Generalizing Beyond the Training Data: New Theory and Algorithms for Optimal Transfer Learning

Reese Pathak

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2025-137
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2025/EECS-2025-137.html

June 18, 2025

Generalizing Beyond the Training Data: New Theory and Algorithms for Optimal Transfer Learning

by

Reese Pathak

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael I. Jordan, Co-chair
Professor Martin J. Wainwright, Co-chair
Professor Adityanand Guntuboyina
Professor Jiantao Jiao

Summer 2025

Abstract

Generalizing Beyond the Training Data:
New Theory and Algorithms for Optimal Transfer Learning

by

Reese Pathak

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Michael I. Jordan, Co-chair

Professor Martin J. Wainwright, Co-chair

Traditional machine learning often assumes that training (source) data closely resembles the testing (target) data. However, in many contemporary applications, this is unrealistic: in e-commerce, consumer behavior is time-varying; in medicine, patient populations can exhibit more or less heterogeneity; in autonomous driving, models are rolled out to new environments. Ignoring these "distribution shifts" can lead to costly, harmful, and even dangerous outcomes. This thesis tackles these challenges by developing an algorithmic and statistical toolkit for addressing distribution shifts. Specifically, this work focuses on covariate shift, a form of distribution shift where the source and target distributions have different covariate laws.

I demonstrate that for a large class of problems, transfer learning is possible, even when the source and target data have non-overlapping support. We study covariate shift in the case of kernel classes, Hölder smoothness classes, and sparsity classes. We demonstrate how a suitably defined notion of defect or dissimilarity in the problem instance can be leveraged algorithmically, leading to methods with optimal learning guarantees.

Our final chapter contains results where we provide instance-optimal learning guarantees. We introduce a new method: penalized risk minimization with a non-traditional choice of regularization which is chosen via semidefinite programming. We show that our method has performance which is optimal with respect to the particular covariate shift instance. To our knowledge, these are the first instance-optimal guarantees for transfer learning. Moreover, our results are assumption-light: we impose essentially no restrictions on the underlying covariate laws, thereby broadening the applicability of our theory.

Contents

Chapter 1

Covariate shift in RKHS-based nonparametric regression

1.1 Introduction

A widely adopted assumption in supervised learning [118, 54] is that the training and test data are sampled from the same distribution. Such a no-distribution-shift assumption, however, is frequently violated in practice. For instance, in medical image analysis [52, 67], distribution mismatch is widespread across the hospitals due to inconsistency in medical equipment, scanning protocols, subject populations, etc. As another example, in natural language processing [61], the training data are often collected from domains with abundant labels (e.g., Wall Street Journal), while the test data may well arise from a different domain (e.g., arXiv which is mainly composed of scientific articles).

In this chapter, we focus on a special and important case of distribution mismatch, known as *covariate shift*. In this version, the marginal distributions over the input covariates may vary from the training (or source) to test (or target) data[1], while the conditional distribution of the output label given the input covariates is shared across training and testing. Motivating applications include image, text, and speech classification in which the input covariates determine the output labels [111]. Despite its importance in practice, the covariate shift problem is underexplored in theory, when compared to supervised learning without distribution mismatch—a subject that has been well studied in the past decades [54].

This chapter aims to bridge this gap by addressing several fundamental theoretical questions regarding covariate shift. First, what is the statistical limit of estimation in the presence of covariate shift? And how does this limit depend on the "amount" of covariate shift between the source and target distributions? Second, does nonparametric least-squares estimation—a dominant (and often optimal) approach in the no-distribution-shift case—achieve the optimal rate of estimation with covariate shift? If not, what is the optimal way of tackling covariate shift?

[1]Hereafter, we use source (resp. target) and training (resp. testing) interchangeably.

1.1.1 Contributions and overview

We address the aforementioned theoretical questions regarding covariate shift in the context of nonparametric regression over reproducing kernel Hilbert spaces (RKHSs) [104]. That is, we assume that under both the source and target distributions, the regression function (i.e., the conditional mean function of the output label given the input covariates) belongs to an RKHS. In this chapter, we focus on two broad families of source-target pairs depending on the configuration of the likelihood ratios between them.

We first consider the uniformly B-bounded family in which the likelihood ratios are uniformly bounded by a quantity B. In this case, we present general performance upper bounds for the kernel ridge regression (KRR) estimator in Theorem 1.1. Instantiations of this general bound on various RKHSs with regular eigenvalues are provided in Corollary 1.1. It is also shown in Theorem 1.2 that KRR—with an optimally chosen regularization parameter that depends on the largest likelihood ratio B—achieves the minimax lower bound for covariate shift over this uniformly B-bounded family. It is worth noting that the optimal regularization parameter shrinks as the likelihood ratio bound increases.

We further show—via a constructive argument—that the nonparametric least-squares estimator, which minimizes the empirical risk on the training data over the specified RKHS, falls short of achieving the lower bound; see Theorem 1.3. This marks a departure from the classical no-covariate-shift setting, where the constrained estimator (i.e., the nonparametric least-squares estimator) and the regularized estimator (i.e., the KRR estimator) can both attain optimal rates of estimation [120]. In essence, the failure arises from the misalignment between the projections under the source and target covariate distributions. Loosely speaking, nonparametric least-squares estimation projects the data onto an RKHS according to the geometry induced by the *source* distribution. Under covariate shift, the resulting projection can be extremely far away from the projection under the *target* covariate distributions.

In the second part of the chapter, we turn to a more general setting, where the likelihood ratios between the target and source distributions may not be bounded. Instead, we only require the target and source covariate distributions to have a likelihood ratio with bounded second moment. We propose a variant of KRR that weights samples based on a careful truncation of the likelihood ratios. We are able to show in Theorem 1.4 that this estimator is rate-optimal over this larger class of covariate shift problems.

1.1.2 Related work

There is a large body of work on distribution mismatch and, in particular, on covariate shift. Below we review the work that is directly relevant to ours, and refer the interested reader to the book [111] and the survey [93] for additional references.

Shimodaira [105] first studied the covariate shift problem from a statistical perspective, and established the asymptotic consistency of the importance-reweighted maximum likelihood estimator (without truncation). However, no finite-sample guarantees were provided therein. Similar to our work, Cortes and coauthors [27] analyzed the importance-reweighted

estimator when the density ratio is either bounded or has a finite second moment. However, their analysis applies to the function class with finite pseudodimension (cf. the book [97]), while the RKHS considered herein does not necessarily obey this assumption. Moreover, even when the RKHS has a finite rank D, their result (e.g., Theorem 8) is sub-optimal—with a rate of $\sqrt{V^2 D/n}$ compared to our optimal rate $V^2 D/n$. Here V^2 is the bound on the second moment of the likelihood ratios and n denotes the number of samples. Recently, Kpotufe and Martinet [69] investigated covariate shift for nonparametric classification. They proposed a novel notion called transfer exponent to measure the amount of covariate shift between the source and target distributions. An estimator based on k nearest neighbors was shown to be minimax optimal over the class of covariate shift problems with bounded transfer exponent. Inspired by the work of Kpotufe and Martinet, the current authors [94] proposed a more fine-grained similarity measure for covariate shift and applied to nonparametric regression over the class of Hölder continuous functions. It is worth pointing out that both the transfer exponent and the new fine-grained similarity measure are different and cannot directly be compared to the moment conditions we impose on the likelihood ratios in this work. In particular, there exist instances of covariate shift where the second moment of the likelihood ratios is bounded whereas the transfer exponent is infinite. One such case is when the source and target distributions are both Gaussian with the same mean but different scales. Another significant difference lies in the assumptions on the regression function. Kpotufe and Martinet [69] and Pathak et al. [94] focused on the class of Hölder continuous functions, while we focus on RKHSs. This leads to drastically different optimal estimators. Schmidt-Hieber and Zamolodtchikov [103] recently established the local convergence of the nonparametric least-squares estimator for the specific class of 1-Lipschitz functions over the unit interval $[0, 1]$ and applied it to the covariate shift setting.

Apart from covariate shift, other forms of distribution mismatch have been analyzed from a statistical perspective. Cai et al. [19] analyzed posterior shift and proposed an optimal k-nearest-neighbor estimator. Maity et al. [78] conducted the minimax analysis for the label shift problem. Recently, Reeve et al. [100] studied the general distribution shift problem (also known as transfer learning) which allows both covariate shift and posterior shift.

Notation. Throughout the chapter, we use c, c', c_1, c_2 to denote universal constants, which may vary from line to line.

1.2 Background and problem formulation

In this section, we formulate and provide background on the problem of covariate shift in nonparametric regression.

1.2.1 Nonparametric regression under covariate shift

The goal of nonparametric regression is to predict a real-valued response Y based on a vector of covariates $X \in \mathcal{X}$. For each fixed x, the optimal estimator in a mean-squared sense is given by the regression function $f^\star(x) := \mathbf{E}[Y \mid X = x]$. In a typical setting, we assume observations of n i.i.d. pairs $\{(x_i, y_i)\}_{i=1}^n$, where each x_i is drawn according to some distribution P over $\mathcal{X}$, and then y_i is drawn according to the law $(Y \mid X = x_i)$. We assume throughout that for each i, the residual $w_i := y_i - f^\star(x_i)$ is a sub-Gaussian random variable with variance proxy σ^2.

We refer to the distribution P over the covariate space as the *source distribution*. In the standard set-up, the performance of an estimator $\widehat{f}$ is measured according to its $L^2(P)$-error:

$$\|\widehat{f} - f^\star\|_P^2 := \mathbf{E}_{X \sim P}\left(\widehat{f}(X) - f^\star(X)\right)^2 \;=\; \int_{\mathcal{X}} \left(\widehat{f}(x) - f^\star(x)\right)^2 p(x)dx,$$

where p is the density of P.

In the covariate shift version of this problem, we have a different goal—that is, we wish to construct an estimate $\widehat{f}$ whose $L^2(Q)$-error is small. Here the *target distribution* Q is different from the source distribution P. In analytical terms, letting q be the density of Q, our goal is to find estimators $\widehat{f}$ such that

$$\|\widehat{f} - f^\star\|_Q^2 = \mathbf{E}_{X \sim Q}\left(\widehat{f}(X) - f^\star(X)\right)^2 \;=\; \int_{\mathcal{X}} \left(\widehat{f}(x) - f^\star(x)\right)^2 q(x)dx$$

is as small as possible. Clearly, the difficulty of this problem should depend on some notion of discrepancy between the source and target distributions.

1.2.2 Conditions on source-target likelihood ratios

The discrepancy between the $L^2(P)$ and $L^2(Q)$ norms is controlled by the *likelihood ratio*

$$\rho(x) := \frac{q(x)}{p(x)},$$

which we assume exists for any $x \in \mathcal{X}$. By imposing different conditions on the likelihood ratio, we can define different families of source-target pairs (P, Q). In this chapter, we focus on two broad families of such pairs.

Uniformly B-bounded families: For a quantity $B \geqslant 1$, we say that the likelihood ratio is *B-bounded* if

$$\sup_{x \in \mathcal{X}} \rho(x) \leqslant B. \tag{1.2}$$

It is worth noting that $B = 1$ recovers the case without covariate shift, i.e., $P = Q$. Our analysis in Section 1.3 works under this condition.

χ^2-bounded families: A uniform bound on the likelihood ratio is a stringent condition, so that it is natural to relax it. One relaxation is to instead bound the second moment: in particular, for a scalar $V^2 \geqslant 1$, we say that the likelihood ratio is V^2-*moment bounded* if

$$\mathbf{E}_{X\sim P}[\rho^2(X)] \leqslant V^2. \tag{1.3}$$

Note that when the uniform bound (1.2) holds, the moment bound (1.3) holds with $V^2 = B$. To see this, we can write $\mathbf{E}_{X\sim P}[\rho^2(X)] = \mathbf{E}_{X\sim Q}[\rho(X)] \leqslant B$. However, the moment bound (1.3) is much weaker in general. It is also worth noting that the χ^2-divergence between Q and P takes the form

$$\chi^2(Q||P) = \mathbf{E}_{X\sim P}[\rho^2(X)] - 1.$$

Therefore, one can understand the quantity $V^2 - 1$ as an upper bound on the χ^2-divergence between Q and P. Our analysis in Section 1.4 applies under this weaker condition on the likelihood ratio.

1.2.3 Unweighted versus likelihood-reweighted estimators

In this chapter, we focus on methods for nonparametric regression that are based on optimizing over a Hilbert space $\mathcal{H}$ defined by a reproducing kernel. The Hilbert norm $\|f\|_{\mathcal{H}}$ is used as a means of enforcing "smoothness" on the solution, either by adding a penalty to the objective function or via an explicit constraint.

In the absence of any knowledge of the likelihood ratio, a naïve approach is to simply compute the *unweighted regularized estimate*

$$\widehat{f}_\lambda := \arg\min_{f\in\mathcal{H}} \Big\{ \frac{1}{n}\sum_{i=1}^n (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \Big\}, \tag{1.4}$$

where $\lambda > 0$ is a user-defined regularization parameter. When $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), then this estimate is known as the *kernel ridge regression* (KRR) estimate. This is a form of empirical risk minimization, but in the presence of covariate shift, the objective involves an empirical approximation to $\mathbf{E}_P[(Y - f(X))^2]$, as opposed to $\mathbf{E}_Q[(Y - f(X))^2]$.

If the likelihood ratio were known, then a natural proposal is to instead compute the *likelihood-reweighted regularized estimate*

$$\widetilde{f}_\lambda^{\mathrm{rw}} := \arg\min_{f\in\mathcal{H}} \Big\{ \frac{1}{n}\sum_{i=1}^n \rho(x_i)(f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \Big\}. \tag{1.5}$$

The introduction of the likelihood ratio ensures that the objective now provides an unbiased estimate of the expectation $\mathbf{E}_Q[(Y - f(X))^2]$. However, reweighting by the likelihood ratio also increases variance, especially in the case of unbounded likelihood ratios. Accordingly, in Section 1.4, we study a suitably truncated form of the estimator (1.5).

1.2.4 Kernels and their eigenvalues

Any reproducing kernel Hilbert space is associated with a positive semidefinite kernel function $\mathscr{K} : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$. Under mild regularity conditions, Mercer's theorem guarantees that this kernel has an eigen-expansion of the form

$$\mathscr{K}(x, x') := \sum_{j=1}^{\infty} \mu_j \phi_j(x) \phi_j(x') \tag{1.6}$$

for a sequence of non-negative eigenvalues $\{\mu_j\}_{j \geqslant 1}$, and eigenfunctions $\{\phi_j\}_{j \geqslant 1}$ taken to be orthonormal in $L^2(Q)$. Given our goal of deriving bounds in the $L^2(Q)$-norm, it is appropriate to expand the kernel in $L^2(Q)$, as we have done here (1.6), in order to assess the richness of the function class.

Given the Mercer expansion, the squared norm in the reproducing kernel Hilbert space takes the form

$$\|f\|_{\mathcal{H}}^2 = \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j}, \qquad \text{where } \theta_j := \textstyle\int_{\mathcal{X}} f(x) \phi_j(x) q(x) dx.$$

Consequently, the Hilbert space itself can be written as

$$\mathcal{H} := \left\{ f = \sum_{j=1}^{\infty} \theta_j \phi_j \;\; | \;\; \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} < \infty \right\}.$$

Our goal is to understand the performance of nonparametric regression under covariate shift when the regression function lies in $\mathcal{H}$.

Throughout this chapter, we impose a standard boundedness condition on the kernel function—namely, there exists some finite $\kappa > 0$ such that

$$\sup_{x \in \mathcal{X}} \mathscr{K}(x, x) \leqslant \kappa^2. \tag{1.7}$$

Note that any continuous kernel over a compact domain satisfies this condition. Moreover, a variety of commonly used kernels, including the Gaussian and Laplacian kernels, satisfy this condition over any domain.

1.3 Analysis for bounded likelihood ratios

We begin our analysis in the case of bounded likelihood ratios. Our first main result is to prove an upper bound on the performance of the unweighted KRR estimate (1.4). First, we prove a family of upper bounds (Theorem 1.1) depending on the regularization parameter λ. By choosing λ so as to minimize this family of upper bounds, we obtain concrete results for different classes of kernels (Corollary 1.1). We then turn to the complementary question

of lower bounds: in Theorem 1.2, we prove a family of lower bounds that establish that for covariate shift with B-bounded likelihood ratios, the KRR estimator is minimax-optimal up to logarithmic factors in the sample size. This optimality guarantee is notable since it applies to the unweighted estimator that does not involve full knowledge of the likelihood ratio (apart from an upper bound).

In the absence of covariate shift, it is well-known that performing empirical risk minimization with an explicit constraint on the function also leads to minimax-optimal results. Indeed, without covariate shift, projecting an estimate onto a convex constraint set containing the true function can never lead to a worse result. In Theorem 1.3, we show that this natural property is no longer true under covariate shift: performing empirical risk minimization over the smallest Hilbert ball containing $f^\star$ can be sub-optimal. Optimal procedures—such as the regularized KRR estimate—are actually operating over Hilbert balls with radius substantially larger than the true norm $\|f^\star\|_{\mathcal{H}}$.

1.3.1 Unweighted kernel ridge regression is near-optimal

We begin by deriving a family of upper bounds on the kernel ridge regression estimator (1.4) under covariate shift. In conjunction with our later analysis, these bounds will establish that the KRR estimate is minimax-optimal up to logarithmic factors for covariate shift with bounded likelihood ratios.

Theorem 1.1. *Consider a covariate-shifted regression problem with likelihood ratio that is B-bounded* (1.2) *over a Hilbert space with a κ-uniformly bounded kernel* (1.7)*. Then for any $\lambda \geqslant 10\kappa^2/n$, the KRR estimate $\widehat{f}_\lambda$ satisfies the bound*

$$\|\widehat{f}_\lambda - f^\star\|_Q^2 \leqslant \underbrace{4\lambda B \|f^\star\|_{\mathcal{H}}^2}_{\mathbf{b}_\lambda^2(B)} + \underbrace{80\sigma^2 B \frac{\log n}{n} \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda B}}_{\mathbf{v}_\lambda(B)} \tag{1.8}$$

with probability at least $1 - 28\,\frac{\kappa^2}{\lambda} e^{-\frac{n\lambda}{16\kappa^2}} - \frac{1}{n^{10}}$.

See Section 1.5.1 for the proof of this theorem. In Section 1.7.5.1, we also present a corollary which provides a corresponding expectation bound for the KRR estimator $\widehat{f}_\lambda$ for such B-bounded covariate shifts.

Note that the upper bound (1.8) involves two terms. The first term $\mathbf{b}_\lambda^2(B)$ corresponds to the squared bias of the KRR estimate, and it grows proportionally with the regularization parameter λ and the likelihood ratio bound B. The second term $\mathbf{v}_\lambda(B)$ represents the variance of the KRR estimator, and it shrinks as λ increases, so that λ controls the bias-variance trade-off. This type of trade-off is standard in nonparametric regression; what is novel of interest to us here is how the shapes of these trade-off curves change as a function of the likelihood ratio bound B.

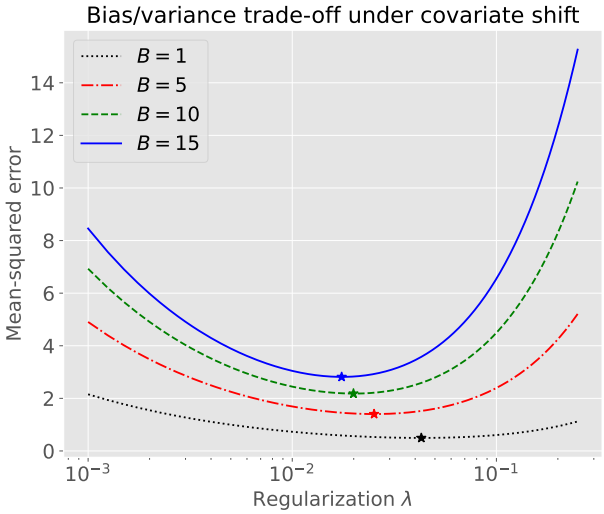


Figure 1.1. Plot of the upper bound (1.8) on the mean-squared error versus the log regularization parameter $\log \lambda$ for four different choices of the likelihood ratio bound B, in all cases with eigenvalues $\mu_j = (1/j)^2$, noise variance $\sigma^2 = 1$ and sample size $n = 8000$. The points marked with $\star$ on each curve corresponds to the choice of $\lambda^*(B)$ that minimizes the upper bound. Note how this minimizing value shifts to the left as B increases above the standard problem without covariate shift ($B = 1$).

Figure 1.1 plots the right-hand side of the upper bound (1.8) as a function of λ for several different choices $B \in \{1, 5, 10, 15\}$. (In all cases, we fixed a kernel with eigenvalues decaying as $\mu_j = j^{-2}$, sample size $n = 8000$, and noise variance $\sigma^2 = 1$.) Of interest to us is the choice $\lambda^*(B)$ that minimizes this upper bound; note how this optimizing $\lambda^*(B)$ shifts leftwards to smaller values as B is increased.

We would like to understand the balancing procedure that leads to an optimal $\lambda^*(B)$ in analytical terms. This balancing procedure is most easily understood for kernels with *regular eigenvalues*, a notion introduced in past work [123] on kernel ridge regression. For a given targeted error level $\delta > 0$, it is natural to consider the first index $d(\delta)$ for which the associated eigenvalue drops below δ^2—that is, $d(\delta) := \min\{j \geqslant 1 \mid \mu_j \leqslant \delta^2\}$. The eigenvalue sequence is said to be regular if[2]

$$\sum_{j=d(\delta)+1}^{\infty} \mu_j \leqslant c\, d(\delta)\delta^2 \tag{1.9}$$

holds for some universal constant $c > 0$. The class of kernels with regular eigenvalues includes kernels of finite-rank and those with various forms of polynomial or exponential decay in their eigenvalues; all are widely used in practice. For kernels with regular eigenvalues, the

[2]In fact, we can relax this to only require the minimizing δ in equation (1.10) to obey the tail bound.

bound (1.8) implies that there is a universal constant c' such that

$$\|\widehat{f}_\lambda - f^\star\|_Q^2 \leqslant c'\Big\{\delta^2\|f^\star\|_{\mathcal{H}}^2 + \sigma^2 B\frac{d(\delta)\log n}{n}\Big\} \qquad \text{where } \delta^2 = \lambda B. \tag{1.10}$$

We verify this claim as part of proving Corollary 1.1 below.

This bound (1.10) enables us to make (near)-optimal choices of δ—and hence $\lambda = \delta^2/B$. Let us summarize the outcome of this procedure for a few kernels of interest. In particular, we say that a kernel has *finite rank* D if the eigenvalues $\mu_j = 0$ for all $j > D$. The kernels that underlie linear regression and polynomial regression more generally are of this type. A richer family of kernels has eigenvalues that exhibit *α-polynomial decay* $\mu_j \leqslant c\, j^{-2\alpha}$ for some $\alpha > 1/2$. This kind of eigenvalue decay is seen in various types of spline and Sobolev kernels, as well as the Laplacian kernel. It is easy to verify that both of these families have regular eigenvalues. To simplify the presentation, we assume $\|f^\star\|_{\mathcal{H}} = 1$.

Corollary 1.1 (Bounds for specific kernels)**.** *We have the following bounds for specific kernel eigenvalues.*

(a) For a kernel with rank D, as long as $\sigma^2 D\log n \geqslant 10\kappa^2$, the choice $\lambda = \frac{\sigma^2 D\log n}{n}$ yields an estimate $\widehat{f}_\lambda$ such that

$$\|\widehat{f}_\lambda - f^\star\|_Q^2 \leqslant c\sigma^2 B\frac{D\log n}{n} \tag{1.11a}$$

with high probability.

(b) For a kernel with α-decaying eigenvalues, suppose that σ^2 is sufficiently large so that $\lambda = B^{-\frac{1}{2\alpha+1}}\big(\frac{\sigma^2\log n}{n}\big)^{\frac{2\alpha}{2\alpha+1}} \geqslant 10\kappa^2/n$. Then the estimate $\widehat{f}_\lambda$ obeys

$$\|\widehat{f}_\lambda - f^\star\|_Q^2 \leqslant c\Big(\frac{\sigma^2 B\log n}{n}\Big)^{\frac{2\alpha}{2\alpha+1}} \tag{1.11b}$$

with high probability.

Proof. We begin by proving the upper bound (1.10). With the shorthand $\delta^2 = \lambda B$, the variance term in our bound (1.8) can be bounded as

$$\frac{1}{80}\mathbf{v}_\lambda(B) = \sigma^2 B\frac{\log n}{n}\sum_{j=1}^{\infty}\frac{\mu_j}{\mu_j+\delta^2} \leqslant \sigma^2 B\frac{\log n}{n}\Big\{\sum_{j=1}^{d(\delta)}1 + \sum_{j>d(\delta)}^{\infty}\frac{\mu_j}{\mu_j+\delta^2}\Big\},$$

where, by the definition of $d(\delta)$, we have split the eigenvalues into those that are larger than δ^2, and those that are smaller than δ^2. By the definition of a regular kernel, the second term can be upper bounded

$$\sum_{j>d(\delta)}^{\infty}\frac{\mu_j}{\mu_j+\delta^2} \leqslant \frac{1}{\delta^2}c'\,d(\delta)\delta^2 \;=\; c'\,d(\delta).$$

Putting together the pieces yields $\frac{1}{80}\mathbf{v}_\lambda(B) \leqslant c_2\sigma^2 B\frac{\log n}{n}d(\delta)$, for some universal constant c_2. Combining with the bias term yields the claim (1.10).

We now prove claims (1.11a) and (1.11b). For a finite-rank kernel, using the fact that $d(\delta) \leqslant D$ for any $\delta > 0$, we can set $\lambda = \frac{\sigma^2 D\log n}{n}$ to obtain the claimed bound (1.11a). Now suppose that the kernel has α-polynomial decay—that is, $\mu_j \leqslant cj^{-2\alpha}$ for some $c > 0$. For any $\delta > 0$, we then have $d(\delta) \leqslant c'\,(1/\delta)^{1/\alpha}$, and hence

$$\delta^2 + \sigma^2 B\frac{d(\delta)\log n}{n} \leqslant \delta^2 + c'\sigma^2 B\frac{\log n}{n}\big(\frac{1}{\delta}\big)^{1/\alpha}.$$

By equating the two terms, we can solve for near-optimal δ: in particular, we set $\delta^2 = \left(\frac{\sigma^2 B\log n}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$ to obtain the claimed result. Notice that this choice of δ^2 corresponds to

$$\lambda = \delta^2/B \;=\; B^{-\frac{1}{2\alpha+1}}\,\big(\frac{\sigma^2\log n}{n}\big)^{\frac{2\alpha}{2\alpha+1}},$$

as claimed in the corollary. □

1.3.2 Lower bounds with covariate shift for regular kernels

Thus far, we have established a family of upper bounds on the unweighted KRR estimate, and derived concrete results for various classes of regular kernels. We now establish that, for the class of regular eigenvalues, the bounds achieved by the unweighted KRR estimator are minimax-optimal. Recall the definition $d(\delta) = \min\{j \geqslant 1 \mid \mu_j \leqslant \delta^2\}$, and the notion of regular eigenvalues (1.9). For a Hilbert space $\mathcal{H}$, we let $\mathcal{B}_{\mathcal{H}}(1)$ denote the Hilbert norm ball of radius one.

Theorem 1.2. *For any $B \geqslant 1$, there exists a pair (P, Q) with B-bounded likelihood ratio* (1.2) *and an orthonormal basis $\{\phi_j\}_{j\geqslant 1}$ of $L^2(Q)$ such that for any regular sequence of kernel eigenvalues $\{\mu_j\}_{j\geqslant 1}$, we have*

$$\inf_{\widehat{f}}\sup_{f^\star\in\mathcal{B}_{\mathcal{H}}(1)}\mathbf{E}[\|\widehat{f} - f^\star\|_Q^2] \geqslant c\,\inf_{\delta>0}\Big\{\delta^2 + \sigma^2 B\frac{d(\delta)}{n}\Big\}. \tag{1.12}$$

See Section 1.7.1 for the proof of this claim.

Comparing the lower bound (1.12) to our achievable result (1.10) for the unweighted KRR estimate, we see that—with an appropriate choice of the regularization parameter λ—the KRR estimator is minimax optimal up to a $\log n$ term. In particular, it is straightforward to derive the following consequences of Theorem 1.2, which parallel the guarantees in Corollary 1.1:

- For a finite-rank kernel, the minimax risk for B-bounded covariate shift satisfies the lower bound

$$\inf_{\widehat{f}}\sup_{f^\star\in\mathcal{B}_{\mathcal{H}}(1)}\mathbf{E}[\|\widehat{f} - f^\star\|_Q^2] \geqslant c\,\sigma^2 B\frac{D}{n}.$$

- For a kernel with α-polynomial eigenvalues, the minimax risk for B-bounded covariate shift satisfies the lower bound

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{B}_{\mathcal{H}}(1)} \mathbf{E}[\|\widehat{f} - f^\star\|_Q^2] \geqslant c \left(\frac{\sigma^2 B}{n}\right)^{\frac{2\alpha}{2\alpha+1}}.$$

Note that both of these minimax lower bounds reduce to the known lower bounds [123] in the case of no covariate shift (i.e., $B = 1$).

1.3.3 Constrained kernel regression is sub-optimal

In the absence of covariate shift, procedures based on empirical risk minimization with explicit constraints are also known to be minimax-optimal. In the current setting, one such estimator is the constrained kernel regression estimate

$$\widehat{f}_{\text{erm}} := \arg \min_{f \in \mathcal{B}_{\mathcal{H}}(1)} \Big\{ \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 \Big\}. \tag{1.13}$$

Without covariate shift and for any regular kernel, this constrained empirical risk minimization procedure is minimax-optimal over all functions $f^\star$ with $\|f^\star\|_{\mathcal{H}} \leqslant 1$.

In the presence of covariate shift, this minimax-optimality turns out to be false. In particular, suppose that the eigenvalues decay as $\mu_j = (1/j)^2$, so that our previous results show that the minimax risk for B-bounded likelihood ratios scales as $\left(\frac{B\sigma^2}{n}\right)^{2/3}$. It turns out that there exists B-bounded pair (P, Q) and an associated kernel class with the prescribed eigendecay for which the constrained estimator (1.13) is sub-optimal for a broad range of (B, n) pairs. In the following statement, we use c_1, c_2 to denote universal constants.

Theorem 1.3. *Assume that $\|f^\star\|_{\mathcal{H}} = 1$, and that $\sigma^2 = 1$. For any $B \in [c_1(\log n)^2, c_2 n^{2/3}]$, there exists a B-bounded pair (P, Q) and RKHS with eigenvalues $\mu_j \leqslant (1/j^2)$ such that*

$$\sup_{f^\star \in \mathcal{B}_{\mathcal{H}}(1)} \mathbf{E}\Big[\|\widehat{f}_{\text{erm}} - f^\star\|_Q^2\Big] \geqslant c_3 \frac{B^3}{n^2}. \tag{1.14}$$

See Section 1.7.2 for the proof of this negative result.

In order to appreciate some implications of this theorem, suppose that we use it to construct ensembles with $B_n \asymp n^{2/3}$. The lower bound (1.14) then implies that over this sequence of problems, the maximal risk of $\widehat{f}_{\text{erm}}$ is bounded below by a universal constant. On the other hand, if we apply the unweighted KRR procedure, then we obtain consistent estimates, in particular with $L^2(Q)$-error that decays as

$$\left(\frac{B_n}{n}\right)^{2/3} = \left(\frac{n^{2/3}}{n}\right)^{2/3} = n^{-2/9}.$$

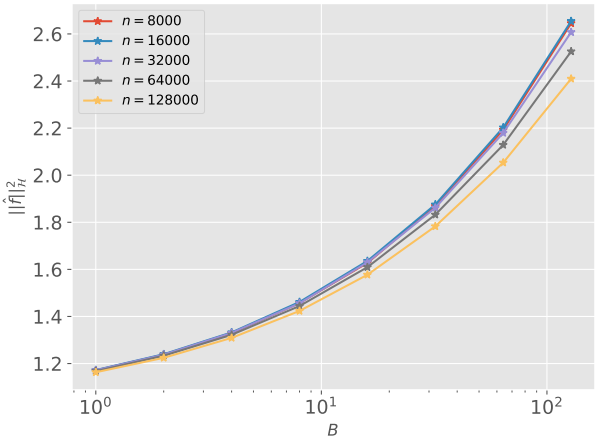


Figure 1.2. Results based on computing the regularized KRR estimate for the "bad" problems, indexed by the pair (n, B), that underlie the proof of Theorem 1.3. Each curve shows the squared Hilbert norm of the regularized KRR estimate $\|\widehat{f}_\lambda\|_{\mathcal{H}}^2$, computed with $\lambda = 4^{2/3} n^{-2/3} B^{-1/3}$, versus the likelihood ratio bound B. Each curve corresponds to a different choice of sample size n as indicated in the legend.

It is worth understanding why the constrained form of KRR is sub-optimal, while the regularized form is minimax-optimal. Recall from Corollary 1.1 that achieving minimax-optimal rates with KRR requires particular choices of the regularization parameter $\lambda^*(B)$, ones that decrease as B increases (see Figure 1.1). This behavior suggests that the Hilbert norm $\|\widehat{f}_\lambda\|_{\mathcal{H}}$ of the regularized KRR estimate with optimal choice of λ should grow significantly above $\|f^\star\|_{\mathcal{H}} = 1$ when we apply this method.

In order to confirm this intuition, we performed some illustrative simulations over the ensembles, indexed by the pair (B, n), that underlie the proof of Theorem 1.3; see Section 1.7.2 for the details. With $\sigma^2 = 1$ remaining fixed, for each given pair (B, n), we simulated regularized kernel ridge regression with the choice $\lambda = 4^{2/3} n^{-2/3} B^{-1/3}$, as suggested by Corollary 1.1. In Figure 1.2, for each fixed n, we plot the squared Hilbert norm $\|\widehat{f}_\lambda\|_{\mathcal{H}}^2$ of the regularized KRR estimate versus the parameter B. We vary the choice of sample size $n \in \{8000, 16000, 32000, 64000, 128000\}$, as indicated in the figure legend. In all of these curves, we see that the squared Hilbert norm is increasing as a polynomial function of B. This behavior is to be expected, given the sub-optimality of the constrained KRR estimate with a fixed radius.

1.4 Unbounded likelihood ratios

Thus far, our analysis imposed the B-bound (1.2) on the likelihood ratio. In practice, however, it is often the case that the likelihood ratio is unbounded. As a simple univariate example, suppose that the target distribution Q is standard normal $\mathsf{N}(0, 1)$, whereas the

source distribution P takes the form $\mathsf{N}(0, 0.9)$. It is easy to see that the likelihood ratio $\rho(x)$ tends to ∞ as $|x| \to \infty$. On the other hand, the second moment of the likelihood ratio under P remains bounded, so that χ^2-condition (1.3) applies.

The key to the success of the *unweighted* KRR estimator (1.4) in the bounded likelihood ratio case is the nice relationship between the covariance $\Sigma_P := \mathbf{E}_{X\sim P}[\phi(X)\phi(X)^\mathsf{T}]$ of the source distribution and the covariance I of the target distribution, namely $\Sigma_P \succcurlyeq \frac{1}{B} I$. In contrast, such a nice relationship (with B replaced by V^2) does not appear to hold with unbounded likelihood ratios. It is therefore natural to consider the likelihood-reweighted estimate (1.5), as previously introduced in Section 1.2.3, that ensures the nice identity $\mathbf{E}_{X\sim P}[\rho(X)\phi(X)\phi(X)^\mathsf{T}] = I$. In contrast to the unweighted KRR estimator, it requires knowledge of the likelihood ratio, but we will see that—when combined with a suitable form of truncation—it achieves minimax-optimal rates (up to logarithmic factors) over the much larger classes of χ^2-bounded source-target pairs.

As noted before, one concern with likelihood-reweighted estimators is that they can lead to substantial inflation of the variance, in particular due to the multiplication by the potentially unbounded quantity $\rho(x)$. For this reason, it is natural to consider truncation: more precisely, for a given $\tau_n > 0$, we define the *truncated likelihood ratio*

$$\rho_{\tau_n}(x) := \begin{cases} \rho(x), & \text{if } \rho(x) \leqslant \tau_n, \\ \tau_n, & \text{otherwise.} \end{cases}$$

We then consider the family of estimators

$$\widehat{f}^{\mathrm{rw}}_\lambda := \arg\min_{f\in\mathcal{H}} \Big\{ \frac{1}{n}\sum_{i=1}^n \rho_{\tau_n}(x_i)(f(x_i) - y_i)^2 + \lambda \|f\|^2_{\mathcal{H}} \Big\}, \tag{1.15}$$

where $\lambda > 0$, along with the truncation level τ_n, are parameters to be specified.

We analyze the behavior of this estimator for kernels whose eigenfunctions are 1-uniformly bounded in sup-norm, meaning that

$$\|\phi_j\|_\infty := \sup_{x\in\mathcal{X}} |\phi_j(x)| \leqslant 1 \quad \text{for all } j = 1, 2, \ldots. \tag{1.16}$$

Our choice of the constant 1 is for notational simplicity. Although there exist kernels whose eigenfunctions are not uniformly bounded, there are many kernels for which this condition does hold. Whenever the domain $\mathcal{X}$ is compact and the eigenfunctions are continuous, this condition will hold. Another class of examples is given by convolutional kernels (i.e., kernels of the form $\mathscr{K}(x, z) = \Psi(x - z)$ for some $\Psi : \mathcal{X} \to \mathbf{R}$), which have sinusoids as their eigenfunctions, and thus satisfy this condition.

Our theorem on the truncated-reweighted KRR estimate (1.15) involves the kernel complexity function $\Psi(\delta, \mu) := \sum_{j=1}^\infty \min\{\delta^2, \mu_j \|f^\star\|^2_{\mathcal{H}}\}$, and works for any solution $\delta_n > 0$ to the inequality $\mathcal{M}(\delta) \leqslant \delta^2/2$, where

$$\mathcal{M}(\delta) := c_0 \sqrt{\tfrac{\sigma^2 V^2 \log^3(n)}{n} \Psi(\delta, \mu)}.$$

Here c_0 is a universal constant, whose value is specified via the proof.

Below, we present the performance guarantee of $\widehat{f}_\lambda^{\mathrm{rw}}$ in the large noise regime (i.e., when $\sigma^2 \geqslant \kappa^2 \|f^\star\|_{\mathcal{H}}^2$) to simplify the statement. Theoretical guarantees for all ranges of σ^2 can be found in Section 1.7.4.

Theorem 1.4. *Consider a kernel with sup-norm bounded eigenfunctions* (1.16)*, and a source-target pair with* $\mathbf{E}_P[\rho^2(X)] \leqslant V^2$*. Further assume that the noise level obeys* $\sigma^2 \geqslant \kappa^2 \|f^\star\|_{\mathcal{H}}^2$*. Then the estimate* $\widehat{f}_\lambda^{\mathrm{rw}}$ *with truncation* $\tau_n = \sqrt{nV^2}$ *and regularization* $\lambda \|f^\star\|_{\mathcal{H}}^2 \geqslant \delta_n^2/3$ *satisfies the bound*

$$\|\widehat{f}_\lambda^{\mathrm{rw}} - f^\star\|_Q^2 \leqslant \delta_n^2 + 3\lambda \|f^\star\|_{\mathcal{H}}^2$$

with probability at least $1 - cn^{-10}$*. Here, we recall that* $\delta_n > 0$ *is any solution to the inequality* $\mathcal{M}(\delta) \leqslant \delta^2/2$*, where*

$$\mathcal{M}(\delta) = c_0 \sqrt{\tfrac{\sigma^2 V^2 \log^3(n)}{n} \Psi(\delta, \mu)}.$$

See Section 1.5.2 for the proof of this claim. In Section 1.7.5.3, we also present a corollary which provides a corresponding expectation bound for the reweighted estimator $\widehat{f}_\lambda^{\mathrm{rw}}$ for such V^2-bounded covariate shifts.

To appreciate the connection to our previous analysis, in the proof of Corollary 1.2 below, we show that for any regular sequence of eigenvalues and $\|f^\star\|_{\mathcal{H}} = 1$, we have

$$\Psi(\delta, \mu) \leqslant c' \, d(\delta) \delta^2 \tag{1.17}$$

for some universal constant c'. Moreover, the condition $\mathcal{M}(\delta) \leqslant \delta^2/2$ can be verified by checking the inequality

$$\sqrt{\frac{\sigma^2 V^2 \log^3(n)}{n} d(\delta)} \leqslant c_1 \delta. \tag{1.18}$$

This further allows us to obtain the rates of estimation over specific kernel classes.

Corollary 1.2. *Consider kernels with sup-norm bounded eigenfunctions* (1.16)*.*

(a) For a kernel with rank D, the truncated-reweighted estimator with $\lambda = c \frac{DV^2 \log^3(n) \sigma^2}{n}$ *achieves*

$$\|\widehat{f}_\lambda^{\mathrm{rw}} - f^\star\|_Q^2 \leqslant c' \frac{DV^2 \log^3(n) \sigma^2}{n} \tag{1.19}$$

with high probability.

(b) For a kernel with α-polynomial eigenvalues, we have with high probability

$$\|\widehat{f}_\lambda^{\mathrm{rw}} - f^\star\|_Q^2 \leqslant c'\Big(\frac{V^2\log^3(n)}{n}\sigma^2\Big)^{\frac{2\alpha}{2\alpha+1}}, \tag{1.20}$$

provided that $\lambda = c\big(\frac{V^2\log^3(n)}{n}\sigma^2\big)^{\frac{2\alpha}{2\alpha+1}}$.

Proof. We begin by verifying the claim (1.17). Recalling the definition of $d(\delta)$ as the smallest integer for which $\mu_j \leqslant \delta^2$, we can write

$$\Psi(\delta,\mu) = \sum_{j=1}^{d(\delta)} \min\{\delta^2,\mu_j\} + \sum_{j=d(\delta)+1}^{\infty} \min\{\delta^2,\mu_j\} \;\leqslant\; d(\delta)\delta^2 + cd(\delta)\delta^2$$

where the bound on the second sum follows from the regularity condition. This completes the proof of the bound (1.17).

Given our bound (1.17), it is straightforward to verify the claim (1.18).

We now prove the bounds (1.19) and (1.20). For the finite rank case, we have $\Psi(\delta,\mu) \leqslant D\delta^2$, which implies $\delta_n^2 \leqslant c\frac{DV^2\log^3(n)\sigma^2}{n}$ for some universal constant c. Apply Theorem 1.4 to obtain the desired rate. Now we move on to the kernel with α-polynomial eigenvalues. We know from the proof of Corollary 1.1 that $d(\delta) \leqslant c(1/\delta)^{1/\alpha}$, and hence $\Psi(\delta,\mu) \leqslant c'\delta^{2-1/\alpha}$. This implies $\delta_n^2 \leqslant c\Big(\frac{V^2\log^3 n}{n}\sigma^2\Big)^{\frac{2\alpha}{2\alpha+1}}$, which together with Theorem 1.4 yields the claim. □

Corollary 1.2 showcases that the reweighted KRR estimator is minimax optimal (up to log factors) over this more general χ^2-bounded family. This can be seen from the lower bound established in Theorem 1.2 and the fact that the χ^2-bounded family is a larger family compared to the uniformly bounded family.

1.5 Proofs

In this section, we provide the proofs of our two sets of upper bounds on different estimators. Section 1.5.1 is devoted to the proof of Theorem 1.1 on upper bounds on unweighted KRR for B-bounded likelihood ratios, whereas Section 1.5.2 is devoted to the proof of Theorem 1.4 on the performance of LR-reweighted KRR with truncation.

1.5.1 Proof of Theorem 1.1

Define the empirical covariance operator[3]

$$\widehat{\Sigma}_P := \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)\phi(x_i)^{\mathsf{T}},$$

[3]In this proof, all the operators are defined with respect to the space $\ell^2(\mathbb{N})$.

the population covariance operator $\Sigma_P := \mathbf{E}_{X\sim P}[\phi(X)\phi(X)^\intercal]$, and the diagonal operator $M := \mathbf{diag}(\{\mu_j\}_{j\geqslant 1})$.

Before we embark on the proof, we single out two important properties regarding Σ_P and $\widehat{\Sigma}_P$ that will be useful in later proofs. For a given $\lambda > 0$, we define the event

$$\mathcal{E}(\lambda) := \Big\{ M^{1/2}\widehat{\Sigma}_P M^{1/2} + \lambda I \succcurlyeq \frac{1}{2}\big(M^{1/2}\Sigma_P M^{1/2} + \lambda I\big)\Big\}, \tag{1.21}$$

where I is the identity operator on $\ell^2(\mathbb{N})$.

Lemma 1.1. *For any B-bounded source-target pair* (1.2)*, we have the deterministic lower bound*

$$\Sigma_P \succcurlyeq \tfrac{1}{B} I. \tag{1.22a}$$

If, in addition, the kernel is κ-uniformly bounded (1.7)*, then whenever $n\lambda \geqslant 10\kappa^2$, the event $\mathcal{E}(\lambda)$ defined in equation* (1.21) *satisfies*

$$\mathbf{P}[\mathcal{E}(\lambda)] \geqslant 1 - 28\, \tfrac{\kappa^2}{\lambda} e^{-\frac{n\lambda}{16\kappa^2}}.$$

See Section 1.5.1.3 for the proof of this claim.

Equipped with Lemma 1.1, we now proceed to the proof of the theorem. In terms of the basis $\{\phi_j\}_{j\geqslant 1}$, the KRR estimate has the expansion $\widehat{f}_\lambda = \sum_{j=1}^\infty \widehat{\theta}_j \phi_j$, where $\widehat{\theta} = \{\widehat{\theta}_j\}_{j\geqslant 1}$ is a sequence of coefficients in $\ell^2(\mathbf{N})$. By the optimality conditions for the KRR problem, we have

$$\widehat{\theta} - \theta^\star = -\lambda(\widehat{\Sigma}_P + \lambda M^{-1})^{-1} M^{-1}\theta^\star + (\widehat{\Sigma}_P + \lambda M^{-1})^{-1}\big(\frac{1}{n}\sum_{i=1}^n \xi_i \phi(x_i)\big). \tag{1.23}$$

By the triangle inequality, we have the upper bound $\|\widehat{\theta} - \theta^\star\|_2^2 \leqslant 2\big(T_1 + T_2\big)$, where

$$T_1 := \|\lambda(\widehat{\Sigma}_P + \lambda M^{-1})^{-1} M^{-1}\theta^\star\|_2^2, \quad \text{and} \quad T_2 := \|(\widehat{\Sigma}_P + \lambda M^{-1})^{-1}\big(\frac{1}{n}\sum_{i=1}^n \xi_i\phi(x_i)\big)\|_2^2.$$

In terms of this decomposition, it suffices to establish that the following bounds

$$T_1 \overset{(a)}{\leqslant} 2\lambda B \|f^\star\|_{\mathcal{H}}^2, \quad \text{and} \quad T_2 \overset{(b)}{\leqslant} \frac{40(\log n)\sigma^2}{n}\sum_{j=1}^\infty \frac{\mu_j}{\mu_j/B + \lambda}, \tag{1.24}$$

hold with probability at least $1 - 28\,\frac{\kappa^2}{\lambda} e^{-\frac{n\lambda}{16\kappa^2}} - n^{-10}$.

1.5.1.1 Proof of the bound (1.24)(a)

We establish that this bound holds conditionally on the event $\mathcal{E}(\lambda)$. Following some algebraic manipulations, we have

$$\begin{aligned} T_1 &= \lambda^2 \|M^{1/2}(M^{1/2}\widehat{\Sigma}_P M^{1/2} + \lambda I)^{-1} M^{-1/2}\theta^\star\|_2^2 \\ &\overset{(i)}{\leqslant} \lambda^2 \|f^\star\|_{\mathcal{H}}^2 |||M^{1/2}(M^{1/2}\widehat{\Sigma}_P M^{1/2} + \lambda I)^{-1}|||_{\mathrm{op}}^2 \\ &\overset{(ii)}{\leqslant} \lambda \|f^\star\|_{\mathcal{H}}^2 |||M^{1/2}(M^{1/2}\widehat{\Sigma}_P M^{1/2} + \lambda I)^{-1/2}|||_{\mathrm{op}}^2 \\ &\overset{(iii)}{\leqslant} 2\lambda \|f^\star\|_{\mathcal{H}}^2 |||M^{1/2}(M^{1/2}\Sigma_P M^{1/2} + \lambda I)^{-1} M^{1/2}|||_{\mathrm{op}}. \end{aligned}$$

Here inequality (i) follows from the fact that $\|M^{-1/2}\theta^\star\|_2 = \|f^\star\|_{\mathcal{H}}$; the second step (ii) uses the fact that $M^{1/2}\widehat{\Sigma}_P M^{1/2} + \lambda I \succcurlyeq \lambda I$, and step (iii) follows from the fact that we are conditioning on the event $\mathcal{E}(\lambda)$.

Lemma 1.1 also guarantees that $\Sigma_P \succcurlyeq \frac{1}{B} I$, whence

$$T_1 \leqslant 2\lambda \|f^\star\|_{\mathcal{H}}^2 |||M^{1/2}(\frac{1}{B}M + \lambda I)^{-1} M^{1/2}|||_{\mathrm{op}} = 2\lambda \cdot \max_{j \geqslant 1} \left\{ \frac{\mu_j}{\frac{\mu_j}{B} + \lambda} \right\} \leqslant 2\lambda B \|f^\star\|_{\mathcal{H}}^2.$$

This establishes the claim (1.24)(a).

1.5.1.2 Proof of the bound (1.24)(b)

Define the random vector $W := (\widehat{\Sigma}_P + \lambda M^{-1})^{-1}\big(\frac{1}{n}\sum_{i=1}^n \xi_i \phi(x_i)\big)$. Conditioned on the covariates $\{x_i\}_{i=1}^n$, W is a zero-mean sub-Gaussian random variable with covariance operator

$$\Lambda := \frac{\sigma^2}{n}(\widehat{\Sigma}_P + \lambda M^{-1})^{-1}\widehat{\Sigma}_P(\widehat{\Sigma}_P + \lambda M^{-1})^{-1}.$$

Consequently, by the Hanson-Wright inequality in the RKHS (cf. Theorem 2.6 in the chapter [26]), we have

$$\mathbf{P}\left[T_2 \geqslant 20(\log n)\ \mathbf{Tr}(\Lambda) \mid \{x_i\}_{i=1}^n\right] \leqslant \frac{1}{n^{10}}, \tag{1.25}$$

where the probability is taken over the noise variables.

It remains to upper bound the trace. We have $\mathbf{Tr}(\Lambda) = \mathbf{Tr}\left(\frac{\sigma^2}{n}(\widehat{\Sigma}_P + \lambda M^{-1})^{-1}\widehat{\Sigma}_P(\widehat{\Sigma}_P + \lambda M^{-1})^{-1}\right)$, so that

$$\begin{aligned} \mathbf{Tr}(\Lambda) &\leqslant \mathbf{Tr}\left(\frac{\sigma^2}{n}(\widehat{\Sigma}_P + \lambda M^{-1})^{-1}(\widehat{\Sigma}_P + \lambda M^{-1})(\widehat{\Sigma}_P + \lambda M^{-1})^{-1}\right) \\ &= \mathbf{Tr}\left(\frac{\sigma^2}{n}(\widehat{\Sigma}_P + \lambda M^{-1})^{-1}\right) = \mathbf{Tr}\left(\frac{\sigma^2}{n}(M^{1/2}(M^{1/2}\widehat{\Sigma}_P M^{1/2} + \lambda I)^{-1} M^{1/2}\right). \end{aligned}$$

Under the event $\mathcal{E}(\lambda)$, we have $M^{1/2}\widehat{\Sigma}_P M^{1/2} + \lambda I \succcurlyeq \frac{1}{2}\left(M^{1/2}\Sigma_P M^{1/2} + \lambda I\right)$, which implies

$$\begin{aligned}\mathbf{Tr}(\Lambda) &\leqslant 2\frac{\sigma^2}{n}\,\mathbf{Tr}\left(M^{1/2}(M^{1/2}\Sigma_P M^{1/2} + \lambda I)^{-1}M^{1/2}\right) \\ &\overset{(i)}{\leqslant} 2\frac{\sigma^2}{n}\,\mathbf{Tr}\left(M^{1/2}(\frac{1}{B}M + \lambda I)^{-1}M^{1/2}\right) \\ &\overset{(ii)}{=} 2\frac{\sigma^2}{n}\sum_{j=1}^{\infty}\frac{\mu_j}{\frac{\mu_j}{B}+\lambda}.\end{aligned}$$

Here step (i) follows since $\Sigma_P \succcurlyeq \frac{1}{B}I$, and step (ii) follows from a direct calculation. Substituting this upper bound on the trace into the tail bound (1.25) yields the claimed bound (1.24)(b).

1.5.1.3 Proof of Lemma 1.1

We begin with the proof of the lower bound (1.22a). Since $\{\phi_j\}_{j\geqslant 1}$ is an orthonormal basis of $L^2(Q)$, we have

$$\mathbf{E}_{X\sim Q}\left[\phi(X)\phi(X)^{\mathsf{T}}\right] = \mathbf{E}_{X\sim P}\left[\rho(X)\phi(X)\phi(X)^{\mathsf{T}}\right] = I.$$

Thus, the B-boundedness of the likelihood ratio (1.2) implies that

$$I \preccurlyeq \mathbf{E}_{X\sim P}[B\phi(X)\phi(X)^{\mathsf{T}}] = B\Sigma_P,$$

which is equivalent to the claim (1.22a).

Next we prove the lower bound (1.21). We introduce the shorthand notation

$$\widehat{\Sigma}_\lambda := M^{1/2}\widehat{\Sigma}_P M^{1/2} + \lambda I, \quad \text{and} \quad \Sigma_\lambda := M^{1/2}\Sigma_P M^{1/2} + \lambda I$$

along with the matrix $\Delta := \Sigma_\lambda^{-1/2}(\widehat{\Sigma}_\lambda - \Sigma_\lambda)\Sigma_\lambda^{-1/2}$. Recalling that $|||\cdot|||_{\mathrm{op}}$ denotes the ℓ_2-operator norm of a matrix, we first observe that $\{|||\Delta|||_{\mathrm{op}} \leqslant \frac{1}{2}\} \subset \mathcal{E}$. Consequently, it suffices to show that $|||\Delta|||_{\mathrm{op}} \leqslant \frac{1}{2}$ with high probability.

The matrix Δ can be written as the normalized sum $\Delta = \frac{1}{n}\sum_{i=1}^n Z_i$, where the random operators

$$Z_i := \Sigma_\lambda^{-1/2}M^{1/2}\left(\phi(x_i)\phi(x_i)^{\mathsf{T}} - \Sigma_P\right)M^{1/2}\Sigma_\lambda^{-1/2}$$

are i.i.d. The operator norm of each term can be bounded as

$$\begin{aligned}|||Z_i|||_{\mathrm{op}} \leqslant 2\sup_{x\in\mathcal{X}}|||\Sigma_\lambda^{-1/2}M^{1/2}\phi(x)\phi(x)^{\mathsf{T}}M^{1/2}\Sigma_\lambda^{-1/2}|||_{\mathrm{op}} &= 2\sup_{x\in\mathcal{X}}\|\Sigma_\lambda^{-1/2}M^{1/2}\phi(x)\|_2^2 \\ &\leqslant 2\kappa^2|||\Sigma_\lambda^{-1/2}|||_{\mathrm{op}}^2 \leqslant \frac{2\kappa^2}{\lambda} =: L, \quad (1.26)\end{aligned}$$

where the final inequality follows from the assumption that $\sup_{x\in\mathcal{X}}\|M^{1/2}\phi(x)\|_2^2 \leqslant \kappa^2$, and the fact that $\Sigma_\lambda \succcurlyeq \lambda I$.

On the other hand, the variance of Z_i can be bounded as

$$\begin{aligned}\mathbf{E}[Z_i^2] &\preccurlyeq \mathbf{E}[(\Sigma_\lambda^{-1/2} M^{1/2}\phi(X)\phi(X)^\mathsf{T} M^{1/2}\Sigma_\lambda^{-1/2})^2] \\ &= \mathbf{E}[\Sigma_\lambda^{-1/2} M^{1/2}\phi(X)\phi(X)^\mathsf{T} M^{1/2}\Sigma_\lambda^{-1} M^{1/2}\phi(X)\phi(X)^\mathsf{T} M^{1/2}\Sigma_\lambda^{-1/2}] \\ &\preccurlyeq \mathbf{E}\left[\Sigma_\lambda^{-1/2} M^{1/2}\phi(X)\phi(X)^\mathsf{T} M^{1/2}\Sigma_\lambda^{-1/2}\right] \cdot \sup_{x\in\mathcal{X}} \left\{\phi(x)^\mathsf{T} M^{1/2}\Sigma_\lambda^{-1} M^{1/2}\phi(x)\right\} \\ &\preccurlyeq \frac{\kappa^2}{\lambda}\Sigma_\lambda^{-1/2} M^{1/2}\Sigma_P M^{1/2}\Sigma_\lambda^{-1/2} =: V,\end{aligned}$$

where the last inequality follows by applying the bound (1.26) on $\sup_{x\in\mathcal{X}} \|\Sigma_\lambda^{-1/2} M^{1/2}\phi(x)\|_2^2$.

Suppose that we can show that

$$\mathbf{Tr}(V) \preccurlyeq \frac{\kappa^2}{\lambda}\cdot\frac{\kappa^2}{\lambda}; \tag{1.27a}$$

$$|||V|||_{\mathrm{op}} \preccurlyeq \frac{\kappa^2}{\lambda}. \tag{1.27b}$$

We can then apply a dimension-free matrix Bernstein inequality (see Lemma 1.8) with $t = 1/2$ to obtain the tail bound

$$\mathbf{P}\left[|||\Delta|||_{\mathrm{op}} \geqslant \tfrac{1}{2}\right] \preccurlyeq 28\frac{\kappa^2}{\lambda}\exp\Big(-\frac{n\lambda}{16\kappa^2}\Big),$$

as long as $n\lambda \geqslant 10\kappa^2$. Thus, the only remaining detail is to prove the bounds (1.27a) and (1.27b).

Proof of the bound (1.27a)**:** Using the definition of V, we have

$$\begin{aligned}\mathbf{Tr}(V) = \frac{\kappa^2}{\lambda}\,\mathbf{Tr}\left(\Sigma_\lambda^{-1/2} M^{1/2}\Sigma_P M^{1/2}\Sigma_\lambda^{-1/2}\right) &= \frac{\kappa^2}{\lambda}\,\mathbf{E}_P[\mathbf{Tr}\left(\Sigma_\lambda^{-1/2} M^{1/2}\phi(x)\phi(x)^\mathsf{T} M^{1/2}\Sigma_\lambda^{-1/2}\right)] \\ &\preccurlyeq \frac{\kappa^2}{\lambda}\cdot\frac{\kappa^2}{\lambda}.\end{aligned}$$

Here we have again applied the bound $\sup_{x\in\mathcal{X}} \|\Sigma_\lambda^{-1/2} M^{1/2}\phi(x)\|_2^2 \preccurlyeq \kappa^2/\lambda$.

Proof of the bound (1.27b)**:** Recalling the definition of Σ_λ, we see that

$$|||\Sigma_\lambda^{-1/2} M^{1/2}\Sigma_P M^{1/2}\Sigma_\lambda^{-1/2}|||_{\mathrm{op}} \preccurlyeq 1,$$

and hence

$$|||V|||_{\mathrm{op}} = \frac{\kappa^2}{\lambda}|||\Sigma_\lambda^{-1/2} M^{1/2}\Sigma_P M^{1/2}\Sigma_\lambda^{-1/2}|||_{\mathrm{op}} \preccurlyeq \frac{\kappa^2}{\lambda},$$

which is the claimed upper bound on $|||V|||_{\mathrm{op}}$.

1.5.2 Proof of Theorem 1.4

We now turn to the proof of our guarantee on the truncated LR-reweighted estimator. At the core of the proof is a uniform concentration result, one that holds within a local ball

$$\mathcal{G}(r) := \{f \in \mathcal{H} \mid \|f - f^\star\|_Q \leqslant r, \text{ and } \|f - f^\star\|_{\mathcal{H}} \leqslant 3\|f^\star\|_{\mathcal{H}}\}$$

around the true regression function $f^\star$.

Lemma 1.2. *Fixing any $r > 0$, we have*

$$\sup_{g \in \mathcal{G}(r)} \Big\{ \|g - f^\star\|_Q^2 + \frac{1}{n}\sum_{i=1}^n \rho_{\tau_n}(x_i)\big[\big(f^\star(x_i) - y_i\big)^2 - \big(g(x_i) - y_i\big)^2\big]\Big\} \leqslant \mathcal{M}(r) \tag{1.28}$$

with probability at least $1 - cn^{-10}$.

See Section 1.5.2.1 for the proof of this lemma.

Taking this lemma as given, we now complete the proof of the theorem. Define the regularized radius $\delta_\lambda := \sqrt{\delta_n^2 + 3\lambda\|f^\star\|_{\mathcal{H}}^2}$, and denote by $\mathcal{E}(\delta_\lambda)$ the "good" event that the relation (1.28) holds at radius δ_λ. We immediately point out an important property of the regularized radius δ_λ, namely $\mathcal{M}(\delta_\lambda) \leqslant (1/2) \cdot \delta_\lambda^2$. To see this, note that $r \mapsto \mathcal{M}(r)/r$ is non-increasing in r, and hence

$$\frac{\mathcal{M}(\delta_\lambda)}{\delta_\lambda} \leqslant \frac{\mathcal{M}(\delta_n)}{\delta_n} \leqslant \frac{1}{2}\delta_n \leqslant \frac{1}{2}\delta_\lambda.$$

Suppose that conditioned on $\mathcal{E}(\delta_\lambda)$, the following inequality holds

$$\inf_{f \in \mathcal{H}, f \notin \mathcal{G}(\delta_\lambda)} \frac{1}{n}\sum_{i=1}^n \rho_{\tau_n}(x_i)\Big\{\big(f(x_i) - y_i\big)^2 - \big(f^\star(x_i) - y_i\big)^2\Big\} + \lambda\|f\|_{\mathcal{H}}^2 - \lambda\|f^\star\|_{\mathcal{H}}^2 > 0. \tag{1.29}$$

It then follows that that $\|\widehat{f} - f^\star\|_Q \leqslant \delta_\lambda$, as desired. Consequently, the remainder of our proof is devoted to establishing that inequality (1.29) holds conditioned on $\mathcal{E}(\delta_\lambda)$.

Given any function $f \in \mathcal{H}$ and $f \notin \mathcal{G}(\delta_\lambda)$, there exists an $\alpha \geqslant 1$ such that $\widetilde{f} := f^\star + \frac{1}{\alpha}(f - f^\star)$ lies in the set $\mathcal{H}$, and more importantly $\widetilde{f}$ lies on the boundary of $\mathcal{G}(\delta_\lambda)$. This follows from the convexity of the two sets $\mathcal{H}$ and $\mathcal{G}(\delta_\lambda)$. Since $\widetilde{f}$ is a convex combination of f and $f^\star$, Jensen's inequality implies that

$$\begin{aligned}
&\rho_{\tau_n}(x_i)\Big\{\big(\widetilde{f}(x_i) - y_i\big)^2 - \big(f^\star(x_i) - y_i\big)^2\Big\} + \lambda\|\widetilde{f}\|_{\mathcal{H}}^2 - \lambda\|f^\star\|_{\mathcal{H}}^2 \\
&\quad\leqslant \tfrac{1}{\alpha}\Big\{\rho_{\tau_n}(x_i)\Big\{\big(f(x_i) - y_i\big)^2 - \big(f^\star(x_i) - y_i\big)^2\Big\} + \lambda\|f\|_{\mathcal{H}}^2 - \lambda\|f^\star\|_{\mathcal{H}}^2\Big\}.
\end{aligned}$$

Consequently, in order to establish the claim (1.29), it suffices to prove that the quantity

$$T := \frac{1}{n}\sum_{i=1}^n \rho_{\tau_n}(x_i)\big\{\big(f^\star(x_i) - y_i\big)^2 - \big(\widetilde{f}(x_i) - y_i\big)^2\big\} + \lambda\|f^\star\|_{\mathcal{H}}^2 - \lambda\|\widetilde{f}\|_{\mathcal{H}}^2$$

is negative. Since $\widetilde{f}$ lies on the boundary of $\mathcal{G}(\delta_\lambda)$, we can split the proof into two cases: (1) $\|\widetilde{f} - f^\star\|_Q = \delta_\lambda$, while $\|\widetilde{f} - f^\star\|_{\mathcal{H}} \leqslant 3\|f^\star\|_{\mathcal{H}}$, and (2) $\|\widetilde{f} - f^\star\|_Q \leqslant \delta_\lambda$, while $\|\widetilde{f} - f^\star\|_{\mathcal{H}} = 3\|f^\star\|_{\mathcal{H}}$.

Case 1: $\|\widetilde{f} - f^\star\|_Q = \delta_\lambda$, while $\|\widetilde{f} - f^\star\|_{\mathcal{H}} \leqslant 3\|f^\star\|_{\mathcal{H}}$. By adding and subtracting terms, we have

$$\begin{aligned} T &= \left[\frac{1}{n}\sum_{i=1}^n \rho_{\tau_n}(x_i)\Big\{\big(f^\star(x_i) - y_i\big)^2 - \big(\widetilde{f}(x_i) - y_i\big)^2\Big\} + \|\widetilde{f} - f^\star\|_Q^2\right] - \|\widetilde{f} - f^\star\|_Q^2 + \lambda\|f^\star\|_{\mathcal{H}}^2 - \lambda\|\widetilde{f}\|_{\mathcal{H}}^2 \\ &\overset{(i)}{\leqslant} \mathcal{M}(\delta_\lambda) - \delta_\lambda^2 + \lambda\|f^\star\|_{\mathcal{H}}^2 \overset{(ii)}{<} -\frac{1}{2}\delta_\lambda^2 + \lambda\|f^\star\|_{\mathcal{H}}^2 \overset{(iii)}{<} 0, \end{aligned}$$

where step (i) follows from conditioning on the event $\mathcal{E}(\delta_\lambda)$, the equality $\|\widetilde{f} - f^\star\|_Q^2 = \delta_\lambda^2$, and non-positivity of $\lambda\|\widetilde{f}\|_{\mathcal{H}}^2$; step (ii) follows from the property $\mathcal{M}(\delta_\lambda) \leqslant (1/2)\cdot\delta_\lambda^2$ and step (iii) uses the definitions of δ_λ and λ.

Case 2: $\|\widetilde{f} - f^\star\|_Q \leqslant \delta_\lambda$, while $\|\widetilde{f} - f^\star\|_{\mathcal{H}} = 3\|f^\star\|_{\mathcal{H}}$. By the same addition and subtraction as above, we have

$$\begin{aligned} T &= \left[\frac{1}{n}\sum_{i=1}^n \rho_{\tau_n}(x_i)\Big\{\big(f^\star(x_i) - y_i\big)^2 - \big(\widetilde{f}(x_i) - y_i\big)^2\Big\} + \|\widetilde{f} - f^\star\|_Q^2\right] - \|\widetilde{f} - f^\star\|_Q^2 + \lambda\|f^\star\|_{\mathcal{H}}^2 - \lambda\|\widetilde{f}\|_{\mathcal{H}}^2 \\ &\overset{(i)}{\leqslant} \mathcal{M}(\delta_\lambda) + \lambda\|f^\star\|_{\mathcal{H}}^2 - \lambda\|\widetilde{f}\|_{\mathcal{H}}^2 \\ &\overset{(ii)}{<} \frac{1}{2}\delta_\lambda^2 - 3\lambda\|f^\star\|_{\mathcal{H}}^2, \end{aligned}$$

Here, step (i) again follows from the conditioning on the event $\mathcal{E}(\delta_\lambda)$ and the assumption that $\|\widetilde{f} - f^\star\|_Q \leqslant \delta_\lambda$. Step (ii) relies on the facts that $\mathcal{M}(\delta_\lambda) \leqslant (1/2)\cdot\delta_\lambda^2$, $\|f^\star\|_{\mathcal{H}} = \|f^\star\|_{\mathcal{H}}$, and that $\|\widetilde{f}\|_{\mathcal{H}} \geqslant 2\|f^\star\|_{\mathcal{H}}$. The latter is a simple consequence of $\|\widetilde{f} - f^\star\|_{\mathcal{H}} = 3\|f^\star\|_{\mathcal{H}}$ and the triangle inequality. Substitute in the definitions of δ_λ and λ to see the negativity of T.

Combine the two cases to finish the proof of the claim (1.29).

1.5.2.1 Proof of Lemma 1.2

Define the shifted function class $\mathcal{F}^\star := \mathcal{H} - f^\star$, along with its r-localized version

$$\mathcal{F}^\star(r) := \big\{h \in \mathcal{F}^\star \mid \|h\|_Q \leqslant r, \quad \text{and} \quad \|h\|_{\mathcal{H}} \leqslant 3\|f^\star\|_{\mathcal{H}}\big\}.$$

We begin by observing that

$$\big(f^\star(x_i) - y_i\big)^2 - \big(g(x_i) - y_i\big)^2 = 2\xi_i[g(x_i) - f^\star(x_i)] - \big(g(x_i) - f^\star(x_i)\big)^2,$$

which yields the following equivalent formulation of the claim in Lemma 1.2:

$$\sup_{h\in\mathcal{F}^\star(r)}\left\{\frac{1}{n}\sum_{i=1}^n\Big[2\xi_i\rho_{\tau_n}(x_i)h(x_i) + \|h\|_Q^2 - \rho_{\tau_n}(x_i)h^2(x_i)\Big]\right\} \leqslant \mathcal{M}(r). \tag{1.30}$$

By the triangle inequality, it suffices to show that $T_1 + T_2 \leqslant \mathcal{M}(r)$, where

$$T_1 \coloneqq \sup_{h \in \mathcal{F}^\star(r)} \Big| \frac{2}{n} \sum_{i=1}^n \xi_i \rho_{\tau_n}(x_i) h(x_i) \Big|, \quad \text{and} \quad T_2 \coloneqq \sup_{h \in \mathcal{F}^\star(r)} \Big| \frac{1}{n} \sum_{i=1}^n \big\{ \|h\|_Q^2 - \rho_{\tau_n}(x_i) h^2(x_i) \big\} \Big|.$$

More precisely, the core of our proof involves establishing the following two bounds:

$$T_1 \leqslant c\sigma \sqrt{\frac{V^2 \log^3(n)}{n}} \Big\{ \sum_{j=1}^\infty \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\} \Big\}^{1/2} \qquad \text{with probability at least } 1 - n^{-10}, \text{ and} \tag{1.31a}$$

$$T_2 \leqslant c \sqrt{\frac{V^2 \log^3(n)}{n}} \cdot \sum_{j=1}^\infty \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\} \qquad \text{with probability at least } 1 - n^{-10}. \tag{1.31b}$$

In conjunction, these two bounds ensure that

$$T_1 + T_2 \leqslant c \sqrt{\frac{V^2 \log^3(n)}{n}} \cdot \sum_{j=1}^\infty \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\} + c \sqrt{\sum_{j=1}^\infty \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\} \frac{V^2 \log^3(n)}{n}} \tag{1.32}$$

Since the kernel function is κ^2-bounded, we have $\sum_{j=1}^\infty \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\} \leqslant \|f^\star\|_{\mathcal{H}}^2 \sum_{j=1}^\infty \mu_j \leqslant \kappa^2 \|f^\star\|_{\mathcal{H}}^2$, which together with the assumption $\sigma^2 \geqslant \kappa^2 \|f^\star\|_{\mathcal{H}}^2$ implies that

$$T_1 + T_2 \leqslant 2c \sqrt{\sum_{j=1}^\infty \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\} \frac{V^2 \log^3(n)}{n} \sigma^2}.$$

Therefore the bound (1.30) holds.

It remains to prove the bounds (1.31a) and (1.31b). The proofs make use of some elementary properties of the localized function class $\mathcal{F}^\star(r)$, which we collect here. For any $h \in \mathcal{F}^\star(r)$, we have

$$|h(x)| \leqslant \sqrt{10 \sum_{j=1}^\infty \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\}}, \qquad \text{and} \tag{1.33a}$$

$$\sum_{j=1}^\infty \frac{\theta_j^2}{\min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\}} \leqslant 10, \qquad \text{where } h = \textstyle\sum_{j=1}^\infty \theta_j \phi_j. \tag{1.33b}$$

See Section 1.7.3 for the proof of these elementary claims.

1.5.2.2 Proof of inequality (1.31b)

We begin by analyzing the term T_2. By the triangle inequality, we have the upper bound $T_2 \leqslant T_{2a} + T_{2b}$, where

$$T_{2a} := \sup_{h \in \mathcal{F}^\star(r)} \Big| \|h\|_Q^2 - \mathbf{E}_P[\rho_{\tau_n}(X) h^2(X)] \Big|, \quad \text{and}$$

$$T_{2b} := \sup_{h \in \mathcal{F}^\star(r)} \Big| \mathbf{E}_P[\rho_{\tau_n}(X) h^2(X)] - \frac{1}{n} \sum_{i=1}^n \rho_{\tau_n}(x_i) h^2(x_i) \Big\} \Big|.$$

Note that T_{2a} is a deterministic quantity, measuring the bias induced by truncation, whereas T_{2b} is the supremum of an empirical process. We split our proof into analysis of these two terms. In particular, we establish the following bounds:

$$T_{2a} \leqslant c \sqrt{\frac{V^2}{n}} \sum_{j=1}^\infty \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\}, \quad \text{and} \tag{1.34a}$$

$$T_{2b} \leqslant c \sqrt{\frac{V^2 \log^2(n)}{n}} \sum_{j=1}^\infty \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\} \qquad \text{with probability at least } 1 - n^{-10} \tag{1.34b}$$

Combining these two bounds yields the claim (1.31b).

Proof of inequality (1.34a)**:** We begin by proving the claimed upper bound on T_{2a}. Note that

$$\begin{aligned}
T_{2a} &\leqslant \sup_{h \in \mathcal{F}^\star(r)} \Big| \|h\|_Q^2 - \mathbf{E}_Q[\mathbf{1}\{\rho(X) \leqslant \tau_n\} h^2(X)] \Big| + \tau_n \cdot \sup_{h \in \mathcal{F}^\star(r)} \Big| \mathbf{E}_P[\mathbf{1}\{\rho(X) > \tau_n\} h^2(X)] \Big| \\
&= \sup_{h \in \mathcal{F}^\star(r)} \mathbf{E}_Q \Big[\mathbf{1}\{\rho(X) > \tau_n\} h^2(X) \Big] + \tau_n \cdot \sup_{h \in \mathcal{F}^\star(r)} \Big| \mathbf{E}_P[\mathbf{1}\{\rho(X) > \tau_n\} h^2(X)] \Big| \\
&\leqslant \mathbf{E}_Q \Big[\mathbf{1}\{\rho(X) > \tau_n\} \Big] \cdot \sup_{h \in \mathcal{F}^\star(r)} \|h\|_\infty^2 + \tau_n \cdot \mathbf{E}_P[\mathbf{1}\{\rho(X) > \tau_n\}] \cdot \sup_{h \in \mathcal{F}^\star(r)} \|h\|_\infty^2 \\
&\leqslant \frac{V^2}{\tau_n} \cdot 10 \sum_{j=1}^\infty \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\} + \tau_n \cdot \frac{V^2}{(\tau_n)^2} \cdot 10 \sum_{j=1}^\infty \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\},
\end{aligned}$$

where the last step follows from a combination of Markov's inequality, Chebyshev's inequality, and the ℓ_∞-norm bound (1.33a). Recalling that $\tau_n = \sqrt{nV^2}$, the bound (1.34a) follows.

Proof of the bound (1.34b)**:** We prove the claimed bound on T_{2b} by first bounding its mean $\mathbf{E}[T_{2b}]$, and then providing a high-probability bound on the deviation $T_{2b} - \mathbf{E}[T_{2b}]$.

Bound on the mean: By a standard symmetrization argument (see e.g., Chapter 4 in the book [120]), we have the upper bound

$$\mathbf{E}[T_{2b}] \leqslant \frac{2}{n} \, \mathbf{E} \Big[\sup_{h \in \mathcal{F}^\star(r)} \big| \sum_{i=1}^n \varepsilon_i \rho_{\tau_n}(x_i) h^2(x_i) \big| \Big],$$

where $\{\varepsilon_i\}_{i=1}^n$ is an i.i.d. sequence of Rademacher variables. Now observe that

$$\sup_{h\in\mathcal{F}^\star(r)}\Big|\sum_{i=1}^n \varepsilon_i\rho_{\tau_n}(x_i)h^2(x_i)\Big| \leqslant \sup_{\widetilde{h},h\in\mathcal{F}^\star(r)} Z(h,\widetilde{h}), \quad \text{where} \quad Z(h,\widetilde{h}) := \Big|\sum_{i=1}^n \varepsilon_i\rho_{\tau_n}(x_i)\widetilde{h}(x_i)h(x_i)\Big|.$$

Writing $\widetilde{h} = \sum_{j=1}^\infty \widetilde{\theta}_j\phi_j$, we have

$$\begin{aligned}
Z(h,\widetilde{h}) &= \Big|\sum_{j=1}^\infty \widetilde{\theta}_j\Big\{\sum_{i=1}^n \varepsilon_i\rho_{\tau_n}(x_i)\phi_j(x_i)h(x_i)\Big\}\Big| \\
&= \Big|\sum_{j=1}^\infty \frac{\widetilde{\theta}_j}{\sqrt{\min\{r^2,\mu_j\|f^\star\|_{\mathcal{H}}^2\}}}\cdot\sqrt{\min\{r^2,\mu_j\|f^\star\|_{\mathcal{H}}^2\}}\Big\{\sum_{i=1}^n \varepsilon_i\rho_{\tau_n}(x_i)\phi_j(x_i)h(x_i)\Big\}\Big| \\
&\leqslant \sqrt{10}\Big\{\sum_{j=1}^\infty \min\{r^2,\mu_j\|f^\star\|_{\mathcal{H}}^2\}\cdot\Big\{\sum_{i=1}^n \varepsilon_i\rho_{\tau_n}(x_i)\phi_j(x_i)h(x_i)\Big\}^2\Big\}^{1/2},
\end{aligned}$$

where the final step follows by combining the Cauchy–Schwarz inequality with the bound (1.33b). We now repeat the same argument to upper bound the inner term involving h; in particular, we have

$$\begin{aligned}
\Big\{\sum_{i=1}^n \varepsilon_i\rho_{\tau_n}(x_i)\phi_j(x_i)h(x_i)\Big\}^2 &= \Big\{\sum_{k=1}^\infty \theta_k\Big(\sum_{i=1}^n \varepsilon_i\rho_{\tau_n}(x_i)\phi_j(x_i)\phi_k(x_i)\Big)\Big\}^2 \\
&\leqslant 10\cdot\sum_{k=1}^\infty\Big\{\min\{r^2,\mu_k\|f^\star\|_{\mathcal{H}}^2\}\Big(\sum_{i=1}^n \varepsilon_i\rho_{\tau_n}(x_i)\phi_j(x_i)\phi_k(x_i)\Big)^2\Big\}.
\end{aligned}$$

Putting together the pieces now leads to the upper bound

$$\begin{aligned}
\frac{2}{n}\sup_{h\in\mathcal{F}^\star(r)}\Big|\sum_{i=1}^n \varepsilon_i\rho_{\tau_n}(x_i)h^2(x_i)\Big| &\leqslant \frac{2}{n}\sup_{h,\widetilde{h}\in\mathcal{F}^\star(r)} Z(h,\widetilde{h}) \\
&\leqslant \frac{20}{n}\Big\{\sum_{j=1}^\infty \min\{r^2,\mu_j\|f^\star\|_{\mathcal{H}}^2\} \\
&\qquad\cdot\sum_{k=1}^\infty \min\{r^2,\mu_k\|f^\star\|_{\mathcal{H}}^2\}\Big(\sum_{i=1}^n \varepsilon_i\rho_{\tau_n}(x_i)\phi_j(x_i)\phi_k(x_i)\Big)^2\Big\}^{1/2}.
\end{aligned}$$

By taking expectations of both sides and applying Jensen's inequality, we find that

$$\begin{aligned}
\mathbf{E}[T_{2b}] \leqslant \frac{20}{n}\Big\{\sum_{j=1}^\infty \min\{r^2,\mu_j\|f^\star\|_{\mathcal{H}}^2\} \\
\cdot\sum_{k=1}^\infty \min\{r^2,\mu_k\|f^\star\|_{\mathcal{H}}^2\}\,\mathbf{E}_{X,\varepsilon}\Big(\sum_{i=1}^n \varepsilon_i\rho_{\tau_n}(x_i)\phi_j(x_i)\phi_k(x_i)\Big)^2\Big\}^{1/2}. \quad (1.35)
\end{aligned}$$

We now observe that

$$\mathbf{E}_{X,\varepsilon}\Big[\big(\sum_{i=1}^n \varepsilon_i \rho_{\tau_n}(x_i)\phi_j(x_i)\phi_k(x_i)\big)^2\Big] = \sum_{i=1}^n \mathbf{E}_{X,\varepsilon}\Big[\varepsilon_i^2(\rho_{\tau_n}(x_i))^2\phi_j^2(x_i)\phi_k^2(x_i)\Big] \leqslant \sum_{i=1}^n \mathbf{E}_{X,\varepsilon}[\rho^2(x_i)] \;\leqslant\; nV^2,$$

where we have used the fact that $\|\phi_j\|_\infty \leqslant 1$ for all $j \geqslant 1$, and that $\rho_{\tau_n}(x_i) \leqslant \rho(x_i)$. Substituting this upper bound into our earlier inequality (1.35) yields

$$\mathbf{E}[T_{2b}] \leqslant 20\sqrt{\frac{V^2}{n}} \cdot \sum_{j=1}^\infty \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\}. \tag{1.36}$$

Bounding the deviation term: Recall that for any $h \in \mathcal{F}^\star$, we have

$$\|h\|_\infty \leqslant \sqrt{10\sum_{j=1}^\infty \min\{r^2, \mu_j\|f^\star\|_{\mathcal{H}}^2\}}.$$

Consequently, we have

$$\sup_{h\in\mathcal{F}^\star(r)} \Big|\mathbf{E}_Q[\mathbf{1}\{\rho(X) \leqslant \tau_n\}h^2(X)] - \rho_{\tau_n}(x_i)h^2(x_i)\Big| \leqslant 10\tau_n \sum_{j=1}^\infty \min\{r^2, \mu_j\|f^\star\|_{\mathcal{H}}^2\} = 10\sqrt{nV^2}\sum_{j=1}^\infty \min\{r^2, \mu_j\|f^\star\|_{\mathcal{H}}^2\}.$$

In addition, we have

$$\sup_{h\in\mathcal{F}^\star(r)} \sum_{i=1}^n \mathbf{E}\Big[\big\{\mathbf{E}_Q[\mathbf{1}\{\rho(X) \leqslant \tau_n\}h^2(X)] - \rho_{\tau_n}(x_i)h^2(x_i)\big\}^2\Big] \leqslant \sup_{h\in\mathcal{F}^\star(r)} \sum_{i=1}^n \mathbf{E}\big[(\rho_{\tau_n}(x_i))^2 h^4(x_i)\big] \leqslant 100nV^2\big(\sum_{j=1}^\infty \min\{r^2, \mu_j\|f^\star\|_{\mathcal{H}}^2\}\big)^2,$$

where we have applied the ℓ_∞-norm bound (1.33a) as well as the V^2-condition on the likelihood ratio. These two facts together allow us to apply Talagrand's concentration results (cf. Lemma 1.9) and obtain

$$\mathbf{P}\Big[T_{2b} \geqslant \mathbf{E}[T_{2b}] + \tfrac{t}{n}\Big] \leqslant \exp\left(-\frac{t^2}{3000nV^2\big(\sum_{j=1}^\infty \min\{r^2, \mu_j\|f^\star\|_{\mathcal{H}}^2\}\big)^2 + 900\sqrt{nV^2}\sum_{j=1}^\infty \min\{r^2, \mu_j\|f^\star\|_{\mathcal{H}}^2\}t}\right). \tag{1.37}$$

Completing the proof of the bound (1.34b): We now have the ingredients to complete the proof of the claim (1.34b). In particular, by combining the upper bound (1.36) on the mean with the deviation bound (1.37), we find that

$$T_{2b} \leqslant c\sqrt{\frac{V^2 \log^2(n)}{n}} \sum_{j=1}^{\infty} \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\} \qquad \text{with probability at least } 1 - n^{-10},$$

as claimed in equation (1.34b).

1.5.2.3 Proof of inequality (1.31a)

Now we focus on the first term $T_1 = \sup_{h \in \mathcal{F}^\star(r)} \left| \frac{1}{n} \sum_{i=1}^n \xi_i \rho_{\tau_n}(x_i) h(x_i) \right|$. Repeating the same strategy as in the proof of the bound (1.34b), we see that

$$T_1 \leqslant \frac{1}{n} \Big\{ 10 \sum_{j=1}^{\infty} \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\} \cdot \big(\sum_{i=1}^n \xi_i \rho_{\tau_n}(x_i) \phi_j(x_i) \big)^2 \Big\}^{1/2}. \tag{1.38}$$

Fix $\{x_i\}_{i=1}^n$. We see that $\big(\sum_{i=1}^n \xi_i \rho_{\tau_n}(x_i) \phi_j(x_i) \big)^2$ is a quadratic form of independent sub-Gaussian random variables. Apply the Hanson-Wright inequality (e.g., Theorem 6.2.1 in the book [119]) to obtain that with probability at least $1 - n^{-10}$,

$$\big(\sum_{i=1}^n \xi_i \rho_{\tau_n}(x_i) \phi_j(x_i) \big)^2 \leqslant c_3 \sigma^2 \sum_{i=1}^n \big[\rho_{\tau_n}(x_i) \phi_j(x_i) \big]^2. \tag{1.39}$$

It remains to control the term $\sum_{i=1}^n \big[\rho_{\tau_n}(x_i) \phi_j(x_i) \big]^2$. To this end, we invoke Bernstein's inequality to arrive at

$$\sum_{i=1}^n \big[\rho_{\tau_n}(x_i) \phi_j(x_i) \big]^2 \leqslant \mathbf{E} \left[\sum_{i=1}^n \big[\rho_{\tau_n}(x_i) \phi_j(x_i) \big]^2 \right] + c_4 \sqrt{\alpha \log n} + c_5 \beta \log n$$

with probability exceeding $1 - n^{-10}$. Here,

$$\alpha := \mathbf{E} \sum_{i=1}^n \mathsf{Var}\Big(\big[\rho_{\tau_n}(x_i) \phi_j(x_i) \big]^2 \Big) \leqslant (nV^2)^2$$
$$\beta := \sup_x |\big[\rho_{\tau_n}(x) \phi_j(x) \big]^2| \leqslant {\tau_n}^2 = nV^2,$$

are the variance and range statistics, respectively. This together with the upper bound $\mathbf{E} \left[\sum_{i=1}^n \big[\rho_{\tau_n}(x_i) \phi_j(x_i) \big]^2 \right] \leqslant nV^2$ implies

$$\sum_{i=1}^n \big[\rho_{\tau_n}(x_i) \phi_j(x_i) \big]^2 \leqslant c_6 n V^2 \log n. \tag{1.40}$$

Combine the inequalities (1.38), (1.39), and (1.40) to complete the proof of the inequality (1.31a).

1.6 Discussion

In this chapter, we study RKHS-based nonparametric regression under covariate shift. In particular, we focus on two broad families of covariate shift problems: (1) the uniformly B-bounded family, and (2) the χ^2-bounded family. For the uniformly B-bounded family, we prove that the unweighted KRR estimate—with properly chosen regularization parameter—achieves optimal rate convergence for a large family of RKHSs with regular eigenvalues. In contrast, the naïve constrained kernel regression estimator is provably suboptimal under covariate shift. In addition, for the χ^2-bounded family, we propose a likelihood-ratio-reweighted KRR with proper truncation that attains the minimax lower bound over this larger family of covariate shift problems.

Our study is an initial step towards understanding the statistical nature of covariate shift. Below we single out several interesting directions to pursue in the future. First, it is of great importance to extend the study to other classes of regression functions, e.g., high dimensional linear regression, decision trees, etc. Second, while it is natural to measure discrepancy between source-target pairs using likelihood ratio, this is certainly not the only possibility. Various measures of discrepancy have been proposed in the literature, and it is interesting to see what the corresponding optimal procedures are. Thirdly, our upper and lower bounds match for regular kernels. It is standard in the kernel regression literature to make an assumption regarding the decay of the kernel eigenvalue sequence [22, 123]. As highlighted by the corollaries to our main upper bound, the assumption of a regular kernel is general enough to capture the main examples of kernels used in practice. Additionally, we emphasize that in this chapter, we have adopted a worst-case perspective where we study the minimax rate of estimation for a sequence of regular kernel eigenvalues, over all B-bounded covariate shifts. A more instance-dependent perspective which studies these minimax rates for a fixed B-bounded covariate shift pair is very interesting and left for future work. Lastly, on a technical end, it is also interesting to see whether one can remove the uniform boundedness of the eigenfunctions in the unbounded likelihood ratio case, and retain the optimal rate of convergence. In the current proof, we mainly use it to develop a localization bound (1.33a) which guarantees that any function $h \in \mathcal{H}$ that is r-close to $f^\star$ in ℓ_2 sense (roughly) enjoys an ℓ_∞ bound that also scales with r.

1.7 Deferred proofs

1.7.1 Proof of Theorem 1.2

Let δ_n be the smallest positive solution to the inequality $c'\sigma^2 B\frac{d(\delta)}{n} \leqslant \delta^2$, where $c' > 0$ is some large constant. We decompose the proof into two steps. First, we construct the lower bound instance, namely the source, target distributions, and the corresponding orthonormal basis. Second, we apply the Fano method to prove the lower bound.

Step 1: Constructing the lower bound instance. Let Q be a uniform distribution on $\{\pm 1\}^{+\infty}$. For the source distribution P, we set it as follows: with probability $1/B$, we sample x uniformly on $\{\pm 1\}^{+\infty}$, and with probability $1 - 1/B$, we set $x = 0$. It can be verified that the pair (P, Q) has B-bounded likelihood ratio. Corresponding to the target distribution Q, we take $\phi_j(x) = x_j$ for every $j \geqslant 1$. In other words, we consider a linear kernel.

Step 2: Establishing the lower bound. In order to apply Fano's method, we first need to construct a packing set of the function class $\mathcal{B}_{\mathcal{H}}(1)$. For a given radius $r > 0$, consider the r-localized ellipse

$$\mathscr{E}(r) := \Big\{\theta \mid \sum_{j=1}^{\infty} \frac{\theta_j^2}{\min\{r^2, \mu_j\}} \leqslant 1\Big\}.$$

It is straightforward to check that for any $\theta \in \mathscr{E}(r)$, the function $f = \sum_{j=1}^{\infty} \theta_j \phi_j$ lies in $\mathcal{B}_{\mathcal{H}}(1)$. This set $\mathscr{E}(r)$ admits a large packing set in the ℓ_2-norm, as claimed in the following lemma.

Lemma 1.3. *For any $r \in (0, \delta_n]$, there exists a set $\{\theta^1, \theta^2, \ldots, \theta^M\} \subset \mathscr{E}(r)$ with $\log M = d_n/64$ such that*

$$\|\theta^j - \theta^k\|_2^2 \geqslant \tfrac{r^2}{4} \qquad \textit{for any distinct pair of indices } j \neq k.$$

See Lemma 4 in the chapter [123].

Having constructed the packing, we then need to control the pairwise KL divergence. Fix an index $j \in [M]$. Let $P \times \mathscr{L}_j$ denote the joint distribution over the observed data $\{(x_i, y_i)\}_{1 \leqslant i \leqslant n}$ when the true function arises from θ^j. Then for any pair of distinct indices $j \neq k$, we have the upper bound

$$\begin{aligned} \mathsf{KL}(P \times \mathscr{L}_j \| P \times \mathscr{L}_k) = \frac{n}{2\sigma^2} \cdot \mathbf{E}_{X \sim P}\Big[\big((\theta^j - \theta^k)^{\mathsf{T}} \phi(X)\big)^2\Big] &\overset{(i)}{=} \frac{n}{2\sigma^2 B} \|\theta^j - \theta^k\|_2^2 \\ &\overset{(ii)}{\leqslant} \frac{2nr^2}{\sigma^2 B}, \end{aligned}$$

where step (i) follows from the definition of P; and step (ii) follows from applying the triangle inequality, and the fact that $\|\theta\|_2 \leqslant r$ for all $\theta \in \mathscr{E}(r)$.

Consequently, we arrive at the lower bound $\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{F}} \mathbf{E}[\|\widehat{f} - f^\star\|_Q^2] \geqslant \frac{r^2}{8}$, valid for any sample size satisfying the condition

$$\frac{2nr^2}{\sigma^2 B} + \log 2 \leqslant \frac{1}{2} \log M = \frac{d_n}{128}. \tag{1.41}$$

By the definition of a regular kernel, we have $d_n \geqslant c \frac{n\delta_n^2}{B\sigma^2}$ for a universal constant c. Furthermore, since δ_n satisfies the lower bound $\delta_n^2 \geqslant c' \frac{\sigma^2 B}{n}$, the condition (1.41) is met by setting $r^2 = c_1 \delta_n^2$ for some sufficiently small constant $c_1 > 0$.

1.7.2 Proof of Theorem 1.3

Let the sample size $n \geqslant 1$ and likelihood ratio bound $B \geqslant 1$ be given. Our failure instance relies on a function class $\mathcal{F}_n$, together with a pair of distributions (P, Q). The function class $\mathcal{F}_n$ is the unit ball of a RKHS with finite-rank kernel, over the hypercube $\{-1, +1\}^n$. The kernel is given by $\mathscr{K}(x, z) := \sum_{j=1}^n \mu_j \phi_j(x)\phi_j(z)$. The eigenfunctions and eigenvalues are

$$\phi_j(x) = x_j, \quad \text{and} \quad \mu_j = \frac{1}{j^2}, \quad \text{for } j = 1, \ldots, n.$$

To be clear, the function class is given by

$$\mathcal{F}_n := \{\, f := \sum_{j=1}^n \theta_j \phi_j \mid \sum_{j=1}^n \tfrac{\theta_j^2}{\mu_j} \leqslant 1 \,\}.$$

The target distribution, Q, is the uniform distribution on $\{-1, +1\}^n$. The source distribution is a product distribution, $P = \otimes_{j=1}^n P_j$. We take P_j to be uniform on $\{+1, -1\}$, when $j > 1$. On the other hand, the first coordinate follows the distribution

$$P_1 := \Big(1 - \frac{1}{B}\Big)\delta_0 + \frac{1}{B}\mathrm{Unif}(\{-1, +1\}).$$

It is immediate that (P, Q) have B-bounded likelihood ratio.

Given this set-up, our first step is to reduce the lower bound to the separation of a single coordinate of the parameter associated with the empirical risk minimizer and a single coordinate of the parameter associated with a hard instance in the function class of interest $\mathcal{F}_n$. We introduce a one-dimensional minimization problem that governs this separation problem and allows us to establish our result.

1.7.2.1 Reduction to a one dimensional separation problem

To establish our lower bound it suffices to consider the following "hard" function

$$f^\star_{\mathrm{hard}}(x) = x_1 = \sum_{j=1}^n (\theta^\star_{\mathrm{hard}})_j \phi_j(x), \quad \text{where} \quad \theta^\star_{\mathrm{hard}} = (1, 0, \ldots, 0) \in \mathbb{R}^n.$$

Since $\phi_j(x) = x_j$ and $\mu_j = j^{-2}$, it follows that $f^\star_{\mathrm{hard}} \in \mathcal{F}_n$. We can write $\widehat{f}_{\mathrm{erm}}(x) = \sum_{j=1}^n (\widehat{\theta}_{\mathrm{erm}})_j x_j$, where we defined

$$\widehat{\theta}_{\mathrm{erm}} := \arg\min \Big\{ \sum_{i=1}^n \big(\sum_{j=1}^n \theta_j x_{ij} - y_i \big)^2 \mid \sum_{j=1}^n \tfrac{\theta_j^2}{\mu_j} \leqslant 1 \Big\}. \tag{1.42}$$

Putting these pieces together, we see that

$$\sup_{f^\star \in \mathcal{F}_n} \mathbf{E}\Big[\|\widehat{f}_{\mathrm{erm}} - f^\star\|_Q^2 \Big] \geqslant \mathbf{E}\Big[\|\widehat{f}_{\mathrm{erm}} - f^\star_{\mathrm{hard}}\|_Q^2 \Big] \overset{\text{(i)}}{=} \mathbf{E}\Big[\|\widehat{\theta}_{\mathrm{erm}} - \theta^\star_{\mathrm{hard}}\|_2^2 \Big] \overset{\text{(ii)}}{\geqslant} \mathbf{E}\Big[\big((\widehat{\theta}_{\mathrm{erm}})_1 - \theta^\star_1 \big)^2 \Big]. \tag{1.43}$$

Above, the relation (i) is a consequence of Parseval's theorem, along with the orthonormality of $\{\phi_j\}_{j=1}^n$ in $L^2(Q)$. Inequality (ii) follows by dropping terms corresponding to indices indices $j > 1$. Therefore, in view of display (1.43), it suffices to show that:

$$\mathbf{P}\left\{\left((\widehat{\theta}_{\mathrm{erm}})_1 - 1\right)^2 \geqslant c_3 \frac{B^3}{n^2}\right\} \geqslant \frac{1}{2}. \tag{1.44}$$

1.7.2.2 Proof of one-dimensional separation bound (1.44)

We begin with a proof outline.

Proof outline To establish (1.44), we can assume $(\widehat{\theta}_{\mathrm{erm}})_1 \in [0,1]$; otherwise, the lower bound follows trivially, provided c_2 is sufficiently small, in particular, $c_2 \leqslant \sqrt[3]{1/c_3}$. We introduce a bit of notation:

$$\widehat{\Sigma}_P := \frac{1}{n}\sum_{i=1}^n x_i x_i^{\mathsf{T}} \quad \text{and} \quad v := \frac{1}{n}\sum_{i=1}^n w_i x_i.$$

Thus, we can further restrict the empirical risk minimization problem (1.42) to

$$\begin{aligned}\widetilde{\theta} &:= \arg\min\left\{ \sum_{i=1}^n \big(\sum_{j=1}^n \theta_j x_{ij} - y_i\big)^2 \mid \sum_{j=1}^n \frac{\theta_j^2}{\mu_j} \leqslant 1, \quad \theta_1 \in [0,1] \right\} \\ &= \arg\min\left\{ (\theta - \theta^\star)^{\mathsf{T}}\widehat{\Sigma}_P(\theta - \theta^\star) - 2v^{\mathsf{T}}(\theta - \theta^\star) \mid \sum_{j=1}^n \frac{\theta_j^2}{\mu_j} \leqslant 1, \quad \theta_1 \in [0,1] \right\}. \end{aligned} \tag{1.45}$$

Indeed, in order to prove inequality (1.44), it suffices to show that

$$\mathbf{P}\left\{\left(\widetilde{\theta}_1 - 1\right)^2 \geqslant c_3 \frac{B^3}{n^2}\right\} \geqslant \frac{1}{2}. \tag{1.46}$$

Let us define an auxiliary function $g\colon [0,1] \to \mathbb{R}$, given by

$$g(t) := \inf\left\{(\theta - \theta^\star)^{\mathsf{T}}\widehat{\Sigma}_P(\theta - \theta^\star) - 2v^{\mathsf{T}}(\theta - \theta^\star) \mid \sum_{j=1}^n \frac{\theta_j^2}{\mu_j} \leqslant 1, \quad \theta_1 = t\right\}. \tag{1.47}$$

By definition (1.45), the choice $\widetilde{\theta}$ minimizes this objective, and therefore $\inf_{t\in[0,1]} g(t) = g(\widetilde{\theta}_1)$. The next two lemmas concern the minimum value and minimizer of g. Lemma 1.4, which we prove in section 1.7.2.5, bounds the minimal value from above. Lemma 1.5, demonstrates that there is an interval of length order $\sqrt{B^3/n^2}$ on which the function g is bounded away from the minimal value. We prove this result in Section 1.7.2.6.

Lemma 1.4 (Minimal value of empirical objective)**.** *There is a constant $c^* > 0$ such that*

$$g(\widetilde{\theta}_1) \leqslant -c^* \frac{\sqrt{B}}{n}$$

holds with probability at least 3/4.

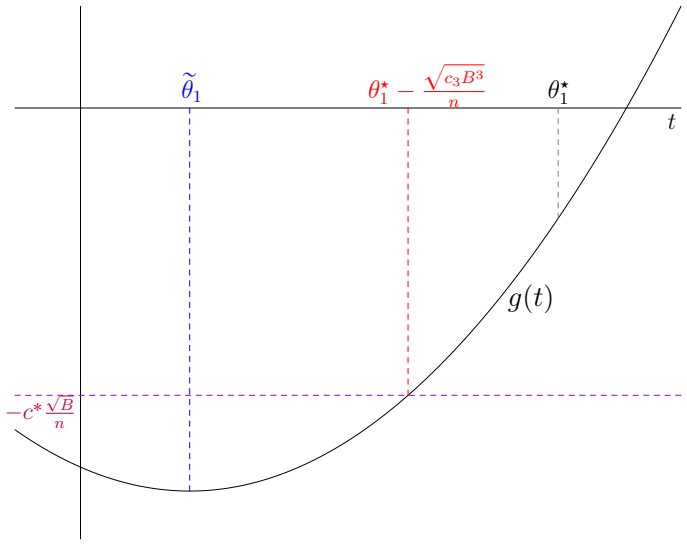


Figure 1.3. Pictorial representation of lower bound argument, separating the first coordinate of empirical risk minimizer, $\widetilde{\theta}_1$, from the true population minimizer $\theta_1^\star$. Lemma 1.4 establishes the upper bound, depicted in purple above, on the minimal value of g. Lemma 1.5 establishes an interval, shown between the red dashed line and $\theta_1^\star$ above, which excludes $\widetilde{\theta}_1$. This allows us to ensure that $\theta_1^\star$ and $\widetilde{\theta}_1$ are sufficiently separated.

Lemma 1.5 (Separation from $\theta_1^\star$)**.** *There exists a constant $c_3 > 0$ such that*

$$\inf_{\substack{t \in [0,1] \\ (1-t)^2 \leqslant c_3 B^3/n^2}} g(t) > -c^* \frac{\sqrt{B}}{n}.$$

where probability at least 3/4.

Note that the constant c^* used in Lemmas 1.4 and 1.5 is the same. Thus—after union bounding over the two error events—with probability at least 1/2,

$$g(\widetilde{\theta}_1) < \inf_{\substack{t \in [0,1] \\ (1-t)^2 \leqslant c_3 B^3/n^2}} g(t).$$

Recalling that $\widetilde{\theta}_1 \in [0, 1]$, we conclude on this event that $(1 - \widetilde{\theta}_1)^2 \geqslant c_3 \frac{B^3}{n^2}$, which furnishes (1.46), and thereby establishes the required result. To complete the proof, it then remains to establish the auxiliary lemmas stated above. Before doing so, we record a useful lemma, which will be used multiple times later.

1.7.2.3 A useful lemma

Lemma 1.6. *For any quantity $\alpha \in (\frac{B}{4n^2}, \frac{B}{4})$, with probability at least $1 - c_1 \exp\big(-c_2 \frac{B^{1/2}}{\alpha^{1/2}}\big)$, one has*

$$c \frac{B^{1/2}}{\alpha^{1/2}} \frac{1}{n} \leqslant \sum_{j=2}^{n} \frac{(v_j)^2}{1 + \frac{\alpha}{B\mu_j}} \leqslant C \frac{B^{1/2}}{\alpha^{1/2}} \frac{1}{n}.$$

Here $c_1, c_2, C, c > 0$ are absolute constants.

Proof. For each $j \geqslant 1$, define $\eta_j := \big(1 + \frac{\alpha}{B\mu_j}\big)^{-1}$. We focus on controlling the term

$$\sum_{j=2}^{n} \eta_j \left[(\sqrt{n} v_j)^2 - 1 \right].$$

Recall from the definition of v that $v_j = \frac{1}{n} \sum_{i=1}^{n} \xi_i x_{ij}$. Under the construction of the lower bound instance, we have $\sqrt{n} v_j \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, 1)$. Therefore $\sqrt{n} v_j)^2 - 1$ is a mean-zero sub-exponential random variable. This allows us to invoke Bernstein's inequality to obtain

$$\mathbf{P}\left(\left| \sum_{j=2}^{n} \eta_j \left[(\sqrt{n} v_j)^2 - 1 \right] \right| \geqslant t \right) \leqslant 2 \exp\left\{ -c \min\left(\frac{t^2}{\sum_{j \geqslant 2} \eta_j^2}, \frac{t}{\max_j \eta_j} \right) \right\},$$

where $c > 0$ is some universal constant.

We claim that there exist three constants $C_1, C_2, C_3 > 0$ such that

$$\max_{j=2,\ldots,n} \eta_j \leqslant 1; \tag{1.48a}$$

$$\sum_{j=2}^{n} \eta_j^2 \leqslant C_1 \frac{B^{1/2}}{\alpha^{1/2}}; \tag{1.48b}$$

$$C_2 \frac{B^{1/2}}{\alpha^{1/2}} \leqslant \sum_{j=2}^{n} \eta_j \leqslant C_3 \frac{B^{1/2}}{\alpha^{1/2}}. \tag{1.48c}$$

As a result, we can $t = c_0 \frac{B^{1/2}}{\alpha^{1/2}}$ with c_0 sufficiently small to arrive at the desired conclusion.

We are left with proving the claimed relations (1.48). The first relation (1.48a) is trivial. We provide the proof of the third inequalities (1.48c); the proof of the middle one (cf. relation (1.48b)) follows by a similar argument. Since $\alpha \in (\frac{B}{4n^2}, \frac{B}{4})$, we can decompose the sum into

$$\sum_{j=2}^{n} \eta_j = \sum_{j=2}^{\lfloor \sqrt{B/\alpha} \rfloor} \frac{1}{1 + \frac{\alpha}{B\mu_j}} + \sum_{j=\lfloor \sqrt{B/\alpha} \rfloor + 1}^{n} \frac{1}{1 + \frac{\alpha}{B\mu_j}}.$$

Recall that $\mu_j = j^{-2}$. We thus have $1 \geqslant \frac{\alpha}{B\mu_j}$ for $j \leqslant \lfloor\sqrt{B/\alpha}\rfloor$ and $1 \leqslant \frac{\alpha}{B\mu_j}$ for $j \geqslant \lfloor\sqrt{B/\alpha}\rfloor$. These allow us to upper bound $\sum_{j=2}^{n} \eta_j$ as

$$\sum_{j=2}^{n} \eta_j \leqslant \lfloor\sqrt{B/\alpha}\rfloor + \frac{B}{\alpha} \sum_{j=\lfloor\sqrt{B/\alpha}\rfloor+1}^{n} \frac{1}{j^2} \leqslant C_3 \frac{B^{1/2}}{\alpha^{1/2}}.$$

Similarly, we have the lower bound

$$\sum_{j=2}^{n} \eta_j \geqslant \sum_{j=2}^{\lfloor\sqrt{B/\alpha}\rfloor} \frac{1}{1+\frac{\alpha}{B\mu_j}} \geqslant \frac{1}{2}\lfloor\sqrt{B/\alpha}\rfloor \geqslant C_2 \frac{B^{1/2}}{\alpha^{1/2}}.$$

This finishes the proof. □

1.7.2.4 Proof of auxiliary lemmas

In order to facilitate the proofs of these lemmas, it is useful to decompose $\theta = (\theta_1, \theta_{\mathsf{R}}) \in \mathbb{R} \times \mathbb{R}^{n-1}$. Additionally, we consider the constraint set

$$\mathcal{C}(t) := \Big\{\theta_{\mathsf{R}} \in \mathbb{R}^{n-1} \mid \sum_{j=2}^{n} \frac{\theta_j^2}{\mu_j} \leqslant 1 - t^2\Big\}, \quad \text{where } t \in [0,1].$$

This set plays a key role. In view of definition (1.47), we can write

$$g(t) = \inf_{\theta_{\mathsf{R}} \in \mathcal{C}(t)} \left\{ \begin{bmatrix} t-1 \\ \theta_{\mathsf{R}} \end{bmatrix}^{\top} \widehat{\Sigma}_P \begin{bmatrix} t-1 \\ \theta_{\mathsf{R}} \end{bmatrix} - 2 \begin{bmatrix} t-1 \\ \theta_{\mathsf{R}} \end{bmatrix}^{\mathsf{T}} v \right\},$$

where above we have used $\theta^\star = (1, 0, \ldots, 0)$. Finally, we will use the diagonal matrix of kernel eigenvalues $M := \mathbf{diag}(\mu_1, \mu_2, \ldots, \mu_n)$, repeatedly.

1.7.2.5 Proof of Lemma 1.4

We show that with probability at least $3/4$,

$$g(\omega) \leqslant -c^* \frac{\sqrt{B}}{n}, \quad \text{where} \quad \omega := \sqrt{1 - \frac{B^{3/2}}{n}}. \tag{1.49}$$

When $n^2 \geqslant B^3$, we have $\omega \in [0,1]$. Since $\inf_{t\in[0,1]} g(t) \leqslant g(\omega)$, the display (1.49) implies the result.

Proof of bound (1.49)**:** From the proof of Lemma 1.1, if we set $\lambda := C\frac{\log n}{n}$ for some constant $C > 0$, then we have

$$\tfrac{1}{2}(\Sigma_P + \lambda M^{-1}) \preccurlyeq \widehat{\Sigma}_P + \lambda M^{-1} \preccurlyeq \tfrac{3}{2}(\Sigma_P + \lambda M^{-1}), \tag{1.50}$$

with probability at least $1 - \frac{1}{n}$. Consequently, for any vector θ obeying $\theta^\top M^{-1}\theta \leqslant 1$, we have the upper bound

$$\begin{aligned}
(\theta - \theta^\star)^\top \widehat{\Sigma}_P (\theta - \theta^\star) &= (\theta - \theta^\star)^\top \left(\widehat{\Sigma}_P + \lambda M^{-1}\right)(\theta - \theta^\star) - \lambda (\theta - \theta^\star)^\top M^{-1} (\theta - \theta^\star) \\
&\leqslant \frac{3}{2}(\theta - \theta^\star)^\top (\Sigma_P + \lambda M^{-1})(\theta - \theta^\star) - \lambda (\theta - \theta^\star)^\top M^{-1} (\theta - \theta^\star) \\
&= \frac{3}{2}(\theta - \theta^\star)^\top \Sigma_P (\theta - \theta^\star) + \frac{\lambda}{2}(\theta - \theta^\star)^\top M^{-1} (\theta - \theta^\star) \\
&\leqslant \frac{3}{2}(\theta - \theta^\star)^\top \Sigma_P (\theta - \theta^\star) + 2\lambda,
\end{aligned}$$

where the final inequality holds since $(\theta - \theta^\star)^\top M^{-1} (\theta - \theta^\star) \leqslant 4$. Applying this result with the vector $\theta = (\omega, \theta_{\mathsf{R}})^{\mathsf{T}}$ yields

$$\begin{aligned}
g(\omega) &\leqslant \min_{\theta_{\mathsf{R}} \in \mathcal{C}} \left\{ \frac{3}{2} \begin{bmatrix} \omega - 1 \\ \theta_{\mathsf{R}} \end{bmatrix}^\top \Sigma_P \begin{bmatrix} \omega - 1 \\ \theta_{\mathsf{R}} \end{bmatrix} - 2 \begin{bmatrix} \omega - 1 \\ \theta_{\mathsf{R}} \end{bmatrix}^\top v + 2\lambda \right\} \\
&= T_1(\omega) + T_2(\omega) + 2\lambda + \min_{\theta_{\mathsf{R}} \in \mathcal{C}} T_3(\theta_{\mathsf{R}}).
\end{aligned} \tag{1.51}$$

Above, we have defined

$$T_1(\omega) := \frac{3}{2}\frac{(\omega - 1)^2}{B}, \quad T_2(\omega) := -2v_1(\omega - 1), \quad \text{and} \quad T_3(\theta_{\mathsf{R}}) := \frac{3}{2}\|\theta_{\mathsf{R}}\|_2^2 - 2v_{\mathsf{R}}^\top \theta_{\mathsf{R}}, \tag{1.52}$$

and we have used the decomposition $v = (v_1, v_{\mathsf{R}})^{\mathsf{T}}$. We now bound each of these three terms in turn.

Controlling the term $T_1(\omega)$**:** Recall $\omega \in [0, 1]$ satisfies the equality $1 - \omega^2 = \frac{B^{3/2}}{n}$. Consequently, we have

$$T_1(\omega) = \frac{3}{2}\frac{(1 - \omega)^2}{B} \leqslant \frac{3}{2}\frac{(1 - \omega^2)^2}{B} = \frac{3}{2}\frac{B^2}{n^2}. \tag{1.53}$$

Controlling the term $T_2(\omega)$**:** For the second term, by definition of ω, we have

$$T_2(\omega) = 2v_1(1 - \omega) \leqslant 2|v_1|(1 - \omega) \leqslant 2|v_1|(1 - \omega^2) = 2|v_1| \cdot \frac{B^{3/2}}{n}. \tag{1.54}$$

We have the following lemma to control the size of $|v_1|$.

Lemma 1.7. *The following holds true with probability at least 0.99*

$$\left| \frac{1}{n} \sum_{i=1}^{n} \xi_i x_{i1} \right| \leqslant \frac{10}{\sqrt{nB}}.$$

Proof. In view of the construction of the lower bound instance, we can calculate

$$\mathbf{E}\left[\left(\frac{1}{n} \sum_{i=1}^{n} \xi_i x_{i1} \right)^2 \right] = \frac{1}{n^2} \sum_{i=1}^{n} \mathbf{E}\left[\xi_i^2 x_{i1}^2 \right] = \frac{1}{nB}.$$

The claim then follows from Chebyshev's inequality. □

Lemma 1.7 demonstrates that $|v_1| \leqslant \frac{10}{\sqrt{nB}}$, with probability at least 99/100. Therefore, on this event, the bound (1.54) guarantees

$$T_2(\omega) \leqslant 20 \frac{B}{n^{3/2}}. \tag{1.55}$$

Controlling the term $T_3(\theta_r)$: Our final step is to upper bound the constrained minimum $\min_{\theta_{\mathsf{R}} \in \mathcal{C}} T_3(\theta_{\mathsf{R}})$. Since this minimization problem is strictly feasible, Lagrange duality guarantees that

$$\begin{aligned} \min_{\theta_{\mathsf{R}} \in \mathcal{C}} T_3(\theta_{\mathsf{R}}) &= \min_{\theta_{\mathsf{R}}} \max_{\xi \geqslant 0} \left\{ \frac{3}{2} \|\theta_{\mathsf{R}}\|_2^2 - 2 v_{\mathsf{R}}^\top \theta_{\mathsf{R}} + \xi (\theta_{\mathsf{R}}^\top M_{\mathsf{R}}^{-1} \theta_{\mathsf{R}} - \frac{B^{3/2}}{n}) \right\} \\ &= \max_{\xi \geqslant 0} \min_{\theta_{\mathsf{R}}} \left\{ \frac{3}{2} \|\theta_{\mathsf{R}}\|_2^2 - 2 v_{\mathsf{R}}^\top \theta_{\mathsf{R}} + \xi (\theta_{\mathsf{R}}^\top M_{\mathsf{R}}^{-1} \theta_{\mathsf{R}} - \frac{B^{3/2}}{n}) \right\}. \end{aligned}$$

The inner minimum is achieved at $\theta_{\mathsf{R}} = \left[\frac{3}{2} I + \xi M_{\mathsf{R}}^{-1} \right]^{-1} v_{\mathsf{R}}$, so that we have established the equality

$$\min_{\theta_{\mathsf{R}} \in \mathcal{C}} T_3(\theta_{\mathsf{R}}) = \max_{\xi \geqslant 0} \left\{ -\xi \tfrac{B^{3/2}}{n} - v_{\mathsf{R}}^\top \left[\tfrac{3}{2} I + \xi M_{\mathsf{R}}^{-1} \right]^{-1} v_{\mathsf{R}} \right\} \; = \; \max_{\xi \geqslant 0} \left\{ -\xi \frac{B^{3/2}}{n} - \sum_{j=2}^{n} \frac{(v_j)^2}{\frac{3}{2} + \frac{\xi}{\mu_j}} \right\}.$$

It remains to analyze the maximum over the dual variable ξ, and we split the analysis into two cases.

- Case 1: First, suppose that the maximum is achieved at some $\xi^\star \geqslant \frac{1}{B}$. In this case, we have

$$\max_{\xi \geqslant 0} \left\{ -\xi \tfrac{B^{3/2}}{n} - \sum_{j=2}^{n} \frac{(v_j)^2}{\frac{3}{2} + \frac{\xi}{\mu_j}} \right\} \leqslant -\xi^\star \frac{B^{3/2}}{n} \leqslant -\frac{B^{1/2}}{n}.$$

- Case 2: Otherwise, we may assume that the maximum achieved at some $\xi^\star \in [0, \frac{1}{B}]$, in which case we have

$$\max_{\xi \geqslant 0} \left\{ -\xi \frac{B^{3/2}}{n} - \sum_{j=2}^{n} \frac{(v_j)^2}{\frac{3}{2} + \frac{\xi}{\mu_j}} \right\} \leqslant - \sum_{j=2}^{n} \frac{(v_j)^2}{\frac{3}{2} + \frac{\xi^\star}{\mu_j}} \leqslant - \sum_{j=2}^{n} \frac{(v_j)^2}{\frac{3}{2} + \frac{1}{B\mu_j}} \leqslant -c \frac{B^{1/2}}{n},$$

  where $c > 0$ is a constant. Here in view of Lemma 1.6, the last inequality holds with probability at least 0.9 as long as B is sufficiently large.

Combining the two cases, we arrive at the conclusion that as long as B is sufficiently large, with probability at least 0.9,

$$\min_{\theta_{\mathsf{R}} \in \mathcal{C}} T_3(\theta_{\mathsf{R}}) \leqslant -c_1 \frac{B^{1/2}}{n} \tag{1.56}$$

for some constant $c_1 > 0$.

Completing the proof: We can now combine bounds (1.53), (1.55), and (1.56) on the terms T_1, T_2, T_3, respectively. Note that when $n \geqslant 7B^{3/2} \geqslant 7$, all three events and the upper bound (1.51) hold simultaneously, with probability $1 - (\frac{1}{n} + \frac{1}{100} + \frac{1}{10}) \geqslant 3/4$. Therefore, we obtain

$$\begin{aligned} g(\omega) &\leqslant \frac{3}{2}\frac{B^2}{n^2} + 20\frac{B}{n^{3/2}} - c_1 \frac{B^{1/2}}{n} + C\frac{\log n}{n} \\ &\leqslant -\frac{c_1}{2}\frac{B^{1/2}}{n}. \end{aligned}$$

The final inequality above holds, since $B \geqslant c_1 (\log n)^2$ and $n \geqslant 7B^{3/2}$, for sufficiently large $c_1 > 0$.

1.7.2.6 Proof of Lemma 1.5

We will prove the slightly stronger claim that with probability at least 3/4, we have

$$\inf_{\substack{t \in [0,1] \\ 1 - t^2 \leqslant \beta B^{3/2}/n}} g(t) > -c^* \frac{\sqrt{B}}{n} \tag{1.57}$$

To see that this proves the claim, note that $\sup_{t \in [0,1]} \frac{(1-t^2)^2}{(1-t)^2} = 4$ Therefore, if $(1-t)^2 \leqslant \frac{\beta^2}{4}\frac{B^3}{n^2}$, then $(1-t^2)^2 \leqslant \beta^2 \frac{B^3}{n^2}$. Hence, (1.57) proves the claim as soon as $c_3 = \beta^2/4$.

Proof of bound (1.57)**:** On the event (1.50), if $\theta = (\theta_1, \theta_{\mathsf{R}})^\top$ obeys $\theta^\top M^{-1}\theta \leqslant 1$, then we have the lower bound

$$\begin{aligned}
(\theta-\theta^\star)^\top \widehat{\Sigma}_P (\theta-\theta^\star) &= (\theta-\theta^\star)^\top \left(\widehat{\Sigma}_P + \lambda M^{-1}\right)(\theta-\theta^\star) - \lambda (\theta-\theta^\star)^\top M^{-1}(\theta-\theta^\star) \\
&\geqslant \frac{1}{2}(\theta-\theta^\star)^\top (\Sigma_P + \lambda M^{-1})(\theta-\theta^\star) - \lambda (\theta-\theta^\star)^\top M^{-1}(\theta-\theta^\star) \\
&= \frac{1}{2}(\theta-\theta^\star)^\top \Sigma_P (\theta-\theta^\star) - \frac{\lambda}{2}(\theta-\theta^\star)^\top M^{-1}(\theta-\theta^\star) \\
&\geqslant \frac{1}{2}(\theta-\theta^\star)^\top \Sigma_P (\theta-\theta^\star) - 2\lambda,
\end{aligned}$$

valid when $\lambda = C\frac{\log n}{n}$ for some constant $C > 0$. Consequently, we have

$$\begin{aligned}
g(\theta_1) &\geqslant \min_{\theta_{\mathsf{R}} \in \mathcal{C}(\theta_1)} \left\{ \frac{1}{2}(\theta-\theta^\star)^\top \Sigma_P (\theta-\theta^\star) - 2(\theta-\theta^\star)^\top v - 2\lambda \right\} \\
&= \min_{\theta_{\mathsf{R}} \in \mathcal{C}(\theta_1)} \left\{ \frac{1}{2}\frac{(\theta_1-1)^2}{B} - 2v_1(\theta_1-1) + \frac{1}{2}\|\theta_{\mathsf{R}}\|_2^2 - 2v_{\mathsf{R}}^\top \theta_{\mathsf{R}} - 2\lambda \right\} \\
&\geqslant -T_2(\theta_1) - 2\lambda + \min_{\theta_{\mathsf{R}} \in \mathcal{C}(\theta_1)} \{\frac{1}{2}\|\theta_{\mathsf{R}}\|_2^2 - 2v_{\mathsf{R}}^\top \theta_{\mathsf{R}}\}. \qquad (1.58)
\end{aligned}$$

where the last line identifies $-2v_1(\theta_1 - 1)$ with $T_2(\theta_1)$ (cf. definition (1.52)).

We separate the proof into two cases—mainly to get around the duality issue.

Case 1: $\theta_1 = 1$**.** In this case, we have

$$g(\theta_1) \geqslant -2\lambda = -\frac{2C\log n}{n}.$$

Case 2: $\theta_1 \in [0, 1)$**.** We lower bound the terms in equation (1.58) in turn.

- **Lower bounding** $T_2(\theta_1)$**.** For any $0 < 1 - \theta_1^2 \leqslant \beta\frac{B^{3/2}}{n}$, the following relation

$$T_2(\theta_1) \geqslant -2|v_1| \cdot |\theta_1 - 1| \overset{(i)}{\geqslant} -2|v_1| \cdot \left(1 - \theta_1^2\right) \overset{(ii)}{\geqslant} -20\beta\frac{B}{n^{3/2}}$$

  holds with probability at least 0.99. Here step (i) uses the fact that

$$|\theta_1 - 1| = |1 - \sqrt{1 - (1-\theta_1^2)}| \leqslant 1 - \theta_1^2 \qquad \text{for all} \quad \theta_1 \in [0, 1],$$

  and step (ii) relies on Lemma 1.7 and the constraint $1 - \theta_1^2 \leqslant \beta\frac{B^{3/2}}{n}$.

- **Lower bounding** $\min_{\theta_{\mathsf{R}} \in \mathcal{C}'(\theta_1)} \{\frac{1}{2}\|\theta_{\mathsf{R}}\|_2^2 - 2v_{\mathsf{R}}^\top \theta_{\mathsf{R}}\}$**.** When $\theta_1 \in [0,1)$, the constraint set $\mathcal{C}'(\theta_1)$ has non-empty interior, and the minimization over θ_{R} is strictly feasible. In this case, strict duality holds so that

$$\begin{aligned}\min_{\theta_{\mathsf{R}} \in \mathcal{C}'(\theta_1)} \left\{\tfrac{1}{2}\|\theta_{\mathsf{R}}\|_2^2 - 2v_{\mathsf{R}}^\top \theta_{\mathsf{R}}\right\} &= \max_{\xi \geqslant 0} \left\{ -\xi(1-\theta_1^2) - \sum_{j=2}^n \frac{(v_j)^2}{\frac{1}{2} + \frac{\xi}{\mu_j}} \right\} \\ &\geqslant -\left[n(1-\theta_1^2)\right]^{-2/3}(1-\theta_1^2) - \sum_{j=2}^n \frac{(v_j)^2}{\frac{1}{2} + \frac{\left(n\left(1-\theta_1^2\right)\right)^{-2/3}}{\mu_j}} \\ &= -\frac{\left(1-\theta_1^2\right)^{1/3}}{n^{2/3}} - \sum_{j=2}^n \frac{(v_j)^2}{\frac{1}{2} + \frac{\left(n\left(1-\theta_1^2\right)\right)^{-2/3}}{\mu_j}}.\end{aligned}$$

  Here the second line arises from a particular choice of ξ, namely $\xi = \left(n(1-\theta_1^2)\right)^{-2/3}$. Since $1-\theta_1^2 \leqslant \beta\frac{B^{3/2}}{n}$, we further have

$$\begin{aligned}-\frac{(1-\theta_1^2)^{1/3}}{n^{2/3}} - \sum_{j=2}^n \frac{(v_j)^2}{\frac{1}{2} + \frac{\left(n\left(1-\theta_1^2\right)\right)^{-2/3}}{\mu_j}} &\geqslant -\frac{\beta^{1/3}B^{1/2}}{n} - \sum_{j=2}^n \frac{(v_j)^2}{\frac{1}{2} + \frac{\left(\beta B^{3/2}\right)^{-2/3}}{\mu_j}} \\ &= -\frac{\beta^{1/3}B^{1/2}}{n} - \sum_{j=2}^n \frac{(v_j)^2}{\frac{1}{2} + \frac{1}{\beta^{2/3}B\mu_j}} \\ &\geqslant -\widetilde{C}\frac{\beta^{1/3}B^{1/2}}{n},\end{aligned}$$

  where $\widetilde{C} > 0$ is a constant. Here, since B is sufficiently large, Lemma 1.6 guarantees that the last inequality holds with probability at least 0.9.

Combining the two cases above, we arrive at the conclusion that for any $1-\theta_1^2 \leqslant \beta\frac{B^{3/2}}{n}$,

$$g(\theta_1) \geqslant -20\beta\frac{B}{n^{3/2}} - 2C\frac{\log n}{n} - \widetilde{C}\frac{\beta^{1/3}B^{1/2}}{n}.$$

Under the assumptions that $B \geqslant C_1(\log n)^2$ and $n \geqslant C_2 B^{3/2}$ for some sufficiently large constants $C_1, C_2 > 0$, we can choose β sufficiently small so as to make sure that

$$g(\theta_1) \geqslant -c^*\frac{B^{1/2}}{n} \qquad \text{for all} \quad 1-\theta_1^2 \leqslant \beta\frac{B^{3/2}}{n}.$$

1.7.3 Proofs of the bounds (1.33)

By definition, any function $h \in \mathcal{F}^\star$ obeys $\|h\|_{\mathcal{H}} \leqslant 3\|f^\star\|_{\mathcal{H}}$. In terms of the expansion $h = \sum_{j=1}^\infty \theta_j\phi_j$, this constraint is equivalent to the bound $\sum_{j=1}^\infty \theta_j^2/\mu_j \leqslant 9\|f^\star\|_{\mathcal{H}}^2$. In addition, the

constraint $\|h\|_Q \leqslant r$ implies that $\sum_{j=1}^{\infty} \theta_j^2 \leqslant r^2$. In conjunction, these two inequalities imply that

$$\sum_{j=1}^{\infty} \frac{\theta_j^2}{\min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\}} \leqslant 10,$$

as claimed in inequality (1.33b).

We now use this inequality to establish the bound (1.33a). For any $x \in \mathcal{X}$, we have

$$\begin{aligned} |h(x)| = \Big| \sum_{j=1}^{\infty} \theta_j \phi_j(x) \Big| &= \Big| \sum_{j=1}^{\infty} \frac{\theta_j}{\sqrt{\min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\}}} \cdot \sqrt{\min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\}} \phi_j(x) \Big| \\ &\overset{(i)}{\leqslant} \sqrt{\sum_{j=1}^{\infty} \frac{\theta_j^2}{\min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\}}} \cdot \sqrt{\sum_{j=1}^{\infty} \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\} \phi_j^2(x)} \\ &\overset{(ii)}{\leqslant} \sqrt{10 \sum_{j=1}^{\infty} \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\}}. \end{aligned}$$

Here step (i) uses the Cauchy–Schwarz inequality, whereas step (ii) follows from the previous claim (1.33b) and the assumption that $|\phi_j(x)| \leqslant 1$ for all $j \geqslant 1$.

1.7.4 Performance guarantees for LR-reweighted KRR

In this section, we present the performance guarantee for the LR-reweighted KRR estimate with truncation for all ranges of σ^2.

Similar to the large noise regime, we define

$$\mathcal{M}^{\mathsf{new}}(\delta) := c_0 \sqrt{\tfrac{\sigma^2 V^2 \log^3(n)}{n} \Psi(\delta, \mu)} \left(\sqrt{\tfrac{\Psi(\delta,\mu)}{\sigma^2}} + 1 \right).$$

Our theorem applies to any solution $\delta_n^{\mathsf{new}} > 0$ to the inequality $\mathcal{M}^{\mathsf{new}}(\delta) \leqslant \delta^2/2$.

Theorem 1.5. *Consider a kernel with sup-norm bounded eigenfunctions* (1.16)*, and a source-target pair with* $\mathbf{E}_P[\rho^2(X)] \leqslant V^2$*. Then the estimate* $\widehat{f}_\lambda^{\mathrm{rw}}$ *with truncation* $\tau_n = \sqrt{nV^2}$ *and regularization* $\lambda \|f^\star\|_{\mathcal{H}}^2 = \delta_n^2/3$ *satisfies the bound*

$$\|\widehat{f}_\lambda^{\mathrm{rw}} - f^\star\|_Q^2 \leqslant \delta_n^2$$

with probability at least $1 - c\, n^{-10}$*.*

Proof. Inspecting the proof of Theorem 1.4 (in particular, equation (1.32)), one has with high probability that

$$\sup_{g \in \mathcal{G}(\delta_n)} \Big\{ \|g - f^\star\|_Q^2 + \frac{1}{n} \sum_{i=1}^{n} \rho_{\tau_n}(x_i) \big[\big(f^\star(x_i) - y_i\big)^2 - \big(g(x_i) - y_i\big)^2 \big] \Big\} \leqslant \mathcal{M}^{\mathsf{new}}(\delta_n).$$

Repeating the analysis in Section 1.5.2 with $\delta_\lambda = \delta_n$ yields the desired claim. □

1.7.5 Expectation bounds for KRR estimates

In this section, we derive expectation bounds as counterparts to our previous high probability upper bounds on the KRR estimates. In Section 1.7.5.1, we present an expectation bound for instances with bounded likelihood ratios, essentially as a consequence of our previous high probability statement, given in Theorem 1.1. Similarly, in Section 1.7.5.3, we present an expectation bound for instances which have possibly unbounded likelihood ratios, but for which the second moment of the likelihood ratios is bounded. Again, this can be seen as an extension of our previous high-probability statement on the truncated, reweighted KRR estimator, as stated in Theorem 1.4.

1.7.5.1 Bounded likelihood ratio

Theorem 1.6. *Consider a covariate-shifted regression problem with likelihood ratio that is B-bounded* (1.2) *over a Hilbert space with a κ-uniformly bounded kernel* (1.7)*. There are universal constants $c_1, c_2 > 0$ such that if $\lambda \geqslant c_1 \frac{\kappa^2 \log n}{n}$, the KRR estimate $\widehat{f}_\lambda$ satisfies the bound*

$$\mathbf{E}\left[\|\widehat{f}_\lambda - f^\star\|_Q^2\right] \leqslant c_2 \Big\{ \lambda B \|f^\star\|_{\mathcal{H}}^2 + \frac{\sigma^2 B}{n} \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda B} + \frac{\sigma^2}{n} \Big\}.$$

Inspecting the proof, one may take $c_1 = 32, c_2 = \frac{519}{256}$. The proof of this result is presented in Section 1.7.5.2.

An immediate consequence is the following result for regular kernels. Note that it matches our lower bound (see Theorem 1.2), apart from logarithmic factors.

Corollary 1.3. *Suppose $\sigma^2 \geqslant \kappa^2$ and $\|f^\star\|_{\mathcal{H}} = 1$. For any $B \geqslant 1$ and any pair (P, Q) with B-bounded likelihood ratio* (1.2)*, any orthonormal basis $\{\phi_j\}_{j \geqslant 1}$ of $L^2(Q)$, and any regular sequence of kernel eigenvalues $\{\mu_j\}_{j \geqslant 1}$, there exist a universal constant $C > 0$ such that*

$$\mathbf{E}\left[\|\widehat{f}_\lambda - f^\star\|_Q^2\right] \leqslant C \inf_{\delta > 0} \Big\{ \delta^2 + \sigma^2 B d(\delta) \frac{\log n}{n} \Big\},$$

where above $\lambda = \delta_n^2$ where $\delta_n^2 = c \frac{\sigma^2 B d(\delta_n) \log n}{n}$ for a universal constant $c > 0$.

Proof. Following the proof of Corollary 1.1, we obtain from the KRR risk bound of Theorem 1.6,

$$\mathbf{E}\left[\|\widehat{f}_\lambda - f^\star\|_Q^2\right]\Big] \leqslant C_1 \Big\{ \delta^2 + \sigma^2 B d(\delta) \frac{\log n}{n} \Big\}, \quad \text{where} \quad \delta^2 = \lambda B,$$

for any $\delta^2 \geqslant c_1 B \kappa^2 \frac{\log n}{n}$. Adjusting constants so that $c \geqslant c_1$, our choice of δ_n^2 is valid since $\sigma^2 \geqslant \kappa^2$ and $d(\delta_n) \geqslant 1$. Moreover, since δ^2 is an increasing function of δ, whereas $d(\delta)$ is nonincreasing, under the choice of $\delta_n^2 = c \frac{\sigma^2 B d(\delta_n) \log n}{n}$, we have

$$\Big\{ \delta_n^2 + \sigma^2 B d(\delta_n) \frac{\log n}{n} \Big\} \leqslant C_2 \inf_{\delta > 0} \Big\{ \delta^2 + \sigma^2 B d(\delta) \frac{\log n}{n} \Big\},$$

for a universal constant $C_2 > 0$. Note that this inequality completes the proof of the result, with $C = C_1 C_2$. □

1.7.5.2 Proof of Theorem 1.6

Using Parseval's theorem and the optimality conditions for the KRR problem as given in equation (1.23), we have $\mathbf{E}[\|\widehat{f}_\lambda - f^\star\|_Q^2] \leqslant \mathbf{E}[T_1] + \mathbf{E}[T_2]$ where

$$T_1 \coloneqq \|\lambda(\widehat{\Sigma}_P + \lambda M^{-1})^{-1} M^{-1}\theta^\star\|_2^2, \quad \text{and} \quad T_2 \coloneqq \|(\widehat{\Sigma}_P + \lambda M^{-1})^{-1}\big(\frac{1}{n}\sum_{i=1}^n \xi_i \phi(x_i)\big)\|_2^2.$$

Recall the event

$$\mathcal{E}(\lambda) \coloneqq \Big\{ M^{1/2}\widehat{\Sigma}_P M^{1/2} + \lambda I \succcurlyeq \frac{1}{2}\big(M^{1/2}\Sigma_P M^{1/2} + \lambda I\big)\Big\},$$

as defined in equation (1.21). We use this event to bound the two terms.

Bound for T_1 Inspecting the proof of Theorem 1.1 (specifically, see the proof of bound (1.24)(a)), it follows that $\mathbf{E}[T_1 \mathbf{1}_{\mathcal{E}(\lambda)}] \leqslant 2\lambda B \|f^\star\|_{\mathcal{H}}^2$. On the other hand, from inequality (ii) of the proof of (1.24)(a), it also holds that

$$\mathbf{E}[T_1 \mathbf{1}_{\mathcal{E}(\lambda)^c}] \leqslant \|f^\star\|_{\mathcal{H}}^2 |||M|||_{\mathrm{op}}\, \mathbf{P}(\mathcal{E}(\lambda)^c) \leqslant \|f^\star\|_{\mathcal{H}} \kappa^2\, \mathbf{P}(\mathcal{E}(\lambda)^c).$$

The final inequality holds since $|||M|||_{\mathrm{op}} \leqslant \mathbf{Tr}(M) = \mathbf{E}_Q[\sum_j \mu_j \phi_j^2(x)] \leqslant \kappa^2$. Now, note that whenever $n\lambda \geqslant 32\kappa^2 \log n$, by Lemma 1.1 we have that

$$\begin{aligned}
\mathbf{E}[T_1 \mathbf{1}_{\mathcal{E}(\lambda)^c}] &\leqslant \|f^\star\|_{\mathcal{H}}^2\, \mathbf{P}(\mathcal{E}(\lambda)^c) \\
&\leqslant 28\lambda \|f^\star\|_{\mathcal{H}}^2 \Big[\big(\frac{\kappa^2}{\lambda}\big)^2 \exp\Big(-\frac{n\lambda}{16\kappa^2}\Big)\Big] \\
&\leqslant \frac{7}{256}\lambda \|f^\star\|_{\mathcal{H}}^2.
\end{aligned}$$

Putting the pieces together, we obtain

$$\mathbf{E}[T_1] \leqslant \frac{519}{256}\lambda \|f^\star\|_{\mathcal{H}}^2.$$

Bound for T_2 By considering the expectation over ξ_i conditional on the covariates and following algebraic manipulations similar to the proof of bound (1.24)(b), we have

$$\mathbf{E}[T_2] \leqslant \mathbf{E}[\widetilde{T}_2], \quad \text{where} \quad \widetilde{T}_2 \coloneqq \mathbf{Tr}\Big(\frac{\sigma^2}{n}(\widehat{\Sigma}_P + \lambda M^{-1})^{-1}\Big).$$

Moreover, inspecting the proof of bound (1.24)(b), we also have

$$\mathbf{E}[\widetilde{T}_2 \mathbf{1}_{\mathcal{E}(\lambda)}] \leqslant 2\frac{\sigma^2 B}{n} \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda B}.$$

On the other hand, by bounding $(\widehat{\Sigma}_P + \lambda M^{-1})^{-1} \preccurlyeq \lambda^{-1} M$,

$$\mathbf{E}[\widetilde{T}_2 \mathbf{1}_{\mathcal{E}(\lambda)^c}] \leqslant \frac{\sigma^2}{n}\frac{\kappa^2}{\lambda}\,\mathbf{P}(\mathcal{E}(\lambda)^c) \leqslant \frac{7}{256}\frac{\sigma^2}{n}.$$

The final inequality above is established in the same manner as in the proof of the bound for T_1 above, when $n\lambda \geqslant 32\kappa^2 \log n$. Thus, combining the two bounds,

$$\mathbf{E}[T_2] \leqslant 2\frac{\sigma^2 B}{n} \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda B} + \frac{7}{256}\frac{\sigma^2}{n}.$$

1.7.5.3 Unbounded likelihood ratio

Theorem 1.7. *Suppose $\sigma^2 \geqslant \kappa^2$ and $\|f^\star\|_{\mathcal{H}} = 1$. Consider a kernel with sup-norm bounded eigenfunctions (1.16), and a source-target pair with $\mathbf{E}_P[\rho^2(X)] \leqslant V^2$. Then, for any orthonormal basis $\{\phi_j\}_{j\geqslant 1}$ of $L^2(Q)$ and any regular sequence of kernel eigenvalues $\{\mu_j\}_{j\geqslant 1}$, there exists a universal constant $C > 0$ such that,*

$$\mathbf{E}\left[\|\widehat{f}_\lambda^{\mathrm{rw}} - f^\star\|_Q^2\right] \leqslant C \inf_{\delta>0}\Big\{\delta^2 + V^2 d(\delta)\frac{\log^3 n}{n}\Big\}.$$

Above, $3\lambda = \delta_n^2$ where δ_n^2 satisfies the equation $\delta^2 = c\frac{\sigma^2 V^2 \log^3 n}{n}$ for a universal constant $c > 0$.

Before giving the proof, we emphasize that—apart from logarithmic factors—this bound is minimax optimal.

Proof. By Theorem 1.4, there is an event $\mathcal{E}$ which has probability at least $1 - cn^{-10}$ such that the truncated, reweighted estimator $\widehat{f}_\lambda^{\mathrm{rw}}$ satisfies

$$\|\widehat{f}_\lambda^{\mathrm{rw}} - f^\star\|_Q^2 \leqslant c_1\delta^2,$$

provided we select $\lambda \asymp \delta^2 \asymp \frac{\sigma^2 V^2 \log^3(n) d(\delta)}{n}$. Note that under this choice of δ^2, we have

$$\delta^2 \asymp \inf_{\delta>0}\Big\{\delta^2 + \frac{\sigma^2 V^2 \log^3(n) d(\delta)}{n}\Big\}.$$

Consequently, there is a constant $c_2 > 0$ such that

$$\mathbf{E}\left[\|\widehat{f}_\lambda^{\mathrm{rw}} - f^\star\|_Q^2\right] \leqslant c_2 \inf_{\delta>0}\Big\{\delta^2 + \frac{\sigma^2 V^2 \log^3(n) d(\delta)}{n}\Big\} + \mathbf{E}\left[\|\widehat{f}_\lambda^{\mathrm{rw}} - f^\star\|_\infty^2 \mathbf{1}_{\mathcal{E}^c}\right]. \tag{1.59}$$

By Cauchy-Schwarz,

$$\|\widehat{f}_\lambda^{\mathrm{rw}} - f^\star\|_\infty^2 \leqslant \kappa^2 \|\widehat{f}_\lambda^{\mathrm{rw}} - f^\star\|_{\mathcal{H}}^2 \leqslant 2\kappa^2(1 + \|\widehat{f}_\lambda^{\mathrm{rw}}\|_{\mathcal{H}}^2).$$

Applying the optimality condition of the reweighted estimator $\widehat{f}_\lambda^{\mathrm{rw}}$, we have

$$\lambda \|\widehat{f}_\lambda^{\mathrm{rw}}\|_{\mathcal{H}}^2 \leqslant \lambda + \sqrt{nV^2}\frac{1}{n}\sum_{i=1}^n \xi_i^2.$$

Therefore, combining the previous two displays,

$$\|\widehat{f}_\lambda^{\mathrm{rw}} - f^\star\|_\infty^2 \leqslant 2\kappa^2\Big(2 + \frac{\sqrt{nV^2}}{\lambda}\frac{1}{n}\sum_{i=1}^n \xi_i^2\Big).$$

It then follows by Cauchy-Schwarz and the sub-Gaussianity of ξ_i, that for some constant $c_3 > 0$,

$$\begin{aligned}
\mathbf{E}\Big[\|\widehat{f}_\lambda^{\mathrm{rw}} - f^\star\|_\infty^2 \mathbf{1}_{\mathcal{E}^c}\Big] &\leqslant c_3\Big(\frac{\kappa^2}{n^{10}} + \frac{\sigma^2V^2}{\lambda n^4}\kappa^2\Big) \\
&\overset{\text{(i)}}{\leqslant} c_3\frac{\sigma^2V^2}{n}\Big(\frac{1}{n^9} + \frac{\kappa^2}{\lambda}\frac{1}{n^3}\Big) \\
&\overset{\text{(ii)}}{\leqslant} c_3\frac{\sigma^2V^2}{n}\Big(\frac{1}{n^9} + \frac{c_4}{n^2}\Big) \\
&\overset{\text{(iii)}}{\leqslant} c_5\frac{\sigma^2V^2}{n}
\end{aligned}$$

Above, inequality (i) uses $\sigma^2 \geqslant \kappa^2$ and $V^2 \geqslant 1$. Inequality (ii) uses the fact that $\lambda \asymp \delta^2 \asymp \frac{\sigma^2V^2\log^3(n)d(\delta)}{n} \gtrsim \frac{\kappa^2}{n}$. Finally, inequality (iii) follows by defining $c_5 \geqslant c_3(1 + c_4)$. This bound furnishes the result, since by applying it to the inequality (1.59), we obtain the result with $C = c_2 + c_5$. □

1.7.6 Performance of unweighted KRR with unbounded likelihood ratios

In this section, we present the performance guarantee of the unweighted KRR estimator when the likelihood ratios are unbounded.

Theorem 1.8. *Consider a covariate-shifted regression problem with likelihood ratios obeying* $\mathbf{E}_P[\rho^2(X)] \leqslant V^2$. *Then for any* $\lambda \geqslant 10\kappa^2/n$, *the KRR estimate* $\widehat{f}_\lambda$ *satisfies the bound*

$$\|\widehat{f}_\lambda - f^\star\|_Q^2 \leqslant 2\sqrt{\lambda V^2\kappa^2}\|f^\star\|_{\mathcal{H}}^2 + 40\frac{\sigma^2\log n}{n}\cdot\frac{\kappa^2}{\lambda}$$

with probability at least $1 - 28\,\frac{\kappa^2}{\lambda}e^{-\frac{n\lambda}{16\kappa^2}} - \frac{1}{n^{10}}$.

Simple algebra shows that the unweighted KRR estimator is still consistent for estimation under covariate shift, with a rate of $(\frac{\sigma^2V^2}{n})^{1/3}$ (ignoring κ^2 and log factors). However, unfortunately, this is far from optimal.

1.7.7 Proof of Theorem 1.8

In view of the proof of Theorem 1.1, we know that

$$\|\widehat{f}_\lambda - f^\star\|_Q^2 \leqslant 4\lambda \|f^\star\|_{\mathcal{H}}^2 \|\!| M^{1/2}(M^{1/2}\Sigma_P M^{1/2} + \lambda I)^{-1} M^{1/2} |\!\|_{\mathrm{op}} + 40\frac{\sigma^2 \log n}{n} \mathbf{Tr}\left(M^{1/2}(M^{1/2}\Sigma_P M^{1/2} + \lambda I)^{-1} M^{1/2}\right)$$

holds with probability at least $1 - 28\,\frac{\kappa^2}{\lambda} e^{-\frac{n\lambda}{16\kappa^2}} - \frac{1}{n^{10}}$. The proof is finished with the help of the following two bounds:

$$\|\!| M^{1/2}(M^{1/2}\Sigma_P M^{1/2} + \lambda I)^{-1} M^{1/2} |\!\|_{\mathrm{op}} \leqslant \frac{1}{2}\sqrt{\frac{V^2\kappa^2}{\lambda}}; \tag{1.60a}$$

$$\mathbf{Tr}\left(M^{1/2}(M^{1/2}\Sigma_P M^{1/2} + \lambda I)^{-1} M^{1/2}\right) \leqslant \frac{\kappa^2}{\lambda}. \tag{1.60b}$$

Proof of the bound (1.60b)**:** Note that $M^{1/2}(M^{1/2}\Sigma_P M^{1/2} + \lambda I)^{-1} M^{1/2} \preccurlyeq \lambda^{-1} M$. We therefore have

$$\mathbf{Tr}\left(M^{1/2}(M^{1/2}\Sigma_P M^{1/2} + \lambda I)^{-1} M^{1/2}\right) \leqslant \mathbf{Tr}(\lambda^{-1} M) \leqslant \frac{\kappa^2}{\lambda},$$

where the last relation uses the fact that $\mathbf{Tr}(M) \leqslant \kappa^2$.

Proof of the bound (1.60a)**:** We first make the observation that the bound (1.60a) is equivalent to

$$\Sigma_P + \lambda M^{-1} \succcurlyeq 2\sqrt{\frac{\lambda}{V^2\kappa^2}} I. \tag{1.61}$$

Therefore from now on, we focus on establishing the bound (1.61). Take an arbitrary vector θ with $\|\theta\|_2 = 1$. We have

$$\begin{aligned} 1 = \|\theta\|_2^2 &\overset{(i)}{=} \mathbf{E}_Q[(\theta^\mathsf{T}\phi(X))^2] \overset{(ii)}{=} \mathbf{E}_P[\rho(X)\cdot(\theta^\mathsf{T}\phi(X))^2] \\ &\overset{(iii)}{\leqslant} \sqrt{\mathbf{E}_P[\rho^2(X)]}\cdot\sqrt{\mathbf{E}_P[(\theta^\mathsf{T}\phi(X))^4]} \\ &\overset{(iv)}{=} \sqrt{V^2}\cdot\sqrt{\mathbf{E}_P[(\theta^\mathsf{T}\phi(X))^4]}. \end{aligned}$$

Here, the identity (i) follows from the fact that $\mathbf{E}_Q[\phi(X)\phi(X)^\mathsf{T}] = I$, the relation (ii) changes the measure from Q to P, the inequality (iii) is due to Cauchy-Schwarz, and the equality (iv) uses the definition of V^2. Apply the Cauchy-Schwarz inequality again to obtain

$$(\theta^\mathsf{T}\phi(X))^2 \leqslant \|M^{-1/2}\theta\|_2^2 \cdot \|M^{1/2}\phi(X)\|_2^2 \leqslant \kappa^2 \|M^{-1/2}\theta\|_2^2,$$

where the second inequality relies on the fact that $\sup_x \|M^{1/2}\phi(x)\|_2^2 \leqslant \kappa^2$. Take the above inequalities together to yield

$$\mathbf{E}_P[(\theta^\mathsf{T}\phi(X))^2] \geqslant \frac{1}{V^2\kappa^2 \cdot (\theta^\mathsf{T} M^{-1}\theta)} \qquad \text{for any } \theta \text{ with } \|\theta\|_2 = 1.$$

As a result, one has

$$\theta^\mathsf{T}(\Sigma_P + \lambda M^{-1})\theta \geqslant \frac{1}{V^2\kappa^2 \cdot (\theta^\mathsf{T} M^{-1}\theta)} + \lambda\theta^\mathsf{T} M^{-1}\theta \geqslant 2\sqrt{\frac{\lambda}{V^2\kappa^2}}.$$

Since this inequality holds for any unit-norm θ, we establish the claim (1.61).

1.7.8 Auxiliary lemmas

The following lemma provides concentration inequalities for the sum of independent self-adjoint operators, which appeared in the work [85].

Lemma 1.8. *Let $Z_1, Z_2, \ldots, Z_n$ be i.i.d. self-adjoint operators on a separable Hilbert space. Assume that $\mathbf{E}[Z_1] = 0$, and $\|\!|Z_1|\!\|_{\mathrm{op}} \leqslant L$ for some $L > 0$. Let V be a positive trace-class operator such that $\mathbf{E}[Z_1^2] \preccurlyeq V$, and $\|\!|\mathbf{E}[Z_1^2]|\!\|_{\mathrm{op}} \leqslant R$. Then one has*

$$\mathbf{P}\Big(\|\!|\frac{1}{n}\sum_{i=1}^n Z_i|\!\|_{\mathrm{op}} \geqslant t\Big) \leqslant \frac{28\,\mathbf{Tr}(V)}{R} \cdot \exp\Big(-\frac{nt^2/2}{R + Lt/3}\Big), \qquad \text{for all } t \geqslant \sqrt{R/n} + L/(3n).$$

Next, we turn attention to bounding the maxima of empirical processes. Let $X_1, X_2, \ldots, X_n$ be independent random variables. Let $\mathcal{F}$ be a countable class of functions uniformly bounded by b. Assume that for all i and all $f \in \mathcal{F}$, $\mathbf{E}[f(X_i)] = 0$. We are interested in controlling the random variable $Z := \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)$, for which the variance statistics $v^2 := \sup_{f\in\mathcal{F}} \mathbf{E}[\sum_{i=1}^n (f(X_i))^2]$ is crucial. Now we are in position to state the classical Talagrand's concentration inequalities; see the paper [66].

Lemma 1.9. *For all $t > 0$, we have*

$$\mathbf{P}(Z \geqslant \mathbf{E}[Z] + t) \leqslant \exp\Big(-\frac{t^2}{2(v^2 + 2v\,\mathbf{E}[Z]) + 3vt}\Big).$$

Chapter 2

Covariate shift over Hölder smoothness classes

2.1 Introduction

In the standard formulation of prediction or classification, future data (as represented by a test set) is assumed to be drawn from the same distribution as the training data. This assumption, while theoretically convenient, may fail to hold in many real-world scenarios. For instance, training data might be collected only from a sub-group within a broader population (such as in medical trials), or the environment might change over time as data are collected. Such scenarios result in a distribution mismatch between the training and test data.

In this chapter, we study an important case of such distribution mismatch—namely, the covariate shift problem (e.g., [105, 98]). Suppose that a statistician observes covariate-response pairs (X, Y), and wishes to build a prediction rule. In the problem of covariate shift, the distribution of the covariates X is allowed to change between the training and test data, while the posterior distribution of the responses (namely, $Y \mid X$) remains fixed. Compared to the usual i.i.d. setting, this serves as a more accurate model for a variety of real-world applications, including image classification [102], biomedical engineering [74], sentiment analysis [13], and audio processing [56], among many others.

More formally, suppose that the statistician observes n_P covariates $\{X_i\}_{i=1}^{n_P}$ from a *source distribution* P, and n_Q covariates $\{X_i\}_{i=n_P+1}^{n_Q+n_P}$ from a *target distribution* Q. For each observed X_i, she also observes a response Y_i drawn from the same conditional distribution. The *regression function* $f^\star(x) = \mathbf{E}[Y \mid x]$ defined by this conditional distribution is assumed to lie in some function class $\mathcal{F}$. The statistician uses these samples to produce an estimate $\widehat{f}$, which will be evaluated on the target distribution, with a fresh sample $X \sim Q$, yielding the mean-squared error

$$\|\widehat{f} - f^\star\|_{L^2(Q)}^2 \coloneqq \mathbf{E}\left[\left(\widehat{f}(X) - f^\star(X)\right)^2\right].$$

When there is no covariate shift, the fundamental (minimax) risks for this problem are well-understood [55, 60, 110]. The goal of this chapter is to understand how, for nonparametric

function classes $\mathcal{F}$, this minimax risk changes as a function of the "amount" of covariate shift between P and Q.

2.1.1 Contributions and related work

Let us summarize the main contributions of this chapter, and put them in the context of related work.

Contributions. We introduce a similarity measure[1] ρ_h between two probability measures P, Q on a common metric space $(\mathcal{X}, d)$. For any level $h > 0$, it is defined as

$$\rho_h(P, Q) \coloneqq \int_{\mathcal{X}} \frac{1}{P\big(\mathsf{B}(x, h)\big)} \, \mathrm{d}Q(x), \tag{2.1}$$

where $\mathsf{B}(x, h) \coloneqq \{ x' \in \mathcal{X} \mid d(x, x') \leqslant h \}$ is the closed ball of radius h centered around x. We substantiate the significance of this similarity measure via the following contributions:

(i) For regression functions that are Hölder continuous, we demonstrate a performance guarantee for the Nadaraya-Watson kernel estimator under covariate shift that is fully determined by the scaling of the similarity measure $\rho_h(P, Q)$ with respect to the radius h.

(ii) We complement these upper bounds with matching lower bounds—in a minimax sense—demonstrating that the best achievable rate of estimation in Hölder classes is also determined by the scaling of this similarity measure.

(iii) We show how the similarity measure ρ_h can be controlled based on the metric properties of the space $\mathcal{X}$. In addition, we compare ρ_h with existing notions for covariate shift (e.g., bounded likelihood ratios, transfer exponents), thereby showcasing some of its advantages.

Related work. The problem of covariate shift was studied in the seminal work by Shimodaira [105], who provided asymptotic guarantees for a weighted maximum likelihood estimator under covariate shift. Since then, a plethora of work has analyzed covariate shift, or the general distribution mismatch problem (also referred to as domain adaptation or transfer learning).

For general distribution mismatch, one line of work provides rates that depend on distance metrics between the source-target pair (e.g., [6, 7, 45, 79, 28, 86]). These results hold under fairly general conditions, but do not necessarily guarantee consistency as the sample size n increases. In contrast, our guarantees for covariate shift do guarantee consistency, and

[1] To be clear, this quantity actually serves as a *dis*-similarity measure: as shown in the sequel, source-target pairs (P, Q) with larger values $\rho_h(P, Q)$ lead to "harder" estimation problems in terms of covariate shift.

moreover, we provide explicit nonasymptotic, optimal nonparametric rates. As pointed out in the paper [69], the distribution mismatch problem is asymmetric in the sense that it may be easier to estimate accurately when dealing with covariate shift from P to Q than from Q to P. Our results also corroborate this intuition. It is worth noting that these prior distance metrics fall short of capturing the inherent asymmetry between P and Q.

Another line of work addresses covariate shift under conditions on the likelihood ratio $\mathrm{d}Q/\mathrm{d}P$. For instance, some authors have obtained results for bounded likelihood ratios [113, 68] or in terms of information-theoretic divergences between the source-target pair [112, 80]. Our work is inspired in part by the work of Kpotufe and Martinet [69], who introduced the notion of the *transfer exponent*. It is a condition that bounds the mass placed by the pair (P, Q) on balls of varying radii; using this notion, they analyzed various problems of nonparametric classification. Our work, focusing instead on nonparametric regression problems and using the measure ρ_h, provides sharper rates than those obtainable by considering the transfer exponent; see Section 2.3.2 for details. Thus, the similarity measure ρ_h provides a more fine-grained control on the effect of covariate shift on nonparametric regression.

Finally, it is worth mentioning other recent works that give risk bounds for covariate shift problems, including on linear models [73], as well as linear models and one-layer neural networks [90]. Although these results deal with covariate shift, the rates obtained are parametric ones, and hence not directly comparable to the nonparametric rates that are the focus of our inquiry.

2.1.2 Notation

Here we collect notation used throughout the chapter. We use $\mathbf{R}$ to denote the real numbers. We use $(\mathcal{X}, d)$ to denote a metric space, and we equip it with the usual Borel σ-algebra. We let $\mathsf{B}(x, r) \coloneqq \{ x' \in \mathcal{X} \mid d(x, x') \leqslant r \}$ be the closed ball of radius r centered at x. We reserve the capital letters X, Y, possibly with subscripts, for a pair of random variables arising from a regression model. Similarly, we reserve P, Q for a pair of two probability measures on $(\mathcal{X}, d)$. For $h > 0$, we denote by $N(h)$ the covering number of $\mathcal{X}$ at resolution h in the metric d. This is the minimal number of balls of radius at most $h > 0$ required to cover the space $\mathcal{X}$.

The remainder of this chapter is organized as follows. We begin in Section 2.2 by setting up the problem more precisely, and stating and discussing our main results on covariate shift: namely, upper bounds in Theorem 2.1, accompanied by matching lower bounds in Theorem 2.2. These results establish that the similarity measure (2.1) provides a useful measure of the "difficulty" of source-target pairs in covariate shift; accordingly, Section 2.3 is devoted to a comparison and discussion of this measure relevant to concepts from past work, including likelihood ratio bounds and transfer exponents. The proofs of all our results are given in Section 2.4, and we conclude with a discussion in Section 2.5.

2.2 Characterizing Hölder-smooth regression under covariate shift

In this section, we use the similarity measure introduced in equation (2.1) to characterize how covariate shift can change the minimax risks of estimation for certain classes of nonparametric regression models. We begin in Section 2.2.1 by setting up the observation model to be considered, along with some associated assumptions on the regression function $f^\star$, the conditional distribution of $Y \mid X$, and the covariate shift. In Section 2.2.2, we derive an achievable result (Theorem 2.1) for nonparametric regression in the presence of covariate shift, in particular via a careful analysis of the classical Nadaraya-Watson estimator. Our upper bound in this section is general, and illustrates the key role of the similarity measure ρ_h. In Section 2.2.3, we introduce the α-families of source-target pairs (P, Q), and use Theorem 2.1 to derive achievable results for these families. In Section 2.2.4, we state some complementary lower bounds for α-families (Theorem 2.2), showing that our achievable results are, in fact, unimprovable.

2.2.1 Observation model and assumptions

Suppose that we observe covariate-response pairs $\{(X_i, Y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathbf{R}$ that are drawn from nonparametric regression model of the following type. The conditional distribution of $Y \mid X$ is the same for all $i = 1, \ldots, n$, and our goal is to estimate the regression function $f^\star(x) \coloneqq \mathbf{E}[Y \mid X = x]$. In terms of the "noise" variables, $w_i \coloneqq Y_i - f^\star(X_i)$, the observations can be written in the form

$$Y_i = f^\star(X_i) + w_i, \quad i = 1, \ldots, n. \tag{2.2}$$

In our analysis, we impose three types of regularity conditions: (i) Hölder continuity of the regression function; (ii) the type of covariate shift allowed; and (iii) tail conditions on the noise variables $\{w_i\}_{i=1}^n$.

Assumption 2.1 (Hölder continuity)**.** For some $L > 0$ and $\beta \in (0, 1]$, the function $f^\star \colon \mathcal{X} \to \mathbf{R}$ is (β, L)-Hölder continuous, meaning that

$$\left|f^\star(z) - f^\star(z')\right| \leqslant L \left[d(z, z')\right]^\beta, \quad \text{for any } z, z' \in \mathcal{X}.$$

We note that in the special case $\beta = 1$, the function $f^\star$ is L-Lipschitz.

Assumption 2.2 (Covariate shift)**.** The covariates $X_1, \ldots, X_n$ are independent, and drawn as

$$X_1, \ldots, X_{n_P} \overset{\text{i.i.d.}}{\sim} P \quad \text{and} \quad X_{n_P+1}, \ldots X_{n_P+n_Q} \overset{\text{i.i.d.}}{\sim} Q \qquad \text{where } n = n_P + n_Q.$$

Assumption 2.3 (Noise assumption)**.** The variables $\{w_i\}_{i=1}^n$ satisfy the second moment bound

$$\sup_x \mathbf{E}\left[w_i^2 \mid X_i = x\right] \leqslant \sigma^2 \qquad \text{for } i = 1, \ldots, n.$$

Note that by construction, the variables w_i are (conditionally) centered. Assumption 2.3 also allows w_i to depend on X_i, as long as the variance is uniformly bounded above.

2.2.2 Achievable performance via the Nadaraya-Watson estimator

We first exhibit an achievable result for the problem of nonparametric regression in the presence of covariate shift. We do so by analyzing a classical and simple method for nonparametric estimation, namely the Nadaraya-Watson estimator [91, 122], or NW for short. The main result of this section is to show that the mean-squared error (MSE) of the NW estimator is upper bounded by a bias-variance decomposition that also involves the similarity measure ρ_h.

We begin by recalling the definition of the NW estimator, focusing here on the version in which the underlying kernel is uniform over a ball of a given bandwidth $h_n > 0$. In particular, define the set

$$\mathcal{G}_n := \bigcup_{i=1}^{n} \mathsf{B}(X_i, h_n),$$

corresponding to the set of points in $\mathcal{X}$ within distance h_n of the observed covariates. In terms of this set, the *Nadaraya-Watson estimator* $\widehat{f}$ takes the form

$$\widehat{f}(x) := \begin{cases} \frac{\sum_{i=1}^n Y_i \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}} & \text{for } x \in \mathcal{G}_n \\ 0 & \text{otherwise.} \end{cases}$$

Our first main result provides an upper bound on the MSE of the NW estimator under covariate shift; this bound exhibits the significance of the similarity measure (2.1). It involves the distribution $\mu_n := \frac{n_P}{n} P + \frac{n_Q}{n} Q$, which is a convex combination of the source and target distributions weighted by their respective fractions of samples.

Theorem 2.1. *Suppose that Assumptions 2.1, 2.2, and 2.3 hold. For any $h_n > 0$, the Nadaraya-Watson estimator $\widehat{f}$ with bandwidth h_n has MSE bounded as*

$$\mathbf{E} \left\| \widehat{f} - f^\star \right\|_{L^2(Q)}^2 \leqslant c_u \Big\{ L^2 h_n^{2\beta} + \frac{\|f^\star\|_\infty^2 + \sigma^2}{n} \rho_{h_n}(\mu_n, Q) \Big\}, \tag{2.3}$$

where $c_u > 0$ is a numerical constant.

See Section 2.4.1 for a proof of this result.

Note that the bound (2.3) exhibits a type of bias-variance trade-off, one that controls the optimal choice of bandwidth h_n. The quantity $h_n^{2\beta}$ in the first term is familiar from the classical analysis of the NW estimator; it corresponds to the bias induced by smoothing over balls of radius h_n, and hence is an increasing function of bandwidth. In the second term, the bandwidth appears in the similarity measure $\rho_{h_n}(\mu_n, Q)$, which is a non-increasing function

of the bandwidth. The optimal choice of bandwidth arises from optimizing this tradeoff; note that it depends on the pair (P, Q), as well as the sample sizes (n_P, n_Q), via the similarity measure applied to the convex combination μ_n and Q.

No covariate shift: As a sanity check, it is worth checking that the bound (2.3) recovers known results in the case of no covariate shift ($P = Q$ and hence $\mu_n = Q$). As a concrete example, if Q is uniform on the hypercube $[0,1]^k$, it can be verified that $\rho_h(Q, Q) \asymp h^{-k}$ as $h \to 0^+$. (See Example 2.2 in the sequel for a more general calculation that implies this fact.) Thus, if we track only the sample size, the optimal bandwidth is given by $h_n^* = n^{-\frac{1}{2\beta+k}}$, and with this choice, the bound (2.3) implies that the NW estimator has MSE bounded as $n^{-\frac{2\beta}{2\beta+k}}$. Thus, we recover the classical and known results in this special case. As we will see, more interesting tradeoffs arise in the presence of covariate shift, so that $\mu_n \neq Q$.

2.2.3 Consequences for α-families of source-target pairs

In order to better understand the bias-variance tradeoff in the bound (2.3) in the presence of covariate shift, it is helpful to derive some explicit consequences of Theorem 2.1 for a particular function class $\mathcal{F}$, along with certain families of source-target pairs (P, Q). The latter families are indexed by a parameter $\alpha > 0$ that controls the amount of covariate shift; accordingly, we refer to them as α-families.

So as to simplify our presentation, we assume that $\mathcal{X}$ is the unit interval $[0, 1]$. For a given pair $\beta \in (0, 1]$ and $L > 0$, consider the class of regression functions

$$\mathcal{F}(\beta, L) = \Big\{ f \colon [0,1] \to \mathbf{R} \mid |f(x) - f(x')| \leqslant L|x - x'|^\beta, \text{for all } x, x' \in \mathcal{X},\ f(0) = 0 \Big\}.$$

This is a special case of β-Hölder continuous functions when the underlying metric space is the unit interval $[0, 1]$ equipped with the absolute value norm. The additional constraint $f(0) = 0$ ensures that this class has finite metric entropy.

Next we introduce some interesting families of source-target pairs.

α-families of (P, Q) pairs: For a given parameter $\alpha \geqslant 1$ and radius $C \geqslant 1$, we define the set of source-target pairs[2]

$$\mathcal{D}(\alpha, C) \coloneqq \Big\{ (P, Q) \mid \sup_{0<h\leqslant 1} h^\alpha \rho_h(P, Q) \leqslant C \Big\}. \tag{2.4a}$$

In words, these are source target pairs for which the growth of the similarity as $h \to 0^+$ is at most $h^{-\alpha}$. In the case $\alpha \in (0, 1]$, we define the related set

$$\mathcal{D}'(\alpha, C) \coloneqq \Big\{ (P, Q) \mid \sup_{0<h\leqslant \Delta} h^\alpha \rho_h(P, Q) \leqslant C, \sup_{0<h\leqslant 1} \rho_h(Q, Q) \leqslant C \Big\},$$

[2]Note that the restriction of the supremum to $h \in [0, 1]$ is necessary, as $\rho_h(P, Q) = 1$ for all $h \geqslant 1$. Note also that since $\rho_1(P, Q) = 1$, one necessarily has $C \geqslant 1$.

where the additional condition is added to address the fact that even without covariate shift, the rate $n^{-2\beta/(2\beta+1)}$ is unimprovable for some distributions [110]. Taking into account the first part of the next corollary, it is necessary to impose some condition on the target distribution in order to obtain significantly faster rates such as $n^{-\frac{2\beta}{2\beta+\alpha}}$, when $\alpha < 1$.

Corollary 2.1. *Suppose that $\sigma \geqslant L$, and that Assumptions 2.2 and 2.3 hold. Then there exists a constant $c_u' > 0$, independent of n, n_P, n_Q, σ^2, and an integer $n_u \coloneqq n_u(\sigma, \beta, L, \alpha, C)$ such that, provided that $\max\{n_P, n_Q\} \geqslant n_u$:*

(a) *For $\alpha \geqslant 1$ and $C \geqslant 1$, we have*

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{F}(\beta, L)} \mathbf{E} \left\| \widehat{f} - f^\star \right\|_{L^2(Q)}^2 \leqslant c_u' \left\{ \left(\frac{n_P}{\sigma^2}\right)^{\frac{2\beta+1}{2\beta+\alpha}} + \left(\frac{n_Q}{\sigma^2}\right) \right\}^{-\frac{2\beta}{2\beta+1}} \quad \text{for any } (P, Q) \in \mathcal{D}(\alpha, C). \tag{2.5a}$$

(b) *For $\alpha \in (0, 1]$ and $C \geqslant 1$, we have*

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{F}(\beta, L)} \mathbf{E} \left\| \widehat{f} - f^\star \right\|_{L^2(Q)}^2 \leqslant c_u' \left\{ \left(\frac{n_P}{\sigma^2}\right)^{\frac{2\beta}{2\beta+\alpha}} + \left(\frac{n_Q}{\sigma^2}\right) \right\}^{-1} \quad \text{for any } (P, Q) \in \mathcal{D}'(\alpha, C).$$

See Section 2.4.2 for a proof of this corollary.

Let us discuss the bound (2.5a) to gain some intuition. The special case of no covariate shift can be captured by setting $n_P = 0$ and $n_Q > 0$, and we recover the familiar $n^{-\frac{2\beta}{2\beta+k}}$ rate previously discussed. At the other extreme, suppose that $n_Q = 0$ so that all of our samples are from the shifted distribution (i.e., $n = n_P$); in this case, the MSE is bounded as $(\sigma^2/n)^{-\frac{2\beta}{2\beta+\alpha}}$. As α increases, our set-up allows for more severe form of covariate shift, and its deleterious effect is witnessed by the exponent $\frac{2\beta}{2\beta+\alpha}$ shrinking towards zero. Thus, the NW estimator—with an appropriate choice of bandwidth—remains consistent but with an arbitrarily slow rate as α diverges to $+\infty$.

There are many papers in the literature (e.g., [113, 68]) that discuss the covariate shift problem when the likelihood ratio is bounded—that is, when Q is absolutely continuous with respect to P and $\sup_{x \in \mathcal{X}} \frac{\mathrm{d}Q}{\mathrm{d}P}(x) \leqslant b$ for some $b \geqslant 1$. We say that the pair (P, Q) are b-bounded in this case.

Example 2.1 (Bounded likelihood ratio)**.** Suppose that $\mathcal{X} = [0, 1]^d$ with the Euclidean metric, and consider a pair (P, Q) with b-bounded likelihood ratio. In this special case, our general theory yields bounds in terms of the b-weighted *effective sample size*

$$n_{\text{eff}}(b) \coloneqq \frac{n_P}{b} + n_Q.$$

In particular, it follows from the proof of Corollary 2.1 that in the regime $\sigma^2 \geqslant L^2$, we have the upper bound

$$\mathbf{E}\,\|\widehat{f} - f^\star\|^2_{L^2(Q)} \leqslant c'_u \Big(\frac{\sigma^2}{n_{\text{eff}}(b)}\Big)^{\frac{2\beta}{2\beta+d}},$$

provided that $n_{\text{eff}}(b)$ is large enough. Consequently, the effect of covariate shift with b-bounded pairs is to reduce n_P to n_P/b. Again, we recover the standard rate $(\frac{\sigma^2}{n})^{\frac{2\beta}{2\beta+d}}$ in the case of no covariate shift (or equivalently, when $b = 1$). This recovers a known result and is minimax optimal. ♣

2.2.4 Matching lower bounds for α-families

Thus far, we have seen that the similarity measure ρ_h plays a central role in determining the estimation error of the NW estimator under covariate shift. However, this is just one of many possible estimators in nonparametric regression. Does this similarity measure play a more fundamental role? In this section, we answer this question in the affirmative by proving minimax lower bounds for covariate shift problems parameterized in terms of bounds on ρ_h. In order to do so, we consider the metric space $\mathcal{X} = [0, 1]$ equipped with the absolute value as the metric.

The main result of this section provides lower bounds on the mean-squared error of any estimator, when measured uniformly over functions in the Hölder class $\mathcal{F}(\beta, L)$, along with target-source pairs (P, Q) belonging to the class $\mathcal{D}(\alpha, C)$ when $\alpha \geqslant 1$ and the class $\mathcal{D}'(\alpha, C)$ when $\alpha < 1$.

Theorem 2.2. *Suppose that Assumptions 2.2 and 2.3 hold. Then there is a constant $c_\ell > 0$, independent of n, n_P, n_Q, σ^2, and an integer $n_\ell := n_\ell(\sigma, L, C, \alpha, \beta)$ such that for all sample sizes $\max\{n_P, n_Q\} \geqslant n_\ell$:*

(a) *For $\alpha > 1$ and $C \geqslant 1$, there is a pair of distributions $(P, Q) \in \mathcal{D}(\alpha, C)$ such that*

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{F}(\beta,L)} \mathbf{E}\,\|\widehat{f} - f^\star\|^2_{L^2(Q)} \geqslant c_\ell \Big\{ \big(\frac{n_P}{\sigma^2}\big)^{\frac{2\beta+1}{2\beta+\alpha}} + \big(\frac{n_Q}{\sigma^2}\big) \Big\}^{-\frac{2\beta}{2\beta+1}}. \tag{2.6a}$$

(b) *For $\alpha \leqslant 1$ and $C \geqslant 1$, there is a pair of distributions $(P, Q) \in \mathcal{D}'(\alpha, C)$ such that*

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{F}(\beta,L)} \mathbf{E}\,\|\widehat{f} - f^\star\|^2_{(Q)} \geqslant c_\ell \Big\{ \big(\frac{n_P}{\sigma^2}\big)^{\frac{2\beta}{2\beta+\alpha}} + \big(\frac{n_Q}{\sigma^2}\big) \Big\}^{-1}.$$

See Sections 2.4.3 and 2.4.4 for the proof of this result.

These lower bounds should be compared to Corollary 2.1. This comparison shows that the MSE bounds achieved by the NW estimator are actually optimal in the minimax sense over families defined by the similarity measure ρ_h.

2.3 Properties of the similarity measure

In the previous sections, we have seen that the similarity measure ρ_h controls both the behavior of the NW estimator, as well as fundamental (minimax) risks applicable to any estimator. Thus, it is natural to explore the similarity measure in some more detail, and in particular to draw some connections to existing notions in the literature.

2.3.1 Controlling ρ_h via covering numbers

We start with a general way of controlling the similarity measure ρ_h, which is based on the covering number of the metric space $(\mathcal{X}, d)$. In particular, for any $h > 0$, the *covering number* $N(h)$ is defined to be the smallest number of balls of radius h needed to cover the space $\mathcal{X}$. See Chapter 5 in the book [120] for more background.

Proposition 2.1 (Covering number bounds for the similarity measure)**.** *Suppose that P, Q are two probability measures on the same metric space $(\mathcal{X}, d)$. Suppose that for some $h > 0$, there is a $\lambda > 0$ such that*

$$P(\mathsf{B}(x, h)) \geqslant \lambda \, Q(\mathsf{B}(x, h)) \quad \textit{for all } x \in \mathcal{X}. \tag{2.7}$$

Then the similarity at scale h is upper bounded as $\rho_h(P, Q) \leqslant N(\frac{h}{2})/\lambda$.

See Section 2.4.5 for the proof of this claim.

It is worth emphasizing that—due to the order of quantifiers above—the quantity $\lambda > 0$ is allowed to depend on $h > 0$. We exploit this fact in subsequent uses of the bound (2.7).

One straightforward application of Proposition 2.1 is in bounding the similarity measure when there is no covariate shift, as we now discuss.

Example 2.2 (No covariate shift)**.** Suppose that we compute the similarity measure in the case $P = Q$; intuitively, this models a scenario where there is no covariate shift. In this case, we clearly may apply Proposition 2.1 with $\lambda = 1$, which reveals that $\rho_h(P, P) \leqslant N(h/2)$. To give one concrete bound, suppose that $\mathcal{X} \subset \mathbf{R}^d$ is a compact set, with diameter D. Then—owing to standard bounds on covering number [120, chap. 5]—we obtain $\rho_h(P, P) \leqslant (1 + \frac{2D}{h})^d$. Note that this bound holds for any metric, so long as the diameter D is computed with the same metric as the balls in the definition of the similarity measure. ♣

We give another application of Proposition 2.1 in the following subsection.

2.3.2 Comparison to previous notions of distribution mismatch

Next, we show how the mapping $h \mapsto \rho_h(P, Q)$ can be bounded naturally using previously proposed notions of distribution mismatch for covariate shift. Again, Proposition 2.1 plays a central role.

Example 2.3 (Bounded likelihood ratio)**.** Suppose that P, Q are such that $Q \ll P$ and the likelihood ratio $\frac{\mathrm{d}Q}{\mathrm{d}P}(x) \leqslant b$, for all $x \in \mathcal{X}$. Then note that by a simple integration argument $P(\mathsf{B}(x,h)) \geqslant \frac{1}{b} Q(\mathsf{B}(x,h))$. Therefore, we conclude $\rho_h(P,Q) \leqslant bN(h/2)$. ♣

As noted previously, our work was inspired by the transfer exponent introduced by Kpotufe and Martinet [69] in the context of covariate shift for nonparametric regression. It is worth comparing these notions so as to understand in what sense the similarity measure ρ_h is a refinement of the transfer exponent. In order to simplify this discussion, we focus here on the special case $\mathcal{X} = [0,1]$.

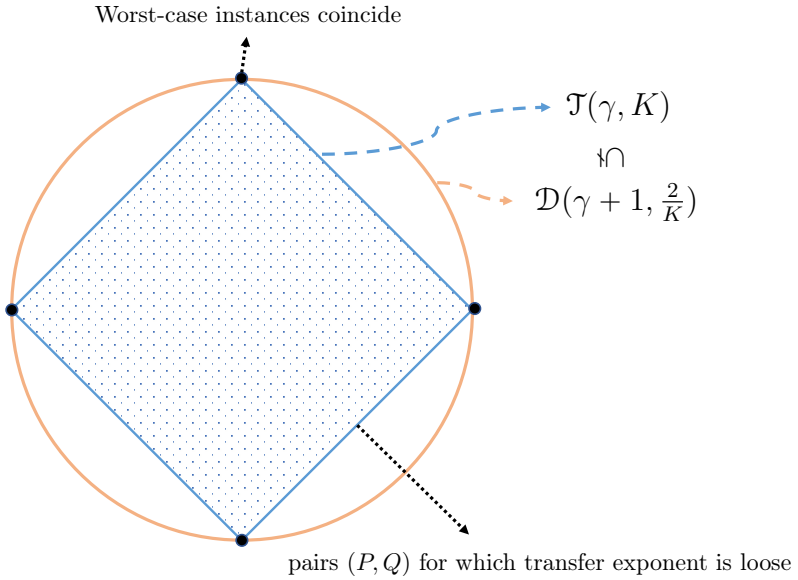


Figure 2.1. The yellow circle depicts the contour for the class $\mathcal{D}(\gamma+1, \frac{2}{K})$, while the blue square plots the contour for the class $\mathcal{T}(\gamma, K)$. It can be seen from Lemma 2.1 and Example 5 that $\mathcal{T}(\gamma, K)$ is strict subset of $\mathcal{D}(\gamma+1, \frac{2}{K})$. In addition, our lower bound shows that under covariate shift, the worst-case instances for both classes coincide with each other. However, there exist instances (P, Q) where the characterization using transfer exponent is intrinsically loose.

We begin by providing the definition of transfer exponent:

Definition 2.1 (Transfer exponent [69])**.** The distributions (P, Q) have transfer exponent $\gamma \geqslant 0$ with constant $K \in (0, 1]$ if

$$P(\mathsf{B}(x,h)) \geqslant K h^{\gamma} Q(\mathsf{B}(x,h)) \qquad \text{for all } x \text{ in the support of } Q.$$

We denote by $\mathcal{T}(\gamma, K)$ the set of all pairs (P, Q) with this property.

It is natural to ask how the set $\mathcal{T}(\gamma, K)$ is related to the α-family previously defined in equation (2.4a). The following result establishes an inclusion:

Lemma 2.1. *For $\mathcal{X} = [0,1]$ and any $\gamma \geqslant 0$ and $K \in (0,1]$, we have the inclusion*

$$\mathcal{T}(\gamma, K) \subset \mathcal{D}(\gamma + 1, \tfrac{2}{K}). \tag{2.8}$$

The proof of this inclusion is given in Section 2.4.6. At a high level, it exploits Proposition 2.1 to show that for any $(P,Q) \in \mathcal{T}(\gamma, K)$, we have the bound $\rho_h(P,Q) \leqslant \frac{1}{Kh^\gamma} N(h/2)$.

From the inclusion (2.8), it follows that any covariate shift instance (P,Q) with finite transfer exponent $\gamma \geqslant 0$ belongs to an α-similarity family with $\alpha = \gamma + 1$. In fact, following a proof similar to that of Theorem 2.2, we can show that for $\gamma \geqslant 0$, there is pair (P,Q) in the class $\mathcal{T}(\gamma, K)$ such that the minimax risk for β-Hölder-continous functions scales as $n_P^{-\frac{2\beta}{2\beta+\gamma+1}}$. Note that this risk bound coincides with the minimax risk associated with the class $\mathcal{D}(\gamma+1, \frac{2}{K})$. In other words, from a *worst case* point of view, the source-target class $\mathcal{T}(\gamma, K)$ is equally as hard as the class $\mathcal{D}(\gamma + 1, \frac{2}{K})$ for nonparametric regression under covariate shift. However, this worst case equivalence does not capture the full picture: there are many covariate shift families for which the transfer exponent provides an overly conservative prediction, and so does not capture the fundamental difficulty of the problem. Let us consider a concrete example to illustrate.

Example 2.4 (Separation between transfer exponent and ρ_h)**.** Let the target distribution Q be a uniform distribution on the interval $[0,1]$, and for some $\kappa \geqslant 1$, suppose that the source distribution P has density $p(x) = (\kappa+1)x^\kappa$ for $x \in [0,1]$. With these definitions, it can be verified that $(P,Q) \in \mathcal{T}(\kappa, K)$ for some constant $K \in (0,1]$, and moreover, that the quantity κ is the *smallest possible* transfer exponent for this pair. In contrast, another direct computation shows that the pair (P,Q) belongs to the class $\mathcal{D}(\kappa, C')$ for some constant $C' > 0$. These two inclusions establish a separation between the rates predicted by the transfer exponent and the similarity ρ_h. Indeed, as shown by our theory, the difficulty of estimation over $\mathcal{D}(\kappa, C')$ is smaller than that prescribed by $\mathcal{T}(\kappa, K)$. Indeed, if one observe n samples from the source distribution, the worst-case rate indicated by the computation from the transfer exponent is $n^{-\frac{2\beta}{2\beta+\kappa+1}}$, whereas the rate guaranteed by the similarity measure ρ_h is $n^{-\frac{2\beta}{2\beta+\kappa}}$. As an explicit example, Lipschitz functions ($\beta = 1$) and $\kappa = 1$, we obtain the slower rate $n^{-1/2}$ versus the faster rate $n^{-2/3}$, so that the ratio between the two rates diverges as $n^{1/6}$ as the sample size grows. ♣

See also Figure 2.1 for an illustration of the connections and differences between the similarity measure and the transfer exponent.

2.4 Proofs

We now turn to the proofs of the results stated in the previous section.

2.4.1 Proof of Theorem 2.1

Recall that the estimate $\widehat{f}$ depends on the observations $\{(X_i, Y_i)\}_{i=1}^n$, and so should be understood as a random function. The core of the proof involves proving that, for each $x \in \mathcal{X}$, we have

$$\mathbf{E}\left[\left(\widehat{f}(x) - f^\star(x)\right)^2\right] \leqslant L^2 h_n^{2\beta} + \frac{4\sigma^2 + \|f^\star\|_\infty^2}{n} \frac{1}{\mu_n(\mathsf{B}(x, h_n))}, \tag{2.9}$$

where the expectation is taking over the observations $\{(X_i, Y_i)\}_{i=1}^n$. Given this inequality, the claim (2.3) of Theorem 2.1 follows, since by Fubini's theorem, we can write

$$\mathbf{E}\left[\|\widehat{f} - f^\star\|_{L^2(Q)}^2\right] = \int_{\mathcal{X}} \mathbf{E}\left[\left(\widehat{f}(x) - f^\star(x)\right)^2\right] \mathrm{d}Q(x).$$

Applying inequality (2.9) and recalling the definition of the similarity measure yields the claim (2.3).

We now focus on establishing the bound (2.9). Our proof makes use of the conditional expectation of $\widehat{f}$ given the covariates

$$\overline{f}(x) \coloneqq \mathbf{E}[\widehat{f}(x) \mid X_1, \ldots, X_n], \quad \text{for any } x \in \mathcal{X}.$$

To be explicit, the expectation is taken over $Y_i \mid X_i$, $i = 1, \ldots, n$. With this definition, our first result provides a bound on the conditional bias and variance.

Lemma 2.2. *For each $x \in \mathcal{X}$ almost surely, the Nadaraya-Watson estimator $\widehat{f}$ satisfies the bounds*

$$\left(\overline{f}(x) - f^\star(x)\right)^2 \leqslant \|f^\star\|_\infty^2 \mathbf{1}\{x \notin \mathcal{G}_n\} + L^2 h_n^{2\beta} \mathbf{1}\{x \in \mathcal{G}_n\} \qquad \textit{and} \tag{2.10a}$$

$$\mathbf{E}[(\overline{f}(x) - \widehat{f}(x))^2 \mid X_1, \ldots, X_n] \leqslant \frac{\sigma^2}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}} \mathbf{1}\{x \in \mathcal{G}_n\}. \tag{2.10b}$$

We prove this auxiliary claim at the end of this section.

Taking the results of Lemma 2.2 as given, we continue our proof of the bound (2.9). For any fixed $x \in \mathcal{X}$, a conditioning argument yields

$$\mathbf{E}\left[(\widehat{f}(x) - f^\star(x))^2\right] = \mathbf{E}\left[\left(\overline{f}(x) - f^\star(x)\right)^2\right] + \mathbf{E}\left[\mathbf{E}[(\overline{f}(x) - \widehat{f}(x))^2 \mid X_1, \ldots, X_n]\right].$$

By applying the bounds (2.10a) and (2.10b) to the two terms above, respectively, we arrive at the upper bound $\mathbf{E}\left[(\widehat{f}(x) - f^\star(x))^2\right] \leqslant T_1 + T_2$, where

$$T_1 \coloneqq \|f^\star\|_\infty^2 \mathbf{E}[\mathbf{1}\{x \notin \mathcal{G}_n\}] + L^2 h_n^{2\beta}, \quad \text{and} \quad T_2 \coloneqq \mathbf{E}\left[\mathbf{1}\{x \in \mathcal{G}_n\} \frac{\sigma^2}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}}\right].$$

We bound each of these terms in turn.

Bounding T_1: By definition, the set $\mathcal{G}_n$ involves n independent random variables, so that for any $x \in \mathcal{X}$, we have

$$\mathbf{E}\big[\mathbf{1}\{x \notin \mathcal{G}_n\}\big] = \Big(1 - P\big(\mathsf{B}(x, h_n)\big)\Big)^{n_P}\Big(1 - Q\big(\mathsf{B}(x, h_n)\big)\Big)^{n_Q} \overset{\text{(i)}}{\leqslant} \frac{1}{n\,\mu_n(\mathsf{B}(x, h_n))},$$

where step (i) follows from the elementary inequality

$$(1-p)^n(1-q)^m \leqslant \exp(-(np+mq)) \leqslant \frac{1}{np+mq},$$

valid for $p, q \in (0,1)$ and nonnegative integers n, m. Consequently, the first term is upper bounded as

$$T_1 \leqslant \|f^\star\|_\infty^2 \frac{1}{n\,\mu_n(\mathsf{B}(x, h_n))} + L^2 h_n^{2\beta}. \tag{2.11a}$$

Bounding T_2: For a fixed $x \in \mathcal{X}$, and for each $i = 1, \ldots, n$, define the Bernoulli random variable $Z_i = \mathbf{1}[X_i \in \mathsf{B}(x, h_n)] \in \{0, 1\}$, along with the binomial random variables $U = \sum_{i=1}^{n_P} Z_i$ and $V = \sum_{i=n_P+1}^{n} Z_i$. With these definitions, we can write

$$\sum_{i=1}^{n} \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\} = U + V, \quad \text{and} \quad \mathbf{1}\{x \in \mathcal{G}_n\} = \mathbf{1}\big\{U + V > 0\big\}.$$

Consequently, by an elementary bound for binomial random variables (see Lemma 2.5), it follows that

$$T_2 = \mathbf{E}\Big[\mathbf{1}\{U+V>0\}\frac{1}{U+V}\Big] \leqslant \frac{4}{n\,\mu_n(\mathsf{B}(x, h_n))}. \tag{2.11b}$$

Combining inequalities (2.11a) and (2.11b) yields the claim (2.9).

The only remaining detail is to prove the auxiliary lemma used in the proof.

Proof of Lemma 2.2. Recall that by definition, we have

$$\overline{f}(x) = \begin{cases} \frac{\sum_{i=1}^n f^\star(X_i)\mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}} & x \in \mathcal{G}_n \\ 0 & x \notin \mathcal{G}_n \end{cases}$$

Proof of the bound (2.10a)**:** By a direct expansion, we have

$$\begin{aligned} \big(\overline{f}(x) - f^\star(x)\big)^2 \mathbf{1}\{x \in \mathcal{G}_n\} &= \Big(\frac{\sum_{i=1}^n (f^\star(x) - f^\star(X_i))\mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}}\Big)^2 \mathbf{1}\{x \in \mathcal{G}_n\} \\ &\overset{\text{(i)}}{\leqslant} \frac{\sum_{i=1}^n (f^\star(x) - f^\star(X_i))^2 \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x, h_n)\}} \mathbf{1}\{x \in \mathcal{G}_n\} \\ &\overset{\text{(ii)}}{\leqslant} L^2 h_n^{2\beta} \mathbf{1}\{x \in \mathcal{G}_n\}, \end{aligned}$$

where step (i) follows from Jensen's inequality; and step (ii) makes use of Assumption 2.1. The bound (2.10a) is an immediate consequence.

Proof of the bound (2.10b)**:** In order to prove this claim, note that by independence among $\{(X_i, w_i)\}_{i=1}^n$,

$$\begin{aligned}\mathbf{E}[(\overline{f}(x) - \widehat{f}(x))^2 \mid X_1, \ldots, X_n] &= \sum_{i=1}^n \mathbf{E}[w_i^2 \mid X_i] \big(\tfrac{\mathbf{1}\{X_i \in \mathsf{B}(x,h_n)\}}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x,h_n)\}}\big)^2 \mathbf{1}\{x \in \mathcal{G}_n\} \\ &\overset{\text{(iii)}}{\leqslant} \sigma^2 \sum_{i=1}^n \big(\tfrac{\mathbf{1}\{X_i \in \mathsf{B}(x,h_n)\}}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x,h_n)\}}\big)^2 \mathbf{1}\{x \in \mathcal{G}_n\} \\ &= \frac{\sigma^2}{\sum_{i=1}^n \mathbf{1}\{X_i \in \mathsf{B}(x,h_n)\}} \mathbf{1}\{x \in \mathcal{G}_n\},\end{aligned}$$

which proves the claim. Here step (iii) is a consequence of Assumption 2.3. □

2.4.2 Proof of Corollary 2.1

Fix some $h \in (0, 1]$, and introduce the indicator variable $\eta = \mathbf{1}\{\alpha \geqslant 1\}$. We then have

$$\begin{aligned}\int_{\mathcal{X}} \frac{1}{n_P P(\mathsf{B}(x,h)) + n_Q Q(\mathsf{B}(x,h))} \, \mathrm{d}Q(x) &\leqslant \min\Big\{\frac{1}{n_P}\rho_h(P,Q), \frac{1}{n_Q}\rho_h(Q,Q)\Big\} \\ &\leqslant 3^\eta C \min\Big\{\frac{1}{n_P h^\alpha}, \frac{1}{n_Q h^\eta}\Big\} \\ &\leqslant 2 \cdot 3^\eta C \frac{1}{n_P h^\alpha + n_Q h^\eta}.\end{aligned}$$

The last inequality follows from (2.1) and standard covering number bounds (note $h \leqslant 1$). Thus the final performance bound is

$$2 \cdot 3^\eta C L^2 \Big\{h^{2\beta} + \frac{L^2 + \sigma^2}{n_P h^\alpha + n_Q h^\eta}\Big\}.$$

We choose the bandwidth $h^\star$ so as to trade off between two terms in this risk bound; more precisely, we set

$$h^\star = \Big(\big(\tfrac{n_Q}{L^2 + \sigma^2}\big) + \big(\tfrac{n_P}{L^2 + \sigma^2}\big)^{\frac{2\beta+\eta}{2\beta+\alpha}}\Big)^{-\frac{1}{2\beta+\eta}}$$

This choice is valid, since $\sigma^2 \geqslant L^2$ and $\max\{n_P, n_Q\} \geqslant 4\sigma^2$ by assumption. Substituting this choice of bandwidth into the risk bound (2.3) yields the claim.

2.4.3 Proof of Theorem 2.2(a)

Before giving the complete proof, we outline the main steps involved.

1. We first construct a hard instance $(P, Q) \in \mathcal{D}(\alpha, C)$. This instance is designed such that the integral quantity $\rho_h(P, Q)$ must scale as $Ch^{-\alpha}$.

2. Then we select a family of hard regression functions contained within $\mathcal{F}(\beta, L)$ that guarantees the worst-case expected error for our pair of distributions, (P, Q).

3. Finally, we apply Fano's method over this set of regression functions to show that the expected error must scale as the righthand side of inequality (2.6a).

It is worth commenting on our proof strategy in relation to past work. On one hand, in the case $\alpha \geqslant 1$, our construction of the distributions (P, Q) is adapted from the lower bound argument introduced by Kpotufe and Martinet [69]. The technical work involves constructing pairs of densities of P, Q, and establishing their membership in the class $\mathcal{D}(\alpha, C)$. As for the case $\alpha \in (0, 1)$, as stated in Theorem 2.2(b), we use a different construction of the distribution pair (P, Q), one that is new (to the best of our knowledge). We combine these constructions of "hard" source-target pairs, in particular by packing the interval $[0, 1]$ with a variable number of small intervals (e.g., [124, 115, 120]). By adapting the number of intervals (and constructing a packing set of the function class $\mathcal{F}(\beta, L)$ appropriately over these intervals), one can adapt the hardness of the lower bound instance to change with the number of samples. In this case, we are able to do this such that the hardness scales appropriately with the critical parameters that govern the final minimax lower bound: $n_P, n_Q, \sigma, \alpha, \beta$. With this high-level overview in place, we now proceed to the technical content of the proof.

Constructing "hard" source-target pairs: For scalars $S, r \in (0, 1]$, define $M = \frac{S}{6r}$ along with the intervals

$$I_j \coloneqq (z_j - 3r, z_j + 3r], \quad \text{where} \quad z_j \coloneqq 6jr - 3r, \qquad j = 1, \ldots, M.$$

We specify P and Q on each interval I_j as follows:

subinterval	density of P	density of Q
$(z_j - 3r, z_j - r]$	$\frac{1}{4Mr}(1 - \frac{\varepsilon}{3}(\frac{r}{S})^{\alpha-1})$	0
$(z_j - r, z_j + r]$	$\frac{\varepsilon}{6Mr}(\frac{r}{S})^{\alpha-1}$	$\frac{1}{2Mr}$
$(z_j + r, z_j + 3r]$	$\frac{1}{4Mr}(1 - \frac{\varepsilon}{3}(\frac{r}{S})^{\alpha-1})$	0

Table 2.1. Specification of densities for lower bound pair of distributions (P, Q) on the interval I_j.

By construction, both P and Q assign probability $1/M$ to the entire interval I_j. The following proposition verifies that (P, Q) lies in $\mathcal{D}(\alpha, C)$ for proper choices of the ε and S.

Proposition 2.2. *Let $\alpha \geqslant 1$ and $C \geqslant 1$. Define P and Q as in Table 2.1, with the following choice of parameters ε, S:*

(a) *if $C > 6$, set $\varepsilon = 6/C$, and $S = 1/4$;*

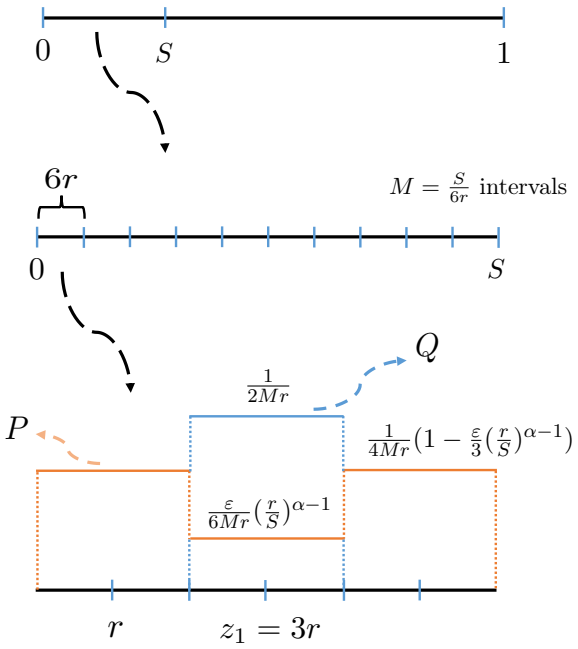


Figure 2.2. An illustration of the distributions (P, Q) constructed as a hard pair in our lower bound.

(b) *if* $1 \leqslant C \leqslant 6$, *set* $\varepsilon = 1$, *and* $S = \frac{1}{4}(C/6)^{1/\alpha}$.

Then for any choice of $M, r > 0$ *satisfying* $S = 6Mr$, *the pair* (P, Q) *lies in* $\mathcal{D}(\alpha, C)$.

See Section 2.4.3.1 for the proof of this claim.

Construction of "hard" regression functions. Next we construct a packing of the function class of $\mathcal{F}(\beta, L)$. We do so by summing together scaled and shifted copies of base function $\Psi\colon [-1, 1] \to \mathbf{R}$ that satisfies the boundary conditions $\Psi(-1) = \Psi(1) = 0$, along with

$$\big|\Psi(x) - \Psi(y)\big| \leqslant |x - y|^{\beta}, \quad \text{for all } x, y \in [-1, 1], \quad \text{and,} \tag{2.12a}$$

$$\int_{-1}^{1} \Psi^2(x)\, \mathrm{d}x =: C_{\Psi}^2 > 0. \tag{2.12b}$$

There are many possible choices of Ψ; see Chapter 2 in the book [115] for details. For our proof, we also require the bound $C_{\Psi}^2 \leqslant 1/6$, so that we make the explicit choice

$$\Psi(x) := e^{-1/(1-x^2)} \mathbf{1}\{|x| \leqslant 1\}.$$

We now form a class of functions using sums of the form

$$f_b(x) := \sum_{j=1}^{M} b_j \phi_j(x), \quad \text{where} \quad \phi_j(x) := L r^{\beta} \Psi\Big(\frac{x - z_j}{r}\Big),$$

and $b = (b_1, \ldots, b_M) \in \{0,1\}^M$ is a Boolean sequence. Our construction makes use of the Gilbert-Varshamov lemma (e.g. [115, Lemma 2.9]), which for $M \geqslant 8$, guarantees the existence of a subset $\mathcal{B} \subset \{0,1\}^M$ of cardinality at least $2^{M/8}$ such that

$$\|b - b'\|_1 \geqslant M/8 \qquad \text{for all distinct } b, b' \in \mathcal{B}. \tag{2.13}$$

Lemma 2.3. *The function class* $\mathcal{H} \coloneqq \Big\{ f_b \mid b \in \mathcal{B} \Big\}$ *has the following properties:*

(a) *It is contained within the Hölder class—*$\mathcal{H} \subset \mathcal{F}(\beta, L)$.

(b) *Pairs of functions are well-separated: for each distinct* $f, g \in \mathcal{H}$, *we have*

$$\|f - g\|_{L^2(Q)}^2 \geqslant \tfrac{C_\Psi^2}{16} L^2 r^{2\beta}.$$

(c) *Its elements satisfy the following* $L^2(P)$ *and* $L^2(Q)$ *bounds:*

$$\|f\|_{L^2(Q)}^2 \leqslant \frac{C_\Psi^2 M}{2S} L^2 r^{2\beta+1} \quad \textit{and} \quad \|f\|_{L^2(P)}^2 \leqslant \frac{\varepsilon C_\Psi^2 M}{6S^\alpha} L^2 r^{2\beta+\alpha},$$

for all $f \in \mathcal{H}$.

Applying Fano's method. We now combine the preceding constructions with a Fano argument to complete the proof of the lower bound. For any function $f \in \mathcal{H}$, let ν_f be the distribution $\{(X_i, Y_i)\}_{i=1}^n$ where (X, Y) pairs are related by our nonparametric regression model (2.2) with $f = f^\star$. For proving our lower bound, it suffices to consider Gaussian noise: in particular, $w_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma^2)$ for $i = 1, \ldots, n$. These variables satisfy Assumption 2.3.

With these choices, Kullback-Leibler divergence between any given pair (ν_f, ν_g) can be bounded as

$$\begin{aligned} &D_{\mathrm{kl}}(\nu_f \parallel \nu_g) \\ &\quad = \frac{1}{2\sigma^2}\Big(n_P \|f - g\|_{L^2(P)}^2 + n_Q \|f - g\|_{L^2(Q)}^2\Big) \leqslant \frac{2}{\sigma^2}\Big(n_P \max_{f \in \mathcal{H}} \|f\|_{L^2(P)}^2 + n_Q \max_{f \in \mathcal{H}} \|f\|_{L^2(Q)}^2\Big). \end{aligned}$$

Now applying part (c) of Lemma 2.3 yields

$$\begin{aligned} D_{\mathrm{kl}}(\nu_f \parallel \nu_g) &\leqslant M C_\Psi^2 \Big\{ n_P \frac{L^2}{3\sigma^2} \frac{\varepsilon}{S^\alpha} r^{2\beta+\alpha} + n_Q \frac{L^2}{\sigma^2} \frac{1}{S} r^{2\beta+1} \Big\} \\ &\leqslant M \Big\{ \frac{4^\alpha}{C} \frac{L^2}{\sigma^2} n_P r^{2\beta+\alpha} + \frac{4^\alpha}{C} \frac{L^2}{\sigma^2} n_Q r^{2\beta+1} \Big\} \end{aligned}$$

The final inequality arises by using $C_\Psi^2 \leqslant 1/6$. Suppose we take

$$r = \Big(\big(64 \frac{4^\alpha}{C} \frac{L^2 n_P}{\sigma^2}\big)^{\frac{2\beta+1}{2\beta+\alpha}} + \big(64 \frac{4^\alpha}{C} \frac{L^2 n_Q}{\sigma^2}\big) \Big)^{-\frac{1}{2\beta+1}}$$

Then for any distinct $f, g \in \mathcal{H}$, we obtain

$$D_{\mathrm{kl}}(\nu_f \parallel \nu_g) \leqslant M/32.$$

By a standard reduction to hypothesis testing [120, chap. 15] along with part (a),

$$\inf_{\widehat{f}} \sup_{f^\star \mathcal{F}(\beta,L)} \mathbf{E}\Big[\|\widehat{f} - f^\star\|_{L^2(Q)}^2\Big] \geqslant \frac{\min_{(f,g)\in\binom{\mathcal{H}}{2}} \|f-g\|_{L^2(Q)}^2}{4}\Big\{1 - \frac{\log 2 + \max_{(f,g)\in\binom{\mathcal{H}}{2}} D_{\mathrm{kl}}(\nu_f \parallel \nu_g)}{\log|\mathcal{H}|}\Big\}$$

Thus, after applying part (b) of Lemma 2.3, we obtain

$$\inf_{\widehat{f}} \sup_{f^\star \mathcal{F}(\beta,L)} \mathbf{E}\Big[\|\widehat{f} - f^\star\|_{L^2(Q)}^2\Big] \geqslant \frac{C_\Psi^2}{64} L^2 r^{2\beta}\Big(1 - \frac{8}{M} - \frac{1}{4}\Big) \geqslant \frac{C_\Psi^2 L^2}{256}\Big(\big(64\frac{4^\alpha}{C}\frac{L^2 n_P}{\sigma^2}\big)^{\frac{2\beta+1}{2\beta+\alpha}} + \big(64\frac{4^\alpha}{C}\frac{L^2 n_Q}{64\sigma^2}\big)\Big)^{-\frac{2\beta}{2\beta+1}},$$

provided that $M \geqslant 32$. Equivalently, $r \leqslant S/192$. It suffices that $r \leqslant \frac{1}{4608}$, this is ensured by having

$$\max\{n_P, n_Q\} \geqslant \Big(72\frac{\sigma^2}{L^2}\frac{C}{4^\alpha}\Big)^{2\beta+\alpha}.$$

2.4.3.1 Proof of Proposition 2.2

We will show that for a general choice of $\varepsilon, S \in (0,1]$, the following holds:

$$P\big(\mathsf{B}(x,h)\big) \geqslant \frac{\varepsilon}{3}\big(\frac{h}{4S}\big)^{\alpha-1} Q\big(\mathsf{B}(x,h)\big), \quad \text{for all } x \in \mathrm{supp}(Q), \text{ and any } h > 0. \tag{2.14}$$

For the moment let us take this bound as given. By Lemma 2.1, note that bound (2.14) implies that $(P,Q) \in \mathcal{D}(\alpha, \mathcal{C}(\varepsilon,S))$, with $\mathcal{C}(\varepsilon,S) = \frac{6}{\varepsilon}(4S)^{\alpha-1}$, for any $\varepsilon, S \in (0,1]$. Note that the parameter choices given in the statement of the result ensure that $\varepsilon, S \in (0,1]$. When $C \geqslant 6$, we have $\mathcal{C}(\varepsilon,S) = 6(C/6)^{1-1/\alpha} = C(6/C)^{1/\alpha} \leqslant 6 \leqslant C$. Otherwise $C \leqslant 6$ and $\mathcal{C}(\varepsilon,S) = C$. Therefore, checking the two cases $C > 6$ and $C \leqslant 6$ verifies $\mathcal{C}(\varepsilon,S) = C$ in both regimes, which furnishes the claim.

We now turn to establish bound (2.14). Let $h > 0$. First observe that the support of Q is the disjoint union of intervals $\cup_{j=1}^M (z_j - r, z_j + r]$. Thus, fix x in the support of Q, and let z_j denote the center of the interval to which x belongs. Suppose that $h \in [0, 4r]$, in which

case, we have the inclusion $\mathsf{B}(x,h) \subset I_j$, whence the lower bound

$$\begin{aligned} P(\mathsf{B}(x,h)) &\geqslant P\big(\mathsf{B}(x,h) \cap \mathsf{B}(z_j,r)\big) \\ &\overset{\text{(i)}}{=} \frac{\varepsilon}{3}\Big(\frac{r}{S}\Big)^{\alpha-1} Q\big(\mathsf{B}(x,h) \cap \mathsf{B}(z_j,r)\big) \\ &\overset{\text{(ii)}}{\geqslant} \frac{\varepsilon}{3}\Big(\frac{h}{4S}\Big)^{\alpha-1} Q\big(\mathsf{B}(x,h) \cap \mathsf{B}(z_j,r)\big) \\ &\overset{\text{(iii)}}{=} \frac{\varepsilon}{3}\Big(\frac{h}{4S}\Big)^{\alpha-1} Q\big(\mathsf{B}(x,h)\big) \end{aligned} \tag{2.15}$$

Above, step (i) follows from the construction of P, Q; step (ii) follows from $h \leqslant 4r$, whereas step (iii) follows since $\mathsf{B}(x,h) \subset I_j$ and Q assigns no mass to the set $I_j \backslash \mathsf{B}(z_j,r)$.

Otherwise, we may assume that $h \in [4r, S]$, in which case we have the inclusion $\mathsf{B}(x,h) \supset I_j$. Denote by $N \geqslant 1$ the number of intervals of the form I_j that are included within $\mathsf{B}(x,h)$. Note that since $\mathsf{B}(x,h)$ is connected, it is always contained in at most $N+2$ intervals (by considering partial intervals on the left and right). Thus,

$$\frac{P(\mathsf{B}(x,h))}{Q(\mathsf{B}(x,h))} \overset{\text{(iii)}}{\geqslant} \frac{N \cdot P(I_j)}{(N+2) \cdot Q(I_j)} \overset{\text{(iv)}}{\geqslant} \frac{1}{3}. \tag{2.16}$$

Here step (iii) follows since $\mathsf{B}(x,h)$ is contained in a collection of at most $(N+2)$ intervals and contains at least N intervals, and the intervals are disjoint and have the same mass under both P and Q. On the other hand, step (iv) uses the equivalence $P(I_j) = Q(I_j)$, along with the fact that the function $x \mapsto \frac{x}{x+2}$ is increasing on the set $\{x \geqslant 1\}$.

Therefore, combining inequalities (2.15) and (2.16), we conclude that

$$P(\mathsf{B}(x,h)) \geqslant \frac{1}{3}\Big[\varepsilon\Big(\frac{h}{4S}\Big)^{\alpha-1} \wedge 1\Big] Q(\mathsf{B}(x,h)) \geqslant \frac{\varepsilon}{3}\big(\frac{h}{4S}\big)^{\alpha-1} Q(\mathsf{B}(x,h))$$

for every x in the support of Q, the final inequality follows since $\alpha \geqslant 1$. Since $h > 0$ was arbitrary, this establishes bound (2.14) and completes the proof.

2.4.3.2 Proof of Lemma 2.3

We prove each of the three parts in turn.

Proof of part (a): Fix a Boolean vector $b \in \{0,1\}^M$. Note that the function ϕ_j is supported on the interval I_j, which is disjoint from any other interval $I_k, k \neq j$. Since Ψ satisfies the continuity condition (2.12a), it follows that ϕ_j is (β, L)-Hölder. Finally, we have $f_\varepsilon(0) = 0$ by definition. Taking these properties together, we have shown that $f_\varepsilon \in \mathcal{F}(\beta, L)$, as required.

Proof of part (b): For any distinct pair $b, b' \in \mathcal{B}$, we have

$$\begin{aligned}\int_0^1 (f_b(x) - f_{b'}(x))^2 \, \mathrm{d}Q(x) &= \int_0^1 \Big(\sum_{j=1}^M (b_j - b'_j)\phi_j(x) \Big)^2 \, \mathrm{d}Q(x) \\ &\overset{\text{(i)}}{=} \frac{1}{2Mr} \sum_{j=1}^M (b_j - b'_j)^2 \int_{z_j - 3r}^{z_j + 3r} \phi_j^2(x) \, \mathrm{d}x \\ &\overset{\text{(ii)}}{=} \frac{C_\Psi^2}{2M} L^2 r^{2\beta} \|b - b'\|_1 \\ &\overset{\text{(iii)}}{\geqslant} \frac{C_\Psi^2}{16} L^2 r^{2\beta}.\end{aligned}$$

Here step (i) follows from the definition of Q along with the disjointedness of the supports of ϕ_j. Step (ii) follows from equation (2.12b) and the fact that $b, b' \in \mathcal{B} \subset \{0,1\}^M$. Finally, step (iii) follows from the Gilbert-Varshamov separation (2.13).

Proof of part (c): For any $b \in \mathcal{B}$, by following the calculations above, for $\mu \in \{P, Q\}$, we have by symmetry

$$\int_0^1 f_b^2(x) \, \mathrm{d}\mu(x) = \sum_{j=1}^M b_j^2 \int_{I_j} \phi_j^2(x) \, \mathrm{d}\mu(x) \leqslant M \int_{I_1} \phi_1^2(x) \, \mathrm{d}\mu(x).$$

Now observe that $\int_0^{6r} \phi_1^2(x) \, \mathrm{d}Q(x) = \frac{C_\Psi^2}{2M} L^2 r^{2\beta}$, and consequently, $\|f_b\|_{L^2(Q)}^2 \leqslant L^2 r^{2\beta} C_\Psi^2 / 2$. Additionally, we can compute

$$\int_0^{6r} \phi_1^2(x) \, \mathrm{d}P(x) = \frac{\varepsilon}{6rM^\alpha} \int_{2r}^{4r} \phi_1^2(x) \, \mathrm{d}x = \frac{\varepsilon}{6S^\alpha} L^2 r^{2\beta+\alpha} C_\Psi^2.$$

Thus, we have established the upper bound $\|f_b\|_{L^2(P)}^2 \leqslant \varepsilon L^2 r^{2\beta+\alpha-1} / (6S^{\alpha-1})$.

2.4.4 Proof of Theorem 2.2(b)

Given the inclusion $\mathcal{D}'(\alpha, 1) \subset \mathcal{D}'(\alpha, C)$, it suffices to prove a lower bound for $C = 1$.

Construction of "hard" distributions. Let $Q = \delta_1$, and let P_α be the distribution supported on $[0,1]$ with density $p_\alpha(x) \coloneqq \alpha(1-x)^{\alpha-1} \mathbf{1}\{x \in [0,1]\}$. By construction, we then have

$$\rho_h(P_\alpha, Q) = \frac{1}{P_\alpha(B(1,h))} = h^{-\alpha} \quad \text{for all } h \in (0,1],$$

which implies that $(P_\alpha, Q) \in \mathcal{D}'(\alpha, 1)$. From herein, we adopt the shorthand $P \coloneqq P_\alpha$ so as to lighten notation.

Construction of two point alternative. If the regression function is f, we denote the resulting joint distribution of $\{(X_i, Y_i)\}_{i=1}^n$ by ν_f. We consider the two point alternatives $\{f_t, g\}$ with $g \equiv 0$ and $f_t(x) \coloneqq L(x-t)_+^\beta$. The next result demonstrates the validity of this choice:

Lemma 2.4. *For any $t \in [0,1]$, the function f_t belongs to $\mathcal{F}(\beta, L)$.*

See Section 2.4.4.1 for the proof.

Moreover, by straightforward calculations, we find that $\|f_t\|_{L^2(Q)}^2 = L^2(1-t)^{2\beta}$, and

$$\begin{aligned}\|f_t\|_{L^2(P)}^2 &= L^2 \int_t^1 \alpha(1-x)^{\alpha-1}(x-t)^{2\beta} \, \mathrm{d}x \\ &\leqslant L^2(1-t)^{2\beta} \int_0^{1-t} \alpha s^{\alpha-1} \, \mathrm{d}s = L^2(1-t)^{2\beta+\alpha}.\end{aligned}$$

Applying Le Cam's method. We are now equipped to apply Le Cam's two point bound. In particular, we have

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{F}(\beta,L)} \mathbf{E}\Big[\|\widehat{f} - f^\star\|_{L^2(Q)}^2\Big] \geqslant \frac{L^2(1-t)^{2\beta}}{16} \exp\big(-D_{\mathrm{kl}}(\nu_{f_t} \,\|\, \nu_g)\big)$$

By standard KL calculations (using $\mathsf{N}(0, \sigma^2)$ noises)

$$D_{\mathrm{kl}}(\nu_{f_t} \,\|\, \nu_g) = \frac{L^2}{2\sigma^2}\Big\{n_P(1-t)^{2\beta+\alpha} + n_Q(1-t)^{2\beta}\Big\}$$

Finally, we make the

$$1 - t = \left(\Big(\frac{L^2 n_P}{2\sigma^2}\Big)^{\frac{1}{2\beta+\alpha}} + \Big(\frac{L^2 n_Q}{2\sigma^2}\Big)^{\frac{1}{2\beta}}\right)^{-1}$$

A little bit of algebra shows that this choice guarantees that $D_{\mathrm{kl}}(\nu_{f_t} \,\|\, \nu_g) \leqslant 2$, which completes the proof.

2.4.4.1 Proof of Lemma 2.4

We begin by observing that $f_t(0) = 0$. Thus, in order to prove the claim, it suffices to show that

$$f_t(y) - f_t(x) \leqslant L(y-x)^\beta \quad \text{for any pair } x, y \text{ such that } 0 \leqslant t < x < y \leqslant 1.$$

In order to prove this bound, consider an arbitrary point $x \in (t, 1)$, and define the function

$$\phi_x(y) \coloneqq L(y^\beta - x^\beta) - L(y-x)^\beta \qquad \text{for } y \in [x, 1].$$

We can compute the derivative $\phi_x'(y) = L\beta(y^{\beta-1} - (y-x)^{\beta-1})$. Since $y \geqslant y - x > 0$ and $\beta \leqslant 1$, we have $y^{\beta-1} \leqslant (y-x)^{\beta-1}$, and hence $\phi_x'(y) \leqslant 0$. Consequently, the function ϕ_x is non-increasing, and since $y > x$, it follows that $\phi_x(y) \leqslant \phi_x(x) = 0$. Putting together the pieces completes the proof.

2.4.5 Proof of Proposition 2.1

Starting with the assumed bound (2.7), we have

$$\int_{\mathcal{X}} \frac{1}{P(\mathsf{B}(x,h))} \, \mathrm{d}Q(x) \leqslant \frac{1}{\lambda} \int_{\mathcal{X}} \frac{1}{Q(\mathsf{B}(x,h))} \, \mathrm{d}Q(x). \tag{2.17}$$

By definition of the covering number $N := N(h/2)$, there is a collection $\{z^j\}_{j=1}^N$ such that the set $\mathcal{X}$ is contained within the union $\bigcup_{j=1}^N \mathsf{B}(z^j, \frac{h}{2})$. This fact, combined with our previous bound (2.17), implies that

$$\int_{\mathcal{X}} \frac{1}{P(\mathsf{B}(x,h))} \, \mathrm{d}Q(x) \leqslant \frac{1}{\lambda} \sum_{j=1}^{N} \int_{\mathsf{B}(z_j,h/2)} \frac{1}{Q(\mathsf{B}(x,h))} \, \mathrm{d}Q(x). \tag{2.18}$$

Note by the triangle inequality, for each $j \in [N]$ and $x \in \mathsf{B}(z_j, h/2)$, we have $\mathsf{B}(z_j, h/2) \subset \mathsf{B}(x,h)$. This inclusion implies that

$$\int_{\mathsf{B}(z_j,h/2)} \frac{1}{Q(\mathsf{B}(x,h))} \, \mathrm{d}Q(x) \leqslant \int_{\mathsf{B}(z_j,h/2)} \frac{1}{Q(\mathsf{B}(z_j,h/2))} \, \mathrm{d}Q(x) = 1,$$

for each $j \in [N]$. Combining this inequality with the bound (2.18) yields the claim.

2.4.6 Proof of Lemma 2.1

By assumption, we have the upper bound

$$\int_0^1 \frac{1}{P(\mathsf{B}(x,h))} \, \mathrm{d}Q(x) \leqslant \frac{1}{Kh^\gamma} \int_0^1 \frac{1}{Q(\mathsf{B}(x,h))} \, \mathrm{d}Q(x)$$

Moreover, we can find a collection of $N := \lceil 1/h \rceil$ balls with centers $\{z_j\}_{j=1}^N$ of radius $h/2$ that cover the interval $[0,1]$, whence

$$\int_0^1 \frac{1}{Q(\mathsf{B}(x,h))} \, \mathrm{d}Q(x) \leqslant \sum_{j=1}^{N} \int_{x \in \mathsf{B}(z_j,h/2)} \frac{1}{Q(\mathsf{B}(x,h))} \, \mathrm{d}Q(x) \leqslant N.$$

The final inequality follows from the inclusion $\mathsf{B}(x,h) \supset \mathsf{B}(z_j, h/2)$.

Now define the function $g(t) := \lceil t \rceil / t$, and observe that $g(t) \leqslant 2$ whenever $t \geqslant 1$. Consequently, we can write

$$h^{\gamma+1} \rho_h(P,Q) \leqslant \frac{1}{K} g(1/h) \leqslant \frac{2}{K}, \quad \text{for any } h \leqslant 1.$$

Passing to the supremum over $h \in (0,1]$ yields the claim.

2.5 Discussion

In this chapter, we have studied the problem of covariate shift in the context of nonparametric regression. We have shown that a measure of (dis)-similarity ρ_h between the source and target distributions, as defined in equation (2.1), can be used to characterize how minimax risks change as the source-target pair are varied. In particular, we proved upper bounds on the Nadaraya-Watson estimator over Hölder classes that are an explicit function of the similarity ρ_h, and also established matching lower bounds over classes constrained in terms of the similarity. We also discussed how the measure ρ_h is related to other characterizations of covariate shift from past work, including likelihood ratio bounds and transfer exponents. Our work shows that similarity measure ρ_h provides a more fine-grained characterization of how covariate shift changes the difficulty of non-parametric regression.

Our work leaves open a number of open questions. First, our lower bounds for covariate shift (cf. Theorem 2.2) are obtained within a global minimax framework, which involves worst-case assessments over a certain function class. These lower bounds match our upper bound on the NW estimator (cf. Theorem 2.1) for certain source-target pairs (P, Q). But the upper bound actually depends explicitly on the source-target pair. Is this upper bound always optimal? Or are there instances of covariate shift for which Nadaraya-Watson is suboptimal for some Hölder continuous function? In general, this question appears non-trivial: even without the (interesting) complication of covariate shift, there are few results that give distribution-dependent results for nonparametric regression outside of the uniform distribution and fixed-design problems.

2.6 Elementary bound for binomial variables

In this section, we state and prove an elementary bound for binomial random variables, used in the proof of Theorem 2.1.

Lemma 2.5. *Let n, m be positive integers and $p, q \in (0, 1)$. Suppose that $U \sim \mathsf{Bin}(n, p)$ and $V \sim \mathsf{Bin}(m, q)$. Then*

$$\mathbf{E}\Big[\frac{1}{U+V}\mathbf{1}\{U+V>0\}\Big] \leqslant \frac{4}{np+mq}.$$

Proof. We begin by observing that conditionally on the event $\{U + V > 0\}$, we have the lower bound

$$U + V \geqslant \frac{U+V+1}{2} \geqslant \frac{U+1}{2} \vee \frac{V+1}{2}.$$

These lower bounds allow us to write

$$\mathbf{E}\Big[\frac{1}{U+V}\mathbf{1}\{U+V>0\}\Big] \leqslant \mathbf{E}\,\frac{2}{U+1} \wedge \mathbf{E}\,\frac{2}{V+1} \leqslant \frac{2}{(np \vee mq)} \leqslant \frac{4}{np+mq}.$$

Here the penultimate inequality is a consequence of known results for binomial random variables [25, equation (3.4)]. □

Chapter 3

Failure of the Lasso under anisotropic design

3.1 Introduction

In this chapter, we consider the standard linear regression model

$$y = X\theta^\star + w, \tag{3.1}$$

where $\theta^\star \in \mathbf{R}^d$ is the unknown parameter, $X \in \mathbf{R}^{n\times d}$ is the design matrix, and $w \sim \mathsf{N}(0, \sigma^2 I_n)$ denotes the stochastic noise. Such linear regression models are pervasive in statistical analysis [72]. To improve model selection and estimation, it is often desirable to impose a *sparsity* assumption on $\theta^\star$—for instance, we might assume that $\theta^\star$ has few nonzero entries or that it has few large entries. This amounts to assuming that for some $p \in [0, 1]$

$$\|\theta^\star\|_p \leqslant R, \tag{3.2}$$

where $\|\cdot\|_p$ denotes the ℓ_p vector (quasi)norm, and $R > 0$ is the radius of the ℓ_p ball. There has been a flurry of research on this sparse linear regression model (3.1)-(3.2) over the last three decades; see the recent books [18, 57, 120, 37, 62] for an overview.

Comparatively less studied, is the effect of the design matrix X on the ability (or inability) to estimate $\theta^\star$ under the sparsity assumption. Intuitively, when X is "close to singular", we would expect that certain directions of $\theta^\star$ would be difficult to estimate. Therefore, in this section we seek to determine the optimal rate of estimation when the *smallest singular value of X is bounded.* More precisely, we consider the following set of design matrices

$$\mathcal{X}_{n,d}(B) \coloneqq \Big\{ X \in \mathbf{R}^{n\times d} : \frac{1}{n} X^\mathsf{T} X \succcurlyeq \frac{1}{B} I_d \Big\}, \tag{3.3}$$

and aim to characterize the corresponding minimax rate of estimation

$$\mathfrak{M}_{n,d}(p, \sigma, R, B) \coloneqq \inf_{\widehat{\theta}} \sup_{\substack{X\in\mathcal{X}_{n,d}(B) \\ \|\theta^\star\|_p\leqslant R}} \mathbf{E}_{y\sim\mathsf{N}(X\theta^\star,\sigma^2 I_n)} \Big[\big\|\widehat{\theta} - \theta^\star\big\|_2^2 \Big].$$

3.1.1 A motivation from learning under covariate shift

We now make a connection between this section and the overall aim of this thesis—specifically, a connection to covariate shift. Although it may seem a bit technical to focus on the dependence of the estimation error on the smallest singular value of the design matrix X, we would like to point out an additional motivation which is more practical and also motivates our problem formulation. This is the problem of linear regression in a well-specified model with covariate shift.

To begin with, recall that under random design, in the standard linear observational model (*i.e.*, without covariate shift) the statistician observes random covariate-label pairs of the form (x, y). Here, the covariate x is drawn from a distribution Q and the label y satisfies $\mathbf{E}[y \mid x] = x^\top \theta^\star$. The goal is to find an estimator $\widehat{\theta}$ that minimizes the out-of-sample excess risk, which takes the quadratic form $\mathbf{E}_{x\sim Q}[((\widehat{\theta} - \theta^\star)^\mathsf{T} x)^2]$. When the covariate distribution Q is isotropic, meaning that $\mathbf{E}_{x\sim Q}[xx^\top] = I$, the out-of-sample excess risk equals the squared ℓ_2 error $\|\widehat{\theta} - \theta^\star\|_2^2$.

Under covariate shift, there is a slight twist to the standard linear regression model previously described, where now the covariates x are drawn from a (source) distribution P that differs from the (target) distribution Q under which we would like to deploy our estimator. Assuming Q is isotropic, the goal is therefore still to minimize the out-of-sample excess risk under Q, which is $\|\theta^\star - \widehat{\theta}\|_2^2$. In general, if $P \neq Q$ and no additional assumptions are made, then learning with covariate shift is impossible in the sense that no estimator can be consistent for the optimal parameter $\theta^\star$. It is therefore common (and necessary) to impose some additional assumptions on the pair (P, Q) to facilitate learning. One popular assumption relates to the likelihood ratio between the source-target pair. It is common to assume that absolute continuity holds so that $Q \ll P$ and that the the likelihood ratio $\frac{\mathrm{d}Q}{\mathrm{d}P}$ is uniformly bounded [77]. Interestingly, it is possible to show that if $\frac{\mathrm{d}Q}{\mathrm{d}P}(x) \leqslant B$ for P-almost every x, then the semidefinite inequality

$$\mathbf{E}_{x\sim P}[xx^\top] \succcurlyeq \frac{1}{B} I \tag{3.4}$$

holds [77, 121]. Comparing the inequality (3.4) to our class $\mathcal{X}_{n,d}(B)$ as defined in display (3.3), we note that our setup can be regarded as a fixed-design variant of linear regression with covariate shift [73, 36, 126].

3.1.2 Determining the minimax rate of estimation

We begin with one of our main results, which precisely characterizes the (order-wise) minimax risk $\mathfrak{M}_{n,d}(p, \sigma, R, B)$ of estimating $\theta^\star$ under the sparsity constraint $\|\theta^\star\|_p \leqslant R$ and over the restricted design class $\mathcal{X}_{n,d}(B)$.

Theorem 3.1. *Let $n \geqslant d \geqslant 1$ and $\sigma, R, B > 0$ be given, and put $\tau_n^2 \coloneqq \frac{\sigma^2 B}{R^2 n}$. There exist two universal constants c_ℓ, c_u satisfying $0 < c_\ell < c_u < \infty$ such that*

(a) if $p \in (0, 1]$ *and* $\tau_n^2 \in [d^{-2/p}, \log^{-1}(\mathrm{e}d)]$*, then*

$$c_\ell \, R^2 \left(\tau_n^2 \log\left(\mathrm{e} d \tau_n^p\right)\right)^{1-p/2} \leqslant \mathfrak{M}_{n,d}(p, \sigma, R, B) \leqslant c_u \, R^2 \left(\tau_n^2 \log\left(\mathrm{e} d \tau_n^p\right)\right)^{1-p/2}, \quad \text{and}$$

(b) if $p = 0$*, we denote* $s = R \in [d]$*, and have*

$$c_\ell \, \frac{\sigma^2 B}{n} s \log\left(\mathrm{e}\frac{d}{s}\right) \leqslant \mathfrak{M}_{n,d}(p, \sigma, s, B) \leqslant c_u \, \frac{\sigma^2 B}{n} s \log\left(\mathrm{e}\frac{d}{s}\right).$$

The proof of Theorem 3.1 relies on a reduction to the Gaussian sequence model [62], and is deferred to Section 3.3.1.

Several remarks on Theorem 3.1 are in order. The first observation is that Theorem 3.1 is sharp, apart from universal constants that do not depend on the tuple of problem parameters $(p, n, d, \sigma, s, R, B)$.

Secondly, it is worth commenting on the sample size restrictions in Theorem 3.1. For all $p \in [0, 1]$, we have assumed the "low-dimensional" setup that the number of observations n dominates the dimension d.[1] Note that this is necessary for the class of designs $\mathcal{X}_{n,d}(B)$ to be nonempty. On the other hand, for $p > 0$ we additionally require that the sample size is "moderate", i.e., $\tau_n^2 \in [d^{-2/p}, \log^{-1}(\mathrm{e}d)]$. We make this assumption so that we can focus on what we believe is the "interesting" regime: where neither ordinary least squares nor constant estimators are optimal. Indeed, when $n \geqslant d$ but $\tau_n^2 \geqslant \log^{-1}(\mathrm{e}d)$, it is easily verified that the optimal rate of estimation is on the order R^2; intuitively the effective noise level is too high and no estimator can dominate $\widehat{\theta} \equiv 0$ uniformly. On the other hand, when $n \geqslant d$ but $\tau_n^2 \leqslant d^{-2/p}$, then the ordinary least squares estimator is minimax optimal; intuitively, the noise level is sufficiently small such that there is, in the worst case, no need to shrink on the basis of the ℓ_p constraint to achieve the optimal rate.

Last but not least, as shown in Theorem 3.1, the optimal rate of estimation depends on the signal-to-noise ratio $\tau_n^{-2} = nR^2/(\sigma^2 B)$. As B increases, the design X becomes closer to singular, estimation of $\theta^\star$, as expected, becomes more challenging. The dependence of our result on B is exactly analagous to the impact of the likelihood ratio bound B appearing in the context of prior work on nonparametric regression under covariate shift [77].

3.1.3 A computationally efficient estimator

The optimal estimator underlying the proof of Theorem 3.1 requires computing a d-dimensional Gaussian integral, and therefore is not computationally efficient in general. In this section we propose an estimator that is both computationally efficient and statistically optimal, up to constant factors.

Our procedure is based on the soft thresholding operator: for $v \in \mathbf{R}^d$ and $\eta > 0$, we define

$$\mathsf{S}_\eta(v) \coloneqq \operatorname*{arg\,min}_{u \in \mathbf{R}^d} \left\{ \, \|u - v\|_2^2 + 2\eta \|u\|_1 \, \right\}.$$

[1] Notably, this still allows n to be proportional to d, *e.g.*, we can tolerate $n = d$.

Note that soft thresholding involves a coordinate-separable optimization problem and has an explicit representation, thus allowing efficient computation. Then we define the *soft thresholded ordinary least squares estimator*

$$\widehat{\theta}_{\eta}^{\mathsf{STOLS}}(X, y) \coloneqq \mathsf{S}_{\eta}\Big(\widehat{\theta}^{\mathsf{OLS}}(X, y)\Big), \tag{3.5}$$

where $\widehat{\theta}^{\mathsf{OLS}}(X, y)$ is the usual ordinary least squares estimate—equal to $(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}y$ in our case. We have the following guarantees for its performance.

Theorem 3.2. *The soft thresholded ordinary least squares estimator* (3.5) *satisfies*

(a) in the case $p \in (0, 1]$, for any $R > 0$, if $\tau_n^2 \in [d^{-2/p}, \log^{-1}(\mathrm{e}d)]$, then

$$\sup_{X \in \mathcal{X}_{n,d}(B)} \sup_{\|\theta^\star\|_p \leqslant R} \mathbf{E}\left[\|\widehat{\theta}_{\eta}^{\mathsf{STOLS}}(X, y) - \theta^\star\|_2^2\right] \leqslant 6\, R^2 \left(\tau_n^2 \log(\mathrm{e}d\tau_n^p)\right)^{1-p/2},$$

with the choice $\eta = \sqrt{2R^2\tau_n^2 \log(\mathrm{e}d\tau_n^p)}$, and

(b) in the case $p = 0$, for any $s \in [d]$,

$$\sup_{X \in \mathcal{X}_{n,d}(B)} \sup_{\|\theta^\star\|_0 \leqslant s} \mathbf{E}\left[\|\widehat{\theta}_{\eta}^{\mathsf{STOLS}}(X, y) - \theta^\star\|_2^2\right] \leqslant 6\, \frac{\sigma^2 B}{n} s \log\Big(\mathrm{e}\frac{d}{s}\Big),$$

with the choice $\eta = \sqrt{2\frac{\sigma^2 B}{n}\log(\frac{\mathrm{e}d}{s})}$.

The proof is presented in Section 3.3.2.

Comparing the guarantee in Theorem 3.2 to the minimax rate in Theorem 3.1, it is immediate to see that the soft thresholded ordinary least squares estimator is minimax optimal apart from constant factors.

Secondly, we would like to point out a (simple) modification to the soft thresholding ordinary least squares procedure that allows it to be adaptive to the hardness of the particular design matrix encountered. To achieve this, note that $X \in \mathcal{X}_{n,d}(\widehat{B})$ for $\widehat{B} \coloneqq \|\!|(X^{\mathsf{T}}X)^{-1}|\!\|_{\mathrm{op}}$. Therefore the results in Theorem 3.2 continue to hold with B replaced by (a possibly smaller) $\widehat{B}$, provided that the thresholding parameter η is properly adjusted. For instance, in the case with $p = 0$, we have

$$\sup_{\|\theta^\star\|_0 \leqslant s} \mathbf{E}\left[\|\widehat{\theta}_{\widehat{\eta}}^{\mathsf{STOLS}}(X, y) - \theta^\star\|_2^2\right] \leqslant 6\, \frac{\sigma^2 \widehat{B}}{n} s \log\Big(\mathrm{e}\frac{d}{s}\Big),$$

provided we take $\widehat{\eta} = \sqrt{2\frac{\sigma^2\widehat{B}}{n}\log(\frac{\mathrm{e}d}{s})}$.

Finally, we note that inspecting our proof, the upper bound for $\widehat{\theta}^{\mathsf{STOLS}}_{\widehat{\eta}}(X, y)$ also holds for a larger set of design matrices

$$\mathcal{X}^{\mathsf{diag}}_{n,d}(B) \coloneqq \Big\{ X \in \mathbf{R}^{n\times d} : \Big(\frac{1}{n}X^{\mathsf{T}}X\Big)^{-1}_{ii} \leqslant B, \quad \text{for } 1 \leqslant i \leqslant d \Big\}.$$

Since $\mathcal{X}_{n,d}(B) \subset \mathcal{X}^{\mathsf{diag}}_{n,d}(B)$, this means after combining the lower bounds in Theorem 3.1 with the guarantees in Theorem 3.2, we additionally have established the minimax rate over this larger family $\mathcal{X}^{\mathsf{diag}}_{n,d}(B)$.

3.1.4 Is Lasso optimal?

Arguably, the Lasso estimator [114] is the most widely used estimator for sparse linear regression. Given a regularization parameter $\lambda > 0$, the Lasso is defined to be

$$\widehat{\theta}_\lambda(X, y) \coloneqq \operatorname*{arg\,min}_{\vartheta\in\mathbf{R}^d} \Big\{ \frac{1}{n}\|X\vartheta - y\|_2^2 + 2\lambda\|\vartheta\|_1 \Big\}. \tag{3.6}$$

Surprisingly, we show that the Lasso estimator—despite its popularity—is provably *suboptimal* for estimating $\theta^\star$ when $B \gg 1$.

Corollary 3.1. *The Lasso is minimax suboptimal by polynomial factors in the sample size when $d = n$ and $B = \sqrt{n}$. More precisely,*

(a) if $p \in (0, 1]$, and $\sigma = R = 1$, then we have

$$\sup_{X\in\mathcal{X}_{n,d}(B)} \sup_{\|\theta^\star\|_p\leqslant R} \mathbf{E}\Big[\inf_{\lambda>0} \|\widehat{\theta}_\lambda(X, y) - \theta^\star\|_2^2 \Big] \gtrsim 1, \quad \textit{and}$$

(b) if $p = 0$, and $\sigma = s = 1$, then we have

$$\sup_{X\in\mathcal{X}_{n,d}(B)} \sup_{\|\theta^\star\|_0\leqslant s} \mathbf{E}\Big[\inf_{\lambda>0} \|\widehat{\theta}_\lambda(X, y) - \theta^\star\|_2^2 \Big] \gtrsim 1.$$

Corollary 3.1 is in fact a special case of a more general theorem (Theorem 3.3) to be provided later.

Applying Theorem 3.1 to the regime considered in Corollary 3.1, we obtain the optimal rate of estimation

$$\Big(\frac{\sqrt{1+\log n}}{n^{1/4}}\Big)^{2-p} \ll 1, \qquad \text{for every } p \in [0, 1].$$

As shown, in the worst-case, the multiplicative gap between the performance of the Lasso and a minimax optimal estimator in this scaling regime is at least polynomial in the sample size. As a result, the Lasso is quite strikingly minimax *suboptimal* in this scaling regime.

In fact, the lower bound against Lasso in Corollary 3.1 is extremely strong. Note that in the lower bound, the Lasso is even allowed to leverage the oracle information $\theta^\star$ to calculate the optimal instance-dependent choice of the regularization parameter (*c.f.*, $\inf_{\lambda>0} \|\widehat{\theta}_\lambda(X,y) - \theta^\star\|_2^2$). As a result, the lower bound applies to any estimator which can be written as the penalized Lasso estimator with data-dependent choice of penalty. Many typical Lasso-based estimators, such as the norm-constrained and cross-validated Lasso, can be written as the penalized Lasso with a data-dependent choice of the penalty parameter λ. For instance, in the case of the norm-constrained Lasso, this holds by convex duality. Thus, we can rule out the minimax optimality of any procedure of this type, in light of Corollary 3.1.

The separation between the oracle Lasso and the minimax optimal estimator can also be demonstrated in experiments, as shown below in Figure 3.1.

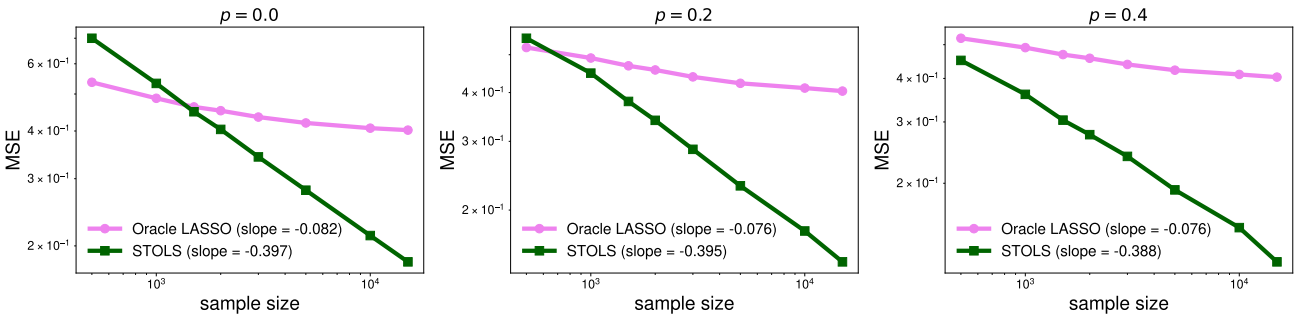


Figure 3.1. Numerical simulation demonstrating suboptimality of the Lasso, with the oracle choice of regularization $\widehat{\lambda} \in \arg\min_{\lambda>0} \|\widehat{\theta}_\lambda(X,y) - \theta^\star\|_2^2$ versus the soft thresholded ordinary least squares (STOLS) procedure as defined in display (3.5). We have simulated the lower bound instance from Corollary 3.1; for each sample size n, we simulate oracle Lasso and STOLS on a pair $(X_n, \theta_n^\star)$, with the dimension $d = n$ and lower singular value bound $B = \sqrt{n}$. Further details on the simulation are provided in Section 3.4.2.

3.1.5 Connections to prior work

In this section, we draw connections and comparisons between our work and existing literature.

Linear regression with elliptical or no constraints. Without any parameter restrictions, the exact minimax rate for linear regression when error is measured in the ℓ_2 norm along with the dependence on the design matrix is known: it is given by $\sigma^2 \mathsf{Tr}((X^\top X)^{-1})$ [72]. These results match our intuition that as the smallest singular value of X decreases, the hardness of estimating $\theta^\star$ increases. It is also worth mentioning that the design matrix does not play a role, apart from being invertible, in determining the optimal rate for the in-sample prediction error. The rate is given uniformly by $\frac{\sigma^2 d}{n}$ when $n \geqslant d$ [58].

On the other hand, with ℓ_2- or elliptical parameter constraints the minimax rate in both fixed and random design was established in the recent paper [95]. Although that work shows the dependence on the design, the rate is not explicit and the achievable result requires potentially solving a semidefinite program. More explicit results were previously derived under proportional asymptotics in the paper [31], in the restricted setting of Gaussian isotropic random design. The author is able to establish the asymptotic minimaxity of a particular ridge regression estimator. These type of results are not immediately useful for our work, since they are based on linear shrinkage procedures which are known to be minimax-suboptimal even in orthogonal design, in the ℓ_p setting for $p < 2$ [33].

Gaussian sequence model and sparse linear regression. In the case of orthogonal design, *i.e.*, when $\frac{1}{n}X^\mathsf{T}X = I_d$, the minimax risk of estimating sparse $\theta^\star$ is known; see [33]. It can also be shown that Lasso, with optimally-tuned parameter, can achieve the (order-wise) optimal estimation rate [62]. This roughly corresponds to the case with $B = 1$ in our consideration. Our work generalizes this line of research in the sense that we characterize the optimal rate of estimation over the larger family of design matrices $\frac{1}{n}X^\mathsf{T}X \succcurlyeq \frac{1}{B}I_d$. In stark contrast to the Gaussian sequence model, Lasso is no longer optimal, even with the oracle knowledge of the true parameter.

Without assuming an orthogonal design, [20] provides a design-dependent lower bound in the exact sparsity case (*i.e.*, $p = 0$). The lower bound depends on the design through its Frobenius norm $\|X\|_\mathsf{F}^2$. Similarly, in the weak sparsity case, [99] provides lower bounds depending on the maximum column norm of the design matrix. However, matching upper bounds are not provided in this general design case. In contrast, using the minimum singular value of X (*c.f.,* the parameter B) allows us to obtain matching upper and lower bounds in sparse regression.

Suboptimality of Lasso. The suboptimality of Lasso for minimizing the *prediction* error has been noted in the case of exact sparsity (*i.e.*, $p = 0$). To our knowledge, previous studies required a carefully chosen design matrix which was highly-correlated. For instance, it was shown that for certain highly-correlated designs the Lasso can achieve only a slow rate $(1/\sqrt{n})$, while information-theoretically, the optimal rate is faster $(1/n)$; see for instance the papers [117, 65]. Additionally in the paper [21], the authors exhibit the failure of Lasso for a *fixed* regularization parameter, which does not necessarily rule out the optimality of other Lasso variants. Similarly, in the paper [39], it is shown via a correlated design matrix and a 2-sparse vector, that the norm-constrained version of Lasso can only achieve a slow rate in terms of the prediction error. Again, this result does not rule out the optimality of other variants of the Lasso. In addition, in the paper [30], there is an example for which Lasso with any fixed (*i.e.*, independent from the observed data) choice of regularization would fail to achieve the optimal rate. Again, this fails to rule out data-dependent choices of regularization or other variants of the Lasso. In our work, we are able to rule out the optimality of the Lasso by considering a simple diagonal design matrix which exhibits no correlations among

the columns. Nonetheless, for any $p \in [0,1]$, we show that the Lasso will fall short of optimality by polynomial factors in the sample size. Our result also simultaneously rules out the optimality of constrained, penalized, and even data-dependent variants of the Lasso, in contrast to the literature described above.

Covariate shift. As mentioned previously, our work is also related to linear regression under covariate shift [73, 126, 36]. The statistical analysis of covariate shift, albeit with an asymptotic nature, dates back to the seminal work by Shimodaira [105]. Recently, nonasymptotic minimax analysis of covariate shift has gained much attention in unconstrained parametric models [44], nonparametric classification [69], and also nonparametric regression [94, 77, 121].

3.2 A closer look at the failure mode of Lasso

In this section, we take a closer look at the failure instance for Lasso. We will investigate the performance of the Lasso on diagonal design matrices $X_\alpha \in \mathbf{R}^{n \times d}$ which satisfy, when $d = 2k$,

$$\frac{1}{n} X_\alpha^\mathsf{T} X_\alpha = \begin{pmatrix} \frac{\alpha}{B} I_k & 0 \\ 0 & \frac{1}{B} I_k \end{pmatrix}.$$

Thus, this matrix has condition number α and satisfies $X_\alpha \in \mathcal{X}_{n,d}(B)$ for all $\alpha \geqslant 1$. As our proof of Theorem 3.1 reveals, from an information-theoretic perspective, the hardest design matrix X_α is with the choice $\alpha = 1$: when all directions have the worst possible signal-to-noise ratio. Strikingly, this is not the case for the Lasso: there are in fact choices of $\alpha \gg 1$ which are even harder for the Lasso.

Theorem 3.3. *Fix $n \geqslant d \geqslant 2$ and let $\sigma, B > 0$ be given. For $\alpha \geqslant 1$, on the diagonal design X_α,*

(a) if $p \in (0,1]$ and $R > 0$, then there is a vector $\theta^\star \in \mathbf{R}^d$ such that $\|\theta^\star\|_p \leqslant R$ but

$$\mathbf{E}_{y \sim \mathsf{N}(X_\alpha \theta^\star, \sigma^2 I_n)} \Big[\inf_{\lambda > 0} \|\widehat{\theta}_\lambda(X_\alpha, y) - \theta^\star\|_2^2 \Big] \geqslant \frac{9}{20000} \Big(\frac{\sigma^2 B d}{n\alpha} \wedge R^2 \Big(\frac{\sigma^2 B}{R^2 n} \alpha \Big)^{1 - p/2} \wedge R^2 \Big), \quad \text{and}$$

(b) if $p = 0$ and $s \in [d]$, then there is a vector $\theta^\star \in \mathbf{R}^d$ which is s-sparse but

$$\mathbf{E}_{y \sim \mathsf{N}(X_\alpha \theta^\star, \sigma^2 I_n)} \Big[\inf_{\lambda > 0} \|\widehat{\theta}_\lambda(X_\alpha, y) - \theta^\star\|_2^2 \Big] \geqslant \frac{9}{20000} \Big(\frac{\sigma^2 B d}{n\alpha} \wedge \frac{\sigma^2 B s}{n} \alpha \Big).$$

The proof of Theorem 3.3 is presented in Section 3.3.3. We now make several comments on the implications of this result.

We emphasize that the dependence of the Lasso on the parameter α, which governs the condition number of the matrix X_α, is suboptimal, as revealed by Theorem 3.3. At a high-level, large α should only make the ability to estimate $\theta^\star$ *easier*—it effectively increases the signal-to-noise ratio in certain directions. This can also be seen from Theorem 3.1: the conditioning of the design matrix *does not* enter into the worst-case rate of estimation when the bottom signular value of X is bounded. Nonetheless, Theorem 3.3 shows that the Lasso actually can suffer when the condition number α is large.

Proof of Corollary 3.1. We now complete the proof of Corollary 3.1 given Theorem 3.3.

Maximizing over the parameter $\alpha \geqslant 1$ appearing in our result, we can determine a particularly nasty configuration of the conditioning of the design matrix for the Lasso. Doing so, we find that for $p \in (0, 1]$ and $R > 0$ that

$$\sup_{X \in \mathcal{X}_{n,d}(B)} \sup_{\|\theta^\star\|_p \leqslant R} \mathbf{E}\Big[\inf_{\lambda > 0} \|\widehat{\theta}_\lambda(X, y) - \theta^\star\|_2^2\Big] \gtrsim R^2 \bigg(\Big(\frac{\sigma^2 B\sqrt{d}}{R^2 n}\Big)^{\frac{4-2p}{4-p}} \wedge 1\bigg).$$

This is exhibited by considering the lower bound in Theorem 3.3 with the choice $\alpha^\star(p) = (\tau_n^2 d^{2/p})^{p/(4-p)}$. On the other hand, if $p = 0$, we have for $s \in [d]$ that

$$\sup_{X \in \mathcal{X}_{n,d}(B)} \sup_{\|\theta^\star\|_0 \leqslant s} \mathbf{E}\Big[\inf_{\lambda > 0} \|\widehat{\theta}_\lambda(X, y) - \theta^\star\|_2^2\Big] \gtrsim \frac{\sigma^2 B}{n}\sqrt{sd}$$

The righthand side above is exhibited by considering the lower bound with the choice $\alpha^\star(0) = \sqrt{d/s}$.

The proof is completed by setting $d = n$, $B = \sqrt{n}$, and $\sigma = R = 1$.

3.3 Proofs

In this section, we present the proofs for the main results of this section. We start with introducing a few useful notations. For a positive integer k, we define $[k] \coloneqq \{1, \ldots, k\}$. For a real number x, we define $\lfloor x \rfloor$ to be the largest integer less than or equal to x and $\{x\}$ to be the fractional part of x.

3.3.1 Proof of Theorem 3.1

Our proof is based on a decision-theoretic reduction to the Gaussian sequence model. It holds in far greater generality, and so we actually prove a more general claim which could be of interest to other linear regression problems on other parameter spaces or with other loss functions.

To develop the claim, we first need to introduce notation. Let $\Theta \subset \mathbf{R}^d$ denote a parameter space, and let $\ell \colon \Theta \times \mathbf{R}^d \to \mathbf{R}$ be a given loss function. We define two minimax rates,

$$\mathfrak{M}_{\text{seq}}(\Theta, \ell, \nu) \coloneqq \inf_{\widehat{\mu}} \sup_{\mu^\star \in \Theta} \mathbf{E}_{y \sim \mathsf{N}(\mu^\star, \nu^2 I_d)} \Big[\ell\big(\mu^\star, \widehat{\mu}(y)\big) \Big], \quad \text{and} \tag{3.7a}$$

$$\mathfrak{M}_{\text{reg}}(\Theta, \ell, \sigma, B, n) \coloneqq \inf_{\widehat{\theta}} \sup_{X \in \mathcal{X}_{n,d}(B)} \sup_{\theta^\star \in \Theta} \mathbf{E}_{y \sim \mathsf{N}(X\theta^\star, \sigma^2 I_n)} \mathbf{E} \Big[\ell\big(\theta^\star, \widehat{\theta}(X, y)\big) \Big]. \tag{3.7b}$$

The definitions above correspond to the minimax rates of estimation over the ℓ_p ball of radius $R > 0$ in $\mathbf{R}^d$ for the Gaussian sequence model, in the case of definition (3.7a), and for n-sample linear regression with B-bounded design, in the case of definition (3.7b). The infima range over measurable functions of the observation vector y, in both cases.

The main result we need is the following statistical reduction from linear regression to mean estimation in the Gaussian sequence model.

Proposition 3.1 (Reduction to sequence model)**.** *Fix $n, d \geqslant 1$ and $\sigma, B > 0$. Let $\Theta \subset \mathbf{R}^d$, and $\ell \colon \Theta \times \mathbf{R}^d \to \mathbf{R}$ be given. If $\ell(\theta, \cdot) \colon \mathbf{R}^d \to \mathbf{R}$ is a convex function for each $\theta \in \Theta$, then*

$$\mathfrak{M}_{\text{reg}}(\Theta, \ell, \sigma, B, n) = \mathfrak{M}_{\text{seq}}\bigg(\Theta, \ell, \sqrt{\frac{\sigma^2 B}{n}}\bigg).$$

Deferring the proof of Proposition 3.1 to Section 3.3.1.1 for the moment, we note that it immediately implies Theorem 3.1. Indeed, we set $\Theta = \Theta_{d,p}(R)$ and $\ell = \ell_{\text{sq}}$ where

$$\Theta_{d,p}(R) \coloneqq \{\theta \in \mathbf{R}^d : \|\theta\|_p \leqslant R\}, \quad \text{and} \quad \ell_{\text{sq}}(\theta, \widehat{\theta}) = \|\widehat{\theta} - \theta\|_2^2.$$

With these choices, we obtain

$$\mathfrak{M}_{n,d}(p, \sigma, R, B) = \mathfrak{M}_{\text{reg}}(\Theta_{d,p}(R), \ell_{\text{sq}}, \sigma, B, n) = \mathfrak{M}_{\text{seq}}\bigg(\Theta_{d,p}, \ell_{\text{sq}}, \sqrt{\frac{\sigma^2 B}{n}}\bigg),$$

where the final equality follows from Proposition 3.1. The righthand side then corresponds to estimation in the ℓ_2 norm over the Guassian sequence model with parameter space corresponding to an ℓ_p ball in $\mathbf{R}^d$, which is a well-studied problem [33, 12, 62]; we thus immediately obtain the result via classical results (for a precise statement with explicit constants, see Propositions 3.2 and 3.3 presented in Section 3.4.1).

3.3.1.1 Proof of Proposition 3.1

We begin by lower bounding the regression minimax risk by the sequence model minimax risk. Indeed, let $X_\star$ be such that $\frac{1}{n}X_\star^\mathsf{T}X_\star = \frac{1}{B}I_d$. Then we have

$$\begin{aligned}\mathfrak{M}_{\mathrm{reg}}(\Theta, \ell, \sigma, B, n) &\geqslant \inf_{\widehat{\theta}} \sup_{\theta^\star \in \Theta} \mathbf{E}_{y \sim \mathsf{N}(X_\star\theta^\star, \sigma^2 I_n)} \Big[\ell\big(\theta^\star, \widehat{\theta}\big)\Big] \\ &\geqslant \inf_{\widehat{\theta}} \sup_{\theta^\star \in \Theta} \mathbf{E}_{z \sim \mathsf{N}\left(\theta^\star, \frac{\sigma^2 B}{n} I_d\right)} \Big[\ell\big(\theta^\star, \widehat{\theta}\big)\Big] \\ &= \mathfrak{M}_{\mathrm{seq}}\left(\Theta, \ell_{\mathrm{sq}}, \sqrt{\frac{\sigma^2 B}{n}}\right).\end{aligned}$$

The penultimate equality follows by noting that in the regression model, $\mathcal{P}_{X_\star} \coloneqq \{\mathsf{N}(X_\star\theta, \sigma^2 I_d) : \theta \in \Theta\}$, the ordinary least squares (OLS) estimate is a sufficient statistic. Therefore, by the Rao-Blackwell Theorem, there exists a minimax optimal estimator which is only a function of the OLS estimate. For any $\theta^\star$, the ordinary least squares estimator has the distribution, $\mathsf{N}\left(\theta^\star, \frac{\sigma^2 B}{n} I_d\right)$, which provides this equality.

We now turn to the upper bound. Let $\widehat{\theta}^{\mathsf{OLS}}(X, y) = (X^\mathsf{T}X)^{-1}X^\mathsf{T}y$ denote the ordinary least squares estimate. For any estimator $\widehat{\mu}$, we define

$$\widetilde{\theta}(X, y) \coloneqq \mathbf{E}_{\xi \sim \mathsf{N}(0, W)} \Big[\widehat{\mu}\big(\widehat{\theta}^{\mathsf{OLS}}(X, y) + \xi\big)\Big], \qquad \text{where} \quad W = \frac{\sigma^2 B}{n} I_d - \sigma^2 (X^\mathsf{T}X)^{-1}.$$

Note that for any $X \in \mathcal{X}_{n,d}(B)$, we have $W \succcurlyeq 0$. Additionally, by Jensen's inequality, for any $X \in \mathcal{X}_{n,d}(B)$, and any $\theta^\star \in \Theta$,

$$\begin{aligned}\mathbf{E}_{y \sim \mathsf{N}(X\theta^\star, \sigma^2 I_n)} \Big[\ell\big(\theta^\star, \widetilde{\theta}(X, y)\big)\Big] &\leqslant \mathbf{E}_{y \sim \mathsf{N}(X\theta^\star, \sigma^2 I_n)} \mathbf{E}_{\xi \sim \mathsf{N}(0, W)} \Big[\ell\big(\theta^\star, \widehat{\mu}\big(\widehat{\theta}^{\mathsf{OLS}}(X, y) + \xi\big)\big)\Big] \\ &= \mathbf{E}_{z \sim \mathsf{N}\left(\theta^\star, \frac{\sigma^2 B}{n} I_d\right)} \Big[\ell\big(\theta^\star, \widehat{\mu}(z)\big)\Big]\end{aligned}$$

Passing to the supremum over $\theta^\star \in \Theta$ on each side and then taking the infimum over measurable estimators, we immediately see that the above display implies

$$\mathfrak{M}_{\mathrm{reg}}(\Theta, \ell, \sigma, B, n) \leqslant \mathfrak{M}_{\mathrm{seq}}\left(\Theta, \ell, \sqrt{\frac{\sigma^2 B}{n}}\right),$$

as needed.

3.3.2 Proof of Theorem 3.2

We begin by bounding the risk for soft thresholding procedures, based on a rescaling and monotonicity argument and applying results from [62]. To state it, we need to define the

quantitites

$$(\nu_n^\star)^2 \coloneqq \frac{\sigma^2 B}{n} \quad \text{and} \quad \rho_n(\theta,\eta) \coloneqq d\,(\nu_n^\star)^2 \mathrm{e}^{-(\eta/\nu_n^\star)^2/2} + \sum_{i=1}^d \theta_i^2 \wedge (\nu_n^\star)^2 + \sum_{i=1}^d \theta_i^2 \wedge \eta^2.$$

Then we have the following risk bound.

Lemma 3.1. *For any $\theta^\star \in \mathbf{R}^d$ and any $\eta > 0$ we have*

$$\sup_{X \in \mathcal{X}_{n,d}(B)} \mathbf{E}\left[\|\widehat{\theta}_\eta^{\mathsf{STOLS}}(X,y) - \theta^\star\|_2^2\right] \leqslant \rho_n(\theta^\star,\eta).$$

We now define for $\zeta > 0$ and a subset $\Theta \subset \mathbf{R}^d$,

$$T(\zeta,\Theta) \coloneqq \sup_{\theta^\star \in \Theta} \sum_{i=1}^d (\theta_i^\star)^2 \wedge \zeta^2.$$

Lemma 3.1 then yields with the choice $\eta = \gamma\nu_n^\star$ for some $\gamma \geqslant 1$ that

$$\sup_{\theta^\star\in\Theta} \sup_{X \in \mathcal{X}_{n,d}(B)} \mathbf{E}\left[\|\widehat{\theta}_\eta^{\mathsf{STOLS}}(X,y) - \theta^\star\|_2^2\right] \leqslant 3\Big[d(\nu_n^\star)^2 \mathrm{e}^{-\gamma^2/2} \vee T(\gamma\nu_n^\star,\Theta)\Big]. \tag{3.8}$$

We bound the map T for the ℓ_p balls of interest. To state the bound, we use the shorthand Θ_p for the radius-R ℓ_p ball in $\mathbf{R}^d$ centered at the origin for $p \neq 0$, and for $p = 0$, the set of s-sparse vectors in $\mathbf{R}^d$, for $s \in [d]$.

Lemma 3.2. *Let $d \geqslant 1$ be fixed. We have the following relations:*

(a) in the case $p \in (0,1]$, we have for any $\zeta > 0$,

$$T(\zeta,\Theta_p) \leqslant R^2\left[\left(\frac{\zeta}{R}\right)^2 d \wedge \left(\frac{\zeta}{R}\right)^{2-p} \wedge 1\right]$$

for any $R > 0$, and

(b) in the case $p = 0$, we have for any $\zeta > 0$,

$$T(\zeta,\Theta_p) = \zeta^2 s,$$

for any $s \in [d]$.

To complete the argument, we now split into the two cases of hard and weak sparsity.

When $p = 0$: Combining inequality (3.8) together with Lemma 3.2, we find for $\eta = \gamma\nu_n^\star$, $\gamma \geqslant 1$, that

$$\sup_{\theta^\star\in\Theta} \sup_{X \in \mathcal{X}_{n,d}(B)} \mathbf{E}\left[\|\widehat{\theta}_\eta^{\mathsf{STOLS}}(X,y) - \theta^\star\|_2^2\right] \leqslant 3(\nu_n^\star)^2\Big[d\mathrm{e}^{-\gamma^2/2} \vee \gamma^2 s\Big] = 6(\nu_n^\star)^2 s \log\left(\mathrm{e}\frac{d}{s}\right),$$

where the last equality holds with $\gamma^2 = 2\log(\mathrm{e}d/s)$.

When $p \in (0,1]$: Combining inequality (3.8) together with Lemma 3.2, we find for $\eta = \gamma\nu_n^\star$, $\gamma \geqslant 1$

$$\sup_{\theta^\star \in \Theta} \sup_{X \in \mathfrak{X}_{n,d}(B)} \mathbf{E}\left[\|\widehat{\theta}_\eta^{\mathsf{STOLS}}(X,y) - \theta^\star\|_2^2\right] \leqslant 3R^2\left[d\tau_n^2 \mathrm{e}^{-\gamma^2/2} \vee \left(\gamma^{2-p}\tau_n^{2-p} \wedge 1\right)\right] \tag{3.9}$$

Above, we used $\tau_n^2 R^2 = (\nu_n^\star)^2$ and $\gamma^2\tau_n^2 d \geqslant \gamma^{2-p}\tau_n^{2-p}$, which holds since $\gamma \geqslant 1$ and $\tau_n^2 \geqslant d^{-2/p}$. If we take $\gamma^2 = 2\log(\mathrm{e}d\tau_n^p)$, then note $\gamma^2 \geqslant 1$ by $\tau_n^2 \geqslant d^{-2/p}$ and the term in brackets in inequality (3.9) satisfies

$$\begin{aligned} d\tau_n^2 \mathrm{e}^{-\gamma^2/2} \vee \left(\gamma^{2-p}\tau_n^{2-p} \wedge 1\right) &= \frac{\tau_n^{2-p}}{\mathrm{e}} \vee \left((2\tau_n^2 \log(\mathrm{e}d\tau_n^p))^{1-p/2} \wedge 1\right) \\ &\leqslant 2\left[\tau_n^{2-p} \vee \left((\tau_n^2 \log(\mathrm{e}d\tau_n^p))^{1-p/2} \wedge 1\right)\right] \\ &= 2(\tau_n^2 \log(\mathrm{e}d\tau_n^p))^{1-p/2}, \end{aligned}$$

which follows by $\tau_n^2 \in [d^{-2/p}, \log^{-1}(\mathrm{e}d)]$.

Thus, to complete the proof of Theorem 3.2 we only need to provide the proofs of the lemmas used above.

3.3.2.1 Proof of Lemma 3.1

Note that if $z = \widehat{\theta}^{\mathsf{OLS}}(X,y)$ then $\widehat{\theta}_\eta^{\mathsf{STOLS}}(X,y) = \mathsf{S}_\eta(z) = \mathsf{S}_\eta(\theta^\star + \xi)$ where $\xi \sim \mathsf{N}\left(0, \frac{\sigma^2}{n}\Sigma_n^{-1}\right)$, where we recall $\Sigma_n \coloneqq (1/n)X^\mathsf{T}X$. We now recall some classical results regarding the soft thresholding estimator. Let us write for $\lambda > 0$ and $\mu \in \mathbf{R}$,

$$\begin{aligned} r_{\mathrm{S}}(\lambda,\mu) &\coloneqq \mathbf{E}_{y\sim\mathsf{N}(\mu,1)}\left[\left(\mathsf{S}_\lambda(y) - \mu\right)^2\right], \quad \text{and,} \\ \widetilde{r}_{\mathrm{S}}(\lambda,\mu) &\coloneqq \mathrm{e}^{-\lambda^2/2} + \left(1 \wedge \mu^2\right) + \left(\lambda^2 \wedge \mu^2\right). \end{aligned}$$

Using $(a+b)\wedge c \leqslant a \wedge c + b \wedge c$ for nonnegative $a, b, c \geqslant 0$, Lemma 8.3 and the inequalities $r_{\mathrm{S}}(\lambda,0) \leqslant 1 + \lambda^2$ and $r_{\mathrm{S}}(\lambda,0) \leqslant \mathrm{e}^{-\lambda^2/2}$ on page 219 of the monograph [62], we find that $r_{\mathrm{S}}(\lambda,\mu) \leqslant \widetilde{r}_{\mathrm{S}}(\lambda,\mu)$. Define $\nu_i^2 \coloneqq \frac{\sigma^2}{n}(\Sigma_n^{-1})_{ii}$. Using the fact that $(\mathsf{S}_\eta(z))_i = \mathsf{S}_{\frac{\eta}{\nu_i}}\left(\frac{z_i}{\nu_i}\right)$ for $i \in [d]$, we obtain

$$\mathbf{E}[((\mathsf{S}_\eta(z))_i - \theta_i^\star)^2] = \nu_i^2 r_{\mathrm{S}}\left(\frac{\eta}{\nu_i}, \frac{\theta_i^\star}{\nu_i}\right) \leqslant \nu_i^2 \widetilde{r}_{\mathrm{S}}\left(\frac{\eta}{\nu_i}, \frac{\theta_i^\star}{\nu_i}\right).$$

Summing over the coordinates yields

$$\begin{aligned} \mathbf{E}\left[\|\widehat{\theta}_\eta^{\mathsf{STOLS}}(X,y) - \theta^\star\|_2^2\right] &\leqslant \sum_{i=1}^d \nu_i^2 \widetilde{r}_{\mathrm{S}}\left(\frac{\eta}{\nu_i}, \frac{\theta_i^\star}{\nu_i}\right) \\ &= \sum_{i=1}^d \nu_i^2 \mathrm{e}^{-(\eta/\nu_i)^2/2} + \left((\theta_i^\star)^2 \wedge \nu_i^2\right) + \left((\theta_i^\star)^2 \wedge \eta^2\right) \\ &\leqslant \rho_n(\theta^\star, \eta), \end{aligned}$$

where the last inequality follows by noting that both $\nu \mapsto \nu^2 \mathrm{e}^{-(\eta/\nu)^2/2}$ and $\nu \mapsto \theta^2 \wedge \nu^2$ are nondecreasing functions of $\nu > 0$. Noting that this inequality holds uniformly on $X \in \mathcal{X}_{n,d}(B)$ and passing to the supremum yields the claim.

3.3.2.2 Proof of Lemma 3.2

The proof of claim (b) is immediate, so we focus on the case $p \in (0, 1]$, $R > 0$. We consider three cases for the tuple (R, ζ, p, d). Combination of all three cases will yield the claim.

When $R \geqslant \zeta d^{1/p}$: Evidently, for each θ such that $\|\theta\|_p \leqslant R$, we have

$$\sum_{i=1}^{d} \theta_i^2 \wedge \zeta^2 \leqslant \zeta^2 d.$$

When $R \leqslant \zeta$: This case is immediate, since $\theta \in \Theta_p$ implies $\|\theta\|_2 \leqslant \|\theta\|_p \leqslant R \leqslant \zeta$.

When $\zeta \leqslant R \leqslant \zeta d^{1/p}$: In this case, by rescaling and putting $\varepsilon \coloneqq \frac{\zeta^2}{R^2}$, we have

$$\sup_{\|\theta\|_p \leqslant R} \sum_{i=1}^{d} \theta_i^2 \wedge \zeta^2 = R^2 \Bigg[\sup_{\lambda \in \Delta_d} \sum_{i : \lambda_i \geqslant \varepsilon^{p/2}} \varepsilon + \sum_{i : \lambda_i < \varepsilon^{p/2}} \lambda_i^{2/p} \Bigg] = \varepsilon R^2 \Big(\big\lfloor \varepsilon^{-p/2} \big\rfloor + \{\varepsilon^{-p/2}\}^{2/p} \Big)$$

where above Δ_d denotes the probability simplex in $\mathbf{R}^d$. Noting that $\varepsilon \leqslant 1$ and $p \leqslant 1$ we have

$$\Big(\big\lfloor \varepsilon^{-p/2} \big\rfloor + \{\varepsilon^{-p/2}\}^{2/p} \Big) \leqslant \varepsilon^{-p/2},$$

which in combination with the previous display shows that

$$\sup_{\|\theta\|_p \leqslant R} \sum_{i=1}^{d} \theta_i^2 \wedge \zeta^2 \leqslant R^2 \varepsilon^{1-p/2}.$$

To conclude, now note that $R^2 \varepsilon^{1-p/2} = R^p \zeta^{2-p}$.

3.3.3 Proof of Theorem 3.3

Since X_α has nonzero entries only on the diagonal, we can derive an explicit representation of the Lasso estimate, as defined in display (3.6). To develop this, we first recall the notion of the *soft thresholding operator*, which is defined by a parameter $\eta > 0$ and then satisfies

$$\mathsf{S}_\eta(v) \coloneqq \operatorname*{arg\,min}_{u \in \mathbf{R}} \Big\{ (u - v)^2 + 2\eta |u| \Big\}.$$

We then start by stating the following lemma which is crucial for our analysis. It is a straightforward consequence of the observation that

$$\widehat{\theta}_\lambda(X_\alpha, y)_i = \begin{cases} \mathsf{S}_{\lambda B/\alpha}(z_i) & 1 \leqslant i \leqslant k \\ \mathsf{S}_{\lambda B}(z_i) & k+1 \leqslant i \leqslant d \end{cases},$$

where we have defined the independent random variables $z_i \sim \mathsf{N}\left(\theta_i^\star, \frac{\sigma^2 B}{n\alpha}\right)$ if $i \leqslant k$ and $z_i \sim \mathsf{N}\left(\theta_i^\star, \frac{\sigma^2 B}{n}\right)$ otherwise.

Lemma 3.3. *Let $\theta^\star \in \mathbf{R}^d$. Then for the design matrix X_α, we have*

$$\|\widehat{\theta}_\lambda(X_\alpha, y) - \theta^\star\|_2^2 = \sum_{i=1}^{k} \left(\mathsf{S}_{\lambda B/\alpha}(z_i) - \theta_i^\star\right)^2 + \sum_{i=k+1}^{d} \left(\mathsf{S}_{\lambda B}(z_i) - \theta_i^\star\right)^2.$$

We will now focus on vectors $\theta^\star(\eta) = (0_k, \eta, 0_{d-2k})$, which are parameterized by $\eta \in \mathbf{R}^k$. For these vectors, we can further lower bound the best risk as

$$\inf_{\lambda > 0} \|\widehat{\theta}_\lambda(X_\alpha, y) - \theta^\star(\eta)\|_2^2 \geqslant T_1 \wedge T_2(\eta)$$

where we have defined

$$\overline{\lambda} = \sqrt{\frac{\sigma^2}{n}\frac{\alpha}{B}}, \quad T_1 \coloneqq \inf_{\lambda \leqslant \overline{\lambda}} \sum_{i=1}^{k} \left(\mathsf{S}_{\lambda B/\alpha}(z_i)\right)^2 \quad \text{and} \quad T_2(\eta) \coloneqq \inf_{\lambda \geqslant \overline{\lambda}} \sum_{i=k+1}^{2k} \left(\mathsf{S}_{\lambda B}(z_i) - \eta_i\right)^2.$$

We now move to lower bound T_1 and $T_2(\eta)$ by auxiliary, independent random variables.

Lemma 3.4 (Lower bound on T_1)**.** *Then, for any $\eta \in \mathbf{R}^k$ if $\theta^\star = \theta^\star(\eta)$, we have*

$$T_1 \geqslant \frac{1}{4}\frac{\sigma^2 B}{n\alpha} Z \qquad \text{where} \quad Z \coloneqq \sum_{i=1}^{k} \mathbf{1}\Big\{\frac{|z_i|}{\sqrt{\sigma^2 B/(n\alpha)}} \geqslant 3/2\Big\}.$$

Lemma 3.5 (Lower bound on $T_2(\eta)$)**.** *Fix $\eta \in \mathbf{R}^k$ if $\theta^\star = \theta^\star(\eta)$, and suppose that*

$$0 \leqslant \eta_i \leqslant 2\sqrt{\frac{\sigma^2 B\alpha}{n}} \quad \text{for all } i \in [k].$$

Then, we have

$$T_2(\eta) \geqslant \frac{1}{4}\sum_{i=1}^{k} \eta_i^2 W_i \quad \text{where} \quad W_i \coloneqq \mathbf{1}\{z_{k+i} \leqslant \eta_i\}$$

Note that Z is distributed as a Binomial random variable: $Z \sim \mathsf{Bin}(k, p)$ where $p \coloneqq \mathbf{P}\{|\mathsf{N}(0,1)| \geqslant 3/2\}$. Similarly, W_i are Bernoulli: we have $W_i \sim \mathsf{Ber}(1/2)$.

Lower bound for $p > 0$: We consider two choices of η. First suppose that $4\tau_n^2\alpha \leqslant 1$. Then, we will consider $\eta = R\delta\mathbf{1}_\ell$ where

$$\delta \coloneqq 2\tau_n\sqrt{\alpha} \quad \text{and} \quad \ell \coloneqq k \wedge \lfloor \delta^{-p} \rfloor$$

For this choice of η we have by assumption that $\tau_n^2 \geqslant d^{-2/p}$ that $\ell \geqslant (1/2)\delta^{-p}$ and so

$$T_2(\eta) \geqslant \frac{1}{4}R^2\delta^2\ell\overline{W_\ell} \geqslant \frac{R^2}{8}\delta^{2-p}\overline{W_\ell} \geqslant \frac{R^2}{8}\Big(\tau_n^2\alpha\Big)^{1-p/2}\overline{W_\ell}$$

Above, $\overline{W_\ell} \coloneqq (1/\ell)\sum_{i=1}^{\ell} W_i$. On the other hand, if $4\tau_n^2\alpha \geqslant 1$, we take $\eta' = Re_1$, and we consequently obtain

$$T_2(\eta') \geqslant \frac{R^2}{4}W_1$$

Taking $\delta = 1/2$ in Lemma 3.4, let us define

$$c_1 \coloneqq \min_{1\leqslant\ell\leqslant k} \mathbf{P}\left\{\overline{W_\ell} \geqslant \frac{1}{2}\right\} \quad \text{and} \quad c_2 \coloneqq \mathbf{P}\Big\{Z \geqslant kp\Big\}.$$

Let us take

$$\theta^\star_\alpha \coloneqq \begin{cases} \theta^\star(\eta) & 4\tau_n^2\alpha \leqslant 1 \\ \theta^\star(\eta') & 4\tau_n^2\alpha > 1 \end{cases}.$$

Then combining Lemmas 3.4 and 3.5 and the lower bounds on $T_2(\eta), T_2(\eta')$ above, we see that

$$\mathbf{E}_{y\sim\mathsf{N}(X_\alpha\theta^\star_\alpha,\sigma^2 I_n)}\Big[\inf_{\lambda>0}\|\widehat{\theta}_\lambda(X_\alpha, y) - \theta^\star_\alpha\|_2^2\Big] \geqslant \frac{c_2c_1p}{16}\Big(\frac{\sigma^2Bd}{n\alpha} \wedge R^2\Big(\frac{\sigma^2B}{R^2n}\alpha\Big)^{1-p/2} \wedge R^2\Big) \tag{3.10}$$

where above we have used $k \geqslant d/4$.

Lower bound when $p = 0$: In this case, we let $s' = s \wedge k$. Note that $s' \geqslant s/4$. We then set $\eta = 2\sqrt{\frac{\sigma^2B\alpha}{n}}\mathbf{1}_{s'}$, and this yields the lower bound

$$T_2(\eta) \geqslant \frac{1}{8}\frac{\sigma^2Bs}{n}\alpha\overline{W_{s'}}$$

In this case, we have, after combining this bound with the bound on T_1 that for $\theta^\star_\alpha \coloneqq \theta^\star(\eta)$ as defined above,

$$\mathbf{E}_{y\sim\mathsf{N}(X_\alpha\theta^\star_\alpha,\sigma^2 I_n)}\Big[\inf_{\lambda>0}\|\widehat{\theta}_\lambda(X_\alpha, y) - \theta^\star_\alpha\|_2^2\Big] \geqslant \frac{c_2c_1p}{16}\Big(\frac{\sigma^2Bd}{n\alpha} \wedge \frac{\sigma^2Bs}{n}\alpha\Big) \tag{3.11}$$

The proof of Theorem 3.3 is complete after combining inequalities (3.10) (3.11), and the following lemma.

Lemma 3.6. *The constant factor $c \coloneqq \frac{c_1c_2p}{16}$ is lower bounded as $c \geqslant \frac{9}{20000}$.*

We conclude this section by proving the lemmas above.

3.3.3.1 Proof of Lemma 3.4

For the first term, T_1, we note that for any $\lambda \leqslant \overline{\lambda}$ we evidently have for each $i \in [k]$ and for any $\zeta > 0$, that

$$\Big(\mathsf{S}_{\lambda B/\alpha}(z_i)\Big)^2 = (|z_i| - \lambda B/\alpha)_+^2 \geqslant (|z_i| - \overline{\lambda} B/\alpha)_+^2 \geqslant \zeta^2 \frac{\sigma^2 B}{n\alpha} \mathbf{1}\Big\{\frac{|z_i|}{\sqrt{\sigma^2 B/(n\alpha)}} \geqslant 1 + \zeta\Big\}$$

Summing over $i \in [k]$, and taking $\zeta = 1/2$, we thus obtain the claimed almost sure lower bound.

3.3.3.2 Proof of Lemma 3.5

Fix any i such that $1 \leqslant i \leqslant k$. For any fixed $\lambda \geqslant \overline{\lambda}$, note

$$S_{\lambda B}(z_{k+i}) \notin [\eta_i/2, 3\eta_i/2] \quad \text{implies} \quad \Big|S_{\lambda B}(z_{k+i}) - \eta_i\Big| \geqslant \frac{\eta_i}{2}.$$

Note that the condition $S_{\lambda B}(z_{k+i}) \notin [\eta_i/2, 3\eta_i/2]$ is equivalent to $z_{k+i} \notin [\eta_i/2 + \lambda B, 3\eta_i/2 + \lambda B]$, Therefore, if $z_{k+i} \leqslant \eta_i/2 + \overline{\lambda} B$, then then for all $\lambda \geqslant \overline{\lambda}$ we have $|S_{\lambda B}(z_{k+i}) - \eta_i| \geqslant \frac{\eta_i}{2}$. Equivalently, we have that

$$T_2 \geqslant \frac{1}{4} \sum_{i=1}^{k} \eta_i^2 \mathbf{1}\{z_{k+i} \leqslant \eta_i/2 + \overline{\lambda} B\} \geqslant \frac{1}{4} \sum_{i=1}^{k} \eta_i^2 \mathbf{1}\{z_{k+i} \leqslant \eta_i\}$$

The final relation uses the distribution of z_{k+i} and

$$\frac{-\eta_i/2 + \overline{\lambda} B}{\sqrt{\sigma^2 B/n}} = \sqrt{\alpha} - \frac{1}{2}\sqrt{\alpha}\sqrt{\frac{n\eta_i^2}{\alpha\sigma^2 B}} \geqslant 0$$

which holds by assumption that $\eta_i^2 \leqslant 4\frac{\sigma^2 B}{n}\alpha$.

3.3.3.3 Proof of Lemma 3.6

Evidently $c_1 \geqslant 1/2$ by symmetry. On the other hand, since $p \leqslant 1/2$, we have by anticoncentration results for Binomial random variables [51, Theorem 6.4] that $c_2 \geqslant p$. Therefore all together, $c \geqslant p^2/32$. Note that by standard lower bounds for the Gaussian tail [35, Theorem 1.2.6], we have

$$p \geqslant \frac{10}{27} \mathrm{e}^{-9/8} \geqslant \frac{3}{25},$$

which provides our claimed bound.

3.4 Deferred results

3.4.1 Results in the Gaussian sequence model

In this section, we collect classical results regarding the nonasymptotic minimax rate of estimation for Gaussian sequence model over the origin-centered ℓ_p balls, $p \in [0, 1]$. All of the results in this section are based on the monograph [62]. We use the following notation to specify the minimax rate of interest,

$$\mathfrak{M}\Big(p, d, R, \varepsilon\Big) \coloneqq \inf_{\widehat{\mu}} \sup_{\substack{\mu^\star \in \mathbf{R}^d \\ \|\mu^\star\|_p \leqslant R}} \mathbf{E}_{y \sim \mathsf{N}(\mu^\star, \varepsilon^2)} \Big[\|\widehat{\mu}(y) - \mu^\star\|_2^2 \Big].$$

As usual, the infimum ranges over measurable estimators from the observation vector $y \in \mathbf{R}^d$ to an estimate $\widehat{\mu}(y) \in \mathbf{R}^d$. Throughout, we use the notation $\tau \coloneqq \frac{\varepsilon}{R}$ for the inverse signal-to-noise ratio.

Proposition 3.2 (Minimax rate of estimation when $0 < p \leqslant 1$)**.** *Fix an integer $d \geqslant 1$. Let $p \in (0, 1]$. If $R, \varepsilon > 0$ satisfy*

$$\frac{1}{d^{2/p}} \leqslant \tau^2 \leqslant \frac{1}{1 + \log d}, \qquad \textit{where} \quad \tau = \frac{\varepsilon}{R},$$

then

$$\frac{7}{2000} R^2 \left(\tau^2 \log(\mathrm{e} d \tau^p)\right)^{1-p/2} \leqslant \mathfrak{M}(p, d, R, \varepsilon) \leqslant 1203\, R^2 \left(\tau^2 \log(\mathrm{e} d \tau^p)\right)^{1-p/2}.$$

The upper and lower bounds are taken from Theorem 11.7 in the monograph [62]. Although the constants are not made explicit in their theorem statement, the upper bound constant is obtained via their Theorem 11.4, setting their parameters as $\zeta = \frac{23}{4}, \gamma = 2\mathrm{e}, \beta = 0$. Similarly, the lower bound constant is implicit in their proof of Theorem 11.7.

We now turn to the minimax rate in the special case that $p = 0$.

Proposition 3.3 (Minimax rate of estimation when $p = 0$)**.** *Suppose that $d \geqslant 1$ and $s \in [d]$. Then for any $\varepsilon > 0$ we have*

$$\frac{3}{500} \varepsilon^2 s \log\Big(\mathrm{e} \frac{d}{s}\Big) \leqslant \mathfrak{M}(p, d, s, \varepsilon) \leqslant 2\, \varepsilon^2 s \log\Big(\mathrm{e} \frac{d}{s}\Big),$$

provided that $p = 0$.

The proof of the above claim is omitted as it is a straightforward combination of the standard minimax rate $\varepsilon^2 k$ for the unconstrained Normal location model in a k-dimensional problem (this provides a useful lower bound when $s \geqslant d/2$ or when $d = 1$) and the result in Proposition 8.20 in the monograph [62].

3.4.2 Details for experiments in Figure 3.1

For each choice of p, we simulate the oracle Lasso and STOLS procedures on instances $(X_n, \theta_n^\star)$ indexed by the sample size $n \in \{1000, 2000, 3000, 5000, 10000, 15000\}$. The matrix $X_n \in \mathbf{R}^{n \times n}$ is block diagonal and given by

$$\frac{1}{\sqrt{n}} X_n = \begin{pmatrix} I_{n/2 \times n/2} & 0 \\ 0 & n^{-1/4} I_{n/2 \times n/2} \end{pmatrix},$$

When $p = 0$, $\theta_n^\star = 2e_{n/2+1}$ and when $p \neq 0$, $\theta_n^\star = e_{n/2+1}$. In the figures, we are plotting the average performance of the oracle Lasso and STOLS procedures, as measured by ℓ_2 error, when applied to the data (X_n, y), where $y \sim \mathsf{N}(X_n \theta_n^\star, I_n)$. The average is taken over 1000 trials for $n < 10,000$. In the case $n \geqslant 10,000$ due to memory constraints we only run 300 trials.

The STOLS procedure is implemented as described in Section 3.1.3. On the other hand, the oracle Lasso procedure is implemented by a slightly more involved procedure. Our goal is to compute

$$\widehat{\theta}_{\widehat{\lambda}}(X, y) \quad \text{where} \quad \widehat{\lambda} \in \arg\min_{\lambda > 0} \|\widehat{\theta}_\lambda(X, y) - \theta^\star\|_2^2,$$

where the Lasso is defined as in display (3.6). To do this, we can use the fact that the Lasso regularization path is piecewise linear. That is, there exist knot points $0 = \lambda_0 < \lambda_1 < \lambda_2 < \cdots < \lambda_m$ such that the knot points $\widehat{\theta}_i \coloneqq \widehat{\theta}_{\lambda_i}(X, y)$ satisfy $\|\theta_i\|_0 > \|\theta_{i+1}\|_0$. Moreover, we have

$$\{\widehat{\theta}_\lambda(X, y) : \lambda \in (\lambda_i, \lambda_{i+1})\} = \{\widehat{\theta}_i + \alpha(\widehat{\theta}_{i+1} - \widehat{\theta}_i) : \alpha \in (0, 1)\}.$$

That is, we can compute the set of Lasso solutions between the knot points by taking all convex combinations of knot points. Therefore the distance between the oracle Lasso solution and the true parameter $\theta^\star$ satisfies,

$$\|\widehat{\theta}_{\widehat{\lambda}}(X, y) - \theta^\star\|_2^2 = \min_i \min_{\alpha \in [0,1]} \|\widehat{\theta}_i + \alpha(\widehat{\theta}_{i+1} - \widehat{\theta}_i) - \theta^\star\|_2^2.$$

We are able to compute the righthand side of the display above by noting that for each i the inner minimization problem is a quadratic function of the univariate parameter α and therefore can be minimized explicitly.

Code: The code has been released at the following public repository,

`https://github.com/reesepathak/lowerlassosim`.

In particular, the repository contains a `Python` program which runs simulations of STOLS and oracle Lasso on the lower bound instance described above for any desired choice of $p \in [0, 1]$.

Chapter 4

Noisy recovery in linear observational models under elliptical constraints

4.1 Introduction

In this chapter, 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

$$y = T_\xi(\theta^\star) + w. \tag{4.1}$$

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 (4.1) captures a broad range of statistical estimation problems.

Example 4.1 (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

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

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 4.2 (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 $f^\star(x) \coloneqq \mathbf{E}[Y \mid X = x]$, 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

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

so that this problem can be written in the form of our general model (4.1). Observe that the randomness in the observation operator T_ξ arises via the randomness in sampling the covariates $\{x_i\}_{i=1}^n$. ♣

Example 4.3 (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 4.2, 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 4.4 (Error-in-variables)**.** Consider the Berkson variant [8, 23] 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

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

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 4.1.2) motivate our study of the operator model (4.1). 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 $\theta^\star$.

4.1.1 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.

4.1.1.1 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 4.3.2, 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 4.1.1.1 requires that w is conditionally centered, and Assumption 4.1.1.1 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 (4.1).

4.1.1.2 Decision-theoretic formulation

In this chapter, 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

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

with radius R, as defined by the matrix K_c.

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

$$\mathfrak{M}(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c) \coloneqq \inf_{\widehat{\theta}} \sup_{\substack{\theta^\star \in \Theta(\varrho, K_c) \\ \nu \in \mathcal{P}(\Sigma_w)}} \mathbf{E}_{(\xi,w) \sim \mathbb{P} \times \nu} \left[\|\widehat{\theta} - \theta^\star\|_{K_e}^2 \right], \tag{4.2}$$

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.

4.1.2 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 4.5 (Covariate shift in linear regression)**.** Recall the set-up for linear regression, as introduced in Example 4.1. 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 [75, 67] 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 $\theta \mapsto \widehat{y} \coloneqq \langle \theta, x \rangle$ 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

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

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 $K_e = \Sigma_P \coloneqq \mathbf{E}_P[x \otimes x]$.

Among the claims in the paper of Mourtada [88] 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

$$\inf_{\widehat{\theta}} \sup_{\theta^\star \in \mathbf{R}^d} \mathbf{E}\left[L_Q(\widehat{\theta}, \theta^\star)\right] = \frac{\sigma^2}{n} \mathbf{E}[\mathbf{Tr}(\Sigma_n^{-1}\Sigma_Q)], \tag{4.3}$$

where Σ_n denotes the sample covariance matrix $(1/n)\sum_{i=1}^n x_i \otimes x_i$, and the expectation is over $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} P$. 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 chapter, 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 4.3.1.3 that these more general results allow us to recover the known relation (4.3) via a simple limiting argument where the constraint radius tends to infinity. ♣

Example 4.6 (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 4.2. 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 [40, 42, 41, 43, 53, 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 (4.1), 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 4.3.2. ♣

Example 4.7 (Covariate shift in nonparametric regression)**.** Combining the scenarios in Examples 4.5 and 4.6, 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 4.6, and the error metric is determined by the new distribution of covariates on which the estimates must be deployed, analogously to Example 4.5. Some recent work has studied general conditions on the pair (P, Q) and the corresponding optimal rates of estimation [69, 46, 94, 77, 103, 121, 106, 47]. 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 4.3.2.3 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 (4.1). 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 chapter 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 4.3, we give concrete consequences of our general results for these types of problems.

4.1.3 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 [54, 115, 58] and references therein for an overview. The recent paper [88] 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 [88, Theorem 1]. In Section 4.3.1.3, 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 [62, Theorem 5.7] for further details.

Finally, we also comment on the similarity of our results to the paper [63]. Specifically, our main results can be compared to their Theorem 2.1. There are a few differences: first, in the paper [63], 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 the results in this chapter. 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 [31] 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 [87, 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 [108, 16, 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 $\theta^\star$ [32].

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 [96, 50, 34, 33, 5, 48, 49]; see also the monograph [62] for a pedagogical overview of this topic. These works extend the classical line of work on estimating a constrained (possibly multivariate) Gaussian mean [24, 11, 83, 9, 81]. We refer the reader to references [82, 38], 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.

4.2 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 (4.2), matching up to a constant pre-factor. These bounds are presented in Section 4.2.1.

4.2.1 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,

$$\begin{aligned}&\Phi(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c) \\ &\quad := \sup_{\Omega} \Big\{ \mathbf{E}\,\mathbf{Tr}\left(K_e^{1/2}(\Omega^{-1} + T_\xi^{\mathsf{T}}\Sigma_w^{-1}T_\xi)^{-1}K_e^{1/2}\right) : \Omega \succ 0,\ \mathbf{Tr}(K_c^{-1/2}\Omega K_c^{-1/2}) \leqslant \varrho^2 \Big\}. \quad (4.4)\end{aligned}$$

Our first main result is a general upper bound.

Theorem 4.1 (General minimax upper bound)**.** *The minimax risk is upper bounded as*

$$\mathfrak{M}(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c) \leqslant \Phi(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c). \quad (4.5)$$

See Section 4.4.1 for the proof.

Our second result is a complementary lower bound.

Theorem 4.2 (Lower bound)**.** *The minimax risk is lower bounded as*

$$\mathfrak{M}(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c) \geqslant \Phi(T, \mathbb{P}, \Sigma_w, \tfrac{\varrho}{2}, K_e, K_c) \geqslant \frac{1}{4}\Phi(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c). \quad (4.6)$$

See Section 4.4.2 for the proof.

Note that the functional on the righthand side of the display (4.6) above matches the quantity appearing in our minimax upper bound (4.5). 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 (4.6) 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

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

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,

$$\mathfrak{M}(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c) \geqslant \mathbf{E}\,\mathbf{Tr}\left(K_e^{1/2}(\frac{1}{c}\Omega^{-1} + T_\xi^{\mathsf{T}}\Sigma_w^{-1}T_\xi)^{-1}K_e^{1/2}\right), \tag{4.7}$$

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 (4.7) can lead to pre-factors that are much closer to 1, and in some cases, converge to one under various scalings. In Section 4.3.1.1, we give one illustration of how the family of bounds (4.7) can be exploited to obtain an improvement of this type.

Form of an optimal procedure Inspecting the proof of Theorem 4.1—specifically, as a consequence of Proposition 4.3—if the supremum on the righthand side of (4.4) is attained at the matrix $\Omega_\star$, then the following estimator, in view of the lower bound (4.6), is near minimax-optimal,

$$\widehat{\theta}(T_\xi, y) \coloneqq \left(\Omega_\star^{-1} + T_\xi^{\mathsf{T}}\Sigma_w^{-1}T_\xi\right)^{-1}T_\xi^{\mathsf{T}}\Sigma_w^{-1}y. \tag{4.8}$$

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

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

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 4.1 and 4.2. If the supremum in (4.4) 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.

4.2.2 Independent and identically distributed regression models

An important application of our general result is for independent and identically distributed (i.i.d.) regression models of the form

$$y_i = \langle \theta^\star, \psi(x_i)\rangle + \sigma z_i, \quad \text{for } i = 1, \ldots, n. \tag{4.9}$$

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

$$\mathfrak{M}_n^{\mathrm{IID}}\Big(\psi, P, \varrho, \sigma^2, K_c, K_e\Big) \coloneqq \inf_{\widehat{\theta}} \sup_{\substack{\theta^\star \in \Theta(\varrho, K_c) \\ \nu \in \mathcal{P}(\sigma^2 I_n)}} \mathbf{E}\left[\left\| \widehat{\theta}(y_1^n, x_1^n) - \theta^\star \right\|_{K_e}^2 \right]. \tag{4.10}$$

Note that this problem can be formulated as an instance of our general operator formulation (4.1) 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:

$$\Sigma_n \coloneqq \frac{1}{n} \sum_{i=1}^n \psi(x_i) \otimes \psi(x_i). \tag{4.11}$$

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,

$$\begin{aligned} &d_n(\psi, P, \varrho, \sigma^2, K_e, K_c) \\ &\qquad \coloneqq \sup_{\Omega} \Big\{ \mathbf{Tr}\, \mathbf{E}_{P^n}\big[K_e^{1/2} (\Sigma_n + \Omega^{-1})^{-1} K_e^{1/2} \big] : \Omega \succ 0, \mathbf{Tr}(K_c^{-1/2} \Omega K_c^{-1/2}) \leqslant \frac{n\varrho^2}{\sigma^2} \Big\}. \end{aligned} \tag{4.12}$$

Then an immediate corollary to Theorems 4.1 and 4.2 is the following pair of inequalities for the IID minimax risk.[1]

Corollary 4.1. *Under the IID regression model* (4.9)*, the minimax rate of estimation as defined in equation* (4.10) *satisfies the following inequalities,*

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

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 $\mathfrak{M}_n^{\mathrm{IID}}$.

[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.

4.2.3 Some properties of the functional appearing in Theorems 4.1 and 4.2

As indicated by Theorem 4.1 and the subsequent discussion, the extremal quantity

$$\sup_{\Omega} \left\{ \mathbf{E}\,\mathbf{Tr}\left(K_e^{1/2}(\Omega^{-1} + T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi)^{-1}K_e^{1/2}\right) : \Omega \succ 0,\ \mathbf{Tr}(K_c^{-1/2}\Omega K_c^{-1/2}) \leqslant \varrho^2 \right\} \tag{4.13}$$

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 (4.8)). Conveniently, it turns out that the maximization problem implied by the display (4.13) is concave.

Proposition 4.1 (Concavity of functional)**.** *The optimization problem*

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

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 Section 4.6.1 for a 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 [15].

The functional (4.13) 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 [10, Exercise 1.5.1] implies that it is always lower bounded by

$$\sup_{\Omega} \left\{ \mathbf{Tr}\left(K_e^{1/2}(\Omega^{-1} + \mathbf{E}\,T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi)^{-1}K_e^{1/2}\right) : \Omega \succ 0,\ \mathbf{Tr}(K_c^{-1/2}\Omega K_c^{-1/2}) \leqslant \varrho^2 \right\} \tag{4.14}$$

Comparing displays (4.13) and (4.14), we have simply moved the expectation over ξ into the inverse. For certain IID regression models, as described in Section 4.2.2, we can give a complementary upper bound. To state our result, we define

$$\begin{aligned} &\overline{d}_n(P, \varrho, \sigma^2, K_e, K_c) \\ &\qquad \coloneqq \sup_{\Omega}\left\{ \mathbf{Tr}\left(K_e^{1/2}(\mathbf{E}_{P^n}\,\Sigma_n + \Omega^{-1})^{-1}K_e^{1/2}\right) : \Omega \succ 0, \mathbf{Tr}(K_c^{-1/2}\Omega K_c^{-1/2}) \leqslant \frac{n\varrho^2}{\sigma^2}\right\}. \end{aligned}$$

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

Proposition 4.2 (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 Section 4.6.2.

4.2.4 Asymptotics for a diverging radius

In this section, we develop an asymptotic limit relation for the minimax risk (4.2) 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 4.2 can actually be made quite close to 1 for large radii.

Corollary 4.2. *Suppose that $T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi$ is $\mathbb{P}$-almost surely nonsingular. Then the minimax risk* (4.2) *satisfies*

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

See Section 4.6.3 for a proof of this claim.

An immediate consequence is that for IID regression settings as in Section 4.2.2, we have the following limit relation.

Corollary 4.3. *Suppose that that the empirical covariance matrix Σ_n from equation* (4.11) *is P^n-almost surely invertible. Then, the minimax risk for an IID observation model* (4.9) *satisfies the relation*

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

4.3 Consequences of main results

In this section, we demonstrate consequences of our main results for a variety of estimation problems. In Section 4.3.1, we develop consequences of our main results for problems where the underlying parameter to be estimated is finite-dimensional. In Section 4.3.2, 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.

4.3.1 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 4.3.1.1, 4.3.1.2, and 4.3.1.3 concern IID regression settings of the form described in Section 4.2.2. In Section 4.3.1.4, we consider a non-IID regression setting.

4.3.1.1 Linear regression with Gaussian covariates

As in the prior work [31], consider a random design IID regression setting of the form presented in the display (4.9), but with Gaussian data. Formally, we assume Gaussian noise, so that $z_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,1)$, and Gaussian covariates, so that $x_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, I_d)$ and $\psi(x) = x$. Here x and z are assumed independent. Then we define

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

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 [31, Corollary 1] on the sphere $\{\|\theta\|_2 = \varrho\}$.

Dicker [31, 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

$$\lim_{d/n \to \gamma} \left| r(n,d,\varrho,\sigma) - \frac{\sigma^2}{n} d_{\text{Dicker}}(n,d,\varrho,\sigma) \right| = 0, \tag{4.15}$$

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 (4.15).

Corollary 4.4. *For linear regression over the ϱ-radius Euclidean sphere with Gaussian covariates, the minimax risk satisfies the sandwich relation*

$$\begin{aligned} c_d \frac{\sigma^2}{n} d_{\text{Dicker}}(n,d,\varrho,\sigma) \leqslant \frac{\sigma^2}{n} d_{\text{Dicker}}(n,d,\sqrt{c_d}\varrho,\sigma) \\ \leqslant r(n,d,\varrho,\sigma) \leqslant \frac{\sigma^2}{n} d_{\text{Dicker}}(n,d,\varrho,\sigma), \end{aligned} \tag{4.16a}$$

where

$$c_d \coloneqq \begin{cases} (1 - \frac{1}{2d-1})(1 - \exp(-\frac{d^{3/2}}{4})) & d \geqslant 2 \\ 1/4 & d = 1 \end{cases}. \tag{4.16b}$$

Note that since $c_d = (1 - O(1/d))$ as $d \to \infty$, the inequalities (4.16a) allow us to immediately recover Dicker's result. It should be emphasized, however, that Corollary 4.4, 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 Section 4.6.4.1, our functional for this problem satisfies

$$d_n(\mathsf{N}(0, I_d), \varrho, \sigma^2, I_d, I_d) = d_{\text{Dicker}}(n, d, \varrho, \sigma). \tag{4.17a}$$

Hence, our Corollary 4.1 implies the following characterization of the minimax risk,[2]

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

To establish our sharper result (4.16a), we leverage the stronger lower bound (4.7). The details of this calculation are presented in Section 4.6.4.2. Note that in Section 4.5.1.1, we simulate this problem and find that as suggested by Corollary 4.4, that, indeed, the gap between our upper and lower bounds is tiny, even for problems with small dimension (see Figure 4.1).

4.3.1.2 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 (4.9), 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 (4.11) 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 4.1 demonstrates that

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

[2] Although Corollary 4.1 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.

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,

$$\frac{\sigma^2}{n} \operatorname{\mathbf{Tr}} \mathbf{E}_{P^n}\big[(\Sigma_n + \tfrac{\sigma^2}{\varrho^2}\tfrac{d}{n} I_d)^{-1}\big] \\ \asymp \underbrace{\mathbf{E}\Big[\sum_{i=1}^{d} \tfrac{\sigma^2}{n} \tfrac{1}{\lambda_i(\Sigma_n)} \mathbf{1}\{\lambda_i(\Sigma_n) \geqslant \tfrac{\sigma^2}{n}\tfrac{d}{\varrho^2}\}\Big]}_{\substack{\text{Estimation error from} \\ \text{large eigenvalues of } \Sigma_n}} + \underbrace{\mathbf{E}\Big[\sum_{i=1}^{d} \tfrac{\varrho^2}{d} \mathbf{1}\{\lambda_i(\Sigma_n) < \tfrac{\sigma^2}{n}\tfrac{d}{\varrho^2}\}\Big]}_{\substack{\text{Approximation error due} \\ \text{to small eigenvalues of } \Sigma_n}}. \quad (4.18)$$

The lower bound (4.18) is sharp in certain cases. For instance, when $x_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, I_d)$ 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 4.3.1.1.

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 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 $(0, \frac{\sigma^2}{n}\frac{d}{\varrho^2})$.

4.3.1.3 Linear regression with an unrestricted parameter space

In recent work, Mourtada [88] 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 (4.9) 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 (4.18) from Section 4.3.1.2, 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

$$\inf_{\widehat{\theta}} \sup_{\substack{\theta^\star \in \mathbf{R}^d \\ \nu \in \mathcal{P}(\sigma^2 I_n)}} \mathbf{E}\Big[\big\|\widehat{\theta} - \theta^\star\big\|_{\Sigma_P}^2\Big] = \frac{\sigma^2}{n} \mathbf{E}\big[\operatorname{\mathbf{Tr}}(\Sigma_n^{-1}\Sigma_P)\big], \quad (4.19)$$

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

$$\inf_{\widehat{\theta}} \sup_{\substack{\theta^\star \in \mathbf{R}^d \\ \nu \in \mathcal{P}(\sigma^2 I_n)}} \mathbf{E}\left[\left\| \widehat{\theta} - \theta^\star \right\|_{\Sigma_P}^2 \right] \geqslant \sup_{\varrho > 0} \left\{ \inf_{\widehat{\theta}} \sup_{\substack{\|\theta^\star\|_2 \leqslant \varrho \\ \nu \in \mathcal{P}(\sigma^2 I_n)}} \mathbf{E}\left[\left\| \widehat{\theta} - \theta^\star \right\|_{\Sigma_P}^2 \right] \right\} \tag{4.20a}$$

$$= \frac{\sigma^2}{n} \lim_{\varrho \to \infty} d_n(P, \varrho, \sigma^2, \Sigma_P, I_d). \tag{4.20b}$$

In order to obtain the relation (4.20b), 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 4.3. A short calculation, which we defer to Section 4.6.4.3, demonstrates that

$$\lim_{\varrho \to \infty} d_n(P, \varrho, \sigma^2, \Sigma_P, I_d) = \mathbf{E}\left[\mathbf{Tr}(\Sigma_n^{-1} \Sigma_P) \right]. \tag{4.21}$$

Thus, after combining displays (4.20b) and (4.21), we have obtained the lower bound in Mourtada's result (4.19). One consequence of this argument is that the inequality (4.20a) is, as may be expected, an equality. That is, we have

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

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.

4.3.1.4 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

$$x_t = \sqrt{r_t}\, x_{t-1} + \sqrt{1 - r_t}\, z_t \quad \text{for } t = 1, \ldots, T, \tag{4.22}$$

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 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

$$y_t = x_t \theta^\star + \sigma w_t, \qquad \text{for } t = 1, \ldots, T, \tag{4.23}$$

[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 (4.8), 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 $\varrho \to \infty$.

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 (4.22). It has entries

$$M_{ss'} = \sum_{t=s \vee s'}^{T} \sqrt{c_{st} c_{s't}}, \quad \text{where} \quad c_{st} \coloneqq (1 - r_s) \prod_{\tau=s+1}^{t} r_\tau. \tag{4.24}$$

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 AR(1) process (*i.e.*, the Kac–Murdock–Szegö matrix [64]); it is easily checked to be symmetric positive definite.

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

Corollary 4.5. *The minimax risk for the Markovian observation model described above satisfies*

$$\inf_{\widehat{\theta}} \sup_{|\theta^\star| \leqslant \varrho} \mathbf{E}\left[(\widehat{\theta} - \theta^\star)^2\right] \asymp \Phi_T(\varrho, \sigma) \coloneqq \mathbf{E}\left[\left(\frac{1}{\varrho^2} + \frac{z^\mathsf{T} M z}{\sigma^2}\right)^{-1}\right]. \tag{4.25}$$

See Section 4.6.4.4 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 (4.22) 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 4.5 gives one example of how our general results can even establish sharp rates for regression problems of the form described in Section 4.2.2, 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.

4.3.2 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 4.3.2.1. We then turn, in Sections 4.3.2.2 and 4.3.2.3, 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.

4.3.2.1 Gaussian sequence model

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

$$y_i = \theta_i^\star + \varepsilon_i z_i, \qquad \text{for } i = 1, 2, 3, \ldots \tag{4.26}$$

Here the variables $\{z_i\}$ are a sequence of IID standard Gaussian variates, and $\varepsilon \coloneqq \{\varepsilon_i\}$ 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 $a \coloneqq \{a_i\}$ and radius $C > 0$ such that

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

The minimax risk for this problem is then defined by

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

where the expectation is over y according to the observation model (4.26).

Let us define a k-dimensional truncation,

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

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

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

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 (4.26). Then, this sequence of minimax risks satisfies the limit relation

$$\lim_{k\to\infty} \mathfrak{M}_k\Big(\varepsilon, a, C\Big) = \mathfrak{M}\Big(\{\varepsilon_j\}_{j=1}^{\infty}, \Theta(a, C)\Big). \tag{4.27}$$

See Section 4.6.5.1 for justification. The k-dimensional problem can be seen as a special case of our operator model (4.1), with parameters $T^{(k)}, \Sigma_w^{(k)}, {K_e}^{(k)}, \varrho^{(k)}, {K_c}^{(k)}$ defined as,

$$\begin{aligned} T^{(k)}(\xi) \equiv I_k, \qquad \Sigma_w^{(k)} = \mathbf{diag}(\varepsilon_1^2, \ldots, \varepsilon_k^2), \qquad K_e^{(k)} = I_k, \\ K_c^{(k)} = \mathbf{diag}\left(\frac{1}{a_1^2}, \ldots, \frac{1}{a_k^2}\right), \quad \text{and,} \quad \varrho^{(k)} = C. \end{aligned} \tag{4.28}$$

Computing the functional (4.13) for the k-dimensional problem, we find it is equal to

$$R_k^\star\Big(\varepsilon, a, C\Big) \coloneqq \sup_{\tau_1,\ldots,\tau_k} \Big\{ \sum_{j=1}^{k} \frac{\tau_j^2 \varepsilon_j^2}{\tau_j^2 + \varepsilon_j^2} : \sum_{j=1}^{k} \tau_j^2 a_j^2 \leqslant C^2 \Big\}. \tag{4.29}$$

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

$$R^\star(\varepsilon, a, C) \coloneqq \sup_{\tau=\{\tau_j\}_{j=1}^{\infty}} \Big\{ \sum_{j=1}^{\infty} \frac{\tau_j^2 \varepsilon_j^2}{\tau_j^2 + \varepsilon_j^2} : \sum_{j=1}^{\infty} \tau_j^2 a_j^2 \leqslant C^2 \Big\}. \tag{4.30}$$

Then our main results, Theorems 4.1 and 4.2 imply the sandwich relation

$$\frac{1}{4} R^\star(\varepsilon, a, C) \leqslant \mathfrak{M}\Big(\varepsilon, a, C\Big) \leqslant R^\star(\varepsilon, a, C). \tag{4.31}$$

See Section 4.6.5.2 for verification of this relation as a consequence of our results. Note that this recovers a well-known result for the Gaussian sequence model [115, 62]. Some previous work [34] 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.

4.3.2.2 Nonparametric regression over reproducing kernel Hilbert spaces (RKHSs)

In this section, we consider a nonparametric regression model of the form

$$y_i = f^\star(x_i) + w_i, \quad \text{for } i = 1, \ldots, n. \tag{4.32}$$

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 4.1.1.1 and 4.1.1.1 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 $f^\star$.

Relating the RKHS observation model (4.32) **with the model** (4.9) We now show that the observation model when $f^\star \in \mathcal{H}$ is an infinite-dimensional version of the observation model (4.9), 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

$$\mathcal{H} = \Big\{ f \coloneqq \sum_{j=1}^{\infty} \theta_j \sqrt{\mu_j} \phi_j \mid \sum_{j=1}^{\infty} \theta_j^2 < \infty \Big\}. \tag{4.34}$$

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 (4.32) in the form

$$y_i = \big\langle \theta^\star, \Phi(x_i) \big\rangle + w_i, \quad \text{for } \; i = 1, \ldots, n. \tag{4.35}$$

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

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

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

[5]The elliptical representation (4.34) 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 [109, see Lemma 2.3] that in this case, $\mathcal{H}$ compactly embeds into $L^2(\nu)$ and that there is a series expansion

$$k(x, x') = \sum_{j=1}^{\infty} \mu_j \phi_j(x) \phi_j(x'), \quad \text{for any } x, x' \in \mathcal{X}. \tag{4.33}$$

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 (4.34) of $\mathcal{H}$ follows from the series expansion of the underlying kernel (4.33); see Cucker and Smale [29] for details.

With these definitions, note that the inner product in equation (4.35) is taken in the sequence space $\ell^2(\mathbf{N})$. From the display (4.35), we see that the RKHS observation model (4.32) is in fact an infinite-dimensional version of the observation model (4.9). 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

$$\mathfrak{M}_n(\varrho, \sigma^2, P) \coloneqq \inf_{\widehat{f}} \sup_{\substack{f^\star \in \mathsf{B}_{\mathcal{H}}(\varrho) \\ \nu \in \mathcal{P}(\sigma^2 I_n)}} \mathbf{E}\Big[\big\| \widehat{f} - f^\star \big\|_{L^2(\nu)}^2 \Big]. \tag{4.36}$$

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 4.3.2.3 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

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

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

$$\mathfrak{M}_n^{(k)}(\varrho, \sigma^2, P) \coloneqq \inf_{\widehat{f}} \sup_{\substack{f^\star \in \mathsf{B}_{\mathcal{H}}(\varrho) \cap \mathcal{H}_k \\ \nu \in \mathcal{P}(\sigma^2 I_n)}} \mathbf{E}\Big[\big\| \widehat{f} - f^\star \big\|_{L^2(\nu)}^2 \Big]. \tag{4.37}$$

In analogy to the limit relation (4.27) for the Gaussian sequence model, we can show that

$$\lim_{k \to \infty} \mathfrak{M}_n^{(k)}(\varrho, \sigma^2, P) = \mathfrak{M}_n(\varrho, \sigma^2, P). \tag{4.38}$$

See Section 4.6.5.3 for a proof of this relation. The k-dimensional problem associated with the risk (4.37) can be seen, using the representation (4.35), as a special case of our IID observation model (4.9), with parameters, P, ϱ, σ and

$$\psi(x) = \Phi_k(x) \coloneqq \big(\sqrt{\mu_j} \phi_j(x) \big)_{j=1}^k, \quad K_e = M_k \coloneqq \mathbf{diag}(\mu_1, \ldots, \mu_k), \quad \text{and} \quad K_c = I_k. \tag{4.39}$$

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

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

Then the using (4.39), we see that the functional (4.12) for the k-dimensional problem is equal to

$$d_n^{(k)} \coloneqq \sup_{\Omega \succ 0} \Big\{ \mathbf{Tr}\, \mathbf{E}_{P^n}\big[M_k^{1/2} (\Sigma_n^{(k)} + \Omega^{-1})^{-1} M_k^{1/2} \big] : \mathbf{Tr}(\Omega) \leqslant \frac{n\varrho^2}{\sigma^2} \Big\} \tag{4.40}$$

Characterizations of RKHS minimax risks of estimation We now state the consequence of our results for the rate of estimation (4.36).

Corollary 4.6. *Define* $d_n^\star = \limsup_{k\to\infty} d_n^{(k)}$, *where the sequence* $\{d_n^{(k)}\}_{k\geqslant 1}$ *is defined in display* (4.40). *Then the RKHS minimax risk satisfies satisfies the inequalities,*

$$\frac{1}{4}\frac{\sigma^2}{n} d_n^\star \leqslant \mathfrak{M}_n(\varrho, \sigma^2, P) \leqslant \frac{\sigma^2}{n} d_n^\star. \tag{4.41}$$

Note that this result is an immediate consequence of Theorems 4.1 and 4.2, together with the limit relation (4.38).

We comment that Corollary 4.6 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

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

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

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

Let us now further simplify the characterization (4.41) 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 4.7. *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*

$$\mathfrak{M}_n(\varrho,\sigma^2,P) \asymp \frac{\sigma^2}{n} k_n, \tag{4.42}$$

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

See Section 4.6.5.4 for a proof of this claim.

We note that Corollaries 4.6 and 4.7 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 4.6 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 (4.42)—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 [123]. To emphasize this point, we now provide one concrete example of an irregular kernel for which Corollary 4.7 provides, to our knowledge, a new result.

Example 4.8 (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 [123], 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 4.7—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

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

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 4.7 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 4.7.

Example 4.9 (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 4.7, we find $k_n \asymp (\frac{\sigma^2}{\varrho^2 n})^{-\frac{d}{2\beta+d}}$, and consequently

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

provided that $\sigma^2 \gtrsim \varrho^2$. The above relation recovers a classical result [60, 110]. ♣

4.3.2.3 Kernel regression under covariate shift

We now discuss one important case in which we have $P \neq \nu$ in the RKHS model (4.32). In the setting of covariate shift, the model (4.32) 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 4.3.2, we find that

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathsf{B}_{\mathcal{H}}(\varrho)} \mathbf{E}\Big[\big\|\widehat{f} - f^\star\big\|_{L^2(Q)}^2\Big] \asymp \frac{\sigma^2}{n} \limsup_{k \to \infty} d_n^{(k)}, \tag{4.43}$$

where the quantity $d_n^{(k)}$ is defined as in display (4.40). Above, the expectation on the lefthand side is over the noise and the covariates drawn from P as described by the model (4.32).

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. [77], 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 [77] 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

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

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 (4.43).

To give one simple and concrete illustration of this, we will show how one can derive Theorem 2 in the paper [77]. By Jensen's inequality, we have

$$d_n^{(k)} \geqslant \sup_{\Omega \succ 0} \Big\{ \mathbf{Tr}(\mathbf{E}_{P^n} M_k^{-1/2} \Sigma_n^{(k)} M_k^{-1/2} + \Omega^{-1})^{-1} : \mathbf{Tr}(M_k^{-1}\Omega) \leqslant \frac{n\varrho^2}{\sigma^2} \Big\}. \tag{4.44}$$

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

$$\mathbf{E}_{P^n} M_k^{-1/2} \Sigma_n^{(k)} M_k^{-1/2} \succcurlyeq \frac{1}{B} I_k. \tag{4.45}$$

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

$$x_0 \in \bigcap_{j \geqslant 1} \phi_j^{-1}(\{0\}), \tag{4.46}$$

then the bound (4.45) 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 (4.46) 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 (4.43) and (4.44) with the choice of $P = P_{x_0}$ given above, we have

$$\begin{aligned} \sup_{P : d_\infty(P,Q) \leqslant B} \inf_{\widehat{f}} \sup_{f^\star \in \mathsf{B}_{\mathcal{H}}(\varrho)} \mathbf{E}\Big[\big\| \widehat{f} - f^\star \big\|_{L^2(Q)}^2 \Big] &\gtrsim \frac{\sigma^2}{n} \sup_{\omega > 0} \Big\{ \sum_{j=1}^\infty \frac{B\omega_j}{\omega_j + B} : \sum_{j=1}^\infty \frac{\omega_j}{\mu_j} = \frac{n\varrho^2}{\sigma^2} \Big\} \\ &\asymp \varrho^2 \sup_{\lambda} \Big\{ \sum_{j=1}^\infty \frac{\sigma^2 B}{n\varrho^2} \wedge \lambda_j \mu_j : \lambda_j \geqslant 0, \ \sum_{j=1}^\infty \lambda_j = 1 \Big\}. \end{aligned} \tag{4.47}$$

Suppose, following the paper [77], 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

$$\sup_{\delta > 0} \frac{\sum_{j > d(\delta)} \mu_j}{\delta^2 d(\delta)} \leqslant c, \quad \text{where} \quad d(\delta) \coloneqq \inf\{j \geqslant 1 : \mu_j \leqslant \delta^2\}. \tag{4.48}$$

Under this condition, we can further lower bound (4.47), up to universal constants, by

$$\varrho^2 \inf_{\delta > 0} \Big\{ \delta^2 + \frac{\sigma^2 B}{\varrho^2 n} d(\delta) \Big\}. \tag{4.49}$$

The details of this calculation can be found in Section 4.6.5.6. Note that by establishing the lower bound (4.49), we have recovered Theorem 2 from the paper [77]. We remark that—as seen from the steps taken to arrive at this lower bound—our more general determination of the minimax rate (4.43) 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 [77], requires fewer regularity assumptions on the underlying kernel and its diagonalization in the target Hilbert space $L^2(Q)$. In fact, as demonstrated in Section 4.6.5.6, the regularity condition (4.48) is *not* necessary for us to establish the lower bound (4.49).

4.4 Proofs of Theorems 4.1 and 4.2

In this section, we present the proofs of our main results. In Section 4.4.1, we provide the proof of our minimax upper bound (cf. Theorem 4.1). In Section 4.4.2, we provide the proof of our minimax lower bound. Some calculations and routine verifications are deferred to Section 4.6.

4.4.1 Proof of Theorem 4.1

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

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

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

$$\mathfrak{M}(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c) = \inf_{\widehat{\theta}} \sup_{\theta^\star \in \Theta(\varrho, K_c)} r(\widehat{\theta}, \theta^\star). \tag{4.50}$$

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

$$\widehat{\theta}_C(T_\xi, y) \coloneqq C(T_\xi) T_\xi^{\mathsf{T}} \Sigma_w^{-1} y, \tag{4.51}$$

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 (4.51).

(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 Section 4.6.6.

4.4.1.1 Supremum risk of estimator $\widehat{\theta}_C$

Starting with the definition (4.51), for any matrix C, we have

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

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

$$\begin{aligned} r(\widehat{\theta}_C, \theta^\star) &\coloneqq \sup_{\nu\in\mathcal{P}(\Sigma_w)} \mathbf{E}\left[\|\widehat{\theta}_C(X,y) - \theta^\star\|_{K_e}^2\right] \\ &= \mathbf{Tr}\left\{ {K_e}^{1/2}\, \mathbf{E}_\xi \left[(C(T_\xi)T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi - I_d)\theta^\star \otimes \theta^\star (C(T_\xi)T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi - I_d)^\mathsf{T} \right.\right. \\ &\qquad \left.\left. + C(T_\xi)T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi C(T_\xi)^\mathsf{T}\right] {K_e}^{1/2}\right\}. \end{aligned} \tag{4.52}$$

The equality above uses the property 4.1.1.1 of distributions $\nu \in \mathcal{P}(\Sigma_w)$; note that it is achieved by the Gaussian distribution $\nu = \mathsf{N}(0, \Sigma_w)$.

4.4.1.2 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

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

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

$$\begin{aligned} r(\widehat{\theta}_C, \Omega) \coloneqq \mathbf{Tr}\left\{ {K_e}^{1/2}\, \mathbf{E}_\xi \left[(C(T_\xi)T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi - I_d)\Omega(C(T_\xi)T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi - I_d)^\mathsf{T} \right.\right. \\ \left.\left. + C(T_\xi)T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi C(T_\xi)^\mathsf{T}\right] {K_e}^{1/2}\right\}. \end{aligned} \tag{4.53}$$

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 4.1. *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 (4.51).

See Section 4.6.6.1 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 4.2. *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 Section 4.6.6.2 for the proof of this claim.

Our next claim determines the minimizer of $r(\cdot, \Omega)$ over estimators $\widehat{\theta}_C$ of the form (4.51), provided that Ω is strictly positive definite.

Proposition 4.3. *Let Ω be a symmetric positive definite matrix. Then*

$$\inf_C r(\widehat{\theta}_C, \Omega) = \mathbf{Tr}\left\{ K_e^{1/2} \, \mathbf{E}_\xi (\Omega^{-1} + T_\xi^{\mathsf{T}} \Sigma_w^{-1} T_\xi)^{-1} K_e^{1/2} \right\} \tag{4.54}$$

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

See Section 4.6.6.3 for the proof.

4.4.1.3 Proof of Theorem 4.1

We now piece together the previous lemmas to establish our main upper bound, as claimed in Theorem 4.1. In view of the relation (4.50) and the bound (4.52), we find that

$$\begin{aligned}
\mathfrak{M}(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c) &\leqslant \inf_C \sup_{\theta^\star \in \Theta(\varrho, K_c)} r(\widehat{\theta}_C, \theta^\star) && \text{(4.55a)} \\
&= \inf_C \sup_{\Omega \in \mathcal{K}(\varrho, K_c)} r(\widehat{\theta}_C, \Omega) && \text{(4.55b)} \\
&= \sup_{\Omega \in \mathcal{K}(\varrho, K_c)} \inf_C r(\widehat{\theta}_C, \Omega) && \text{(4.55c)} \\
&= \sup_{\substack{\Omega > 0 \\ \mathbf{Tr}(K_c^{-1}\Omega) \leqslant \varrho^2}} \mathbf{E}\, \mathbf{Tr}\left(K_e^{1/2} (\Omega^{-1} + T_\xi^{\mathsf{T}} \Sigma_w^{-1} T_\xi)^{-1} K_e^{1/2} \right). && \text{(4.55d)}
\end{aligned}$$

To clarify, in the first display (4.55a) and below, the infimum over C denotes an infimum over all $\mathbf{R}^{d\times d}$-valued measurable functions of T_ξ. In display (4.55b), we have applied Lemma 4.1. Relation (4.55c) follows from the generalized Ky Fan min-max theorem [14, Theorem A] together with Lemma 4.2. Note that the set $\mathcal{K}(\varrho, K_c)$ is evidently a compact convex subset of $\mathbf{R}^{d\times d}$. The final equality (4.55d) is essentially an application of Proposition 4.3; see Section 4.6.6.4 for the details of this verification.

4.4.2 Proof of lower bound, Theorem 4.2

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

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

Evidently, the Gaussian minimax risk lower bounds the general minimax risk, so that we have $\mathfrak{M}^{\mathrm{G}} \leqslant \mathfrak{M}$. In Section 4.4.2.1, 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 4.4 in Section 4.4.2.2. This result is the bulk of the proof of the lower bound, and it quickly allows us to establish our main result, Theorem 4.2. In Section 4.4.2.3, we then complete the proof of Proposition 4.4.

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

$$\Upsilon = VV^{\mathsf{T}}\eta^\star + V\Lambda^{-1/2}z, \quad \text{where} \quad z \sim \mathsf{N}(0, I_r). \tag{4.56}$$

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 $\omega \coloneqq (r, V, \Lambda)$ is distributed according to $\mathbb{Q}$, we denote the minimax risk for this observation model as

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

Above, the expectation indexed by (ω, Υ) is over $\omega \sim \mathbb{Q}$ and Υ as in (4.56). 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 4.3. *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 Section 4.6.7.1 for a proof of this claim.

4.4.2.2 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 4.4. *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*

$$\mathfrak{M}^{\mathrm{G}}_{\mathrm{red}}(\mathbb{Q}, K) \geqslant \mathbf{E}\, \mathbf{Tr}\left((\tfrac{1}{c(\tau, \Pi)} \Pi^{-1} + V \Lambda V^{\mathsf{T}})^{-1} \right), \tag{4.57}$$

where the constant $c(\tau, \Pi)$ is defined in Lemma 4.6. Moreover, we have the lower bounds

$$\begin{aligned} &\mathfrak{M}^{\mathrm{G}}_{\mathrm{red}}(\mathbb{Q}, K) \\ &\quad \geqslant \sup_{\Pi} \left\{ \mathbf{E}\, \mathbf{Tr}\left((\Pi^{-1} + V \Lambda V^{\mathsf{T}})^{-1} \right) : \Pi \succ 0,\ \mathbf{Tr}(K^{-1/2} \Pi K^{-1/2}) \leqslant 1/4 \right\} \quad (4.58a) \\ &\quad \geqslant \frac{1}{4} \sup_{\Pi} \left\{ \mathbf{E}\, \mathbf{Tr}\left((\Pi^{-1} + V \Lambda V^{\mathsf{T}})^{-1} \right) : \Pi \succ 0,\ \mathbf{Tr}(K^{-1/2} \Pi K^{-1/2}) \leqslant 1 \right\}. \quad (4.58b) \end{aligned}$$

Proof of Theorem 4.2 We take the claim of Proposition 4.4 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 4.3 combined with the stronger lower bound (4.58a) in Proposition 4.4 gives us

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

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

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

which furnishes the first inequality in Theorem 4.2. With similar manipulations to the weaker lower bound (4.58b) in Proposition (4.4), or by arguing directly from the display above, the second inequality in Theorem 4.2 follows. In order to establish the more detailed lower bound (4.7), we repeat the argument above but use (4.57).

4.4.2.3 Proof of Proposition 4.4

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 Section 4.6.7.

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

$$\mathfrak{M}_{\mathrm{red}}^{\mathrm{G}}(\mathbb{Q}, K) \;=\; \inf_{\widehat{\eta}} \sup_{\eta \in \Theta(K)} \mathbf{E}_{(\omega,\Upsilon)} \left[\|\widehat{\eta}(\omega, \Upsilon) - \eta\|_2^2 \right] \;\geqslant\; \inf_{\widehat{\eta}} \mathbf{E}_{\eta\sim\pi} \mathbf{E}_{(\omega,\Upsilon)} \left[\|\widehat{\eta} - \eta\|_2^2 \right] \;=:\; B(\pi). \tag{4.59}$$

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

$$B(\pi) = \inf_{\widehat{\eta}} \mathbf{E}_{\eta\sim\pi} \mathbf{E}_{(\omega,\Upsilon)\sim\mathcal{D}_\eta} \left[\left\| \widehat{\eta} - \eta \right\|_2^2 \right] \geqslant \mathbf{E}_{\omega\sim\mathbb{Q}} \left[\inf_{\widehat{\eta}_\omega} \mathbf{E}_{\eta\sim\pi} \mathbf{E}_{\Upsilon} \left\| \widehat{\eta}_\omega(\Upsilon) - \eta \right\|_2^2 \right]. \tag{4.60}$$

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 $\widehat{\eta}_\omega(\Upsilon) \coloneqq \widehat{\eta}(\omega, \Upsilon)$ 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 $\omega = \omega_0$,

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

Using this definition, along with the two inequalities (4.59) and (4.60), we have demonstrated

$$\mathfrak{M}\mathfrak{l}_{\mathrm{red}}^{\mathrm{G}}(\mathbb{Q}, K) \geqslant \mathbf{E}_{\omega \sim \mathbb{Q}} \left[B(\pi \mid \omega) \right], \qquad \text{for any prior } \pi \text{ on } \Theta(K). \tag{4.61}$$

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 $z \sim \mathsf{N}(0, I_d)$

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

where the last equality holds in distribution. Define $\Sigma_\lambda \coloneqq V \Lambda^{-1} V^{\mathsf{T}} + \lambda V_\perp$; 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,

$$B_\lambda(\pi \mid \omega) \coloneqq \inf_{\widehat{\eta}} \left\{ B_\lambda(\widehat{\eta}, \pi \mid \omega) \coloneqq \mathbf{E} \left[\|\widehat{\eta}(\Upsilon_\lambda) - \eta\|_2^2 \right] \right\}.$$

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

Lemma 4.4. *For any ω and $\lambda > 0$, we have*

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

See Section 4.6.7.2 for a proof of this claim.

[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 $\widehat{\eta}(\omega, \Upsilon) \coloneqq \widehat{\eta}_\omega(\Upsilon)$ is measurable.

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) [116, 101, 17].

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

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

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

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

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 4.5. *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*

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

See Section 4.6.7.3 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

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

Recall also that the Fisher information associated with a Gaussian distribution $\mathsf{N}(\mu, \Pi)$ for nonsingular Π is given by Π^{-1} [72, Example 6.3]. Therefore, applying well-known results for the Fisher information [125, eqn. (8) and Corollary 1]

$$\mathcal{I}(\Upsilon_\lambda) \preccurlyeq (X_\lambda \mathcal{I}(\pi)^{-1} X_\lambda + \Sigma_\lambda)^{-1}. \tag{4.62}$$

Next, we select a prior distribution and calculate the Fisher information $\mathcal{I}(\Upsilon_\lambda)$ 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

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

In other words, $\pi_{\tau,\Pi}$ denotes the probability measure $\mathsf{N}(0, \tau^2\Pi)$ 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 4.6. *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 Section 4.6.7.4 for the proof of this claim.

Step 5: Putting the pieces together Combining Lemmas 4.4 and 4.5 along with the inequality (4.62) and Lemma 4.6, 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

$$B(\pi \mid \omega) \geqslant \lim_{\lambda\to 0^+} \mathbf{Tr}\left((\tfrac{1}{c(\tau,\Pi)}\Pi^{-1} + V\Lambda V^{\mathsf{T}} + \lambda V_\perp)^{-1}\right) = \mathbf{Tr}\left((\tfrac{1}{c(\tau,\Pi)}\Pi^{-1} + V\Lambda V^{\mathsf{T}})^{-1}\right). \quad (4.63)$$

Taking the expectation over ω, and applying our minimax lower bound (4.61), we have established lower bound (4.57). 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 (4.63) with our minimax lower bound (4.61), we obtain

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

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

Lemma 4.7. *The constant $c_\ell(K)$ is lower bounded, for any symmetric positive definite K, as*

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

See Section 4.6.7.5 for a proof of this claim.

4.5 Discussion

In this work, we determined the minimax risk of estimation for observation models of the form (4.1), 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 this chapter by presenting some simulation results; see Section 4.5.1

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, 84, 59, 71, 92], 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 [76] for a overview of work in this direction.

4.5.1 Some illustrative simulations

We conclude this chapter 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 4.5.1.1, 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 4.5.1.2, 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.

4.5.1.1 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 (4.9) 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

$$P_\lambda \coloneqq \lambda\delta_0 + (1-\lambda)\mathsf{N}\left(0, \frac{1}{1-\lambda} I_d\right). \tag{4.64}$$

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

$$\mathbf{E}_{P_\lambda}[x] = 0 \quad \text{and} \quad \mathrm{Cov}_{P_\lambda}(x) = \mathbf{E}_{P_\lambda}[x \otimes x] = I_d, \quad \text{for all } \lambda \in [0, 1]. \tag{4.65}$$

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 4.3.1.1, in this case, the minimax risk is upper and lower bounded as

$$\frac{\sigma^2}{n}\mathbf{E}_{P_\lambda^n}[\mathbf{Tr}((\Sigma_n + \tfrac{c_d\sigma^2 d}{n\varrho^2} I_d)^{-1})] \leqslant \mathfrak{M}_n^{\mathrm{IID}}\Big(P_\lambda, \varrho, \sigma^2, I_d, I_d\Big) \leqslant \frac{\sigma^2}{n}\mathbf{E}_{P_\lambda^n}[\mathbf{Tr}((\Sigma_n + \tfrac{\sigma^2 d}{n\varrho^2} I_d)^{-1})]. \tag{4.66}$$

Above, the lower bound constant c_d is defined in display (4.16b).

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 (4.66)—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 (4.66), by introducing the following quantities which only depend on the parameters τ, γ and the sample size n and the mixture parameter λ,

$$\begin{aligned}
\mathfrak{m}_n(\lambda, \tau, \gamma) &\coloneqq \frac{\mathfrak{M}_n^{\mathrm{IID}}\Big(P_\lambda, \tau\sigma, \sigma^2, I_{\lceil \gamma n \rceil}, I_{\lceil \gamma n \rceil}\Big)}{\tau^2\sigma^2}, \\
u_n(\lambda, \tau, \gamma) &\coloneqq \frac{1}{\tau^2 n} \mathbf{E}_{P_\lambda^n}[\mathbf{Tr}((\Sigma_n + \tfrac{\lceil \gamma n \rceil}{n\tau^2} I_{\lceil \gamma n \rceil})^{-1})], \\
\ell_n(\lambda, \tau, \gamma) &\coloneqq \frac{1}{\tau^2 n} \mathbf{E}_{P_\lambda^n}[\mathbf{Tr}((\Sigma_n + \tfrac{c_d \lceil \gamma n \rceil}{n\tau^2} I_{\lceil \gamma n \rceil})^{-1})].
\end{aligned}$$

Then, the relations (4.66), 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.

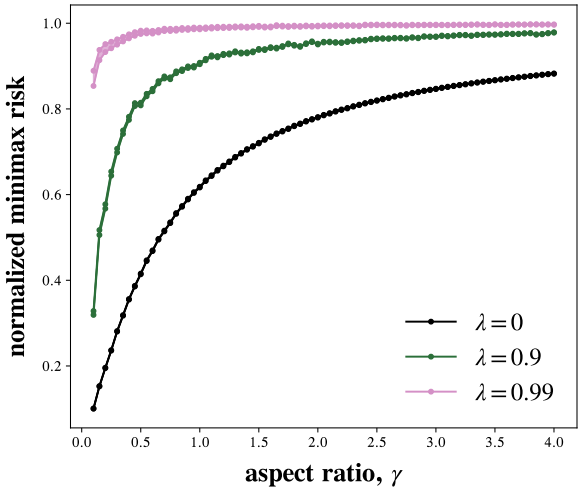
In our simulations, we take $\lambda \in \{0, 0.9, 0.99\}$ and vary $\gamma \in [0.05, 4]$. The results of these simulations are presented in Figure 4.1; see the caption for a detailed description and commentary. The general pattern should be clear: the covariate law can have a dramatic impact on the overall rate of estimation, even when restricting some moments such as we have with the relations (4.65).

4.5.1.2 Mixing time effects in Markovian linear regression

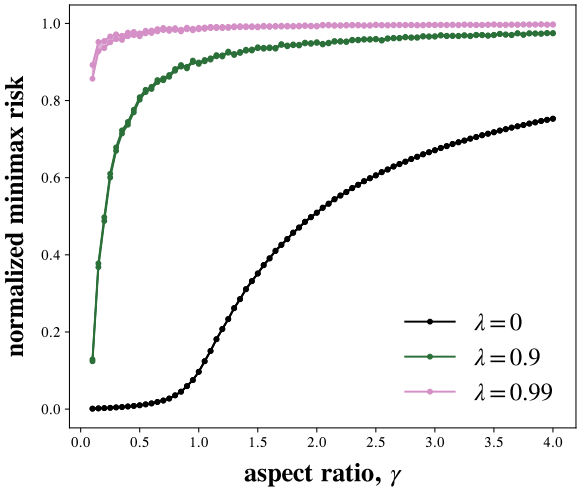
Covariates need not be drawn in an IID manner, and any dependencies can be expected to affect the minimax risk. Here we illustrate this general phenomena via some simulations for the Markov regression example as outlined in Section 4.3.1.4. We seek to study a wide range of possible mixing conditions for the Markovian covariate model. In order to do so, we consider covariates generated from the Markovian model (4.22) with

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

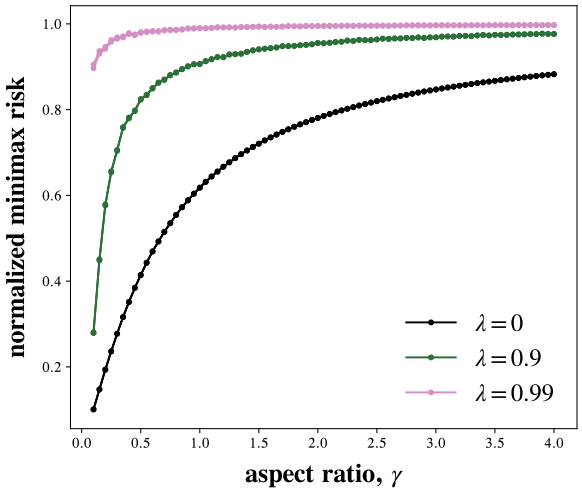
[7]Specifically, we take $d = \lceil \gamma n \rceil$.



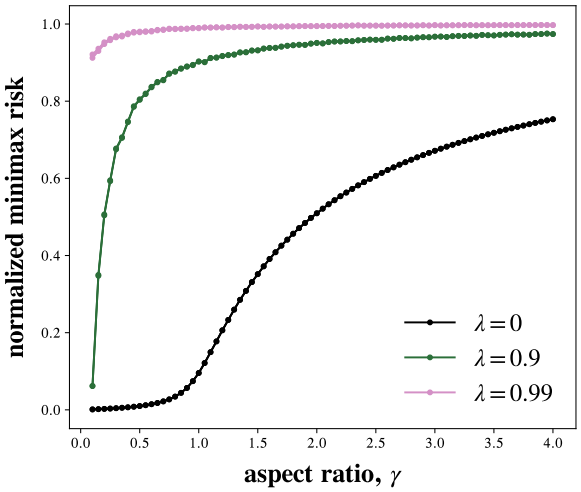
(a) $n = 128,\ \tau = 1$



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



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



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

Figure 4.1. Simulations of random design regression for three covariate laws, P_λ as defined in equation (4.64) 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 (4.67). In panels (4.1a) and (4.1b), we set the sample size $n = 128$, and set the SNR as $\tau = 1, 10$, respectively. In panels (4.1c) and (4.1d), 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 (4.1d)—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 (4.65)), 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 $d > 5$.

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 4.5, for this problem behaves for different choices of the function ψ and the signal-to-noise ratio (SNR). As in Section 4.5.1.1, 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 4.5.

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 4.2; see the caption for more details and commentary.

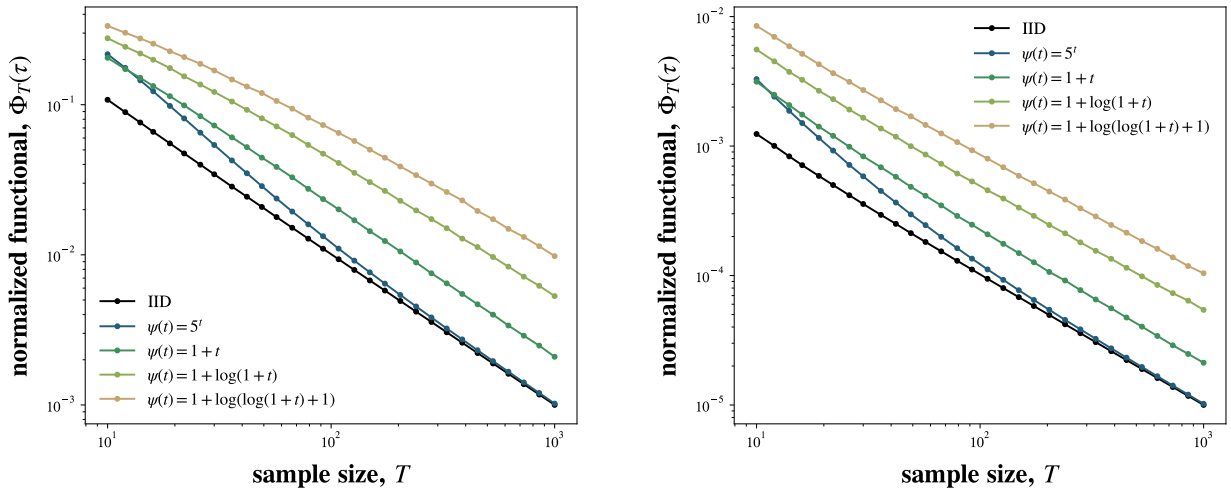
4.6 Deferred proofs

4.6.1 Proof of Proposition 4.1

The constraint set is evidently convex, as it is formed by the intersection of of two convex sets: the $d \times d$ real, symmetric positive definite matrices with the hyperplane $\{\Omega : \mathbf{Tr}({K_c}^{-1}\Omega) \leqslant \varrho^2\}$.

We claim that the objective function f is concave over the set of symmetric positive definite matrices. It can be expressed as

$$f(\Omega) = \mathbf{E}_\xi[g(T_\xi^{\mathsf{T}} \Sigma_w^{-1} T_\xi, \Omega)], \quad \text{where} \quad g(X, \Omega) \coloneqq \mathbf{Tr}({K_e}^{1/2}(X + \Omega^{-1})^{-1} {K_e}^{1/2}).$$



(a) $\tau = 1$ (b) $\tau = 10$

Figure 4.2. Simulations for five distributions of Markovian covariates. In panel (4.2a), we set the SNR parameter as $\tau = 1$, and in panel (4.2b), 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 $T \mapsto \Phi_T(\tau)$.

Evidently to establish that f is concave, it is enough to show that $g(X, \cdot)$ is concave for every symmetric positive semidefinite X. In order to establish this claim, let us fix some $\varepsilon > 0$, and define $X(\varepsilon) \coloneqq X + \varepsilon I_d$. By the joint concavity of the harmonic mean of positive operators [107, Corollary 37.2], it follows that for any pair of positive definite matrices Ω, Ω', we have

$$\left(X(\varepsilon) + \left(\frac{\Omega + \Omega'}{2}\right)^{-1}\right)^{-1} \succcurlyeq \frac{1}{2}\left(X(\varepsilon) + \Omega^{-1}\right)^{-1} + \frac{1}{2}\left(X(\varepsilon) + (\Omega')^{-1}\right)^{-1}.$$

Passing to the limit as $\varepsilon \to 0$ yields

$$\left(X + \left(\frac{\Omega + \Omega'}{2}\right)^{-1}\right)^{-1} \succcurlyeq \frac{1}{2}\left(X + \Omega^{-1}\right)^{-1} + \frac{1}{2}\left(X + (\Omega')^{-1}\right)^{-1}.$$

Since the trace is a monotone mapping on positive definite matrices, and g is continuous in its second argument, we obtain the claimed concavity of g.

4.6.2 Proof of Proposition 4.2

To establish the upper bound, it suffices to show that for each positive definite $\Omega \succ 0$ with $\mathbf{Tr}(K_c^{-1/2} \Omega K_c^{-1/2}) \leqslant \frac{n\varrho^2}{\sigma^2}$ that the following inequality holds

$$\mathbf{Tr}\left(\mathbf{E}\left[\left(\Sigma_n + \Omega^{-1}\right)^{-1} K_e\right]\right) \leqslant \left(1 + \frac{\varrho^2 \kappa^2}{\sigma^2}\right) \mathbf{Tr}\left(\left(\Sigma_P + \Omega^{-1}\right)^{-1} K_e\right). \tag{4.68}$$

To prove inequality (4.68), we begin by stating a more general result: a multiplicative positive operator inequality. We use the notation

$$\mathrm{Var}(W) = \mathbf{E}[W^2] - (\mathbf{E}\, W)^2$$

whenever W is a random positive, self-adjoint operator.

Theorem 4.3 (Random positive operator inequality)**.** *Let Y denote a random positive definite matrix. Suppose that there exists a (deterministic) positive definite $Z \succ 0$ such that*

$$\widetilde{Y} \coloneqq (\mathbf{E}\, Y)^{-1/2} Y (\mathbf{E}\, Y)^{-1/2} \succcurlyeq Z, \quad \text{almost surely.}$$

Then, the following sandwich relation holds,

$$(\mathbf{E}\, Y)^{-1} \preccurlyeq \mathbf{E}[Y^{-1}] \preccurlyeq \Big(1 + \|\!| Z^{1/2} \mathrm{Var}(Z^{-1/2} \widetilde{Y} Z^{-1/2}) Z^{1/2} |\!\|_{\mathrm{op}}\Big) (\mathbf{E}\, Y)^{-1}.$$

For proof, see Section 4.6.2.1. Note that this result can be viewed as a strengthening and generalization of Lemma 2 in the paper [89].

We can instantiate Theorem 4.3 in the special case where the random matrix Y arises as an average of IID summands. This immediately yields the following consequence.

Corollary 4.8. *Let $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$ denote an average of IID , random, positive definite matrices. Suppose that there exists a (deterministic) positive definite matrix $W \succ 0$ such that*

$$\widetilde{X}_1 \coloneqq (\mathbf{E}\, X_1)^{-1/2} X_1 (\mathbf{E}\, X_1)^{-1/2} \succcurlyeq W, \quad \text{almost surely.}$$

Then, the following sandwich relation holds,

$$(\mathbf{E}\, X_1)^{-1} \preccurlyeq \mathbf{E}[\overline{X_n}^{-1}] \preccurlyeq \left(1 + \frac{1}{n} \|\!| W^{1/2} \mathrm{Var}(W^{-1/2} \widetilde{X}_1 W^{-1/2}) W^{1/2} |\!\|_{\mathrm{op}}\right) (\mathbf{E}\, X_1)^{-1}.$$

We now demonstrate how Corollary 4.8 establishes inequality (4.68).

Proof of bound (4.68) In Corollary 4.8, we can ensure that $\overline{X_n} = \Sigma_n + \Omega^{-1}$ by taking $X_i = \psi(x_i) \otimes \psi(x_i) + \Omega^{-1}$ and $W = (\Sigma_P + \Omega^{-1})^{-1/2} \Omega^{-1} (\Sigma_P + \Omega^{-1})^{-1/2}$. Then we have

$$\begin{aligned}
&W^{1/2} \mathrm{Var}(W^{-1/2} \widetilde{X}_1 W^{-1/2}) W^{1/2} \\
&\quad = (\Sigma_P + \Omega^{-1})^{-1/2} \Big[\mathbf{E}\left[\|\Omega^{1/2} \psi(x)\|_2^2 \psi(x) \otimes \psi(x) \right] - \Sigma_P \Omega \Sigma_P \Big] (\Sigma_P + \Omega^{-1})^{-1/2} \\
&\quad \preccurlyeq \frac{n \kappa^2 \varrho^2}{\sigma^2} I.
\end{aligned}$$

The final inequality uses the P-almost sure inequality

$$\|\Omega^{1/2} \varphi(x)\|_2 \leqslant \|\!| \Omega^{1/2} {K_c}^{-1/2} |\!\|_{\mathrm{op}} \|{K_c}^{1/2} \varphi(x)\|_2 \leqslant \sqrt{n} \frac{\kappa \varrho}{\sigma}.$$

Since the inequality above implies $\|\!| W^{1/2} \mathrm{Var}(W^{-1/2} \widetilde{X}_1 W^{-1/2}) W^{1/2} |\!\|_{\mathrm{op}} \leqslant \frac{n \kappa^2 \varrho^2}{\sigma^2}$, bound (4.68) follows from Corollary 4.8.

4.6.2.1 Proof of Theorem 4.3

The lower bound is immediate by Jensen's inequality the operator convexity of the inverse over the space of positive definite matrices. By rescaling, it suffices to prove the upper bound under the assumption that $\mathbf{E}\, Y = I$. We can then write

$$Y^{-1} - I + (Y - I) = (Y - I)Y^{-1}(Y - I) \preccurlyeq (Y - I)Z^{-1}(Y - I).$$

Here, we used that $\widetilde{Y} = Y \succcurlyeq Z \succ 0$. Rearranging the above display and taking expectations, we find

$$\mathbf{E}[Y^{-1}] \preccurlyeq I + \mathbf{E}(Y - I)Z^{-1}(Y - I) \preccurlyeq \left(1 + |||Z^{1/2}\mathrm{Var}(Z^{-1/2}YZ^{-1/2})Z^{1/2}|||_{\mathrm{op}}\right) I,$$

thereby establishing our upper bound.

4.6.3 Proof of Corollary 4.2

Combining Theorems 4.1 and 4.2, we find that

$$\Phi(T, \mathbb{P}, \Sigma_w, \tfrac{\varrho}{2}, K_e, K_c) \preccurlyeq \mathfrak{M}(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c) \preccurlyeq \Phi(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c). \tag{4.69}$$

Evidently, by definition of the functional Φ (see definition (4.4)), the map

$$\varrho \to \Phi(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c)$$

is nondecreasing. Moreover since $T_\xi^{\mathsf{T}}\Sigma_w^{-1}T_\xi$ is invertible with probability 1, it is a bounded function. Therefore,

$$\lim_{\varrho\to\infty} \frac{\Phi(T, \mathbb{P}, \Sigma_w, \varrho, K_e, K_c)}{\Phi(T, \mathbb{P}, \Sigma_w, \varrho/2, K_e, K_c)} = 1,$$

which in view of the sandwich relation (4.69), furnishes the claim.

4.6.4 Proof and calculations from Section 4.3.1

4.6.4.1 Proof of equation (4.17a)

From the definition of the functional (4.12), we have

$$d_n(\mathsf{N}(0, I_d), \varrho, \sigma^2, I_d, I_d) = \sup\Big\{ \mathbf{E}[\mathbf{Tr}((\Sigma_n + \tfrac{\sigma^2 d}{n\varrho^2}M^{-1})^{-1})] : M \succ 0, \quad \mathbf{Tr}(M) = d \Big\}.$$

In this section, all expectations are over $x_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, I_d)$. We claim that the supremum above is achieved at $M = I_d$.

Lemma 4.8. *For any positive definite matrix $M \succ 0$ such that $\mathbf{Tr}(M) = d$, we have*

$$\mathbf{E}[\mathbf{Tr}((\Sigma_n + \tfrac{\sigma^2 d}{n\varrho^2}M^{-1})^{-1})] \preccurlyeq \mathbf{E}[\mathbf{Tr}((X^{\mathsf{T}}X + \tfrac{d\sigma^2}{\varrho^2}I_d)^{-1})]$$

Assuming Lemma 4.8, we then have

$$d_n(\mathsf{N}(0, I_d), \varrho, \sigma^2, I_d, I_d) = \mathbf{E}[\mathbf{Tr}((\Sigma_n + \tfrac{\sigma^2 d}{n\varrho^2}I_d)^{-1})] = d_{\mathrm{Dicker}}(n, d, \varrho, \sigma),$$

which establishes (4.17a), as needed.

Proof of Lemma 4.8 Define the function $\phi\colon (\Sigma, M) \mapsto (\Sigma + \frac{d\sigma^2}{n\varrho^2} M^{-1})^{-1}$, where Σ, M are assumed symmetric positive semidefinite and M is nonsingular. For each $\Sigma \succcurlyeq 0$, it is well known that $\phi(\Sigma, \cdot)$ is operator concave [107, Corollary 37.2]—for any collection $\{M_i\}_{i=1}^d$ of symmetric positive definite matrices, one has

$$\frac{1}{d}\sum_{i=1}^{d} \phi(\Sigma, M_i) \preccurlyeq \phi(\Sigma, \tfrac{1}{d}\sum_{i=1}^{d} M_i), \qquad \text{for any } \Sigma \in \mathbb{S}_+^d. \tag{4.70}$$

Now let $M \succ 0$ satisfying $\mathbf{Tr}(M) = d$ be given. Diagonalize M so that $M = U\Lambda U^{\mathsf{T}}$, where $\Lambda = \mathbf{diag}(\lambda) \succ 0$, and U is orthogonal. Consider the cyclic permutations of Λ, given by

$$\Lambda^{(j)} = \mathbf{diag}(\lambda^{(j)}), \quad \text{where} \quad \lambda_i^{(j)} = \lambda_{i+j}.$$

Above, the arithmetic $i + j$ occurs modulo d. By rotational invariance of the Gaussian and the fact that x_i has iid coordinates, we have

$$\begin{aligned}
\mathbf{E}\,\mathbf{Tr}((\Sigma_n + \tfrac{d\sigma^2}{n\varrho^2} M^{-1})^{-1}) &= \mathbf{E}\,\mathbf{Tr}((\Sigma_n + \tfrac{d\sigma^2}{n\varrho^2} \Lambda^{-1})^{-1}) \\
&= \mathbf{E}\Big[\frac{1}{d}\sum_{j=1}^{d} \mathbf{Tr}((\Sigma_n + \tfrac{d\sigma^2}{n\varrho^2}(\Lambda^{(j)})^{-1})^{-1})\Big] \\
&= \mathbf{Tr}\Big\{\mathbf{E}\Big[\frac{1}{d}\sum_{j=1}^{d} \phi(\Sigma_n, \Lambda^{(j)})\Big]\Big\} \\
&\leqslant \mathbf{Tr}\Big\{\mathbf{E}\Big[\phi(\Sigma_n, \overline{\Lambda})\Big]\Big\} \qquad \text{where} \quad \overline{\Lambda} \coloneqq \tfrac{1}{d}\sum_{j=1}^{d} \Lambda^{(j)},
\end{aligned}$$

The final inequality above uses the concavity inequality (4.70), where we have taken $M_i = \Lambda^{(i)}$. Now note that

$$\overline{\Lambda} = \frac{\mathbf{Tr}(\Lambda)}{d} I_d = \frac{\mathbf{Tr}(M)}{d} I_d = I_d.$$

Combining the preceding displays furnishes the claim.

4.6.4.2 Proof of the lower bound in equation (4.16a)

We apply our our sharp lower bound in Theorem 4.2 with $\Omega = \frac{\varrho^2}{d} I_d$ and $\tau^2 = 1 - \frac{1}{2d-1}$. Let us define $u = (1 - \frac{1}{2d-1})(1 - \mathbf{P}\{Z > 2d^2 - d\})$, where Z is a χ^2-random variable with d-degrees of freedom. Note that $d(d-1) \geqslant \sqrt{dt} + t$ for $t = \frac{d^{3/2}}{4}$ for all $d \geqslant 2$. Therefore by standard tail bounds for χ^2-variates [70, pp. 1325], we have $u \leqslant \exp(-d^{3/2}/4)$. Applying the sharp lower bound (4.7) in Theorem 4.2 then yields the claim.

4.6.4.3 Proof of equation (4.21)

Using the semidefinite inequality

$$\left(\Sigma_n + \Omega^{-1}\right)^{-1} \preccurlyeq \Sigma_n^{-1},$$

and the choice $\Omega = \frac{n}{\sigma^2}\frac{\varrho^2}{d} I_d$, we have the sandwich relation

$$\mathbf{Tr}\,\mathbf{E}_{P^n}\left[\Sigma_P^{1/2}(\Sigma_n + \tfrac{\sigma^2}{n}\tfrac{d}{\varrho^2} I_d)^{-1}\Sigma_P^{1/2}\right] \leqslant d_n(P, \varrho, \sigma^2, I_d, \Sigma_P) \leqslant \mathbf{Tr}\,\mathbf{E}_{P^n}\left[\Sigma_P^{1/2}\Sigma_n^{-1}\Sigma_P^{1/2}\right],$$

for all $\varrho > 0$. Since $\varrho \mapsto d_n(P, \varrho, \sigma^2, I_d, \Sigma_P)$ is nondecreasing, the display above also demonstrates that this map has a limit. Now, note that by continuity, P^n-almost surely we have

$$\lim_{\varrho\to\infty} \mathbf{Tr}(\Sigma_P^{1/2}(\Sigma_n + \tfrac{\sigma^2}{n}\tfrac{d}{\varrho^2} I_d)^{-1}\Sigma_P^{1/2}) = \mathbf{Tr}(\Sigma_P^{1/2}\Sigma_n^{-1}\Sigma_P^{1/2}).$$

Thus, using the sandwich relation (4.21) and Fatou's lemma, we have

$$\begin{aligned}\mathbf{Tr}\,\mathbf{E}_{P^n}\left[\Sigma_P^{1/2}\Sigma_n^{-1}\Sigma_P^{1/2}\right] &\leqslant \liminf_{\varrho\to\infty} \mathbf{Tr}\,\mathbf{E}_{P^n}\left[\Sigma_P^{1/2}(\Sigma_n + \tfrac{\sigma^2}{n}\tfrac{d}{\varrho^2} I_d)^{-1}\Sigma_P^{1/2}\right] \\ &\leqslant \lim_{\varrho\to\infty} d_n(P, \varrho, \sigma^2, I_d, \Sigma_P) \leqslant \mathbf{Tr}\,\mathbf{E}_{P^n}\left[\Sigma_P^{1/2}\Sigma_n^{-1}\Sigma_P^{1/2}\right],\end{aligned}$$

which establishes relation (4.21), as required.

4.6.4.4 Proof of minimax relation (4.25)

Let us state the claim corresponding to relation (4.25) somewhat more precisely. We define the functional

$$\Phi_T(\varrho, \sigma) \coloneqq \mathbf{E}\left[\left(\frac{1}{\varrho^2} + \frac{z^\mathsf{T} M z}{\sigma^2}\right)^{-1}\right]$$

Then the following lemma corresponds to the claim underlying relation (4.25).

Lemma 4.9. *The minimax risk under the Markovian observation model defined by the displays* (4.22) *and* (4.23) *satisfies*

$$\frac{1}{4}\,\Phi_T(\varrho, \sigma) \leqslant \inf_{\widehat{\theta}} \sup_{|\theta^\star| \leqslant \varrho} \mathbf{E}\left[(\widehat{\theta} - \theta^\star)^2\right] \leqslant \Phi_T(\varrho, \sigma).$$

The remainder of this section is devoted to the proof of this claim

Note that if we define $\xi = (x_1, \ldots, x_T)$, and $T_\xi = x$, then the observation model (4.23) can be written

$$y = T_\xi \theta^\star + \Sigma_w^{1/2} w,$$

where $w \sim \mathsf{N}(0, I_T)$ and $\Sigma_w = \sigma^2 I_T$. We have $K_c = 1 = K_e$, since we are considering a univariate estimation problem. Therefore, since the functional (4.4) is attained at $\Omega = \varrho^2$, in order to establish Lemma 4.9, it is sufficient to show that

$$T_\xi^\mathsf{T} \Sigma_w^{-1} T_\xi = \frac{x^\mathsf{T} x}{\sigma^2} = \frac{z^\mathsf{T} M z}{\sigma^2}. \tag{4.71}$$

However, from display (4.22), by induction we can establish that

$$x_t = \sum_{s=1}^{t} \sqrt{c_{st}}\, z_s,$$

where the coefficients $\{c_{st}\}$ are defined as in display (4.24). Then, it follows that

$$x^\mathsf{T} x = \sum_{t=1}^{T} \sum_{s,s'=1}^{t} \sqrt{c_{st} c_{s't}} z_s z_{s'} = \sum_{s,s'=1}^{T} \underbrace{\sum_{t=s \vee s'} \sqrt{c_{st} c_{s't}}}_{= M_{ss'}} z_s z_{s'}.$$

Using the display above, we establish the relation (4.71), which in turn establishes Lemma 4.9, as needed.

4.6.5 Proof and calculations from Section 4.3.2

4.6.5.1 Proof of limit relation (4.27)

To lighten notation in this section, let us define the shorthands

$$\mathfrak{M}_k \coloneqq \mathfrak{M}_k\Big(\{\varepsilon_j\}_{j=1}^k, \Theta_k(a, C)\Big), \quad \text{and,} \tag{4.72a}$$

$$\mathfrak{M} \coloneqq \mathfrak{M}\Big(\{\varepsilon_j\}_{j=1}^\infty, \Theta(a, C)\Big) \coloneqq \inf_{\widehat{\theta}} \sup_{\theta^\star \in \Theta(a,C)} \mathbf{E}\Big[\sum_{j=1}^{\infty} (\widehat{\theta}_j(y) - \theta_j^\star)^2\Big]. \tag{4.72b}$$

We begin by stating the following sandwich relation for the minimax risks.

Lemma 4.10. *The sequence of minimax risks $\{\mathfrak{M}_k\}$ and infinite-dimensional risk $\mathfrak{M}$ satisfies the sandwich relation*

$$\mathfrak{M}_k \leqslant \mathfrak{M} \leqslant \mathfrak{M}_k + \frac{C^2}{a_{k+1}^2}, \tag{4.73}$$

for all $k \geqslant 1$.

Assuming Lemma 4.10 for the moment, note that it implies for any divergent sequence $a_k \to \infty$ that

$$\lim_{k \to \infty} \mathfrak{M}_k = \mathfrak{M}.$$

In view of the shorthands (4.72), the display above establishes our desired limit relation (4.27).

Proof of Lemma 4.10 We begin by establishing the lower bound. Note that $\Theta_k(a, C) \subset \Theta(a, C)$, hence we have

$$\begin{aligned}\mathfrak{M} &\geqslant \inf_{\widehat{\theta}} \sup_{\theta^\star \in \Theta_k(a,C)} \mathbf{E}\Big[\sum_{j=1}^{\infty} (\widehat{\theta}_j((y_i)_{i=1}^{\infty}) - \theta_j^\star)^2\Big] \\ &\geqslant \inf_{\widehat{\theta}} \sup_{\theta^\star \in \Theta_k(a,C)} \mathbf{E}\Big[\sum_{j=1}^{k} (\widehat{\theta}_j((y_i)_{i=1}^{\infty}) - \theta_j^\star)^2\Big],\end{aligned}$$

where the last equation arises since $\theta_j^\star = 0$ for $j > k$ and thus any minimax optimal estimator over $\Theta_k(a, C)$ satisfies $\widehat{\theta}_j \equiv 0$ for all $j > k$. The righthand side differs from $\mathfrak{M}_k$ in that $\widehat{\theta}$ is a function of the full sequence $y = (y_i)_{i=1}^{\infty}$. However, note that due to the independence of the noise variables z_i, for the observation model (4.26) restricted to $\Theta_k(a, C)$, the vector $y^{(k)} = (y_i)_{i=1}^k$ is a sufficient statistic. Hence we have for each $k \geqslant 1$,

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

which establishes the lower bound in relation (4.73).

To establish the upper bound, note that we certainly may restrict the infimum in the definition of $\mathfrak{M}$ to those estimators taking values in $\mathbf{R}^k$ which only are a function of $y^{(k)}$. Indeed, we then find

$$\begin{aligned}\mathfrak{M} &\leqslant \inf_{\widehat{\theta} \in \mathbf{R}^k} \sup_{\theta^\star \in \Theta(a,C)} \mathbf{E}\Big[\sum_{j=1}^{k} (\widehat{\theta}_j(y^{(k)}) - \theta_j^\star)^2 + \sum_{j>k} (\theta_j^\star)^2\Big] && (4.74) \\ &\leqslant \mathfrak{M}_k + \sup_{\theta^\star \in \Theta(a,C)} \sum_{j>k} (\theta_j^\star)^2. && (4.75)\end{aligned}$$

The inequality (4.75) arises by taking the supremum over the two terms of the risk in display (4.74), and noting the first term only depends on the first k coordinate of $\theta^\star \in \Theta(a, C)$, and hence the supremum may be taken over $\Theta_k(a, C)$ in the first term so as to obtain $\mathfrak{M}_k$.

Now observe by Hölder's inequality, and the membership $\theta^\star \in \Theta(a, C)$,

$$\sum_{j>k} (\theta_j^\star)^2 = \sum_{j>k} \frac{1}{a_j^2}(a_j^2(\theta_j^\star)^2) \leqslant \Big(\max_{j>k} \frac{1}{a_j^2}\Big) C^2 = \frac{C^2}{a_{k+1}^2},$$

with the last equality arising because $j \mapsto a_j^2$ is assumed nondecreasing. Combining the display above with inequality (4.75) establishes the upper bound in (4.73), and thus establishes Lemma 4.10 as needed.

4.6.5.2 Proof of relation (4.31)

Let us continue to adopt the shorthands $\mathfrak{M}_k$ and $\mathfrak{M}$ defined, respectively, in the displays (4.72a) and (4.72b). Moreover, we also use the shorthands

$$R_k^\star \coloneqq R_k^\star\Big(\{\varepsilon\}_{j=1}^k, \{a_j\}_{j=1}^k, C\Big), \quad \text{and} \quad R^\star \coloneqq R^\star(\varepsilon, a, C),$$

corresponding to the functionals (4.29) and (4.30), respectively.

We prove the following lemma.

Lemma 4.11. *The functionals $R_k^\star$, $R^\star$ and minimax risks $\mathfrak{M}_k$ satisfy*

$$\frac{1}{4}R_k^\star \leqslant \mathfrak{M}_k \leqslant R_k^\star \quad \textit{for all } k \geqslant 1, \textit{ and,} \tag{4.76a}$$

$$\lim_{k\to\infty} R_k^\star = R^\star. \tag{4.76b}$$

Assuming Lemma 4.11 for the moment, note that the two inequalities immediately imply the sandwich relation (4.31), simply by applying the sandwich (4.76a) to the terms $\mathfrak{M}_k$ and then applying the limit relations (4.27) and (4.76b). Consequently, it suffices to establish Lemma 4.11.

Proof of Lemma 4.11 Recall the settings of the parameters $T^{(k)}, \Sigma_w^{(k)}, {K_e}^{(k)}, \varrho^{(k)}, {K_c}^{(k)}$, corresponding to the k dimensional minimax risk $\mathfrak{M}_k$, as given in (4.28). We claim that

$$\Phi(T^{(k)}, \mathbb{P}, \Sigma_w^{(k)}, \varrho^{(k)}, {K_e}^{(k)}, {K_c}^{(k)}) = R_k^\star. \tag{4.77}$$

(Note by our construction of $T^{(k)}$ the choice of $\mathbb{P}$ is irrelevant.) Then the sandwich relation (4.76a) follows by applying Theorems 4.1 and 4.2 to the minimax risk $\mathfrak{M}_k$.

To see that relation (4.77) holds, note that by definition 4.4, we have

$$\Phi(T^{(k)}, \mathbb{P}, \Sigma_w^{(k)}, \varrho^{(k)}, {K_e}^{(k)}, {K_c}^{(k)}) = \sup_{\Omega\succ 0}\Big\{ \mathbf{Tr}\left((\Omega^{-1} + (\Sigma_w^{(k)})^{-1})^{-1}\right) : \sum_{j=1}^k a_j^2\Omega_{jj} \leqslant C^2 \Big\}.$$

We claim that the supremum above can be reduced to diagonal Ω. To see why, first note that for every nonzero $\lambda \in \mathbf{R}$

$$\left(\Omega^{-1} + (\Sigma_w^{(k)})^{-1}\right)^{-1} \preccurlyeq \lambda^2\Omega + (1-\lambda)^2\Sigma_w^{(k)}.$$

This follows from Lemma 4.14, with the choices

$$A = \Sigma_w^{(k)}, \quad B = \Omega^{-1}, \quad \text{and} \quad D = \lambda I.$$

Consequently, we have for every nonzero $u \in \mathbf{R}^k$, that

$$u^\mathsf{T}\left(\Omega^{-1} + (\Sigma_w^{(k)})^{-1}\right)^{-1}u \leqslant \inf_{\lambda\in\mathbf{R}} \lambda^2 u^\mathsf{T}\Omega u + (1-\lambda)^2 u^\mathsf{T}\Sigma_w^{(k)}u = \left(\frac{1}{u^\mathsf{T}\Omega u} + \frac{1}{u^\mathsf{T}\Sigma_w^{(k)}u}\right)^{-1} \tag{4.78}$$

Hence taking u to be elements of the standard basis e_i, and summing over $i = 1, \ldots, k$, we obtain,

$$\mathbf{Tr}\left(\left(\Omega^{-1} + (\Sigma_w^{(k)})^{-1}\right)^{-1}\right) \leqslant \sum_{i=1}^{k}\left(\frac{1}{\Omega_{ii}} + \frac{1}{\varepsilon_i^2}\right)^{-1} = \sum_{i=1}^{k}\frac{\Omega_{ii}\varepsilon_i^2}{\Omega_{ii}+\varepsilon_i^2}.$$

Moreover, by taking Ω to be diagonal, the inequality above holds with equality. Thus,

$$\begin{aligned}\Phi(T^{(k)}, \mathbb{P}, \Sigma_w^{(k)}, \varrho^{(k)}, {K_e}^{(k)}, {K_c}^{(k)}) &= \sup_{\Omega_{jj}>0}\Big\{\sum_{j=1}^{k}\frac{\Omega_{jj}\varepsilon_j^2}{\Omega_{jj}+\varepsilon_j^2} : \sum_{j=1}^{k} a_j^2\Omega_{jj} \leqslant C^2\Big\} \\ &= \sup_{\tau_j^2>0}\Big\{\sum_{j=1}^{k}\frac{\tau_j^2\varepsilon_j^2}{\tau_j^2+\varepsilon_j^2} : \sum_{j=1}^{k} a_j^2\tau_j^2 \leqslant C^2\Big\} \\ &= R_k^\star,\end{aligned}$$

which establishes the relation (4.77). Note that in the last equality, we have dropped the inequality constraints $\tau_j^2 > 0$, due to the continuity of the map $\tau \mapsto \sum_{i=1}^{k}\frac{\tau_j^2\varepsilon_j^2}{\tau_j^2+\varepsilon_j^2}$ over $\tau \in \mathbf{R}^k$.

We now turn to establishing the relation (4.76b). Note that for any $\tau \in \mathbf{R}^{\mathbf{N}}$ with $\sum_{j=1}^{\infty} a_j^2\tau_j^2 \leqslant C^2$, we have

$$\sum_{j=1}^{k}\frac{\tau_j^2\varepsilon_j^2}{\tau_j^2+\varepsilon_j^2} \leqslant \sum_{j=1}^{\infty}\frac{\tau_j^2\varepsilon_j^2}{\tau_j^2+\varepsilon_j^2} \leqslant \sum_{j=1}^{k}\frac{\tau_j^2\varepsilon_j^2}{\tau_j^2+\varepsilon_j^2} + \sup_{\tau\in\mathbf{R}^{\mathbf{N}}:\sum_{j=1}^{\infty}a_j^2\tau_j^2\leqslant C^2}\sum_{j>k}^{\infty}\tau_j^2$$

By Hölder's inequality, the second term is bounded above by C^2/a_{k+1}^2, hence in view of definitions (4.29) and (4.30), we have the sandwich relation

$$R_k^\star \leqslant R^\star \leqslant R_k^\star + \frac{C^2}{a_{k+1}^2},$$

which holds for all $k \geqslant 1$. Since $a_k \to \infty$, the limit relation (4.76b) follows.

4.6.5.3 Proof of limit relation (4.38)

We claim that the following sandwich relation holds for the minimax risks in this case.

Lemma 4.12. *For all $k \geqslant 1$, we have*

$$\mathfrak{M}_n^{(k)}(\varrho, \sigma^2, P) \leqslant \mathfrak{M}_n(\varrho, \sigma^2, P) \leqslant \mathfrak{M}_n^{(k)}(\varrho, \sigma^2, P) + \varrho^2\mu_{k+1}. \tag{4.79}$$

Assuming Lemma 4.12, note that since $\mu_k \to 0$ as $k \to \infty$, it immediately implies limit relation (4.38)

Proof of Lemma 4.12 The proof is quite similar to Lemma 4.10. We now prove inequality (4.79). We begin by defining the sets

$$\mathcal{B}(\varrho) = \{\theta \in \ell^2(\mathbf{N}) : \|\theta\|_2 \leqslant \varrho\}, \quad \text{and} \quad \mathcal{B}_k(\varrho) = \{\theta \in \mathcal{B}_k(\varrho) : \theta_j = 0, \quad \text{for all } j > k\}.$$

By Parseval's identity, we may rewrite the minimax risks in the following form

$$\mathfrak{M}_k \equiv \mathfrak{M}_n^{(k)}(\varrho, \sigma^2, P) = \inf_{\widehat{\theta}} \sup_{\substack{\theta^\star \in \mathcal{B}_k(\varrho) \\ \nu \in \mathcal{P}(\sigma^2 I_n)}} \mathbf{E}\Big[\sum_{j=1}^{k} \mu_j (\widehat{\theta}_j(y_1, \ldots, y_n, \Phi_k(x_1), \ldots, \Phi_k(x_n)) - \theta_j^\star)^2 \Big],$$

$$\mathfrak{M} \equiv \mathfrak{M}_n(\varrho, \sigma^2, P) = \inf_{\widehat{\theta}} \sup_{\substack{\theta^\star \in \mathcal{B}(\varrho) \\ \nu \in \mathcal{P}(\sigma^2 I_n)}} \mathbf{E}\Big[\sum_{j=1}^{\infty} \mu_j (\widehat{\theta}_j(y_1, \ldots, y_n, \Phi(x_1), \ldots, \Phi(x_n)) - \theta_j^\star)^2 \Big].$$

Evidently, we have $\mathfrak{M} \geqslant \mathfrak{M}_k$, since $\mathcal{B}_k(\varrho) \subset \mathcal{B}(\varrho)$ and $(y, \Phi_k(x))$ are sufficient in this submodel. Similarly, the upper bound follows since by restricting to those estimators $\widehat{\theta}$ with $\widehat{\theta}_j = 0$ for all $j > k$ that are functions of $(y, \Phi_k(x))$, we have

$$\mathfrak{M} \leqslant \mathfrak{M}_k + \sup_{\theta \in \mathcal{B}(\varrho)} \sum_{j>k} \mu_j \theta_j^2 = \mathfrak{M}_k + \varrho^2 \mu_{k+1},$$

which establishes the upper bound.

4.6.5.4 Proof of Corollary 4.6

Using the fact that $P = \nu$, we can define

$$\overline{d_{n,k}} \coloneqq \sup_{\Omega > 0} \Big\{ \mathbf{Tr}(I_k + M_k^{-1/2} \Omega^{-1} M_k^{-1/2})^{-1} : \mathbf{Tr}(\Omega) \leqslant \frac{n\varrho^2}{\sigma^2} \Big\}.$$

Let $\overline{d_n^\star} = \limsup_{k \to \infty} \overline{d_{n,k}}$. Then following Corollary 4.6 and Proposition 4.2, we obtain

$$\frac{1}{4} \frac{\sigma^2}{n} \overline{d_n^\star} \leqslant \mathfrak{M}(\varrho, \sigma^2, P) \leqslant 2 \frac{\sigma^2}{n} \overline{d_n^\star}. \tag{4.81}$$

We now simply the quantities $\overline{d_{n,k}}$. Using an argument analagous to the proof of inequality (4.78), we can write

$$\overline{d_{n,k}} = \sup_{\omega_1, \ldots, \omega_k > 0} \Big\{ \sum_{j=1}^{k} \frac{\omega_j \mu_j}{1 + \omega_j \mu_j} : \sum_{j=1}^{k} \omega_j \leqslant \frac{n\varrho^2}{\sigma^2} \Big\}.$$

Since $\frac{1}{2}(x \wedge 1) \leqslant \frac{x}{x+1} \leqslant x \wedge 1$ for any $x > 0$, we can then introduce

$$D_k \coloneqq \sup_{\omega_1, \ldots, \omega_k > 0} \Big\{ \sum_{j=1}^{k} \omega_j \mu_j \wedge 1 : \sum_{j=1}^{k} \omega_j \leqslant \frac{n\varrho^2}{\sigma^2} \Big\}.$$

Evidently $\frac{1}{2}D_k \leqslant \overline{d_{n,k}} \leqslant D_k$. By inspection, we have $k \wedge k_n \leqslant D_k \leqslant 2(k \wedge k_n)$, in which case it follows after passing to the superior limit that

$$k_n \leqslant \overline{d_n^\star} \leqslant 2k_n,$$

which upon combination with inequality (4.81) yields the claim.

4.6.5.5 Proof of relation (4.41)

Applying Corollary 4.1 to the minimax risk $\mathfrak{M}_k(\varrho, \sigma^2, P)$, we find that

$$\frac{1}{4}\frac{\sigma^2}{n} d_n^{(k)} \leqslant \mathfrak{M}_k(\varrho, \sigma^2, P) \leqslant \frac{\sigma^2}{n} d_n^{(k)},$$

since the quantity $d_n^{(k)}$ equals the functional for this minimax risk (see equation (4.40)). Therefore passing to the superior limit and applying the limit relation (4.38), we obtain the result.

4.6.5.6 Proof of relation (4.49)

Note that the kernel regularity condition is not necessary for our lower bound. Indeed, note that we first have

$$\inf_{\delta > 0} \left\{ \delta^2 + \frac{\sigma^2 B}{n\varrho^2} d(\delta) \right\} = \inf_{d \geqslant 1} \left\{ \mu_d + \frac{\sigma^2 B d}{n\varrho^2} \right\}$$

Let $d_n^\star$ be the largest integer d such that $\mu_d \geqslant \frac{\sigma^2 B d}{n\varrho^2}$; this must exist since $\mu_d \to 0$. As the two sequences are nonincreasing and strictly increasing, respectively, the display above is bounded above by

$$4\Big(\mu_{d_n^\star} \wedge \frac{\sigma^2 B d_n^\star}{n\varrho^2}\Big) \leqslant 4\frac{\sigma^2 B d_n^\star}{n\varrho^2}.$$

Hence, it suffices to establish that the lower bound $\frac{\sigma^2 B d_n^\star}{n\varrho^2}$ can be obtained from our result (4.47).

Note that if $\mu_d \geqslant \frac{\sigma^2 B d}{n\varrho^2}$ then the choice of λ in the lower bound (4.47), given by

$$\lambda_j = \frac{\sigma^2 B}{n\varrho^2} \frac{1}{\mu_j} \mathbf{1}\{j \leqslant d\}, \quad \text{for } j = 1, 2, 3, \ldots,$$

satisfies $\sum_j \lambda_j \leqslant 1$. Evaluating the corresponding lower bound, with the maximal choice $d = d_n^\star$ yields the lower bound $\frac{\sigma^2 B d}{n\varrho^2}$, as needed.

4.6.6 Deferred proofs from Section 4.4.1

In this section, we collect proofs of the results underlying the argument establishing our upper bound in Section 4.4.1.

4.6.6.1 Proof of Lemma 4.1

Clearly the lefthand side is less than the right hand side as for $\theta \in \Theta(\varrho, K_c)$ we have $\theta \otimes \theta \succcurlyeq 0$, and $\mathbf{Tr}(K_c^{-1/2}\theta \otimes \theta K_c^{-1/2}) = \|\theta\|^2_{K_c^{-1}} \leqslant \varrho^2$.

For the reverse inequality, fix $\Omega \in \mathcal{K}(\varrho, K_c)$. We diagonalize the positive semidefinite matrix $K_c^{-1/2}\Omega K_c^{-1/2} = UDU^T$, and define $\theta(\varepsilon) = K_c^{1/2}UD^{1/2}\varepsilon$, where $\varepsilon \in \{\pm 1\}^d$. Evidently,

$$\|\theta(\varepsilon)\|^2_{K_c^{-1}} = \|UD^{1/2}\varepsilon\|_2^2 = \mathbf{Tr}(D) = \mathbf{Tr}(K_c^{-1/2}\Omega K_c^{-1/2}) \leqslant \varrho^2.$$

Thus, for all $\varepsilon \in \{\pm 1\}^d$, the vector $\theta(\varepsilon)$ lies in the set $\Theta(\varrho, K_c)$. Consequently, we have

$$\begin{aligned} \sup_{\theta \in \Theta(\varrho, K_c)} r(\widehat{\theta}_C, \theta) &\geqslant \max_{\varepsilon \in \{\pm 1\}^d} r(\widehat{\theta}_C, \theta(\varepsilon)) \\ &\geqslant \mathbf{E}_\varepsilon \, r(\widehat{\theta}_C, \theta(\varepsilon)) \qquad (4.82) \\ &= r(\widehat{\theta}_C, \Omega). \qquad (4.83) \end{aligned}$$

Note that $\Omega \in \mathcal{K}(\varrho, K_c)$ was arbitrary in this argument, and hence passing to supremum over Ω gives us the desired reverse inequality. Above, display (4.82) follows by lower bounding the maximum over $\varepsilon \in \{\pm 1\}^d$ by the expectation over ε where ε_i are IID Rademacher variables. The relation (4.83) follows by noting that $r(\widehat{\theta}_C, \theta(\varepsilon)) = r(\widehat{\theta}_C, \theta(\varepsilon) \otimes \theta(\varepsilon))$, and moreover this latter quantity is linear in the rank-one matrix $\theta(\varepsilon) \otimes \theta(\varepsilon)$, as justified by Lemma 4.2. By linearity of expectation we can bring the expectation inside, and use the fact that

$$\mathbf{E}_\varepsilon[\theta(\varepsilon) \otimes \theta(\varepsilon)] = K_c^{1/2}UDU^T K_c^{1/2} = \Omega.$$

4.6.6.2 Proof of Lemma 4.2

Inspecting the definition of r (see equation (4.53)), we see that it is affine in Ω. To verify that it is convex in C, note that r can be equivalently expressed as

$$r(\widehat{\theta}_C, \Omega) \quad = \quad \mathbf{E}_\xi \Big[\|\!| K_e^{1/2}(C(T_\xi)T_\xi^T \Sigma_w^{-1} T_\xi \;-\; I_d)\Omega^{1/2} |\!\|_{\mathrm{F}}^2 \;+\; \|\!| K_e^{1/2}(C(T_\xi)T_\xi^T \Sigma_w^{-1/2} |\!\|_{\mathrm{F}}^2 \Big].$$

Evidently, the display above is convex in C.

4.6.6.3 Proof of Proposition 4.3

In order to prove Proposition 4.3, we need two results regarding the harmonic mean of positive (semi)definite matrices. For our results, it is important to allow once of these matrices to be (possibly) singular, and so we study (twice) the harmonic mean of A and the Moore-Penrose pseudoinverse $B^\dagger$—that is, the quantity $(A^{-1} + B)^{-1}$, where $B \succcurlyeq 0$ and $A \succ 0$. Note that since $(B^\dagger)^\dagger = B$, these results also imply bounds for the mean $(A^{-1} + B^\dagger)^{-1}$. See the reference [10, chap. 4] for additional details about the harmonic mean of positive matrices.

Lemma 4.13. *Suppose that A, B are two symmetric positive semidefinite matrices, and that A is nonsingular. For any $x \in \mathbf{R}^d$ and any y in the range of B, we have*

$$(x-y)^\mathsf{T} A(x-y) + y^\mathsf{T} B^\dagger y \geqslant x^\mathsf{T}(A^{-1}+B)^{-1}x,$$

where $B^\dagger$ denotes the Moore-Penrose pseudoinverse associated with B.

Proof. Using $BB^\dagger B = B$, the claim is equivalent to showing that $\inf_{x,u} g(x,u) \geqslant 0$ where

$$g(x,u) \coloneqq (x-Bu)^\mathsf{T} A(x-Bu) + u^\mathsf{T} Bu - x^\mathsf{T}(A^{-1}+B)^{-1}x.$$

Define $f(u) = \inf_x g(x,u)$. A calculation demonstrates that

$$\begin{aligned} f(u) &= u^\mathsf{T}\Big[B + BAB - BA(A-(A^{-1}+B)^{-1})^\dagger AB\Big]u \\ &= u^\mathsf{T} BA^{1/2}\Big[K^\dagger + I - (I-(I+K)^{-1})^\dagger\Big]A^{1/2}Bu. \end{aligned} \tag{4.84}$$

Above, $K \coloneqq A^{1/2}BA^{1/2}$. Diagonalizing K, we may write $K = UDU^\mathsf{T}$ and therefore $K^\dagger = UD^\dagger U^T$. Applying the similarity transformation under U, we have

$$U^\mathsf{T}(K^\dagger + I - (I-(I+K)^{-1})^\dagger)U = D^\dagger + I - (I-(I+D)^{-1})^\dagger = I - D^\dagger D \geqslant 0. \tag{4.85}$$

Therefore, combining displays (4.84) with (4.85), we obtain

$$\inf_{x,u} g(x,u) = \inf_u f(u) \geqslant 0,$$

which establishes the desired claim. □

Lemma 4.14. *Suppose that A, B are two symmetric positive semidefinite matrices, and that A is nonsingular. If $D^\mathsf{T} \in \mathbf{R}^{d\times d}$ has range included in the range of B, then*

$$(I-D)A(I-D)^\mathsf{T} + DB^\dagger D^\mathsf{T} \succcurlyeq (A^{-1}+B)^{-1}.$$

Moreover equality holds with the choice $D = (A^{-1}+B)^{-1}B$.

Proof. Let $x \in \mathbf{R}^d$ and note that if $y \coloneqq D^\mathsf{T} x$, then

$$\begin{aligned} x^\mathsf{T}\Big[(I-D)A(I-D)^\mathsf{T} + DB^\dagger D^\mathsf{T}\Big]x &= (x-y)^\mathsf{T} A(x-y) + y^\mathsf{T} B^\dagger y \\ &\geqslant x^\mathsf{T}(A^{-1}+B)^{-1}x, \end{aligned}$$

where the final inequality follows from Lemma 4.13, since y lies in the range of B. As the inequality holds for arbitrary $x \in \mathbf{R}^d$, we have established the desired matrix inequality. To see the attainment at $D = (A^{-1}+B)^{-1}B$, first note that $D^\mathsf{T} = B(A^{-1}+B)^{-1}$. Therefore the range of D^T is exactly the range of B. Additionally, since $I - D = (A^{-1}+B)^{-1}A^{-1}$, we have

$$(I-D)A(I-D)^\mathsf{T} + DB^\dagger D^\mathsf{T} = (A^{-1}+B)^{-1}(A^{-1}+BB^\dagger B)(A^{-1}+B)^{-1} = (A^{-1}+B)^{-1},$$

as required. □

We are now in a situation to prove Proposition 4.3.

Proof of Proposition 4.3 From display (4.53), to establish the claim, it suffices to lower bound the following matrix in the semidefinite ordering,

$$(C(T_\xi)T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi - I_d)\Omega(C(T_\xi)T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi - I_d)^\mathsf{T} + C(T_\xi)T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi C(T_\xi)^\mathsf{T}. \quad (4.86)$$

This matrix can be written as $(I-D)\Omega(I-D)^\mathsf{T} + DB^\dagger D^\mathsf{T}$ where we defined

$$B \coloneqq T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi, \quad \text{and,} \quad D \coloneqq C(T_\xi)T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi.$$

Evidently, the range of D^T is included in the range of B, and so it follows from Lemma 4.14 that the matrix in equation (4.86) is lower bounded in the semidefinite ordering by

$$(\Omega^{-1} + T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi)^{-1}. \quad (4.87)$$

Moreover, Lemma 4.14 also demonstrates this is established by taking

$$D = (\Omega^{-1} + T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi)^{-1}T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi,$$

which arises from taking $C(T_\xi) = (\Omega^{-1} + T_\xi^\mathsf{T}\Sigma_w^{-1}T_\xi)^{-1}$, as claimed. Evaluating this lower bound matrix (4.87) in (4.53) establishes equality (4.54).

4.6.6.4 Proof of equation (4.55d)

Let us formally state our claim, equivalent to equation (4.55d), as a lemma.

Lemma 4.15. *Let $\mathcal{K}_+(\varrho, K_c)$ denote the subset of nonsingular matrices in $\mathcal{K}(\varrho, K_c)$—that is, the set $\{\Omega \succ 0 : \Omega \in \mathcal{K}(\varrho, K_c)\}$. Then, we have*

$$\sup_{\Omega\in\mathcal{K}(\varrho,K_c)} \inf_C r(\widehat{\theta}_C, \Omega) = \sup_{\Omega\in\mathcal{K}_+(\varrho,K_c)} \inf_C r(\widehat{\theta}_C, \Omega).$$

We prove this claim now. Evidently, since $\mathcal{K}_+(\varrho, K_c) \subset \mathcal{K}(\varrho, K_c)$ it suffices to show that the lefthand side is less than or equal to the righthand side. To begin, we note that for each $\lambda > 0$, we have

$$\sup_{\Omega\in\mathcal{K}(\varrho,K_c)} \inf_C r(\widehat{\theta}_C, \Omega) \overset{(a)}{\leqslant} \sup_{\Omega\in\mathcal{K}(\varrho,K_c)} \inf_C r(\widehat{\theta}_C, \Omega + \tfrac{(\varrho+\lambda)^2-\varrho^2}{d}K_c) \overset{(b)}{\leqslant} \sup_{\Omega\in\mathcal{K}_+(\varrho+\lambda,K_c)} \inf_C r(\widehat{\theta}_C, \Omega) =: f(\lambda).$$

Inequality (a) above follows since $r(\widehat{\theta}_C, \Omega) \leqslant r(\widehat{\theta}_C, \Omega')$ for any $\Omega \preccurlyeq \Omega'$—this follows immediately from display (4.53). Here we have taken $\Omega' \coloneqq \Omega + \frac{(\varrho+\lambda)^2-\varrho^2}{d}K_c \succcurlyeq \Omega$. Inequality (b) then follows by noting that Ω' is symmetric positive (strictly) definite, and $\mathbf{Tr}(K_c^{-1/2}\Omega' K_c^{-1/2}) \leqslant (\varrho+\lambda)^2$, since $\Omega \in \mathcal{K}(\varrho, K_c)$. Since the displayed relation above holds for any $\lambda > 0$, it suffices to show that

$$\inf_{\lambda>0} f(\lambda) = f(0). \quad (4.88)$$

By Proposition 4.3, we have

$$\begin{aligned}
f(\lambda) &= \sup_{\Omega} \Big\{ \mathbf{E}\,\mathbf{Tr}\left(K_e^{1/2}(\Omega^{-1} + T_\xi^{\mathsf{T}}\Sigma_w^{-1}T_\xi)^{-1}K_e^{1/2}\right) : \\
&\qquad\qquad \Omega \succ 0, \mathbf{Tr}(K_c^{-1/2}\Omega K_c^{-1/2}) \leqslant (\varrho+\lambda)^2 \Big\} \\
&= \sup_{\Omega} \Big\{ \mathbf{E}\,\mathbf{Tr}\left(K_e^{1/2}((\tfrac{\varrho+\lambda}{\varrho})^{-2}\Omega^{-1} + T_\xi^{\mathsf{T}}\Sigma_w^{-1}T_\xi)^{-1}K_e^{1/2}\right) : \\
&\qquad\qquad \Omega \succ 0, \mathbf{Tr}(K_c^{-1/2}\Omega K_c^{-1/2}) \leqslant \varrho^2 \Big\} \\
&\leqslant \Big(\frac{\varrho+\lambda}{\varrho}\Big)^2 \sup_{\Omega} \Big\{ \mathbf{E}\,\mathbf{Tr}\left(K_e^{1/2}((\Omega^{-1} + T_\xi^{\mathsf{T}}\Sigma_w^{-1}T_\xi)^{-1}K_e^{1/2}\right) : \\
&\qquad\qquad \Omega \succ 0, \mathbf{Tr}(K_c^{-1/2}\Omega K_c^{-1/2}) \leqslant \varrho^2 \Big\} \\
&= \Big(\frac{\varrho+\lambda}{\varrho}\Big)^2 f(0).
\end{aligned}$$

Hence we have established the sandwich relation

$$f(0) \leqslant f(\lambda) \leqslant \Big(\frac{\varrho+\lambda}{\varrho}\Big)^2 f(0), \qquad \text{for all } \lambda > 0.$$

Note that $f(0) \leqslant f(\lambda') \leqslant f(\lambda)$ whenever $0 < \lambda' \leqslant \lambda$. Thus, $\inf_{\lambda>0} = \lim_{\lambda\to 0^+} f(\lambda) = f(0)$, which establishes (4.88), completing the proof of the claim.

4.6.7 Deferred proofs from Section 4.4.2

In this section, we collect proofs of the results underlying the argument establishing our lower bound in Section 4.4.2.

4.6.7.1 Proof of Lemma 4.3

By parameterizing $\theta^\star = K_e^{-1/2}\eta^\star$, we have

$$\begin{aligned}
&\mathfrak{M}^{\mathrm{G}}(T, \mathbb{P}, \Sigma_w, \varrho, K_c, K_e) \\
&\quad= \inf_{\widehat{\eta}} \sup_{\eta^\star \in \Theta(\varrho^2 K_e^{1/2} K_c K_e^{1/2})} \mathbf{E}_{\xi, w\sim\mathsf{N}(0,I_n)} \left[\left\| \widehat{\eta}(T_\xi K_e^{-1/2}, T_\xi K_e^{-1/2}\eta^\star + \Sigma_w^{1/2} w) - \eta^\star \right\|_2^2 \right] \\
&\quad= \inf_{\widehat{\eta}} \sup_{\eta^\star \in \Theta(\varrho^2 K_e^{1/2} K_c K_e^{1/2})} \mathbf{E}_{\xi, z\sim\mathsf{N}(0,I_{r(\xi)})} \left[\left\| \widehat{\eta}(Q_\xi, Q_\xi\eta^\star + V_\xi \Lambda_\xi^{1/2} z) - \eta^\star \right\|_2^2 \right] \qquad (4.89) \\
&\quad= \inf_{\widehat{\eta}} \sup_{\eta^\star \in \Theta(\varrho^2 K_e^{1/2} K_c K_e^{1/2})} \mathbf{E}_{\omega\sim\widetilde{\mathbb{P}}, z\sim\mathsf{N}(0,I_{r(\xi)})} \left[\left\| \widehat{\eta}(\omega, V_\xi V_\xi^{\mathsf{T}}\eta^\star + V_\xi \Lambda_\xi^{-1/2} z) - \eta^\star \right\|_2^2 \right] \qquad (4.90) \\
&\quad= \mathfrak{M}^{\mathrm{G}}_{\mathrm{red}}(\widetilde{\mathbb{P}}, \varrho^2 K_e^{1/2} K_c K_e^{1/2}).
\end{aligned}$$

We justify some of the relations in the display above. Since the density of $v = T_\xi K_e^{-1/2}\eta^\star + \Sigma_w^{1/2} w$ is, up to constants independent of $\eta^\star$, proportional to

$$\exp\Big(-\frac{1}{2}\big\{\langle \eta^\star, K_e^{-1/2} T_\xi^\mathsf{T} \Sigma_w^{-1} T_\xi K_e^{-1/2} \eta^\star\rangle - 2\langle v, \Sigma_w^{-1} T_\xi K_e^{-1/2}\eta^\star\rangle\big\}\Big),$$

factorization arguments imply $Q_\xi \coloneqq K_e^{-1/2} T_\xi^\mathsf{T} \Sigma_w^{-1} T_\xi K_e^{-1/2}$ and $v' \coloneqq K_e^{-1/2} T_\xi^\mathsf{T} \Sigma_w^{-1} v$ are sufficient statistics for $\eta^\star$. Note that v' is distributed $\mathsf{N}(Q_\xi \eta^\star, Q_\xi)$. Thus, as consequence of the Rao-Blackwell theorem, any minimax optimal estimator is a function of (Q_ξ, v'), and hence display (4.89) follows. Similarly, any optimal estimator function is a function of any bijective function of (Q_ξ, v'). Evidently one can construct Q_ξ from $\omega \coloneqq (r(\xi), V_\xi, \Lambda_\xi)$, and vice versa. On the other hand, v' lies in the range of $G(\xi) \coloneqq K_e^{-1/2} T_\xi^\mathsf{T} \Sigma_w^{-1/2}$, which is the same as the range of $G(\xi)G(\xi)^\mathsf{T} = Q_\xi$; consequently one may replace v' with $Q_\xi^\dagger v' \equiv V_\xi(\Lambda_\xi)^{-1} V_\xi^\mathsf{T} v'$, which is distributed $\mathsf{N}\big(V_\xi V_\xi^\mathsf{T}\eta^\star, V_\xi(\Lambda_\xi)^{-1} V_\xi^\mathsf{T}\big)$, and so that display (4.90) follows.

4.6.7.2 Proof of Lemma 4.4

In this argument, we use the notation $B(\widehat{\eta}, \pi \mid \omega)$ to denote the Bayes risk of estimator $\widehat{\eta}$, conditional on ω, for the original observation Υ. Formally, it is the expectation $\mathbf{E}[\|\widehat{\eta}(\Upsilon) - \eta\|_2^2]$, where the expectation is over $\Upsilon \sim \mathsf{N}\big(VV^\mathsf{T}, V\Lambda^{-1}V^\mathsf{T}\big)$.

The main observation is that if we consider the projection of Υ_λ onto the range of V, we will recover a random variable with the same distribution as Υ, and therefore the risks are the same. Formally, let $\widehat{\eta}$ be any estimator which is constant over the fibers of the operator VV^T. Equivalently, it can be written

$$\widehat{\eta}(y) = \widehat{\eta}_0(VV^\mathsf{T} y), \qquad \text{for some measurable } \widehat{\eta}_0.$$

Let this class of estimators be denoted by $\mathcal{E}_V$. Then we evidently have

$$B_\lambda(\pi \mid \omega) \leqslant \inf_{\widehat{\eta} \in \mathcal{E}_V} B_\lambda(\widehat{\eta}, \pi \mid \omega). \tag{4.91}$$

To complete the proof of the claim, we claim that

$$B_\lambda(\widehat{\eta}, \pi \mid \omega) = B(\widehat{\eta}, \pi \mid \omega), \qquad \text{for any } \widehat{\eta} \in \mathcal{E}_V.$$

This follows immediately from the fact that $VV^\mathsf{T}\Upsilon_\lambda = \Upsilon$ with probability 1. We note that combination with (4.91) furnishes the claim, since it implies that

$$B_\lambda(\pi \mid \omega) \leqslant \inf_{\widehat{\eta} \in \mathcal{E}_V} B(\widehat{\eta}, \pi \mid \omega) = B(\pi \mid \omega).$$

The final equality occurs since for any measurable estimator $\widehat{\eta} \notin \mathcal{E}_V$, we can define $\widehat{\eta}_V(y) = \widehat{\eta}(VV^\mathsf{T} y)$, and since $\Upsilon = VV^\mathsf{T}\Upsilon$ with probability 1, and therefore $B(\widehat{\eta}_V, \pi \mid \omega) = B(\widehat{\eta}, \pi \mid \omega)$, which establishes this claim.

4.6.7.3 Proof of Lemma 4.5

Let $\widehat{\eta}_\pi$ denote the posterior mean $y \mapsto \mathbf{E}[\eta \mid \Upsilon_\lambda = y]$. Then, as the posterior mean $\widehat{\eta}_\pi$ minimizes the Bayes risk $\widehat{\eta} \mapsto B_\lambda(\widehat{\eta}, \pi \mid \omega)$ over all measurable estimators $\widehat{\eta}$, it suffices to compute the risk of $\widehat{\eta}_\pi$. Note that, by definition of conditional expectation, we have

$$\widehat{\eta}_\pi(y) = \frac{1}{p(y)} \int \eta \, p(y \mid \eta) \, \pi(\, \mathrm{d}\eta).$$

We now compute the derivative of $p(y)$. Exchanging integration and differentiation,[8]

$$\Sigma_\lambda \nabla p(y) = \int (X_\lambda \eta - y) \, p(y \mid \eta) \, \pi(\, \mathrm{d}\eta).$$

Therefore, we conclude that

$$\widehat{\eta}_\pi(y) = X_\lambda^{-1} \Big(y + \Sigma_\lambda \nabla \log p(y) \Big).$$

Finally, to compute risk of the posterior mean $\widehat{\eta}_\pi(\Upsilon_\lambda) \coloneqq \mathbf{E}[\eta \mid \Upsilon_\lambda]$, we add and subtract the observation $X_\lambda^{-1} \Upsilon_\lambda$, and find that

$$\begin{aligned} &\mathbf{E}_{(\eta, \Upsilon_\lambda)} \Big[(\eta - \widehat{\eta}_\pi(\Upsilon_\lambda)) \otimes (\eta - \widehat{\eta}_\pi(\Upsilon_\lambda)) \Big] \\ &\qquad = X_\lambda^{-1} \Sigma_\lambda X_\lambda^{-1} - X_\lambda^{-1} \Sigma_\lambda \, \mathbf{E}[\nabla \log p(\Upsilon_\lambda) \otimes \nabla \log p(\Upsilon_\lambda)] \Sigma_\lambda X_\lambda^{-1}. \end{aligned}$$

Identifying the Fisher information in the display above, factoring the expression, and taking the trace yields the desired result.

4.6.7.4 Proof of Lemma 4.6

Note that $\pi_{\tau,\Pi}$ is evidently absolutely continuous with respect to Lebesgue measure. In particular, on the interior of $\Theta(K)$, $\pi_{\tau,\Pi}$ and $\pi_{\tau,\Pi}^{\mathrm{G}}$ have the same Lebesgue density up to rescaling by $\pi_{\tau,\Pi}^{\mathrm{G}}(\Theta(K))$. Denote this density by $f_{\tau,\Pi}$. Therefore, we have

$$\begin{aligned} \mathcal{I}\left(\pi_{\tau,\Pi}^{\mathrm{G}}\right) &= \mathbf{E}_{\eta \sim \pi_{\tau,\Pi}^{\mathrm{G}}} \mathbf{1}_{\Theta(K)}(\eta) \nabla \log f_{\tau,\Pi}(\eta) \otimes \nabla \log f_{\tau,\Pi}(\eta) \\ &\quad + \mathbf{E}_{\eta \sim \pi_{\tau,\Pi}^{\mathrm{G}}} \mathbf{1}_{\Theta(K)^c}(\eta) \nabla \log f_{\tau,\Pi}(\eta) \otimes \nabla \log f_{\tau,\Pi}(\eta) \\ &\succcurlyeq \mathbf{E}_{\eta \sim \pi_{\tau,\Pi}^{\mathrm{G}}} \mathbf{1}_{\Theta(K)}(\eta) \nabla \log f_{\tau,\Pi}(\eta) \otimes \nabla \log f_{\tau,\Pi}(\eta) \\ &= \pi_{\tau,\Pi}^{\mathrm{G}}(\Theta(K)) \mathcal{I}\left(\pi_{\tau,\Pi}\right). \end{aligned}$$

[8] This is valid since $y \mapsto p(y \mid \eta)$ is differentiable for each η, and for each y, we have $\eta \mapsto p(y \mid \eta)$ and $\eta \mapsto \nabla_y p(y \mid \eta) = \Sigma_\lambda^{-1}(X_\lambda \eta - y)$ are π-integrable (since $0 \leqslant p(y \mid \eta) \leqslant 1$, and the gradient is an affine function of η).

The final equality arises since the boundary of $\Theta(K)$ has Lebesgue measure zero. Using the well known relation $\mathcal{I}\left(\pi^{\mathrm{G}}_{\tau,\Pi}\right) = (\tau^2\Pi)^{-1}$ [72, Example 6.3], the above display implies that

$$\mathcal{I}\left(\pi_{\tau,\Pi}\right)^{-1} \succcurlyeq \pi^{\mathrm{G}}_{\tau,\Pi}(\Theta(K))\tau^2\Pi = \tau^2(1 - \pi^{\mathrm{G}}_{\tau,\Pi}(\Theta(K)^c))\Pi.$$

To ensure that $\eta \sim \pi^{\mathrm{G}}_{\tau,\Pi}$ lies in $\Theta(K)$ with decent probability, we take Π to satisfy the relation $\mathbf{Tr}(K^{-1}\Pi) \leqslant 1$. Then defining

$$c(\tau, \Pi) \coloneqq \tau^2(1 - \pi^{\mathrm{G}}_{\tau,\Pi}(\Theta(K)^c)),$$

completes the proof of the claim.

4.6.7.5 Proof of Lemma 4.7

Fix $\Pi \succ 0$ such that $\mathbf{Tr}(\Pi^{1/2}K^{-1}\Pi^{1/2}) \leqslant 1$. Let $\lambda = (\lambda_1, \ldots, \lambda_d)$ denote the eigenvalues of $\Pi^{1/2}K^{-1}\Pi^{1/2}$. The vector satisfies the inequalities $\lambda \succ 0, \lambda^{\mathsf{T}}\mathbf{1} \leqslant 1$. Moreover, by the rotational invariance of the Gaussian, we have for $g \sim \mathsf{N}(0, I_d)$, that

$$\pi^{\mathrm{G}}_{\tau,\Pi}(\Theta(K)^c) = \mathbf{P}\left\{\tau^2 g^{\mathsf{T}}\Pi^{1/2}K^{-1}\Pi^{1/2}g > 1\right\} = \mathbf{P}\left\{\tau^2\sum_{i=1}^{d}\lambda_i g_i^2 > 1\right\}.$$

Let us make the choice $\tau^2 = 1/2$. Then, note for any $\lambda \succ 0, \lambda^{\mathsf{T}}\mathbf{1} \leqslant 1$, by Markov's inequality,

$$\mathbf{P}\left\{\sum_{i=1}^{d}\lambda_i g_i^2 > 2\right\} \leqslant \frac{\sum_{i=1}^{d}\lambda_i\,\mathbf{E}[g_i^2]}{2} = \frac{1}{2}.$$

Hence, using this bound in the definition of $c(\tau, \Pi)$, we find

$$c_\ell(K) \geqslant \inf_{\lambda\succ 0, \lambda^{\mathsf{T}}\mathbf{1}\leqslant 1} c(1/2, \mathbf{diag}(\lambda)) \geqslant \frac{1}{4},$$

which completes the proof of the claim.

Bibliography

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

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

[3] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[4] J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794, 2011.

[5] E. N. Belitser and B. Y. Levit. On minimax filtering over ellipsoids. *Math. Methods Statist.*, 4(3):259–273, 1995.

[6] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[7] S. Ben-David, T. Lu, T. Luu, and D. Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136, 2010.

[8] J. Berkson. Are there two regressions? *Journal of the American Statistical Association*, 45:164–180, 1950.

[9] J. C. Berry. Minimax estimation of a bounded normal mean vector. *J. Multivariate Anal.*, 35(1):130–139, 1990.

[10] R. Bhatia. *Positive definite matrices*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2007.

[11] P. J. Bickel. Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Statist.*, 9(6):1301–1309, 1981.

[12] L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3:203–268, 2001.

[13] J. Blitzer, M. Dredze, and F. Pereira. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[14] J. M. Borwein and D. Zhuang. On Fan's minimax theorem. *Math. Programming*, 34(2):232–234, 1986.

[15] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

[16] L. Breiman and D. Freedman. How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.*, 78(381):131–136, 1983.

[17] L. D. Brown. Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.*, 42:855–903, 1971.

[18] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[19] T. T. Cai and H. Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100–128, 2021.

[20] E. J. Candes and M. A. Davenport. How well can we estimate a sparse vector? *Applied and Computational Harmonic Analysis*, 34(2):317–323, 2013.

[21] E. J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.*, 37(5A):2145–2177, 2009.

[22] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[23] R. J. Carroll, D. Ruppert, and L. A. Stefanski. *Measurement Error in Nonlinear Models*. Chapman and Hall, 1995.

[24] G. Casella and W. E. Strawderman. Estimating a bounded normal mean. *Ann. Statist.*, 9(4):870–878, 1981.

[25] M. T. Chao and W. E. Strawderman. Negative moments of positive random variables. *J. Amer. Statist. Assoc.*, 67(338):429–431, 1972.

[26] X. Chen and Y. Yang. Hanson–Wright inequality in Hilbert spaces with application to K-means clustering for non-Euclidean data. *Bernoulli*, 27(1):586–614, 2021.

[27] C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems*, volume 23, pages 442–450, 2010.

[28] C. Cortes, M. Mohri, and A. M. Medina. Adaptation based on generalized discrepancy. *J. Mach. Learn. Res.*, 20(1):1–30, 2019.

[29] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49, 2002.

[30] A. S. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the Lasso. *Bernoulli*, 23(1):552–581, 2017.

[31] L. H. Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1–37, 2016.

[32] D. L. Donoho. Statistical estimation and optimal recovery. *Ann. Statist.*, 22(1):238–270, 1994.

[33] D. L. Donoho and I. M. Johnstone. Minimax risk over ℓ_p-balls for ℓ_q-error. *Probab. Theory Related Fields*, 99(2):277–303, 1994.

[34] D. L. Donoho, R. C. Liu, and B. MacGibbon. Minimax risk over hyperrectangles, and implications. *Ann. Statist.*, 18(3):1416–1437, 1990.

[35] R. Durrett. *Probability—theory and examples*, volume 49 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019.

[36] B. Eyre, E. Creager, D. Madras, V. Papyan, and R. Zemel. Out of the ordinary: Spectrally adapting regression for covariate shift. *arXiv preprint arXiv:2312.17463*, 2023.

[37] J. Fan, R. Li, C.-H. Zhang, and H. Zou. *Statistical foundations of data science*. CRC press, 2020.

[38] D. Fourdrinier, W. E. Strawderman, and M. T. Wells. *Shrinkage estimation*. Springer Series in Statistics. Springer, Cham, 2018.

[39] R. Foygel and N. Srebro. Fast-rate and optimistic-rate error bounds for ℓ_1-regularized regression. *arXiv preprint arXiv:1108.0373*, 2011.

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

[41] S. Gaïffas. On pointwise adaptive curve estimation based on inhomogeneous data. *ESAIM Probab. Stat.*, 11:344–364, 2007.

[42] S. Gaïffas. Sharp estimation in sup norm with random design. *Statist. Probab. Lett.*, 77(8):782–794, 2007.

[43] S. Gaïffas. Uniform estimation of a signal based on inhomogeneous data. *Statist. Sinica*, 19(2):427–447, 2009.

[44] J. Ge, S. Tang, J. Fan, C. Ma, and C. Jin. Maximum likelihood estimation is all you need for well-specified covariate shift. *arXiv preprint arXiv:2311.15961*, 2023.

[45] P. Germain, A. Habrard, F. Laviolette, and E. Morvant. A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In *International Conference on Machine Learning*, pages 738–746, 2013.

[46] D. Gogolashvili. Importance weighting correction of regularized least-squares for covariate and target shifts, 2022.

[47] D. Gogolashvili, M. Zecchin, M. Kanagawa, M. Kountouris, and M. Filippone. When is importance weighting correction needed for covariate shift adaptation?, 2023.

[48] A. Goldenshluger and A. Tsybakov. Adaptive prediction and estimation in linear regression with infinitely many parameters. *Ann. Statist.*, 29(6):1601–1619, 2001.

[49] A. Goldenshluger and A. Tsybakov. Optimal prediction for linear regression with infinitely many parameters. *J. Multivariate Anal.*, 84(1):40–60, 2003.

[50] G. K. Golubev. Quasilinear estimates for signals in L_2. *Problemy Peredachi Informatsii*, 26(1):19–24, 1990.

[51] S. Greenberg and M. Mohri. Tight lower bound on the probability of a binomial exceeding its expectation. *Statist. Probab. Lett.*, 86:91–98, 2014.

[52] H. Guan and M. Liu. Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, pages 1–1, 2021.

[53] A. Guillou and N. Klutchnikoff. Minimax pointwise estimation of an anisotropic regression function with unknown density of the design. *Math. Methods Statist.*, 20(1):30–57, 2011.

[54] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[55] G. Halász. Statistical interpolation. *Mat. Lapok*, 23:71–87 (1973), 1972.

[56] A. Hassan, R. Damper, and M. Niranjan. On acoustic emotion recognition: Compensating for covariate shift. *IEEE Trans. Audio Speech Lang. Process.*, 21(7):1458–1468, 2013.

[57] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the Lasso and generalizations*. CRC press, 2015.

[58] D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. *Found. Comput. Math.*, 14(3):569–600, 2014.

[59] D. Hsu and S. Sabato. Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.*, 17:Paper No. 18, 40, 2016.

[60] I. A. Ibragimov and R. Z. Khas′ minskiĭ. Nonparametric regression estimation. *Dokl. Akad. Nauk SSSR*, 252(4):780–784, 1980.

[61] J. Jiang and C. Zhai. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[62] I. M. Johnstone. Gaussian estimation: Sequence and wavelet models. Book manuscript, September 2019.

[63] A. Juditsky and A. Nemirovski. Near-optimality of linear recovery in Gaussian observation scheme under $\|\cdot\|_2^2$-loss. *Ann. Statist.*, 46(4):1603–1629, 2018.

[64] M. Kac, W. L. Murdock, and G. Szegö. On the eigenvalues of certain Hermitian forms. *J. Rational Mech. Anal.*, 2:767–800, 1953.

[65] J. A. Kelner, F. Koehler, R. Meka, and D. Rohatgi. On the power of preconditioning in sparse linear regression. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science—FOCS 2021*, pages 550–561. 2022.

[66] T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005.

[67] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[68] S. Kpotufe. Lipschitz density-ratios, structured data, and data-driven tuning. In *Artificial Intelligence and Statistics*, pages 1320–1328, 2017.

[69] S. Kpotufe and G. Martinet. Marginal singularity and the benefits of labels in covariate-shift. *Ann. Statist.*, 49(6):3299–3323, 2021.

[70] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

[71] G. Lecué and S. Mendelson. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534, 2016.

[72] E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 1998.

[73] Q. Lei, W. Hu, and J. Lee. Near-optimal linear regression under distribution shift. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proc. Mach. Learn. Res.*, pages 6164–6174. PMLR, 18–24 Jul 2021.

[74] Y. Li, H. Kambara, Y. Koike, and M. Sugiyama. Application of covariate shift adaptation techniques in brain–computer interfaces. *IEEE Trans. Biomed. Eng.*, 57(6):1318–1324, 2010.

[75] M. Liu, Y. Zhang, K. P. Liao, and T. Cai. Augmented transfer regression learning with semi-non-parametric nuisance models, 2020.

[76] G. Lugosi and S. Mendelson. Mean estimation and regression under heavy-tailed distributions: a survey. *Found. Comput. Math.*, 19(5):1145–1190, 2019.

[77] C. Ma, R. Pathak, and M. J. Wainwright. Optimally tackling covariate shift in RKHS-based nonparametric regression, 2022.

[78] S. Maity, Y. Sun, and M. Banerjee. Minimax optimal approaches to the label shift problem. *arXiv preprint arXiv:2003.10443*, 2020.

[79] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

[80] Y. Mansour, M. Mohri, and A. Rostamizadeh. Multiple source adaptation and the rényi divergence. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 367–374. AUAI Press, 2009.

[81] E. Marchand. Estimation of a multivariate mean with constraints on the norm. *Canad. J. Statist.*, 21(4):359–366, 1993.

[82] E. Marchand and W. E. Strawderman. Estimation in restricted parameter spaces: a review. In *A festschrift for Herman Rubin*, volume 45 of *IMS Lecture Notes Monogr. Ser.*, pages 21–44. Inst. Math. Statist., Beachwood, OH, 2004.

[83] A. A. Melkman and Y. Ritov. Minimax estimation of the mean of a general distribution when the parameter space is restricted. *Ann. Statist.*, 15(1):432–442, 1987.

[84] S. Mendelson. Learning without concentration. *J. ACM*, 62(3):Art. 21, 25, 2015.

[85] S. Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 127:111–119, 2017.

[86] M. Mohri and A. M. Medina. New analysis and algorithm for learning with drifting distributions. In *International Conference on Algorithmic Learning Theory*, pages 124–138. Springer, 2012.

[87] J. Mourtada. *Contributions à l'apprentissage statistique : estimation de densité, agrégation d'experts et forêts aléatoires*. Theses, Institut Polytechnique de Paris, June 2020.

[88] J. Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *Ann. Statist.*, page to appear, 2022.

[89] J. Mourtada and L. Rosasco. An elementary analysis of ridge regression with random design. *C. R. Math. Acad. Sci. Paris*, page to appear, 2022.

[90] M. Mousavi Kalan, Z. Fabian, S. Avestimehr, and M. Soltanolkotabi. Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1959–1969. Curran Associates, Inc., 2020.

[91] E. A. Nadaraya. On estimating regression. *Theory Probab. Appl.*, 9(1):141–142, 1964.

[92] R. I. Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probab. Theory Related Fields*, 166(3-4):1175–1194, 2016.

[93] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[94] R. Pathak, C. Ma, and M. Wainwright. A new similarity measure for covariate shift with applications to nonparametric regression. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17517–17530. PMLR, 17–23 Jul 2022.

[95] R. Pathak, M. J. Wainwright, and L. Xiao. Noisy recovery from random linear observations: Sharp minimax rates under elliptical constraints, 2023.

[96] M. S. Pinsker. Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Inf. Transm.*, 16(2):52–68, 1980.

[97] D. Pollard. *Convergence of stochastic processes*. Springer Science & Business Media, 2012.

[98] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

[99] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory*, 57(10):6976–6994, 2011.

[100] H. W. Reeve, T. I. Cannings, and R. J. Samworth. Adaptive transfer learning. *arXiv preprint arXiv:2106.04455*, 2021.

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

[102] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision – ECCV 2010*, pages 213–226, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[103] J. Schmidt-Hieber and P. Zamolodtchikov. Local convergence rates of the nonparametric least squares estimator with applications to transfer learning. *Bernoulli*, 30(3):1845–1877, 2024.

[104] B. Schölkopf, A. J. Smola, F. Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[105] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference*, 90(2):227–244, 2000.

[106] M. Simchowitz, A. Ajay, P. Agrawal, and A. Krishnamurthy. Statistical learning under heterogenous distribution shift, 2023.

[107] B. Simon. *Loewner's theorem on monotone matrix functions*, volume 354 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham, 2019.

[108] C. Stein. Multiple regression. In *Contributions to probability and statistics*, pages 424–443. Stanford Univ. Press, Stanford, Calif., 1960.

[109] I. Steinwart and C. Scovel. Mercer's theorem on general domains: on the interaction between measures, kernels, and RKHSs. *Constr. Approx.*, 35(3):363–417, 2012.

[110] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 1982.

[111] M. Sugiyama and M. Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

[112] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, pages 1433–1440, 2008.

[113] M. Sugiyama, T. Suzuki, and T. Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[114] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[115] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

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

[117] S. van de Geer. On tight bounds for the Lasso. *J. Mach. Learn. Res.*, 19:Paper No. 46, 48, 2018.

[118] V. N. Vapnik. *The nature of statistical learning theory*. Statistics for Engineering and Information Science. Springer-Verlag, New York, second edition, 2000.

[119] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[120] M. J. Wainwright. *High-dimensional statistics*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019. A non-asymptotic viewpoint.

[121] K. Wang. Pseudo-labeling for kernel ridge regression under covariate shift, 2023.

[122] G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

[123] Y. Yang, M. Pilanci, and M. J. Wainwright. Randomized sketches for kernels: fast and optimal nonparametric regression. *Ann. Statist.*, 45(3):991–1023, 2017.

[124] B. Yu. Assouad, Fano and Le Cam. In *Festschrift in Honor of L. Le Cam on his 70th Birthday*. Springer New York, 1993.

[125] R. Zamir. A proof of the Fisher information inequality via a data processing argument. *IEEE Trans. Inform. Theory*, 44(3):1246–1250, 1998.

[126] X. Zhang, J. Blanchet, S. Ghosh, and M. S. Squillante. A class of geometric structures in transfer learning: Minimax bounds and optimality. In *International Conference on Artificial Intelligence and Statistics*, pages 3794–3820. PMLR, 2022.