The Annals of Statistics
2023, Vol. 51, No. 2, 738–761
https://doi.org/10.1214/23-AOS2268

OPTIMALLY TACKLING COVARIATE SHIFT IN RKHS-BASED NONPARAMETRIC REGRESSION

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

[1]*Department of Statistics, University of Chicago,* [a]*congm@uchicago.edu*
[2]*Department of Electrical Engineering and Computer Sciences, University of California, Berkeley,* [b]*pathakr@berkeley.edu*
[3]*EECS and Mathematics, Statistics and Data Science Center, Massachusetts Institute of Technology,* [c]*wainwrigwork@gmail.com*

We study the covariate shift problem in the context of nonparametric regression over a reproducing kernel Hilbert space (RKHS). We focus on two natural families of covariate shift problems defined using the likelihood ratios between the source and target distributions. When the likelihood ratios are uniformly bounded, we prove that the kernel ridge regression (KRR) estimator with a carefully chosen regularization parameter is minimax rate-optimal (up to a log factor) for a large family of RKHSs with regular kernel eigenvalues. Interestingly, KRR does not require full knowledge of the likelihood ratio apart from an upper bound on it. In striking contrast to the standard statistical setting without covariate shift, we also demonstrate that a naïve estimator, which minimizes the empirical risk over the function class, is strictly suboptimal under covariate shift as compared to KRR. We then address the larger class of covariate shift problems where likelihood ratio is possibly unbounded yet has a finite second moment. Here, we propose a reweighted KRR estimator that weights samples based on a careful truncation of the likelihood ratios. Again, we are able to show that this estimator is minimax optimal, up to logarithmic factors.

1. Introduction. A widely adopted assumption in supervised learning [6, 20] 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 [5, 8], 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 [7], 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 paper, 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 [19]. 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 [6].

This paper 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

Received May 2022; revised October 2022.
MSC2020 subject classifications. Primary 62C20; secondary 62G08.
Key words and phrases. Covariate shift, nonparametric regression, reproducing kernel Hilbert spaces, kernel ridge regression, transfer learning.

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

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?

1.1. *Contributions and overview.* We address the aforementioned theoretical questions regarding covariate shift in the context of nonparametric regression over reproducing kernel Hilbert spaces (RKHSs) [17]. 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 paper, 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 3.1. Instantiations of this general bound on various RKHSs with regular eigenvalues are provided in Corollary 3.2. It is also shown in Theorem 3.3 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 3.4. 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 [22]. 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 paper, 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 4.1 that this estimator is rate-optimal over this larger class of covariate shift problems.

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 [19] and the survey [12] for additional references.

Shimodaira [18] 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 [4] 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 [14]), 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 suboptimal—with a rate of $\sqrt{V^2D/n}$ compared to our optimal rate V^2D/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 [9] 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 [13] 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 [9] and Pathak et al. [13] focused on the class of Hölder continuous functions, while we focus on RKHSs. This leads to drastically different optimal estimators. Schmidt-Hieber and Zamolodtchikov [16] 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. [1] analyzed posterior shift and proposed an optimal k-nearest-neighbor estimator. Maity et al. [11] conducted the minimax analysis for the label shift problem. Recently, Reeve et al. [15] studied the general distribution shift problem (also known as transfer learning) which allows both covariate shift and posterior shift.

Notation. Throughout the paper, we use c, c', c_1, c_2 to denote universal constants, which may vary from line to line. Also, $f(n) \lesssim h(n)$ (or $f(n) = O(h(n))$) means $|f(n)| \le c_1|h(n)|$ for some constant $c_1 > 0$, $f(n) \gtrsim h(n)$ means $|f(n)| \ge c_2|h(n)|$ for some constant $c_2 > 0$, $f(n) \asymp h(n)$ means $c_2|h(n)| \le |f(n)| \le c_1|h(n)|$ for some constants $c_1, c_2 > 0$, and $f(n) = o(h(n))$ means $\lim_{n\to\infty} f(n)/h(n) = 0$.

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

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 $f^\star(x) := \mathbb{E}[Y|X = x]$. 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 $(Y|X = x_i)$. 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:

$$\|\widehat{f} - f^\star\|_P^2 := \mathbb{E}_{X\sim P}\big(\widehat{f}(X) - f^\star(X)\big)^2 = \int_{\mathcal{X}} \big(\widehat{f}(x) - f^\star(x)\big)^2 p(x)\,dx, \tag{1a}$$

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

$$\|\widehat{f} - f^\star\|_Q^2 = \mathbb{E}_{X\sim Q}\big(\widehat{f}(X) - f^\star(X)\big)^2 = \int_{\mathcal{X}} \big(\widehat{f}(x) - f^\star(x)\big)^2 q(x)\,dx \tag{1b}$$

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

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*

$$\rho(x) := \frac{q(x)}{p(x)}, \tag{2}$$

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 paper, we focus on two broad families of such pairs.

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

$$\sup_{x \in \mathcal{X}} \rho(x) \leq B. \tag{3}$$

It is worth noting that $B = 1$ recovers the case without covariate shift, that is, $P = Q$. Our analysis in Section 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 $V^2 \geq 1$, we say that the likelihood ratio is *V^2-moment bounded* if

$$\mathbb{E}_{X \sim P}[\rho^2(X)] \leq V^2. \tag{4}$$

Note that when the uniform bound (3) holds, the moment bound (4) holds with $V^2 = B$. To see this, we can write $\mathbb{E}_{X \sim P}[\rho^2(X)] = \mathbb{E}_{X \sim Q}[\rho(X)] \leq B$. However, the moment bound (4) is much weaker in general. It is also worth noting that the χ^2-divergence between Q and P takes the form

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

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 4 applies under this weaker condition on the likelihood ratio.

2.3. *Unweighted versus likelihood-reweighted estimators.* In this paper, 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*

$$\widehat{f}_\lambda := \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}, \tag{5}$$

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 $\mathbb{E}_P[(Y - f(X))^2]$, as opposed to $\mathbb{E}_Q[(Y - f(X))^2]$.

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

$$\widetilde{f}_\lambda^{\mathrm{rw}} := \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho(x_i)\big(f(x_i) - y_i\big)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}. \tag{6}$$

The introduction of the likelihood ratio ensures that the objective now provides an unbiased estimate of the expectation $\mathbb{E}_Q[(Y - f(X))^2]$. However, reweighting by the likelihood ratio also increases variance, especially in the case of unbounded likelihood ratios. Accordingly, in Section 4, we study a suitably truncated form of the estimator (6).

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

$$\mathscr{K}(x, x') := \sum_{j=1}^{\infty} \mu_j \phi_j(x) \phi_j(x') \tag{7}$$

for a sequence of nonnegative eigenvalues $\{\mu_j\}_{j \geq 1}$, and eigenfunctions $\{\phi_j\}_{j \geq 1}$ 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 (7), 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

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

Consequently, the Hilbert space itself can be written as

$$\mathcal{H} := \left\{ f = \sum_{j=1}^{\infty} \theta_j \phi_j \,\Big|\, \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} < \infty \right\}. \tag{8}$$

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

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

$$\sup_{x \in \mathcal{X}} \mathscr{K}(x, x) \leq \kappa^2. \tag{9}$$

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.

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 (5). First, we prove a family of upper bounds (Theorem 3.1) 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 3.2). We then turn to the complementary question of lower bounds: in Theorem 3.3, 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 3.4, 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 suboptimal. Optimal procedures—such as the regularized KRR estimate—are actually operating over Hilbert balls with radius substantially larger than the true norm $\|f^\star\|_{\mathcal{H}}$.

3.1. *Unweighted kernel ridge regression is near-optimal.* We begin by deriving a family of upper bounds on the kernel ridge regression estimator (5) 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 3.1. *Consider a covariate-shifted regression problem with likelihood ratio that is B-bounded (3) over a Hilbert space with a κ-uniformly bounded kernel (9). Then for any $\lambda \geq 10\kappa^2/n$, the KRR estimate $\widehat{f}_\lambda$ satisfies the bound*

$$\|\widehat{f}_\lambda - f^\star\|_Q^2 \leq \underbrace{4\lambda B\|f^\star\|_{\mathcal{H}}^2}_{\mathbf{b}_\lambda^2(B)} + \underbrace{80\sigma^2 B\frac{\log n}{n}\sum_{j=1}^{\infty}\frac{\mu_j}{\mu_j+\lambda B}}_{\mathbf{v}_\lambda(B)} \tag{10}$$

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

See Section 5.1 for the proof of this theorem. In Appendix 5.1 of the Supplementary Material [10], 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 (10) 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 plots the right-hand side of the upper bound (10) 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 [23] 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, $d(\delta) := \min\{j \geq 1 | \mu_j \leq \delta^2\}$. The eigenvalue sequence is said to be regular if[2]

$$\sum_{j=d(\delta)+1}^{\infty} \mu_j \leq c d(\delta)\delta^2 \tag{11}$$

[2]In fact, we can relax this to only require the minimizing δ in equation (12) to obey the tail bound.

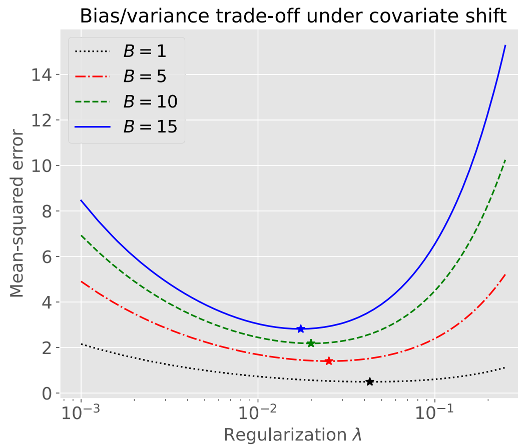

FIG. 1. *Plot of the upper bound* (10) *on the mean-squared error versus the log regularization parameter* $\log\lambda$ *for four different choices of the likelihood ratio bound* B, *in all cases with eigenvalues* $\mu_j = (1/j)^2$, *noise variance* $\sigma^2 = 1$ *and sample size* $n = 8000$. *The 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* ($B = 1$).

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 bound (10) implies that there is a universal constant c' such that

$$\|\widehat{f}_\lambda - f^\star\|_Q^2 \le c'\left\{\delta^2\|f^\star\|_{\mathcal{H}}^2 + \sigma^2 B\frac{d(\delta)\log n}{n}\right\} \quad \text{where } \delta^2 = \lambda B. \tag{12}$$

We verify this claim as part of proving Corollary 3.2 below.

This bound (12) 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* $\mu_j \le cj^{-2\alpha}$ 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 3.2 (Bounds for specific kernels).

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

$$\|\widehat{f}_\lambda - f^\star\|_Q^2 \le c\sigma^2 B\frac{D\log n}{n} \tag{13a}$$

with high probability.

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

$$\|\widehat{f}_\lambda - f^\star\|_Q^2 \le c\left(\frac{\sigma^2 B\log n}{n}\right)^{\frac{2\alpha}{2\alpha+1}} \tag{13b}$$

with high probability.

PROOF. We begin by proving the upper bound (12). With the shorthand $\delta^2 = \lambda B$, the variance term in our bound (10) can be bounded as

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

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

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

Putting together the pieces yields $\frac{1}{80}\mathbf{v}_\lambda(B) \le c_2 \sigma^2 B \frac{\log n}{n} d(\delta)$, for some universal constant c_2. Combining with the bias term yields the claim (12).

We now prove claims (13a) and (13b). For a finite-rank kernel, using the fact that $d(\delta) \le D$ for any $\delta > 0$, we can set $\lambda = \frac{\sigma^2 D \log n}{n}$ to obtain the claimed bound (13a). Now suppose that the kernel has α-polynomial decay—that is, $\mu_j \le c j^{-2\alpha}$ for some $c > 0$. For any $\delta > 0$, we then have $d(\delta) \le c'(1/\delta)^{1/\alpha}$, and hence

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

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

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

as claimed in the corollary. □

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 $d(\delta) = \min\{j \ge 1 | \mu_j \le \delta^2\}$, and the notion of regular eigenvalues (11). For a Hilbert space $\mathcal{H}$, we let $\mathcal{B}_{\mathcal{H}}(1)$ denote the Hilbert norm ball of radius one.

THEOREM 3.3. *For any $B \ge 1$, there exists a pair (P, Q) with B-bounded likelihood ratio (3) and an orthonormal basis $\{\phi_j\}_{j\ge1}$ of $L^2(Q)$ such that for any regular sequence of kernel eigenvalues $\{\mu_j\}_{j\ge1}$, we have*

$$\inf_{\widehat{f}} \sup_{f^\star \in \mathcal{B}_{\mathcal{H}}(1)} \mathbb{E}[\|\widehat{f} - f^\star\|_Q^2] \ge c \inf_{\delta>0} \left\{ \delta^2 + \sigma^2 B \frac{d(\delta)}{n} \right\}. \tag{14}$$

See Appendix 1 for the proof of this claim.

Comparing the lower bound (14) to our achievable result (12) 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 3.3, which parallel the guarantees in Corollary 3.2:

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

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

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

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

Note that both of these minimax lower bounds reduce to the known lower bounds [23] in the case of no covariate shift (i.e., $B = 1$).

3.3. *Constrained kernel regression is suboptimal.* 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

$$\widehat{f}_{\text{erm}} := \arg\min_{f \in \mathcal{B}_{\mathcal{H}}(1)} \left\{\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2\right\}. \tag{15}$$

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

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 $(\frac{B\sigma^2}{n})^{2/3}$. 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 (15) is suboptimal for a broad range of (B, n) pairs. In the following statement, we use c_1, c_2 to denote universal constants.

THEOREM 3.4. *Suppose $\|f^\star\|_{\mathcal{H}} = 1$ and $\sigma^2 = 1$. 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 $\mu_j \leq (1/j^2)$ such that*

$$\sup_{f^\star \in \mathcal{B}_{\mathcal{H}}(1)} \mathbb{E}[\|\widehat{f}_{\text{erm}} - f^\star\|_Q^2] \geq c_3 \frac{B^3}{n^2}. \tag{16}$$

See Appendix 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 (16) 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

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

It is worth understanding why the constrained form of KRR is suboptimal, while the regularized form is minimax-optimal. Recall from Corollary 3.2 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). 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.

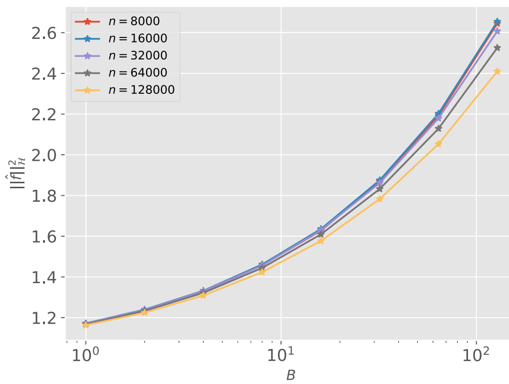

FIG. 2. *Results based on computing the regularized KRR estimate for the "bad" problems, indexed by the pair (n, B), that underlie the proof of Theorem 3.4. Each curve shows the squared Hilbert norm of the regularized KRR estimate $\|\widehat{f}_\lambda\|^2_{\mathcal{H}}$, computed with $\lambda = (\frac{4}{n\sqrt{B}})^{2/3}$, versus the likelihood ratio bound B. Each curve corresponds to a different choice of sample size n as indicated in the legend.*

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 3.4; see Appendix 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 $\lambda = (\frac{4}{n\sqrt{B}})^{2/3}$, as suggested by Corollary 3.2. In Figure 2, for each fixed n, we plot the squared Hilbert norm $\|\widehat{f}_\lambda\|^2_{\mathcal{H}}$ 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 suboptimality of the constrained KRR estimate with a fixed radius.

4. Unbounded likelihood ratios. Thus far, our analysis imposed the B-bound (3) 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 $\mathcal{N}(0, 1)$, whereas the source distribution P takes the form $\mathcal{N}(0, 0.9)$. 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 (4) applies.

The key to the success of the *unweighted* KRR estimator (5) in the bounded likelihood ratio case is the nice relationship between the covariance $\boldsymbol{\Sigma}_P := \mathbb{E}_{X\sim P}[\phi(X)\phi(X)^\top]$ of the source distribution and the covariance $\boldsymbol{I}$ of the target distribution, namely $\boldsymbol{\Sigma}_P \succeq \frac{1}{B}\boldsymbol{I}$. 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 (6), as previously introduced in Section 2.3, that ensures the nice identity $\mathbb{E}_{X\sim P}[\rho(X)\phi(X)\phi(X)^\top] = \boldsymbol{I}$. 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*

$$\rho_{\tau_n}(x) := \begin{cases} \rho(x) & \text{if } \rho(x) \le \tau_n, \\ \tau_n & \text{otherwise.} \end{cases} \tag{17}$$

We then consider the family of estimators

$$\widehat{f}_\lambda^{\mathrm{rw}} := \arg\min_{f\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}\rho_{\tau_n}(x_i)\big(f(x_i)-y_i\big)^2+\lambda\|f\|_{\mathcal{H}}^2\right\}, \tag{18}$$

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

$$\|\phi_j\|_\infty := \sup_{x\in\mathcal{X}}\big|\phi_j(x)\big| \le 1 \quad \text{for all } j = 1, 2, \dots. \tag{19}$$

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 $\Psi : \mathcal{X} \to \mathbb{R}$), which have sinusoids as their eigenfunctions, and thus satisfy this condition.

Our theorem on the truncated-reweighted KRR estimate (18) involves the kernel complexity function $\Psi(\delta,\mu) := \sum_{j=1}^{\infty}\min\{\delta^2, \mu_j\|f^\star\|_{\mathcal{H}}^2\}$, and works for any solution $\delta_n > 0$ to the inequality $\mathcal{M}(\delta) \le \delta^2/2$, where

$$\mathcal{M}(\delta) := c_0\sqrt{\frac{\sigma^2 V^2\log^3(n)}{n}\Psi(\delta,\mu)}. \tag{20}$$

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 $\sigma^2 \ge \kappa^2\|f^\star\|_{\mathcal{H}}^2$) to simplify the statement. Theoretical guarantees for all ranges of σ^2 can be found in Appendix 4.

THEOREM 4.1. *Consider a kernel with sup-norm bounded eigenfunctions* (19), *and a source-target pair with $\mathbb{E}_P[\rho^2(X)] \le V^2$. Further assume that the noise level obeys $\sigma^2 \ge \kappa^2\|f^\star\|_{\mathcal{H}}^2$. Then the estimate $\widehat{f}_\lambda^{\mathrm{rw}}$ with truncation $\tau_n = \sqrt{nV^2}$ and regularization $\lambda\|f^\star\|_{\mathcal{H}}^2 \ge \delta_n^2/3$ satisfies the bound*

$$\|\widehat{f}_\lambda^{\mathrm{rw}} - f^\star\|_Q^2 \le \delta_n^2 + 3\lambda\|f^\star\|_{\mathcal{H}}^2 \tag{21}$$

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

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

See Section 5.2 for the proof of this claim. In Appendix 5.2, 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 4.2 below, we show that for any regular sequence of eigenvalues and $\|f^\star\|_{\mathcal{H}} = 1$, we have

$$\Psi(\delta,\mu) \le c' d(\delta)\delta^2 \tag{22}$$

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

$$\sqrt{\frac{\sigma^2 V^2\log^3(n)}{n}d(\delta)} \le c_1\delta. \tag{23}$$

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

COROLLARY 4.2. *Consider kernels with sup-norm bounded eigenfunctions* (19).

(a) *For a kernel with rank D, the truncated-reweighted estimator achieves with high probability*

$$\|\widehat{f}_\lambda^{\mathrm{rw}} - f^\star\|_Q^2 \le c' \frac{DV^2 \log^3(n)\sigma^2}{n} \tag{24}$$

provided that $\lambda = c\frac{DV^2\log^3(n)\sigma^2}{n}$.

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

$$\|\widehat{f}_\lambda^{\mathrm{rw}} - f^\star\|_Q^2 \le c' \left(\frac{V^2\log^3(n)}{n}\sigma^2\right)^{\frac{2\alpha}{2\alpha+1}}, \tag{25}$$

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

PROOF. We begin by verifying the claim (22). Recalling the definition of $d(\delta)$ as the smallest integer for which $\mu_j \le \delta^2$, we can write

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

where the bound on the second sum follows from the regularity condition. This completes the proof of the bound (22).

Given our bound (22), it is straightforward to verify the claim (23).

We now prove the bounds (24) and (25). For the finite rank case, the kernel complexity measure is bounded as $\Psi(\delta, \mu) \le D\delta^2$, which implies $\delta_n^2 \le c\frac{DV^2\log^3(n)\sigma^2}{n}$ for some universal constant c. Apply Theorem 4.1 to obtain the desired rate. Now we move on to the kernel with α-polynomial eigenvalues. We know from the proof of Corollary 3.2 that $d(\delta) \le c(1/\delta)^{1/\alpha}$, and hence $\Psi(\delta, \mu) \le c'\delta^{2-1/\alpha}$. This implies $\delta_n^2 \le c(\frac{V^2\log^3 n}{n}\sigma^2)^{\frac{2\alpha}{2\alpha+1}}$, which together with Theorem 4.1 yields the claim. $\square$

Corollary 4.2 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 3.3 and the fact that the χ^2-bounded family is a larger family compared to the uniformly bounded family.

5. Proofs. In this section, we provide the proofs of our two sets of upper bounds on different estimators. Section 5.1 is devoted to the proof of Theorem 3.1 on upper bounds on unweighted KRR for B-bounded likelihood ratios, whereas Section 5.2 is devoted to the proof of Theorem 4.1 on the performance of LR-reweighted KRR with truncation.

5.1. *Proof of Theorem* 3.1. Define the empirical covariance operator[3]

$$\widehat{\mathbf{\Sigma}}_P := \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)\phi(x_i)^\top, \tag{26}$$

the population covariance operator $\mathbf{\Sigma}_P := \mathbb{E}_{X\sim P}[\phi(X)\phi(X)^\top]$, and the diagonal operator $\boldsymbol{M} := \mathsf{diag}(\{\mu_j\}_{j\ge 1})$.

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

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

$$\mathcal{E}(\lambda) := \left\{ \boldsymbol{M}^{1/2}\widehat{\mathbf{\Sigma}}_P \boldsymbol{M}^{1/2} + \lambda \boldsymbol{I} \succeq \frac{1}{2}(\boldsymbol{M}^{1/2}\mathbf{\Sigma}_P \boldsymbol{M}^{1/2} + \lambda \boldsymbol{I}) \right\}, \tag{27}$$

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

LEMMA 5.1. *For any B-bounded source-target pair* (3), *we have the deterministic lower bound*

$$\mathbf{\Sigma}_P \succeq \frac{1}{B}\boldsymbol{I}. \tag{28a}$$

If, in addition, the kernel is κ-uniformly bounded (9), *then whenever $n\lambda \geq 10\kappa^2$, the event $\mathcal{E}(\lambda)$ defined in equation* (27) *satisfies*

$$\mathbb{P}[\mathcal{E}(\lambda)] \geq 1 - 28\frac{\kappa^2}{\lambda}e^{-\frac{n\lambda}{16\kappa^2}}. \tag{28b}$$

See Section 5.1.3 for the proof of this claim.

Equipped with Lemma 5.1, we now proceed to the proof of the theorem. In terms of the basis $\{\phi_j\}_{j\geq 1}$, the KRR estimate has the expansion $\widehat{f}_\lambda = \sum_{j=1}^\infty \widehat{\theta}_j \phi_j$, where $\widehat{\theta} = \{\widehat{\theta}_j\}_{j\geq 1}$ is a sequence of coefficients in $\ell^2(\mathbb{N})$. By the optimality conditions for the KRR problem, we have

$$\widehat{\theta} - \theta^\star = -\lambda(\widehat{\mathbf{\Sigma}}_P + \lambda \boldsymbol{M}^{-1})^{-1}\boldsymbol{M}^{-1}\theta^\star + (\widehat{\mathbf{\Sigma}}_P + \lambda \boldsymbol{M}^{-1})^{-1}\left(\frac{1}{n}\sum_{i=1}^n w_i \phi(x_i)\right). \tag{29}$$

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

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

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

$$T_1 \overset{(a)}{\leq} 2\lambda B\|f^\star\|_{\mathcal{H}}^2, \quad \text{and} \quad T_2 \overset{(b)}{\leq} \frac{40(\log n)\sigma^2}{n}\sum_{j=1}^\infty \frac{\mu_j}{\mu_j/B + \lambda}, \tag{30}$$

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

5.1.1. *Proof of the bound* (30)(*a*). We establish that this bound holds conditionally on the event $\mathcal{E}(\lambda)$. Following some algebraic manipulations, we have

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

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

Lemma 5.1 also guarantees that $\boldsymbol{\Sigma}_P \succeq \frac{1}{B}\boldsymbol{I}$, whence

$$T_1 \le 2\lambda\|f^\star\|_{\mathcal{H}}^2 \left\| \boldsymbol{M}^{1/2}\left(\frac{1}{B}\boldsymbol{M} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{M}^{1/2}\right\|_2 = 2\lambda \cdot \max_{j\ge 1}\left\{\frac{\mu_j}{\frac{\mu_j}{B} + \lambda}\right\} \le 2\lambda B\|f^\star\|_{\mathcal{H}}^2.$$

This establishes the claim (30)(a).

5.1.2. *Proof of the bound* (30)(*b*). Define the random vector

$$W := (\widehat{\boldsymbol{\Sigma}}_P + \lambda\boldsymbol{M}^{-1})^{-1}\left(\frac{1}{n}\sum_{i=1}^n w_i\phi(x_i)\right).$$

Conditioned on the covariates $\{x_i\}_{i=1}^n$, W is a zero-mean sub-Gaussian random variable with covariance operator

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

Consequently, by the Hanson–Wright inequality in the RKHS (cf. Theorem 2.6 in the paper [3]), we have

$$\mathbb{P}\big[T_2 \ge 20(\log n)\mathrm{trace}(\boldsymbol{\Lambda}) \mid \{x_i\}_{i=1}^n\big] \le \frac{1}{n^{10}}, \tag{31}$$

where the probability is taken over the noise variables.

It remains to upper bound the trace. We have the relation

$$\mathrm{trace}(\boldsymbol{\Lambda}) = \mathrm{trace}\left(\frac{\sigma^2}{n}(\widehat{\boldsymbol{\Sigma}}_P + \lambda\boldsymbol{M}^{-1})^{-1}\widehat{\boldsymbol{\Sigma}}_P(\widehat{\boldsymbol{\Sigma}}_P + \lambda\boldsymbol{M}^{-1})^{-1}\right),$$

so that

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

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

$$\begin{aligned}
\mathrm{trace}(\boldsymbol{\Lambda}) &\le 2\frac{\sigma^2}{n}\mathrm{trace}\big(\boldsymbol{M}^{1/2}(\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_P\boldsymbol{M}^{1/2} + \lambda\boldsymbol{I})^{-1}\boldsymbol{M}^{1/2}\big)\\
&\overset{\text{(i)}}{\le} 2\frac{\sigma^2}{n}\mathrm{trace}\left(\boldsymbol{M}^{1/2}\left(\frac{1}{B}\boldsymbol{M} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{M}^{1/2}\right)\\
&\overset{\text{(ii)}}{=} 2\frac{\sigma^2}{n}\sum_{j=1}^{\infty}\frac{\mu_j}{\frac{\mu_j}{B} + \lambda}.
\end{aligned}$$

Here step (i) follows since $\boldsymbol{\Sigma}_P \succeq \frac{1}{B}\boldsymbol{I}$, and step (ii) follows from a direct calculation. Substituting this upper bound on the trace into the tail bound (31) yields the claimed bound (30)(b).

5.1.3. *Proof of Lemma* 5.1. We begin with the proof of the lower bound (28a). Since $\{\phi_j\}_{j\geq 1}$ is an orthonormal basis of $L^2(Q)$, we have

$$\mathbb{E}_{X\sim Q}\big[\phi(X)\phi(X)^\top\big] = \mathbb{E}_{X\sim P}\big[\rho(X)\phi(X)\phi(X)^\top\big] = \boldsymbol{I}.$$

Thus, the B-boundedness of the likelihood ratio (3) implies that

$$\boldsymbol{I} \preceq \mathbb{E}_{X\sim P}\big[B\phi(X)\phi(X)^\top\big] = B\boldsymbol{\Sigma}_P,$$

which is equivalent to the claim (28a).

Next, we prove the lower bound (27). We introduce the shorthand notation

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

along with the matrix $\boldsymbol{\Delta} := \boldsymbol{\Sigma}_\lambda^{-1/2}(\widehat{\boldsymbol{\Sigma}}_\lambda - \boldsymbol{\Sigma}_\lambda)\boldsymbol{\Sigma}_\lambda^{-1/2}$. Recalling that $|||\cdot|||_2$ denotes the ℓ_2-operator norm of a matrix, we first observe that $\{|||\boldsymbol{\Delta}|||_2 \leq \frac{1}{2}\} \subseteq \mathcal{E}$. Consequently, it suffices to show that $|||\boldsymbol{\Delta}|||_2 \leq \frac{1}{2}$ with high probability.

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

$$\boldsymbol{Z}_i := \boldsymbol{\Sigma}_\lambda^{-1/2}\boldsymbol{M}^{1/2}\big(\phi(x_i)\phi(x_i)^\top - \boldsymbol{\Sigma}_P\big)\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_\lambda^{-1/2}$$

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

$$\begin{aligned} |||\boldsymbol{Z}_i|||_2 &\leq 2\sup_{x\in\mathcal{X}} |||\boldsymbol{\Sigma}_\lambda^{-1/2}\boldsymbol{M}^{1/2}\phi(x)\phi(x)^\top\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_\lambda^{-1/2}|||_2 \\ &= 2\sup_{x\in\mathcal{X}} \|\boldsymbol{\Sigma}_\lambda^{-1/2}\boldsymbol{M}^{1/2}\phi(x)\|_2^2 \qquad (32)\\ &\leq 2\kappa^2|||\boldsymbol{\Sigma}_\lambda^{-1/2}|||_2^2 \leq \frac{2\kappa^2}{\lambda} =: L, \end{aligned}$$

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

On the other hand, the variance of $\boldsymbol{Z}_i$ can be bounded as

$$\begin{aligned} \mathbb{E}[\boldsymbol{Z}_i^2] &\preceq \mathbb{E}\big[\big(\boldsymbol{\Sigma}_\lambda^{-1/2}\boldsymbol{M}^{1/2}\phi(X)\phi(X)^\top\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_\lambda^{-1/2}\big)^2\big] \\ &= \mathbb{E}\big[\boldsymbol{\Sigma}_\lambda^{-1/2}\boldsymbol{M}^{1/2}\phi(X)\phi(X)^\top\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_\lambda^{-1}\boldsymbol{M}^{1/2}\phi(X)\phi(X)^\top\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_\lambda^{-1/2}\big] \\ &\preceq \mathbb{E}\big[\boldsymbol{\Sigma}_\lambda^{-1/2}\boldsymbol{M}^{1/2}\phi(X)\phi(X)^\top\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_\lambda^{-1/2}\big]\cdot\sup_{x\in\mathcal{X}}\{\phi(x)^\top\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_\lambda^{-1}\boldsymbol{M}^{1/2}\phi(x)\} \\ &\preceq \frac{\kappa^2}{\lambda}\boldsymbol{\Sigma}_\lambda^{-1/2}\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_P\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_\lambda^{-1/2} =: \boldsymbol{V}, \end{aligned}$$

where the last inequality follows by applying the bound (32) on $\sup_{x\in\mathcal{X}}\|\boldsymbol{\Sigma}_\lambda^{-1/2}\boldsymbol{M}^{1/2}\phi(x)\|_2^2$.

Suppose that we can show that

$$\operatorname{trace}(\boldsymbol{V}) \leq \frac{\kappa^2}{\lambda}\cdot\frac{\kappa^2}{\lambda}; \qquad (33a)$$

$$|||\boldsymbol{V}|||_2 \leq \frac{\kappa^2}{\lambda}. \qquad (33b)$$

We can then apply a dimension-free matrix Bernstein inequality (see Lemma 7.1) with $t = 1/2$ to obtain the tail bound

$$\mathbb{P}\Big[|||\boldsymbol{\Delta}|||_2 \geq \frac{1}{2}\Big] \leq 28\frac{\kappa^2}{\lambda}\exp\Big(-\frac{n\lambda}{16\kappa^2}\Big),$$

as long as $n\lambda \geq 10\kappa^2$. Thus, the only remaining detail is to prove the bounds (33a) and (33b).

Proof of the bound (33a). Using the definition of $\boldsymbol{V}$, we have

$$\begin{aligned}\mathsf{trace}(\boldsymbol{V}) &= \frac{\kappa^2}{\lambda}\mathsf{trace}\big(\boldsymbol{\Sigma}_\lambda^{-1/2}\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_P\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_\lambda^{-1/2}\big)\\ &= \frac{\kappa^2}{\lambda}\mathbb{E}_P\big[\mathsf{trace}\big(\boldsymbol{\Sigma}_\lambda^{-1/2}\boldsymbol{M}^{1/2}\phi(x)\phi(x)^\top\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_\lambda^{-1/2}\big)\big]\\ &\le \frac{\kappa^2}{\lambda}\cdot\frac{\kappa^2}{\lambda}.\end{aligned}$$

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

Proof of the bound (33b). Recalling the definition of $\boldsymbol{\Sigma}_\lambda$, we see that $|||\boldsymbol{\Sigma}_\lambda^{-1/2}\boldsymbol{M}^{1/2}\times\boldsymbol{\Sigma}_P\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_\lambda^{-1/2}|||_2 \le 1$, and hence

$$|||\boldsymbol{V}|||_2 = \frac{\kappa^2}{\lambda}|||\boldsymbol{\Sigma}_\lambda^{-1/2}\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_P\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_\lambda^{-1/2}|||_2 \le \frac{\kappa^2}{\lambda},$$

which is the claimed upper bound on $|||\boldsymbol{V}|||_2$.

5.2. *Proof of Theorem* 4.1. 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

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

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

LEMMA 5.2. *Fixing any $r>0$, we have*

$$\sup_{g\in\mathcal{G}(r)}\left\{\|g-f^\star\|_Q^2 + \frac{1}{n}\sum_{i=1}^n \rho_{\tau_n}(x_i)\big[\big(f^\star(x_i)-y_i\big)^2 - \big(g(x_i)-y_i\big)^2\big]\right\} \le \mathcal{M}(r) \tag{34}$$

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

See Section 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 (34) holds at radius δ_λ. We immediately point out an important property of the regularized radius δ_λ, namely $\mathcal{M}(\delta_\lambda) \le (1/2)\cdot\delta_\lambda^2$. To see this, note that $r\mapsto\mathcal{M}(r)/r$ is nonincreasing in r, and hence

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

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

$$\inf_{f\in\mathcal{H}, f\notin\mathcal{G}(\delta_\lambda)} \frac{1}{n}\sum_{i=1}^n \rho_{\tau_n}(x_i)\big\{\big(f(x_i)-y_i\big)^2 - \big(f^\star(x_i)-y_i\big)^2\big\} + \lambda\|f\|_{\mathcal{H}}^2 - \lambda\|f^\star\|_{\mathcal{H}}^2 > 0. \tag{35}$$

It then follows that that $\|\widehat{f}-f^\star\|_Q \le \delta_\lambda$, as desired. Consequently, the remainder of our proof is devoted to establishing that inequality (35) holds conditioned on $\mathcal{E}(\delta_\lambda)$.

Given any function $f\in\mathcal{H}$ and $f\notin\mathcal{G}(\delta_\lambda)$, there exists an $\alpha\ge 1$ such that the shifted function

$$\widetilde{f} := f^\star + \frac{1}{\alpha}(f-f^\star)$$

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

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

Consequently, in order to establish the claim (35), it suffices to prove that the quantity

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

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 $\|\widetilde{f}-f^\star\|_{\mathcal{H}}\le 3\|f^\star\|_{\mathcal{H}}$, and (2) $\|\widetilde{f}-f^\star\|_Q\le\delta_\lambda$, while $\|\widetilde{f}-f^\star\|_{\mathcal{H}}=3\|f^\star\|_{\mathcal{H}}$.

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

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

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 nonpositivity of $\lambda\|\widetilde{f}\|_{\mathcal{H}}^2$; step (ii) follows from the property $\mathcal{M}(\delta_\lambda)\le(1/2)\cdot\delta_\lambda^2$ and step (iii) uses the definitions of δ_λ and λ.

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

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

Here, step (i) again follows from the conditioning on the event $\mathcal{E}(\delta_\lambda)$ and the assumption that $\|\widetilde{f}-f^\star\|_Q\le\delta_\lambda$. Step (ii) relies on the facts that $\mathcal{M}(\delta_\lambda)\le(1/2)\cdot\delta_\lambda^2$, $\|f^\star\|_{\mathcal{H}}=\|f^\star\|_{\mathcal{H}}$, and that $\|\widetilde{f}\|_{\mathcal{H}}\ge 2\|f^\star\|_{\mathcal{H}}$. 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 (35).

5.2.1. *Proof of Lemma* 5.2. Define the shifted function class $\mathcal{F}^\star:=\mathcal{H}-f^\star$, along with its r-localized version

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

We begin by observing that

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

which yields the following equivalent formulation of the claim in Lemma 5.2:

$$\sup_{h \in \mathcal{F}^\star(r)} \left\{ \frac{1}{n} \sum_{i=1}^{n} \big[2w_i \rho_{\tau_n}(x_i) h(x_i) + \|h\|_Q^2 - \rho_{\tau_n}(x_i) h^2(x_i)\big] \right\} \leq \mathcal{M}(r). \tag{36}$$

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

$$T_1 := \sup_{h \in \mathcal{F}^\star(r)} \left| \frac{2}{n} \sum_{i=1}^{n} w_i \rho_{\tau_n}(x_i) h(x_i) \right|, \quad \text{and}$$

$$T_2 := \sup_{h \in \mathcal{F}^\star(r)} \left| \frac{1}{n} \sum_{i=1}^{n} \big\{\|h\|_Q^2 - \rho_{\tau_n}(x_i) h^2(x_i)\big\} \right|.$$

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

$$\begin{aligned} T_1 &\leq c\sigma \sqrt{\frac{V^2 \log^3(n)}{n}} \\ &\quad \cdot \left\{ \sum_{j=1}^{\infty} \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\} \right\}^{1/2} \quad \text{with probability at least } 1 - n^{-10}, \text{ and} \end{aligned} \tag{37a}$$

$$\begin{aligned} T_2 &\leq c \sqrt{\frac{V^2 \log^3(n)}{n}} \\ &\quad \cdot \sum_{j=1}^{\infty} \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\} \quad \text{with probability at least } 1 - n^{-10}. \end{aligned} \tag{37b}$$

In conjunction, these two bounds ensure that

$$\begin{aligned} T_1 + T_2 &\leq c \sqrt{\frac{V^2 \log^3(n)}{n}} \cdot \sum_{j=1}^{\infty} \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\} \\ &\quad + c \sqrt{\sum_{j=1}^{\infty} \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\} \frac{V^2 \log^3(n)}{n} \sigma^2}. \end{aligned} \tag{38}$$

Since the kernel function is κ^2-bounded, we have

$$\sum_{j=1}^{\infty} \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\} \leq \|f^\star\|_{\mathcal{H}}^2 \sum_{j=1}^{\infty} \mu_j \leq \kappa^2 \|f^\star\|_{\mathcal{H}}^2,$$

which together with the assumption $\sigma^2 \geq \kappa^2 \|f^\star\|_{\mathcal{H}}^2$ implies that

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

Therefore the bound (36) holds.

It remains to prove the bounds (37a) and (37b). 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

$$|h(x)| \le \sqrt{10 \sum_{j=1}^{\infty} \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\}}, \quad \text{and} \tag{39a}$$

$$\sum_{j=1}^{\infty} \frac{\theta_j^2}{\min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\}} \le 10, \quad \text{where } h = \sum_{j=1}^{\infty} \theta_j \phi_j. \tag{39b}$$

See Appendix 3 for the proof of these elementary claims.

5.2.2. *Proof of inequality* (37b). We begin by analyzing the term T_2. By the triangle inequality, we have the upper bound $T_2 \le T_{2a} + T_{2b}$, where

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

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

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:

$$T_{2a} \le c \sqrt{\frac{V^2}{n}} \sum_{j=1}^{\infty} \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\}, \quad \text{and} \tag{40a}$$

$$T_{2b} \le c \sqrt{\frac{V^2 \log^2(n)}{n}} \sum_{j=1}^{\infty} \min\{r^2, \mu_j \|f^\star\|_{\mathcal{H}}^2\} \quad \text{with probability at least } 1 - n^{-10}. \tag{40b}$$

Combining these two bounds yields the claim (37b).

Proof of inequality (40a). We begin by proving the claimed upper bound on T_{2a}. Note that

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

where the last step follows from a combination of Markov's inequality, Chebyshev's inequality, and the ℓ_∞-norm bound (39a). Recalling that $\tau_n = \sqrt{nV^2}$, the bound (40a) follows.

Proof of the bound (40b). We prove the claimed bound on T_{2b} by first bounding its mean $\mathbb{E}[T_{2b}]$, and then providing a high-probability bound on the deviation $T_{2b} - \mathbb{E}[T_{2b}]$.

Bound on the mean: By a standard symmetrization argument (see, e.g., Chapter 4 in the book [22]), we have the upper bound

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

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

$$\sup_{h\in\mathcal{F}^\star(r)}\left|\sum_{i=1}^n \varepsilon_i \rho_{\tau_n}(x_i)h^2(x_i)\right| \le \sup_{\widetilde{h},h\in\mathcal{F}^\star(r)} Z(h,\widetilde{h}),$$

where above we have defined

$$Z(h,\widetilde{h}) := \left|\sum_{i=1}^n \varepsilon_i \rho_{\tau_n}(x_i)\widetilde{h}(x_i)h(x_i)\right|.$$

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

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

where the final step follows by combining the Cauchy–Schwarz inequality with the bound (39b). We now repeat the same argument to upper bound the inner term involving h; in particular, we have

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

Putting together the pieces now leads to the upper bound

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

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

$$\mathbb{E}[T_{2b}] \le \frac{20}{n}\Bigg\{\sum_{j=1}^{\infty}\min\{r^2, \mu_j\|f^\star\|_{\mathcal{H}}^2\} \cdot \sum_{k=1}^{\infty}\min\{r^2, \mu_k\|f^\star\|_{\mathcal{H}}^2\}\mathbb{E}_{X,\varepsilon}\Bigg(\sum_{i=1}^{n}\varepsilon_i\rho_{\tau_n}(x_i)\phi_j(x_i)\phi_k(x_i)\Bigg)^2\Bigg\}^{1/2}. \tag{41}$$

We now observe that

$$\begin{aligned}\mathbb{E}_{X,\varepsilon}\Bigg[\Bigg(\sum_{i=1}^{n}\varepsilon_i\rho_{\tau_n}(x_i)\phi_j(x_i)\phi_k(x_i)\Bigg)^2\Bigg] &= \sum_{i=1}^{n}\mathbb{E}_{X,\varepsilon}\big[\varepsilon_i^2\big(\rho_{\tau_n}(x_i)\big)^2\phi_j^2(x_i)\phi_k^2(x_i)\big] \\ &\le \sum_{i=1}^{n}\mathbb{E}_{X,\varepsilon}\big[\rho^2(x_i)\big] \le nV^2,\end{aligned}$$

where we have used the fact that $\|\phi_j\|_\infty \le 1$ for all $j \ge 1$, and that $\rho_{\tau_n}(x_i) \le \rho(x_i)$. Substituting this upper bound into our earlier inequality (41) yields

$$\mathbb{E}[T_{2b}] \le 20\sqrt{\frac{V^2}{n}} \cdot \sum_{j=1}^{\infty}\min\{r^2, \mu_j\|f^\star\|_{\mathcal{H}}^2\}. \tag{42}$$

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

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

Consequently, we have

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

In addition, we have

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

where we have applied the ℓ_∞-norm bound (39a) 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 7.2) and obtain

$$\begin{aligned}&\mathbb{P}\Big[T_{2b} \ge \mathbb{E}[T_{2b}] + \frac{t}{n}\Big] \\ &\le \exp\Bigg(-\frac{t^2}{3000nV^2(\sum_{j=1}^{\infty}\min\{r^2, \mu_j\|f^\star\|_{\mathcal{H}}^2\})^2 + 900\sqrt{nV^2}\sum_{j=1}^{\infty}\min\{r^2, \mu_j\|f^\star\|_{\mathcal{H}}^2\}t}\Bigg).\end{aligned} \tag{43}$$

Completing the proof of the bound (40b)*:* We now have the ingredients to complete the proof of the claim (40b). In particular, by combining the upper bound (42) on the mean with the deviation bound (43), we find that

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

as claimed in equation (40b).

5.2.3. *Proof of inequality* (37a). Now we focus on the first term

$$T_1 = \sup_{h\in\mathcal{F}^\star(r)} \left|\frac{1}{n}\sum_{i=1}^{n} w_i \rho_{\tau_n}(x_i)h(x_i)\right|.$$

Repeating the same strategy as in the proof of the bound (40b), we see that

$$T_1 \le \frac{1}{n}\left\{10\sum_{j=1}^{\infty}\min\{r^2, \mu_j\|f^\star\|_{\mathcal{H}}^2\}\cdot\left(\sum_{i=1}^{n} w_i\rho_{\tau_n}(x_i)\phi_j(x_i)\right)^2\right\}^{1/2}. \tag{44}$$

Fix $\{x_i\}_{i=1}^n$. We see that $(\sum_{i=1}^n w_i\rho_{\tau_n}(x_i)\phi_j(x_i))^2$ is a quadratic form of independent sub-Gaussian random variables. Apply the Hanson–Wright inequality (e.g., Theorem 6.2.1 in the book [21]) to obtain that with probability at least $1-n^{-10}$,

$$\left(\sum_{i=1}^{n} w_i\rho_{\tau_n}(x_i)\phi_j(x_i)\right)^2 \le c_3\sigma^2\sum_{i=1}^{n}[\rho_{\tau_n}(x_i)\phi_j(x_i)]^2. \tag{45}$$

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

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

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

$$\alpha := \mathbb{E}\sum_{i=1}^{n}\mathsf{Var}([\rho_{\tau_n}(x_i)\phi_j(x_i)]^2) \le (nV^2)^2,$$

$$\beta := \sup_x\left|[\rho_{\tau_n}(x)\phi_j(x)]^2\right| \le {\tau_n}^2 = nV^2,$$

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

$$\sum_{i=1}^{n}[\rho_{\tau_n}(x_i)\phi_j(x_i)]^2 \le c_6 nV^2\log n. \tag{46}$$

Combine the inequalities (44), (45), and (46) to complete the proof of the inequality (37a).

6. Discussion. In this paper, 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, for example, 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 [2, 23]. 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 paper, 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 (39a) 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.

Acknowledgments. The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper. The authors would also like to thank Samory Kpotufe for helpful discussions on transfer learning. MJW is currently on leave from UC Berkeley EECS and Statistics.

Funding. This work was partially supported by NSF-DMS Grant 2015454, NSF-IIS Grant 1909365, as well as Office of Naval Research Grant DOD-ONR-N00014-18-1-2640 to MJW. RP was partially supported by a Berkeley Fellowship and an ARCS Fellowship.

SUPPLEMENTARY MATERIAL

Additional proofs (DOI: 10.1214/23-AOS2268SUPP; .pdf). Additional proofs of the results in the paper can be found in the Supplementary Material.

REFERENCES

[1] CAI, T. T. and WEI, H. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *Ann. Statist.* **49** 100–128. MR4206671 https://doi.org/10.1214/20-AOS1949

[2] CAPONNETTO, A. and DE VITO, E. (2007). Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.* **7** 331–368. MR2335249 https://doi.org/10.1007/s10208-006-0196-8

[3] CHEN, X. and YANG, Y. (2021). Hanson–Wright inequality in Hilbert spaces with application to K-means clustering for non-Euclidean data. *Bernoulli* **27** 586–614. MR4177382 https://doi.org/10.3150/20-BEJ1251

[4] CORTES, C., MANSOUR, Y. and MOHRI, M. (2010). Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems* **23** 442–450.

[5] GUAN, H. and LIU, M. (2021). Domain adaptation for medical image analysis: A survey. *IEEE Trans. Biomed. Eng.* 1–1.

[6] GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. *Springer Series in Statistics*. Springer, New York. MR1920390 https://doi.org/10.1007/b97848

[7] JIANG, J. and ZHAI, C. (2007). Instance weighting for domain adaptation in NLP. In *Proceedings of the* 45*th Annual Meeting of the Association of Computational Linguistics* 264–271. Association for Computational Linguistics, Prague, Czech Republic.

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

[9] KPOTUFE, S. and MARTINET, G. (2021). Marginal singularity and the benefits of labels in covariate-shift. *Ann. Statist.* **49** 3299–3323. MR4352531 https://doi.org/10.1214/21-aos2084

[10] MA, C., PATHAK, R. and WAINWRIGHT, M. J. (2023). Supplement to "Optimally tackling covariate shift in RKHS-based nonparametric regression." https://doi.org/10.1214/23-AOS2268SUPP

[11] MAITY, S., SUN, Y. and BANERJEE, M. (2020). Minimax optimal approaches to the label shift problem. ArXiv preprint. Available at arXiv:2003.10443.

[12] PAN, S. J. and YANG, Q. (2009). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22** 1345–1359.

[13] PATHAK, R., MA, C. and WAINWRIGHT, M. J. (2022). A new similarity measure for covariate shift with applications to nonparametric regression. In *Proceedings of the* 39*th International Conference on Machine Learning*, *PMLR* **162** 17517–17530.

[14] POLLARD, D. (2012). *Convergence of Stochastic Processes*. Springer, Berlin.

[15] REEVE, H. W. J., CANNINGS, T. I. and SAMWORTH, R. J. (2021). Adaptive transfer learning. *Ann. Statist.* **49** 3618–3649. MR4352543 https://doi.org/10.1214/21-aos2102

[16] SCHMIDT-HIEBER, J. and ZAMOLODTCHIKOV, P. (2022). Local convergence rates of the least squares estimator with applications to transfer learning. ArXiv preprint. Available at arXiv:2204.05003.

[17] SCHÖLKOPF, B., SMOLA, A. J., BACH, F. et al. (2002). *Learning with Kernels*: *Support Vector Machines*, *Regularization*, *Optimization*, *and Beyond*. MIT Press, Cambridge.

[18] SHIMODAIRA, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference* **90** 227–244. MR1795598 https://doi.org/10.1016/S0378-3758(00)00115-4

[19] SUGIYAMA, M. and KAWANABE, M. (2012). *Machine Learning in Non-stationary Environments*: *Introduction to Covariate Shift Adaptation*. MIT Press, Cambridge.

[20] VAPNIK, V. (1999). *The Nature of Statistical Learning Theory*. Springer, Berlin.

[21] VERSHYNIN, R. (2018). *High-Dimensional Probability*: *An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge Univ. Press, Cambridge. MR3837109 https://doi.org/10.1017/9781108231596

[22] WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics*: *A Non-asymptotic Viewpoint*. *Cambridge Series in Statistical and Probabilistic Mathematics* **48**. Cambridge Univ. Press, Cambridge. MR3967104 https://doi.org/10.1017/9781108627771

[23] YANG, Y., PILANCI, M. and WAINWRIGHT, M. J. (2017). Randomized sketches for kernels: Fast and optimal nonparametric regression. *Ann. Statist.* **45** 991–1023. MR3662446 https://doi.org/10.1214/16-AOS1472

ADDITIONAL PROOFS FOR THE PAPER "”OPTIMALLY TACKLING COVARIATE SHIFT IN RKHS-BASED NONPARAMETRIC REGRESSION”

BY CONG MA[1,a], REESE PATHAK[2,b] AND MARTIN J. WAINWRIGHT[2,3,4,c]

[1]*Department of Statistics, University of Chicago,* [a]*congm@uchicago.edu*

[2]*Department of Electrical Engineering and Computer Sciences, UC Berkeley,* [b]*pathakr@berkeley.edu;* [c]*wainwrig@berkeley.edu*

[3]*Department of Statistics, UC Berkeley,* [c]*wainwrig@berkeley.edu*

[4]*Laboratory for Information and Decision Systems, Statistics and Data Science Center, EECS and Mathematics, Massachusetts Institute of Technology*

1. Proof of Theorem 3.3. Let δ_n be the smallest positive solution to the inequality $c'\sigma^2 B\frac{d(\delta)}{n} \leq \delta^2$, 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 $j \geq 1$. 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

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

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

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

See Lemma 4 in the paper [3].

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

MSC2020 subject classifications: Primary 62C20; secondary 62G08.

Keywords and phrases: covariate shift, nonparametric regression, reproducing kernel Hilbert spaces, kernel ridge regression, transfer learning.

$\{(x_i, y_i)\}_{1\le i\le n}$ when the true function arises from θ^j. Then for any pair of distinct indices $j \neq k$, we have the upper bound

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

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

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

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

By the definition of a regular kernel, we have $d_n \ge c\frac{n\delta_n^2}{B\sigma^2}$ for a universal constant c. Furthermore, since δ_n satisfies the lower bound $\delta_n^2 \ge c'\frac{\sigma^2 B}{n}$, the condition (1) is met by setting $r^2 = c_1\delta_n^2$ for some sufficiently small constant $c_1 > 0$.

2. Proof of Theorem 3.4. Let the sample size $n \ge 1$ and likelihood ratio bound $B \ge 1$ 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

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

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.

2.1. *Reduction to a one dimensional separation problem.* To establish our lower bound it suffices to consider the following "hard" function

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

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

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

Putting these pieces together, we see that

$$\sup_{f^\star \in \mathcal{F}_n} \mathbb{E}\Big[\|\widehat{f}_{\text{erm}} - f^\star\|_Q^2\Big] \geq \mathbb{E}\Big[\|\widehat{f}_{\text{erm}} - f^\star_{\text{hard}}\|_Q^2\Big] \overset{\text{(i)}}{=} \mathbb{E}\Big[\|\widehat{\theta}_{\text{erm}} - \theta^\star_{\text{hard}}\|_2^2\Big] \overset{\text{(ii)}}{\geq} \mathbb{E}\Big[\big((\widehat{\theta}_{\text{erm}})_1 - \theta_1^\star\big)^2\Big]. \tag{3}$$

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 (3), it suffices to show that:

$$\mathbb{P}\Big\{ \big((\widehat{\theta}_{\text{erm}})_1 - 1\big)^2 \geq c_3 \frac{B^3}{n^2} \Big\} \geq \frac{1}{2}. \tag{4}$$

2.2. *Proof of one-dimensional separation bound* (4). We begin with a proof outline.

Proof outline

To establish (4), we can assume $(\widehat{\theta}_{\text{erm}})_1 \in [0, 1]$; otherwise, the lower bound follows trivially, provided c_2 is sufficiently small, in particular, $c_2 \leq \sqrt[3]{1/c_3}$. We introduce a bit of notation:

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

Thus, we can further restrict the empirical risk minimization problem (2) to

$$\tilde{\theta} := \arg\min \Big\{ \sum_{i=1}^n \big(\sum_{j=1}^n \theta_j x_{ij} - y_i\big)^2 \mid \sum_{j=1}^n \frac{\theta_j^2}{\mu_j} \leq 1, \;\; \theta_1 \in [0,1] \Big\}$$

$$= \arg\min \Big\{ (\theta - \theta^\star)^T \widehat{\mathbf{\Sigma}}_P (\theta - \theta^\star) - 2v^T(\theta - \theta^\star) \mid \sum_{j=1}^n \frac{\theta_j^2}{\mu_j} \leq 1, \;\; \theta_1 \in [0,1] \Big\}. \tag{5}$$

Indeed, in order to prove inequality (4), it suffices to show that

$$\mathbb{P}\Big\{ \big(\tilde{\theta}_1 - 1\big)^2 \geq c_3 \frac{B^3}{n^2} \Big\} \geq \frac{1}{2}. \tag{6}$$

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

$$g(t) := \inf \Big\{ (\theta - \theta^\star)^T \widehat{\mathbf{\Sigma}}_P (\theta - \theta^\star) - 2v^T(\theta - \theta^\star) \mid \sum_{j=1}^n \frac{\theta_j^2}{\mu_j} \leq 1, \;\; \theta_1 = t \Big\}. \tag{7}$$

By definition (5), the choice $\tilde{\theta}$ minimizes this objective, and therefore $\inf_{t \in [0,1]} g(t) = g(\tilde{\theta}_1)$. The next two lemmas concern the minimum value and minimizer of g. Lemma 2.1, which we prove in section 2.4.1, bounds the minimal value from above. Lemma 2.2, 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 2.4.2.

LEMMA 2.1 (Minimal value of empirical objective). *There is a constant $c^* > 0$ such that*

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

holds with probability at least $3/4$.

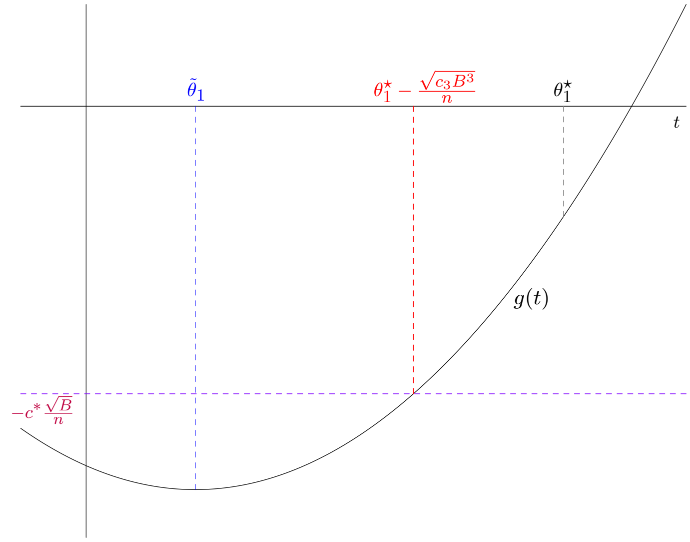

FIG 1. *Pictorial representation of lower bound argument, separating the first coordinate of empirical risk minimizer, $\tilde{\theta}_1$, from the true population minimizer $\theta_1^\star$. Lemma 2.1 establishes the upper bound, depicted in purple above, on the minimal value of g. Lemma 2.2 establishes an interval, shown between the red dashed line and $\theta_1^\star$ above, which excludes $\tilde{\theta}_1$. This allows us to ensure that $\theta_1^\star$ and $\tilde{\theta}_1$ are sufficiently separated.*

LEMMA 2.2 (Separation from $\theta_1^\star$). *There exists a constant $c_3 > 0$ such that*

$$\inf_{\substack{t\in[0,1]\\(1-t)^2\le c_3 B^3/n^2}} g(t) > -c^* \frac{\sqrt{B}}{n}. \tag{8}$$

where probability at least $3/4$.

Note that the constant c^* used in Lemmas 2.1 and 2.2 is the same. Thus—after union bounding over the two error events—with probability at least $1/2$,

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

Recalling that $\tilde{\theta}_1 \in [0,1]$, we conclude on this event that $(1-\tilde{\theta}_1)^2 \ge c_3 \frac{B^3}{n^2}$, which furnishes (6), 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.

2.3. *A useful lemma.*

LEMMA 2.3. *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*

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

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

PROOF. For each $j \geq 1$, define $\eta_j := \left(1 + \frac{\alpha}{B\mu_j}\right)^{-1}$. 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 $v_j = \frac{1}{n} \sum_{i=1}^{n} w_i x_{ij}$. Under the construction of the lower bound instance, we have $\sqrt{n} v_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. 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

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

where $c > 0$ is some universal constant.

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

$$\max_{j=2,\ldots,n} \eta_j \leq 1; \tag{9a}$$

$$\sum_{j=2}^{n} \eta_j^2 \leq C_1 \frac{B^{1/2}}{\alpha^{1/2}}; \tag{9b}$$

$$C_2 \frac{B^{1/2}}{\alpha^{1/2}} \leq \sum_{j=2}^{n} \eta_j \leq C_3 \frac{B^{1/2}}{\alpha^{1/2}}. \tag{9c}$$

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 (9). The first relation (9a) is trivial. We provide the proof of the third inequalities (9c); the proof of the middle one (cf. relation (9b)) 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 $1 \geq \frac{\alpha}{B\mu_j}$ for $j \leq \lfloor \sqrt{B/\alpha} \rfloor$ and $1 \leq \frac{\alpha}{B\mu_j}$ for $j \geq \lfloor \sqrt{B/\alpha} \rfloor$. These allow us to upper bound $\sum_{j=2}^{n} \eta_j$ as

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

Similarly, we have the lower bound

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

This finishes the proof. □

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

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

This set plays a key role. In view of definition (7), we can write

$$g(t) = \inf_{\theta_{\mathsf{R}} \in \mathcal{C}(t)} \left\{ \begin{bmatrix} t-1 \\ \theta_{\mathsf{R}} \end{bmatrix}^\top \widehat{\mathbf{\Sigma}}_P \begin{bmatrix} t-1 \\ \theta_{\mathsf{R}} \end{bmatrix} - 2 \begin{bmatrix} t-1 \\ \theta_{\mathsf{R}} \end{bmatrix}^\top v \right\}, \tag{10}$$

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

2.4.1. *Proof of Lemma 2.1.* We show that with probability at least $3/4$,

$$g(\omega) \le -c^* \frac{\sqrt{B}}{n}, \quad \text{where} \quad \omega := \sqrt{1 - \frac{B^{3/2}}{n}}. \tag{11}$$

When $n^2 \ge B^3$, we have $\omega \in [0,1]$. Since $\inf_{t \in [0,1]} g(t) \le g(\omega)$, the display (11) implies the result.

Proof of bound (11)**:** From the proof of Lemma 5.1, if we set $\lambda := C \frac{\log n}{n}$ for some constant $C > 0$, then we have

$$\tfrac{1}{2}(\mathbf{\Sigma}_P + \lambda \boldsymbol{M}^{-1}) \preceq \widehat{\mathbf{\Sigma}}_P + \lambda \boldsymbol{M}^{-1} \preceq \tfrac{3}{2}(\mathbf{\Sigma}_P + \lambda \boldsymbol{M}^{-1}), \tag{12}$$

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

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

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

$$\begin{aligned}
g(\omega) &\le \min_{\theta_{\mathsf{R}} \in \mathcal{C}} \left\{ \frac{3}{2} \begin{bmatrix} \omega - 1 \\ \theta_{\mathsf{R}} \end{bmatrix}^\top \mathbf{\Sigma}_P \begin{bmatrix} \omega - 1 \\ \theta_{\mathsf{R}} \end{bmatrix} - 2 \begin{bmatrix} \omega - 1 \\ \theta_{\mathsf{R}} \end{bmatrix}^\top v + 2\lambda \right\} \\
&= T_1(\omega) + T_2(\omega) + 2\lambda + \min_{\theta_{\mathsf{R}} \in \mathcal{C}} T_3(\theta_{\mathsf{R}}).
\end{aligned} \tag{13}$$

Above, we have defined

$$T_1(\omega) := \frac{3}{2} \frac{(\omega - 1)^2}{B}, \quad T_2(\omega) := -2 v_1 (\omega - 1), \quad \text{and} \quad T_3(\theta_{\mathsf{R}}) := \frac{3}{2} \|\theta_{\mathsf{R}}\|_2^2 - 2 v_{\mathsf{R}}^\top \theta_{\mathsf{R}}, \tag{14}$$

and we have used the decomposition $v = (v_1, v_{\mathsf{R}})^\top$. 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

$$T_1(\omega) = \frac{3}{2} \frac{(1 - \omega)^2}{B} \le \frac{3}{2} \frac{(1 - \omega^2)^2}{B} = \frac{3}{2} \frac{B^2}{n^2}. \tag{15}$$

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

$$T_2(\omega) = 2 v_1 (1 - \omega) \le 2|v_1|(1 - \omega) \le 2|v_1|(1 - \omega^2) = 2|v_1| \cdot \frac{B^{3/2}}{n}. \tag{16}$$

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

LEMMA 2.4. *The following holds true with probability at least 0.99*

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

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

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

The claim then follows from Chebyshev's inequality. □

Lemma 2.4 demonstrates that $|v_1| \leq \frac{10}{\sqrt{nB}}$, with probability at least $99/100$. Therefore, on this event, the bound (16) guarantees

$$T_2(\omega) \leq 20 \frac{B}{n^{3/2}}. \tag{17}$$

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

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

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

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

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 $\xi^\star \geq \frac{1}{B}$. In this case, we have

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

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

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

  where $c > 0$ is a constant. Here in view of Lemma 2.3, 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,

$$\min_{\theta_{\mathsf{R}}\in\mathcal{C}} T_3(\theta_{\mathsf{R}}) \le -c_1\frac{B^{1/2}}{n} \tag{18}$$

for some constant $c_1 > 0$.

Completing the proof: We can now combine bounds (15), (17), and (18) on the terms T_1, T_2, T_3, respectively. Note that when $n \ge 7B^{3/2} \ge 7$, all three events and the upper bound (13) hold simultaneously, with probability $1-(\frac{1}{n}+\frac{1}{100}+\frac{1}{10}) \ge 3/4$. Therefore, we obtain

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

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

2.4.2. *Proof of Lemma 2.2.* We will prove the slightly stronger claim that with probability at least $3/4$, we have

$$\inf_{\substack{t\in[0,1] \\ 1-t^2\le\beta B^{3/2}/n}} g(t) > -c^*\frac{\sqrt{B}}{n} \tag{19}$$

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 $(1-t)^2 \le \frac{\beta^2}{4}\frac{B^3}{n^2}$, then $(1-t^2)^2 \le \beta^2\frac{B^3}{n^2}$. Hence, (19) proves the claim as soon as $c_3 = \beta^2/4$.

Proof of bound (19): On the event (12), if $\theta = (\theta_1, \theta_{\mathsf{R}})^\top$ obeys $\theta^\top \boldsymbol{M}^{-1}\theta \le 1$, then we have the lower bound

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

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

$$\begin{aligned} g(\theta_1) &\ge \min_{\theta_{\mathsf{R}}\in\mathcal{C}(\theta_1)}\left\{\frac{1}{2}(\theta-\theta^\star)^\top\boldsymbol{\Sigma}_P(\theta-\theta^\star) - 2(\theta-\theta^\star)^\top v - 2\lambda\right\} \\ &= \min_{\theta_{\mathsf{R}}\in\mathcal{C}(\theta_1)}\left\{\frac{1}{2}\frac{(\theta_1-1)^2}{B} - 2v_1(\theta_1-1) + \frac{1}{2}\|\theta_{\mathsf{R}}\|_2^2 - 2v_{\mathsf{R}}^\top\theta_{\mathsf{R}} - 2\lambda\right\} \\ &\ge -T_2(\theta_1) - 2\lambda + \min_{\theta_{\mathsf{R}}\in\mathcal{C}(\theta_1)}\{\frac{1}{2}\|\theta_{\mathsf{R}}\|_2^2 - 2v_{\mathsf{R}}^\top\theta_{\mathsf{R}}\}. \qquad (20) \end{aligned}$$

where the last line identifies $-2v_1(\theta_1-1)$ with $T_2(\theta_1)$ (cf. definition (14)).

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

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

$$g(\theta_1) \geq -2\lambda = -\frac{2C\log n}{n}. \tag{21}$$

Case 2: $\theta_1 \in [0,1)$**.** We lower bound the terms in equation (20) in turn.

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

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

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

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

and step (ii) relies on Lemma 2.4 and the constraint $1-\theta_1^2 \leq \beta\frac{B^{3/2}}{n}$.

- **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

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

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

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

where $\tilde{C} > 0$ is a constant. Here, since B is sufficiently large, Lemma 2.3 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 $1-\theta_1^2 \leq \beta\frac{B^{3/2}}{n}$,

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

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

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

3. Proofs of the bounds (39). By definition, any function $h \in \mathcal{F}^\star$ obeys $\|h\|_{\mathcal{H}} \leq 3\|f^\star\|_{\mathcal{H}}$. In terms of the expansion $h = \sum_{j=1}^\infty \theta_j \phi_j$, this constraint is equivalent to the bound $\sum_{j=1}^\infty \theta_j^2/\mu_j \leq 9\|f^\star\|_{\mathcal{H}}^2$. In addition, the constraint $\|h\|_Q \leq r$ implies that $\sum_{j=1}^\infty \theta_j^2 \leq r^2$. In conjunction, these two inequalities imply that

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

as claimed in inequality (39b).

We now use this inequality to establish the bound (39a). For any $x \in \mathcal{X}$, we have

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

Here step (i) uses the Cauchy–Schwarz inequality, whereas step (ii) follows from the previous claim (39b) and the assumption that $|\phi_j(x)| \leq 1$ for all $j \geq 1$.

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

$$\mathcal{M}^{\mathsf{new}}(\delta) := c_0 \sqrt{\tfrac{\sigma^2 V^2 \log^3(n)}{n} \Psi(\delta, \mu)} \left(\sqrt{\tfrac{\Psi(\delta,\mu)}{\sigma^2}} + 1 \right). \tag{22}$$

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

THEOREM 4.1. *Consider a kernel with sup-norm bounded eigenfunctions* (19)*, and a source-target pair with* $\mathbb{E}_P[\rho^2(X)] \leq V^2$. *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*

$$\|\widehat{f}_\lambda^{\mathrm{rw}} - f^\star\|_Q^2 \leq \delta_n^2 \tag{23}$$

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

PROOF. Inspecting the proof of Theorem 4.1 (in particular, equation (38)), one has with high probability that

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

Repeating the analysis in Section 5.2 with $\delta_\lambda = \delta_n$ yields the desired claim. □

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 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 3.1. Similarly, in Section 5.2, 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 4.1.

5.1. *Bounded likelihood ratio.*

THEOREM 5.1. *Consider a covariate-shifted regression problem with likelihood ratio that is B-bounded* (3) *over a Hilbert space with a κ-uniformly bounded kernel* (9). *There are universal constants $c_1, c_2 > 0$ such that if $\lambda \geq c_1 \frac{\kappa^2 \log n}{n}$, the KRR estimate $\widehat{f}_\lambda$ satisfies the bound*

$$\mathbb{E}\big[\|\widehat{f}_\lambda - f^\star\|_Q^2\big] \leq c_2 \Big\{ \lambda B \|f^\star\|_{\mathcal{H}}^2 + \frac{\sigma^2 B}{n} \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda B} + \frac{\sigma^2}{n} \Big\}. \tag{24}$$

Inspecting the proof, one may take $c_1 = 32, c_2 = \frac{519}{256}$. The proof of this result is presented in Section 5.1.1.

An immediate consequence is the following result for regular kernels. Note that it matches our lower bound (see Theorem 3.3), apart from logarithmic factors.

COROLLARY 5.2. *Suppose $\sigma^2 \geq \kappa^2$ and $\|f^\star\|_{\mathcal{H}} = 1$. For any $B \geq 1$ and any pair (P, Q) with B-bounded likelihood ratio* (3)*, any orthonormal basis $\{\phi_j\}_{j \geq 1}$ of $L^2(Q)$, and any regular sequence of kernel eigenvalues $\{\mu_j\}_{j \geq 1}$, there exist a universal constant $C > 0$ such that*

$$\mathbb{E}\big[\|\widehat{f}_\lambda - f^\star\|_Q^2\big] \leq C \inf_{\delta > 0} \Big\{ \delta^2 + \sigma^2 B d(\delta) \frac{\log n}{n} \Big\}, \tag{25}$$

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 3.2, we obtain from the KRR risk bound of Theorem 5.1,

$$\mathbb{E}\Big[\|\widehat{f}_\lambda - f^\star\|_Q^2\Big]\Big] \leq C_1 \Big\{ \delta^2 + \sigma^2 B d(\delta) \frac{\log n}{n} \Big\}, \quad \text{where} \quad \delta^2 = \lambda B, \tag{26}$$

for any $\delta^2 \geq c_1 B \kappa^2 \frac{\log n}{n}$. Adjusting constants so that $c \geq c_1$, our choice of δ_n^2 is valid since $\sigma^2 \geq \kappa^2$ and $d(\delta_n) \geq 1$. 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

$$\Big\{ \delta_n^2 + \sigma^2 B d(\delta_n) \frac{\log n}{n} \Big\} \leq C_2 \inf_{\delta > 0} \Big\{ \delta^2 + \sigma^2 B d(\delta) \frac{\log n}{n} \Big\}, \tag{27}$$

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

5.1.1. *Proof of Theorem 5.1.* Using Parseval's theorem and the optimality conditions for the KRR problem as given in equation (29), we have $\mathbb{E}[\|\widehat{f}_\lambda - f^\star\|_Q^2] \le \mathbb{E}[T_1] + \mathbb{E}[T_2]$ where

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

Recall the event

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

as defined in equation (27). We use this event to bound the two terms. **Bound for** T_1 Inspecting the proof of Theorem 3.1 (specifically, see the proof of bound (30)(a)), it follows that $\mathbb{E}[T_1 \mathbf{1}_{\mathcal{E}(\lambda)}] \le 2\lambda B \|f^\star\|_{\mathcal{H}}^2$. On the other hand, from inequality (ii) of the proof of (30)(a), it also holds that

$$\mathbb{E}[T_1 \mathbf{1}_{\mathcal{E}(\lambda)^c}] \le \|f^\star\|_{\mathcal{H}}^2 \|\boldsymbol{M}\|_2 \mathbb{P}(\mathcal{E}(\lambda)^c) \le \|f^\star\|_{\mathcal{H}} \kappa^2 \mathbb{P}(\mathcal{E}(\lambda)^c).$$

The final inequality holds since $\|\boldsymbol{M}\|_2 \le \mathsf{trace}(\boldsymbol{M}) = \mathbb{E}_Q[\sum_j \mu_j \phi_j^2(x)] \le \kappa^2$. Now, note that whenever $n\lambda \ge 32\kappa^2 \log n$, by Lemma 5.1 we have that

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

Putting the pieces together, we obtain

$$\mathbb{E}[T_1] \le \frac{519}{256}\lambda \|f^\star\|_{\mathcal{H}}^2. \tag{28}$$

Bound for T_2 By considering the expectation over w_i conditional on the covariates and following algebraic manipulations similar to the proof of bound (30)(b), we have

$$\mathbb{E}[T_2] \le \mathbb{E}[\widetilde{T}_2], \quad \text{where} \quad \widetilde{T}_2 := \mathsf{trace}\Big(\frac{\sigma^2}{n}(\widehat{\boldsymbol{\Sigma}}_P + \lambda \boldsymbol{M}^{-1})^{-1}\Big).$$

Moreover, inspecting the proof of bound (30)(b), we also have

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

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

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

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

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

5.2. *Unbounded likelihood ratio.*

THEOREM 5.3. *Suppose $\sigma^2 \geq \kappa^2$ and $\|f^\star\|_{\mathcal{H}} = 1$. Consider a kernel with sup-norm bounded eigenfunctions* (19)*, and a source-target pair with $\mathbb{E}_P[\rho^2(X)] \leq V^2$. Then, for any orthonormal basis $\{\phi_j\}_{j\geq 1}$ of $L^2(Q)$ and any regular sequence of kernel eigenvalues $\{\mu_j\}_{j\geq 1}$, there exists a universal constant $C > 0$ such that,*

$$\mathbb{E}\Big[\|\widehat{f}^{\mathrm{rw}}_\lambda - f^\star\|_Q^2\Big] \leq C \inf_{\delta>0}\Big\{\delta^2 + V^2 d(\delta)\frac{\log^3 n}{n}\Big\}. \tag{29}$$

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

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

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

$$\mathbb{E}\Big[\|\widehat{f}^{\mathrm{rw}}_\lambda - f^\star\|_Q^2\Big] \leq c_2 \inf_{\delta>0}\Big\{\delta^2 + \frac{\sigma^2 V^2 \log^3(n) d(\delta)}{n}\Big\} + \mathbb{E}\Big[\|\widehat{f}^{\mathrm{rw}}_\lambda - f^\star\|_\infty^2 \mathbf{1}_{\mathcal{E}^c}\Big]. \tag{30}$$

By Cauchy-Schwarz,

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

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

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

Therefore, combining the previous two displays,

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

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

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

Above, inequality (i) uses $\sigma^2 \geq \kappa^2$ and $V^2 \geq 1$. Inequality (ii) uses the fact that $\lambda \asymp \delta^2 \asymp \frac{\sigma^2 V^2 \log^3(n) d(\delta)}{n} \gtrsim \frac{\kappa^2}{n}$. Finally, inequality (iii) follows by defining $c_5 \geq c_3(1+c_4)$. This bound furnishes the result, since by applying it to the inequality (30), we obtain the result with $C = c_2 + c_5$. □

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

$$\|\widehat{f}_\lambda - f^\star\|_Q^2 \leq 2\sqrt{\lambda V^2 \kappa^2}\|f^\star\|_{\mathcal{H}}^2 + 40\frac{\sigma^2 \log n}{n} \cdot \frac{\kappa^2}{\lambda} \tag{31}$$

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 $(\frac{\sigma^2 V^2}{n})^{1/3}$ (ignoring κ^2 and $\log$ factors). However, unfortunately, this is far from optimal.

6.1. *Proof of Theorem 6.1.* In view of the proof of Theorem 3.1, we know that

$$\begin{aligned} \|\widehat{f}_\lambda - f^\star\|_Q^2 &\leq 4\lambda\|f^\star\|_{\mathcal{H}}^2 \|\!|\boldsymbol{M}^{1/2}(\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_P\boldsymbol{M}^{1/2} + \lambda\boldsymbol{I})^{-1}\boldsymbol{M}^{1/2}|\!\|_2 \\ &\quad + 40\frac{\sigma^2 \log n}{n}\mathsf{trace}\Big(\boldsymbol{M}^{1/2}(\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_P\boldsymbol{M}^{1/2} + \lambda\boldsymbol{I})^{-1}\boldsymbol{M}^{1/2}\Big) \end{aligned} \tag{32}$$

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:

$$\|\!|\boldsymbol{M}^{1/2}(\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_P\boldsymbol{M}^{1/2} + \lambda\boldsymbol{I})^{-1}\boldsymbol{M}^{1/2}|\!\|_2 \leq \frac{1}{2}\sqrt{\frac{V^2\kappa^2}{\lambda}}; \tag{33a}$$

$$\mathsf{trace}\Big(\boldsymbol{M}^{1/2}(\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_P\boldsymbol{M}^{1/2} + \lambda\boldsymbol{I})^{-1}\boldsymbol{M}^{1/2}\Big) \leq \frac{\kappa^2}{\lambda}. \tag{33b}$$

Proof of the bound (33b)**:** Note that $\boldsymbol{M}^{1/2}(\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_P\boldsymbol{M}^{1/2} + \lambda\boldsymbol{I})^{-1}\boldsymbol{M}^{1/2} \preceq \lambda^{-1}\boldsymbol{M}$. We therefore have

$$\mathsf{trace}\Big(\boldsymbol{M}^{1/2}(\boldsymbol{M}^{1/2}\boldsymbol{\Sigma}_P\boldsymbol{M}^{1/2} + \lambda\boldsymbol{I})^{-1}\boldsymbol{M}^{1/2}\Big) \leq \mathsf{trace}(\lambda^{-1}\boldsymbol{M}) \leq \frac{\kappa^2}{\lambda},$$

where the last relation uses the fact that $\mathsf{trace}(\boldsymbol{M}) \leq \kappa^2$.

Proof of the bound (33a)**:** We first make the observation that the bound (33a) is equivalent to

$$\boldsymbol{\Sigma}_P + \lambda\boldsymbol{M}^{-1} \succeq 2\sqrt{\frac{\lambda}{V^2\kappa^2}}\boldsymbol{I}. \tag{34}$$

Therefore from now on, we focus on establishing the bound (34). Take an arbitrary vector θ with $\|\theta\|_2 = 1$. We have

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

Here, the identity (i) follows from the fact that $\mathbb{E}_Q[\phi(X)\phi(X)^\top] = \boldsymbol{I}$, 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

$$(\theta^\top \phi(X))^2 \le \|\boldsymbol{M}^{-1/2}\theta\|_2^2 \cdot \|\boldsymbol{M}^{1/2}\phi(X)\|_2^2 \le \kappa^2 \|\boldsymbol{M}^{-1/2}\theta\|_2^2,$$

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

$$\mathbb{E}_P[(\theta^\top \phi(X))^2] \ge \frac{1}{V^2\kappa^2 \cdot (\theta^\top \boldsymbol{M}^{-1}\theta)} \qquad \text{for any } \theta \text{ with } \|\theta\|_2 = 1.$$

As a result, one has

$$\theta^\top (\boldsymbol{\Sigma}_P + \lambda \boldsymbol{M}^{-1})\theta \ge \frac{1}{V^2\kappa^2 \cdot (\theta^\top \boldsymbol{M}^{-1}\theta)} + \lambda \theta^\top \boldsymbol{M}^{-1}\theta \ge 2\sqrt{\frac{\lambda}{V^2\kappa^2}}.$$

Since this inequality holds for any unit-norm θ, we establish the claim (34).

7. Auxiliary lemmas.

The following lemma provides concentration inequalities for the sum of independent self-adjoint operators, which appeared in the work [2].

LEMMA 7.1. *Let $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n$ be i.i.d. self-adjoint operators on a separable Hilbert space. Assume that $\mathbb{E}[\boldsymbol{Z}_1] = \boldsymbol{0}$, and $\|\|\boldsymbol{Z}_1\|\|_2 \le L$ for some $L > 0$. Let $\boldsymbol{V}$ be a positive trace-class operator such that $\mathbb{E}[\boldsymbol{Z}_1^2] \preceq \boldsymbol{V}$, and $\|\|\mathbb{E}[\boldsymbol{Z}_1^2]\|\|_2 \le R$. Then one has*

$$\mathbb{P}\Big(\Big\|\Big|\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\Big\|\Big|_2 \ge t\Big) \le \frac{28\mathsf{trace}(\boldsymbol{V})}{R} \cdot \exp\Big(-\frac{nt^2/2}{R + Lt/3}\Big), \qquad \text{for all } t \ge \sqrt{R/n} + L/(3n).$$

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 $f \in \mathcal{F}$, $\mathbb{E}[f(X_i)] = 0$. 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 $v^2 := \sup_{f\in\mathcal{F}} \mathbb{E}[\sum_{i=1}^n (f(X_i))^2]$ is crucial. Now we are in position to state the classical Talagrand's concentration inequalities; see the paper [1].

LEMMA 7.2. *For all $t > 0$, we have*

$$\mathbb{P}(Z \ge \mathbb{E}[Z] + t) \le \exp\Big(-\frac{t^2}{2(v^2 + 2v\mathbb{E}[Z]) + 3vt}\Big). \tag{35}$$

REFERENCES

[1] KLEIN, T. and RIO, E. (2005). Concentration around the mean for maxima of empirical processes. *The Annals of Probability* **33** 1060–1077.

[2] MINSKER, S. (2017). On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters* **127** 111–119.

[3] YANG, Y., PILANCI, M. and WAINWRIGHT, M. J. (2017). Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics* **45** 991–1023.