Bernoulli **30**(3), 2024, 1845–1877
https://doi.org/10.3150/23-BEJ1655

Local convergence rates of the nonparametric least squares estimator with applications to transfer learning

JOHANNES SCHMIDT-HIEBER[a] and PETR ZAMOLODTCHIKOV[b]

Department of Applied Mathematics, University of Twente, Enschede, The Netherlands,
[a]*a.j.schmidt-hieber@utwente.nl,* [b]*p.zamolodtchikov@utwente.nl*

Convergence properties of empirical risk minimizers can be conveniently expressed in terms of the associated population risk. To derive bounds for the performance of the estimator under covariate shift, however, pointwise convergence rates are required. Under weak assumptions on the design distribution, it is shown that least squares estimators (LSE) over 1-Lipschitz functions are also minimax rate optimal with respect to a weighted uniform norm, where the weighting accounts in a natural way for the non-uniformity of the design distribution. This implies that although least squares is a global criterion, the LSE adapts locally to the size of the design density. We develop a new indirect proof technique that establishes the local convergence behavior based on a carefully chosen local perturbation of the LSE. The obtained local rates are then applied to analyze the LSE for transfer learning under covariate shift.

Keywords: Covariate shift; domain adaptation; local rates; mean squared error; minimax estimation; nonparametric least squares; nonparametric regression; transfer learning

1. Introduction

Consider the nonparametric regression model with random design supported on $[0,1]$, that is, we observe n i.i.d. pairs $(X_1,Y_1),\dots,(X_n,Y_n)\in[0,1]\times\mathbb{R}$, with

$$Y_i = f_0(X_i)+\varepsilon_i, \quad i=1,\dots,n \tag{1}$$

and independent measurement noise variables $\varepsilon_1,\dots,\varepsilon_n\sim\mathcal{N}(0,1)$. The design distribution is the marginal distribution of X_1 and is denoted by $\mathrm{P_X}$. Throughout this paper, we assume that $\mathrm{P_X}$ has a Lebesgue density p. The least squares estimator (LSE) for the nonparametric regression function f taken over a function class $\mathcal{F}$ is given by

$$\widehat{f}_n \in \operatorname*{arg\,min}_{f\in\mathcal{F}} \sum_{i=1}^n \big(Y_i - f(X_i)\big)^2.$$

In the case of non-uniqueness, the subsequent discussion and analysis applies to any minimizer $\widehat{f}_n$. If the class $\mathcal{F}$ is convex, computing the estimator $\widehat{f}_n$ results in a convex optimization problem, which can also be written as a quadratic programming problem, see [3]. For a fixed function f, the law of large numbers implies that the least squares objective $\sum_i(Y_i - f(X_i))^2$ is close to its expectation $n\,\mathrm{E}[(Y_1-f(X_1))^2]=n+n\,\mathrm{E}[\int_0^1(f_0(x)-f(x))^2p(x)\,dx]$. It is therefore natural that the standard analysis of LSEs based on empirical process methods and metric entropy bounds for the function class $\mathcal{F}$ leads to convergence rates with respect to the empirical L^2-loss $\|\widehat{f}-f_0\|_n^2 := \frac{1}{n}\sum_{i=1}^n(\widehat{f}_n(X_i)-f_0(X_i))^2$ and

1350-7265 © 2024 ISI/BS

the associated population version $\mathrm{E}[\int_0^1 (\widehat{f}_n(x) - f_0(x))^2 p(x)\,dx]$, see, for instance, [16,26,42,48]. The latter risk is the expected squared loss if a new X is sampled from the design distribution $\mathrm{P_X}$ and $f_0(X)$ is estimated by $\widehat{f}_n(X)$.

A widely observed phenomenon is that the distribution of the new X is different from the design distribution of the training data. For example, assume that we want to predict the response Y of a patient to a drug based on a measurement X summarizing the patient's health status. To learn such a relationship, data are collected in one hospital resulting in an estimator $\widehat{f}_n$. Later $\widehat{f}_n$ will be applied to patients from a different hospital. It is conceivable that the distribution of X in the other hospital is different. For instance, there could be a different age distribution, or patients have a different socio-economic status due to variations in the imposed treatment costs.

Therefore, an important problem is to evaluate the estimator's expected squared risk if a new observation X is sampled from a different design distribution $\mathrm{Q_X}$ with density q. The associated prediction error under the new distribution is then

$$\int_0^1 \big(\widehat{f}_n(x) - f_0(x)\big)^2 q(x)\,dx. \tag{2}$$

If $\mathrm{P_X}$ and $\mathrm{Q_X}$ are similar enough so that for some finite constant C and any $x \in [0,1]$, $q(x) \le Cp(x)$, then, the prediction error under the design density q is of the same order as under p. However, in machine learning applications, there are often subsets of the domain with very few data points. This motivates the relevance of the problematic case, where the density q is large in a low-density region of p. Differently speaking, we are more likely to see a covariate X in a region with few training data based on the sample $(X_1,Y_1),\dots,(X_n,Y_n)$. Since the lack of data in such a region means that the LSE will not fit the true regression function f_0 well, this could lead to a large prediction error under the new design distribution.

An extension of this problem setting is transfer learning under covariate shift. Here we know the least-squares estimator $\widehat{f}_n$ and the sample size n based on the sample $(X_1,Y_1),\dots,(X_n,Y_n)$ with design density p. On top of that, we have a second, smaller dataset with $m \ll n$ new i.i.d. data points $(X_1',Y_1'),\dots,(X_m',Y_m')$ where $Y_i' = f_0(X_i') + \varepsilon_i'$, $i = 1,\dots,m$ and $X_1' \sim \mathrm{Q_X}$. In the framework of the hospital data above, this means that we also have data from a small study with m patients from the second hospital. In other words, the regression function f_0 remains unchanged, but the design distribution changes. Since the number of extra training data points m is small compared to the original sample size n, we want to quantify how well an estimator combining $\widehat{f}_n$ and the new sample can predict under the new design distribution with associated prediction error (2). Establishing convergence rates for the loss in (2), given a sample with design density p, is, however, a hard problem. To our knowledge, no simple modification of the standard least squares analysis allows to obtain optimal rates for this loss.

To address this problem, we study the case where the LSE is selected within the function class $\mathcal{F}$ consisting of all 1-Lipschitz functions. For this setting, we prove under weak assumptions that for a sufficiently large constant K and any $0 \le x \le 1$,

$$\big|\widehat{f}_n(x) - f_0(x)\big| \le K t_n(x) \tag{3}$$

with high probability, where the local convergence rate t_n is the function that is, pointwise, the solution to the equation

$$t_n(x)^2\ \mathrm{P_X}([x \pm t_n(x)]) = \frac{\log n}{n}.$$

The article [13] shows under slightly different assumptions on the design density that t_n is locally the optimal estimation rate and constructs a wavelet thresholding estimator that is specifically designed to attain this local convergence rate. We prove that the LSE achieves this optimal local rate without any tuning. This is surprising since the LSE is based on minimization of the (global) empirical L^2-distance, and convergence in L^2 is weaker than convergence in the weighted sup-norm loss underlying the statement in (3).

To establish (3), we only assume a local doubling property of the design distribution. By imposing more regularity on the design density, we can prove that $t_n(x) \asymp (\log n/(np(x))^{1/3}$. For this result, p is also allowed to depend on the sample size n such that the $p(x)$ in the denominator does not only change the constant but also the local convergence rate. This quantifies how the local convergence rate varies depending on the density p and how small-density regions increase the local convergence rate. In Section 2, we argue that kernel smoothing with fixed bandwidth has a slower convergence rate than the LSE. Therefore, the least squares fit can better recover the regression function if the values of the density p range over different orders of magnitude. This property is particularly important for machine learning applications.

Based on (3), we can then obtain a high-probability bound for the prediction error in (2) by

$$\int_0^1 \big(\widehat{f}_n(x) - f_0(x)\big)^2 q(x)\, dx \le K^2 \int_0^1 t_n(x)^2 q(x)\, dx.$$

In many cases, simpler expressions for the convergence rate can be derived from the right-hand side. For instance in the case $t_n(x) \asymp (\log n/(np(x))^{1/3}$, the convergence rate is $(\log n/n)^{2/3}$ if $\int_0^1 q(x)/p(x)^{2/3}\, dx$ is bounded by a finite constant.

A major contribution of this paper is the proof strategy to establish local convergence rates. For that, we argue by contradiction, first assuming that the LSE has a slower local rate. Afterwards, we construct a local perturbation with smaller least squares loss. This means that the original estimator was not the LSE, leading to the desired contradiction. While a similar strategy has been followed for shape-constrained estimation in [12,14], the construction of the local perturbation and the verification of a smaller least squares loss for Lipschitz functions are both non-standard and involved. We believe that these arguments can be generalized to various extensions beyond Lipschitz function classes.

The paper is structured as follows. In Section 2, we state the new upper and lower bounds on the local convergence rate. This section precedes a discussion on the imposed doubling condition and examples in Section 3. Section 4 gives a high-level overview of the new proof strategy to establish local convergence rates. The full proof can be found in Section 7. Applications to transfer learning are discussed in Section 5. Section 6 provides a brief literature review and an outlook. The remaining proofs are deferred to the supplement [39].

Notation: For two real numbers a, b, we write $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$. For any real number x, we denote by $\lceil x \rceil$ the smallest integer m such that $m \ge x$ and by $\lfloor x \rfloor$ the greatest integer m such that $m \le x$. Furthermore, for any set S, we denote by $x \mapsto \mathbb{1}(x \in S)$ the indicator function of the set S. To increase readability of the formulas, we define $[a \pm b] := [a-b, a+b]$. For any two positive sequences $\{a_n\}_n, \{b_n\}_n$, we say that $a_n \lesssim b_n$ if there exists a constant $0 < c < \infty$, and a positive integer N such that for all $n \ge N, a_n \le cb_n$. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Finally, if for all $\varepsilon > 0$, there exists a positive integer N such that for all $n \ge N, a_n \le \varepsilon b_n$, then we write $a_n \ll b_n$. For a random variable X and a (measurable) set A, $\mathrm{P}_X(A)$ stands for $\mathrm{P}(X \in A)$. For any function h for which the integral is finite, we set $\|h\|_{L^2(\mathrm{P})} := \big(\int h^2(x)p(x)\, dx\big)^{1/2}$. We also write $\|h\|_n := \big(\frac{1}{n}\sum_{i=1}^n h^2(X_i)\big)^{1/2}$.

2. Main results

In this section, we state the local convergence results for the LSE. Set

$$\mathcal{M} := \{\text{probability measures that are both supported on } [0,1] \text{ and admit a Lebesgue density}\}.$$

The local convergence rate t_n turns out to be the functional solution to an equation that depends on the design distribution $\mathrm{P_X}$.

Lemma 1. *If* $\mathrm{P_X} \in \mathcal{M}$*, then, for any* $n > 1$ *and any* $x \in [0,1]$*, there exists a unique solution* $t_n(x)$ *of the equation*

$$t_n(x)^2 \, \mathrm{P_X}\left(\left[x \pm t_n(x)\right]\right) = \frac{\log n}{n}.$$

Therefore the function $x \mapsto t_n(x)$ *is well defined on* $[0,1]$*. From now on, we refer to* t_n *as the spread function (associated to* $\mathrm{P_X}$*).*

The spread function can be viewed as a measure for the local mass of the distribution $\mathrm{P_X}$ around x. The more mass $\mathrm{P_X}$ has around x, the smaller $t_n(x)$ is. Whenever necessary, the spread function associated to a probability distribution P is denoted by t_n^{P}.

To derive a local convergence rate of the least-squares estimator taken over Lipschitz functions, one has to exclude the possibility that the design distribution $\mathrm{P_X}$ is completely erratic. Interestingly, no Hölder smoothness has to be imposed on the design density, and it is sufficient to consider design distributions satisfying the following weak regularity assumption.

Definition 2. For $n \geq 3$ and $D \geq 2$, define $\mathcal{P}_n(D)$ as the class of all design distributions $\mathrm{P_X} \in \mathcal{M}$, such that for any $0 < \eta \leq \sqrt{\log n} \sup_{x \in [0,1]} t_n(x)$,

$$\sup_{x \in [0,1]} \frac{\mathrm{P_X}\left([x - 2\eta, x + 2\eta]\right)}{\mathrm{P_X}\left([x - \eta, x + \eta]\right)} \leq D. \tag{LDP}$$

We call (LDP) the local D-doubling property, or local doubling property when the constant D is irrelevant or unambiguous. A design distribution $\mathrm{P_X}$ satisfies the (global) doubling property if (LDP) holds for all $\eta > 0$. Denote by $\mathcal{P}_G(D)$ the space of all globally doubling measures in $\mathcal{M}$.

The restriction $x \in [0,1]$ allows to include distributions with Lebesgue densities that are discontinuous at 0 or 1. For instance the uniform distribution on $[0,1]$ is 2-doubling, but since $\mathrm{P_X}[-3\eta, \eta] / \mathrm{P_X}[-2\eta, 0] = \infty$, (LDP) does not hold if the supremum includes $x = -\eta$.

Since the uniform distribution on $[0,1]$ is contained in $\mathcal{P}_n(2) \subseteq \mathcal{P}_n(D)$ for $D \geq 2$, we see that these classes are non-empty. Inequality (LDP) states that doubling the size of a small interval cannot inflate its probability by more than a factor D. The next result shows that the maximum interval size $\sqrt{\log n} \sup_{x \in [0,1]} t_n(x)$ tends to zero as n becomes large.

Lemma 3. *Let* $\mathrm{P_X} \in \mathcal{P}_n(D)$ *with* $D \geq 2$*. If* $\varepsilon > 0$*, then there exists an* $N = N(\varepsilon, D)$ *such that for all* $n \geq N$, $\sqrt{\log n} \sup_{x \in [0,1]} t_n(x) < \varepsilon$*.*

The local doubling condition allows us to consider sample size-dependent design distributions. See Section 3 for a more in-depth discussion and some examples.

We now show that the spread function is indeed the minimax rate. Denote by $\mathrm{Lip}(\kappa)$ the set of functions $f\colon [0,1] \to \mathbb{R}$ that are Lipschitz, with Lipschitz constant at most κ, that is, $f \in \mathrm{Lip}(\kappa)$ iff $|f(x) - f(y)| \le \kappa|x - y|$ for all $x, y \in [0,1]$. If $f \in \mathrm{Lip}(\kappa)$, then we also say that f is κ-Lipschitz. Recall that P_{f_0} is the distribution of the data in the nonparametric regression model (1) if the true regression function is f_0 and that $\mathrm{P_X}$ denotes the distribution of the design X.

Theorem 4. *Consider the nonparametric regression model* (1)*. Let* $0 < \delta < 1$*, and* $D \ge 2$*. If* $\widehat{f}_n$ *denotes the LSE taken over the class of* 1*-Lipschitz functions* $\mathrm{Lip}(1)$*, then, for a sufficiently large constant* $K = K(D, \delta)$*,*

$$\sup_{\mathrm{P_X} \in \mathcal{P}_n(D)} \ \sup_{f_0 \in \mathrm{Lip}(1-\delta)} \mathrm{P}_{f_0}\left(\sup_{x \in [0,1]} t_n(x)^{-1} |\widehat{f}_n(x) - f_0(x)| > K \right) \to 0 \quad \text{as } n \to \infty.$$

The proof reveals that if the constant K is chosen as the value K_* defined in (34), the right-hand side of Theorem 4 converges to zero with a polynomial rate in the sample size n. For $\delta \to 0$, $K_* \asymp \delta^{-1/2 - 3\log_2(D)/4}$. Consequently, the constant K will become large for small δ and large doubling constant D. We want to stress that no attempt has been made to optimize the constants and that further refinements of the inequalities in the proof will likely improve the constant K considerably.

Since the previous result is uniform over design distributions $\mathrm{P_X} \in \mathcal{P}_n(D)$, we can also consider sequences P_X^n. While, at first sight, it might appear unnatural to consider for every sample size n a different design distribution, this constitutes a useful statistical concept to study the effect of low-density regions on the convergence rate. Indeed, the influence of a small density region disappears in the constant for a fixed density, while the dependence on the sample size makes the effect visible in the convergence rate. Moreover, sample size-dependent quantities are widely studied in mathematical statistics, most prominently in high-dimensional statistics, where the number of parameters typically grows with the sample size.

One key question is to identify conditions for which the local convergence rate t_n has a more explicit expression. One such instance is the case of Hölder-smooth design densities. Let $\lfloor \beta \rfloor$ denote the largest integer that is strictly smaller than β. The Hölder-β semi-norm of a function $g : \mathbb{R} \to \mathbb{R}$ is defined as

$$|g|_\beta := \sup_{x,y \in \mathbb{R},\, x \neq y} \frac{|g^{(\lfloor \beta \rfloor)}(x) - g^{(\lfloor \beta \rfloor)}(y)|}{|x - y|^{\beta - \lfloor \beta \rfloor}}. \tag{4}$$

For $\beta = 1$, $|g|_\beta$ is the Lipschitz constant of g.

Corollary 5. *Consider the nonparametric regression model* (1)*. Let* $0 < \delta < 1$ *and* $\widehat{f}_n$ *be the LSE taken over the class of* 1*-Lipschitz functions* $\mathrm{Lip}(1)$*. For* $\beta \in (0, 2]$*, let* P_X^n *be a sequence of distributions with corresponding Lebesgue densities* p_n*. If for any* n*, there exists a non-negative function* h_n *with* $p_n(x) = h_n(x)$ *for all* $x \in [0,1]$*,* $\max_n |h_n|_\beta \le \kappa$ *and* $\min_{x \in [0,1]} p_n(x) \ge n^{-\beta/(3+\beta)} \log n$*, then, for all* $n \ge \exp(4\kappa) \vee 9$*,*

$$\left(\frac{\log n}{3 n p_n(x)} \right)^{1/3} \le t_n(x) \le \left(\frac{2 \log n}{n p_n(x)} \right)^{1/3}, \tag{5}$$

$\mathrm{P}_\mathrm{X}^n \in \mathcal{P}_n(2 + 2^{\beta/3} 3^\beta \kappa + 2^{1/3} 3 \kappa^{1/\beta})$*, and there exists a finite constant* K' *independent of the sequence* P_X^n*, such that*

$$\sup_{f_0 \in \mathrm{Lip}(1-\delta)} \mathrm{P}_{f_0}\left(\sup_{x \in [0,1]} p_n(x)^{1/3} \left| \widehat{f}_n(x) - f_0(x) \right| \ge K' \left(\frac{\log n}{n} \right)^{1/3} \right) \to 0 \quad \text{as } n \to \infty.$$

In the previous result, the regression function is assumed to be Lipschitz, and β denotes the smoothness index of the design densities p_n. The convergence rate $(\log n/n)^{1/3}$ is known to be the optimal nonparametric rate for Lipschitz regression functions, sup-norm loss and uniform fixed design, cf. [44], Corollary 2.5.

The rate $(\log n/(np_n(x))^{1/3}$ is natural, since $np_n(x)$ can be viewed as local effective sample size around x.

For $\beta \in (0,1]$, we can always choose $h_n(x) = p_n(0)$ for $x < 0$, $h_n(x) = p_n(x)$ for $x \in [0,1]$, and $h_n(x) = p_n(1)$ for $x > 1$. While the rate is independent of the smoothness index β, we can allow faster decaying low density regions if β gets larger. The fastest possible decay is $n^{-2/5}\log n$ if $\beta = 2$.

To extend the result to $\beta > 2$ and to allow for even smaller densities, it is widely believed that imposing Hölder smoothness is insufficient. One way around this is to use Hölder smoothness plus some extra flatness constraint. See [34,35] for more on this topic.

The lower bound on the small density regions in Corollary 5 ensures that the local doubling property (LDP) is satisfied. A lower bound is also necessary, as otherwise $p_n(x) \ll \log n/n$ would imply that the rate $t_n(x) \asymp (\log n/(np_n(x))^{1/3}$ diverges. The next lemma shows how the spread function behaves at a point with vanishing Lebesgue density p.

Lemma 6. *Let* $\mathrm{P_X} \in \mathcal{P}_G(D)$ *with* $D \geq 2$ *and density* p. *Suppose that* $p(x_0) = 0$ *for some* $x_0 \in [0,1]$. *If there exists some* $A, \alpha > 0$ *and an open neighbourhood* U *of* x_0 *such that for any* $x \in U$, $1/A \leq |p(x) - p(x_0)|/|x - x_0|^\alpha \leq A$, *then, there exists* $N > 0$, *depending only on* U *and* D, *such that for any* $n > N$,

$$\left(\frac{(\alpha+1)\log n}{An}\right)^{1/(\alpha+3)} \leq t_n(x_0) \leq \left(\frac{(\alpha+1)A\log n}{n}\right)^{1/(\alpha+3)}.$$

An immediate consequence is that if p is k times differentiable, $p^{(\ell)}(x_0) = 0$ for all $\ell < k$, and $p^{(k)}(x_0) \neq 0$, then $t_n(x_0) \asymp (\log n/n)^{1/(k+3)}$.

We complement Theorem 4 with a matching minimax lower bound. A closely related result is Theorem 2 in [13].

Theorem 7. *If* C_∞ *is a positive constant, then there exists a positive constant* c, *such that for any sufficiently large* n, *and any sequence of design distribution* $\mathrm{P}_\mathrm{X}^n \in \mathcal{M}$ *with corresponding Lebesgue densities* p_n *all upper bounded by* C_∞, *we have*

$$\inf_{\widehat{f}_n}\ \sup_{f_0 \in \mathrm{Lip}(1)} \mathrm{P}_{f_0}\left(\sup_{x\in[0,1]} t_n(x)^{-1}|\widehat{f}_n(x) - f_0(x)| \geq \frac{1}{12}\right) \geq c,$$

where the infimum is taken over all estimators.

The proof is deferred to Appendix D in the Supplement [39].

Corollary 5 states that $t_n(x) \asymp (\log n/(np_n(x)))^{1/3}$. Combined with the lower bound, this shows that the local minimax estimation rate in this framework is $(\log n/(np_n(x)))^{1/3}$.

It is known that for Lipschitz functions and squared L^2-loss, the LSE achieves the minimax estimation rate $n^{-2/3}$. Summarizing the statements on the convergence rates above shows that the LSE is also minimax rate optimal with respect to the stronger weighted sup-norm loss.

Next, we discuss how the derived local rates imply several advantages of the LSE if compared to kernel smoothing estimators. In the case of uniform design $p(x) = \mathbb{1}(x \in [0,1])$, the LSE achieves the convergence rate $n^{-2/3}$ with respect to squared L^2-loss and Corollary 5 gives the rate $(\log n/n)^{1/3}$ with respect to sup-norm loss. To our knowledge, it is impossible to obtain these two rates simultaneously for kernel smoothing estimators. The squared L^2 rate $n^{-2/3}$ can be achieved for a kernel bandwidth $h \asymp n^{-1/3}$ and the sup-norm rate $(\log n/n)^{1/3}$ requires more smoothing in the sense that the bandwidth should be of the order $(\log n/n)^{1/3}$, see Corollary 1.2 and Theorem 1.8 in [44]. Any bandwidth choice in the range $n^{-1/3} \lesssim h \lesssim (\log n/n)^{1/3}$ will incur an additional $\log n$-factor in at least one of these two convergence rates of the kernel smoothing estimator. Although the suboptimality in the rate is only a $\log n$-factor, it is surprising that the LSE does not suffer from this issue.

Secondly, we argue that kernel smoothing estimators with fixed global bandwidth cannot achieve the local convergence rate $(\log n/(np_n(x)))^{1/3}$ in the setting of Corollary 5. Denote the bandwidth by h and the kernel smoothing estimator by $\widehat{f}_{nh}$. The decomposition in stochastic error and bias leads to an inequality of the form

$$\left|\widehat{f}_{nh}(x) - f_0(x)\right| \lesssim \underbrace{\sqrt{\frac{\log n}{nhp_n(x)}}}_{\text{stochastic error}} + \underbrace{h}_{\text{deterministic error}}, \quad \text{for all } x \in [0,1], \tag{6}$$

with high probability. Since the dependence on the density $p_n(x)$ is typically ignored, we provide a heuristic for this bound in Appendix B in the supplement [39]. To balance the two errors, one would have to choose as a bandwidth $h \asymp (\log n/(np_n(x)))^{1/3}$. In this case, the local convergence rate would also be $(\log n/(np_n(x)))^{1/3}$. But this requires choosing the bandwidth locally depending on x. From that, one can deduce that any global choice for h in (6) leads to suboptimal local rates. It is, therefore, surprising that although the LSE is based on a global criterion, it changes the amount of smoothing locally to adapt to the amount of data points in each regime. This is a clear advantage of the least squares method over smoothing procedures. This benefit seems particularly advantageous for machine learning problems that typically have high- and low-density regions in the design distribution.

To draw uniform confidence bands, but also for the application to transfer learning discussed later, it is important to estimate the spread function t_n from data. For $\widehat{\mathrm{P}}^n_{\mathrm{X}}(A) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i \in A)$ the empirical design distribution, a natural estimator is

$$\widehat{t}_n(x) := \inf\left\{t : t^2\widehat{\mathrm{P}}^n_{\mathrm{X}}([x \pm t]) \geq \frac{\log n}{n}\right\}. \tag{7}$$

Theorem 8. *If* $\mathrm{P}_{\mathrm{X}} \in \mathcal{P}_G(D)$ *for some* $D \geq 2$ *and* $\|p\|_\infty < \infty$, *then*

$$\max_{n>1} \sup_{x\in[0,1]} \sqrt{\log n}\left|\frac{\widehat{t}_n(x)}{t_n(x)} - 1\right| < \infty, \quad \textit{almost surely,}$$

where $t_n(x)$ *is the spread function associated to* P_{X}.

The result implies that for any $\varepsilon > 0$ and all sufficiently large n, we have $(1-\varepsilon)t_n(x) \leq \widehat{t}_n(x) \leq (1+\varepsilon)t_n(x)$ for all $x \in [0,1]$.

3. Local doubling property and examples of local rates

By Definition 2, $\mathcal{P}_n(D)$ is the class of all locally doubling distributions, and $\mathcal{P}_G(D)$ is the class of all globally doubling distributions in $\mathcal{M}$. It follows from the definitions that $\mathcal{P}_G(D) \subseteq \mathcal{P}_n(D)$. A converse statement is

Lemma 9. *If* $\mathrm{P_X} \in \mathcal{P}_n(D)$*, then* $\mathrm{P_X} \in \mathcal{P}_G(D_n(\mathrm{P_X}))$ *for a finite number* $D_n(\mathrm{P_X})$*. In particular, if* $\mathrm{P_X}$ *does not depend on the sample size n, neither does* $D_n(\mathrm{P_X})$*.*

This means that the distinction between local and global doubling is only relevant in the case where we study sequences of design distributions, such as in the setup of Corollary 5. In Example 3, a sequence P_X^n is constructed such that $\mathrm{P}_\mathrm{X}^n \in \mathcal{P}_n(D)$ for all n and $\mathrm{P}_\mathrm{X}^n \in \mathcal{P}_G(D_n)$ necessarily requires $D_n \longrightarrow \infty$ as $n \longrightarrow \infty$.

Doubling is known to be a weak regularity assumption and does not even imply that $\mathrm{P_X}$ has a Lebesgue density [8,21]. It can, moreover, be easily verified for a wide range of distributions. All distributions with continuous Lebesgue density bounded away from zero and all densities of the form $p(x) \propto x^\alpha$ for $\alpha \geq 0$ are doubling.

Examples of non-doubling measures are distributions $\mathrm{P_X}$ with $\mathrm{P_X}([a,b]) = 0$ for some $0 \leq a < b \leq 1$, see Lemma 29 in Appendix C in the supplement [39]. If a density verifies $p(x_0) = 0$ for some $x_0 \in [0,1]$ and behaves like an inverse exponential around x_0, then it is not in $\mathcal{P}_n$ for any constant. The density $x \mapsto x^{-2}e^{1-1/x}\mathbb{1}(x \in [0,1])$ with corresponding cumulative distribution function (c.d.f.) $x \mapsto e^{1-1/x}\mathbb{1}(x \in [0,1])$ provides an example of such a behaviour. To see this, observe that $P([-2\eta, 2\eta]) = e^{1-1/(2\eta)} = e^{1/(2\eta)}e^{1-1/\eta} = e^{1/(2\eta)}P([-\eta,\eta])$. Since $e^{1/(2\eta)} \longrightarrow \infty$ as $\eta \longrightarrow 0$, (LDP) cannot hold.

We now derive explicit expressions for the local convergence rates and verify the (LDP) for different design distributions by proving that $\mathrm{P_X} \in \mathcal{P}_G(D)$ or $\mathrm{P_X} \in \mathcal{P}_n(D)$.

Example 1. Assume that the design density p is bounded from below and above, in the sense that

$$0 < \underline{p} := \inf_{x\in[0,1]} \leq \overline{p} := \sup_{x\in[0,1]} p(x) < \infty. \tag{8}$$

The following result shows that in this case Theorem 4 is applicable and the local convergence rate is $t_n(x) \asymp (\log n/n)^{1/3}$.

Lemma 10. *Assume that the design distribution* $\mathrm{P_X}$ *admits a Lebesgue density satisfying* (8)*. Then,* $\mathrm{P_X} \in \mathcal{P}_G(4\overline{p}/\underline{p})$ *and for any* $0 \leq x \leq 1$*,*

$$\left(\frac{\log n}{2n\overline{p}}\right)^{1/3} \leq t_n(x) \leq \left(\frac{\log n}{n\underline{p}}\right)^{1/3}. \tag{9}$$

As a second example, we consider densities that vanish at $x = 0$.

Example 2 (Density vanishing with polynomial speed at zero). Assume that, for some $\alpha > 0$, the design distribution $\mathrm{P_X}$ has Lebesgue density

$$p(x) = (\alpha + 1)x^{\alpha+1}\mathbb{1}(x \in [0,1]).$$

This means that there is a low-density regime near zero with rather few observed design points. In this regime, it is more difficult to estimate the regression function, which is reflected in a slower local convergence rate.

Lemma 11. *If* $n > 9, \alpha > 0$ *and* $a_n := (\log n/n2^{\alpha+1})^{1/(\alpha+3)}$*, then, the distribution with density* $p\colon x \mapsto (\alpha+1)x^\alpha\mathbb{1}(x \in [0,1])$ *is in* $\mathcal{P}_G(D)$ *for some D depending only on* α*. Thus, Theorem 4 is applicable and*

$$\left(\frac{\log n}{2^{\alpha+1}n}\right)^{1/(\alpha+3)} \leq t_n(x) \leq \left(\frac{\log n}{n}\right)^{1/(\alpha+3)}, \quad \text{for } 0 \leq x \leq a_n, \tag{10}$$

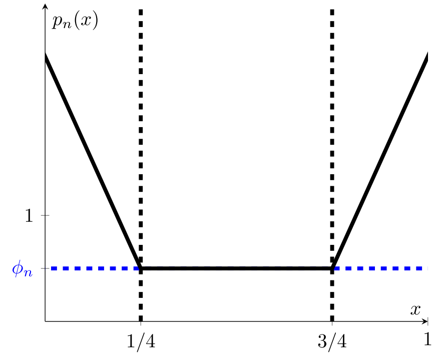

Figure 1. The density p_n.

and

$$\left(\frac{\log n}{2^{\alpha+1}(\alpha+1)nx^{\alpha}}\right)^{1/3} \leq t_n(x) \leq \left(\frac{\log n}{nx^{\alpha}}\right)^{1/3}, \quad \textit{for } a_n \leq x \leq 1. \tag{11}$$

By rewriting the expression for the spread function, we find that the local convergence rate is $t_n(x) \asymp (\log n/(n(x \vee a_n)^{\alpha}))^{1/3}$. The behavior of $t_n(0)$ can also directly be deduced from Lemma 6.

As a last example, we consider a sequence of design distributions with decreasing densities on $[1/4, 3/4]$.

Example 3 (Sequence of distributions with low-density region). For $\phi_n = 1 \wedge n^{-1/4}\log n$, consider the sequence of distributions $\mathrm{P}_{\mathrm{X}}^{n}$ with associated Lebesgue densities

$$p_n(x) := \phi_n + 16(1-\phi_n)\max\left(\frac{1}{4} - x, 0, x - \frac{3}{4}\right). \tag{12}$$

It is easy to check that this defines Lebesgue densities on $[0,1]$. See Figure 1 for a graph of p_n. According to Lemma 10, these distributions are globally doubling. Since $\mathrm{P}_{\mathrm{X}}^{n}([0,1])/\mathrm{P}_{\mathrm{X}}^{n}([1/4,3/4]) = 1/\phi_n$, the doubling constants are $\geq 1/\phi_n$ and hence tend to infinity as n grows. Therefore there is no $D > 0$ such that $p_n \in \mathcal{P}_G(D)$ for all n. On the contrary, for all n, $p_n \in \mathrm{Lip}(16)$ and since $\phi_n \geq n^{-1/4}\log n$, the assumptions of Corollary 5 are satisfied with $\beta = 1$ and $\kappa = 16$. Therefore, $p_n \in \mathcal{P}_n(8)$ for all n large enough and the local convergence rate is $(\log n/(np_n(x)))^{1/3}$. In particular, in the regime $[1/4, 3/4]$, the local convergence rate becomes $n^{-1/4}$.

4. Proof strategy

As the new proof strategy to establish local rates for least squares estimation is the main mathematical contribution of this work, we outline it here. Consider the LSE

$$\widehat{f}_n \in \operatorname*{arg\,min}_{f \in \mathrm{Lip}(1)} \sum_{i=1}^{n} (Y_i - f(X_i))^2 .$$

The definition of the estimator ensures that for any $g \in \mathrm{Lip}(1)$, the so called *basic inequality* $\sum_{i=1}^n (Y_i - \widehat{f}_n(X_i))^2 \le \sum_{i=1}^n (Y_i - g(X_i))^2$ holds. Assume that the function g satisfies

$$\big(\widehat{f}_n(X_i) - g(X_i)\big)\big(g(X_i) - f_0(X_i)\big) \geq 0, \quad \text{for all } i = 1, \dots, n, \tag{13}$$

which is the same as saying that at all data points g should lie between $\widehat{f}_n$ and the true regression function f_0. Together with the basic inequality and using $Y_i = f_0(X_i) + \varepsilon_i$, we obtain

$$\begin{aligned} \sum_{i=1}^n \big(\widehat{f}_n(X_i) - g(X_i)\big)^2 &\leq \sum_{i=1}^n \big(\widehat{f}_n(X_i) - f_0(X_i)\big)^2 - \sum_{i=1}^n \big(g(X_i) - f_0(X_i)\big)^2 \\ &\leq 2 \sum_{i=1}^n \varepsilon_i \big(\widehat{f}_n(X_i) - g(X_i)\big). \end{aligned} \tag{14}$$

We prove that $t_n(x)$ is a local convergence rate by contradiction. Assume that the LSE $\widehat{f}_n$ is more than $Kt_n(x^*)$ away from the true regression function f_0 for some $x^* \in [0,1]$ and a sufficiently large constant K. Then, we choose g as a specific local perturbation of $\widehat{f}_n$ (in the sense that g differs from $\widehat{f}_n$ only on a small interval) such that the previous inequality (14) is violated, resulting in the desired contradiction.

Denote the space of all possible functions $\widehat{f}_n - g$ by $\mathcal{F}^*$. Since $\widehat{f}_n \in \mathrm{Lip}(1)$ and $g \in \mathrm{Lip}(1)$, we have $\widehat{f}_n - g \in \mathrm{Lip}(2)$ and thus, $\mathcal{F}^* \subseteq \mathrm{Lip}(2)$. In fact, by choosing g as a local perturbation of $\widehat{f}_n$, the function class $\mathcal{F}^*$ will be much smaller than Lip(2). Due to the small support of $\widehat{f} - g$, we have $\widehat{f}(X_i) - g(X_i) = 0$ for most X_i. It is conceivable that one can remove these indices from (14) and that the effective sample size $m = m(X_1, Y_1, \dots, X_n, Y_n)$ is the number of indices for which $\widehat{f}_n(X_i) - g(X_i) \neq 0$. Denote by $N(r, \mathcal{F}^*, \|\cdot\|_\infty)$ the covering number of $\mathcal{F}^*$ with sup-norm balls of radius r and assume moreover that $\mathcal{F}^*$ is star-shaped, that is, if $h \in \mathcal{F}^*$ and $\alpha \in [0,1]$, then also $\alpha h \in \mathcal{F}^*$. We now argue similarly as in [48]. Replacing f^* by g in their inequality (13.18, p. 452) and then following exactly the same steps as in the proofs for their Theorem 13.1 and Corollary 13.1, one can now show that if there exists a sequence η_n with $0 \le \eta_n \le 1$ satisfying

$$\frac{16}{\sqrt{m}} \int_{\eta_n^2/4}^{\eta_n} \sqrt{\log N\big(r, \mathcal{F}^*, \|\cdot\|_\infty\big)}\, dr \leq \frac{\eta_n^2}{4}, \tag{15}$$

then,

$$\mathrm{P}\Big(\frac{1}{m} \sum_{i=1}^n \big(\widehat{f}_n(X_i) - g(X_i)\big)^2 \geq 16\eta_n^2\Big) \leq e^{-m\eta_n^2/2}. \tag{16}$$

To derive a contradiction assume that there exists a point x^* such that $|\widehat{f}_n(x^*) - f_0(x^*)| > Kt_n(x^*)$. Suppose moreover, that for all K large enough, we can find a function $g \in \mathrm{Lip}(1)$ satisfying (13) and $|\widehat{f}_n(x^*) - g(x^*)| > Kt_n(x^*)$, and support of $\widehat{f}_n - g$ with length of the order $Kt_n(x^*)$. The assumed properties of such a function g are plausible due to $\widehat{f}_n - g \in \mathrm{Lip}(2)$. Because we can also choose $K \geq 4$, another consequence of $\widehat{f}_n - g \in \mathrm{Lip}(2)$ is that $|\widehat{f}_n(x) - g(x)| > Kt_n(x^*)/2$ for all $x \in [x^* \pm t_n(x^*)]$. Thus,

$$\sum_{i=1}^n \big(\widehat{f}_n(X_i) - g(X_i)\big)^2 \geq \frac{K^2}{4} t_n(x^*)^2 \sum_{i=1}^n \mathbb{1}\big(X_i \in [x^* \pm t_n(x^*)]\big).$$

The right-hand side should be close to its expectation $\frac{1}{4}K^2 t_n(x^*)^2 nP_X([x^* \pm t_n(x^*)]) = \frac{1}{4}K^2 \log n$, where we used the definition of $t_n(x^*)$. Thus, up to approximation errors, we obtain the lower bound

$$\sum_{i=1}^{n} \big(\widehat{f}_n(X_i) - g(X_i)\big)^2 \geq \frac{K^2}{4} \log n. \tag{17}$$

We now explain the choice of η_n in (16) that leads to the upper bound for $\sum_{i=1}^{n}(\widehat{f}_n(X_i) - g(X_i))^2$ in (14). Since the perturbation is supported on an interval with length of the order $Kt_n(x^*)$, one can bound the metric entropy $\log N\big(r, \mathcal{F}^*, \|\cdot\|_\infty\big) \lesssim Kt_n(x^*)/r$, with proportionality constant independent of K. Therefore, (15) holds for $\eta_n \propto (Kt_n(x^*)\log n/m)^{1/3}$. The additional $\log n$-factor in η_n is necessary to obtain uniform statements in x. For this choice of η_n, the probability in (16) converges to zero. Consequently, on an event with large probability, we have that

$$\sum_{i=1}^{n} \big(\widehat{f}_n(X_i) - g(X_i)\big)^2 \leq 16m\eta_n^2 \lesssim \big(mK^2 t_n(x^*)^2 \log^2 n\big)^{1/3}. \tag{18}$$

Recall that the support of $\widehat{f}_n - g$ is contained in $[x^* \pm CKt_n(x^*)]$ for some constant C. Moreover, m is the number of observations in the support of $\widehat{f}_n - g$. Now m should be close to its expectation which can be upper bounded by $n\mathrm{P}_{\mathrm{X}}([x^* \pm CKt_n(x^*)])$. Invoking the local doubling property (LDP), m can also be upper bounded by $C_K n\mathrm{P}_{\mathrm{X}}([x^* \pm t_n(x^*)])$, for a constant C_K depending on K. Using the definition of $t_n(x)$, (18) can be further bounded by

$$\sum_{i=1}^{n} \big(\widehat{f}_n(X_i) - g(X_i)\big)^2 \lesssim \Big(C_K K^2 nt_n(x^*)^2\, \mathrm{P}_{\mathrm{X}}([x^* \pm t_n(x^*)]) \log^2 n\Big)^{1/3} \leq (C_K K^2)^{1/3} \log n.$$

Comparing this with the lower bound (17) and dividing both sides by $\log n$, we conclude that on an event with large probability, $\frac{1}{4}K^2 \lesssim (C_K K^2)^{1/3}$, where the proportionality constant does not depend on K. A technical argument that links the upper and lower bound more tightly and that we omit here shows that one can even avoid the dependence of C_K on K, such that we finally obtain $K^2 \lesssim K^{2/3}$. Taking K large and since the proportionality constant is independent of K, we finally obtain a contradiction. This means that on an event with large probability and for all sufficiently large K, there cannot be a point x^*, such that $|\widehat{f}_n(x^*) - f_0(x^*)|/t_n(x^*) > K$, proving $\sup_x |\widehat{f}_n(x) - f_0(x)|/t_n(x) \leq K$ with large probability.

There is still a major technical obstacle in the proof strategy, namely the choice of the local perturbation g. This construction appears to be one of the main difficulties of the proof. In fact, the empirical risk minimizer over 1-Lipschitz functions will typically lie somehow on the boundary of the space Lip(1) in the sense that on small neighbourhoods, the Lipschitz constant of the estimator is exactly one.

To see this, assume the statement would be false. Then we could build tiny perturbations around the estimator that are 1-Lipschitz and lead to a smaller least squares loss, which contradicts the fact that the original estimator is a least squares minimizer (see Figure 2). This makes it tricky to construct a local perturbation g of $\widehat{f}$ that also lies in Lip(1) and satisfies the required conditions. To find a suitable perturbation, our approach is to introduce first x^* as above and then define another point $\tilde{x}$ in the neighbourhood of x^* with some specific properties. The full construction is explained in Figure 3 and Lemma 22 in Section 7.

5. Applications to transfer learning

Transfer Learning (TL) aims to exploit that an estimator achieving good performance on a certain task should also work well on similar tasks. This allows to emulate a bigger dataset and to save computational time by relying on previously trained models. In the supervised learning framework, we have access to training data generated from a distribution $\mathrm{Q}_{\mathrm{X},\mathrm{Y}}$. Observing X from a pair $(X,Y) \sim \mathrm{Q}_{\mathrm{X},\mathrm{Y}}$ we want to predict the corresponding value of Y. To do so, we compute an estimate based on observing m i.i.d. copies sampled from $\mathrm{Q}_{\mathrm{X},\mathrm{Y}}$. Assume now that we also have access to $n > m$ i.i.d. copies sampled from another distribution $\mathrm{P}_{\mathrm{X},\mathrm{Y}}$. The transfer learning paradigm states that, depending on some similarity criterion between P_{X} and Q_{X}, fitting an estimator using both samples improves the predictive power. In other words, $\mathrm{P}_{\mathrm{X},\mathrm{Y}}$ contains information about $\mathrm{Q}_{\mathrm{X},\mathrm{Y}}$ that can be transferred to improve the fit. Two standard settings within TL are posterior drift and covariate shift. For posterior drift, one assumes that the marginal distributions are the same, that is, $\mathrm{P}_{\mathrm{X}} = \mathrm{Q}_{\mathrm{X}}$, but $\mathrm{P}_{\mathrm{Y}|\mathrm{X}}$ and $\mathrm{Q}_{\mathrm{Y}|\mathrm{X}}$ may be different. On the contrary, TL with covariate shift assumes that $\mathrm{P}_{\mathrm{Y}|\mathrm{X}} = \mathrm{Q}_{\mathrm{Y}|\mathrm{X}}$, while P_{X} and Q_{X} can differ. Here, we address the covariate shift paradigm within the nonparametric regression framework. This means, we observe $n + m$ independent pairs $(X_1,Y_1),\ldots,(X_{n+m},Y_{n+m}) \in [0,1]\times\mathbb{R}$ with

$$\begin{aligned} X_i &\sim \mathrm{P}_{\mathrm{X}}, && \text{for } i = 1,\ldots,n, \\ X_i &\sim \mathrm{Q}_{\mathrm{X}}, && \text{for } i = n+1,\ldots,n+m, \\ Y_i &= f_0(X_i) + \varepsilon_i, && \text{for } i = 1,\ldots,n+m, \end{aligned} \tag{19}$$

and independent noise variables $\varepsilon_1,\ldots,\varepsilon_{n+m} \sim \mathcal{N}(0,1)$.

We now discuss estimation in this model, treating the cases $m = 0$ and $m > 0$, separately. In both cases, the risk is the prediction error under the target distribution. For readability, we omit the subscript X and write P and Q for P_{X} and Q_{X}, respectively. Throughout the section, we assume global doubling, that is, $\mathrm{P},\mathrm{Q} \in \mathcal{P}_G(D)$ for some $D \geq 2$.

5.1. Using LSE from source distribution to predict under target distribution

As before, let q denote the density of the target design distribution Q. Recall that we are considering the covariate shift model (19) with $m = 0$ and Lipschitz continuous regression functions. If the n training

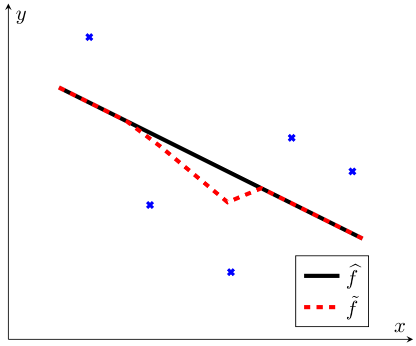

Figure 2. If the LSE $\widehat{f}$ would not have locally slope = 1, then one could construct a perturbed version $\tilde{f}$ that better fits the data, implying that $\widehat{f}$ cannot be a least squares fit.

data were generated from the target distribution $\mathrm{Q}_{\mathrm{X,Y}}$ instead, the classical empirical risk theory would lead to the standard nonparametric rate $n^{-2\beta/(2\beta+1)}$ with $\beta = 1$. More precisely, the statement would be that with probability tending to one as $n \to \infty$, $\int_0^1 (\widehat{f}_n(x) - f_0(x))^2 q(x)\,dx \lesssim n^{-2/3}$.

For n training samples from the source distribution $\mathrm{P}_{\mathrm{X,Y}}$, Theorem 4 shows that $|\widehat{f}_n(x) - f_0(x)| \leq K t_n^{\mathrm{P}}(x)$ for all x, with high probability, where t_n^{P} denotes the spread function associated to the distribution P. This means that the prediction risk under the target marginal distribution Q_{X} with density q is bounded by

$$\int_0^1 \big(\widehat{f}_n(x) - f_0(x)\big)^2 q(x)\,dx \leq K^2 \int_0^1 t_n^{\mathrm{P}}(x)^2 q(x)\,dx, \tag{20}$$

with probability converging to one as $n \longrightarrow \infty$. For a given source density p, the main question is whether the right-hand side is of the order $n^{-2/3}$, up to $\log n$-factors. This would imply that there is no loss in terms of convergence rate (ignoring $\log n$-factors) due to the different sampling scheme. To get at least close to the $n^{-2/3}$-rate, some conditions on p are needed. If p is, for instance, zero on $[0, 1/2]$, we have no information about the regression function f on this interval and any estimator will be inconsistent on $[0, 1/2]$. If we then try to predict with Q the uniform distribution, it is clear that $\int_0^1 (\widehat{f}_n(x) - f_0(x))^2\,dx \geq \int_0^{1/2} (\widehat{f}_n(x) - f_0(x))^2\,dx \gtrsim 1$.

In the setting of sample size dependent densities p_n, Corollary 5 shows that under the imposed conditions, there exists a constant K' that does not depend on n, such that

$$\int_0^1 \big(\widehat{f}_n(x) - f_0(x)\big)^2 q(x)\,dx \leq (K')^2 \Big(\frac{\log n}{n}\Big)^{2/3} \int_0^1 \frac{q(x)}{p_n(x)^{2/3}}\,dx,$$

with probability tending to one as $n \longrightarrow \infty$. For instance, for the sequence of densities $p_n(x) := \phi_n(x) + 16(1-\phi_n)\max(1/4 - x, 0, x - 3/4)$ with $\phi_n = 1 \wedge n^{-1/4}\log(n)$, as considered in (12), the right-hand side in the previous display is of the order $(\log n/n)^{2/3}\phi_n^{-2/3} \leq n^{-1/2}$.

For distributions satisfying the conditions of Theorem 4, we need to bound the more abstract integral $\int_0^1 t_n^{\mathrm{P}}(x)^2 q(x)\,dx$. The next result provides a different, sometimes simpler formulation.

Lemma 12. *In the same setting and for the same conditions as in Theorem 4, there exists a constant K'', such that*

$$\begin{aligned}\int_0^1 \big(\widehat{f}_n(x) - f_0(x)\big)^2 q(x)\,dx &\leq K'' \frac{\log n}{n} \int_0^1 \frac{\mathrm{Q}([x \pm t_n^{\mathrm{P}}(x)])}{t_n^{\mathrm{P}}(x)\mathrm{P}([x \pm t_n^{\mathrm{P}}(x)])}\,dx \\ &\leq 2^{1/3} K'' \Big(\frac{\log n}{n}\Big)^{2/3} \|p\|_\infty^{1/3} \int_0^1 \frac{\mathrm{Q}([x \pm t_n^{\mathrm{P}}(x)])}{\mathrm{P}([x \pm t_n^{\mathrm{P}}(x)])}\,dx,\end{aligned}$$

with probability tending to one as $n \longrightarrow \infty$.

In [25], a pair (P, Q) is said to have transfer exponent γ, if there exists a constant $0 < C \leq 1$, such that for all $0 \leq x \leq 1$ and $0 < \eta \leq 1$, we have $\mathrm{P}([x \pm \eta]) \geq C\eta^\gamma\,\mathrm{Q}([x \pm \eta])$. Combined with the previous lemma, we get for transfer exponent γ,

$$\int_0^1 \big(\widehat{f}_n(x) - f_0(x)\big)^2 q(x)\,dx \leq 2^{1/3} \frac{K''}{C} \Big(\frac{\log n}{n}\Big)^{2/3} \|p\|_\infty^{1/3} \int_0^1 t_n^{\mathrm{P}}(x)^{-\gamma}\,dx,$$

with probability tending to one as $n \to \infty$. Interestingly, the right-hand side does not depend on the target distribution Q.

The next lemma provides an example of convergence rates.

Lemma 13. *Work in the nonparametric covariate shift model* (19) *with* $m = 0$. *Let* $\alpha > 0$. *For source design density* $p(x) = (\alpha + 1)x^{\alpha}\mathbb{1}(x \in [0,1])$ *and uniform target design density* $q(x) = \mathbb{1}(x \in [0,1])$, *we have that*

$$\int_0^1 \Big(\widehat{f}_n(x) - f_0(x)\Big)^2 q(x)\,dx \lesssim (\log n)^{\mathbb{1}(\alpha=3/2)}\bigg[\Big(\frac{\log n}{n}\Big)^{3/(3+\alpha)} \vee \Big(\frac{\log n}{n}\Big)^{2/3}\bigg] \tag{21}$$

with probability tending to one as $n \to \infty$.

The proof shows that the result follows for $0 < \alpha \leq 1$ by a direct application of Lemma 12. For general $\alpha > 0$, we prove the lemma by a more sophisticated analysis based on the bounds derived in Example 2.

The convergence rate is $(\log n/n)^{3/(\alpha+3)}$ if $\alpha > 3/2$ and $(\log n/n)^{2/3}$ if $\alpha < 3/2$. For $\alpha = 3/2$ an additional $\log n$-factor appears. This result shows that the low-density region near zero causes a slower convergence for $\alpha > 3/2$.

5.2. Combining both samples to predict under the target distribution

We now consider the nonparametric regression model under covariate shift (19) with a second sample, that is, $m > 0$.

In the first step, we construct an estimator combining the information from both samples. The main idea is to consider the LSEs for the first and second part of the sample and, for a given x, pick the LSE with the smallest estimated local rate. For a proper definition of the estimator, some notation is required. Restricting to the first and second part of the sample, let $\widehat{f}_n^{(1)}$ and $\widehat{f}_m^{(2)}$ denote the corresponding LSEs taken over 1-Lipschitz functions. Because the spread function is the local convergence rate, it is now natural to study $\widetilde{f}_{n,m}(x) = \widehat{f}_n^{(1)}(x)\mathbb{1}(t_n^{\mathrm{P}}(x) \leq t_m^{\mathrm{Q}}(x)) + \widehat{f}_m^{(2)}(x)\mathbb{1}(t_n^{\mathrm{P}}(x) > t_m^{\mathrm{Q}}(x))$. Because the spread functions are unknown, $\widetilde{f}_{n,m}(x)$ is not yet an estimator. Replacing $t_n^{\mathrm{P}}(x)$ and $t_m^{\mathrm{Q}}(x)$ by the estimators

$$\begin{aligned}
\widehat{t}_n^{\mathrm{P}}(x) &:= \inf\Big\{t : t^2\widehat{\mathrm{P}}^n([x \pm t]) \geq \frac{\log n}{n}\Big\}, \quad \widehat{\mathrm{P}}^n([x \pm t]) := \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_i \in [x \pm t]),\\
\widehat{t}_m^{\mathrm{Q}}(x) &:= \inf\Big\{t : t^2\widehat{\mathrm{Q}}^m([x \pm t]) \geq \frac{\log m}{m}\Big\}, \quad \widehat{\mathrm{Q}}^m([x \pm t]) := \frac{1}{m}\sum_{i=n+1}^{n+m}\mathbb{1}(X_i \in [x \pm t]),
\end{aligned} \tag{22}$$

leads to the definition of our nonparametric regression estimator under covariate shift,

$$\widehat{f}_{n,m}(x) := \widehat{f}_n^{(1)}(x)\mathbb{1}\big(\widehat{t}_n^{\mathrm{P}}(x) \leq \widehat{t}_m^{\mathrm{Q}}(x)\big) + \widehat{f}_m^{(2)}(x)\mathbb{1}(\widehat{t}_n^{\mathrm{P}}(x) > \widehat{t}_m^{\mathrm{Q}}(x)). \tag{23}$$

Let p and q be the respective Lebesgue densities of P and Q. We omit the dependence on n,m and define by P_f the distribution of the data in model (19).

Theorem 14. *Consider the nonparametric regression model under covariate shift* (19). *Let* $0 < \delta < 1$ *and* $D > 0$. *If* $\mathrm{P},\mathrm{Q} \in \mathcal{P}_G(D)$ *and the estimator* $\widehat{f}_{n,m}$ *is as in* (23), *then, for a sufficiently large constant*

K,

$$\sup_{f_0 \in \mathrm{Lip}(1-\delta)} \mathrm{P}_{f_0}\left(\sup_{x\in[0,1]} \frac{|\widehat{f}_{n,m}(x) - f_0(x)|}{t_n^{\mathrm{P}}(x) \wedge t_m^{\mathrm{Q}}(x)} > K\right) \to 0 \quad \textit{as } n \to \infty \textit{ and } m \to \infty.$$

The proof shows that to achieve the rate $t_n^{\mathrm{P}}(x) \wedge t_m^{\mathrm{Q}}(x)$, it is actually enough to estimate $t_n^{\mathrm{P}}(x)$ using N data points $X_1, \ldots, X_N \sim \mathrm{P}$, where N is a sufficiently large number. Thus, instead of observing the full first dataset $(X_1,Y_1), \ldots, (X_n,Y_n) \sim \mathrm{P}$, the estimator only needs the LSE $\widehat{f}_n^{(1)}$ and N i.i.d. observations from the design distribution P.

In the next step, we show that the rate $t_n^{\mathrm{P}}(x) \wedge t_m^{\mathrm{Q}}(x)$ is the local minimax rate. The design distributions $\mathrm{P}_{\mathrm{X}}^n, \mathrm{Q}_{\mathrm{X}}^n$ are allowed to depend on the sample size. The corresponding spread functions are denoted by $t_n^{\mathrm{P}}(x)$ and $t_m^{\mathrm{Q}}(x)$.

Theorem 15. *Consider the nonparametric regression model under covariate shift* (19). *If C_∞ is a positive constant, then there exists a positive constant c, such that for any sufficiently large n, and any sequences of design distribution $\mathrm{P}_{\mathrm{X}}^n, \mathrm{Q}_{\mathrm{X}}^n \in \mathcal{M}$ with corresponding Lebesgue densities p_n, q_n all upper bounded by C_∞, we have*

$$\inf_{\widehat{f}_{n,m}} \sup_{f_0 \in \mathrm{Lip}(1)} \mathrm{P}_{f_0}\left(\sup_{x\in[0,1]} \frac{|\widehat{f}_{n,m}(x) - f_0(x)|}{t_n^{\mathrm{P}}(x) \wedge t_m^{\mathrm{Q}}(x)} \geq \frac{1}{12}\right) \geq c,$$

where the infimum is taken over all estimators and P_{f_0} is the distribution of the data in model (19).

Given the full dataset in model (19), an alternative procedure is to use the LSE over all data, that is,

$$\widehat{f}_{n+m} \in \operatorname*{arg\,min}_{f \in \mathrm{Lip}(1)} \sum_{i=1}^{n+m} \big(Y_i - f(X_i)\big)^2.$$

Instead of analyzing this estimator in model (19), the risk can rather easily be controlled in the related model, where we observe $n+m$ i.i.d. observations $(X_1,Y_1), \ldots, (X_{n+m},Y_{n+m})$ with X_i drawn from the mixture distribution $\widetilde{\mathrm{P}} := \frac{m}{m+n}\mathrm{Q} + \frac{n}{m+n}\mathrm{P}$ and $Y_i = f_0(X_i) + \varepsilon_i$. In this model, we draw in average n observations from P and m observations from Q. Since $\mathcal{P}_G(D)$ is convex, Theorem 4 applies and, consequently, $t_{n+m}^{\widetilde{\mathrm{P}}}(x)$ is a local convergence rate. The spread function can be bounded as follows.

Lemma 16. *If* $\widetilde{\mathrm{P}} := \frac{n}{n+m}\mathrm{P} + \frac{m}{n+m}\mathrm{Q}$, *then,*

$$t_{n+m}^{\widetilde{\mathrm{P}}}(x) \leq t_n^{\mathrm{P}}(x)\sqrt{\frac{\log(m+n)}{\log n}} \wedge t_m^{\mathrm{Q}}(x)\sqrt{\frac{\log(m+n)}{\log m}}.$$

If there are positive constants C, κ such that $\sup_x t_n^{\mathrm{P}}(x) \leq Cn^{-\kappa}$ for all n, then, there exists a constant C', such that

$$t_{n+m}^{\widetilde{\mathrm{P}}}(x) \leq C'\big(t_n^{\mathrm{P}}(x) \wedge t_m^{\mathrm{Q}}(x)\big), \quad \textit{for all } n \geq m > 2.$$

One can see that the rate is at most a log-factor larger than the local minimax rate $t_n^{\mathrm{P}}(x) \wedge t_m^{\mathrm{Q}}(x)$. Moreover, this additional log-factor can be avoided in the relevant regime where the local rate $t_n^{\mathrm{P_X}}(x)$ decays with some polynomial rate uniformly over $[0,1]$.

We now return to our leading example with source density $p(x) = (\alpha + 1)x^{\alpha+1}\mathbb{1}(x \in [0,1])$ and target density $q(x) = \mathbb{1}(x \in [0,1])$. Lemma 10 and Lemma 11 show that if $\alpha > 0$, then, the assumptions of Theorem 14 are satisfied and the local convergence rate of the combined estimator $\widehat{f}_{n,m}$ is $t_n^{\mathrm{P}}(x) \wedge t_m^{\mathrm{Q}}(x)$. From (21), we see that as long as $\alpha < 3/2$, the first sample is enough to achieve the rate $(\log n/n)^{2/3}$. We therefore focus on the regime $3/2 < \alpha$.

Lemma 17. *Consider the nonparametric regression model under covariate shift* (19). *For* $3/2 < \alpha$, *let* P *be the distribution with Lebesgue density* $p(x) = (\alpha + 1)x^{\alpha+1}\mathbb{1}(x \in [0,1])$ *and* Q *be the uniform distribution on* $[0,1]$. *If* $n^{3/(3+\alpha)} \log^{\alpha/(3+\alpha)} n \ll m \le n$, *then, we have that for any* $f_0 \in \mathrm{Lip}(1-\delta)$,

$$\int_0^1 \big(\widehat{f}_{n,m}(x) - f_0(x)\big)^2 q(x)\,dx \lesssim \Big(\frac{\log n}{m}\Big)^{2/3}\Big(\frac{m}{n}\Big)^{1/\alpha}, \tag{24}$$

with probability converging to one as $n \to \infty$.

We have $m \lesssim n^{3/(\alpha+3)}(\log n)^{\alpha/(\alpha+3)}$ if and only if $(\log n/m)^{2/3}(m/n)^{1/\alpha} \lesssim (\log n/n)^{3/(3+\alpha)}$. Since the right-hand side is the rate obtained without a second sample in (21), the additional data $(X_{n+1}, Y_{n+1}), \dots, (X_{n+m}, Y_{n+m})$ improve the convergence rate if $m \gg n^{3/(3+\alpha)}(\log n)^{\alpha/(3+\alpha)}$.

6. Discussion

6.1. A brief review of convergence results for the least squares estimator in nonparametric regression

The standard strategy to derive convergence rates with respect to (empirical) L^2-type losses is based on empirical process theory and covering bounds. The field is well-developed, see e.g. [16,23,45,47,48]. At the same time, it remains a topic of active research. A recent advance is to establish convergence rates of the LSE under heavy-tailed noise distributions [19,26].

Some convergence results are with respect to the squared Hellinger distance, see for instance [5,46]. This is slightly weaker but essentially the same as convergence with respect to the prediction risk $\mathrm{E}[(\widehat{f}_n(X) - f_0(X))^2]$. To see this, recall that for two probability measures P, Q defined on the same measurable space, the squared Hellinger distance is defined as $H^2(P,Q) = \frac{1}{2}\int(\sqrt{dP} - \sqrt{dQ})^2$ (some authors do not use the factor 1/2). Denote by Q_f the distribution of (X_1, Y_1) in the nonparametric regression model (1) with regression function f. It can be shown that

$$H^2(\mathrm{Q}_f, \mathrm{Q}_g) = 1 - \int e^{-\frac{1}{8}(f(x)-g(x))^2} p(x)\,dx.$$

In view of the formula $1 - e^{-u} \le u$, it follows that $H^2(\mathrm{Q}_f, \mathrm{Q}_g) \le \frac{1}{8}\mathrm{E}[(f(X) - g(X))^2]$. Thus, the squared Hellinger loss is weaker than the squared prediction loss.

Concerning estimation rates, the LSE achieves the rate $n^{-2\beta/(2\beta+d)} \vee n^{-\beta/d}$ over balls of β-smooth Hölder functions. To see this, observe that if $\mathcal{F}$ denotes a Hölder ball and $\|g\|_n := (\frac{1}{n}\sum_{i=1}^n g(X_i)^2)^{1/2}$ is the empirical L^2 norm, the metric entropy is $\log N\big(r, \mathcal{F}, \|\cdot\|_n\big) \lesssim r^{-d/\beta}$, see Corollary 2.7.2 in [47]. Any solution δ^2 of the inequality $\int_{\delta^2}^{\delta} \sqrt{\log N\big(r, \mathcal{F}, \|\cdot\|_n\big)}\,dr \lesssim \delta^2\sqrt{n}$ is then a rate for the LSE, see Corollary 13.1 in [48]. It is now straightforward to check that this yields the convergence rate $\delta^2 \asymp n^{-2\beta/(2\beta+d)} \vee n^{-\beta/d}$. While $n^{-2\beta/(2\beta+d)}$ is the optimal convergence rate, Theorem 4 in [5] shows that for $d = 1$ and a specifically designed subset of the Hölder ball with index $\beta < 1/2$, the LSE cannot achieve a faster rate than $n^{-\beta/2}$ (up to a possibly non-optimal logarithmic factor in n). The (sub)optimality of the LSE over Hölder balls in the non-Donsker regime $\beta < d/2$ remains an open problem. Considering shape-constrained problems, [27] shows that for different classes of convex functions, the LSE is suboptimal for dimensions $d \geq 5$, while [17] proves that the LSE can still achieve near-optimal convergence rates in the non-Donsker regime.

To the best of our knowledge, the only sup-norm rate result for the LSE is [37]. In this work, the LSE is studied for $\mathcal{F}$ the linear space spanned by a nearly orthogonal function system. For this setting, the LSE has an explicit representation that can be exploited to prove sup-norm rates.

For a number of other settings, a more explicit characterization of the LSE is available from which local properties can be inferred. [30] shows this for least squares penalized regression with total variation penalties. In this case, the LSE can be linked to splines, and this is exploited in their Proposition 8 to provide a local characterization of the LSE.

More explicit characterizations of the LSE are also available for several shape-constrained estimation problems. In isotonic regression, the regression function is non-decreasing, and the LSE admits an explicit expression. Let $(X_{(1)},Y_{(1)}),\dots,(X_{(n)},Y_{(n)})$ be a reordered version of the dataset such that $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ and for all $0 \leq k \leq n$, define the k-th partial sum of Y as $S_k := \sum_{i=1}^{k} Y_{(i)}$. Additionally, for $x \in [0,1]$, set $k_-(x) := \max\{0 \leq k \leq n : X_{(k)} < x\}$ and $k_+(x) := \min\{0 \leq k \leq n : X_{(k)} \geq x\}$. The LSE for isotonic regression is then piecewise constant on $[0,1]$ and is given by

$$\widehat{f}_n(x) := \min_{k_+(x) \leq i \leq n} \max_{0 \leq j \leq k_-(x)} \frac{S_i - S_j}{i - j},$$

see for instance [6,7,20,49] and Lemma 2.1 in [41]. Moreover, the pointwise limiting distributional results are available for the isotonic LSE. For the purposes of this discussion, we provide the following, adapted version of the main theorem in [49].

Theorem 18. *Let $A > 0, \alpha > 1/2$ and $x_0 \in [0,1]$ with $p(x_0) > 0$. Suppose that there exists an open neighbourhood U of 0 and a continuous function $g \colon U \mapsto \mathbb{R}$ such that $\lim_{x\to 0} g(x)/x^\alpha = 0$ and for all $x \in U$, $|f_0(x) - f_0(x_0)| = A|x - x_0|^\alpha + g(x - x_0)$. Then*

$$\left(\frac{\alpha+1}{A}\right)^{1/(2\alpha+1)} \big(np(x_0)\big)^{\frac{\alpha}{2\alpha+1}} \big(\widehat{f}_n(x_0) - f_0(x_0)\big) \xrightarrow{d} Z,$$

with Z a random variable distributed as the slope at zero of the greatest convex minorant of $W_t + |t|^{\alpha+1}$ and $(W_t)_t$ a two-sided Brownian motion.

If f_0 is Lipschitz continuous, then $\alpha = 1$ and Z is known to follow Chernoff's distribution. Assuming moreover that $p(x_0) > 0$, leads to $|\widehat{f}_n(x_0) - f(x_0)| \asymp (np(x_0))^{-1/3}$. This agrees with the local rate $t_n(x_0) \asymp (\log n/(np(x_0)))^{1/3}$ obtained in Corollary 5 up to the $\log n$-factor that emerges due to the uniformity of the local rates.

For isotonic regression in d dimensions, the recent article [18] shows that the LSE achieves the minimax estimation rate $n^{-\min(2/(d+2),1/d)}$ up to $\log n$-factors. For $d \geq 3$, is it known that $\log N(r, \mathcal{F}, \|\cdot\|_2) \asymp r^{-2(d-1)}$. Since for uniform design, the norms $\|\cdot\|_2$ and $\|\cdot\|_n$ are close, the standard approach to derive convergence rates via the entropy integral is then expected to yield no convergence rate faster than $n^{-1/(2d-2)}$. Since this rate is slower than the actual convergence rate of the LSE, this provides another instance where the entropy integral approach is suboptimal. Interestingly, [18] proves, moreover, that if the isotonic function is piecewise constant with k pieces, the LSE adapts to the number of pieces and attains the optimal adaptive rate $(k/n)^{\min(1,2/d)}$ up to $\log n$-factors. More on adaptation and the pointwise behaviour of the LSE in isotonic regression and other shape-constrained estimation problems can be found in the survey article [15].

For penalized LSEs, an alternative to the concentration bounds in empirical process methods is the recently developed proof strategy based on the small ball method, [11,24,28,29,31].

6.2. Related work on transfer learning

From a theory perspective, the key problem in TL is to quantify the information that can be carried over from one task to another [2,4,10,32]. Among the mathematical statistics articles, [43] proposes unbiased model selection procedures and [40] considers re-weighting to improve the predictive power of models based on likelihood maximization. The nonparametric TL literature mainly focuses on classification. Minimax rates are derived under posterior drift by [9], under covariate shift by [25], and in a general setting by [36].

The closest related work is the recent preprint [33]. While we consider the LSE, this article proves minimax convergence rates for the Nadaraya-Watson estimator in nonparametric regression under covariate shift. The proofs differ, as one can use the closed-form formula for the Nadaraya-Watson estimator (NW). The rates are proven uniformly over two different sets of distribution pairs. Let $\rho_\eta(\mathrm{P_X},\mathrm{Q_X}) := \int_0^1 \mathrm{P_X}([x \pm \eta])^{-1}\, d\,\mathrm{Q_X}(x)$, $\gamma, C \geq 1$ and denote by $\mathcal{S}(\gamma,C)$ the set of all pairs $(\mathrm{P_X},\mathrm{Q_X})$, such that $\sup_{\eta\in(0,1]} \eta^\gamma \rho_\eta(\mathrm{P_X},\mathrm{Q_X}) \leq C$.

To discuss the connection of this class to our approach, observe that, in our framework, bounding the prediction risk of the LSE with regards to some target distribution amounts to bounding the quantity $\int_0^1 t_n^{\mathrm{P_X}}(x)^2\, d\,\mathrm{Q_X}(x)$. Using the definition of the spread function and assuming $(\mathrm{P_X},\mathrm{Q_X}) \in S(\gamma,C)$, we obtain

$$\int_0^1 t_n^{\mathrm{P_X}}(x)^2\, d\,\mathrm{Q_X}(x) = \frac{\log n}{n}\int_0^1 \frac{d\,\mathrm{Q_X}(x)}{\mathrm{P_X}([x \pm t_n^{\mathrm{P_X}}(x)])} \leq \frac{\log n}{n}\Big(\inf_{x\in[0,1]} t_n^{\mathrm{P_X}}(x)\Big)^{-\gamma}.$$

In some cases, faster rates for the prediction error can be obtained for the LSE using our results. As an example, consider again the case that the source density is $p(x) = (\alpha+1)x^\alpha \mathbb{1}(x \in [0,1])$ and the target distribution is uniform on $[0,1]$. For the nonparametric regression model with covariate shift (19), Lemma 13 shows that for the LSE $\widehat{f}_n$,

$$\int_0^1 \Big(\widehat{f}_n(x) - f_0(x)\Big)^2 q(x)\,dx \lesssim (\log n)^{\mathbb{1}(\alpha=3/2)}\Big[\Big(\frac{\log n}{n}\Big)^{3/(3+\alpha)} \vee \Big(\frac{\log n}{n}\Big)^{2/3}\Big] \tag{25}$$

with probability tending to one as $n \to \infty$. For the Nadaraya-Watson estimator, Lemma 31 in Appendix E of the supplement [39] shows that if $\alpha \geq 1$, there exists a $C > 0$, such that for any $\varepsilon \in (0,\alpha)$, $(\mathrm{P_X},\mathrm{Q_X}) \in \mathcal{S}(\alpha,C) \setminus \mathcal{S}(\alpha-\varepsilon,C)$. According to Corollary 1 in [33], for $\widehat{f}_{NW}$ the Nadaraya-Watson estimator with suitable bandwidth choice, we then have for any $\alpha \geq 1$,

$$E\Big[\int_0^1 \big(\widehat{f}_{NW}(x) - f(x)\big)^2 q(x)\,dx\Big] \lesssim n^{-2/(2+\alpha)}.$$

This is a slower rate than (25). The convergence rate of the LSE becomes slower than $(\log n/n)^{2/3}$ for $\alpha > 3/2$, while for the Nadaraya-Watson estimator, this happens already for $\alpha > 1$. We believe that the loss in the rate is due to the lack of local adaptivity of kernel smoothing with fixed bandwidth, as discussed in Section 2.

6.3. Extensions and open problems

For machine learning applications, we are, of course, interested in multivariate nonparametric regression with d-dimensional design vectors X_i and arbitrary Hölder smoothness β. To extend the result, the definition of the spread function has to be adjusted. If $\beta > d/2$, the LSE converges with the rate $n^{-2\beta/(2\beta+d)}$ (see the discussion above) and we believe that the local rate t_n is now determined by the solution of the equation

$$t_n(x)^2 \, \mathrm{P_X}\left(y : |x-y|_\infty \le t_n(x)^{1/\beta}\right) = \frac{\log n}{n}, \tag{26}$$

where $|v|_\infty$ denotes the largest absolute value of the components of the vector v. In the case $d = 1$, this coincides with the minimax rate found in [13]. Observe moreover that for the uniform design distribution, $\mathrm{P_X}(y : |x-y|_\infty \le t_n(x)^{1/\beta}) \asymp t_n(x)^{d/\beta}$ and we obtain $t_n(x) \asymp (\log n/n)^{\beta/(2\beta+d)}$. To show that t_n is a lower bound on the local convergence rate, we believe that the lower bound in Theorem 7 can be generalized without too much additional effort. But the upper bound is considerably harder than the case $\beta = d = 1$ we considered in this work. The main reason is that the local perturbation in the proof also needs to be $\beta > 1$ smooth, thus a piecewise approach as in (31) does not work anymore, and one needs to have tight control of the derivatives of the LSE.

It is unclear what the local convergence rate is in the regime $\beta < d/2$.

Throughout this work, we assumed data from the nonparametric regression model $Y_i = f_0(X_i) + \varepsilon_i$ with independent noise variables $\varepsilon_i \sim \mathcal{N}(0,1)$. If instead, we have $\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$, the spread function should be determined by

$$t_n(x)^2 \, \mathrm{P_X}\left(\left[x \pm t_n(x)\right]\right) = \frac{\sigma^2 \log n}{n}.$$

That this is the right scaling has already been observed in the article [13].

In Theorem 4, we assume that the regression function is $(1-\delta)$-Lipschitz for some positive δ. Another interesting question is whether the local convergence result can be extended to a regression function that is itself 1-Lipschitz. Again, the main complication arises in constructing the local perturbation in Lemma 22. One might also wonder whether the same rates can be achieved if, instead of the global minimizer, we take any estimator $\widehat{f}$ satisfying

$$\sum_{i=1}^n \left(Y_i - \widehat{f}(X_i)\right)^2 \le \inf_{f \in \mathrm{Lip}(1-\delta)} \sum_{i=1}^n \left(Y_i - f(X_i)\right)^2 + \tau_n$$

for a pre-defined rate τ_n. In particular, it is of interest to determine the largest τ_n such that the optimal local rates can still be obtained.

The proposed approach might be used to derive theoretical guarantees for transfer learning based on deep ReLU networks, extending the earlier analysis of the prediction risk [1,22,38]. Here we briefly illustrate this using shallow ReLU networks of the form

$$f(x) = \sum_{i=1}^N a_i(w_i x - v_i)_+ \quad \text{with } (u)_+ := \max(u,0) \quad \text{and } a_i, w_i, v_i \in \mathbb{R}.$$

Denote by $\mathrm{ReLU}_N(1-\delta)$ the function class of all such shallow ReLU networks that are moreover $(1-\delta)$-Lipschitz.

Lemma 19. *If* $N \geq n-1 \geq 1$, *then,*

$$\widehat{f}_n \in \mathop{\arg\min}_{f \in \mathrm{ReLU}_N(1-\delta)} \sum_{i=1}^n \big(Y_i - f(X_i)\big)^2,$$

implies that $\widehat{f}_n$ *is also a minimizer over all* $\mathrm{Lip}(1-\delta)$ *functions, that is,*

$$\widehat{f}_n \in \mathop{\arg\min}_{f \in \mathrm{Lip}(1-\delta)} \sum_{i=1}^n \big(Y_i - f(X_i)\big)^2.$$

The result shows that a global minimizer over all ReLU networks in the class $\mathrm{ReLU}(1-\delta)$ is also an empirical risk minimizer over all $\mathrm{Lip}(1-\delta)$-functions. In particular, this means that all the results derived in this paper can be immediately applied, leading to local convergence rates and theoretical guarantees in the case of transfer learning.

Another possible future direction is to use the refined analysis and the local convergence of the LSE to prove distributional properties similar to those established for the least squares procedure under shape constraints. See also the discussion in Section 6.1.

7. Proof of the local convergence rate for the LSE

We now describe the construction of the local perturbation and give the proof of Theorem 4.

Preliminary: Concentration of histogram

For sufficiently large sample size n, we can find an integer sequence $(N_n)_n$ satisfying

$$1 \leq \frac{N_n}{16}\sqrt{\frac{\log n}{n}} \leq 2.$$

For discretization step size

$$\Delta_n := \frac{1}{N_n} \tag{27}$$

we show that $n^{-1}\sum_{i=1}^n \mathbb{1}(X_i \in [j\Delta_n, k\Delta_n])$ concentrates around its expectation $\int_{j\Delta_n}^{k\Delta_n} p(u)\,du$. For this purpose, we first recall the classical Bernstein inequality for Bernoulli random variables.

Lemma 20 (Bernstein inequality). *Let* $p \in [0,1]$ *and* $V_1, \ldots, V_n$ *be n independent Bernoulli variables with success probability p, then,*

$$\mathrm{P}\left(\frac{1}{2}p \leq \frac{1}{n}\sum_{i=1}^n V_i \leq \frac{3}{2}p\right) \geq 1 - 2\exp\left(-\frac{np}{10}\right).$$

Proof. Let $U_i = V_i - p$. We have, $|U_i| \leq 1$, $\mathrm{E}[U_i] = 0$ and $\mathrm{E}[U_i^2] = \mathrm{Var}(V_i) = p(1-p) \leq p$. By Bernstein's inequality,

$$\mathrm{P}\left(\left|\sum_{i=1}^n U_i\right| \geq \frac{n}{2}p\right) \leq 2\exp\left(-\frac{n^2p^2/8}{np + np/6}\right) \leq 2\exp\left(-\frac{np}{8+4/3}\right) < 2\exp\left(-\frac{np}{10}\right). \qquad \square$$

Set $\overline{p}_{j,k} := \int_{[j\Delta_n, k\Delta_n]} p(u)\,du$ and define $\Gamma_n(\alpha)$ as the event

$$\Gamma_n(\alpha) := \bigcup_{\substack{j,k=1,\dots,N_n \\ \overline{p}_{j,k} \geq \alpha \frac{\log^2 n}{n}}} \left\{ X_1,\dots,X_n : \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \in [j\Delta_n, k\Delta_n]) - \overline{p}_{j,k} \right| > \frac{\overline{p}_{j,k}}{2} \right\}. \tag{28}$$

Roughly speaking, this set consists of all samples for which the histogram does not concentrate well around its expectation.

Lemma 21. *If $\alpha > 0$, then, the probability of the event $\Gamma_n(\alpha)$ vanishes as the sample size grows. More precisely,*

$$\mathrm{P}(\Gamma_n(\alpha)) \leq \frac{1024n}{\log n} \exp\left(-\frac{\alpha \log^2 n}{10} \right) \to 0 \quad \textit{as } n \to \infty.$$

Proof. Use the union bound and apply Lemma 20. Since $N_n \leq 32\sqrt{n/\log n}$, we obtain the inequality. The convergence to zero follows from $\alpha > 0$. □

The previous result allows us to work on the event $\Gamma_n(\alpha)^c$. On this event, the random quantity $n^{-1} \sum_{i=1}^{n} \mathbb{1}(X_i \in [j\Delta_n, k\Delta_n])$ is the same as its expectation $\int_{j\Delta_n}^{k\Delta_n} p(u)\,du$ up to a factor two. In particular, we will apply this to random integers j,k depending on the sample $(X_1,Y_1),\dots,(X_n,Y_n)$. We frequently use that for an X that is independent of the data,

$$\mathrm{P}_{\mathrm{X}}(X \in A) = \mathrm{P}\left(X \in A \,|\, (X_1,Y_1),\dots,(X_n,Y_n)\right). \tag{29}$$

The proof outline in Section 4 assumes that there exists a function g lying at all data points between $\widehat{f}_n$ and f_0 in the sense that $(\widehat{f}_n(X_i) - g(X_i))(g(X_i) - f_0(X_i)) \geq 0$ for all $i = 1,\dots,n$. The next lemma guarantees the existence of a sufficiently large local perturbation g of $\widehat{f}_n$ with this property. This will allow us later to carry out the proof strategy sketched in Section 4.

Lemma 22 (Construction of a local perturbation). *Let $\psi \in \mathrm{Lip}(1)$ and $f \in \mathrm{Lip}(1-\delta)$. Define $x^* \in \arg\max_{x\in[0,1]}(\psi(x) - f(x))/t_n(x)$ and $\tilde{x} \in \arg\max_{x\in[0,1]} \psi(x) - f(x) - \frac{\delta}{2}|x - x^*|$. Assume the existence of some $K > 0$, such that*

$$\frac{\psi(x^*) - f(x^*)}{t_n(x^*)} \geq K,$$

and set

$$s_n := 2Kt_n(\tilde{x}) \wedge \left(2Kt_n(x^*) + \frac{\delta}{2}|x^* - \tilde{x}| \right). \tag{30}$$

Then there exists a function g_n and two real numbers $0 \leq x_\ell \leq x_u \leq 1$, such that

- (i) $g_n \in \mathrm{Lip}(1)$ *and* $\mathrm{supp}(\psi - g_n) = [x_\ell, x_u]$.
- (ii) $f \leq g_n \leq \psi$ *on* $[x_\ell, x_u]$.
- (iii) $\tilde{x} - s_n/\delta \leq x_\ell \leq \left(\tilde{x} - s_n/4\right) \vee 0 \leq \left(\tilde{x} + s_n/4\right) \wedge 1 \leq x_u \leq \tilde{x} + s_n/\delta$.
- (iv) *the inequality $\psi(x) - g_n(x) \geq s_n/4$ holds for all $x \in [\tilde{x} \pm s_n/8] \cap [0,1]$.*

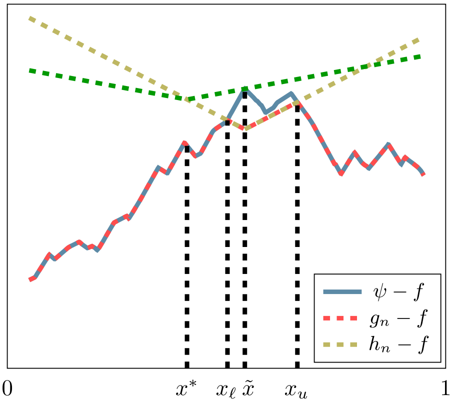

Figure 3. (Construction of the local perturbation) The variables x^* and $\tilde{x}$ are as in Lemma 22. From the construction of $\tilde{x}$, we know that the function $\psi - f$ (plotted in blue) cannot lie above the green dashed curve with slope $\delta/2$. The yellow function is $h_n - f$. Since this function has slope δ, it will hit the green dashed curve in a neighbourhood of $\tilde{x}$. This also implies that $h_n - f$ intersects for the first time with $\psi - f$ (blue curve) in this neighbourhood and provides us with control for the hitting points x_ℓ and x_u. The (shifted) perturbation $g_n - f$ is given by the red curve.

Proof. We construct a function g_n satisfying all claimed properties. The construction requires several steps and can be understood best through the visualization in Figure 3. Recall that $\tilde{x} \in \arg\max_{x\in[0,1]} \psi(x) - f(x) - \frac{\delta}{2}|x - x^*|$. Hence, $\psi(\tilde{x}) - f(\tilde{x}) - \frac{\delta}{2}|\tilde{x} - x^*| \geq \psi(x^*) - f(x^*)$. By assumption, we have $\psi(x^*) - f(x^*) > t_n(x^*)K$ and thus

$$\psi(\tilde{x}) - f(\tilde{x}) > t_n(x^*)K + \frac{\delta}{2}|\tilde{x} - x^*| \geq \frac{s_n}{2}.$$

Define the function

$$h_n(x) := \psi(\tilde{x}) - f(\tilde{x}) + \delta|x - \tilde{x}| + f(x) - \frac{s_n}{2}.$$

Since $f \in \mathrm{Lip}(1-\delta)$ and $\delta|\cdot| \in \mathrm{Lip}(\delta)$, we have that $h_n \in \mathrm{Lip}(1)$. By construction, $h_n(\tilde{x}) = \psi(\tilde{x}) - s_n/2 < \psi(\tilde{x})$. Denote by x_ℓ the largest x below $\tilde{x}$ satisfying $h_n(x) = \psi(x)$. If no such x exists, set $x_\ell := 0$. Similarly, define x_u as the smallest x above $\tilde{x}$ satisfying $h_n(x) = \psi(x)$ and set $x_u := 1$ if this does not exist. Define

$$g_n(x) := \psi(x)\mathbb{1}(x \in [0,1] \setminus [x_\ell, x_u]) + h_n(x)\mathbb{1}(x \in [x_\ell, x_u]). \tag{31}$$

By construction, $g_n \in \mathrm{Lip}(1)$ and $\mathrm{supp}(\psi - g_n) = [x_\ell, x_u]$. Thus (i) holds. Also, (ii) follows directly from the inequalities above.

We now prove (iii). Applying triangle inequality yields

$$\begin{aligned} h_n(x) - f(x) &= \psi(\tilde{x}) - f(\tilde{x}) - \frac{\delta}{2}|\tilde{x} - x^*| + \frac{\delta}{2}|\tilde{x} - x^*| + \delta|x - \tilde{x}| - \frac{1}{2}s_n \\ &\geq \psi(x) - f(x) - \frac{\delta}{2}|x - x^*| + \frac{\delta}{2}|\tilde{x} - x^*| + \delta|x - \tilde{x}| - \frac{1}{2}s_n \\ &\geq \psi(x) - f(x) + \frac{\delta}{2}|x - \tilde{x}| - \frac{1}{2}s_n. \end{aligned}$$

From the last inequality, we deduce that for any $x \in [0,1]$ with $|x - \tilde{x}| \geq s_n/\delta$, we have $h_n(x) - f(x) \geq \psi(x) - f(x)$. Thus,

$$\tilde{x} - \frac{s_n}{\delta} \leq x_\ell \leq x_u \leq \tilde{x} + \frac{s_n}{\delta}, \tag{32}$$

proving the first and last inequality in (iii).

To prove the remaining inequalities in (iii), we use $\psi - f \in \mathrm{Lip}(2-\delta)$ to deduce $\psi(x) - f(x) \geq \psi(\tilde{x}) - f(\tilde{x}) - (2-\delta)|x - \tilde{x}|$. By definition, we have moreover $h_n(x) - f(x) = \psi(\tilde{x}) - f(\tilde{x}) + \delta|x - \tilde{x}| - s_n/2$ and therefore $\psi(x) - f(x) + 2|x - \tilde{x}| \geq h_n(x) - f(x) + s_n/2$, which can be rewritten into

$$\psi(x) - h_n(x) \geq \frac{s_n}{2} - 2|x - \tilde{x}|. \tag{33}$$

The right-hand side of this inequality is > 0 for all $x \in [\tilde{x} \pm s_n/4] \cap [0,1]$. The definition of x_ℓ and x_u implies then that $x_\ell \leq 0 \vee \left(\tilde{x} - s_n/4\right) \leq 1 \wedge \left(\tilde{x} + s_n/4\right) \leq x_u$.

We now establish (iv). One can use the lower bound from Equation (33) to obtain that for any $x \in [\tilde{x} \pm s_n/8] \cap [0,1]$,

$$\psi(x) - f(x) \geq \frac{s_n}{2} - 2|x - \tilde{x}| + g_n(x) - f(x) \geq \frac{s_n}{4} + g_n(x) - f(x) \geq \frac{s_n}{4},$$

applying (ii) for the last inequality. This proves (iv). □

Lemma 23. *For given $K > 1/2$ and $0 < \delta < 1$, let s_n and $\tilde{x}$ be as defined in Lemma 22. Moreover, let $\mathrm{P_X} \in \mathcal{P}_n(D)$ for some $D \geq 2$. If $n \geq \exp(4K^2 \vee 36(1 \vee \log^2 D))$ and $0 < c\delta \leq 2$ then,*

- (i) $P_X\left([\tilde{x} \pm cs_n]\right) > D^{-\lceil \log_2(1/(\delta c)) \rceil - 1} \frac{\log^2 n}{n}$,
- (ii) $s_n^2 \mathrm{P_X}\left([\tilde{x} \pm cs_n]\right) \geq D^{-\lceil \log_2(1/(\delta c)) \rceil - 1} 4K^2 \frac{\log n}{n}$.

Proof. Recall that $s_n = 2Kt_n(\tilde{x}) \wedge (2Kt_n(x^*) + \delta|x^* - \tilde{x}|/2)$. Since $n \geq \exp(4K^2)$, $s_n \leq 2Kt_n(\tilde{x}) \leq \sqrt{\log n}\, t_n(\tilde{x}) \leq \sqrt{\log n} \sup_{x \in [0,1]} t_n(x)$. This allows to apply now the local doubling property of $\mathrm{P_X}$ to intervals of length up to s_n.

By assumption $K \geq 1/4$ and $\delta < 1$. Hence $2Kt_n(\tilde{x}) \geq \delta t_n(\tilde{x})/2$. Moreover, using the fact that t_n is 1-Lipschitz from Lemma 27 (Appendix A of the supplement [39]), $2Kt_n(x^*) + \delta|x^* - \tilde{x}|/2 \geq \frac{\delta}{2}\left[t_n(x^*) + |x^* - \tilde{x}|\right] \geq \delta t_n(\tilde{x})/2$. Combining the two previous bounds, we obtain $cs_n \geq \delta ct_n(\tilde{x})/2$. Using the latter and applying (LDP) in total $\lceil \log_2(1/(\delta c)) \rceil + 1$ times to increase the constant $\delta c/2$ to $\delta c 2^{\lceil \log_2(1/(\delta c)) \rceil} \in [1,2]$, we find whenever $\delta c \leq 2$,

$$\begin{aligned}
\mathrm{P_X}\left(\left[\tilde{x} \pm cs_n\right]\right) &\geq \mathrm{P_X}\left(\left[\tilde{x} \pm \delta ct_n(\tilde{x})/2\right]\right) \\
&\geq D^{-\lceil \log_2(1/(\delta c)) \rceil - 1} \mathrm{P_X}\left(\left[\tilde{x} \pm \delta c 2^{\lceil \log_2(1/(\delta c)) \rceil} t_n(\tilde{x})\right]\right) \\
&\geq D^{-\lceil \log_2(1/(\delta c)) \rceil - 1} \mathrm{P_X}\left(\left[\tilde{x} \pm t_n(\tilde{x})\right]\right).
\end{aligned}$$

Applying Remark 2 (Appendix A of the supplement [39]) completes the proof of (i). To prove the second claim, we once again lower bound cs_n.

We first consider the case $s_n = 2Kt_n(\tilde{x})$. If $c \geq 1$, then, we get $cs_n \geq s_n \geq 2Kt_n(\tilde{x}) \geq t_n(\tilde{x})$ and

$$\begin{aligned} s_n^2 \mathrm{P_X}\left(\left[\tilde{x} \pm cs_n\right]\right]\right) &\geq 4K^2 t_n(\tilde{x})^2 \mathrm{P_X}\left(\left[\tilde{x} \pm t_n(\tilde{x})\right]\right) \\ &= 4K^2 \frac{\log n}{n} \\ &\geq \left(1 \wedge D^{-\lceil \log_2(1/(\delta c))\rceil - 1}\right) 4K^2 \frac{\log n}{n}. \end{aligned}$$

Otherwise, if $c < 1$, then, we can apply (LDP) in total $k = \lceil \log_2(1/c)\rceil$ times, so that $2^k c \geq 1$, and obtain

$$\begin{aligned} s_n^2 \mathrm{P_X}\left(\left[\tilde{x} \pm cs_n\right]\right]\right) &\geq 4K^2 D^{-\lceil \log_2(1/c)\rceil} t_n(\tilde{x})^2 \mathrm{P_X}\left(\left[\tilde{x} \pm 2^k ct_n(\tilde{x})\right]\right) \\ &= 4K^2 D^{-\lceil \log_2(1/c)\rceil} \frac{\log n}{n} \\ &\geq \left(1 \wedge D^{-\lceil \log_2(1/(\delta c))\rceil - 1}\right) 4K^2 \frac{\log n}{n}. \end{aligned}$$

We now consider the case $cs_n = 2Kct_n(x^*) + \delta c|x^* - \tilde{x}|/2$. Suppose without loss of generality that $x^* \leq \tilde{x}$. If $c \geq 2/\delta > 1$, then we get

$$\begin{aligned} \left[\tilde{x} \pm cs_n\right] &\supset \left[\tilde{x} - |x^* - \tilde{x}| - 2Kt_n(x^*), \tilde{x} + |x^* - \tilde{x}| + 2Kt_n(x^*)\right] \\ &\supset \left[x^* - 2Kt_n(x^*), x^* + 2Kt_n(x^*)\right] \\ &\supset \left[x^* \pm t_n(x^*)\right], \end{aligned}$$

which implies

$$\begin{aligned} s_n^2 \mathrm{P_X}\left(\left[\tilde{x} \pm cs_n\right]\right) &\geq 4K^2 t_n(x^*) \mathrm{P_X}\left(\left[x^* \pm t_n(x^*)\right]\right) \\ &= 4K^2 \frac{\log n}{n} \\ &\geq \left(1 \wedge D^{-\lceil \log_2(1/(\delta c))\rceil - 1}\right) 4K^2 \frac{\log n}{n}. \end{aligned}$$

Otherwise, if $c < 2/\delta$, we can apply (LDP) in total $k = \lceil \log_2(1/(\delta c))\rceil + 1$ times, so that $2^k \delta c/2 \geq 1$ to obtain

$$s_n^2 \mathrm{P_X}\left(\left[\tilde{x} \pm cs_n\right]\right]\right) \geq 4K^2 D^{-\lceil \log_2(1/(\delta c))\rceil - 1} t_n(x^*)^2 \mathrm{P_X}\left(\left[\tilde{x} \pm 2^k cs_n\right]\right).$$

Since $2^k c \geq 2/\delta$, we proceed as in the previous case and find

$$s_n^2 \mathrm{P_X}\left(\left[\tilde{x} \pm cs_n\right]\right) \geq 4K^2 D^{-\lceil \log_2(1/(\delta c))\rceil - 1} \frac{\log n}{n} \geq \left(1 \wedge D^{-\lceil \log_2(1/(\delta c))\rceil - 1}\right) 4K^2 \frac{\log n}{n}.$$

Combining both cases, for any $K > 1/2$, any $0 < \delta < 1$ and any $c > 0$,

$$s_n^2 \mathrm{P_X}\left(\left[\tilde{x} \pm cs_n\right]\right) \geq \left(1 \wedge D^{-\lceil \log_2(1/(\delta c))\rceil - 1}\right) 4K^2 \frac{\log n}{n}.$$

Finally, if $c < 4/\delta$, then $\lceil \log_2(1/(\delta c)\rceil \geq -1$ and (ii) follows. □

Proof of the main theorem

Proof of Theorem 4. The proof follows the steps outlined in Section 4. Because of $|z| = \max(z, -z)$ and since all arguments carry over to the other case, it is enough to show that

$$\sup_{\mathrm{P_X} \in \mathcal{P}_n(D)} \ \sup_{f_0 \in \mathrm{Lip}(1-\delta)} \mathrm{P}_{f_0}\left(\sup_{x\in[0,1]} \frac{\widehat{f}_n(x) - f_0(x)}{t_n(x)} \ge K \right) \to 0 \quad \text{as } n \to \infty.$$

We will derive a contradiction by considering

$$K > K_* := \frac{1}{2} \vee 3\big(1 \vee \log(D)\big) \vee \frac{2^{11/2} D^2}{21^3} \delta^{-\log_2(D)/2} \vee (32\chi D^5)^{3/4} D^{1+\frac{1}{2}\log_2(1/(2\delta))} \tag{34}$$

where

$$\chi := 2^{17/3} \cdot 21^2 D^{\lceil \log_2(\delta^{-1}) \rceil/3} \delta^{-2/3}. \tag{35}$$

The condition $K > 3(1 \vee \log(D))$ ensures that if $n \ge \exp(4K^2)$, then, $n \ge \exp(4K^2 \vee (36(1 \vee \log^2(D))))$. Therefore, we can apply Lemma 23 under the simplified condition $n \ge \exp(4K^2)$.

Inequality (14) states that if $g_n \in \mathrm{Lip}(1)$ satisfies $(\widehat{f}_n(X_i) - g_n(X_i))(g_n(X_i) - f_0(X_i)) \ge 0$ for all $i = 1, \dots, n$, then,

$$\sum_{i=1}^n (\widehat{f}_n(X_i) - g_n(X_i))^2 \le 2 \sum_{i=1}^n \varepsilon_i (\widehat{f}_n(X_i) - g_n(X_i)). \tag{36}$$

Let g_n be the local perturbation constructed in Lemma 22 with ψ and f replaced by $\widehat{f}_n$ and f_0, respectively. In particular, Lemma 22 (ii) ensures that $(\widehat{f}_n(x) - g_n(x))(g_n(x) - f_0(x)) \ge 0$ for all $x \in [0,1]$. As indicated in Section 4, we begin by lower bounding the left-hand side of inequality (36).

Lower bound for the left-hand side of (36): Work on the event $\Gamma_n(D^{-4})^c$ defined in (28). Let $x^*, \tilde{x}$ and s_n be defined as in Lemma 22, that is, $x^* \in \arg\max_{x\in[0,1]} t_n(x)^{-1}(\widehat{f}_n(x) - f_0(x))$, $\tilde{x} \in \arg\max_{x\in[0,1]} \widehat{f}_n(x) - f_0(x) - \frac{\delta}{2}|x - x^*|$ and $s_n = 2Kt_n(\tilde{x}) \wedge (2Kt_n(x^*) + \frac{\delta}{2}|x^* - \tilde{x}|)$. It is sufficient to show the result for all sufficiently large n. In particular, it is enough to consider $n \ge \exp(4K^2/\delta^2)$, ensuring that

$$\frac{s_n}{\delta} \le \frac{2K}{\delta} t_n(\tilde{x}) \le \sqrt{\log n} \sup_{x\in[0,1]} t_n(x). \tag{37}$$

Lemma 22 (iii) shows existence of an interval $I := [\tilde{x} \pm s_n/8] \cap [0,1]$ with length $\ge (s_n/8) \wedge 1/2$ such that for all $x \in I$, $\widehat{f}_n(x) - g_n(x) \ge s_n/4$. By restriction of the sum to $\{i : X_i \in I\}$ and using Lemma 22 (iv), we find

$$\sum_{i=1}^n \big(\widehat{f}_n(X_i) - g_n(X_i)\big)^2 \ge \left(\frac{s_n}{4}\right)^2 \sum_{i=1}^n \mathbb{1}(X_i \in I). \tag{38}$$

We now relate the right-hand side of (38) to our discretization of $[0,1]$ with step size Δ_n defined in (27). By (34), $K \ge 1/2$. Thus, by Lemma 27 (ii) (Appendix A of the supplement [39]), $s_n \ge t_n(\tilde{x}) \ge \sqrt{\log n/n}$ and

$$\Delta_n \le \frac{1}{16}\sqrt{\frac{\log n}{n}} \le \frac{s_n}{16}. \tag{39}$$

Hence, there exist two random integers $0 \le \ell_1 < k_1 \le N_n$ satisfying

$$\Big(\tilde{x} - \frac{s_n}{8}\Big) \vee 0 \le \ell_1 \Delta_n \le \Big(\tilde{x} - \frac{s_n}{16}\Big) \vee 0 < \Big(\tilde{x} + \frac{s_n}{16}\Big) \wedge 1 \le k_1 \Delta_n \le \Big(\tilde{x} + \frac{s_n}{8}\Big) \wedge 1.$$

This implies $(k_1 - \ell_1)\Delta_n \ge s_n/16 > 0$ and $[\ell_1\Delta_n, k_1\Delta_n] \subseteq I = [\tilde{x} \pm s_n/8] \cap [0,1]$. Applying the lower bound (i) in Lemma 23, we find

$$\mathrm{P_X}\left([\ell_1\Delta_n, k_1\Delta_n]\right) \ge \mathrm{P_X}\left(\Big[\tilde{x} \pm \frac{s_n}{16}\Big]\right) \ge \frac{\log^2 n}{D^4 n},$$

with $\mathrm{P_X}$ the conditional distribution defined in (29). By the definition of the event $\Gamma_n(D^{-4})^c$ in (28), we have $n^{-1}\sum_{i=1}^n \mathbb{1}(X_i \in I) \ge n^{-1}\sum_{i=1}^n \mathbb{1}(X_i \in [\ell_1\Delta_n, k_1\Delta_n]) \ge \mathrm{P_X}([\ell_1\Delta_n, k_1\Delta_n])/2$. This and our choice of ℓ_1, k_1 means that, on $\Gamma_n(D^{-4})^c$, inequality (38) implies

$$\sum_{i=1}^n \Big(\widehat{f}_n(X_i) - g_n(X_i)\Big)^2 \ge \Big(\frac{s_n}{4}\Big)^2 \frac{n}{2} \mathrm{P_X}\left([\ell_1\Delta_n, k_1\Delta_n]\right) \ge \frac{s_n^2 n}{32} \mathrm{P_X}\left(\Big[\tilde{x} \pm \frac{s_n}{16}\Big]\right).$$

By (37), $s_n \le \sqrt{\log n}\sup_{x\in[0,1]} t_n(x)$. This allows to apply (LDP) in total five times to obtain

$$\sum_{i=1}^n \Big(\widehat{f}_n(X_i) - g_n(X_i)\Big)^2 \ge \frac{s_n^2 n}{32 D^5} \mathrm{P_X}\left(\big[\tilde{x} \pm 2s_n\big]\right). \tag{40}$$

Upper bound for the right-hand side of (36)*:* We now derive an upper bound for $2\sum_{i=1}^n \varepsilon_i\big(\widehat{f}(X_i) - g_n(X_i)\big)$. Since $\widehat{f}_n - g_n$ is supported on a small subset of $[0,1]$, it is advantageous to study the sum over X_i in the support. For $0 \le a < b \le 1$, denote the class of 1-Lipschitz functions supported on the interval $[a,b]$ by

$$\mathcal{E}_{[a,b]} := \{h \in \mathrm{Lip}(1) : \mathrm{supp}(h) \subset [a,b]\}.$$

Define $m(X)$ as the cardinality of the set $\{i \in \{1, \dots n\} : X_i \in [a,b]\}$ and write $Z_1, \dots, Z_{m(X)}$ for the $m(X)$ variables X_i that fall into the interval $[a,b]$. For a function $h \in \mathcal{E}_{[a,b]}$, we define the effective empirical L_2-norm,

$$\|h\|_{m(X)} := \left(\frac{1}{m(X)}\sum_{i=1}^{m(X)} h(Z_i)^2\right)^{1/2}.$$

Here, effective refers to the fact that the semi-norm is computed based on the 'effective' sample $Z_1, \dots, Z_{m(X)}$. The normalization is chosen such that $n\|h\|_n^2 = m(X)\|h\|_{m(X)}^2$ for all $h \in \mathcal{E}_{[a,b]}$. From now on, we follow the same steps as in Chapter 13 from [48] with the sample size n replaced by the effective sample size $m(X)$. The so-called critical inequality with $\sigma = 1$ is

$$\frac{\mathcal{G}_n(\eta, \mathcal{E}_{[a,b]})}{\eta} \le \frac{\eta}{2}, \tag{41}$$

with $\mathcal{G}(\eta, \mathcal{E}_{[a,b]})$ the Gaussian complexity of the set $\mathcal{E}_{[a,b]}$, that is,

$$\mathcal{G}_n(\eta, \mathcal{E}_{[a,b]}) := \mathrm{E}_\varepsilon\left[\sup_{h \in \mathcal{E}_{[a,b]},\, \|h\|_{m(X)} \le \eta} \frac{1}{m(X)}\left|\sum_{i=1}^{m(X)} \varepsilon_i h(Z_i)\right|\right].$$

Recall that a function class $\mathcal{F}$ is called star-shaped if $f \in \mathcal{F}$ implies $\alpha f \in \mathcal{F}$ for all $0 \leq \alpha \leq 1$. Observing that $\mathcal{E}_{[a,b]}$ is star-shaped, one can derive the following modified version of Theorem 13.1 in [48].

Theorem 24. *If η is a solution of the critical inequality (41), then for any $t \geq \eta$,*

$$\mathrm{P}\Big(\big\{\|\widehat{f}_n - g_n\|_{m(X)}^2 \geq 16t\eta\big\} \cap \big\{\operatorname{supp}(\widehat{f}_n - g_n) \subset [a,b]\big\} \,\Big|\, X_1,\dots,X_n\Big) \leq e^{-\frac{m(X)t\eta}{2}}.$$

The covering number of $\mathcal{F}^*$ with sup-norm balls of radius r is denoted by $N(r,\mathcal{F}^*,\|\cdot\|_\infty)$. Along with Theorem 24 comes a modified version of Corollary 13.1 in [48] stating a sufficient condition for η to solve the critical inequality.

Corollary 25. *Set $B_n(\eta,\mathcal{E}_{[a,b]}) := \{h \in \mathcal{E}_{[a,b]} : \|h\|_{m(X)} \leq \eta\}$. Under the conditions of Theorem 24, any $\eta \in (0,1]$ satisfying*

$$\frac{16}{\sqrt{m(X)}} \int_{\frac{\eta^2}{4}}^{\eta} \sqrt{\log N\big(t, B_n(\eta,\mathcal{E}_{[a,b]}, \|\cdot\|_\infty)\big)}\, dt \leq \frac{\eta^2}{4} \tag{42}$$

also satisfies the critical inequality and can be used in the conclusion of Theorem 24.

We now derive a bound for the covering number of the class $\mathcal{E}_{[a,b]}$ by slightly refining the proof of classical results such as Corollary 2.7.10 in [47], see Appendix B of the supplement [39] for a full proof.

Lemma 26. *Given two real numbers $a < b$, let $\mathcal{E}_{[a,b]} := \{u \in \mathrm{Lip}(1) : \operatorname{supp}(u) \subset [a,b]\}$. Then, for any positive r,*

$$\log N\big(r, \mathcal{E}_{[a,b]}, \|.\|_\infty\big) \leq \frac{b-a}{r} \log 3.$$

This allows us to upper bound the left-hand side of (42) by $32\sqrt{m(X)^{-1}\eta(b-a)\log 3}$ and, for $n \geq 9$, by $32\sqrt{m(X)^{-1}\eta(b-a)\log(n)/2}$. Hence, if η satisfies

$$\left(2^{13}\frac{(b-a)\log(n)}{m(X)}\right)^{1/3} \leq \eta,$$

then it also satisfies (42). The latter inequality holds for $\eta = 21\big(\frac{(b-a)\log n}{m(X)}\big)^{1/3}$.

We further work on the classes $\mathcal{E}_{[\ell\Delta_n, k\Delta_n]}$ with $0 \leq \ell < k \leq N_n$ and adapt the notation by defining $m_{k,\ell}(X)$ as the cardinality of the set $\{i \in \{1,\dots,n\} : X_i \in [\ell\Delta_n, k\Delta_n]\}$, and

$$\eta_{n,k,\ell} := 21\Big(\frac{(k-\ell)\Delta_n \log n}{m_{k,\ell}(X)}\Big)^{1/3}.$$

Set

$$\mathcal{S}_{k,\ell} := \big\{(X_1,Y_1),\dots,(X_n,Y_n) : \operatorname{supp}(\widehat{f}_n - g_n) \subseteq [\ell\Delta_n, k\Delta_n]\big\}.$$

On $\mathcal{S}_{k,\ell}$, we have by Lemma 22 (iii), that $[\tilde{x} \pm s_n/4] \subseteq [x_\ell, x_u] = \operatorname{supp}(\widehat{f}_n - g_n) \subseteq [\ell\Delta_n, k\Delta_n]$. Using the first claim of Lemma 23 and the fact that $D \geq 2$, we find that

$$\mathrm{P}_{\mathrm{X}}\left([\ell\Delta_n, k\Delta_n]\right) > \frac{\log^2 n}{D^4 n}. \tag{43}$$

The second claim of Lemma 23 combined with $P([\tilde{x} \pm a]) = 1$ for $a \geq 1$, yields

$$\begin{aligned}
\mathrm{P}_{\mathrm{X}}([\ell\Delta_n, k\Delta_n])(k-\ell)^2\Delta_n^2 &\geq \mathrm{P}_{\mathrm{X}}\left(\left[\tilde{x} \pm \frac{s_n}{4}\right]\right)\left(\frac{s_n}{4} \wedge \frac{1}{2}\right)^2 \\
&\geq \frac{1}{4}\mathrm{P}_{\mathrm{X}}\left(\left[\tilde{x} \pm \frac{s_n}{4}\right]\right)\left(\frac{s_n}{4} \wedge 1\right)^2 \\
&\geq \frac{1}{4} \wedge \left(\frac{s_n^2}{64}\mathrm{P}_{\mathrm{X}}\left(\left[\tilde{x} \pm \frac{s_n}{4}\right]\right)\right) \\
&\geq \frac{1}{4} \wedge \frac{K^2}{16}\frac{\log n}{n} D^{\log_2(\delta)-4}.
\end{aligned} \tag{44}$$

Define the set $\mathcal{T} = \{(k,\ell) : \mathrm{P}_{\mathrm{X}}([\ell\Delta_n, k\Delta_n]) > \log^2 n/(D^4 n)\}$ and note that by (43), $(k,\ell) \notin \mathcal{T}$ implies that $\mathcal{S}_{k,\ell}$ is the empty set. Applying Corollary 25, we obtain

$$\begin{aligned}
\mathrm{P}\Bigg(\Bigg\{\frac{1}{m_{k,\ell}(X)}\sum_{i=1}^n \left(\widehat{f}_n(X_i) - g_n(X_i)\right)^2 \geq 16t\eta_{n,k,\ell}\Bigg\} \cap \mathcal{S}_{k,\ell}\,\Bigg|\, X_1, \dots, X_n\Bigg)& \\
\leq e^{-\frac{m_{k,\ell}(X)t\eta_{n,k,\ell}}{2}}\mathbb{1}\big((k,\ell) \in \mathcal{T}\big)&.
\end{aligned} \tag{45}$$

Recall that we work on the event $\Gamma(D^{-4})^c$. For any pair $(k,\ell) \in \mathcal{T}$, we have by (43), $\frac{n}{2}\mathrm{P}_{\mathrm{X}}([\ell\Delta_n, k\Delta_n]) \leq m_{k,\ell}(X) \leq 2n\,\mathrm{P}_{\mathrm{X}}([\ell\Delta_n, k\Delta_n])$. Multiplying by $\eta_{n,k,\ell}^2$ and rearranging the terms yields

$$\frac{1}{2^{1/3}} \leq \frac{m_{k,\ell}(X)\eta_{n,k,\ell}^2}{21^2 P_{k,\ell}^{1/3}} \leq 2^{1/3}, \quad \text{with } \ P_{k,\ell} := n\,\mathrm{P}_{\mathrm{X}}([\ell\Delta_n, k\Delta_n])(k-\ell)^2\Delta_n^2 \log^2 n. \tag{46}$$

Consider the event

$$\begin{aligned}
\mathcal{D}_{k,\ell} &:= \underbrace{\left\{\sum_{i=1}^n \left(\widehat{f}(X_i) - g(X_i)\right)^2 \geq 16 \cdot 21^2 (2P_{k,\ell})^{1/3}\right\}}_{:=A} \cap \mathcal{S}_{k,\ell} \cap \Gamma_n(D^{-4})^c \\
&\subseteq \left\{\frac{1}{m_{k,\ell}(X)}\sum_{i=1}^n \left(\widehat{f}(X_i) - g(X_i)\right)^2 \geq 16\eta_{n,k,\ell}^2\right\} \cap \mathcal{S}_{k,\ell} \cap \Gamma_n(D^{-4})^c.
\end{aligned}$$

Let A, B be measurable sets and assume that $\mathrm{P}(A|X) \leq a(X)$. If B only depends on X, then, we have that $\mathrm{P}(A \cap B) = \mathrm{E}[\mathrm{P}(A \cap B|X)] = \mathrm{E}[\mathrm{P}(A|X)\mathbb{1}(X \in B)] \leq \mathrm{E}[a(X)\mathbb{1}(X \in B)]$. Below we apply this inequality for X the sample $(X_1, Y_1), \dots, (X_n, Y_n)$, A the event defined in the previous display, $B = S_{k,\ell} \cap \Gamma_n(D^{-4})^c$ and $a(X) = \exp(-\frac{1}{2}m_{k,\ell}(X)\eta_{n,k,\ell}^2)$. By (34), $K > 2^{11/2}21^{-3}D^2\delta^{-\log_2(D)/2}$. Choosing $t = \eta_{n,k,\ell}$ and using (45) as well as (44), we find for any $(k,\ell)\in\mathcal{T}$,

$$
\begin{aligned}
\mathrm{P}(\mathcal{D}_{k,\ell}) &\leq \mathrm{E}\Big[e^{-\frac{m_{k,\ell}(X)\eta^2_{n,k,\ell}}{2}}\mathbb{1}\big(X\in\mathcal{S}_{k,\ell}\cap\Gamma_n(D^{-4})^c\big)\Big]\\
&\leq \mathrm{E}\bigg[\exp\bigg(-\frac{21^2}{2^{4/3}}\Big(n\log^2 n\,\mathrm{P}_{\mathrm{X}}\left([\ell\Delta_n,k\Delta_n]\right)(k-\ell)^2\Delta_n^2\Big)^{1/3}\bigg)\bigg]\\
&\leq \exp\bigg(-\frac{21^2}{2^{4/3}}\bigg[\Big(\frac{n\log^2 n}{4}\Big)^{1/3}\wedge\Big(\frac{K^2}{16}D^{\log_2(\delta)-4}\log^3 n\Big)^{1/3}\bigg]\bigg)\\
&\leq \exp\bigg(-\frac{21^2}{2^{4/3}}\Big(\frac{n\log^2 n}{4}\Big)^{1/3}\bigg)\vee n^{-2},
\end{aligned}
\tag{47}
$$

using $\exp(-a\log n)=n^{-a}$ for the last step. (Choosing the lower bound K_* in (34) large enough, one can modify the previous argument and achieve polynomial decay in n of any order.) If $(k,\ell)\notin\mathcal{T}$, then (45) implies $\mathrm{P}(\mathcal{D}_{k,\ell})=0$.

Define the random variables $0\leq\widehat{\ell}<\widehat{k}\leq N_n$ such that

$$
(\widehat{k},\widehat{\ell})\in\operatorname*{arg\,min}_{(k,\ell):0\leq\ell<k\leq N_n}\big\{(k-\ell)\Delta_n:\operatorname{supp}(\widehat{f}_n-g_n)\subset[\ell\Delta_n,k\Delta_n]\big\}.
$$

With $\mathcal{D}_{k,\ell}$ as defined above, applying (47), $N_n\leq 32\sqrt{n/\log n}$ and the union bound yields

$$
\begin{aligned}
\mathrm{P}(\mathcal{D}_{\widehat{k},\widehat{\ell}}) &\leq \sum_{0\leq\ell<k\leq N_n}\mathrm{P}(\mathcal{D}_{k,\ell})\\
&\leq \sum_{(k,\ell)\in\mathcal{T}}\mathrm{P}(\mathcal{D}_{k,\ell})\\
&\leq \frac{1024n}{\log n}\bigg(\exp\Big(-\frac{21^2}{2^{4/3}}\big(n\log^2 n/4\big)^{1/3}\Big)\vee n^{-2}\bigg)\to 0\quad\text{as } n\to\infty.
\end{aligned}
\tag{48}
$$

The convergence is uniform over $f_0\in\mathrm{Lip}(1)$ and $\mathrm{P}_{\mathrm{X}}\in\mathcal{P}_n(D)$.

In a next step of the proof, we provide a simple upper bound of the least squares distance on the set $\mathcal{D}_{\widehat{k},\widehat{\ell}}$. Inequality (39) shows $\Delta_n\leq s_n/16\leq s_n/\delta$ and Lemma 22 (*iii*) yields

$$
\tilde{x}-2\frac{s_n}{\delta}\leq\tilde{x}-\frac{s_n}{\delta}-\Delta_n\leq\widehat{\ell}\Delta_n\leq x_\ell<x_u\leq\widehat{k}\Delta_n\leq\tilde{x}+\frac{s_n}{\delta}+\Delta_n\leq\tilde{x}+2\frac{s_n}{\delta},
$$

implying $\mathrm{P}_{\mathrm{X}}([\widehat{\ell}\Delta_n,\widehat{k}\Delta_n])\leq\mathrm{P}_{\mathrm{X}}([\tilde{x}\pm 2s_n/\delta])$ and $(\widehat{k}-\widehat{\ell})\Delta_n\leq 4s_n/\delta$. This allows to further upper bound $P_{\widehat{k},\widehat{\ell}}$ by $n\,\mathrm{P}_{\mathrm{X}}([\tilde{x}\pm 2s_n/\delta])(4s_n/\delta)^2$. Using this, rearranging the rightmost inequality in (46) and applying the local doubling property (LDP) $\lceil\log_2(\delta^{-1})\rceil$ times recalling the inequality $s_n/\delta\leq\sqrt{\log n}\sup_x t_n(x)$ derived in (37), shows that, on $\mathcal{D}^c_{\widehat{k},\widehat{\ell}}\cap\Gamma_n(D^{-4})^c$,

$$
\begin{aligned}
\sum_{i=1}^n\big(\widehat{f}_n(X_i)-g_n(X_i)\big)^2 &\leq 16\cdot 21^2\Big(2n\,\mathrm{P}_{\mathrm{X}}\left([\tilde{x}\pm 2s_n/\delta]\right)\Big(\frac{4s_n}{\delta}\Big)^2\log^2 n\Big)^{1/3}\\
&\leq 2^{17/3}\cdot 21^2\delta^{-2/3}\Big(nD^{\lceil\log_2(\delta^{-1})\rceil}\,\mathrm{P}_{\mathrm{X}}\left([\tilde{x}\pm 2s_n]\right)s_n^2\log^2 n\Big)^{1/3}
\end{aligned}
$$

$$\leq \chi\Big(n\,\mathrm{P_X}\left([\tilde{x} \pm 2s_n]\right) s_n^2 \log^2 n\Big)^{1/3},$$

with $\chi = 2^{17/3} \cdot 21^2 D^{\lceil \log_2(\delta^{-1})\rceil/3}\delta^{-2/3}$ as defined in (35).

Combining the bounds for (36)*:* Using the lower bound (40) and the upper bound derived above, (36) implies that on the event $\mathcal{D}^c_{\widehat{k},\widehat{\ell}} \cap \Gamma_n(D^{-4})^c$,

$$\frac{s_n^2 n}{32D^5}\,\mathrm{P_X}\left(\left[\tilde{x} \pm 2s_n\right]\right) \leq \chi\Big(n\,\mathrm{P_X}\left([\tilde{x} \pm 2s_n]\right) s_n^2 \log^2 n\Big)^{1/3}. \tag{49}$$

Rearranging the terms in (49), and raising both sides to the power $3/2$ gives

$$Bs_n^2 n\,\mathrm{P_X}\left([\tilde{x} \pm 2s_n]\right) \leq \log n,$$

with $B := (32\chi D^5)^{-3/2}$. Since χ is an absolute constant independent of K, B is also independent of K. Using the second claim of Lemma 23 and dividing by $\log n$ on both sides, we obtain on the event $\mathcal{D}^c_{\widehat{k},\widehat{\ell}} \cap \Gamma_n(D^{-4})^c$,

$$4K^2 BD^{-\lceil \log_2(1/(2\delta))\rceil - 1} \leq 1.$$

But since by (34), $K > B^{-1/2} D^{1+\frac{1}{2}\log_2(1/(2\delta))}$, we have derived a contradiction. Because K was chosen to be any number larger than K_*, we must have, on the event $\mathcal{D}^c_{\widehat{k},\widehat{\ell}} \cap \Gamma_n(D^{-4})^c$,

$$\sup_{x\in[0,1]} \frac{\widehat{f}_n(x) - f_0(x)}{t_n(x)} \leq K_*.$$

The probability of the exceptional set tends to zero because by (48) and Lemma 21,

$$\mathrm{P}\left((\mathcal{D}^c_{\widehat{k},\widehat{\ell}} \cap \Gamma_n(D^{-4})^c)^c\right) \leq \mathrm{P}(\mathcal{D}_{\widehat{k},\widehat{\ell}}) + \mathrm{P}\left(\Gamma_n(D^{-4})\right) \to 0 \text{ as } n \to \infty.$$

The convergence can be checked to be uniform over $f_0 \in \mathrm{Lip}(1-\delta)$ and $\mathrm{P_X} \in \mathcal{P}_n(D)$. The proof is complete. □

Acknowledgements

The authors would like to thank the editor, the AE and three anonymous referees for helpful comments and suggestions.

Funding

The research has been supported by the NWO/STAR grant 613.009.034b and the NWO Vidi grant VI.Vidi.192.021.

Supplementary Material

Supplement to "Local convergence rates of the nonparametric least squares estimator with applications to transfer learning" (DOI: 10.3150/23-BEJ1655SUPP; .pdf). Additional proofs and technical lemmas.

References

[1] Bauer, B. and Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Statist.* **47** 2261–2285. MR3953451 https://doi.org/10.1214/18-AOS1747

[2] Baxter, J. (1997). A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Mach. Learn.* **28** 7–39. https://doi.org/10.1023/A:1007327622663

[3] Beliakov, G. (2007). Smoothing Lipschitz functions. *Optim. Methods Softw.* **22** 901–916. MR2360803 https://doi.org/10.1080/10556780701393591

[4] Ben-David, S. and Schuller, R. (2003). Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines* (B. Schölkopf and M.K. Warmuth, eds.) 567–580. Berlin, Heidelberg: Springer Berlin Heidelberg.

[5] Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150. MR1240719 https://doi.org/10.1007/BF01199316

[6] Brunk, H.D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Stat.* **26** 607–616. MR0073894 https://doi.org/10.1214/aoms/1177728420

[7] Brunk, H.D. (1958). On the estimation of parameters restricted by inequalities. *Ann. Math. Stat.* **29** 437–454. MR0132632 https://doi.org/10.1214/aoms/1177706621

[8] Buckley, S.M. and MacManus, P. (2000). Singular measures and the key of *G*. *Publ. Mat.* **44** 483–489. MR1800819 https://doi.org/10.5565/PUBLMAT_44200_07

[9] 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

[10] Caruana, R. (1997). Multitask learning. *Mach. Learn.* **28** 41–75. https://doi.org/10.1023/A:1007379606734

[11] Chinot, G., Löffler, M. and van de Geer, S. (2022). On the robustness of minimum norm interpolators and regularized empirical risk minimizers. *Ann. Statist.* **50** 2306–2333. MR4474492 https://doi.org/10.1214/22-aos2190

[12] Dümbgen, L. and Rufibach, K. (2009). Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli* **15** 40–68. MR2546798 https://doi.org/10.3150/08-BEJ141

[13] Gaïffas, S. (2009). Uniform estimation of a signal based on inhomogeneous data. *Statist. Sinica* **19** 427–447. MR2514170

[14] Groeneboom, P., Jongbloed, G. and Wellner, J.A. (2001). Estimation of a convex function: Characterizations and asymptotic theory. *Ann. Statist.* **29** 1653–1698. MR1891742 https://doi.org/10.1214/aos/1015345958

[15] Guntuboyina, A. and Sen, B. (2018). Nonparametric shape-restricted regression. *Statist. Sci.* **33** 568–594. MR3881209 https://doi.org/10.1214/18-STS665

[16] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. *Springer Series in Statistics*. New York: Springer. MR1920390 https://doi.org/10.1007/b97848

[17] Han, Q. (2021). Set structured global empirical risk minimizers are rate optimal in general dimensions. *Ann. Statist.* **49** 2642–2671. MR4338378 https://doi.org/10.1214/21-aos2049

[18] Han, Q., Wang, T., Chatterjee, S. and Samworth, R.J. (2019). Isotonic regression in general dimensions. *Ann. Statist.* **47** 2440–2471. MR3988762 https://doi.org/10.1214/18-AOS1753

[19] Han, Q. and Wellner, J.A. (2019). Convergence rates of least squares regression estimators with heavy-tailed errors. *Ann. Statist.* **47** 2286–2319. MR3953452 https://doi.org/10.1214/18-AOS1748

[20] Hanson, D.L., Pledger, G. and Wright, F.T. (1973). On consistency in monotonic regression. *Ann. Statist.* **1** 401–421. MR0353540

[21] Kahane, J.-P. (1969). Trois notes sur les ensembles parfaits linéaires. *Enseign. Math. (2)* **15** 185–192. MR0245734

[22] Kohler, M. and Langer, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *Ann. Statist.* **49** 2231–2249. MR4319248 https://doi.org/10.1214/20-aos2034

[23] Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. MR2329442 https://doi.org/10.1214/009053606000001019

[24] Koltchinskii, V. and Mendelson, S. (2015). Bounding the smallest singular value of a random matrix without concentration. *Int. Math. Res. Not. IMRN* **2015** 12991–13008. MR3431642 https://doi.org/10.1093/imrn/rnv096

[25] 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

[26] Kuchibhotla, A.K. and Patra, R.K. (2022). On least squares estimation under heteroscedastic and heavy-tailed errors. *Ann. Statist.* **50** 277–302. MR4382017 https://doi.org/10.1214/21-aos2105

[27] Kur, G., Gao, F., Guntuboyina, A. and Sen, B. (2020). Convex regression in multidimensions: suboptimality of least squares estimators. Preprint. Available at arXiv:2006.02044v1.

[28] Lecué, G. and Mendelson, S. (2017). Regularization and the small-ball method II: Complexity dependent error rates. *J. Mach. Learn. Res.* **18** Paper No. 146, 48 pp. MR3763780

[29] Lecué, G. and Mendelson, S. (2018). Regularization and the small-ball method I: Sparse recovery. *Ann. Statist.* **46** 611–641. MR3782379 https://doi.org/10.1214/17-AOS1562

[30] Mammen, E. and van de Geer, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413. MR1429931 https://doi.org/10.1214/aos/1034276635

[31] Mendelson, S. (2014). Learning without concentration. In *Proceedings of the 27th Conference on Learning Theory* (M.F. Balcan, V. Feldman and C. Szepesvári, eds.). *Proceedings of Machine Learning Research* **35** 25–39. Barcelona, Spain: PMLR.

[32] Micchelli, C. and Pontil, M. (2004). Kernels for multi-task learning. In *Advances in Neural Information Processing Systems* (L. Saul, Y. Weiss and L. Bottou, eds.) **17**. Cambridge, MA: MIT Press.

[33] Pathak, R., Ma, C. and Wainwright, M.J. (2022). A new similarity measure for covariate shift with applications to nonparametric regression. Preprint. Available at arXiv:2202.02837.

[34] Patschkowski, T. and Rohde, A. (2016). Adaptation to lowest density regions with application to support recovery. *Ann. Statist.* **44** 255–287. MR3449768 https://doi.org/10.1214/15-AOS1366

[35] Ray, K. and Schmidt-Hieber, J. (2017). A regularity class for the roots of nonnegative functions. *Ann. Mat. Pura Appl. (4)* **196** 2091–2103. MR3714756 https://doi.org/10.1007/s10231-017-0655-2

[36] 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

[37] Saumard, A. (2010). Convergence in sup-norm of least-squares estimators in regression with random design and nonparametric heteroscedastic noise. HAL Id: hal-00528539.

[38] Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Statist.* **48** 1875–1897. MR4134774 https://doi.org/10.1214/19-AOS1875

[39] Schmidt-Hieber, J. and Zamolodtchikov, P. (2024). Supplement to "Local convergence rates of the nonparametric least squares estimator with applications to transfer learning." https://doi.org/10.3150/23-BEJ1655SUPP

[40] 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

[41] Soloff, J.A., Guntuboyina, A. and Pitman, J. (2019). Distribution-free properties of isotonic regression. *Electron. J. Stat.* **13** 3243–3253. MR4010598 https://doi.org/10.1214/19-ejs1594

[42] Stone, C.J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360. MR0594650

[43] Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. and Kawanabe, M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems* (J. Platt, D. Koller, Y. Singer and S. Roweis, eds.) **20**. Curran Associates, Inc.

[44] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. New York: Springer. MR2724359 https://doi.org/10.1007/b13794

[45] van de Geer, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924. MR1056343 https://doi.org/10.1214/aos/1176347632

[46] van de Geer, S.A. (2000). *Applications of Empirical Process Theory*. *Cambridge Series in Statistical and Probabilistic Mathematics* **6**. Cambridge: Cambridge Univ. Press. https://doi.org/1739079

[47] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. *Springer Series in Statistics*. New York: Springer. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

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

[49] Wright, F.T. (1981). The asymptotic behavior of monotone regression estimates. *Ann. Statist.* **9** 443–448. MR0606630

Received August 2022 and revised April 2023