloneqq \Big\{ \theta \in \mathbf{R}^d : \|\theta\|_{K_c^{-1}} \leqslant \varrho \Big\}$$

with radius R, as defined by the matrix K_c.

With this notation in hand, the central object of study in this paper is the *minimax risk*

$$\mathfrak{M}(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c) := \inf_{\widehat{\theta}} \sup_{\substack{\theta^\star \in \Theta(\varrho, K_c) \\ \nu \in \mathcal{P}(\Sigma_w)}} \mathbf{E}_{(\xi, w) \sim \mathbb{P} \times \nu} \Big[\|\widehat{\theta} - \theta^\star\|_{K_e}^2 \Big], \tag{3}$$

where the infimum ranges over all measurable functions $\widehat{\theta} \equiv \widehat{\theta}(T_\xi, y)$ that map the observed pair (T_ξ, y) to $\mathbf{R}^d$. Note that by straightforward rescaling arguments, one can always take one of the three operators (Σ_w, K_e, K_c) to be equal to the identity. Moreover, one can "absorb" the radius ϱ into the constraint matrix K_c so that without loss of generality it is equal to 1. For convenience in deriving results in particular problems, we have presented our main results without making these reductions.

1.2. *Examples of choices of sampling laws, constraints and error norms.* As discussed previously, our general theory accommodates various forms of the random linear operators T_ξ. As might one expect, the sampling law $\mathbb{P}$ for ξ changes the statistical structure of the observations, and so influences the quality of the best possible estimates. Moreover, the interaction between $\mathbb{P}$ and the geometry of the error norm, as defined by the matrix K_e, plays an important role. Finally, both of these factors interact with the geometry of the constraint set, as determined by the matrix K_c.

Below we discuss some examples of these types of interactions. To be clear, each of these statistical settings have been considered separately in the literature previously; one benefit of our approach is that it provides a unifying framework that includes each of these problems as special cases.

EXAMPLE 5 (Covariate shift in linear regression). Recall the set-up for linear regression, as introduced in Example 1. In practice, the *source distribution* from which the covariates x are sampled when constructing an estimate of $\theta^\star$ need not be the same as the *target distribution* of covariates on which the predictor is to be deployed. This phenomenon—a discrepancy between the source and target distributions—is known as *covariate shift*. It is now known to arise in a wide variety of applications (e.g., see the papers [43, 39] and references therein for more details).

As one concrete example, in healthcare applications, the covariate vector $x \in \mathbf{R}^d$ might correspond to various diagnostic measures run on a given patient, and the response $y \in \mathbf{R}$ could correspond to some outcome variable (e.g., blood pressure). Clinicians might use one population of patients to develop a predictive model relating the diagnostic measures x to the outcome y, but then be interested in making predictions for a related but distinct population of patients.

In our setting, suppose that we use the linear model $\theta \mapsto \widehat{y} := \langle \theta, x \rangle$ to make predictions over a collection of covariates with distribution Q. A simple computation shows that the mean-squared prediction error, averaging over both the noise w and random covariates x, takes the form

$$\mathbf{E}\left[(\widehat{y} - y)^2\right] = \underbrace{(\theta - \theta^\star)^{\mathsf{T}} \Sigma_Q (\theta - \theta^\star)}_{=: L_Q(\widehat{\theta}, \theta^\star)} + c, \qquad \text{where} \quad \Sigma_Q := \mathbf{E}_Q[x \otimes x],$$

and c is a constant independent of the pair $(\theta, \theta^\star)$. Thus, the excess prediction error over the new population Q corresponds to taking $K_e = \Sigma_Q$ in our general set-up. Similarly, if one wanted to assess parameter error, then studying the minimax risk with the choice $K_e = I_d$ would be reasonable. Finally, the error in the original population (denoted P) can be assessed with the choice $K_e = \Sigma_P := \mathbf{E}_P[x \otimes x]$.

Among the claims in the paper of Mourtada [51] is the following elegant result: when no constraints are imposed on $\theta^\star$, the minimax risk in the squared metric $L_Q(\widehat{\theta}, \theta^\star) = \|\widehat{\theta} - \theta^\star\|_{\Sigma_Q}^2$ is equal to

$$\inf_{\widehat{\theta}} \sup_{\theta^\star \in \mathbf{R}^d} \mathbf{E}\left[L_Q(\widehat{\theta}, \theta^\star)\right] = \frac{\sigma^2}{n} \mathbf{E}[\mathbf{Tr}(\Sigma_n^{-1}\Sigma_Q)], \tag{4}$$

where Σ_n denotes the sample covariance matrix $(1/n)\sum_{i=1}^n x_i \otimes x_i$, and the expectation is over $x_1, \ldots, x_n \overset{\text{IID}}{\sim} P$. Thus, the fundamental rate of estimation depends on the distribution of the sample covariance matrix, the noise level, and the target distribution Q.

In this paper, we derive related but more general results that allow for many other choices of the error metric and, perhaps more importantly, permit the statistician to incorporate constraints on the parameter $\theta^\star$. We demonstrate in Section 3.1.3 that these more general results allow us to recover the known relation (23) via a simple limiting argument where the constraint radius tends to infinity.

EXAMPLE 6 (Nonparametric regression with non-uniform sampling). Consider observing covariate-target pairs $\{(x_i, y_i)\}_{i=1}^n$ where y_i is modeled as being a noisy realization of a conditional mean function; *i.e.*, we have $y_i = f^\star(x_i) + w_i$ where $f^\star(x) = \mathbf{E}[Y \mid X = x]$, analogously to Example 2. When $f^\star$ is appropriately smooth and the covariates are drawn from a uniform distribution over some compact domain, this problem has been intensively studied, and the minimax risks are well-understood. However, when the sampling of the covariates x_i is non-uniform, the possible rates of estimation can deteriorate drastically—see for instance the papers [22, 23, 24, 25, 31, 2].

Using tools from the theory of reproducing kernel Hilbert spaces (RKHSs), one can formulate this problem as an infinite-dimensional counterpart to our model (1), where the constraint parameters (ϱ, K_c) are determined by the Hilbert radius and the eigenvalues of the integral operator associated with the kernel. Although formally our minimax risk is defined for finite dimensional problems, via limiting arguments, it is straightforward to obtain consequences for the infinite-dimensional problem of the type discussed here, which discuss in Section 3.2.

EXAMPLE 7 (Covariate shift in nonparametric regression). Combining the scenarios in Examples 5 and 6, now consider the problem of covariate shift in a nonparametric setting. We observe samples (x_i, y_i) where the covariates have been drawn according to some law P, and our goal is to construct a predictor with low risk in the squared norm defined by some other covariate law Q.

In our study of this setting, the constraint set is determined by the underlying function class in a manner analogous to Example 6, and the error metric is determined by the new distribution of covariates on which the estimates must be deployed, analogously to Example 5. Some recent work has studied general conditions on the pair (P, Q) and the corresponding optimal rates of estimation [40, 26, 53, 45, 56, 63, 57, 27]. Among the consequences of our work are more refined results that are instance-dependent, in the sense that we characterize optimality for fixed pairs (P, Q), as opposed to optimality over broad classes of (P, Q) pairs. See Section 3.2.3 for a detailed discussion of these refined results.

The examples above share the common feature of being problems where estimating a conditional mean function is able to be formulated within the observation model (1). Additionally, in these examples, the fundamental hardness of the problem depends on both the structure of this function (modelled via assumptions on $\theta^\star$) as well as the distribution of the covariates. The goal of this paper is to build a general theory for these types of observation models, which elucidates how both the structure of $\theta^\star$ as well as the covariate law $\mathbb{P}$ determine the minimax rate of estimation in finite samples. In Section 3, we give concrete consequences of our general results for these types of problems.

1.3. *Relation to prior work.* Let us discuss in more detail some connections and relations between our problem formulation and results, and various branches of the statistics literature.

Connections to random design regression. As shown by the examples discussed so far, our general set-up includes, among other problems, many variants of *random design regression*. This is a classical problem in statistics, with a large literature; see the sources [32, 61, 33] and references therein for an overview. The recent paper [51] also studies the analogous problem studied here when the vector $\theta^\star$ is allowed to be arbitrary; the only assumption made is that $\theta^\star \in \mathbf{R}^d$. In this case, it is possible to use tools from Bayesian decision theory to exhibit the minimax optimality of the ordinary least squares (OLS) estimator [51, Theorem 1]. In Section 3.1.3, we demonstrate how to obtain this result as a corollary of our more general results.

Note that in applications, such as those given by the preceding examples, it is important that there is a constraint on $\theta^\star$. For instance, in a nonparametric regression problem, the parameter $\theta^\star$ denotes the coefficients of a series expansion corresponding to a conditional mean function $f^\star(x) = \mathbf{E}[Y \mid X = x]$ in an appropriate orthonormal family of functions. In this case, constraints are in fact *necessary*: to have consistent estimation, compactness is essential—see the monograph [36, Theorem 5.7] for further details.

Finally, we also comment on the similarity of our results to the paper [37]. Specifically, our main results can be compared to their Theorem 2.1. There are a few differences: first, in the paper [37], they study "fixed design" problems, whereas our formulation allows us to simultaneously treat both random and fixed design problems with the same analysis tools. Secondly, even restricting to the fixed design setup, our results are stronger than theirs, in the case of an ellispoidal constraint set. Their Theorem 2.1 shows that linear estimates only achieve the minimax rate within ellipsoid-dependent logarithmic factors; our result, on the other hand, demonstrates that linear estimates are order-optimal with factors which are *universal*—they depend on neither the dimension nor the ellipsoid under consideration. In fact, to the best of our knowledge, our result—even specialized to fixed design—is the first to treat observation operators and constraint sets given by matrices that do not commute. Previous results requirde stronger assumptions to attain (near) rate-optimality.

Random design and Bayesian priors. When the the norm of the vector $\theta^\star$ is constrained, there are relatively few minimax results in the random design setting. On the other hand, a related Bayesian setting has been studied. In this line of work, the definition of the minimax risk is altered so that the "worst-case" supremum over $\theta^\star$ in the constraint set is replaced with a suitable "average"—namely the expectation over $\theta^\star$ drawn according to a prior distribution over the constraint set.

In addition to the clear differences in the formulation, this line of work exhibits two main qualitative differences from our paper. First, these Bayesian results have primarily been established in the proportional asymptotics framework, in the ratio d/n is assumed to converge towards some aspect ratio $\gamma > 0$ as both (d, n) diverge to infinity. Secondly, by selecting "nice priors", it is possible to leverage certain properties—for instance, equivariance to some group action—that can hold for *both* the prior and covariate law. On the other hand, our setting is somewhat more challenging in that we make no *a priori* assumptions about the covariate law and its relationship to the constraint set.

In more detail, when the covariates are drawn from a multivariate Gaussian, for certain constraint sets, it is possible to find a prior such that the minimax and Bayesian risks coincide. As one example, Dicker [17] studies the asymptotic minimax risk when the ratio d/n is allowed to grow, and by using equivariance arguments, he obtains asymptotically minimax procedure. Proposition 3(b) in his paper gives a prior for which the minimax and Bayesian risks coincide. The thesis [50, Corollary 8.2] provides a matching asymptotic lower bound.

The relation between Bayes and minimax risks in this line of work cannot be expected in general, as the arguments repose critically on the rotation invariance of the standard multivariate Gaussian. Moreover, this and other classical work on random design regression using Gaussian covariates typically hinges on special, closed-form formulae for quantities related to the distribution of the sample covariance matrix (see, e.g., the papers [58, 12, 1]).

Fixed design results. Although we focus on minimax estimation of the unknown parameter $\theta^\star$ in the random design setting, we note that the related fixed design setting is well studied. In fact, in classical work, Donoho studied a very similar operator-based observation model to the one considered here; a key difference is that in that work, the focus is on estimating a (scalar-valued) functional of $\theta^\star$ [18].

By sufficiency arguments, our problem, when instantiated in the setting of fixed design with Gaussian noise, is equivalent to mean estimation on an elliptical parameter set. It is therefore related to classical work on sharp asymptotic minimax estimation in the Gaussian sequence model [54, 30, 20, 19, 5, 28, 29]; see also the monograph [36] for a pedagogical overview of this topic. These works extend the classical line of work on estimating a constrained (possibly multivariate) Gaussian mean [15, 9, 48, 7, 46]. We refer the reader to references [47, 21], which contain a more thorough overview of prior work on minimax estimation of a parameter when a notion of 'signal to noise ratio' is fixed. Of course, applying an optimal fixed design estimator cannot be expected to yield an optimal random design estimator in general. This is because in the fixed design formulation, the worst-case $\theta^\star$ could adapt to a single design matrix, whereas in the random design formulation, the worst-case $\theta^\star$ must adapt to the *random ensemble* of design matrices induced by sampling n samples in an IID fashion from a fixed covariate law.

2. Main results.

We now turn to the presentation of our main results, which are upper and lower bounds on the minimax rate of estimation as defined in display (3), matching up to a constant pre-factor. These bounds are presented in Section 2.1.

2.1. *General upper and lower bounds.*

Our general upper bounds are stated as the following functional of the distribution of the operator T_ξ; the noise covariance Σ_w; the constraint norm, as determined by the pair (ϱ, K_c); and the estimation norm, as defined by the operator K_e,

$$\begin{aligned}(5)\quad &\Phi(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c)\\ &\coloneqq \sup_{\Omega} \Big\{ \mathbf{E}\,\mathbf{Tr}\left(K_e^{1/2}(\Omega^{-1} + T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi)^{-1}K_e^{1/2}\right) : \Omega \succ 0,\ \mathbf{Tr}(K_c{}^{-1/2}\Omega K_c{}^{-1/2}) \leqslant \varrho^2 \Big\}.\end{aligned}$$

Our first main result is a general upper bound.

THEOREM 1 (General minimax upper bound). *The minimax risk is upper bounded as*

$$\mathfrak{M}(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c) \leqslant \Phi(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c). \tag{6}$$

See Section 4.1 for the proof.

Our second result is a complementary lower bound.

THEOREM 2 (Lower bound). *The minimax risk is lower bounded as*

$$\mathfrak{M}(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c) \geqslant \Phi(T, \mathbb{P}, \Sigma_w, \tfrac{\varrho}{2}, K_e, K_c) \geqslant \frac{1}{4}\,\Phi(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c). \tag{7}$$

See Section 4.2 for the proof.

Note that the functional on the righthand side of the display (7) above matches the quantity appearing in our minimax upper bound (6). Thus, in a nonasymptotic fashion, we have determined the minimax risk for this problem up the prefactor $1/4$.

Sharper lower bound constants. The constant appearing in the lower bound (7) can typically be substantially sharpened. To describe how this can be done via our results, fix a scalar $\tau \in (0, 1]$ and a symmetric positive definite matrix Ω, and let $Z \in \mathbf{R}^d$ be vector of IID standard Gaussians. Define the scalar

$$c := \tau^2 (1 - \mathbf{P}\{\tau^2 \sum_{i=1}^{d} \lambda_i Z_i^2 > 1\}),$$

where $\{\lambda_i\}_{i=1}^d$ are the the eigenvalues of the matrix $(1/\varrho^2) {K_e}^{1/2} \Omega {K_e}^{1/2}$. Then, we are able to establish the following minimax lower bound,

$$\mathfrak{M}(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c) \geqslant \mathbf{E}\, \mathbf{Tr}\Big(K_e^{1/2} (\frac{1}{c}\Omega^{-1} + T_\xi^\mathsf{T} \Sigma_w^{-1} T_\xi)^{-1} K_e^{1/2}\Big), \tag{8}$$

provided that the parameter $\tau \in (0, 1]$ and the symmetric positive definite matrix Ω is such that $\mathbf{Tr}({K_c}^{-1/2} \Omega {K_c}^{-1/2}) = \varrho^2$.

With appropriate choices of the pair (τ, Ω), the lower bound (8) can lead to pre-factors that are much closer to 1, and in some cases, converge to one under various scalings. In Section 3.1.1, we give one illustration of how the family of bounds (8) can be exploited to obtain an improvement of this type.

Form of an optimal procedure. Inspecting the proof of Theorem 1—specifically, as a consequence of Proposition 3—if the supremum on the righthand side of (5) is attained at the matrix $\Omega_\star$, then the following estimator, in view of the lower bound (7), is near minimax-optimal,

$$\widehat{\theta}(T_\xi, y) := \big(\Omega_\star^{-1} + T_\xi^\mathsf{T} \Sigma_w^{-1} T_\xi\big)^{-1} T_\xi^\mathsf{T} \Sigma_w^{-1} y. \tag{9}$$

It is perhaps instructive to write this estimator in its "ridge" formulation

$$\widehat{\theta}(T_\xi, y) = \operatorname*{arg\,min}_{\vartheta \in \mathbf{R}^d} \Big\{ \|y - T_\xi \vartheta\|_{\Sigma_w^{-1}}^2 + \|\vartheta\|_{\Omega_\star^{-1}}^2 \Big\}.$$

In the language of Bayesian statistics, our order-optimal procedure is a maximum *a posteriori* (MAP) estimate for $\theta^\star$ when $y \sim \mathsf{N}(T_\xi \theta^\star, \Sigma_w)$ and the parameter follows the prior distribution $\theta^\star \sim \mathsf{N}(0, \Omega_\star)$. The optimal prior is identified via the choice of $\Omega_\star$ which is determined by the functional appearing in Theorems 1 and 2. If the supremum in (5) is not attained, then by selecting a sequence of matrices Ω_k that approach the maximal value of the functional, one can similarly argue there exists a sequence of estimators that approach the order-optimal minimax risk.

2.2. *Independent and identically distributed regression models.* An important application of our general result is for independent and identically distributed (IID) regression models of the form

$$y_i = \langle \theta^\star, \psi(x_i) \rangle + \sigma z_i, \quad \text{for } i = 1, \ldots, n. \tag{10}$$

Above, we assume that x_i are independent and identical draws from a fixed covariate distribution P, on some measurable space $\mathcal{X}$, and that $\psi \colon \mathcal{X} \to \mathbf{R}^d$. The covariates $\{x_i\}_{i=1}^n$ are independent and the conditional distribution of $z \mid x$ is an element of $\mathcal{P}(I_n)$. The parameter $\sigma > 0$ indicates the noise level; it is an upper bound on the conditional standard deviation of $y_i - \langle \theta^\star, \psi(x_i) \rangle$.

For the model described above, the following minimax risk of estimation provides the best achievable performance of any estimator, when $\theta^\star$ lies in a compact ellipse and the error is measured in the quadratic norm

$$\mathfrak{M}_n^{\mathrm{IID}}\Big(\psi, P, \varrho, \sigma^2, K_c, K_e\Big) \coloneqq \inf_{\widehat{\theta}} \sup_{\substack{\theta^\star \in \Theta(\varrho, K_c) \\ \nu \in \mathcal{P}(\sigma^2 I_n)}} \mathbf{E}\Big[\big\|\widehat{\theta}(y_1^n, x_1^n) - \theta^\star\big\|_{K_e}^2\Big]. \tag{11}$$

Note that this problem can be formulated as an instance of our general operator formulation (1) where we take $y = (y_1, \ldots, y_n)$, $w = \sigma(z_1, \ldots, z_n)$, and $\xi = (x_1, \ldots, x_n)$, so that $\mathbb{P} = P^n$. The operator T_ξ is given by the $n \times d$-matrix with rows $\psi(x_i)^{\mathsf{T}}$. In this context the following random matrix, which is a rescaling of the operator $T_\xi^{\mathsf{T}} T_\xi$, plays an important role:

$$\Sigma_n \coloneqq \frac{1}{n} \sum_{i=1}^n \psi(x_i) \otimes \psi(x_i). \tag{12}$$

In order to state the consequence of our more general results for this problem, let us introduce a functional. We denote it by d_n to indicate that it is essentially an "effective statistical dimension" for this problem,

$$d_n(\psi, P, \varrho, \sigma^2, K_e, K_c) \coloneqq \sup\nolimits_\Omega \Big\{ \mathbf{Tr}\, \mathbf{E}_{P^n}\big[K_e^{1/2}(\Sigma_n + \Omega^{-1})^{-1} K_e^{1/2}\big] : \Omega \succ 0, \mathbf{Tr}(K_c^{-1/2} \Omega K_c^{-1/2}) \leqslant \tfrac{n\varrho^2}{\sigma^2} \Big\}\,. \tag{13}$$

Then an immediate corollary to Theorems 1 and 2 is the following pair of inequalities for the IID minimax risk.[1]

COROLLARY 1. *Under the IID regression model* (10)*, the minimax rate of estimation as defined in equation* (11) *satisfies the following inequalities,*

$$\begin{aligned} \frac{1}{4} \frac{\sigma^2}{n} d_n(\psi, P, \varrho, \sigma^2, K_e, K_c) &\leqslant \frac{\sigma^2}{n} d_n(\psi, P, \tfrac{\varrho}{2}, \sigma^2, K_e, K_c) \\ &\leqslant \mathfrak{M}_n^{\mathrm{IID}}\Big(\psi, P, \varrho, \sigma^2, K_e, K_c\Big) \leqslant \frac{\sigma^2}{n} d_n(\psi, P, \varrho, \sigma^2, K_e, K_c). \end{aligned} \tag{14}$$

So as to lighten notation, in the sequel, when the feature map ψ is the identity mapping $\psi(x) = x$, we drop the parameter ψ from the functional d_n and the minimax rate $\mathfrak{M}_n^{\mathrm{IID}}$.

2.3. *Some properties of the functional appearing in Theorems 1 and 2.*

As indicated by Theorem 1 and the subsequent discussion, the extremal quantity

$$\sup_\Omega \Big\{ \mathbf{E}\, \mathbf{Tr}\, \Big(K_e^{1/2}(\Omega^{-1} + T_\xi^{\mathsf{T}} \Sigma_w^{-1} T_\xi)^{-1} K_e^{1/2}\Big) : \Omega \succ 0,\ \mathbf{Tr}(K_c^{-1/2} \Omega K_c^{-1/2}) \leqslant \varrho^2 \Big\} \tag{15}$$

is fundamental in that it determines our minimax risk; moreover when the supremum is attained, the maximizer defines an order-optimal estimation procedure (see equation (9)). Conveniently, it turns out that the maximization problem implied by the display (15) is concave.

[1] Strictly speaking, this result follows immediately if we had defined the minimax risk over estimators which are measurable functions of the variables $\{(y_i, \psi(x_i))\}$. Nonetheless, since our lower bounds use Gaussian noise, the stated inequalities hold even when defining the minimax risk for estimators which operate on $\{(y_i, x_i)\}$, by a standard sufficiency argument.

PROPOSITION 1 (Concavity of functional). *The optimization problem*

$$\begin{aligned} &\textit{maximize} \quad f(\Omega) \coloneqq \mathbf{Tr}\,\mathbf{E}\left[{K_e}^{1/2}\left(\Omega^{-1} + T_\xi^{\mathsf{T}} \Sigma_w^{-1} T_\xi\right)^{-1} {K_e}^{1/2}\right] \\ &\textit{subject to} \quad \Omega \succ 0, \quad \mathbf{Tr}({K_c}^{-1/2} \Omega {K_c}^{-1/2}) \leqslant \varrho^2, \end{aligned} \tag{16}$$

is equivalent to a convex program, with variable Ω. Formally, the constraint set above is convex, and function f is concave over this set.

See Appendix A.1 in the supplementary material for the proof.

Note that this claim implies that, provided oracle access to the objective function f appearing above, one can in principle obtain a maximizer in a computationally tractable manner, by leveraging algorithms for convex optimization [11].

The functional (15) depends on the distribution of $T_\xi^{\mathsf{T}} \Sigma_w^{-1} T_\xi$. In general, Jensen's inequality along with the convexity of the trace of the inverse of positive matrices [8, Exercise 1.5.1] implies that it is always lower bounded by

$$\sup_\Omega \left\{ \mathbf{Tr}\left(K_e^{1/2} (\Omega^{-1} + \mathbf{E}\, T_\xi^{\mathsf{T}} \Sigma_w^{-1} T_\xi)^{-1} K_e^{1/2} \right) : \Omega \succ 0,\ \mathbf{Tr}({K_c}^{-1/2} \Omega {K_c}^{-1/2}) \leqslant \varrho^2 \right\} \tag{17}$$

Comparing displays (15) and (17), we have simply moved the expectation over ξ into the inverse. For certain IID regression models, as described in Section 2.2, we can give a complementary upper bound. To state our result, we define

$$\overline{d}_n(P, \varrho, \sigma^2, K_e, K_c) \coloneqq \sup\nolimits_\Omega \Big\{ \mathbf{Tr}\left(K_e^{1/2} (\mathbf{E}_{P^n} \Sigma_n + \Omega^{-1})^{-1} K_e^{1/2} \right) : \Omega \succ 0, \mathbf{Tr}({K_c}^{-1/2} \Omega {K_c}^{-1/2}) \leqslant \tfrac{n \varrho^2}{\sigma^2} \Big\}\ . \tag{18}$$

Note that this quantity only depends on the distribution P^n through the matrix $\mathbf{E}_{P^n} \Sigma_n$.

PROPOSITION 2 (Comparison of d_n to $\overline{d}_n$). *Define κ to be the P-essential supremum of $x \mapsto \|{K_c}^{1/2} \psi(x)\|_2$. If $\kappa < \infty$, then for any $\varrho > 0, \sigma > 0$, we have*

$$\overline{d}_n(\psi, P, \varrho, \sigma^2, K_e, K_c) \leqslant d_n(\psi, P, \varrho, \sigma^2, K_e, K_c) \leqslant \Big(1 + \frac{\varrho^2 \kappa^2}{\sigma^2}\Big) \overline{d}_n(\psi, P, \varrho, \sigma^2, K_e, K_c).$$

Unpacking this result, when $K_c^{1/2} \psi(x)$ is essentially bounded, we see that the functionals $\overline{d}_n$ and d_n are of the same order when the signal-to-noise ratio satisfies the relation $\frac{\varrho^2}{\sigma^2} \lesssim \frac{1}{\kappa^2}$. As mentioned above, the first inequality is a consequence of a generic lower bound, while the upper bound is a consequence of a new operator inequality for random positive definite matrices, presented as Theorem 3 in Appendix A.2 in the supplementary material.

2.4. *Asymptotics for a diverging radius.* In this section, we develop an asymptotic limit relation for the minimax risk (3) as the radius ϱ of the constraint set $\Theta(\varrho, K_c)$ tends to infinity. The relation reveals that the lower bound constant $1/4$ appearing in the lower bound Theorem 2 can actually be made quite close to 1 for large radii.

COROLLARY 2. *Suppose that $T_\xi^{\mathsf{T}} \Sigma_w^{-1} T_\xi$ is $\mathbb{P}$-almost surely nonsingular. Then the minimax risk* (3) *satisfies*

$$\mathfrak{M}(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c) = \big(1 - o(1)\big)\, \Phi(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c), \quad \textit{as } \varrho \to \infty.$$

See Appendix A.3 in the supplement for a proof of this claim.

An immediate consequence is that for IID regression settings as in Section 2.2, we have the following limit relation.

COROLLARY 3. *Suppose that that the empirical covariance matrix Σ_n from equation* (12) *is P^n-almost surely invertible. Then, the minimax risk for an IID observation model* (10) *satisfies the relation*

$$\mathfrak{M}_n^{\mathrm{IID}}\Big(\psi, P, \varrho, \sigma^2, K_e, K_c\Big) = \big(1 - o(1)\big)\, \frac{\sigma^2}{n} d_n\big(\psi, P, \varrho, \sigma^2, K_e, K_c\big), \quad \text{as } \varrho \to \infty.$$

3. Consequences of main results. In this section, we demonstrate consequences of our main results for a variety of estimation problems. In Section 3.1, we develop consequences of our main results for problems where the underlying parameter to be estimated is finite-dimensional. In Section 3.2, we develop consequences of our main results for problems where the underlying parameter is infinite-dimensional. In both cases, we are able to derive minimax rates of estimation, which to the best of our knowledge, are not yet in the literature. Additionally, we are also able to re-derive classical as well as recent results in a unified fashion via our main theorems.

3.1. *Applications to parametric models.* We begin by developing the consequences of our main results for regression problems where the statistician is aiming to estimate a finite-dimensional parameter. Sections 3.1.1, 3.1.2, and 3.1.3 concern IID regression settings of the form described in Section 2.2. In Section 3.1.4, we consider a non-IID regression setting.

3.1.1. *Linear regression with Gaussian covariates.* As in the prior work [17], consider a random design IID regression setting of the form presented in the display (10), but with Gaussian data. Formally, we assume Gaussian noise, so that $z_i \overset{\mathrm{IID}}{\sim} \mathsf{N}(0, 1)$, and Gaussian covariates, so that $x_i \overset{\mathrm{IID}}{\sim} \mathsf{N}(0, I_d)$ and $\psi(x) = x$. Here x and z are assumed independent. Then we define

$$r(n, d, \varrho, \sigma) := \inf_{\widehat{\theta}} \sup_{\|\theta\|_2 \leqslant \varrho} \mathbf{E}\Big[\|\widehat{\theta} - \theta\|_2^2\Big], \quad \text{and} \quad d_{\mathrm{Dicker}}(n, d, \varrho, \sigma) := \mathbf{Tr}\, \mathbf{E}\Big[(\Sigma_n + \tfrac{\sigma^2}{n}\tfrac{d}{\varrho^2} I_d)^{-1}\Big],$$

where the expectations are over the Gaussian covariates and noise pairs $\{(x_i, z_i)\}_{i=1}^n$. These quantities correspond, respectively, to the minimax risk and the worst-case risk (rescaled by n/σ^2), of a certain ridge estimator [17, Corollary 1] on the sphere $\{\|\theta\|_2 = \varrho\}$.

Dicker [17, Corollary 3] proves the following limiting result. Under the proportional asymptotics $d/n \to \gamma$, where the limiting ratio γ lies in $(0, \infty)$, the minimax risk satisfies

$$\lim_{d/n \to \gamma} \Big| r(n, d, \varrho, \sigma) - \frac{\sigma^2}{n} d_{\mathrm{Dicker}}(n, d, \varrho, \sigma) \Big| = 0, \tag{19}$$

for any radius $\varrho > 0$ and noise level $\sigma > 0$.

Let us now demonstrate that our general theory yields a nonasymptotic counterpart of this claim, and taking limits recovers the asymptotic relation (19).

COROLLARY 4. *For linear regression over the ϱ-radius Euclidean sphere with Gaussian covariates, the minimax risk satisfies the sandwich relation*

(20a)

$$c_d \frac{\sigma^2}{n} d_{\mathrm{Dicker}}(n, d, \varrho, \sigma) \leqslant \frac{\sigma^2}{n} d_{\mathrm{Dicker}}(n, d, \sqrt{c_d}\varrho, \sigma) \leqslant r(n, d, \varrho, \sigma) \leqslant \frac{\sigma^2}{n} d_{\mathrm{Dicker}}(n, d, \varrho, \sigma),$$

where

$$c_d := \begin{cases} (1 - \frac{1}{2d-1})(1 - \exp(-\frac{d^{3/2}}{4})) & d \geqslant 2 \\ 1/4 & d = 1 \end{cases}. \tag{20b}$$

Note that since $c_d = (1 - O(1/d))$ as $d \to \infty$, the inequalities (20a) allow us to immediately recover Dicker's result. It should be emphasized, however, that Corollary 4, holds for *any* quadruple (n, d, ϱ, σ). In particular, it is valid in a completely nonasymptotic fashion and with explicit constants.

We now sketch how this result follows from our main results. As calculated in Appendix B.1.1 in the supplement, our functional for this problem satisfies

$$d_n(\mathsf{N}\left(0, I_d\right), \varrho, \sigma^2, I_d, I_d) = d_{\text{Dicker}}(n, d, \varrho, \sigma). \tag{21a}$$

Hence, our Corollary 1 implies the following characterization of the minimax risk,[2]

$$\frac{1}{4}\frac{\sigma^2}{n} d_{\text{Dicker}}(n, d, \varrho, \sigma) \leqslant r(n, d, \varrho, \sigma) \leqslant \frac{\sigma^2}{n} d_{\text{Dicker}}(n, d, \varrho, \sigma^2). \tag{21b}$$

To establish our sharper result (20a), we leverage the stronger lower bound (8). The details of this calculation are presented in Appendix B.1.2 in the supplementary material. Note that in Section 5.1.1, we simulate this problem and find that as suggested by Corollary 4, that, indeed, the gap between our upper and lower bounds is tiny, even for problems with small dimension (see Figure 1).

3.1.2. *Underdetermined linear regression.* Consider observing samples from a standard linear regression model; that is, we observe pairs $\{(x_i, y_i)\}$ according to the model (10), with $\psi(x) = x$. A practical scenario in which some assumption regarding the norm of the underlying parameter is necessary is when the sample covariance matrix Σ_n, defined in display (12) is singular with positive P^n-probability. This occurs if $n < d$, or if there is a hyperplane $H \subset \mathbf{R}^d$ such that $x \sim P$ lies in H with positive probability.

In this setting, the correct dependence of the minimax risk on the geometry of the constraint set and the distribution of sample covariance matrix is relatively poorly understood. For simplicity—although our results are more general than this—let us assume that error is measured in the Euclidean norm and that it is assumed that the underlying parameter $\theta^\star$ has Euclidean norm bounded by $\varrho > 0$, and that the noise is independent Gaussian with variance σ^2. Then Corollary 1 demonstrates that

$$\inf_{\widehat{\theta}} \sup_{\|\theta\|_2 \leqslant \varrho} \mathbf{E}[\|\widehat{\theta} - \theta\|_2^2] \asymp \frac{\sigma^2}{n} d_n(P, \varrho, \sigma^2, I_d, I_d) = \frac{\sigma^2}{n} \sup_{\Omega \succ 0} \Big\{ \mathbf{Tr}\, \mathbf{E}_{P^n}\big[(\Sigma_n + \Omega^{-1})^{-1}\big] : \mathbf{Tr}(\Omega) \leqslant \frac{n\varrho^2}{\sigma^2} \Big\}.$$

Taking $\Omega = \frac{n}{d}\frac{\varrho^2}{\sigma^2} I_d$, we obtain the following lower bound on the minimax risk for any covariate law P,

$$\frac{\sigma^2}{n} \mathbf{Tr}\, \mathbf{E}_{P^n}\big[(\Sigma_n + \tfrac{\sigma^2}{\varrho^2}\tfrac{d}{n} I_d)^{-1}\big] \asymp \underbrace{\mathbf{E}\Big[\sum_{i=1}^d \tfrac{\sigma^2}{n}\tfrac{1}{\lambda_i(\Sigma_n)} \mathbf{1}\{\lambda_i(\Sigma_n) \geqslant \tfrac{\sigma^2}{n}\tfrac{d}{\varrho^2}\}\Big]}_{\substack{\text{Estimation error from} \\ \text{large eigenvalues of } \Sigma_n}} + \underbrace{\mathbf{E}\Big[\sum_{i=1}^d \tfrac{\varrho^2}{d} \mathbf{1}\{\lambda_i(\Sigma_n) < \tfrac{\sigma^2}{n}\tfrac{d}{\varrho^2}\}\Big]}_{\substack{\text{Approximation error due} \\ \text{to small eigenvalues of } \Sigma_n}}. \tag{22}$$

The lower bound (22) is sharp in certain cases. For instance, when $x_i \overset{\text{IID}}{\sim} \mathsf{N}\left(0, I_d\right)$ but there are fewer samples than the dimension, so that $n < d$, it is equal to the minimax risk up to universal constants, following the same argument as in Section 3.1.1.

Note that above, λ_i denotes the ith largest (nonnegative) eigenvalue of a symmetric positive semidefinite matrix. One possible interpretation of this lower bound is as follows: the

[2] Although Corollary 1 takes the supremum over a larger family of noise distributions, note that our lower bounds are obtained with Gaussian noise, so that the result applies even if we restrict to Gaussian noise.

first term indicates the estimation error incurred in directions where the effective signal-to-noise ratio is high; on the other hand, the second term indicates the bias or approximation error that must be incurred in directions where the effective signal-to-noise ratio is low. In fact, the message of this lower bound is that in these directions, no procedure can do much better than estimating 0 there. One concrete and interesting takeaway is that if Σ_n has an eigenvalue equal to zero, it increases the minimax risk by essentially the same amount as if the eigenvalue were positive and in the interval $(0, \frac{\sigma^2}{n}\frac{d}{\varrho^2})$.

3.1.3. *Linear regression with an unrestricted parameter space.* In recent work, Mourtada [51] characterizes the minimax risk for random design linear regression problem for an *unrestricted* parameter space. Consider observing samples $\{(x_i, y_i)\}_{i=1}^n$ following the IID model (10) with $\psi(x) = x$, where the covariates are drawn from some distribution P on $\mathbf{R}^d$. As argued by Mourtada (see his Proposition 1), or as can be seen by taking $\varrho \to \infty$ in our singular lower bound (22) from Section 3.1.2, if we impose no constraint on the underlying parameter $\theta^\star$, then it is necessary to assume that the sample covariance matrix Σ_n is invertible with probability 1 in order to obtain finite minimax risks. Theorem 1 in Mourtada's paper then asserts that under this condition, we have

$$\inf_{\widehat{\theta}} \sup_{\substack{\theta^\star \in \mathbf{R}^d \\ \nu \in \mathcal{P}(\sigma^2 I_n)}} \mathbf{E}\Big[\|\widehat{\theta} - \theta^\star\|_{\Sigma_P}^2\Big] = \frac{\sigma^2}{n}\,\mathbf{E}\big[\,\mathbf{Tr}(\Sigma_n^{-1}\Sigma_P)\big], \tag{23}$$

where the expectation is over the data $\{(x_i, y_i)\}_{i=1}^n$, and $\Sigma_P \coloneqq \mathbf{E}_P[x \otimes x]$ is the population covariance matrix under P.

We now show that this result, with the exact constants, is a consequence of our more general results. We focus on establishing the lower bound, because it is well-known (and easy to show) that the upper bound is achieved by the ordinary least squares estimator.[3] Thus for the lower bound, our results imply that

$$\inf_{\widehat{\theta}} \sup_{\substack{\theta^\star \in \mathbf{R}^d \\ \nu \in \mathcal{P}(\sigma^2 I_n)}} \mathbf{E}\Big[\|\widehat{\theta} - \theta^\star\|_{\Sigma_P}^2\Big] \geqslant \sup_{\varrho > 0} \Bigg\{ \inf_{\widehat{\theta}} \sup_{\substack{\|\theta^\star\|_2 \leqslant \varrho \\ \nu \in \mathcal{P}(\sigma^2 I_n)}} \mathbf{E}\Big[\|\widehat{\theta} - \theta^\star\|_{\Sigma_P}^2\Big]\Bigg\} \tag{24a}$$

$$= \frac{\sigma^2}{n} \lim_{\varrho \to \infty} d_n(P, \varrho, \sigma^2, \Sigma_P, I_d). \tag{24b}$$

In order to obtain the relation (24b), we have used the fact that the constrained minimax risk over the set $\{\|\theta^\star\|_2 \leqslant \varrho\}$ is nondecreasing in $\varrho > 0$, and have applied our limit relation in Corollary 3. A short calculation, which we defer to Appendix B.1.3 in the supplement, demonstrates that

$$\lim_{\varrho \to \infty} d_n(P, \varrho, \sigma^2, \Sigma_P, I_d) = \mathbf{E}\big[\,\mathbf{Tr}(\Sigma_n^{-1}\Sigma_P)\big]. \tag{25}$$

Thus, after combining displays (24b) and (25), we have obtained the lower bound in Mourtada's result (23). One consequence of this argument is that the inequality (24a) is, as may be expected, an equality. That is, we have

$$\inf_{\widehat{\theta}} \sup_{\substack{\theta^\star \in \mathbf{R}^d \\ \nu \in \mathcal{P}(\sigma^2 I_n)}} \mathbf{E}\Big[\|\widehat{\theta} - \theta^\star\|_{\Sigma_P}^2\Big] = \sup_{\varrho > 0} \Bigg\{ \inf_{\widehat{\theta}} \sup_{\substack{\|\theta^\star\|_2 \leqslant \varrho \\ \nu \in \mathcal{P}(\sigma^2 I_n)}} \mathbf{E}\Big[\|\widehat{\theta} - \theta^\star\|_{\Sigma_P}^2\Big]\Bigg\}.$$

[3] Alternatively, note that if we define $\widehat{\theta}_\varrho$ to be the order-optimal estimator we derive for the constraint set $\{\|\theta^\star\|_2^2 \leqslant \varrho^2\}$ (see equation (9), with $K_c = I_d$, $\Sigma_w = \sigma^2 I_d$, and $T_\xi = X$, where X is the design matrix.), then it converges compactly to the ordinary least squares estimate as $\varrho \to \infty$.

Note that establishing this equality directly is somewhat cumbersome, as it requires essentially applying a form of a min-max theorem, which in turn requires compactness and continuity arguments.

3.1.4. *Regression with Markovian covariates.* We consider a dataset $\{(x_t, y_t)\}_{t=1}^{T}$ comprising of covariate-response pairs. The covariates are initialized with $x_0 = 0$, and then proceed via the recursion

$$x_t = \sqrt{r_t}\, x_{t-1} + \sqrt{1 - r_t}\, z_t \quad \text{for } t = 1, \ldots, T, \tag{26}$$

for some collection of parameters $\{r_t\}_{t=1}^{T} \subset [0, 1]$, and family of independent standard Gaussian variates $\{z_t\}_{t=1}^{T}$. By construction, the samples $\{x_t\}_{t=1}^{T}$ form a Markov chain—a time-varying $\mathrm{AR}(1)$ process with stationary distribution being the standard Gaussian law. At the extreme $r_t \equiv 0$, the sequence $\{x_i\}_{i=1}^{n}$ is IID, whereas for $r_t \in (0, 1)$, is a dependent sequence, and its mixing becomes slower as the parameters $\{r_t\}$ get closer to 1. In addition to these random covariates, suppose that we also observe responses $\{y_t\}_{t=1}^{T}$ from the model

$$y_t = x_t \theta^\star + \sigma w_t, \qquad \text{for } t = 1, \ldots, T, \tag{27}$$

where $\sigma > 0$ is a noise standard deviation, and the noise sequence $\{w_t\}_{t=1}^{T}$ consists of IID standard Gaussian variates. We assume that z_t and x_t are independent for all $t = 1, \ldots, T$.

We now describe how our main results apply to this setting. Let us define a matrix $M \in \mathbf{R}^{T \times T}$ which is associated to the dynamical system (26). It has entries

$$M_{ss'} = \sum_{t = s \vee s'}^{T} \sqrt{c_{st} c_{s't}}, \quad \text{where} \quad c_{st} := (1 - r_s) \prod_{\tau = s+1}^{t} r_\tau. \tag{28}$$

To give one example, in the special case that $r_t \equiv \alpha \in (0, 1)$ for all t, then the matrix M is similar under permutation to the matrix with entries

$$M_{st} = \sqrt{\alpha}^{|s-t|} - \sqrt{\alpha}^{s+t}.$$

Evidently, this matrix is a rank-one update to the covariance matrix for the underlying $\mathrm{AR}(1)$ process (*i.e.*, the Kac–Murdock–Szegö matrix [38]); it is easily checked to be symmetric positive definite.

We now state the consequences of our main results for this problem.

COROLLARY 5. *The minimax risk for the Markovian observation model described above satisfies*

$$\inf_{\widehat{\theta}} \sup_{|\theta^\star| \leqslant \varrho} \mathbf{E}\left[(\widehat{\theta} - \theta^\star)^2\right] \asymp \Phi_T(\varrho, \sigma) := \mathbf{E}\left[\left(\frac{1}{\varrho^2} + \frac{z^{\mathsf{T}} M z}{\sigma^2}\right)^{-1}\right]. \tag{29}$$

See Appendix B.1.5 of the supplement for details of this calculation.

Note that in the result above, the expectation on the lefthand side is over the dataset $\{(x_i, y_i)\}_{i=1}^{T}$, under the Markovian model (26) for the covariates, and the expectation on the righthand size is over the Gaussian vector $z = (z_1, \ldots, z_T) \sim \mathsf{N}(0, I_T)$. Corollary 5 gives one example of how our general results can even establish sharp rates for regression problems of the form described in Section 2.2, but with additional dependence among the covariates.

Additionally, we note that with $\tau^2 = \sigma^2/\varrho^2$, we have by simple integration that

$$\Phi_T(\varrho, \sigma) = \frac{\sigma^2}{2} \int_0^\infty \exp\Big\{ -\frac{u\tau^2 + \sum_{t=1}^{T} \log(1 + u\lambda_t)}{2} \Big\}\, \mathrm{d}u,$$

where $\{\lambda_t\}_{t \in [T]}$ denote the eigenvalues of the matrix M.

3.2. *Applications to infinite-dimensional and nonparametric models.* In this section, we derive some of the consequences of our main results for infinite-dimensional models, such as those arising in nonparametric regression. The basic idea will be to identify an infinite dimensional parameter space Θ, typically lying in the Hilbert space $\ell^2(\mathbf{N})$. We then find a nested sequence of subsets

$$\Theta_1 \subset \Theta_2 \subset \cdots \subset \Theta_k \subset \cdots \subset \Theta,$$

where Θ_k are finite-dimensional truncations of Θ. Under regularity conditions, we can show that the minimax risk for the k-dimensional problems converge to the minimax risk for the infinite dimensional problem as $k \to \infty$. Thus, since we have determined the minimax risk for each subset Θ_k up to universal constants (importantly, constants independent of the underlying dimension), we take the limit of our functional in the limit $k \to \infty$ to obtain a tight characterization of the minimax risk for the infinite-dimensional set Θ.

In the next few sections, we carry this program out in a few examples. We begin with a study of the canonical Gaussian sequence model in Section 3.2.1. We then turn, in Sections 3.2.2 and 3.2.3, to nonparametric regression models arising from reproducing kernel Hilbert spaces. In this setting, we are able to derive some classical results for Sobolev spaces, derive new and sharper forms of bounds on nonparametric regression with covariate shift, and obtain new results for random design nonparametric models with non-uniform covariate laws.

3.2.1. *Gaussian sequence model.* In the canonical Gaussian sequence model, we make a countably infinite sequence of observations of the form

$$y_i = \theta^\star_i + \varepsilon_i z_i, \qquad \text{for } i = 1, 2, 3, \ldots \tag{30}$$

Here the variables $\{z_i\}$ are a sequence of IID standard Gaussian variates, and $\varepsilon := \{\varepsilon_i\}$ indicate the noise level (*i.e.*, the standard deviation) of the entries of the observation y. It is typically assumed that there is a nondecreasing sequence of divergent, nonnegative numbers $a := \{a_i\}$ and radius $C > 0$ such that

$$\theta^\star \in \Theta(a, C) := \Big\{ \theta \in \mathbf{R}^{\mathbf{N}} : \sum_{j \geqslant 1} a_j^2 \theta_j^2 \leqslant C^2 \Big\}.$$

The minimax risk for this problem is then defined by

$$\mathfrak{M}\Big(\varepsilon, a, C\Big) := \inf_{\widehat{\theta}} \sup_{\theta^\star \in \Theta(a,C)} \mathbf{E}\Big[\sum_{j=1}^{\infty} (\widehat{\theta}_j(y) - \theta^\star_j)^2 \Big],$$

where the expectation is over y according to the observation model (30).

Let us define a k-dimensional truncation,

$$\Theta_k(a, C) := \Big\{ \theta \in \Theta(a, C) : \theta_j = 0, \text{ for all } j > k \Big\}.$$

Evidently $\Theta_k(a, C)$ may be regarded as a subset of $\mathbf{R}^k$. Note that the class $\{\Theta_k(a, C)\}_{k \geqslant 1}$ forms a nested sequence of subsets within Θ. Moreover, we can define the minimax risk for the k-dimensional problem

$$\mathfrak{M}_k\Big(\varepsilon, a, C\Big) := \inf_{\widehat{\theta}} \sup_{\theta^\star \in \Theta_k(a,C)} \mathbf{E}\Big[\sum_{j=1}^{k} (\widehat{\theta}_j(y) - \theta^\star_j)^2 \Big].$$

Slightly abusing notation, above we regard $y, \theta^\star \in \mathbf{R}^k$, where y is distributed as the first k components of the observation model (30). Then, this sequence of minimax risks satisfies the limit relation

$$\lim_{k \to \infty} \mathfrak{M}_k\Big(\varepsilon, a, C\Big) = \mathfrak{M}\Big(\{\varepsilon_j\}_{j=1}^\infty, \Theta(a, C)\Big). \tag{31}$$

See Appendix B.2.1 for a proof of this relation. The k-dimensional problem can be seen as a special case of our operator model (1), with parameters $T^{(k)}, \Sigma_w^{(k)}, K_e{}^{(k)}, \varrho^{(k)}, K_c{}^{(k)}$ defined as,

$$\begin{gathered} T^{(k)}(\xi) \equiv I_k, \qquad \Sigma_w^{(k)} = \mathbf{diag}(\varepsilon_1^2, \ldots, \varepsilon_k^2), \qquad K_e^{(k)} = I_k, \\ K_c^{(k)} = \mathbf{diag}\Big(\frac{1}{a_1^2}, \ldots, \frac{1}{a_k^2}\Big), \quad \text{and,} \quad \varrho^{(k)} = C. \end{gathered} \tag{32}$$

Computing the functional (15) for the k-dimensional problem, we find it is equal to

$$R_k^\star\Big(\varepsilon, a, C\Big) := \sup_{\tau_1, \ldots, \tau_k} \Big\{ \sum_{j=1}^k \frac{\tau_j^2 \varepsilon_j^2}{\tau_j^2 + \varepsilon_j^2} : \sum_{j=1}^k \tau_j^2 a_j^2 \leqslant C^2 \Big\}. \tag{33}$$

Hence, define the following functional of $\varepsilon := \{\varepsilon_j\}_{j \geqslant 1}, a := \{a_j\}_{j \geqslant 1}$, and $C > 0$,

$$R^\star(\varepsilon, a, C) := \sup_{\tau = \{\tau_j\}_{j=1}^\infty} \Big\{ \sum_{j=1}^\infty \frac{\tau_j^2 \varepsilon_j^2}{\tau_j^2 + \varepsilon_j^2} : \sum_{j=1}^\infty \tau_j^2 a_j^2 \leqslant C^2 \Big\}. \tag{34}$$

Then our main results, Theorems 1 and 2 imply the sandwich relation

$$\frac{1}{4} R^\star(\varepsilon, a, C) \leqslant \mathfrak{M}\Big(\varepsilon, a, C\Big) \leqslant R^\star(\varepsilon, a, C). \tag{35}$$

See Appendix B.2.2 of the supplement for verification of this relation as a consequence of our results. Note that this recovers a well-known result for the Gaussian sequence model [61, 36]. Some previous work [20] has shown that the lower bound constant can be slightly improved to $\frac{1}{1.25}$ by arguments specific to the Gaussian sequence model. Importantly, the Gaussian sequence model is a "deterministic" operator model in the sense that the operator T_ξ has no dependence on ξ for this problem. The next few examples show some consequences of our theory for infinite-dimensional problems where the corresponding operator T_ξ is truly random.

3.2.2. *Nonparametric regression over reproducing kernel Hilbert spaces (RKHSs).* In this section, we consider a nonparametric regression model of the form

$$y_i = f^\star(x_i) + w_i, \quad \text{for } i = 1, \ldots, n. \tag{36}$$

We assume that $\{x_i\}_{i=1}^n$ are IID samples covariate law P and w_i being conditionally centered with conditional variance bounded above by σ^2. Equivalently, the noise variables are drawn from a conditional distribution satisfying the noise conditions (N1) and (N2) with $\Sigma_w = \sigma^2 I_n$.[4] We will assume that $f^\star$ lies in a reproducing kernel Hilbert space $\mathcal{H}$, and has bounded Hilbert norm $\|f^\star\|_{\mathcal{H}} \leqslant \varrho$. The goal is to estimate $f^\star$.

[4] The discussion below is unaffected by imposing additional structure on the noise, so long as the family of possible noise distributions includes $w \sim \mathsf{N}\Big(0, \sigma^2 I_n\Big)$.

Relating the RKHS observation model (36) *with the model* (10). We now show that the observation model when $f^\star \in \mathcal{H}$ is an infinite-dimensional version of the observation model (10), as can be made precise with RKHS theory. Indeed, fix a measure space $(\mathcal{X}, \mathcal{A}, \nu)$, and a measurable positive definite kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ and let $\mathcal{H}$ denote its reproducing kernel Hilbert space [3]. Under mild regularity assumptions[5], the RKHS $\mathcal{H}$ can be put into one-to-one correspondence with a mapping of $\ell^2(\mathbf{N})$. Formally, we have

$$\mathcal{H} = \Big\{ f \coloneqq \sum_{j=1}^{\infty} \theta_j \sqrt{\mu_j} \phi_j \mid \sum_{j=1}^{\infty} \theta_j^2 < \infty \Big\}. \tag{38}$$

for a nonincreasing sequence $\mu_j \to 0$ as $j \to \infty$, and for an orthonormal sequence $\{\phi_j\}$ in $L^2(\nu)$. This allows us to equivalently write the observations (36) in the form

$$y_i = \langle \theta^\star, \Phi(x_i) \rangle + w_i, \quad \text{for } i = 1, \ldots, n. \tag{39}$$

Above, we have defined the sequence $\theta^\star \coloneqq (\theta_j^\star)_{j=1}^\infty$ and "feature map" $\Phi(x) \in \ell^2(\mathbf{N})$, by the formulas

$$\theta_j^\star \coloneqq \frac{\int_{\mathcal{X}} f^\star(x) \phi_j(x) \, \mathrm{d}\nu(x)}{\sqrt{\mu_j}}, \quad \text{and} \quad \big(\Phi(x)\big)_j \coloneqq \sqrt{\mu_j} \phi_j(x), \qquad \text{for all } j \geqslant 1.$$

With these definitions, note that the inner product in equation (39) is taken in the sequence space $\ell^2(\mathbf{N})$. From the display (39), we see that the RKHS observation model (36) is in fact an infinite-dimensional version of the observation model (10). The remainder of this section is devoted to deriving consequences of our results for this model by various truncation and limiting arguments.

Truncation argument for RKHS minimax risks. Given the RKHS ball $\mathsf{B}_{\mathcal{H}}(\varrho) \coloneqq \big\{ g \in \mathcal{H} : \|g\|_{\mathcal{H}} \leqslant \varrho \big\}$, our goal is to characterize the minimax risk

$$\mathfrak{M}_n(\varrho, \sigma^2, P) \coloneqq \inf_{\hat{f}} \sup_{\substack{f^\star \in \mathsf{B}_{\mathcal{H}}(\varrho) \\ \nu \in \mathcal{P}(\sigma^2 I_n)}} \mathbf{E}\Big[\|\hat{f} - f^\star\|_{L^2(\nu)}^2 \Big]. \tag{40}$$

It should be noted here that the covariates are drawn from P and the error is measured in $L^2(\nu)$. In classical work on estimation over RKHSs, it is typical to assume that $P = \nu$. However, we develop in this section and in Section 3.2.3 some interesting consequences of our theory when $P \neq \nu$, and so this generality is important for our discussion.

To apply our results to this setting, we need to define certain finite-dimensional truncations. We start by defining

$$\mathcal{H}_k \coloneqq \Big\{ f \coloneqq \sum_{j=1}^{\infty} \theta_j \sqrt{\mu_j} \phi_j \mid \theta_j = 0, \ \text{for all } j > k \Big\}.$$

[5] The elliptical representation (38) is available in great generality. Indeed, a sufficient condition is for the map $x \mapsto \sqrt{k(x,x)}$ to lie in $L^2(\nu)$. It can be shown [59, see Lemma 2.3] that in this case, $\mathcal{H}$ compactly embeds into $L^2(\nu)$ and that there is a series expansion

$$k(x, x') = \sum_{j=1}^{\infty} \mu_j \phi_j(x) \phi_j(x'), \quad \text{for any } x, x' \in \mathcal{X}. \tag{37}$$

Here $\{\mu_j\}_{j=1}^\infty$ denotes a summable sequence of non-negative eigenvalues, whereas the sequence $\{\phi_j\}_{j=1}^\infty$ is an orthonormal family of functions $\mathcal{X} \to \mathbf{R}$ that lie in $L^2(\nu)$. Finally, the series converges absolutely, for each $x, x' \in \mathcal{X}$. Note that the infinite-dimensional series representation (38) of $\mathcal{H}$ follows from the series expansion of the underlying kernel (37); see Cucker and Smale [16] for details.

We then define the minimax risk over the the ball $\mathsf{B}_{\mathcal{H}}(\varrho)$ restricted to $\mathcal{H}_k$,

$$\mathfrak{M}_n^{(k)}(\varrho, \sigma^2, P) := \inf_{\hat{f}} \sup_{\substack{f^\star \in \mathsf{B}_{\mathcal{H}}(\varrho) \cap \mathcal{H}_k \\ \nu \in \mathcal{P}(\sigma^2 I_n)}} \mathbf{E}\Big[\|\hat{f} - f^\star\|_{L^2(\nu)}^2\Big]. \tag{41}$$

In analogy to the limit relation (31) for the Gaussian sequence model, we can show that

$$\lim_{k \to \infty} \mathfrak{M}_n^{(k)}(\varrho, \sigma^2, P) = \mathfrak{M}_n(\varrho, \sigma^2, P). \tag{42}$$

See Appendix B.2.3 of the supplement for a proof of this relation. The k-dimensional problem associated with the risk (41) can be seen, using the representation (39), as a special case of our IID observation model (10), with parameters, P, ϱ, σ and

$$\psi(x) = \Phi_k(x) := \big(\sqrt{\mu_j}\phi_j(x)\big)_{j=1}^k, \quad K_e = M_k := \mathbf{diag}(\mu_1, \ldots, \mu_k), \quad \text{and} \quad K_c = I_k. \tag{43}$$

Let us define the $k \times k$ empirical covariance matrix

$$\Sigma_n^{(k)} := \frac{1}{n} \sum_{i=1}^n \Phi_k(x_i) \otimes \Phi_k(x_i).$$

Then the using (43), we see that the functional (13) for the k-dimensional problem is equal to

$$d_n^{(k)} := \sup_{\Omega \succ 0} \Big\{ \mathbf{Tr}\, \mathbf{E}_{P^n}\big[M_k^{1/2}(\Sigma_n^{(k)} + \Omega^{-1})^{-1} M_k^{1/2}\big] : \mathbf{Tr}(\Omega) \leqslant \frac{n\varrho^2}{\sigma^2} \Big\} \tag{44}$$

Characterizations of RKHS minimax risks of estimation. We now state the consequence of our results for the rate of estimation (40).

COROLLARY 6. *Define $d_n^\star = \limsup_{k \to \infty} d_n^{(k)}$, where the sequence $\{d_n^{(k)}\}_{k \geqslant 1}$ is defined in display* (44)*. Then the RKHS minimax risk satisfies satisfies the inequalities,*

$$\frac{1}{4} \frac{\sigma^2}{n} d_n^\star \leqslant \mathfrak{M}_n(\varrho, \sigma^2, P) \leqslant \frac{\sigma^2}{n} d_n^\star. \tag{45}$$

Note that this result is an immediate consequence of Theorems 1 and 2, together with the limit relation (42).

We comment that Corollary 6 can also be written in a more appealing form. Indeed, although we do not make use of it here, we comment that there is an "extrinsic" representation of the rate description provided in this corollary. To define it, let us introduce

$$\mathbb{S}_\nu := \mathbf{E}_{x \sim \nu}[k(x, \cdot) \otimes_{\mathcal{H}} k(x, \cdot)] \quad \text{and} \quad \mathbb{S}_n := \frac{1}{n} \sum_{i=1}^n k(x_i, \cdot) \otimes_{\mathcal{H}} k(x_i, \cdot),$$

which are two positive self-adjoint operators $\mathcal{H} \to \mathcal{H}$. Then, we have

$$\mathfrak{M}_n(\varrho, \sigma^2, P) \asymp \frac{\sigma^2}{n} \sup_{\substack{\Omega \succcurlyeq 0 \\ \mathbf{Tr}_{\mathcal{H}}(\Omega) = 1}} \mathbf{Tr}_{\mathcal{H}}\, \mathbf{E}_{P^n}\Big[\mathbb{S}_\nu\, \Omega^{1/2}(\Omega^{1/2}\mathbb{S}_n\Omega^{1/2} + \tfrac{\sigma^2}{n\varrho^2} I_{\mathcal{H}})^{-1}\Omega^{1/2}\Big]. \tag{46}$$

Let us now further simplify the characterization (45) in the classical situation where the noise level dominates the Hilbert radius, we have $P = \nu$, and the map $x \mapsto k(x, x)$ is P-essentially bounded by a finite number κ under P.

COROLLARY 7. *Suppose that $P = \nu$ and $x \mapsto k(x,x)$ is P-essentially bounded by $\kappa \in (0,\infty)$. If $\sigma^2 \geqslant \kappa^2 \varrho^2$, then the RKHS minimax risk satisfies*

$$\mathfrak{M}_n(\varrho, \sigma^2, P) \asymp \frac{\sigma^2}{n} k_n, \tag{47}$$

where $k_n \equiv k_n(\sigma, \varrho) := \max\{k : \sum_{j=1}^{k} \frac{1}{\mu_j} \leqslant \frac{n\varrho^2}{\sigma^2}\}$.

See Appendix B.2.4 of the supplement for a proof of this claim.

We note that Corollaries 6 and 7 establish the nonasymptotic minimax risk of estimation for the RKHS ball of radius ρ, apart from universal constants, in a fairly general fashion. The latter claim permits easier calculation, at the expense of some slightly stronger assumptions. One advantage to Corollary 6 is that it holds for *any* configuration of the noise level and the Hilbert radius, in contrat to the prior work on the minimax rates for RKHS balls which typically requires that the signal-to-noise ratio is sufficiently small.

Interestingly, we note that our characterizations—even the loosened characterization (47)—does not need the kernel to satisfy an additional eigenvalue decay condition. Indeed, our results hold even if the kernel eigenvalues do not satisfy the requirement of a *regular kernel* as proposed in prior work [64]. To emphasize this point, we now provide one concrete example of an irregular kernel for which Corollary 7 provides, to our knowledge, a new result.

EXAMPLE 8 (Irregular kernel). Suppose that $P = \nu$ and that the kernel eigenvalues satisfy $\mu_j(\alpha) = \frac{1}{(j+1)\log^\alpha(j+1)}$ for some $\alpha > 1$. It is easily verified that the corresponding kernel eigenvalues violates the regularity condition in the paper [64], since an elementary calculation shows for J sufficiently large, we have $\frac{\sum_{j>J}\mu_j}{J\mu_J} \gtrsim \log(J)$, which diverges as $J \to \infty$. Nonetheless, our result—specifically Corollary 7—establishes the optimal rate of estimation. Assuming that $x \mapsto \sum_j \mu_j \phi_j^2(x)$ is P-almost surely less than $\kappa \in (0,\infty)$ and $\sigma^2 \geqslant \kappa^2\varrho^2$, the minimax rate for this kernel satsifies

$$\inf_{\hat{f}} \sup_{\|f^\star\|_{\mathcal{H}_\alpha} \leqslant \varrho} \mathbf{E}\,\|\hat{f} - f^\star\|_{L^2(P)}^2 \asymp R\sqrt{\frac{\sigma^2}{n\log^\alpha(n\varrho^2/\sigma^2)}}$$

where $\mathcal{H}_\alpha$ denotes an RKHS corresponding to kernel eigenvalues $\mu_j(\alpha)$. The relation above follows from a straightforward calculation which shows that the quantity k_n appearing in Corollary 7 is of the order $\sqrt{\frac{n\varrho^2}{\sigma^2\log^\alpha(n\varrho^2/\sigma^2)}}$. To our knowledge, the minimax rate for kernels having eigenvalues of this type was not previously known in the literature.

For a more classical example, we now record yet another consequence of Corollary 7.

EXAMPLE 9 (Minimax rate for nonparametric regression on a Sobolev space). Suppose that $P = \nu$ is the uniform distribution on $[0,1]^d$ and $\mathcal{H}_\beta$ is the order β-Sobolev space with $\beta > d/2$. It is classical that $\mu_j \asymp j^{-2\beta/d}$ for the kernel eigenvalues associated with this setup. Thus, calculating k_n in Corollary 7, we find $k_n \asymp (\frac{\sigma^2}{\varrho^2 n})^{-\frac{d}{2\beta+d}}$, and consequently

$$\inf_{\hat{f}} \sup_{\|f^\star\|_{\mathcal{H}_\beta} \leqslant \varrho} \mathbf{E}\,\|\hat{f} - f^\star\|_{L^2(P)}^2 \asymp \varrho^2 \Big(\frac{\sigma^2}{\varrho^2 n}\Big)^{\frac{2\beta}{2\beta+d}},$$

provided that $\sigma^2 \gtrsim \varrho^2$. The above relation recovers a classical result [35, 60].

3.2.3. *Kernel regression under covariate shift.* We now discuss one important case in which we have $P \neq \nu$ in the RKHS model (36). In the setting of covariate shift, the model (36) comprises of covariates x_i drawn from a *source* distribution P that is different from the *target* distribution Q of covariates on which estimates of the regression function are to be deployed. In this setting, then we take $\nu = Q$ and $P \neq Q$.

For any such pair, following the argument given previously in Section 3.2, we find that

$$\inf_{\hat{f}} \sup_{f^\star \in \mathsf{B}_{\mathcal{H}}(\varrho)} \mathbf{E}\Big[\|\hat{f} - f^\star\|_{L^2(Q)}^2\Big] \asymp \frac{\sigma^2}{n} \limsup_{k \to \infty} d_n^{(k)}, \tag{48}$$

where the quantity $d_n^{(k)}$ is defined as in display (44). Above, the expectation on the lefthand side is over the noise and the covariates drawn from P as described by the model (36). Note that the eigenvalues $\{\mu_j\}_{j \geqslant 1}$ here correspond to the diagonalization of the integral kernel operator under the target distribution Q.

Let us now compare to past work due to Ma et al. [45], who studied the covariate shift problem in RKHSs. In contrast to this work, our result is *source-target distribution-dependent*: it characterizes, apart from universal constants, the minimax risk for any kernel, any radius, any noise level, and any covariate shift pair (P, Q). By contrast, the results in the paper [45] consider a more restrictive setup in which pair (P, Q) satisfy an absolute continuity condition ($Q \ll P$), and moreover, the likelihood ratio is P-essentially bounded, meaning that there exists some $B \in [1, \infty)$ such that

$$\frac{\mathrm{d}Q}{\mathrm{d}P}(x) \leqslant B, \quad \text{for } P\text{-almost every } x. \tag{49}$$

Let $d_\infty(P, Q)$ denote the P-essential supremum of the likelihood ratio $\mathrm{d}Q/\mathrm{d}P$ when $Q \ll P$ and $d_\infty(P, Q) = +\infty$ otherwise. "Uniform" results, where minimax risks of estimation are studied over families of covariate shifts P relative to Q where $d_\infty(P, Q) \leqslant B$ for some parameter B can be derived as a corollary to the sharper rate description (48).

To give one simple and concrete illustration of this, we will show how one can derive Theorem 2 in the paper [45]. By Jensen's inequality, we have

$$d_n^{(k)} \geqslant \sup_{\Omega \succ 0} \Big\{ \mathbf{Tr}(\mathbf{E}_{P^n} M_k^{-1/2} \Sigma_n^{(k)} M_k^{-1/2} + \Omega^{-1})^{-1} : \mathbf{Tr}(M_k^{-1}\Omega) \leqslant \frac{n\varrho^2}{\sigma^2} \Big\}. \tag{50}$$

If P satisfies $d_\infty(P, Q) \leqslant B$, then it follows that we have the ordering

$$\mathbf{E}_{P^n} M_k^{-1/2} \Sigma_n^{(k)} M_k^{-1/2} \succcurlyeq \frac{1}{B} I_k. \tag{51}$$

Moreover, this lower bound can be achieved by a shift P whenever the zero sets of the eigenfunctions ϕ_j in $L^2(Q)$ of the integral operator associated with the kernel k have nontrivial intersection. Equivalently, when there exists

$$x_0 \in \bigcap_{j \geqslant 1} \phi_j^{-1}(\{0\}), \tag{52}$$

then the bound (51) is achieved by the distribution $P_{x_0} := \frac{1}{B} Q + \Big(1 - \frac{1}{B}\Big)\delta_{x_0}$. This choice is evidently a B-bounded shift relative to Q. To give an example where the zero set condition (52) holds, note that in the case of where the kernel k is associated with the periodic β-order Sobolev class on $[0, 1]$ and Q is the uniform law on $[0, 1]$, one can take $x_0 = 0$ as the eigenfunctions are sinusoids.

Now, combining relations (48) and (50) with the choice of $P = P_{x_0}$ given above, we have

$$\sup_{P : d_\infty(P,Q) \leqslant B} \inf_{\hat{f}} \sup_{f^\star \in \mathsf{B}_{\mathcal{H}}(\varrho)} \mathbf{E}\Big[\|\hat{f} - f^\star\|_{L^2(Q)}^2 \Big] \gtrsim \frac{\sigma^2}{n} \sup_{\omega > 0} \Big\{ \sum_{j=1}^\infty \frac{B\omega_j}{\omega_j + B} : \sum_{j=1}^\infty \frac{\omega_j}{\lambda_j} = \frac{n\varrho^2}{\sigma^2} \Big\}$$

$$\asymp \varrho^2 \sup_{\lambda} \Big\{ \sum_{j=1}^\infty \frac{\sigma^2 B}{n\varrho^2} \wedge \lambda_j \mu_j : \lambda_j \geqslant 0, \ \sum_{j=1}^\infty \lambda_j = 1 \Big\}. \tag{53}$$

Suppose, following the paper [45], we additionally impose a regularity condition on the decay of the eigenvalues μ_j of kernel integral operator in $L^2(Q)$. Namely, that there exists a constant $c \in (0, \infty)$ such that

$$\sup_{\delta > 0} \frac{\sum_{j > d(\delta)} \mu_j}{\delta^2 d(\delta)} \leqslant c, \quad \text{where} \quad d(\delta) := \inf\{ j \geqslant 1 : \mu_j \leqslant \delta^2 \}. \tag{54}$$

Under this condition, we can further lower bound (53), up to universal constants, by

$$\varrho^2 \inf_{\delta > 0} \Big\{ \delta^2 + \frac{\sigma^2 B}{\varrho^2 n} d(\delta) \Big\}. \tag{55}$$

The details of this calculation can be found in Appendix B.2.6 of the supplement. Note that by establishing the lower bound (55), we have recovered Theorem 2 from the paper [45]. We remark that—as seen from the steps taken to arrive at this lower bound—our more general determination of the minimax rate (48) is sharper in that it holds for a fixed pair (P, Q) rather than uniformly over the larger class $\{P : d_\infty(P, Q) \leqslant B\}$. Moreover, our result, as compared to the work [45], requires fewer regularity assumptions on the underlying kernel and its diagonalization in the target Hilbert space $L^2(Q)$. In fact, as demonstrated in Appendix B.2.6, the regularity condition (54) is *not* necessary for us to establish the lower bound (55).

4. Proofs of Theorems 1 and 2.

In this section, we present the proofs of our main results. In Section 4.1, we provide the proof of our minimax upper bound (cf. Theorem 1). In Section 4.2, we provide the proof of our minimax lower bound. Some calculations and routine verifications are deferred to Appendix C in the supplement.

4.1. *Proof of Theorem 1.*

In this section, we develop an upper bound on the minimax risk. In order to do so, so, we define the risk function

$$r(\widehat{\theta}, \theta^\star) := \sup_{\nu \in \mathcal{P}(\Sigma_w)} \mathbf{E}_{(\xi, w) \sim \mathbb{P} \times \nu} \mathbf{E} \Big[\|\widehat{\theta}(T_\xi, T_\xi \theta^\star + w) - \theta^\star\|_{K_e}^2 \Big].$$

defined for any measurable estimator $\widehat{\theta}$ of (T_ξ, y), and any $\theta^\star \in \Theta(\varrho, K_c)$. Evidently, the minimax risk we are bounding is then expressible as

$$\mathfrak{M}(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c) = \inf_{\widehat{\theta}} \sup_{\theta^\star \in \Theta(\varrho, K_c)} r(\widehat{\theta}, \theta^\star). \tag{56}$$

In order to derive an upper bound, we restrict our focus to estimators that are *conditionally linear*. Formally, we consider the class of procedures

$$\widehat{\theta}_C(T_\xi, y) := C(T_\xi) T_\xi^\mathsf{T} \Sigma_w^{-1} y, \tag{57}$$

where C is a $\mathbf{R}^{d \times d}$-valued measurable function of T_ξ. Our strategy involves the following three steps:

(i) First, we compute the supremum risk over the parameter set $\Theta(\varrho, K_c)$ and all $\nu \in \mathcal{P}(\Sigma_w)$.
(ii) Second, compute the minimizer of the supremum risk in the choice of C in (57).

(iii) Finally, by using the curvature of the supremum risk and appealing to a min-max theorem, we put the pieces together to determine the final minimax risk.

The following subsections are devoted to the details associated with each of these three steps. In all cases, we defer routine calculations and verification to Appendix C.1 of the supplement.

4.1.1. *Supremum risk of estimator $\widehat{\theta}_C$.* Starting with the definition (57), for any matrix C, we have

$$\widehat{\theta}_C - \theta^\star = (C(T_\xi)T_\xi^\top \Sigma_w^{-1} T_\xi - I_d)\theta^\star + C(T_\xi)T_\xi^\top \Sigma_w^{-1} w.$$

Therefore, the risk $r(\widehat{\theta}_C, \theta^\star)$ associated with $\widehat{\theta}_C$ can be bounded as

$$\begin{aligned} r(\widehat{\theta}_C, \theta^\star) &\coloneqq \sup_{\nu \in \mathcal{P}(\Sigma_w)} \mathbf{E}\Big[\|\widehat{\theta}_C(X,y) - \theta^\star\|_{K_e}^2 \Big] \\ &= \mathbf{Tr}\Big\{ K_e^{1/2}\, \mathbf{E}_\xi \Big[(C(T_\xi)T_\xi^\top \Sigma_w^{-1} T_\xi - I_d)\theta^\star \otimes \theta^\star (C(T_\xi)T_\xi^\top \Sigma_w^{-1} T_\xi - I_d)^\top \\ &\qquad + C(T_\xi)T_\xi^\top \Sigma_w^{-1} T_\xi C(T_\xi)^\top \Big] K_e^{1/2} \Big\}. \end{aligned} \tag{58}$$

The equality above uses the property (N2) of distributions $\nu \in \mathcal{P}(\Sigma_w)$; note that it is achieved by the Gaussian distribution $\nu = \mathsf{N}(0, \Sigma_w)$.

4.1.2. *Curvature and minimizers of the functional $r(\widehat{\theta}_C, \theta^\star)$.* We begin by observing that the function $r(\widehat{\theta}_C, \cdot) \colon \Theta(\varrho, K_c) \to \mathbf{R}_+$ can be replaced by an equivalent mapping—which, with a slight abuse of notation we denote by the same symbol r— on the space of symmetric positive definite matrices of the form

$$\mathcal{K}(\varrho, K_c) \coloneqq \Big\{ \Omega \succcurlyeq 0 \mid \mathbf{Tr}(K_c^{-1/2} \Omega K_c^{-1/2}) \leqslant \varrho^2 \Big\}.$$

We define (in a sense, this is can be regarded as an extension to the set $\mathcal{K}(\varrho, K_c)$)

$$\begin{aligned} r(\widehat{\theta}_C, \Omega) \coloneqq \mathbf{Tr}\Big\{ K_e^{1/2}\, \mathbf{E}_\xi \Big[&(C(T_\xi)T_\xi^\top \Sigma_w^{-1} T_\xi - I_d)\Omega (C(T_\xi)T_\xi^\top \Sigma_w^{-1} T_\xi - I_d)^\top \\ &+ C(T_\xi)T_\xi^\top \Sigma_w^{-1} T_\xi C(T_\xi)^\top \Big] K_e^{1/2} \Big\}. \end{aligned} \tag{59}$$

Note that $r(\widehat{\theta}_C, \theta^\star) = r(\widehat{\theta}_C, \theta^\star \otimes \theta^\star)$ for $\theta^\star \in \Theta(\varrho, K_c)$. We claim that the suprema over $\Theta(\varrho, K_c)$ and $\mathcal{K}(\varrho, K_c)$ are the same.

LEMMA 1. *The suprema of the risk functional r taken over either the set $\Theta(\varrho, K_c)$ or the set $\mathcal{K}(\varrho, K_c)$ are equal—that is, we have*

$$\sup_{\theta^\star \in \Theta(\varrho, K_c)} r(\widehat{\theta}_C, \theta^\star) = \sup_{\Omega \in \mathcal{K}(\varrho, K_c)} r(\widehat{\theta}_C, \Omega),$$

for every conditionally linear estimator $\widehat{\theta}_C$ of the form (57).

See Appendix C.1.1 of the supplement for the proof of this claim. Briefly, the argument underlying this claim shows that the risk functional is affine in Ω and the set $\mathcal{K}(\varrho, K_c)$ can be viewed as the closed convex hull of rank-one outer products $\theta^\star \otimes \theta^\star$.

Our next result characterizes some properties of the mapping $(C, K) \mapsto r(\widehat{\theta}_C, K)$.

LEMMA 2. *Over the set of measurable functions C and matrices $\Omega \in \mathcal{K}(\varrho, K_c)$, the mapping $(C, \Omega) \mapsto r(\widehat{\theta}_C, \Omega)$ is affine in Ω and convex in C.*

See Appendix C.1.2 for the proof of this claim.

Our next claim determines the minimizer of $r(\cdot, \Omega)$ over estimators $\widehat{\theta}_C$ of the form (57), provided that Ω is strictly positive definite.

PROPOSITION 3. *Let Ω be a symmetric positive definite matrix. Then*

$$\inf_C r(\widehat{\theta}_C, \Omega) = \mathbf{Tr}\left\{K_e^{1/2}\, \mathbf{E}_\xi (\Omega^{-1} + T_\xi^{\mathsf{T}} \Sigma_w^{-1} T_\xi)^{-1} K_e^{1/2}\right\} \tag{60}$$

Moreover, the infimum is attained with the choice $C(T_\xi) = (\Omega^{-1} + T_\xi^{\mathsf{T}} \Sigma_w^{-1} T_\xi)^{-1}$.

See Appendix C.1.3 for the proof.

4.1.3. *Proof of Theorem 1.* We now piece together the previous lemmas to establish our main upper bound, as claimed in Theorem 1. In view of the relation (56) and the bound (58), we find that

$$\mathfrak{M}(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c) \leqslant \inf_C \sup_{\theta^\star \in \Theta(\varrho, K_c)} r(\widehat{\theta}_C, \theta^\star) \tag{61a}$$

$$= \inf_C \sup_{\Omega \in \mathcal{K}(\varrho, K_c)} r(\widehat{\theta}_C, \Omega) \tag{61b}$$

$$= \sup_{\Omega \in \mathcal{K}(\varrho, K_c)} \inf_C r(\widehat{\theta}_C, \Omega) \tag{61c}$$

$$= \sup_{\substack{\Omega \succ 0 \\ \mathbf{Tr}(K_c{}^{-1}\Omega) \leqslant \varrho^2}} \mathbf{E}\,\mathbf{Tr}\left(K_e^{1/2} (\Omega^{-1} + T_\xi^{\mathsf{T}} \Sigma_w^{-1} T_\xi)^{-1} K_e^{1/2}\right). \tag{61d}$$

To clarify, in the first display (61a) and below, the infimum over C denotes an infimum over all $\mathbf{R}^{d \times d}$-valued measurable functions of T_ξ. In display (61b), we have applied Lemma 1. Relation (61c) follows from the generalized Ky Fan min-max theorem [10, Theorem A] together with Lemma 2. Note that the set $\mathcal{K}(\varrho, K_c)$ is evidently a compact convex subset of $\mathbf{R}^{d \times d}$. The final equality (61d) is essentially an application of Proposition 3; see Appendix C.1.4 for the details of this verification.

4.2. *Proof of lower bound, Theorem 2.* In this section, we prove our lower bound on the minimax risk. In order to do so, we focus on lower bounding the Gaussian minimax risk

$$\mathfrak{M}^{\mathrm{G}}(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c) := \inf_{\widehat{\theta}} \sup_{\theta^\star \in \Theta(\varrho, K_c)} \mathbf{E}_{(\xi, w) \sim \mathbb{P} \times \mathsf{N}(0, \Sigma_w)} \left[\|\widehat{\theta}(T_\xi, T_\xi \theta^\star + w) - \theta^\star\|_{K_e}^2\right].$$

Evidently, the Gaussian minimax risk lower bounds the general minimax risk, so that we have $\mathfrak{M}^{\mathrm{G}} \leqslant \mathfrak{M}$. In Section 4.2.1, we reduce this Gaussian minimax risk to yet another Gaussian observation model. A minimax lower bound for this auxiliary problem is then presented as Proposition 4 in Section 4.2.2. This result is the bulk of the proof of the lower bound, and it quickly allows us to establish our main result, Theorem 2. In Section 4.2.3, we then complete the proof of Proposition 4.

4.2.1. *Reduction to an alternate observation model.* To establish the lower bound, we first show that the minimax risk associated with our estimation problem is equivalent to another, perhaps simpler, minimax risk.

An auxiliary observation model. This observation model is defined by a random quadruple $(r, V, \Lambda, \Upsilon)$. The triple (r, V, Λ) comprises a random integer r, a random orthogonal matrix $V \in \mathbf{R}^{d \times r}$ satisfying $V^\mathsf{T} V = I_r$, and a random, $r \times r$ diagonal positive definite matrix Λ. Conditional on (r, V, Λ), the observation Υ is a Gaussian random variable, satisfying the equation

$$\Upsilon = VV^\mathsf{T}\eta^\star + V\Lambda^{-1/2}z, \quad \text{where} \quad z \sim \mathsf{N}\left(0, I_r\right). \tag{62}$$

Above, the random vector z is drawn from the multivariate Gaussian with identity covariance in $\mathbf{R}^r$; it is independent of (r, V, Λ). If $\omega := (r, V, \Lambda)$ is distributed according to $\mathbb{Q}$, we denote the minimax risk for this observation model as

$$\mathfrak{M}^{\mathrm{G}}_{\mathrm{red}}(\mathbb{Q}, K) := \inf_{\widehat{\eta}} \sup_{\eta \in \Theta(K)} \mathbf{E}_{(\omega,\Upsilon)}\left[\|\widehat{\eta}(\omega, \Upsilon) - \eta\|_2^2\right].$$

Above, the expectation indexed by (ω, Υ) is over $\omega \sim \mathbb{Q}$ and Υ as in (62). The infimum is over measurable functions of (ω, Υ). The set $\Theta(K)$ is a shorthand for the set $\Theta(1, K) = \{\|\theta\|_K \leqslant 1\}$.

Reduction to the new observation model. We formally reduce the minimax risk $\mathfrak{M}^{\mathrm{G}}$ to the reduction $\mathfrak{M}^{\mathrm{G}}_{\mathrm{red}}$, as follows.

LEMMA 3. *Let $\widetilde{\mathbb{P}}$ denote the distribution of the triple $(r(\xi), V_\xi, \Lambda_\xi)$ under $\mathbb{P}$, where $r(\xi)$ is the (finite) rank of $Q_\xi = {K_e}^{-1/2} T_\xi^\mathsf{T} \Sigma_w^{-1} T_\xi {K_e}^{-1/2}$, and $Q_\xi = V_\xi \Lambda_\xi V_\xi^\mathsf{T}$ denotes the diagonalization of this positive definite matrix. Then, for any $(T, \mathbb{P}, \Sigma_w, \varrho, K_c, K_e)$, we have*

$$\mathfrak{M}^{\mathrm{G}}(T, \mathbb{P}, \Sigma_w, \varrho, K_c, K_e) = \mathfrak{M}^{\mathrm{G}}_{\mathrm{red}}(\widetilde{\mathbb{P}}, \varrho^2 {K_e}^{1/2} K_c {K_e}^{1/2}).$$

See Appendix C.2.1 of the supplement for a proof of this claim.

4.2.2. *Lower bounding the minimax risk.* We now focus on lower bounding $\mathfrak{M}^{\mathrm{G}}_{\mathrm{red}}$. The following result is a formal statement of the lower bound for the "reduced" minimax risk.

PROPOSITION 4. *For any $\tau \in (0, 1]$ and any $\Pi \succ 0$ such that $\mathbf{Tr}(K^{-1/2}\Pi K^{-1/2}) \leqslant 1$, we have*

$$\mathfrak{M}^{\mathrm{G}}_{\mathrm{red}}(\mathbb{Q}, K) \geqslant \mathbf{E}\,\mathbf{Tr}\left(\left(\tfrac{1}{c(\tau,\Pi)}\Pi^{-1} + V\Lambda V^\mathsf{T}\right)^{-1}\right), \tag{63}$$

where the constant $c(\tau, \Pi)$ is defined in Lemma 6. Moreover, we have the lower bounds

$$\mathfrak{M}^{\mathrm{G}}_{\mathrm{red}}(\mathbb{Q}, K) \geqslant \sup_{\Pi}\left\{\, \mathbf{E}\,\mathbf{Tr}\left((\Pi^{-1} + V\Lambda V^\mathsf{T})^{-1}\right) : \Pi \succ 0,\ \mathbf{Tr}(K^{-1/2}\Pi K^{-1/2}) \leqslant 1/4 \,\right\} \tag{64a}$$

$$\geqslant \frac{1}{4}\sup_{\Pi}\left\{\, \mathbf{E}\,\mathbf{Tr}\left((\Pi^{-1} + V\Lambda V^\mathsf{T})^{-1}\right) : \Pi \succ 0,\ \mathbf{Tr}(K^{-1/2}\Pi K^{-1/2}) \leqslant 1 \,\right\}. \tag{64b}$$

Proof of Theorem 2. We take the claim of Proposition 4 as given for the moment, and use it to derive our minimax lower bound. As mentioned, we may restrict to Gaussian noise to establish the lower bound; formally, we have $\mathfrak{M} \geqslant \mathfrak{M}^{\mathrm{G}}$. Additionally, the reduction given in Lemma 3 combined with the stronger lower bound (64a) in Proposition 4 gives us

$$\begin{aligned}&\mathfrak{M}(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c)\\ &\geqslant \sup\nolimits_\Pi \Big\{ \mathbf{E}\,\mathbf{Tr}\Big((\Pi^{-1} + K_e{}^{-1/2} T_\xi^{\mathsf{T}} \Sigma_w^{-1} T_\xi K_e{}^{-1/2})^{-1}\Big) : \Pi \succ 0, \mathbf{Tr}(K_e{}^{-1/2} \Pi K_e{}^{-1/2} K_c{}^{-1}) \leqslant \tfrac{\varrho^2}{4} \Big\}.\end{aligned}$$

Now define the matrix $\Omega = K_e{}^{-1/2} \Pi K_e{}^{-1/2}$. Then, the quantity on the righthand side is equal to

$$\sup_\Omega \Big\{ \mathbf{E}\,\mathbf{Tr}\Big(K_e{}^{1/2} (\Omega^{-1} + T_\xi^{\mathsf{T}} \Sigma_w^{-1} T_\xi)^{-1} K_e{}^{1/2}\Big) : \Omega \succ 0,\ \mathbf{Tr}(K_c{}^{-1/2} \Omega K_c{}^{-1/2}) \leqslant \tfrac{\varrho^2}{4} \Big\},$$

which furnishes the first inequality in Theorem 2. With similar manipulations to the weaker lower bound (64b) in Proposition (4), or by arguing directly from the display above, the second inequality in Theorem 2 follows. In order to establish the more detailed lower bound (8), we repeat the argument above but use (63).

4.2.3. *Proof of Proposition 4.* The lower bound proceeds in five steps:

(i) We first lower bound the minimax risk in terms of the expected conditional Bayesian risk over any prior on the parameter set $\Theta(K)$.

(ii) We then demonstrate that, conditionally, there is a family of auxiliary Bayesian estimation problems, indexed by a parameter $\lambda > 0$, which are all no harder than the Bayesian estimation problem implied by the conditional Bayesian risk.

(iii) We compute, in closed form, the Bayesian risk for any prior and any parameter $\lambda > 0$. We are able to show that the Bayesian risk is a functional of the Fisher information of the marginal distribution of the observed data under the prior and sampling model.

(iv) For each $\lambda > 0$, we then calculate a lower bound on the Fisher information for a prior obtained by conditioning a Gaussian distribution with mean zero and covariance Π to the parameter space.

(v) We put the pieces together: optimizing over all covariance operators Π, and the family of "easier" problems (*i.e.*, optimizing over $\lambda > 0$), we obtain our claimed lower bound.

Next, we present the details of the steps outlined above. Extended calculations and routine verification are deferred to Appendix C.2 of the supplement.

Step 1: Reduction to conditional Bayesian risk. We begin by lower bounding the minimax risk via the Bayes risk. Owing to the standard relation between minimax and Bayesian risks, we have for any prior π on $\Theta(K)$ that

(65)
$$\mathfrak{M}^{\mathrm{G}}_{\mathrm{red}}(\mathbb{Q}, K) = \inf_{\widehat{\eta}} \sup_{\eta \in \Theta(K)} \mathbf{E}_{(\omega, \Upsilon)} \Big[\|\widehat{\eta}(\omega, \Upsilon) - \eta\|_2^2\Big] \geqslant \inf_{\widehat{\eta}} \mathbf{E}_{\eta \sim \pi} \mathbf{E}_{(\omega, \Upsilon)} \big[\|\widehat{\eta} - \eta\|_2^2\big] =: B(\pi).$$

The quantity $B(\pi)$ appearing above is the Bayesian risk when the parameter η is drawn from the prior π. The following observation is key for the lower bound. After moving to Bayesian risks, we can condition on the "design", denoted by the random tuple $\omega = (r, V, \Lambda)$, and consider the conditional Bayesian risk. Formally, we have

$$B(\pi) = \inf_{\widehat{\eta}} \mathbf{E}_{\eta \sim \pi} \mathbf{E}_{(\omega, \Upsilon) \sim \mathcal{D}_\eta} \Big[\big\|\widehat{\eta} - \eta\big\|_2^2\Big] \geqslant \mathbf{E}_{\omega \sim \mathbb{Q}} \Big[\inf_{\widehat{\eta}_\omega} \mathbf{E}_{\eta \sim \pi} \mathbf{E}_{\Upsilon} \big\|\widehat{\eta}_\omega(\Upsilon) - \eta\big\|_2^2\Big]. \tag{66}$$

Above, the inequality follows by observing that if the function $\widehat{\eta}\colon (\omega, \Upsilon) \mapsto \widehat{\eta} \in \mathbf{R}^d$ is measurable, then $\widehat{\eta}_\omega(\Upsilon) := \widehat{\eta}(\omega, \Upsilon)$ is a measurable of Υ. Note that the infimum on the righthand side is restricted to those maps which are measurable function of ω; note that they may depend on ω, and therefore we have included a subscript depending on ω to indicate this.[6] To lighten notation in the subsequent discussion, we define the *conditional Bayesian risk* under π and for a realization of the random variable $\omega = \omega_0$,

$$B(\pi \mid \omega_0) := \inf_{\widehat{\eta}} \mathbf{E}_{\eta \sim \pi} \mathbf{E}_{z \sim \mathsf{N}(0, I_{r_0})} \Big[\big\| \widehat{\eta}(V_0 V_0^{\mathsf{T}} \eta + V_0 \Lambda_0^{-1/2} z) - \eta \big\|_2^2 \Big], \quad \text{where } \omega_0 = (r_0, V_0, \Lambda_0).$$

Using this definition, along with the two inequalities (65) and (66), we have demonstrated

$$\mathfrak{M}^{\mathrm{G}}_{\mathrm{red}}(\mathbb{Q}, K) \geqslant \mathbf{E}_{\omega \sim \mathbb{Q}} \big[B(\pi \mid \omega) \big], \qquad \text{for any prior } \pi \text{ on } \Theta(K). \tag{67}$$

Therefore, it suffices for us to lower bound $B(\pi \mid \omega)$.

Step 2: Reduction to a family of easier problems. In this step, we fix a parameter $\lambda > 0$, which will index yet another auxiliary Bayesian estimation problem. The intuition will be that as $\lambda \to 0^+$, we are "approaching" the difficulty of the original Bayesian estimation problem.

Formally, fix $\omega = (r, V, \Lambda)$. Throughout we will let $V_\perp \colon \mathbf{R}^d \to \mathbf{ran}(V)^\perp$ denote the projection of an element $\eta \in \mathbf{R}^d$ to the orthogonal complement of the closed subspace $\mathbf{ran}(V)$. We now consider the observation, where for an independent random Gaussian variable $z \sim \mathsf{N}(0, I_d)$

$$\Upsilon_\lambda = \underbrace{(V V^{\mathsf{T}} + \lambda V_\perp)}_{=:X_\lambda} \eta + V \Lambda^{-1/2} w + \sqrt{\lambda} V_\perp z = X_\lambda \eta + (V \Lambda^{-1} V^{\mathsf{T}} + \lambda V_\perp)^{1/2} w', \tag{68}$$

where the last equality holds in distribution. Define $\Sigma_\lambda := V \Lambda^{-1} V^{\mathsf{T}} + \lambda V_\perp$; evidently Σ_λ is a symmetric positive definite matrix for any $\lambda > 0$. Then, Υ_λ has distribution $\mathsf{N}(X_\lambda \eta, \Sigma_\lambda)$. We remark that the observation Υ_λ is more convenient than Υ as its covariance is nonsingular and moreover its mean is a nonsingular linear transformation of η—note that neither of these properties hold for Υ.

Our goal is to show that the observation Υ_λ is more "informative" than Υ. To do this, we now define the (conditional) Bayesian risk for Υ_λ,

$$B_\lambda(\pi \mid \omega) := \inf_{\widehat{\eta}} \Big\{ B_\lambda(\widehat{\eta}, \pi \mid \omega) := \mathbf{E} \big[\|\widehat{\eta}(\Upsilon_\lambda) - \eta\|_2^2 \big] \Big\}.$$

The main claim is that this provides a lower bound on our original conditional Bayesian risk.

LEMMA 4. *For any ω and $\lambda > 0$, we have*

$$B(\pi \mid \omega) \geqslant B_\lambda(\pi \mid \omega).$$

See Appendix C.2.2 for a proof of this claim.

Step 3: Calculation of Bayesian risk $B_\lambda(\pi \mid \omega)$, for a fixed prior π and parameter $\lambda > 0$. To compute the Bayesian risk for a fixed prior π and parameter $\lambda > 0$, we develop a variant of Tweedie's formula (also sometimes referred to as Brown's identity, when applied to Bayesian risks) [62, 55, 13].

[6] In some cases, this inequality may hold with equality. However, to be clear, in general the inequality arises since if $\{\widehat{\eta}_\omega\}_\omega$ is a family of measurable functions (of Υ) for each ω in the support of $\mathbb{Q}$, it is not necessarily the case that $\widehat{\eta}(\omega, \Upsilon) := \widehat{\eta}_\omega(\Upsilon)$ is measurable.

To state the result, we need to introduce some notation. We define the marginal and conditional densities of Υ_λ—disregarding normalization constants—as,

$$p(y) \coloneqq \int p(y \mid \eta)\, \pi(\mathrm{d}\eta) \qquad \text{where} \quad p(y \mid \eta) \coloneqq \exp\Big(-\frac{1}{2}\|y - X_\lambda \eta\|^2_{\Sigma_\lambda^{-1}} \Big).$$

Finally we define the Fisher information of the marginal distribution of Υ_λ, which is given by

$$\mathcal{I}\left(\Upsilon_\lambda\right) \coloneqq \mathbf{E}[\nabla \log p(\Upsilon_\lambda) \otimes \nabla \log p(\Upsilon_\lambda)].$$

With this notation in hand, we can now state our formula for the Bayesian risk under the prior π and for parameter $\lambda > 0$.

LEMMA 5. *Fix $\omega = (r, V, \Lambda)$. Define $X_\lambda \coloneqq VV^\mathsf{T} + \lambda V_\perp$ and $\Sigma_\lambda \coloneqq V\Lambda^{-1}V^\mathsf{T} + \lambda V_\perp$. Fix prior π, and parameter $\lambda > 0$. Then the conditional Bayesian risk is given by*

$$B_\lambda(\pi \mid \omega) = \mathbf{Tr}\left(X_\lambda^{-1}\Sigma_\lambda\big[\Sigma_\lambda^{-1} - \mathcal{I}\left(\Upsilon_\lambda\right)\big]\Sigma_\lambda X_\lambda^{-1}\right).$$

See Appendix C.2.3 for a proof of this claim.

Step 4: Lower bound on Fisher information for conditioned Gaussian prior. Consider a prior π which is absolutely continuous with respect to Lebesgue measure on $\mathbf{R}^d$. Furthermore, suppose that its Lebesgue density $f_\pi \coloneqq \frac{\mathrm{d}\pi}{\mathrm{d}\eta}$ has logarithmic gradient almost everywhere. Define

$$\mathcal{I}\left(\pi\right) \coloneqq \int \nabla \log f_\pi(\eta) \otimes \nabla \log f_\pi(\eta)\, \mathrm{d}\pi(\eta).$$

Recall also that the Fisher information associated with a Gaussian distribution $\mathsf{N}\left(\mu, \Pi\right)$ for nonsingular Π is given by Π^{-1} [42, Example 6.3]. Therefore, applying well-known results for the Fisher information [65, eqn. (8) and Corollary 1]

$$\mathcal{I}\left(\Upsilon_\lambda\right) \preccurlyeq (X_\lambda \mathcal{I}\left(\pi\right)^{-1} X_\lambda + \Sigma_\lambda)^{-1}. \tag{69}$$

Next, we select a prior distribution and calculate the Fisher information $\mathcal{I}\left(\Upsilon_\lambda\right)$ for the marginal density under this prior. For a parameter $\tau \in (0, 1]$ and symmetric positive definite covariance matrix Π, we define the probability measures

$$\pi^{\mathrm{G}}_{\tau,\Pi} = \mathsf{N}\left(0, \tau^2\Pi\right) \quad \text{and} \quad \pi_{\tau,\Pi} = \pi^{\mathrm{G}}_{\tau,\Pi}\big(\cdot \mid \Theta(K)\big). \tag{70}$$

In other words, $\pi_{\tau,\Pi}$ denotes the probability measure $\mathsf{N}\left(0, \tau^2\Pi\right)$ conditioned on the constraint set. Formally, it is defined by the relation,

$$\pi_{\tau,\Pi}(A) \coloneqq \frac{\pi^{\mathrm{G}}_{\tau,\Pi}\big(A \cap \Theta(K)\big)}{\pi^{\mathrm{G}}_{\tau,\Pi}\big(\Theta(K)\big)},$$

for any event A. For these priors, we have the following claim.

LEMMA 6. *Let $\tau \in (0, 1]$ and Π be a symmetric positive definite matrix satisfying the relation $\mathbf{Tr}(\Pi^{1/2}K^{-1}\Pi^{1/2}) \leqslant 1$. Then the Fisher information of the conditioned prior $\pi_{\tau,\Pi}$ satisfies the inequality*

$$\mathcal{I}\left(\pi_{\tau,\Pi}\right)^{-1} \succcurlyeq c(\tau, \Pi)\Pi,$$

where $c(\tau, \Pi) = \tau^2(1 - \pi^{\mathrm{G}}_{\tau,\Pi}(\Theta(K)^c)) > 0$.

See Appendix C.2.4 for the proof of this claim.

Step 5: Putting the pieces together. Combining Lemmas 4 and 5 along with the inequality (69) and Lemma 6, we find that for any $\tau \in (0,1]$ and symmetric positive definite matrix Π satisfying $\mathbf{Tr}(\Pi^{1/2}K^{-1}\Pi^{1/2}) \leqslant 1$, that

$$\begin{aligned} B(\pi \mid \omega) &\geqslant \sup_{\lambda>0} \mathbf{Tr}\left(X_\lambda^{-1}\Sigma_\lambda\big[\Sigma_\lambda^{-1} - (c(\tau,\Pi)X_\lambda \Pi X_\lambda + \Sigma_\lambda)^{-1}\big]\Sigma_\lambda X_\lambda^{-1}\right) \\ &= \sup_{\lambda>0} \mathbf{Tr}\left((\tfrac{1}{c(\tau,\Pi)}\Pi^{-1} + X_\lambda \Sigma_\lambda^{-1} X_\lambda)^{-1}\right). \end{aligned}$$

Above, we used the relation $A(A^{-1} - (B+A)^{-1})A = (A^{-1} + B^{-1})^{-1}$, valid for any pair (A,B) of symmetric positive definite matrices. Our particular choice of matrices was $A = \Sigma_\lambda$ and $B = X_\lambda$. Note that

$$X_\lambda \Sigma_\lambda^{-1} X_\lambda = V\Lambda V^{\mathsf{T}} + \lambda V_\perp.$$

Therefore, by continuity, we have

(71)
$$B(\pi \mid \omega) \geqslant \lim_{\lambda\to 0^+} \mathbf{Tr}\left((\tfrac{1}{c(\tau,\Pi)}\Pi^{-1} + V\Lambda V^{\mathsf{T}} + \lambda V_\perp)^{-1}\right) = \mathbf{Tr}\left((\tfrac{1}{c(\tau,\Pi)}\Pi^{-1} + V\Lambda V^{\mathsf{T}})^{-1}\right).$$

Taking the expectation over ω, and applying our minimax lower bound (67), we have established lower bound (63). Note that since $c(\tau,\Pi) \in (0,1]$, we evidently have from the above display that

$$B(\pi \mid \omega) \geqslant c(\tau,\Pi)\, \mathbf{Tr}\left((\Pi^{-1} + V\Lambda V^{\mathsf{T}})^{-1}\right).$$

Let us define the constant

$$c_\ell(K) \coloneqq \inf_{\substack{\Pi \succ 0 \\ \mathbf{Tr}(\Pi K^{-1}) \leqslant 1}} \sup_{\tau \in (0,1]} c(\tau,\Pi).$$

Then combining the conditional lower bound (71) with our minimax lower bound (67), we obtain

$$\begin{aligned} \mathfrak{M}^{\mathrm{G}}_{\mathrm{red}}(\mathbb{Q},K) &\geqslant \sup_{\Pi} \Big\{ \mathbf{E}\,\mathbf{Tr}\left((\Pi^{-1} + V\Lambda V^{\mathsf{T}})^{-1}\right) : \Pi \succ 0,\ \mathbf{Tr}(\Pi^{1/2}K^{-1}\Pi^{1/2}) \leqslant c_\ell(K) \Big\} \\ &= \sup_{\Pi} \Big\{ \mathbf{E}\,\mathbf{Tr}\left((\tfrac{1}{c_\ell(K)}\Pi^{-1} + V\Lambda V^{\mathsf{T}})^{-1}\right) : \Pi \succ 0,\ \mathbf{Tr}(\Pi^{1/2}K^{-1}\Pi^{1/2}) \leqslant 1 \Big\} \\ &\geqslant c_\ell(K) \sup_{\Pi} \Big\{ \mathbf{E}\,\mathbf{Tr}\left((\Pi^{-1} + V\Lambda V^{\mathsf{T}})^{-1}\right) : \Pi \succ 0,\ \mathbf{Tr}(\Pi^{1/2}K^{-1}\Pi^{1/2}) \leqslant 1 \Big\}. \end{aligned}$$

To complete the proof, we simply need to lower bound the constant $c_\ell(K)$ universally.

LEMMA 7. *The constant $c_\ell(K)$ is lower bounded, for any symmetric positive definite K, as*

$$c_\ell(K) \geqslant \frac{1}{4}.$$

See Appendix C.2.5 for a proof of this claim.

5. Discussion.

In this work, we determined the minimax risk of estimation for observation models of the form (1), where one observes the image of a unknown parameter under a random linear operator with additive noise. Our results reveal the dependence of the rate of convergence on the covariate law, the parameter space, the error metric, and the noise level. We conclude our paper by presenting some simulation results; see Section 5.1

Finally, we note that in this work we studied minimax risks of convergence in expectation. This is convenient, as it requires relatively minor assumptions of the distribution of T_ξ. On the other hand, for the setting of random design regression, high-probability results, such as those obtained in the papers [4, 49, 34, 41, 52], typically require stronger assumptions such as the sub-Gaussianity of the covariate distribution. Nonetheless, high-probability guarantees provide a complementary perspective on the problem we consider. Indeed, when the covariate law can be considered "heavy-tailed," it may be more relevant to develop robust estimators that have low risk with high probability. We refer to the survey article [44] for a overview of work in this direction.

5.1. *Some illustrative simulations.* We conclude our paper by presenting the results of some simulations reveal how changes in the distribution of the random operator T_ξ can lead to dramatic changes in the overall minimax risk.

In this section, we present simulation results to illustrate the behavior of the functionals appearing in our main results for two versions of random design linear regression. In Section 5.1.1, we present simulation results for a multivariate, random design linear regression setting with IID covariates. Concretely, we provide two different covariate laws, where the minimax error for the same parameter space differs by at least two orders of magnitude. We emphasize this difference in *entirely* due to the covariate law; the noise, observation model, error metric, and parameter space are fixed in this comparison.

Additionally, in Section 5.1.2, we present simulation results for a univariate regression setting where the covariates are sampled from a Markov chain. In both cases, the functional is able to capture the dependence of the minimax rate of estimation on the underlying covariate distribution.

5.1.1. *Higher-order effects in IID random design linear regression.* For random design linear regression, higher order properties of the covariate distribution over the covariates can have striking effects on the minimax risk. In order to illustrate this phenomenon, we consider the regression model (10) with feature map $\psi(x) = x$, and parameter vector $\theta^\star$ constrained to a ball in the Euclidean norm. We then construct a family of distributions over the covariates that are all zero-mean with identity covariance, but differ in interesting ways in terms of their higher-order moment properties. More precisely, we let δ_0 denote the Dirac measure with unit mass at 0, and for a mixture weight $\lambda \in [0,1]$, we consider covariates generated from the probability distribution

$$P_\lambda := \lambda\delta_0 + (1-\lambda)\mathsf{N}\left(0, \frac{1}{1-\lambda} I_d\right). \tag{72}$$

By construction, all members of the ensemble have the same behavior with respect to their first and second moments,

$$\mathbf{E}_{P_\lambda}[x] = 0 \quad \text{and} \quad \mathrm{Cov}_{P_\lambda}(x) = \mathbf{E}_{P_\lambda}[x \otimes x] = I_d, \quad \text{for all } \lambda \in [0,1]. \tag{73}$$

In the special case $\lambda = 0$, the distribution P_λ corresponds to the standard Gaussian law on $\mathbf{R}^d$, whereas it becomes an increasingly ill-behaved Gaussian mixture distribution as $\lambda \to 1^-$.

Following the argument in Section 3.1.1, in this case, the minimax risk is upper and lower bounded as

$$\frac{\sigma^2}{n}\mathbf{E}_{P_\lambda^n}[\mathbf{Tr}((\Sigma_n + \tfrac{c_d\sigma^2 d}{n\varrho^2} I_d)^{-1})] \leqslant \mathfrak{M}_n^{\mathrm{IID}}\Big(P_\lambda, \varrho, \sigma^2, I_d, I_d\Big) \leqslant \frac{\sigma^2}{n}\mathbf{E}_{P_\lambda^n}[\mathbf{Tr}((\Sigma_n + \tfrac{\sigma^2 d}{n\varrho^2} I_d)^{-1})]. \tag{74}$$

Above, the lower bound constant c_d is defined in display (20b).

To understand the effect of the covariate law, we fix the signal-to-noise ratio such that $\frac{\varrho}{\sigma} = \tau$, for $\tau \in \{1, 10\}$. Note that after renormalizing the minimax risk by ϱ^2, it only depends on τ (and not on the particular choices of (ϱ, σ)). Similarly, this invariance relation holds for the functionals appearing on the left- and righthand sides of the display (74)—after normalization by $1/\varrho^2$, they no longer depend on (ϱ, σ) except via the ratio $\tau = \frac{\varrho}{\sigma}$. Additionally, we fix the aspect ratio $\gamma = \frac{d}{n}$.[7]By varying $\gamma \in [0.05, 4]$ we are able to illustrate the behavior of the minimax risk, as characterized by our functional, for problems which are both under- and overdetermined.

Having fixed the SNR at τ and aspect ratio at γ, we can somewhat simplify the display (74), by introducing the following quantities which only depend on the parameters τ, γ and the sample size n and the mixture parameter λ,

$$\mathfrak{m}_n(\lambda, \tau, \gamma) := \frac{\mathfrak{M}_n^{\mathrm{IID}}\Big(P_\lambda, \tau\sigma, \sigma^2, I_{\lceil \gamma n \rceil}, I_{\lceil \gamma n \rceil}\Big)}{\tau^2 \sigma^2}, \tag{75a}$$

$$u_n(\lambda, \tau, \gamma) := \frac{1}{\tau^2 n} \mathbf{E}_{P_\lambda^n}[\mathbf{Tr}((\Sigma_n + \tfrac{\lceil \gamma n \rceil}{n\tau^2} I_{\lceil \gamma n \rceil})^{-1})], \tag{75b}$$

$$\ell_n(\lambda, \tau, \gamma) := \frac{1}{\tau^2 n} \mathbf{E}_{P_\lambda^n}[\mathbf{Tr}((\Sigma_n + \tfrac{c_d \lceil \gamma n \rceil}{n\tau^2} I_{\lceil \gamma n \rceil})^{-1})]. \tag{75c}$$

Then, the relations (74), can be equivalently expressed as

$$\ell_n(\lambda, \tau, \gamma) \leqslant \mathfrak{m}_n(\lambda, \tau, \gamma) \leqslant u_n(\lambda, \tau, \gamma),$$

and moreover this holds for all $\lambda \in [0, 1], \tau > 0, \gamma > 0$. In our simulation, we use Monte Carlo simulation with 50 trials to estimate the upper and lower bound functionals ℓ_n and u_n.

In our simulations, we take $\lambda \in \{0, 0.9, 0.99\}$ and vary $\gamma \in [0.05, 4]$. The results of these simulations are presented in Figure 1; see the caption for a detailed description and commentary. The general pattern should be clear: the covariate law can have a dramatic impact on the overall rate of estimation, even when restricting some moments such as we have with the relations (73).

5.1.2. *Mixing time effects in Markovian linear regression.* Covariates need not be drawn in an IID manner, and any dependencies can be expected to affect the minimax risk. Here we illustrate this general phenomena via some simulations for the Markov regression example as outlined in Section 3.1.4. We seek to study a wide range of possible mixing conditions for the Markovian covariate model. In order to do so, we consider covariates generated from the Markovian model (26) with

$$r_t = \frac{\psi(t-1)}{\psi(t)},$$

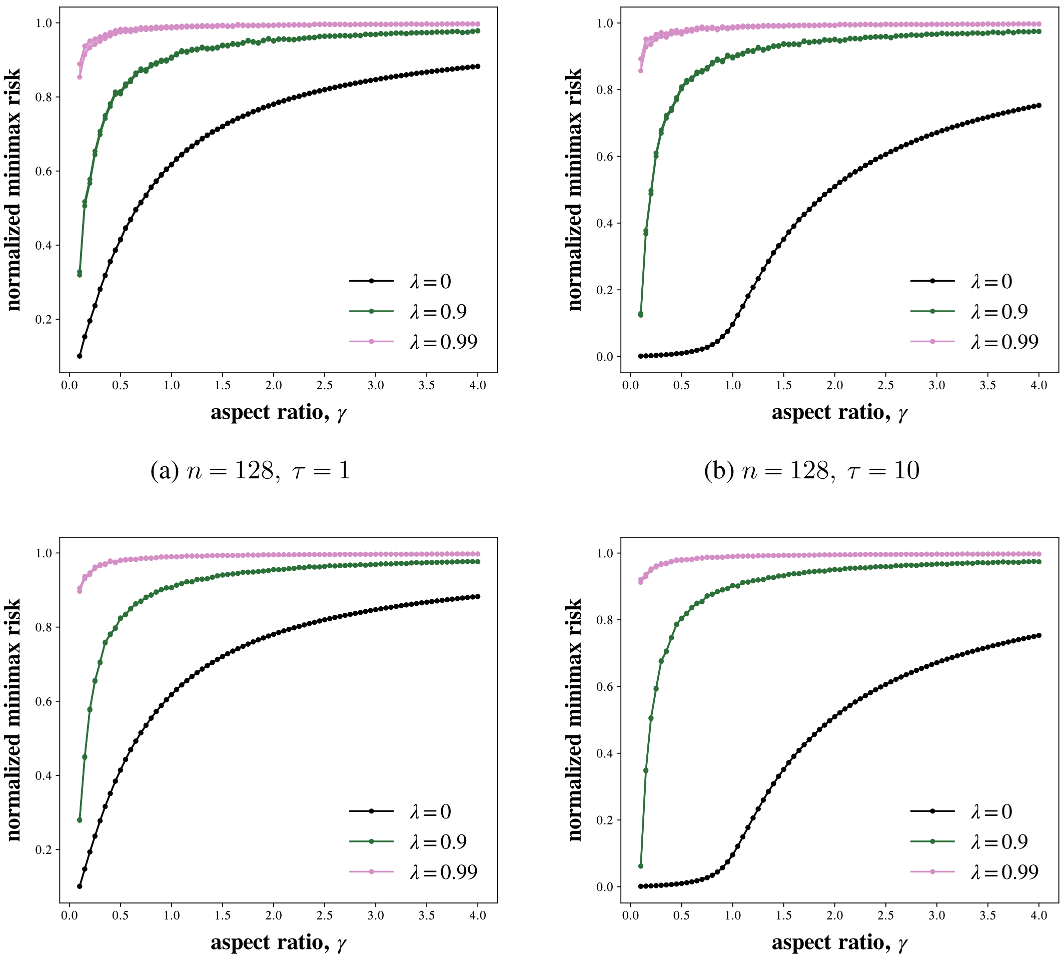
where $\psi \colon \mathbf{N} \cup \{0\} \to \mathbf{R}_+$ is a nondecreasing function satisfying $\psi(0) = 1$ and $\lim_{t \to \infty} \psi(t) = \infty$. With this choice, it is easily checked that, marginally

$$x_t \sim \mathsf{N}\left(0, 1 - \frac{1}{\psi(t)}\right).$$

Therefore, $x_t \to \mathsf{N}(0, 1)$ in distribution as $t \to \infty$, and the rate of convergence is of order $1/\psi(t)$.

We now illustrate how the minimax rate, as determined in Corollary 5, for this problem behaves for different choices of the function ψ and the signal-to-noise ratio (SNR). As in

[7] Specifically, we take $d = \lceil \gamma n \rceil$.



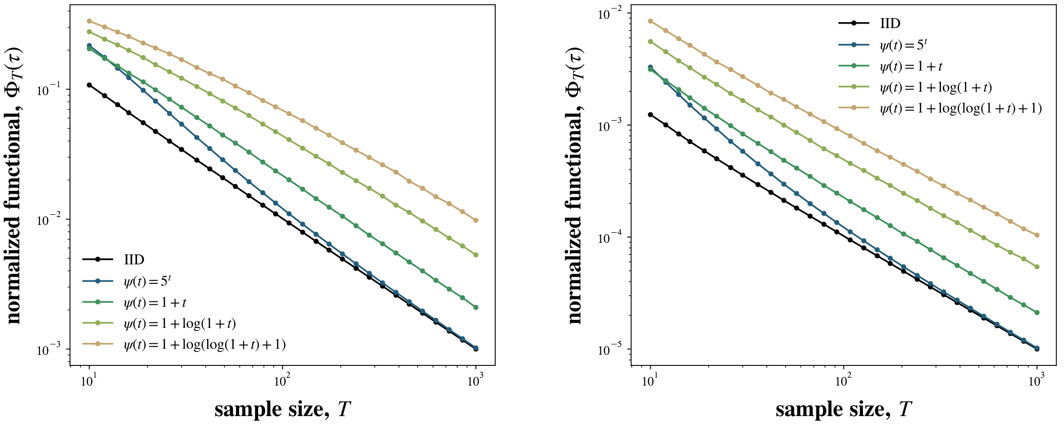
(a) $n = 128,\ \tau = 1$

(b) $n = 128,\ \tau = 10$

(c) $n = 512,\ \tau = 1$

(d) $n = 512,\ \tau = 10$

Fig 1. Simulations of random design regression for three covariate laws, P_λ as defined in equation (72) with $\lambda \in \{0, 0.9, 0.99\}$. For a given choice of the mixture weight λ and signal-to-noise ratio (SNR) τ, we plot the lower bound $\ell_n(\lambda, \tau, \gamma)$ and upper bound $u_n(\lambda, \tau, \gamma)$ as γ varies between 0.05 and 4. The normalized minimax risk $\mathfrak{m}_n$ is then guaranteed to lie in the region whose upper and lower envelopes are given by u_n and ℓ_n, respectively. To facilitate interpretation of these figures, we have shaded this region to highlight where we can guarantee the minimax risk $\mathfrak{m}_n$ must lie. The quantities $u_n, \ell_n, \mathfrak{m}_n$ are all defined in display (75). In panels (1a) and (1b), we set the sample size $n = 128$, and set the SNR as $\tau = 1, 10$, respectively. In panels (1c) and (1d), we set the sample size $n = 512$, and set the SNR as $\tau = 1, 10$, respectively. The plots above demonstrate that as λ increases, the minimax risks are much worse. Numerically, in the setting where $n = 512$ and $\tau = 10$—as depicted in panel (1d)—our upper and lower bounds guarantee that the minimax risk for the isotropic ensemble (depicted with $\lambda = 0$ above) can be over 806 times larger than the minimax risk for the ensemble with $\lambda = 0.99$. It should be noted that in this comparison the first and second moments of the ensemble are held fixed (see equation (73)), and hence the differences between the lines plotted in any given panel can only be explained by differences in higher-order moments within the ensemble $\{P_\lambda\}$. The figures also demonstrate that the gap between our upper and lower bounds is fairly small, particularly whenever $d > 5$.



(a) $\tau = 1$

(b) $\tau = 10$

Fig 2. Simulations for five distributions of Markovian covariates. In panel (2a), we set the SNR parameter as $\tau = 1$, and in panel (2b), we set the SNR parameter as $\tau = 10$. As the scaling function ψ grows more slowly, the chain converges to its stationary distribution more slowly, and the minimax rate decays more slowly, as indicated by the displayed behavior of our functional $T \mapsto \Phi_T(\tau)$.

Section 5.1.1, we normalize the minimax risk by the squared radius so that it only depends on $\tau = \frac{\varrho}{\sigma}$. The quantity we then plot is

$$\Phi_T(\tau) \coloneqq \frac{\Phi_T(\tau, 1)}{\tau^2},$$

where $\Phi_T(\varrho, \sigma)$ is the functional appearing in Corollary 5.

In the simulation, we consider the following choices of scaling function ψ,

$$5^t, \quad t + 1, \quad 1 + \log(t + 1), \quad \text{and} \quad 1 + \log\big(1 + \log(t + 1)\big).$$

With the choice $\psi(t) = 5^t$, the underlying Markov chain converges geometrically to the standard Normal law. On the other hand, the choice $\psi(t) = \log(1 + \log(1 + t)) + 1$ exhibits much slower convergence—the variational distance between the law of x_t and $\mathsf{N}(0, 1)$ is of order $O(1/(\log \log t))$.

We simulate each of these chains, computing the normalized functional $\Phi_T(\tau)$ over the course of 5000 Monte Carlo trials. The sample size T is varied between 10 and 3162. In the simulation we also include the choice $r_t \equiv 0$, which corresponds to IID covariates. The results of the simulation are presented in Figure 2; see the caption for more details and commentary.

Acknowledgments. We thank Jaouad Mourtada for a helpful conversation and useful email exchanges; we also thank Peter Bickel for a helpful discussion regarding his prior work on the Gaussian sequence model.

Funding. RP was partially supported by a UC Berkeley Chancellor's Fellowship via the ARCS Foundation. MJW and RP were partially funded by ONR grant N00014-21-1-2842, National Science Foundation Grant grant CCF-1955450, and National Science Foundation grant DMS-2015454. MJW and RP gratefully acknowledge funding support from Meta via the UC Berkeley AI Research (BAIR) Commons initiative.

SUPPLEMENTARY MATERIAL

Supplementary material

Contains omitted proofs from the main text.

REFERENCES

[1] ANDERSON, T. W. (2003). *An introduction to multivariate statistical analysis*, third ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

[2] ANTONIADIS, A., PENSKY, M. and SAPATINAS, T. (2014). Nonparametric regression estimation based on spatially inhomogeneous data: minimax global convergence rates and adaptivity. *ESAIM Probab. Stat.* **18** 1–41.

[3] ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404. https://doi.org/10.2307/1990404

[4] AUDIBERT, J.-Y. and CATONI, O. (2011). Robust linear least squares regression. *Ann. Statist.* **39** 2766–2794. https://doi.org/10.1214/11-AOS918 MR2906886

[5] BELITSER, E. N. and LEVIT, B. Y. (1995). On minimax filtering over ellipsoids. *Math. Methods Statist.* **4** 259–273.

[6] BERKSON, J. (1950). Are There Two Regressions? *Journal of the American Statistical Association* **45** 164–180.

[7] BERRY, J. C. (1990). Minimax estimation of a bounded normal mean vector. *J. Multivariate Anal.* **35** 130–139. https://doi.org/10.1016/0047-259X(90)90020-I

[8] BHATIA, R. (2007). *Positive definite matrices*. *Princeton Series in Applied Mathematics*. Princeton University Press, Princeton, NJ.

[9] BICKEL, P. J. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Statist.* **9** 1301–1309.

[10] BORWEIN, J. M. and ZHUANG, D. (1986). On Fan's minimax theorem. *Math. Programming* **34** 232–234. https://doi.org/10.1007/BF01580587

[11] BOYD, S. and VANDENBERGHE, L. (2004). *Convex optimization*. Cambridge University Press, Cambridge.

[12] BREIMAN, L. and FREEDMAN, D. (1983). How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.* **78** 131–136.

[13] BROWN, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42** 855–903.

[14] CARROLL, R. J., RUPPERT, D. and STEFANSKI, L. A. (1995). *Measurement Error in Nonlinear Models*. Chapman and Hall.

[15] CASELLA, G. and STRAWDERMAN, W. E. (1981). Estimating a bounded normal mean. *Ann. Statist.* **9** 870–878.

[16] CUCKER, F. and SMALE, S. (2002). On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)* **39** 1–49. https://doi.org/10.1090/S0273-0979-01-00923-5

[17] DICKER, L. H. (2016). Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli* **22** 1–37. https://doi.org/10.3150/14-BEJ609

[18] DONOHO, D. L. (1994). Statistical estimation and optimal recovery. *Ann. Statist.* **22** 238–270.

[19] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields* **99** 277–303. https://doi.org/10.1007/BF01199026

[20] DONOHO, D. L., LIU, R. C. and MACGIBBON, B. (1990). Minimax risk over hyperrectangles, and implications. *Ann. Statist.* **18** 1416–1437. https://doi.org/10.1214/aos/1176347758

[21] FOURDRINIER, D., STRAWDERMAN, W. E. and WELLS, M. T. (2018). *Shrinkage estimation*. *Springer Series in Statistics*. Springer, Cham. https://doi.org/10.1007/978-3-030-02185-6

[22] GAÏFFAS, S. (2005). Convergence rates for pointwise curve estimation with a degenerate design. *Math. Methods Statist.* **14** 1–27.

[23] GAÏFFAS, S. (2007). Sharp estimation in sup norm with random design. *Statist. Probab. Lett.* **77** 782–794.

[24] GAÏFFAS, S. (2007). On pointwise adaptive curve estimation based on inhomogeneous data. *ESAIM Probab. Stat.* **11** 344–364.

[25] GAÏFFAS, S. (2009). Uniform estimation of a signal based on inhomogeneous data. *Statist. Sinica* **19** 427–447.

[26] GOGOLASHVILI, D. (2022). Importance Weighting Correction of Regularized Least-Squares for Covariate and Target Shifts. https://doi.org/10.48550/ARXIV.2210.09709

[27] GOGOLASHVILI, D., ZECCHIN, M., KANAGAWA, M., KOUNTOURIS, M. and FILIPPONE, M. (2023). When is Importance Weighting Correction Needed for Covariate Shift Adaptation? https://doi.org/10.48550/ARXIV.2303.04020

[28] GOLDENSHLUGER, A. and TSYBAKOV, A. (2001). Adaptive prediction and estimation in linear regression with infinitely many parameters. *Ann. Statist.* **29** 1601–1619. https://doi.org/10.1214/aos/1015345956

[29] GOLDENSHLUGER, A. and TSYBAKOV, A. (2003). Optimal prediction for linear regression with infinitely many parameters. *J. Multivariate Anal.* **84** 40–60. https://doi.org/10.1016/S0047-259X(02)00006-4

[30] GOLUBEV, G. K. (1990). Quasilinear estimates for signals in L_2. *Problemy Peredachi Informatsii* **26** 19–24. MR1051584

[31] GUILLOU, A. and KLUTCHNIKOFF, N. (2011). Minimax pointwise estimation of an anisotropic regression function with unknown density of the design. *Math. Methods Statist.* **20** 30–57.

[32] GYÖRFI, L., KOHLER, M., KRZY˙ ZAK, A. and WALK, H. (2002). *A distribution-free theory of nonparametric regression*. *Springer Series in Statistics*. Springer-Verlag, New York. https://doi.org/10.1007/b97848

[33] HSU, D., KAKADE, S. M. and ZHANG, T. (2014). Random design analysis of ridge regression. *Found. Comput. Math.* **14** 569–600. https://doi.org/10.1007/s10208-014-9192-1

[34] HSU, D. and SABATO, S. (2016). Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.* **17** Paper No. 18, 40.

[35] IBRAGIMOV, I. A. and KHAS′ MINSKIĬ, R. Z. (1980). Nonparametric regression estimation. *Dokl. Akad. Nauk SSSR* **252** 780–784.

[36] JOHNSTONE, I. M. (2019). Gaussian estimation: Sequence and wavelet models. Book manuscript.

[37] JUDITSKY, A. and NEMIROVSKI, A. (2018). Near-optimality of linear recovery in Gaussian observation scheme under $\| \cdot \|_2^2$-loss. *Ann. Statist.* **46** 1603–1629.

[38] KAC, M., MURDOCK, W. L. and SZEGÖ, G. (1953). On the eigenvalues of certain Hermitian forms. *J. Rational Mech. Anal.* **2** 767–800.

[39] KOH, P. W., SAGAWA, S., MARKLUND, H., XIE, S. M., ZHANG, M., BALSUBRAMANI, A., HU, W., YASUNAGA, M., PHILLIPS, R. L., GAO, I. et al. (2021). Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning* 5637–5664. PMLR.

[40] KPOTUFE, S. and MARTINET, G. (2021). Marginal singularity and the benefits of labels in covariate-shift. *Ann. Statist.* **49** 3299–3323.

[41] LECUÉ, G. and MENDELSON, S. (2016). Performance of empirical risk minimization in linear aggregation. *Bernoulli* **22** 1520–1534. https://doi.org/10.3150/15-BEJ701

[42] LEHMANN, E. L. and CASELLA, G. (1998). *Theory of point estimation*, second ed. *Springer Texts in Statistics*. Springer-Verlag, New York.

[43] LIU, M., ZHANG, Y., LIAO, K. P. and CAI, T. (2020). Augmented Transfer Regression Learning with Semi-non-parametric Nuisance Models.

[44] LUGOSI, G. and MENDELSON, S. (2019). Mean estimation and regression under heavy-tailed distributions: a survey. *Found. Comput. Math.* **19** 1145–1190. https://doi.org/10.1007/s10208-019-09427-x

[45] MA, C., PATHAK, R. and WAINWRIGHT, M. J. (2022). Optimally tackling covariate shift in RKHS-based nonparametric regression.

[46] MARCHAND, E. (1993). Estimation of a multivariate mean with constraints on the norm. *Canad. J. Statist.* **21** 359–366. https://doi.org/10.2307/3315700

[47] MARCHAND, E. and STRAWDERMAN, W. E. (2004). Estimation in restricted parameter spaces: a review. In *A festschrift for Herman Rubin*. *IMS Lecture Notes Monogr. Ser.* **45** 21–44. Inst. Math. Statist., Beachwood, OH. https://doi.org/10.1214/lnms/1196285377

[48] MELKMAN, A. A. and RITOV, Y. (1987). Minimax estimation of the mean of a general distribution when the parameter space is restricted. *Ann. Statist.* **15** 432–442. https://doi.org/10.1214/aos/1176350278 MR885749

[49] MENDELSON, S. (2015). Learning without concentration. *J. ACM* **62** Art. 21, 25. https://doi.org/10.1145/2699439

[50] MOURTADA, J. (2020). Contributions à l'apprentissage statistique : estimation de densité, agrégation d'experts et forêts aléatoires, Theses, Institut Polytechnique de Paris.

[51] MOURTADA, J. (2022). Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *Ann. Statist.* to appear.

[52] OLIVEIRA, R. I. (2016). The lower tail of random quadratic forms with applications to ordinary least squares. *Probab. Theory Related Fields* **166** 1175–1194. https://doi.org/10.1007/s00440-016-0738-9

[53] PATHAK, R., MA, C. and WAINWRIGHT, M. (2022). A new similarity measure for covariate shift with applications to nonparametric regression. In *Proceedings of the 39th International Conference on Machine Learning* (K. CHAUDHURI, S. JEGELKA, L. SONG, C. SZEPESVARI, G. NIU and S. SABATO, eds.). *Proceedings of Machine Learning Research* **162** 17517–17530. PMLR.

[54] PINSKER, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Inf. Transm.* **16** 52–68.

[55] ROBBINS, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I* 157–163. University of California Press, Berkeley-Los Angeles, Calif.

[56] SCHMIDT-HIEBER, J. and ZAMOLODTCHIKOV, P. (2022). Local convergence rates of the least squares estimator with applications to transfer learning.

[57] SIMCHOWITZ, M., AJAY, A., AGRAWAL, P. and KRISHNAMURTHY, A. (2023). Statistical Learning under Heterogenous Distribution Shift. https://doi.org/10.48550/ARXIV.2302.13934

[58] STEIN, C. (1960). Multiple regression. In *Contributions to probability and statistics* 424–443. Stanford Univ. Press, Stanford, Calif.

[59] STEINWART, I. and SCOVEL, C. (2012). Mercer's theorem on general domains: on the interaction between measures, kernels, and RKHSs. *Constr. Approx.* **35** 363–417. https://doi.org/10.1007/s00365-012-9153-3

[60] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

[61] TSYBAKOV, A. B. (2009). *Introduction to nonparametric estimation*. *Springer Series in Statistics*. Springer, New York Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. https://doi.org/10.1007/b13794 MR2724359

[62] TWEEDIE, M. C. K. (1947). Functions of a statistical variate with given means, with special reference to Laplacian distributions. *Proc. Cambridge Philos. Soc.* **43** 41–49.

[63] WANG, K. (2023). Pseudo-labeling for Kernel Ridge Regression under Covariate Shift. https://doi.org/10.48550/ARXIV.2302.10160

[64] YANG, Y., PILANCI, M. and WAINWRIGHT, M. J. (2017). Randomized sketches for kernels: fast and optimal nonparametric regression. *Ann. Statist.* **45** 991–1023. MR3662446

[65] ZAMIR, R. (1998). A proof of the Fisher information inequality via a data processing argument. *IEEE Trans. Inform. Theory* **44** 1246–1250.