Smoothness Adaptive Hypothesis Transfer Learning

Haotian Lin [1] **Matthew Reimherr** [1]

Abstract

Many existing two-phase kernel-based hypothesis transfer learning algorithms employ the same kernel regularization across phases and rely on the known smoothness of functions to obtain optimality. Therefore, they fail to adapt to the varying and unknown smoothness between the target/source and their offset. This paper introduces *Smoothness Adaptive Transfer Learning* (SATL), a two-phase kernel ridge regression (KRR)-based algorithm to address these limitations. We first demonstrate that employing a misspecified fixed bandwidth Gaussian kernel in target-only KRR learning can achieve minimax optimality when the true function resides in Sobolev spaces. Leveraging this result, SATL enables the estimators to provably and universally adapt to the varying and unknown Sobolev smoothness of the source and offset functions. We derive the minimax lower bound of the learning problem in excess risk and show that SATL achieves a matching upper bound up to logarithmic factors. The optimal statistical rate reveals the factors influencing the transfer dynamics and efficacy, including the source sample size and the relative strength between domains. The theoretical findings and the effectiveness of SATL are confirmed by several experiments.

1. Introduction

Nonparametric regression is one of the most prevalent statistical problems studied in many communities in past decades due to its flexibility in modeling data. A large number of algorithms have been proposed, such as kernel regression, local regression, smoothing splines, and regression trees, to name only a few. However, the effectiveness of all the algorithms in these existing works is based on having sufficient samples drawn from the same target domain. When samples are scarce, either due to costs or other constraints, the performance of these algorithms can suffer empirically and theoretically.

Hypothesis Transfer Learning (HTL) (Li & Bilmes, 2007; Kuzborskij & Orabona, 2013; Du et al., 2017), which leverages models trained on the source domain and uses samples from the target domain to learn the model shift to the target model, is an appealing and promising mechanism. When the parameters of interest are infinite-dimensional (e.g., nonparametric models), Lin & Reimherr (2024) employed the reproducing kernel Hilbert space (RKHS) norm as a metric for assessing similarity in functional regression frameworks, linking the transferred knowledge to the employed RKHS structure. They leveraged offset transfer learning (OTL), which is one of the most popular HTL algorithms, to obtain target estimators in a two-phase manner: first, training a source model on the large sample size source dataset, then estimating the offset model between the target and source using the target dataset and trained source model. However, a noteworthy observation is that both phases of estimating the source model and offset model utilize the same RKHS regularization. This goes against the principle that led to OTL's success, which posits that the offset should ideally have a simpler structure than the target and source. A similar limitation also appears in a series of two-phase HTL algorithms for finite-dimensional models (e.g., multivariate/high-dimensional linear regression) (Bastani, 2021; Li et al., 2022; Tian & Feng, 2022), which typically utilize the ℓ^1 or ℓ^2 norm of the offset parameters as the similarity measure. Since ℓ^1-norm can reflect sparsity and ℓ^2 usually serves to control complexity, using the same norm regularization across both phases is more robust to model structure heterogeneous and more defensible in finite-dimensional models than in infinite-dimensional ones.

In the realm of nonparametric regression, although OTL has shown great success in practice, there are only a few studies that provide theoretical analysis (Wang & Schneider, 2015; Du et al., 2017), and these works are still limited in terms of problem settings, estimation procedures, and theoretical bounds. For example, although Wang & Schneider (2015) noticed the nature of simple offsets, they didn't use any quantity (like Sobolev or Hölder smoothness) to formularize the difference of target/source models and their offset. Their KRR-based OTL algorithm also employed the same kernel

[1] Department of Statistics, Pennsylvania State University, University Park, PA, USA. Correspondence to: Haotian Lin <hzl435@psu.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

to train the source model and the offset and thus has similar limitations as Lin & Reimherr (2024) methodologically. Du et al. (2017) formalized the varying structures via different Hölder smoothness, but the theoretical results derived are under too ideal assumptions, unverifiable in practice. Besides, neither their approaches nor the statistical convergence rates were adaptive, and their upper bound overlooked certain factors, failing to provide deeper insights into the influence of domain properties on the transfer learning dynamics and efficacy. This raises the following fundamental question that motivates our study:

Can we develop an HTL algorithm so that the different structures (smoothness) of the target/source functions and their offset can be adaptively learned?

Main contributions. This work answers the above question positively and makes the following contributions:

We propose *Smoothness Adaptive Transfer Learning* (SATL), building upon the prevalent two-phase offset transfer learning paradigm. Specifically, we study the setting where the target/source function lies in Sobolev space with order m_0 while the offset function lies in Sobolev space with order m (where $m > m_0$). One key feature of SATL is its universal capability to adapt to the unknown and varying smoothness of the target, source, and offset functions.

We first begin by establishing the robustness of the Gaussian kernel in misspecified KRR, i.e., for regression functions belonging to certain fractional Sobolev spaces (or RKHSs that are norm equivalent to such Sobolev spaces), employing a fixed bandwidth Gaussian kernel in target-only KRR yields minimax optimal generalization error. Remarkably, the optimal order of the regularization parameters follows an exponential pattern, which differs from the variable bandwidth setting and we conduct comprehensive experiments to support the finding. Furthermore, we demonstrate that an estimator, developed through standard training and validation methods, achieves the same optimality up to a logarithmic factor without prior knowledge of the true smoothness.

Leveraging these new results of the Gaussian kernel, SATL employs Gaussian kernels in both learning phases, avoiding the saturation effect of KRR and thus ensuring its universal and consistent adaptability to the diverse and unknown smoothness levels m_0 and m. We also establish the minimax statistical lower bound for the learning problem in terms of excess risk and show that SATL achieves minimax optimality since it enjoys a matching upper bound (up to logarithmic factors). Crucially, our results shed light on the impact of signal strength from both domains on the dynamic and efficacy of OTL, which, to the best of our knowledge, has been largely overlooked in the existing literature. This insight enhances our understanding of the contributions of each phase in the transfer learning process.

1.1. Related Literature

Transfer Learning. OTL (a.k.a. bias regularization transfer learning) has been extensively researched in supervised regression. The work in Kuzborskij & Orabona (2013; 2017) focused on OTL in linear regression, establishing generalization bounds through Rademacher complexity. Wang & Schneider (2015) derived generalization bounds for applying KRR on OTL without formularizing the simple offset structure. Wang et al. (2016) assumed target/source regression functions in the Sobolev ellipsoid with order m_0 and the offset in a smoother power Sobolev ellipsoid. They used finite orthonormal basis functions for modeling, which becomes restrictive if the chosen basis is misaligned with the eigenfunctions of the Sobolev ellipsoid. Du et al. (2017) further proposed a transformation function for the offset, thereby integrating many preview OTL studies and offering upper bounds on excess risk for both kernel smoothing and KRR. Apart from regression settings, generalization bounds for classification problems with surrogate losses have been studied in Aghbalou & Staerman (2023) via stability analysis techniques. Other results that study HTL outside OTL can be found in Orabona et al. (2009); Cheng & Shang (2015); Minami et al. (2024). Besides, OTL can also be viewed as a case of representation learning Du et al. (2020); Tripuraneni et al. (2020); Xu & Tewari (2021) by viewing the trained source model as a representation for target tasks.

The idea of OTL has also been recently adopted by the statistics community, which typically involves regularizing the offset via different metrics in parameter spaces. For example, Bastani (2021); Li et al. (2022); Tian & Feng (2022) considered ℓ^1-distance OTL for high-dimensional (generalized) linear regression. Duan & Wang (2022); Tian et al. (2023) considered ℓ^2-distance for general linear models. Lin & Reimherr (2024) utilized RKHS-distance for functional linear models. However, all of these works used the same type of distance while estimating the source model and the offset. For nonparametric regression, another study by Cai & Pu (2024) assumed the target/source models lie in Hölder spaces while the offset can be approximated with any desired accuracy by a polynomial function in L_1. They proposed an algorithm based on local polynomial regression that adapts to Hölder smoothness, but the approach can be computationally intensive in practice. KRR under the covariate shift setting has also been studied in several works. Ma et al. (2022) derived optimal rates of the generalization error under different likelihood ratio bound conditions and proposed rate-optimal estimator based on reweighting KRR. Wang (2023) introduced a pseudo-labeling algorithm to address TL scenarios where the labels in the target domain are unobserved. For nonparametric classification, Kpotufe & Martinet (2021); Cai & Wei (2021); Reeve et al. (2021) developed adaptive classifiers based on K-nearest neighbors that are rate-optimal in different distribution shift settings.

Misspecification in KRR. This line of research focuses on using misspecified kernels in target-only KRR to achieve optimal statistical convergence rates.

In the realm of variable bandwidths, Eberts & Steinwart (2013) derived the convergence rates of the excess risk for KRR using Gaussian kernels when the true regression function lies in a Sobolev space. They found that appropriate choices of regularization parameters and the bandwidth will yield a non-adaptive rate that can be arbitrarily close (but not equal) to the optimal rate under the bounded response assumption. Building upon this work, Hamm & Steinwart (2021) further improved the non-adaptive rate, attaining optimal rates up to logarithmic factors. It is worth noting that their results show that both the optimal regularization parameter and bandwidth should decay in polynomial patterns, which is different from ours. Apart from regression setting, Li & Yuan (2019) studied using variable bandwidth Gaussian kernels to achieve optimality in a series of nonparametric statistical tests.

Another line of research considers fixed bandwidth kernels. For instance, Wang & Jing (2022) investigated the misspecification of Matérn kernel-based KRR. They demonstrated that even when the true regression functions belong to a Sobolev space, utilizing misspecified Matérn kernels can still attain minimax optimal convergence rates or, in some cases, a slower convergence rate (referred to as the saturation effect of KRR). Similarly, several other works have presented similar results on general RKHS with polynomial eigen-decay rate, and the true function resides in the power space of the RKHS, see Steinwart et al. (2009); Dicker et al. (2017); Blanchard & Mücke (2018); Fischer & Steinwart (2020); Lin & Cevher (2020); Zhang et al. (2023) and more references therein.

2. Preliminaries

Problem Formulation. Consider the two nonparametric regression models

$$y_{p,i} = f_p(x_{p,i}) + \epsilon_{p,i}, \quad p \in \{T, S\}$$

where p is the task index (T for target and S for source), f_p are unknown regression functions, $x_{p,i} \in \mathcal{X} \subset \mathbb{R}^d$ is a compact set with positive Lebesgure measure and Lipschitz boundary, and $\epsilon_{p,i}$ are i.i.d. random noise with zero mean. The target and source regression function, f_T and f_S, belong to the (fractional) Sobolev space H^{m_0} with smoothness $m_0 \geq d/2$ over $\mathcal{X}$. The joint probability distribution $\rho_p(x,y)$ is defined on $\mathcal{X} \times \mathcal{Y}$ for the data points $\{(x_{p,i}, y_{p,i})\}_{i=1}^{n_p}$, and μ_p represents the marginal distribution of ρ_p on $\mathcal{X}$. In this work, we assume the model shift (a.k.a. posterior drift) setting, where μ_T is equal to μ_S, while the regression function f_T differs from f_S. The goal of this paper is to find a function $\hat{f}_T$ based on the combined data $\{(x_{T,i}, y_{T,i})\}_{i=1}^{n_T} \cup \{(x_{S,i}, y_{S,i})\}_{i=1}^{n_S}$ that minimizes the generalization error on the target domain, i.e.

$$\mathcal{E}(\hat{f}_T) = \mathrm{E}_{x\sim\mu_T}[(\hat{f}_T(x) - f_T(x))^2].$$

Non-Transfer Scenario. In the absence of source data, recovering f_T using KRR is referred to as target-only learning and has been extensively studied. We now state some of its well-known results.

Proposition 2.1 (Target-only Learning)**.** *For a symmetric and positive semi-definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, let $\mathcal{H}_K$ be the RKHS associated with K (Wendland, 2004). The KRR estimator is*

$$\hat{f}_T = \underset{f\in\mathcal{H}_K}{\operatorname{argmin}} \left\{ \frac{1}{n_T} \sum_{i=1}^{n_T} (y_{T,i} - f(x_{T,i}))^2 + \lambda \|f\|_{\mathcal{H}_K}^2 \right\},$$

and we call the kernel K as the imposed kernel. Then, the convergence rate of the generalization error of $\hat{f}_T$, $\mathcal{E}(\hat{f}_T)$, is given as follows.

1. *(Well-specified Kernel) If $\mathcal{H}_K$ coincides with H^{m_0}, $\mathcal{E}(\hat{f}_T)$ can reach the standard minimax convergence rate in high-probability given $\lambda \asymp n^{-\frac{2m_0}{2m_0+d}}$, i.e.*

$$\mathcal{E}(\hat{f}_T) = O_{\mathbb{P}}\left(n_T^{-\frac{2m_0}{2m_0+d}} \right).$$

2. *(Misspecified Kernel) If the K is the Matérn kernel then its induced space is isomorphic to $H^{m_0'}$ with $m_0' > \frac{d}{2}$. Furthermore, given $\lambda \asymp n^{-\frac{2m_0'}{2m_0+d}}$ and $\gamma = \min\{2, \frac{m_0}{m_0'}\}$, then*

$$\mathcal{E}(\hat{f}_T) = O_{\mathbb{P}}\left(n_T^{-\frac{2\gamma m_0'}{2\gamma m_0'+d}} \right).$$

3. *(Saturation Effect) For $m_0' < \frac{m_0}{2}$ and any choice of parameter $\lambda(n_T)$ satisfying that $n_T \to \infty$, we have*

$$\mathcal{E}(\hat{f}_T) = \Omega_{\mathbb{P}}\left(n_T^{-\frac{4m_0'}{4m_0'+d}} \right).$$

The well-specified result is well-known and can be found in a line of past work (Geer, 2000; Caponnetto & De Vito, 2007). The misspecified kernel result comes from a combination (with a modification) of Theorem 15 and 16 in Wang & Jing (2022). The saturation effect is proved by Li et al. (2023). The Proposition 2.1 indicates that for target-only KRR, even when the smoothness of the imposed RKHS, m_0', disagrees with the smoothness m_0 of the Sobolev space to which f_T belongs, the optimal rate of convergence is still achievable if $m_0' \geq m_0/2$ with the λ appropriately chosen.

However, if $m_0' < m_0/2$, i.e., the true function is much smoother than the estimator itself, the saturation effect occurs, meaning that the information-theoretic lower bound $n_T^{-2m_0/(2m_0+d)}$ seemingly cannot be achieved regardless of the selection of the regularization parameters in KRR (Bauer et al., 2007; Gao et al., 2008).

Transfer Learning Framework. We introduce the KRR-based version of OTL for nonparametric regression, which serves as the backbone of our proposed algorithm. Formally, OTL obtains the estimator for f_T as $\hat{f}_T = \hat{f}_S + \hat{f}_\delta$ via two phases. In the first phase, it obtains $\hat{f}_S$ by KRR with the source dataset $\{(x_{S,i}, y_{S,i})\}_{i=1}^{n_S}$. In the second phase, it generates pseudo offset labels $\{y_{T,i} - \hat{f}_S(x_{T,i})\}_{i=1}^{n_T}$ and then learns the $\hat{f}_\delta$ via KRR by replacing target labels by pseudo offset labels. The main idea of OTL is that the f_S can be learned well given sufficiently large source samples, and the offset f_δ can be learned with much fewer target samples. We formulate the OTL variant of KRR as Algorithm 1.

Algorithm 1 OTL-KRR

```
Input: Target and source training data $\{(x_{T,i}, y_{T,i})\}_{i=1}^{n_T} \cup \{(x_{S,i}, y_{S,i})\}_{i=1}^{n_S}\}$; Self-specified KRR imposed kernel $K$
Output: Target function estimator $\hat{f}_T = \hat{f}_S + \hat{f}_\delta$.
Phase 1:
    $\hat{f}_S = \operatorname*{argmin}_{f \in \mathcal{H}_K} \frac{1}{n_S} \sum_{i=1}^{n_S} (y_{S,i} - f(x_{S,i}))^2 + \lambda_1 \|f\|_{\mathcal{H}_K}^2$
Phase 2:
    $\hat{f}_\delta = \operatorname*{argmin}_{f \in \mathcal{H}_K} \frac{1}{n_T} \sum_{i=1}^{n_T} (y_{T,i} - \hat{f}_S(x_{T,i}) - f(x_{T,i}))^2 + \lambda_2 \|f\|_{\mathcal{H}_K}^2$
```

Model Assumptions. We first state the smoothness assumption on the offset function $f_\delta := f_T - f_S$.

The learning framework (Algorithm 1) reveals a smoothness-agnostic nature: the imposed kernels K (also the associated RKHSs) stay the same regardless of the level of smoothness of f_S and f_δ. More specifically, based on Proposition 2.1, the learning algorithm is rate-optimal when the smoothness of both imposed RKHSs $\mathcal{H}_K$ in both steps matches that of f_S, and f_δ, i.e. the smoothness of f_S and f_δ stay the same. However, in the transfer learning context, such a smoothness condition on the offset function may not be precise enough. One should rather consider the offset function smoother than the target/source functions themselves to represent the similarity between f_S and f_T.

To illustrate this point, consider the following example. Suppose $f_T = f_S + f_\delta$ where f_S is a complex function with low smoothness (less regularized) while f_δ is rather simple (well regularized), e.g. a linear function. Then f_S can be estimated well via larger n_S while f_δ is a highly smooth function and can also be estimated well via small n_T due to its simplicity. In this example, the effectiveness of OTL relies on the similarity between f_T and f_S, i.e., the offset f_δ possessing a "simpler" structure than the target and source models. Such "simpler" offset assumptions have been proven to make OTL effective in other models, e.g., high-dimensional linear regression works (Bastani, 2021; Li et al., 2022; Tian & Feng, 2022) assume the offset coefficient should be sparser than target/source coefficients. This motivates our endeavor to introduce the following smoothness assumptions to quantify the similarity between target and source domains.

Assumption 2.2 (Smoothness of Target/Source)**.** There exists an $m_0 \geq d/2$ such that f_T and f_S belong to H^{m_0}.

Assumption 2.3 (Smoothness of Offset)**.** There exists an $m \geq m_0$ such that $f_\delta := f_T - f_S$ belongs to H^m.

Remark 2.4. The results of this paper are applicable not only to Sobolev spaces but also to those general RKHSs that are norm equivalent to Sobolev spaces. Thus, we can assume that f_S, f_T, and f_δ belong to RKHSs with different regularity. Due to norm equivalency, our discussion is primarily focused on Sobolev spaces, and we refer readers to Appendix B.2 for the analysis pertaining to general RKHSs.

Assumption 2.2 is a very common assumption in nonparametric regression literature, and Assumption 2.3 naturally holds if Assumption 2.2 is satisfied. Compared to the offset assumption in Wang et al. (2016) where f_δ is assumed to belong to the power space of H^{m_0}, our setting presents a unique challenge. Since we consider the offset function in a Sobolev space with higher smoothness, which doesn't necessarily share the same eigenfunctions with H^{m_0}, this renders orthonormal basis modeling less promising. Assumption 2.3 also makes our setting conceptually align with contemporary transfer learning models. For instance, in prevalent pretraining-finetuning neural networks, the pretrained feature extractor tends to encompass a greater number of layers, while the newly added fine-tuning structure typically involves only a few layers. In this analogy, m_0 and m in our setting are akin to the deeper pre-trained layers and the shallow fine-tuned layers.

We also state a standard assumption that frequently appears in KRR literature (Fischer & Steinwart, 2020; Zhang et al., 2023) to establish theoretical results, which controls the noise tail probability decay speed.

Assumption 2.5 (Moment of error)**.** There are constants σ,$L > 0$ such that for any $r \geq 2$, the noise, ϵ, satisfies

$$\mathrm{E}\left[|\epsilon_p|^r \mid x\right] \leq \frac{1}{2} r! \sigma^2 L^{r-2}, \quad \text{for} \quad p \in \{T, S\}.$$

3. Target-Only KRR with Gaussian Kernels

3.1. Motivation for Employing Gaussian Kernel

To achieve optimality in Algorithm 1 under the smoothness assumptions, an applicable approach is to employ distinct kernels that can accurately capture the correct smoothness of f_S and f_δ during both phases. This approach, however, faces the practical challenge of identifying the unknown smoothness m_0 and m, which, in turn, induce different kernels and RKHS; this is an issue also prevalent in the target-only KRR context.

One potential solution is to leverage the robustness of Matérn kernels, i.e., employ a misspecified Matérn kernel as the imposed kernel in KRR. As indicated by Proposition 2.1, the optimal convergence rate is still attainable for some appropriately chosen Matérn kernels and regularization parameters. Nonetheless, this still faces two problems:

(1) The rate with misspecified Matérn kernel in Proposition 2.1 is still non-adaptive, i.e. one still needs to know the true smoothness when tuning λ_1 and λ_2.

(2) The risk of the saturation effect of KRR happening in both phases when a less smooth kernel is chosen.

While the former can be potentially addressed by cross-validation or data-driven adaptive approach, the second one is more fatal as one might end up choosing a less smooth kernel and never be able to achieve the information-theoretic lower bounds because of the saturation effect.

Hence, there is a clear demand for a kernel with a more general and robust property, i.e., in the target-only KRR, for the regression function that lies in H^{m_0} with any $m_0 \geq d/2$, employing such a kernel ensures that there's always an optimal λ such that the optimal convergence rate is achievable. Motivated by the fact that the Gaussian kernel is the limit of Matérn kernel K_ν as $\nu \to \infty$ and the RKHS associated with the Gaussian kernel is contained in the Sobolev space H^ν for any $\nu > d/2$ (Fasshauer & Ye, 2011), we show that the Gaussian kernel indeed possesses this desired property.

3.2. Theoretical Results

Consider the Target-Only learning KRR setting, where $f_0 \in H^{\alpha_0}$ (we use α_0 to denote smoothness in target-only context to distinguish from TL context) and the underlying supervised learning model setup as

$$y_i = f_0(x_i) + \epsilon_i, \quad i = 1, \cdots, n.$$

First, we show the non-adaptive rate for the Gaussian kernel.

Theorem 3.1 (Non-Adaptive Rate). *Under the Assumptions 2.5, let the imposed kernel, K, be the Gaussian kernel with fixed bandwidth and $\hat{f}$ be the corresponding KRR estimator based on data $\{(x_i, y_i)\}_{i=1}^n$. If $f_0 \in H^{\alpha_0}$, by choosing $log(1/\lambda) \asymp n^{\frac{2}{2\alpha_0+d}}$, for any $\delta \in (0,1)$, when n is sufficient large, with probability at least $1-\delta$, we have*

$$\|\hat{f} - f_0\|_{L_2}^2 \leq C \left(\log \frac{4}{\delta}\right)^2 n^{-\frac{2\alpha_0}{2\alpha_0+d}},$$

where C is a constant independent of n and δ.

Remark 3.2. Although Eberts & Steinwart (2013); Hamm & Steinwart (2021) has studied the robustness of Gaussian kernel on misspecified KRR, their results are built on variable bandwidths, and the nearly rate-optimal results are established given both bandwidth and λ decay polynomially in n. In contrast, our result is built on fixed bandwidth Gaussian kernels and achieves the optimal rate with the optimal λ that behaves differently from theirs.

We note to the reader that while the RKHS associated with the Matérn kernel coincides with a Sobolev space (i.e., they are the same space with slightly different, though equivalent, norms), the Gaussian kernel does not, making the behavior of the optimal λ totally different compared to the misspecified Matérn kernel scenarios in Proposition 2.1. Particularly, even if the Gaussian kernel is the limit of Matérn kernel K_ν as $\nu \to \infty$, setting the smoothness parameter m_0' of the imposed Matérn kernel as infinity in misspecified kernel case of Proposition 2.1 will never yield analytical results but only tells the optimal order of λ should converge to 0 faster than polynomial ($\lim_{m_0' \to \infty} n^{-2m_0'/(2m_0+d)} = 0$). On the other side, our result identifies that λ should converge to 0 exponentially in n.

The exponential decay form for the optimal λ originates from managing the approximation error. The standard real interpolation technique (like Proposition 2.1) is inadequate for controlling this error when the intermediate term lies in RKHS of Gaussian kernels. We address this by the Fourier transform of the RKHS, which reveals this exponential form. We refer readers to Appendix C.1 for more details. To further highlight our findings, we compare our results with existing state-of-the-art works on misspecified KRR and refer readers to Appendix C.4 for a detailed discussion.

To develop an adaptive procedure without known α_0, we employ a standard training and validation approach (Steinwart & Christmann, 2008). To this end, we construct a finite set that is an arithmetic sequence, i.e., $\mathcal{A} = \{\alpha_{\min} < \cdots < \alpha_{\max}\}$ where $\{\alpha_i\}$ satisfy $\alpha_{\min} > d/2$, $\alpha_{\max}$ large enough such that $\alpha_0 \leq \alpha_{\max}$ and $\alpha_i - \alpha_{i-1} \asymp 1/\log n$. Split dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ into

$$\mathcal{D}_1 := \{(x_1, y_1), \cdots, (x_j, y_j)\}$$
$$\mathcal{D}_2 := \{(x_{j+1}, y_{j+1}), \cdots, (x_n, y_n)\}$$

The adaptive estimator is obtained by following the training and validation approach.

1. For each $\alpha \in \mathcal{A}$, obtain non-adaptive estimator $\hat{f}_{\lambda_\alpha}$ by KRR with dataset $\mathcal{D}_1$.

2. Obtain the adaptive estimator $\hat{f}_{\lambda_{\hat\alpha}}$ by minimizing empirical L_2 error on $\mathcal{D}_2$, i.e.

$$\hat{f}_{\lambda_{\hat\alpha}} = \operatorname*{argmin}_{\alpha \in \mathcal{A}} \left\{ \frac{1}{n-j} \sum_{i=j+1}^{n} (y_i - \hat{f}_{\lambda_\alpha}(x_i))^2 \right\}.$$

When constructing the collection of non-adaptive estimators over $\mathcal{A}$, Theorem 3.1 suggests choosing the regularization parameter $\lambda = \exp\{-Cn^{2/2\alpha+d}\}$ for some constant C.

The following theorem shows the estimator from the training and validation approach achieves an optimal minimax rate up to a logarithmic factor in n.

Theorem 3.3 (Adaptive Rate)**.** *Under the same conditions of Theorem 3.1 and $\mathcal{A} = \{\alpha_1, \cdots, \alpha_N\}$ with $\alpha_j - \alpha_{j-1} \asymp 1/\log n$. Then, for $\delta \in (0,1)$, when n is sufficient large, with probability $1-\delta$, we have*

$$\mathcal{E}(\hat{f}_{\lambda_{\hat\alpha}}) \leq C \left(\log \frac{4}{\delta}\right)^2 \left(\frac{n}{\log n}\right)^{-\frac{2\alpha_0}{2\alpha_0+d}},$$

where C is a constant independent of n and δ.

Remark 3.4. If the marginal distribution of x, μ, is known, one can also apply the well-known Lepski's method (Lepskii, 1991) to obtain an adaptive estimator without known α_0, which also achieves optimal nonadaptive rate up to a logarithmic factor as training and validation approach does.

4. Smoothness Adaptive Transfer Learning

We formally propose Smoothness Adaptive Transfer Learning in Algorithm 2.

Algorithm 2 *S̲moothness A̲daptive T̲ransfer L̲earning* (SATL)

```
Input: Target and source dataset $\mathcal{D}_T$ and $\mathcal{D}_S$;
1: Let the smoothness candidate set for the source model $f_S$ as $\mathcal{M}_S = \{\frac{Q_1}{\log(n_S)}, \cdots, \frac{Q_1 N_1}{\log(n_S)}\}$ and the candidate set for the offset model $f_\delta$ as $\mathcal{M}_\delta = \{\frac{Q_2}{\log(n_T)}, \cdots, \frac{Q_2 N_2}{\log(n_T)}\}$ for some fixed positive number $Q_1, Q_2$ and integer $N_1, N_2$.
2: Obtain the adaptive source model $\hat{f}_S$ via the training and validation with the Gaussian kernel and $\mathcal{M}_S$.
3: Generate the label $\hat{e}_{T,i} = y_{T,i} - \hat{f}_S(x_{T,i})$ and the offset dataset as $\tilde{\mathcal{D}}_T = \{(x_{T,1}, \hat{e}_{T,1}), \cdots, (x_{T,n_T}, \hat{e}_{T,n_T}))\}$.
4: Using the offset datasets $\tilde{\mathcal{D}}_T$ to obtain the adaptive offset model $\hat{f}_\delta$ via the training and validation with the Gaussian kernel and $\mathcal{M}_\delta$.
```

While SATL can be viewed as a specification of the Algorithm 1, the desirable property exhibited by the Gaussian kernel surpasses all other misspecified kernel choices by allowing estimators from both phases always to be able to adapt to the true Sobolev smoothness of the functions inherently even with unknown the true values of m_0, m.

4.1. Theoretical Analysis

In order to provide concrete theoretical bounds, we assume the offset function of f_T and f_S in the h-ball of H^m, i.e. f_S is said to be h-transferable to f_T if $\|f_\delta\|_{H^m} \leq h$. Hence, the parameter space is defined as

$$\Theta(h, R, m_0, m) = \{(\rho_T, \rho_S) : \\ \|f_S\|_{H^{m_0}} \leq R, \|f_\delta\|_{H^m} \leq h\}$$

for some positive constants R and h. We note that to achieve rigorous optimality in the context of transfer learning under the regression setting, such an upper bound for the distance between parameters from both domains is often required, e.g., ℓ^1 or ℓ^0 distance in high-dimensional setting (Li et al., 2022; Tian & Feng, 2022; He et al., 2024), Fisher-Rao distance in low-dimensional setting (Zhang et al., 2022), RKHS distance in functional setting (Lin & Reimherr, 2024), etc.

Theorem 4.1 (Optimality of SATL)**.** *Suppose the Assumption 2.2, 2.3, and 2.5 hold, and n_S and n_T are sufficiently large but still in transfer learning regime ($n_S \gg n_T$), we have the lower bound for the transfer learning problem and the upper bound of SATL as follows.*

1. *(**Lower bound**) For $\delta \in (0,1)$, with probability $1-\delta$*

$$\inf_{\tilde{f}} \sup_{\Theta(h,R,m_0,m)} \mathbb{P}\left\{\|\tilde{f} - f_T\|_{L_2}^2 \geq C\delta R^2 \left(n_S^{-\frac{2m_0}{2m_0+d}} + n_T^{-\frac{2m}{2m+d}}\xi_L\right)\right\} \geq 1-\delta,$$

   where $\xi_L \propto h^2/R^2$ and C is a constant independent of n_S, n_T, R, h, and δ. The inf *is taken over all possible estimators $\tilde{f}$ based on the target and source data.*

2. *(**Upper bound**) Suppose that $\hat{f}_T$ is the output of SATL. For $\delta \in (0,1)$, with probability $1-\delta$, we have*

$$\|\hat{f}_T - f_T\|_{L_2}^2 \leq C\left(\log\frac{8}{\delta}\right)^2 (R^2 + \sigma_S^2) \left\{\left(\frac{n_S}{\log n_S}\right)^{-\frac{2m_0}{2m_0+d}} + \left(\frac{n_T}{\log n_T}\right)^{-\frac{2m}{2m+d}}\xi_U\right\},$$

   where $\xi_U \propto (h^2+\sigma_T^2)/(R^2+\sigma_S^2)$ and C is a constant independent of n_S, n_T, R, h, and δ.

Theorem 4.1 indicates that the convergence rate of excess risk for SATL consists of two terms: the first term is the rough estimation error, and the second is the offset estimation error. The first term represents the error of learning f_T with the source samples, while the second term is the error of learning the offset function with the target samples. The terms ξ_L and ξ_U control the transfer dynamic and efficacy; see discussion on Section 4.2, and we refer readers to Appendix D for their origin. Besides, the upper bound is tight up to logarithmic factors, which is a price paid for adaptivity. Note that the upper bound can be exactly tight when the Sobolev smoothness m_0 and m are known.

Finally, it is also worth highlighting how the refinement of a "simple" offset provides a better convergence rate compared to homogeneous kernel regularization (Lin & Reimherr, 2024). Based on the saturation effort of KRR, using the same Matérn kernel with smoothness m_0 for both phases in Algoritm 1 will lead the offset estimation error to be $n_T^{-2\gamma m_0/(2\gamma m_0+d)}$, where $\gamma = \min\{2, m/m_0\}$, which is never faster than $n_T^{-2m/(2m+d)}$.

4.2. OTL Transfer Dynamic and Efficacy

In this part, we discuss some insights provided by our upper bound into the transfer dynamic, i.e., how OTL benefits the learning over the target domain, and the transfer efficacy, i.e., whether the OTL is taking effect.

OTL Dynamic. For offset and source models, we term the quantities $h^2 + \sigma_T^2$ and $R^2 + \sigma_S^2$ as the model total strength, i.e., the sum of upper signal strength bound and noise variance. Then, ξ_U quantifies the relative model total strength between the offset and source models. In comparison to the convergence rate of the target-only baseline estimator, $n_T^{-2m_0/(2m_0+d)}$, our results indicate that the transfer learning dynamic depends jointly on the sample size in source domain n_S, and the constant ξ_U. Specifically, when the relative model total strength between offset and source is small, the rough estimation error predominates, and thus, the statistical convergence rate of $\hat{f}_T$ is much faster than the target-only baseline given $n_S \gg n_T$. Conversely, a large ξ_U will make the offset estimation error the dominant term, but the statistical rate keeps the same as the target-only baseline up to a constant.

OTL Efficacy. In earlier theoretical works, Wang et al. (2016); Du et al. (2017) failed to identify the presence of ξ_U in their statistical rates. The bounds in some recent works, e.g., Li et al. (2022); Tian & Feng (2022), only identified $\xi_U \propto h^2$, i.e., the corresponding upper bound should be

$$O_{\mathbb{P}}\left(\left(\frac{n_S}{\log n_S}\right)^{-\frac{2m_0}{2m_0+d}} + \left(\frac{n_T}{\log n_T}\right)^{-\frac{2m}{2m+d}} h^2\right). \quad (1)$$

This bound claimed that the OTL takes effect when the magnitude of h is small while disregarding the influence of the source model's total strength. Conversely, our results reveal a new perspective: the transfer efficacy within the OTL framework jointly depends on the properties of offset and source models. This means that even with the same offset model, whether OTL takes effect can differ given different source models. Thus, the constant ξ_U can be viewed as a similarity measure between source and target domains under the OTL framework.

We further illustrate how our results provide a more accurate interpretation of the OTL efficacy compared to the form (1) via the following example. In Figure 1, we construct two source-target pairs termed as (S^0, T^0) and (S^1, T^1) with identical h. While with identical h, the two pairs possess difference angle θ_0 and θ_1 given different source model's total strength, thus implying the different degree of similarity between source and target domains. Geometrically, one can interpret ξ_U as a factor that approximately represents the angle θ_0 and θ_1 between source and target domains. While the form (1) suggests two pairs have the same OTL efficacy, our upper bound indicates the set (S^0, T^0) has higher efficacy, which aligns with the fact that (S^0, T^0) possesses higher similarity.

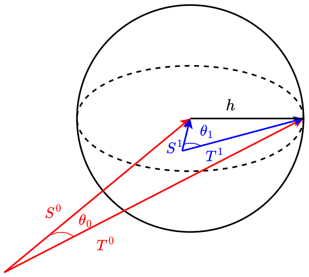


Figure 1. Geometric illustration for how ξ_U will affect the OTL. The circle represents a ball centered around the source with radius h. The length of the red and blue lines represents the magnitude of the model's total strength. Two source-target pairs (denoted by red and blue) possess the same offset while the source models' total strength is different, leading to different angles θ_0 and θ_1 between domains.

Remark 4.2. It should be noted that the above geometric interpretation of ξ_U is somewhat rough due to the use of the upper bound of $\|f_S\|_{H^{m_0}}$ and $\|f_\delta\|_{H^m}$. As a result, ξ_U cannot precisely reflect the exact angle between f_T and f_S but only the angle between domains as we termed above. However, with some additional assumptions, one can obtain a fine-grained angle interpretation. For example, if one uses the Sobolev norm of f_S and f_δ directly and assumes the signal-to-noise ratio of source and offset models are bounded below, then $\xi_U \propto \|f_\delta\|^2_{H^m}/\|f_S\|^2_{H^{m_0}}$ and thereby is able to reflect the exact angle between f_T and f_S.

5. Experiments

This section aims to confirm our theoretical results in the target-only KRR and transfer learning sections. [1]

5.1. Experiments for Target-Only KRR

Let $\mathcal{X} = [0, 1]$ and the marginal distribution of x be the uniform distribution over $[0, 1]$. Our objective is to empirically confirm the adaptability of Gaussian kernels in target-only KRR when $f_0 \in H^\alpha([0, 1])$ for different smoothness α. Specifically, we explore cases where f_0 belongs to H^2 and H^3. To generate such f_0 with the desired Sobolev smoothness, we set f_0 to be the sample path that is generated from the Gaussian process with isotropic Matérn covariance kernels K_ν (Stein, 1999). We set $\nu = 2.01$ and 3.01 to generate the corresponding f_0 with smoothness 2 and 3, see Corollary 4.15 in Kanagawa et al. (2018) for detail discussion about the connection between ν and α. Formally, we consider the following data generation procedure: $y_i = f_0(x_i) + \sigma\epsilon_i$, where ϵ_i are i.i.d. standard Gaussian noise, $\{x_i\}_{i=1}^n \overset{i.i.d.}{\sim} U([0, 1])$ and $\sigma = 0.5$.

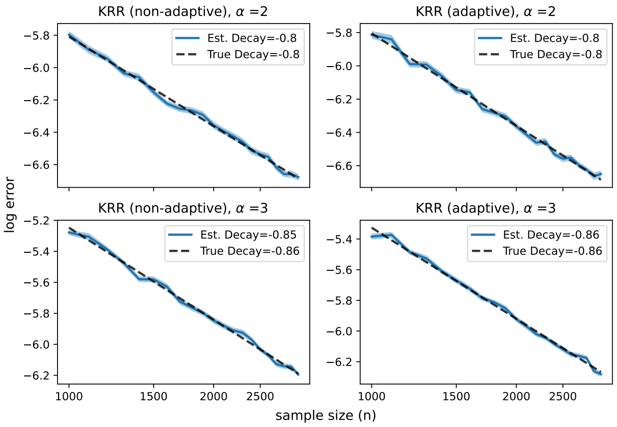


Figure 2. Error decay curves of target-only KRR based on Gaussian kernel, both axes are in log scale. The blue curves denote the average generalization errors over 100 trials. The dashed black lines denote the theoretical decay rates.

We verify both the nonadaptive and adaptive rate presented in Theorem 3.1 and 3.3. The sample size ranges from 1000 to 3000 in intervals of 100. For different α, we set $\lambda = \exp\{-Cn^{\frac{2}{2\alpha+1}}\}$ with a fixed C. To evaluate the adaptivity rate, we set the candidate smoothness as $[1, 2, 3, 4, 5]$ and split the dataset equally in size to implement training and validation. The generalization error $\|\hat{f} - f\|_{L_2}$ is obtained by Simpson's rule. For each combination of n and α, we repeat the experiments 100 times and report the average generalization error. To demonstrate the convergence rate of the error is sharp, we regress the logarithmic average generalization error, i.e. $\log(\|\hat{f} - f\|_{L_2})$, on $\log(n)$ and compare the regression coefficient to its theoretical counterpart $-\frac{2\alpha}{2\alpha+1}$.

We try different values of C lies in the equally spaced sequence $[0.05, 0.1, \cdots, 4]$, and report the optimal curve in Figure 2 under the best choice of C. Remarkably, the empirical data points align closely with the theoretical lines for both nonadaptive and adaptive rates. The estimated regression coefficients also closely agree with the theoretical counterparts. Additionally, we also report the generalization error decay estimation results for other values of C and refer to Appedix E for more details.

5.2. Experiments for Transfer Learning

We now illustrate our theoretical analysis of SATL through two experiments with synthetic data. We generate the target/source functions and the offset function as follows: (i) The target function f_T is a sample path of the Gaussian process with Matérn kernel $K_{1.01}$ such that $f_T \in H^1$; (ii) The offset function f_δ is a sample path of Gaussian process with Matérn kernel K_ν with $\nu = 2.01, 3.01, 4.01$ such that the offset f_δ belongs to H^2, H^3, H^4 respectively. Hence, we consider the following data generation procedure:

$$\{x_{i,T}\}_{i=1}^{n_T}, \{x_{i,S}\}_{i=1}^{n_S} \overset{i.i.d.}{\sim} U([0, 1])$$
$$y_{i,T} = f_T(x_{i,T}) + \sigma\epsilon_{i,T} \quad i = 1, \cdots, n_T$$
$$y_{i,S} = f_T(x_{i,S}) + f_\delta(x_{i,S}) + \sigma\epsilon_{i,S} \quad i = 1, \cdots, n_S$$

where $\epsilon_{i,p}$ are i.i.d. standard Gaussian noise and $\sigma = 0.5$.

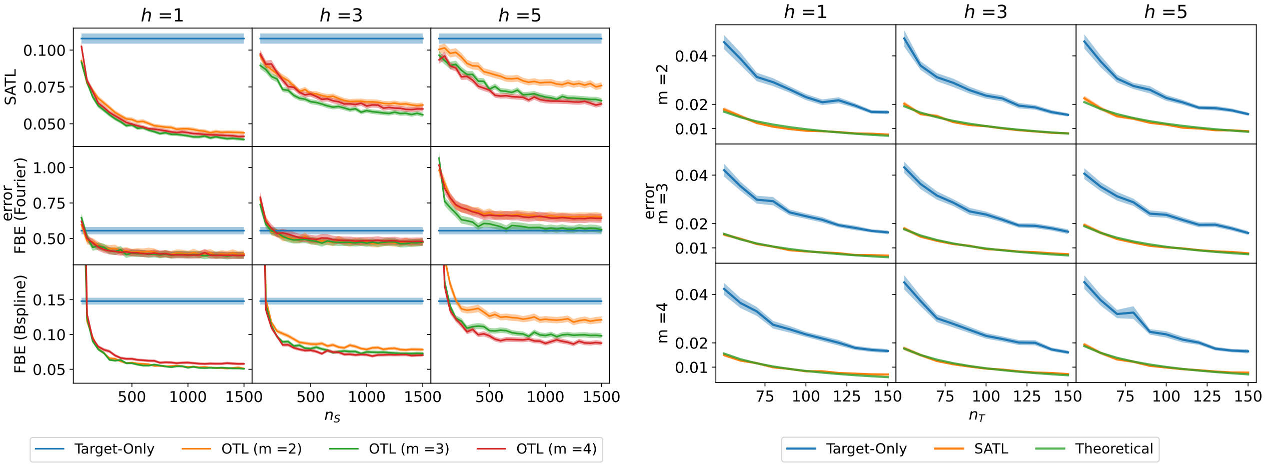
To demonstrate the transfer learning effect, we consider two different settings:

(1) Fixed n_T scenario: Fix n_T as 50 and vary n_S.

(2) Varying n_T scenario: Set $n_S = n_T^{3/2}$ while varying n_T, i.e., the source sample size grows in a polynomial order of target sample size.

In the first scenario, it is expected that the generalization error first decreases and then remains unchanged as n_S increases since the offset estimation error (a constant for fixed n_T) eventually dominates. In the second scenario, the generalization error satisfies $\mathcal{E}(\hat{f}_T) = O(n_T^{-\frac{3m_0}{2m_0+1}} + n_T^{-\frac{2m}{2m+1}}\xi_U) = O(n_T^{-\frac{2m}{2m+1}})$. We consider the finite basis expansion (FBE) TL algorithm proposed in Wang et al. (2016) as a competitor. The authors originally used the Fourier basis in their paper, which produced weak results in our setting. Therefore, we compared SATL to their algorithm with Fourier basis and an additional modification by employing Bspline. We refer to Appendix E for implementation details on different types of Fourier basis and the other experiments' details.

[1] The code to reproduce our experimental results is available at https://github.com/haotianlin/SATL.



(a) Fixed n_T scenario

(b) Varying n_T scenario

Figure 3. Generalization error under different h and smoothness of f_δ. Each curve denotes the average error over 100 trails, and the shadow regions denote one standard error of the mean. The left figure contains results for fixed n_T scenario while the right figure is for varying n_T scenario. In Figure 3b, the green line denotes the theoretical upper bound $n_T^{-2m/(2m+1)}$ (up to constants).

Figure 3a presents the generalization error for the fixed n_T scenario. As n_S increases, the generalization error initially decreases and then gradually levels off, consistent with our expectations. Furthermore, if the smoothness of the offset function f_δ is higher, a smaller error is obtained, which agrees mildly with our theoretical analysis. Finally, compared to the FBE approach, SATL achieves overall smaller errors. Figure 3b presents the generalization error for the varying n_T scenario. Here, the error term is expected to be upper bounded by $n_T^{-\frac{2m}{2m+1}}$. One can see our empirical error is consistent with the theoretical upper bound asymptotically in all settings. Besides, the SATL outperforms the target-only learning KRR baseline in all settings.

6. Future Direction

Beyond Sobolev Space. In developing the optimality of target-only KRR with Gaussian kernels, we use the Fourier transform technique to control the approximation error (see Appendix C.1), which is feasibly applied to RKHS that are norm-equivalent to fractional Sobolev spaces. Although this makes our results quite broadly applicable when one is primarily interested in the smoothness of the functions, it certainly doesn't cover all possible structures of interest (e.g., periodic functions, etc). It is of interest to develop other mathematical tools to extend the current results for Sobolev spaces to more general RKHS.

Few-Shot Transfer Learning. Theorem 4.1 is developed based on the asymptotic rates of Theorem 3.3. Thus, the validity of theoretical results of SATL requires both n_T and n_S sufficiently large to allow some lower order terms vanish (note that although n_T and n_S need to be sufficiently large, $n_S \gg n_T$ still remains, which shows that the settings we consider are still within the regime of transfer learning). In the case when n_T is extremely small, e.g., few-shot transfer learning, one needs a non-asymptotic rate for KRR with fixed bandwidth Gaussian kernel to develop the upper bound.

7. Conclusion

We presented SATL, a kernel-based OTL that uses Gaussian kernels as imposed kernels. This enables the estimators to adapt to the varying and unknown smoothness in their corresponding functions. SATL achieves minimax optimality (up to a logarithmic factor) as the upper bound of SATL matched the lower bound of the OTL problem. Notably, our Gaussian kernels' result in target-only learning also serves as a good supplement to misspecified kernel learning literature.

Impact Statement

This paper aims to theoretically achieve hypothesis transfer learning adaptively under a nonparametric regression setting and provide corresponding statistical guarantees. It provides insights about the transfer dynamic, i.e., when offset transfer learning improves performance compared to target-only learning, and the necessity of adaptive learning in statistical transfer learning. Since the work is focused on a theoretical perspective, there is no present immediate ethical impact or societal implication that we feel must be specifically highlighted here.

References

Aghbalou, A. and Staerman, G. Hypothesis transfer learning with surrogate classification losses: Generalization bounds through algorithmic stability. In *International Conference on Machine Learning*, pp. 280–303. PMLR, 2023.

Bastani, H. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984, 2021.

Bauer, F., Pereverzev, S., and Rosasco, L. On regularization algorithms in learning theory. *Journal of complexity*, 23 (1):52–72, 2007.

Blanchard, G. and Mücke, N. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.

Cai, T. T. and Pu, H. Transfer learning for nonparametric regression: Non-asymptotic minimax analysis and adaptive procedure. *arXiv preprint arXiv:2401.12272*, 2024.

Cai, T. T. and Wei, H. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100–128, 2021.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

Cheng, G. and Shang, Z. Joint asymptotics for semi-nonparametric regression models with partially linear structure. *The Annals of Statistics*, 43(3):1351–1390, 2015.

DeVore, R. A. and Sharpley, R. C. Besov spaces on domains in r^{d}. *Transactions of the American Mathematical Society*, 335(2):843–864, 1993.

Dicker, L. H., Foster, D. P., and Hsu, D. Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators. 2017.

Du, S. S., Koushik, J., Singh, A., and Póczos, B. Hypothesis transfer learning via transformation functions. *Advances in neural information processing systems*, 30, 2017.

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

Duan, Y. and Wang, K. Adaptive and robust multi-task learning. *arXiv preprint arXiv:2202.05250*, 2022.

Eberts, M. and Steinwart, I. Optimal regression rates for SVMs using Gaussian kernels. *Electronic Journal of Statistics*, 7(none):1 – 42, 2013. doi: 10.1214/12-EJS760. URL https://doi.org/10.1214/12-EJS760.

Fasshauer, G. E. and Ye, Q. Reproducing kernels of generalized sobolev spaces via a green function approach with distributional operators. *Numerische Mathematik*, 119: 585–611, 2011.

Fischer, S. and Steinwart, I. Sobolev norm learning rates for regularized least-squares algorithms. *The Journal of Machine Learning Research*, 21(1):8464–8501, 2020.

Gao, J., Fan, W., Jiang, J., and Han, J. Knowledge transfer via multiple model local structure mapping. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 283–291, 2008.

Geer, S. A. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Hamm, T. and Steinwart, I. Adaptive learning rates for support vector machines working on data with low intrinsic dimension. *The Annals of Statistics*, 49(6):3153–3180, 2021.

He, Z., Sun, Y., and Li, R. Transfusion: Covariate-shift robust transfer learning for high-dimensional regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 703–711. PMLR, 2024.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

Kpotufe, S. and Martinet, G. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323, 2021.

Kuzborskij, I. and Orabona, F. Stability and hypothesis transfer learning. In *International Conference on Machine Learning*, pp. 942–950. PMLR, 2013.

Kuzborskij, I. and Orabona, F. Fast rates by transferring from auxiliary hypotheses. *Machine Learning*, 106:171–195, 2017.

Lepskii, O. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35 (3):454–466, 1991.

Li, S., Cai, T. T., and Li, H. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 2022.

Li, T. and Yuan, M. On the optimality of gaussian kernel based nonparametric tests against smooth alternatives. *arXiv preprint arXiv:1909.03302*, 2019.

Li, X. and Bilmes, J. A bayesian divergence prior for classiffier adaptation. In *Artificial Intelligence and Statistics*, pp. 275–282. PMLR, 2007.

Li, Y., Zhang, H., and Lin, Q. On the saturation effect of kernel ridge regression. In *The Eleventh International Conference on Learning Representations*, 2023.

Lin, H. and Reimherr, M. On hypothesis transfer learning of functional linear models. *stat*, 1050:22, 2024.

Lin, J. and Cevher, V. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *Journal of Machine Learning Research*, 21 (147):1–63, 2020.

Ma, C., Pathak, R., and Wainwright, M. J. Optimally tackling covariate shift in rkhs-based nonparametric regression. *arXiv preprint arXiv:2205.02986*, 2022.

Mendelson, S. and Neeman, J. Regularization in kernel learning. 2010.

Minami, S., Fukumizu, K., Hayashi, Y., and Yoshida, R. Transfer learning with affine model transformation. *Advances in Neural Information Processing Systems*, 36, 2024.

Orabona, F., Castellini, C., Caputo, B., Fiorilla, A. E., and Sandini, G. Model adaptation with least-squares svm for adaptive hand prosthetics. In *2009 IEEE international conference on robotics and automation*, pp. 2897–2903. IEEE, 2009.

Reeve, H. W., Cannings, T. I., and Samworth, R. J. Adaptive transfer learning. *The Annals of Statistics*, 49(6):3618–3649, 2021.

Smale, S. and Zhou, D.-X. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

Stein, M. L. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 1999.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Steinwart, I., Hush, D., and Scovel, C. An explicit description of the reproducing kernel hilbert spaces of gaussian rbf kernels. *IEEE Transactions on Information Theory*, 52(10):4635–4643, 2006.

Steinwart, I., Hush, D. R., Scovel, C., et al. Optimal rates for regularized least squares regression. In *COLT*, pp. 79–93, 2009.

Tian, Y. and Feng, Y. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, pp. 1–14, 2022.

Tian, Y., Gu, Y., and Feng, Y. Learning from similar linear representations: Adaptivity, minimaxity, and robustness. *arXiv preprint arXiv:2303.17765*, 2023.

Tripuraneni, N., Jordan, M., and Jin, C. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33: 7852–7862, 2020.

Varshamov, R. R. Estimate of the number of signals in error correcting codes. *Docklady Akad. Nauk, SSSR*, 117: 739–741, 1957.

Wang, K. Pseudo-labeling for kernel ridge regression under covariate shift. *arXiv preprint arXiv:2302.10160*, 2023.

Wang, W. and Jing, B.-Y. Gaussian process regression: Optimality, robustness, and relationship with kernel ridge regression. *Journal of Machine Learning Research*, 23 (193):1–67, 2022.

Wang, X. and Schneider, J. G. Generalization bounds for transfer learning under model shift. In *UAI*, pp. 922–931, 2015.

Wang, X., Oliva, J. B., Schneider, J. G., and Póczos, B. Nonparametric risk and stability analysis for multi-task learning problems. In *IJCAI*, pp. 2146–2152, 2016.

Wendland, H. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

Xu, Z. and Tewari, A. Representation learning beyond linear prediction functions. *Advances in Neural Information Processing Systems*, 34:4792–4804, 2021.

Zhang, H., Li, Y., Lu, W., and Lin, Q. On the optimality of misspecified kernel ridge regression. In *International Conference on Machine Learning*, pp. 41331–41353. PMLR, 2023.

Zhang, X., Blanchet, J., Ghosh, S., and Squillante, M. S. A class of geometric structures in transfer learning: Minimax bounds and optimality. In *International Conference on Artificial Intelligence and Statistics*, pp. 3794–3820. PMLR, 2022.

Appendix

This appendix encompasses more discussions, experiments, and proofs of the theoretical results presented in the main body. Appendix A provides the introduction of the notation we used. Appendix B provides basic concepts of RKHS, Sobolev space, the interpolation space, and the results of norm-equivalence between RKHS and Sobolev space. Appendix C presents the proofs of the results in the Target-Only KRR section, including non-adaptive rate (Theorem 3.1) and adaptive rate (Theorem 3.3). A discussion of the comparison between this work and previous works is also presented. Appendix D contains the proofs of the upper bound and lower bounds of SATL. We present additional experiment details in Appendix E.

A. Notation

The following notations are used throughout the rest of this work and follow standard conventions. For asymptotic notations: $f(n) = O(g(n))$ means for all c there exists $k > 0$ such that $f(n) \leq cg(n)$ for all $n \geq k$; $f(n) \asymp g(n)$ means $f(n) = O(g(n))$ and $g(n) = O(f(n))$; $f(n) = \Omega(g(n))$ means for all c there exists $k > 0$ such that $f(n) \geq cg(n)$ for all $n \geq k$. We use the asymptotic notations in probability $O_{\mathbb{P}}(\cdot)$. That is, for a positive sequence $\{a_n\}_{n\geq 1}$ and a non-negative random variable sequence $\{X_n\}_{n\geq 1}$, we say $X_n = O_{\mathbb{P}}(a_n)$ if for any $\delta > 0$, there exist M_δ and N_δ such that $\mathbb{P}(X_n \leq M_\delta a_n) \geq 1 - \delta, \forall n \geq N_\delta$. The definition of $\Omega_{\mathbb{P}}$ follows similarly.

For a function $f \in L_1(\mathbb{R}^d)$, its Fourier transform is denoted as

$$\mathcal{F}(f)(\omega) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(x) e^{-ix^T\omega} dx.$$

Since we assume $\mu_T = \mu_S$, we use $L_2(\mathcal{X}, d\mu_p)$ for $p \in \{T, S\}$ to represent the Lebesgue L_2 space and abbreviate it as L_2 for simplicity when there is no confusion.

B. Foundation of RKHS

B.1. Basic Concept

In this section, we will present some facts about the RKHS that are useful in our proof and refer readers to Wendland (2004) for more details.

Assume $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a continuous positive definite kernel function defined on a compact set $\mathcal{X} \subset \mathbb{R}^d$ (with positive Lebesgure measure and Lipschitz boundary). Indeed, every positive definite kernel can be associated with a reproducing kernel Hilbert space (RKHS). The RKHS, $\mathcal{H}_K$, of K are usually defined as the closure of linear space span$\{K(\cdot, x), x \in \mathcal{X}\}$. In a special case where the kernel function $K(x, y)$ is equal to a translation invariant (stationary) function $\Phi(x - y) = K(x, y)$ with $\Phi : \mathbb{R}^d \to \mathbb{R}$, we can characterize the RKHS of K in terms of Fourier transforms, i.e.

$$\mathcal{H}_K(\mathbb{R}^d) = \left\{ f \in L_2(\mathbb{R}^d) \cap C(\mathbb{R}^d) : \frac{\mathcal{F}(f)}{\sqrt{\mathcal{F}(\Phi)}} \in L_2(\mathbb{R}^d) \right\}.$$

When $\mathcal{X}$ is a subset of $\mathbb{R}^d$, such a definition still captures the regularity of functions in $\mathcal{H}_K(\mathcal{X})$ via a norm equivalency result that holds as long as X has a Lipschitz boundary.

For an integer m, we introduce the integer-order Sobolev space, $\mathcal{W}^{m,p}(\mathcal{X})$. For vector $\alpha = (\alpha_1, \cdots, \alpha_d)$, define $|\alpha| = \alpha_1 + \cdots + \alpha_d$ and $D^{(\alpha)} = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$ denote the multivariate mixed partial weak derivative. Then

$$\mathcal{W}^{m,p}(\mathcal{X}) = \left\{ f \in L^p(\mathcal{X}) : D^{(\alpha)} f \in L_p(\mathbb{R}^d), \forall |\alpha| \leq m \right\},$$

where m is the smoothness order of the Sobolev space. In this paper, we only consider $p = 2$ and abbreviate $\mathcal{W}^{m,2}(\mathcal{X}) := H^m(\mathcal{X})$. Later, in Appendix B.2, we will see one can define the H^m via Fourier transform of the reproducing kernel instead of weak derivative.

We now introduce the power space of an RKHS. For the reproducing kernel K, we can define its integral operator $T_K : L_2 \to L_2$ as

$$T_K(f)(\cdot) = \int_{\mathcal{T}} K(s, \cdot) f(s) ds.$$

L_K is self-adjoint, positive-definite, and trace class (thus Hilbert-Schmidt and compact). By the spectral theorem for self-adjoint compact operators, there exists an at most countable index set N, a non-increasing summable positive sequence $\{\tau_j\}_{j\geq 1}$ and an orthonormal basis of L_2, $\{e_j\}_{j\geq 1}$ such that the integrable operator can be expressed as

$$T_K(\cdot) = \sum_{j \in N} \tau_j \langle \cdot, e_j \rangle_{L_2} e_j.$$

The sequence $\{\tau_j\}_{j\geq 1}$ and the basis $\{e_j\}_{j\geq 1}$ are referred as the eigenvalues and eigenfunctions. The Mercer's theorem shows that the kernel K itself can be expressed as

$$K(x,x') = \sum_{j\in N} \tau_j e_j(x) e_j(x'), \quad \forall x, x' \in \mathcal{T},$$

where the convergence is absolute and uniform.

We now introduce the fractional power integral operator and the composite integral operator of two kernels. For any $s \geq 0$, the fractional power integral operator $L_K^s : L_2 \to L_2$ is defined as

$$T_K^s(\cdot) = \sum_{j\in N} \tau_j^s \langle \cdot, e_j\rangle_{L_2} e_j.$$

Then the power space $[\mathcal{H}_K]^s$ is defined as

$$[\mathcal{H}_K]^s := \left\{ \sum_{j\in N} a_j \tau_j^{\frac{s}{2}} e_j : (a_j) \in \ell^2(N) \right\}$$

and equipped with the inner product

$$\langle f, g\rangle_{[\mathcal{H}_K]^s} = \left\langle T_K^{-\frac{s}{2}}(f), T_K^{-\frac{s}{2}}(g) \right\rangle_{L_2}.$$

For $0 < s_1 < s_2$, the embedding $[\mathcal{H}_K]^{s_2} \hookrightarrow [\mathcal{H}_K]^{s_1}$ exists and is compact. A higher s indicates the functions in $[\mathcal{H}_K]^s$ have higher regularity. When $\mathcal{H}_K = H^m$ with $m > d/2$, the real interpolation indicates $[H^m]^s \cong H^{ms}, \forall s > 0$.

B.2. Norm Equivalency between RKHS and Sobolev Space

Now, we state the result that connects the general RKHS and Sobolev space.

Lemma B.1. *Let $K(x,x')$ be the translation-invariant kernel and $\tilde{K} \in L^1(\mathbb{R}^d)$. Suppose $\mathcal{X}$ has a Lipschitiz boundary, and the Fourier transform of K has the following spectral density of m, for $m \geq d/2$,*

$$c_1(1 + \|\cdot\|_2^2)^m \leq \mathcal{F}(K)(\cdot) \leq c_2(1 + \|\cdot\|_2^2)^m. \tag{2}$$

for some constant $0 < c_1 \leq c_2$. Then, the associated RKHS of K, $\mathcal{H}_K(\mathcal{X})$, is norm-equivalent to the Sobolev space $H^m(\mathcal{X})$.

Hence, we can naturally define the Sobolev space of order m ($m > \frac{d}{2}$) as

$$H^m(\mathbb{R}^d) = \left\{ f \in L_2(\mathbb{R}^d) \cap C(\mathbb{R}^d) : \mathcal{F}(f)(\cdot)(1 + \|\cdot\|_2^2)^m \in L_2(\mathbb{R}^d) \right\}.$$

One advantage of this definition over the classical way that involves weak derivatives is it does not require m to be an integer, and thus one can consider the fractional Sobolev space, i.e. $m \in \mathbb{R}^+$. Such equivalence also holds on $\mathcal{X}$ by applying the extension theorem (DeVore & Sharpley, 1993). As an implication, let $K_{m,\nu}$ denotes the isotropic Matérn kernel (Stein, 1999), i.e.

$$K_{m,\nu}(x;\rho) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{\|x\|_2}{\rho} \right)^{\nu} K_\nu \left(\sqrt{2\nu} \frac{\|x\|_2}{\rho} \right),$$

then the Fourier transform of $K_{m,\nu}$ satisfies Equation 2 with $m = \nu + \frac{d}{2}$, and thus the RKHS associated with $K_{m,\nu}$ is norm equivalent to Sobolev space $H^{\nu+\frac{d}{2}}$ (Wendland, 2004).

For a reproducing kernel that satisfies (2), we call it a kernel with Fourier decay rate m and denote it as K_m. We further denote its associated RKHS as $\mathcal{H}_{K_m}(\mathcal{X})$. The Fourier decay rate m captures the regularity of $\mathcal{H}_{K_m}(\mathcal{X})$.

Now, we are ready to define the function space of f_S, f_T and f_δ via the kernel regularity

Assumption B.2 (Smoothness of Target/Source)**.** There exists an $m_0 \geq d/2$ such that f_T and f_S belong to $\mathcal{H}_{K_{m_0}}$.

Assumption B.3 (Smoothness of Offset)**.** There exists an $m \geq m_0$ such that $f_\delta := f_T - f_S$ belongs to $\mathcal{H}_{K_m}$.

The proof of all the theoretical results in Section 3 and 4 is built on the assumptions that the true functions are in Sobolev space. Via the norm equivalency (Lemma B.1), the true functions also reside in RKHSs associated with kernel K_{m_0} and K_m. Therefore, all the theoretical results still hold under Assumption B.2 and B.3.

C. Target-Only KRR Learning Results

C.1. Proof Scheme

Before formally proving the theoretical results, we first illustrate the main scheme for the proof of Theorem 3.1.

For the given imposed Gaussian kernel K, we define its corresponding integral operator $T_K : L_2(\mathcal{X}) \to L_2(\mathcal{X})$ as

$$T_K(f)(\cdot) = \int_{\mathcal{X}} f(x)K(x, \cdot)d\rho(x)$$

where $\rho(x)$ is the probability measure over $\mathcal{X}$. We also note that the integral operator T_K can also be viewed as a bounded linear operator on $\mathcal{H}_K$.

We now consider its empirical version when the sample $\{(x_i, y_i)\}_{i=1}^n$ are available. For a $x \in \mathcal{X}$, we define the sampling operator as $K_x : \mathcal{H}_K \to \mathbb{R}$ and its adjoint operator $K_x^* : \mathbb{R} \to \mathcal{H}_K$ as

$$K_x : \mathcal{H}_K \to \mathbb{R}, f \mapsto \langle f, K_x \rangle_{\mathcal{H}_K} \quad \text{and} \quad K_x^* : \mathbb{R} \to \mathcal{H}_K, y \mapsto yK_x.$$

Then we can define the empirical version of T_K, termed sample covariance operator, as $T_{K,n} : \mathcal{H}_K \to \mathcal{H}_K$ as

$$T_{K,n} = \frac{1}{n}\sum_{i=1}^n K_{x_i}^* K_{x_i}.$$

With these notations, the KRR estimator can be written as

$$\hat{f} = (T_{K,n} + \lambda\mathbf{I})^{-1}(\frac{1}{n}\sum_{i=1}^n K_{x_i} y_i) := (T_{K,n} + \lambda\mathbf{I})^{-1} g_n$$

where $g_n = \frac{1}{n}\sum_{i=1}^n K_{x_i} y_i \in \mathcal{H}_K$ and $\mathbf{I}$ is the identical operator. We further define the intermediate term as follows,

$$f_\lambda := \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \left\{ \|(f_0 - f)\|_{L_2}^2 + \lambda \|f\|_{\mathcal{H}_K}^2 \right\}$$

and one can show $f_\lambda = (T_K + \lambda\mathbf{I})^{-1} T_K(f_0) = (T_K + \lambda\mathbf{I})^{-1} g$.

Then by triangle inequality,

$$\left\|\hat{f} - f_0\right\|_{L_2} \leq \underbrace{\left\|\hat{f} - f_\lambda\right\|_{L_2}}_{\text{estimation error}} + \underbrace{\|f_\lambda - f_0\|_{L_2}}_{\text{approximation error}}.$$

The following paragraphs discuss the analysis for each of the error terms and how they differ from previous works.

Approximation error. In classical misspecified kernel methods, one typically controls the approximation error via interpolation/power space technique. Specifically, the intermediate term f_λ is placed in $[H^{\alpha_0}]^s$ while the true function in H^{α_0} (here, s denotes the interpolation index). Therefore, one can expand f_λ and f_0 under the same basis since f_λ lies in the interpolation/power space of H^{α_0}, which typically controls the approximation error in the form of $\lambda^s \|f_0\|_{H^{\alpha_0}}$ and makes the optimal order of λ in n takes polynomial pattern like Proposition 2.1. This technique is widely used in many misspecified kernel literature like Theorem A.2 in Zhang et al. (2023) and etc.

However, since in our case, the intermediate term, f_λ, lies in the RKHS associated with Gaussian kernels (s needs to be ∞), one can't expand f_λ and f_0 under the same basis and thus such a technique no longer holds. Therefore, the techniques we used are the Fourier transform of the Gaussian kernel and Plancherel Theorem, which allows us to prove

$$\|f_\lambda - f_0\|_{L_2}^2 \leq \|f_\lambda - f_0\|_{L_2}^2 + \lambda \|f_\lambda\|_{\mathcal{H}_K}^2 \leq C log\left(\frac{1}{\lambda}\right)^{-\alpha_0} \|f_0\|_{H^{\alpha_0}}^2.$$

Here, the second inequality is proved via the Plancherel Theorem, see Proposition C.2, and $\mathcal{H}_K$ denote the RKHS associated with Gaussian kernels. We also highlight that this is a technical contribution of this paper.

Estimation error. Regarding the estimation error, we use the standard integral operator techniques origin from Smale & Zhou (2007) and follow a similar strategy as Fischer & Steinwart (2020); Zhang et al. (2023). While most of the current work deals with cases where the eigenvalue decay rate is polynomial, we refine the proof to handle the Gaussian kernel case, whose eigenvalues decay exponentially.

C.2. Proof of Non-adaptive Rate (Theorem 3.1)

In the following proof, we will use C, C_1, and C_2 to represent constants that could change from place to place. Unless specifically specified, we also omit the $\mathcal{X}$ in the norms or in the inner product notation. We use $\|\cdot\|_{op}$ to denote the operator norm of a bounded linear operator.

In addition, we denote the effective dimension as

$$\mathcal{N}(\lambda) = tr((T_K + \lambda)^{-1}T_K) = \sum_{j=1}^{\infty} \frac{s_j}{s_j + \lambda}.$$

C.2.1. Proof of the approximation error

For the approximation error, we can directly apply Proposition C.2, which leads to

$$\|f_\lambda - f_0\|_{L_2}^2 \leq log(\frac{1}{\lambda})^{-\alpha_0}\|f_0\|_{H^{\alpha_0}}^2.$$

Then selecting $\log(1/\lambda) \asymp n^{\frac{2}{2\alpha_0+d}}$ leads to

$$\|f_\lambda - f_0\|_{L_2}^2 \leq n^{-\frac{2\alpha_0}{2\alpha_0+d}}\|f_0\|_{H^{\alpha_0}}^2.$$

C.2.2. Proof of the estimator error

Theorem C.1. *Suppose the Assumption (A1) to (A3) hold and $\|f_0\|_{L_q} \leq C_q$ for some q. Then by choosing $\log(1/\lambda) \asymp n^{\frac{2}{2\alpha_0+d}}$, for any fixed $\delta \in (0,1)$, when n is sufficient large, with probability $1-\delta$, we have*

$$\left\|\hat{f} - f_\lambda\right\|_{L_2} \leq ln(\frac{4}{\delta})Cn^{-\frac{\alpha_0}{2\alpha_0+d}}$$

where C is a constant proportional to σ.

Proof. First, we notice that

$$\begin{aligned}\left\|\hat{f} - f_\lambda\right\|_{L_2} &= \left\|T_K^{\frac{1}{2}}\left(\hat{f} - f_\lambda\right)\right\|_{\mathcal{H}_K} \\ &\leq \underbrace{\left\|T_K^{\frac{1}{2}}\left(T_K + \lambda I\right)^{-\frac{1}{2}}\right\|_{op}}_{A_1} \\ &\quad \cdot \underbrace{\left\|\left(T_K + \lambda I\right)^{\frac{1}{2}}\left(T_{K,n} + \lambda I\right)^{-1}\left(T_K + \lambda I\right)^{\frac{1}{2}}\right\|_{op}}_{A_2} \\ &\quad \cdot \underbrace{\left\|\left(T_K + \lambda I\right)^{-\frac{1}{2}}\left(g_n - \left(T_{K,n} + \lambda I\right)f_\lambda\right)\right\|_{\mathcal{H}_K}}_{A_3}\end{aligned}$$

For the first term A_1, we have

$$A_1 = \left\|T_K^{\frac{1}{2}}\left(T_K + \lambda I\right)^{-\frac{1}{2}}\right\|_{op} = \sup_{i\geq 1}\left(\frac{s_j}{s_j + \lambda}\right)^{\frac{1}{2}} \leq 1.$$

For the second term, using Proposition C.3 with sufficient large n, we have

$$v := \frac{\mathcal{N}(\lambda)}{n} ln(\frac{8\mathcal{N}(\lambda)}{\delta} \frac{(\|T_K\|_{op} + \lambda)}{\|T_K\|_{op}}) \leq \frac{1}{8}$$

such that

$$\left\|(T_K + \lambda I)^{-\frac{1}{2}} (T_K - T_{K,n}) (T_K + \lambda I)^{-\frac{1}{2}}\right\|_{op} \leq \frac{4}{3} v + \sqrt{2v} \leq \frac{2}{3}$$

holds with probability $1 - \frac{\delta}{2}$. Thus,

$$\begin{aligned} A_2 &= \left\|(T_K + \lambda I)^{\frac{1}{2}} (T_{K,n} + \lambda I^{-1}) (T_K + \lambda I)^{\frac{1}{2}}\right\|_{op} \\ &= \left\|\left((T_K + \lambda I)^{-\frac{1}{2}} (T_{K,n} + \lambda) (T_K + \lambda I)^{-\frac{1}{2}}\right)^{-1}\right\|_{op} \\ &= \left\|\left(I - (T_K + \lambda I)^{-\frac{1}{2}} (T_{K,n} - T_K) (T_K + \lambda I)^{-\frac{1}{2}}\right)^{-1}\right\|_{op} \\ &\leq \sum_{k=0}^{\infty} \left\|(T_K + \lambda I)^{-\frac{1}{2}} (T_K - T_{K,n}) (T_K + \lambda I)^{-\frac{1}{2}}\right\|_{op}^{k} \\ &\leq \sum_{k=0}^{\infty} \left(\frac{2}{3}\right)^k \leq 3, \end{aligned}$$

For the third term A_3, notice

$$\left\|(T_K + \lambda I)^{-\frac{1}{2}} (g_n - (T_{K,n} + \lambda I) f_\lambda)\right\|_{\mathcal{H}_K} = \left\|(T_K + \lambda I)^{-\frac{1}{2}} \left[(g_n - T_{K,n}(f_\lambda)) - (g - T_K(f_\lambda))\right]\right\|_{\mathcal{H}_K}$$

Using the Proposition C.4, with probability $1 - \frac{\delta}{2}$, we have

$$T_3 = \left\|(T_K + \lambda I)^{-\frac{1}{2}} (g_n - (T_{K,n} + \lambda I) f_\lambda)\right\|_{\mathcal{H}_K} \leq C ln(\frac{4}{\delta}) n^{-\frac{\alpha_0}{2\alpha_0 + d}}$$

where $C \propto \sigma$. Combing the bounds for T_1, T_2 and T_3, we finish the proof. □

C.2.3. Proof of Theorem 3.1

Based on the bounds of the approximation and estimation error, we finish the proof as follows,

$$\begin{aligned} \left\|\hat{f} - f_0\right\|_{L_2} &\leq \left\|\hat{f} - f_\lambda\right\|_{L_2} + \|f_\lambda - f_0\|_{L_2} \\ &= O_{\mathbb{P}}\left\{(\sigma + \|f_0\|_{H^{\alpha_0}}) n^{-\frac{\alpha_0}{2\alpha_0 + d}}\right\}. \end{aligned}$$

C.2.4. Propositions

Proposition C.2. *Suppose f_λ is defined as follows,*

$$f_\lambda = \underset{f \in \mathcal{H}_K(\mathcal{X})}{\operatorname{argmin}} \left\{\|f - f_0\|_{L_2(\mathcal{X})}^2 + \lambda \|f\|_{\mathcal{H}_K(\mathcal{X})}^2\right\}.$$

Then, under the regularized conditions, the following inequality holds,

$$\|f_\lambda - f_0\|_{L_2(\mathcal{X})}^2 + \lambda \|f_\lambda\|_{\mathcal{H}_K(\mathcal{X})}^2 \leq C log\left(\frac{1}{\lambda}\right)^{-\alpha_0} \|f_0\|_{H^{\alpha_0}}^2.$$

Proof. Since $\mathcal{X}$ has Lipschitz boundary, there exists an extension mapping from $L_2(\mathcal{X})$ to $L_2(\mathbb{R}^d)$, such that the smoothness of functions in $L_2(\mathcal{X})$ get preserved. Therefore, there exist constants C_1 and C_2 such that for any function $g \in H^{\alpha_0}(\mathcal{X})$, there exists an extension of g, $g_e \in H^{\alpha_0}(\mathbb{R}^d)$ satisfying

$$C_1\|g_e\|_{H^{\alpha_0}(\mathbb{R}^d)} \le \|g\|_{H^{\alpha_0}(\mathcal{X})} \le C_2\|g_e\|_{H^{\alpha_0}(\mathbb{R}^d)}.$$

Denote

$$f_{\lambda,e} = \operatorname*{argmin}_{f\in\mathcal{H}_K(\mathbb{R}^d)} \left\{\|f-f_{0,e}\|^2_{L_2(\mathbb{R}^d)} + \lambda\|f\|^2_{\mathcal{H}_K(\mathbb{R}^d)}\right\}$$

Then we have,

$$\begin{aligned}\|f_\lambda - f_0\|^2_{L_2(\mathcal{X})} + \lambda\|f_\lambda\|^2_{\mathcal{H}_K(\mathcal{X})} &\le \|f_{\lambda,e}|_{\mathcal{X}} - f_0\|^2_{L_2(\mathcal{X})} + \lambda\|f_{\lambda,e}|_{\mathcal{X}}\|^2_{L_2(\mathcal{X})}\\ &\le C_2\left(\|f_{\lambda,e} - f_{0,e}\|^2_{L_2(\mathbb{R}^d)} + \lambda\|f_{\lambda,e}\|^2_{L_2(\mathbb{R}^d)}\right).\end{aligned}$$

where $f_{\lambda,e}|_{\mathcal{X}}$ is the restriction of $f_{\lambda,e}$ on $\mathcal{X}$. By Fourier transform of the Gaussian kernel and Plancherel Theorem, we have

$$\begin{aligned}&\|f_{\lambda,e} - f_{0,e}\|^2_{L_2(\mathbb{R}^d)} + \lambda\left\|f_{\lambda,e}\right\|^2_{\mathcal{H}_K(\mathbb{R}^d)}\\ =&\int_{\mathbb{R}^d}\left|\mathcal{F}\left(f_{0,e}\right)(\omega) - \mathcal{F}\left(f_{\lambda,e}\right)(\omega)\right|^2 d\omega + \lambda\int_{\mathbb{R}^d}\left|\mathcal{F}\left(f_{\lambda,e}\right)(\omega)\right|^2 exp\{C\|\omega\|_2^2\}d\omega\\ =&\int_{\mathbb{R}^d}\left(\left|\mathcal{F}\left(f_{0,e}\right)(\omega) - \mathcal{F}\left(f_{\lambda,e}\right)(\omega)\right|^2 + \lambda\left|\mathcal{F}\left(f_{\lambda,e}\right)(\omega)\right|^2 exp\{C\|\omega\|_2^2\}\right)d\omega\\ =&\int_{\mathbb{R}^d}\frac{\lambda exp\{C\|\omega\|_2^2\}}{1+\lambda exp\{C\|\omega\|_2^2\}}\left|\mathcal{F}\left(f_{0,e}\right)(\omega)\right|^2 d\omega\\ \le&\int_{\Omega}\frac{\lambda exp\{C(1+\|\omega\|_2^2)\}}{1+\lambda exp\{C(1+\|\omega\|_2^2)\}}\left|\mathcal{F}\left(f_{0,e}\right)(\omega)\right|^2 d\omega + \int_{\Omega^C}\frac{\lambda exp\{C(1+\|\omega\|_2^2)\}}{1+\lambda exp\{C(1+\|\omega\|_2^2)\}}\left|\mathcal{F}\left(f_{0,e}\right)(\omega)\right|^2 d\omega\\ \le&\int_{\Omega}\lambda exp\{C(1+\|\omega\|_2^2)\}\left|\mathcal{F}\left(f_{0,e}\right)(\omega)\right|^2 d\omega + \int_{\Omega^C}\left|\mathcal{F}\left(f_{0,e}\right)(\omega)\right|^2 d\omega\end{aligned}$$

where $\Omega = \{\omega : \lambda exp\{C(1+\|\omega\|_2^2)\} < 1\}$ and $\Omega^C = \mathbb{R}^d\backslash\Omega$, and the third equality follows the definition of f_e^*. Over Ω^C, we notice that

$$(1+\|\omega\|_2^2) \ge \frac{1}{C}log\left(\frac{1}{\lambda}\right) \implies C^{\alpha_0}log\left(\frac{1}{\lambda}\right)^{-\alpha_0}(1+\|\omega\|_2^2)^{\alpha_0} \ge 1.$$

Over Ω, we first note that the function $h(\omega) = exp\{C(1+\|\omega\|_2^2)\}/(1+\|\omega\|_2^2)^{\alpha_0}$ reaches its maximum $C^{\alpha_0}\lambda^{-1}log(\frac{1}{\lambda})^{-\alpha_0}$ if λ satisfies $\lambda < exp\{-\alpha_0\}$ and $\lambda log(\frac{1}{\lambda})^{\alpha_0} \le C^{\alpha_0}exp\{-C\}$. One can verify when $\lambda \to 0$ as $n \to 0$, the two previous inequality holds. Then

$$\lambda exp\{C(1+\|\omega\|_2^2)\} \le C^{\alpha_0}log\left(\frac{1}{\lambda}\right)^{-\alpha_0}(1+\|\omega\|_2^2)^{\alpha_0} \quad \forall\omega\in\Omega.$$

Combining the inequality over Ω and Ω^C,

$$\begin{aligned}&\|f_e^* - f_{0,e}\|^2_{L_2(\mathbb{R}^d)} + \lambda\left\|f_e^*\right\|^2_{\mathcal{H}_K(\mathbb{R}^d)}\\ \le&\int_{\Omega}\lambda exp\{C(1+\|\omega\|_2^2)\}\left|\mathcal{F}\left(f_{0,e}\right)(\omega)\right|^2 d\omega + \int_{\Omega^C}\left|\mathcal{F}\left(f_{0,e}\right)(\omega)\right|^2 d\omega\\ \le&C^{\alpha_0}log\left(\frac{1}{\lambda}\right)^{-\alpha_0}\int_{\mathbb{R}^d}(1+\|\omega\|_2^2)^{\alpha_0}|\mathcal{F}(f_{0,e})(\omega)|^2 d\omega\\ =&C^{\alpha_0}log\left(\frac{1}{\lambda}\right)^{-\alpha_0}\|f_{0,e}\|^2_{H^{\alpha_0}(\mathbb{R}^d)}\\ \le&C^{'}log\left(\frac{1}{\lambda}\right)^{-\alpha_0}\|f_0\|^2_{H^{\alpha_0}(\mathcal{X})}\end{aligned}$$

which completes the proof. □

Proposition C.3. *For all $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\left\|(T_K+\lambda I)^{-\frac{1}{2}}\left(T_K-T_{K,n}\right)(T_K+\lambda I)^{-\frac{1}{2}}\right\|_{op} \leq \frac{4\mathcal{N}(\lambda)B}{3n}+\sqrt{\frac{2\mathcal{N}(\lambda)}{n}B}$$

where

$$B = ln(\frac{4\mathcal{N}(\lambda)}{\delta}\frac{(\|T_K\|_{op}+\lambda)}{\|T_K\|_{op}}).$$

Proof. Denote $A_i = (T_K+\lambda I)^{-\frac{1}{2}}(T_K-T_{K,x_i})(T_K+\lambda I)^{-\frac{1}{2}}$, applying Lemma C.7, we get

$$\begin{aligned}\|A_i\|_{op} &\leq \left\|(T_K+\lambda I)^{-\frac{1}{2}}T_{K,x}(T_K+\lambda I)^{-\frac{1}{2}}\right\|_{op}+\left\|(T_K+\lambda I)^{-\frac{1}{2}}T_{K,x_i}(T_K+\lambda I)^{-\frac{1}{2}}\right\|_{op}\\ &\leq 2E_K^2\mathcal{N}(\lambda)\end{aligned}$$

Notice

$$\begin{aligned}\mathrm{E}\,A_i^2 &\preceq \mathrm{E}\left[(T_K+\lambda I)^{-\frac{1}{2}}T_{K,x_i}(T_K+\lambda I)^{-\frac{1}{2}}\right]^2\\ &\preceq E_K^2\mathcal{N}(\lambda)\,\mathrm{E}\left[(T_K+\lambda I)^{-\frac{1}{2}}T_{K,x_i}(T_K+\lambda I)^{-\frac{1}{2}}\right]\\ &= E_K^2\mathcal{N}(\lambda)(T_K+\lambda I)^{-1}T_K := V\end{aligned}$$

where $A \preceq B$ denotes $B-A$ is a positive semi-definite operator. Notice

$$\|V\|_{op} = \mathcal{N}(\lambda)\frac{\|T_K\|_{op}}{\|T_K\|_{op}+\lambda} \leq \mathcal{N}(\lambda), \quad \text{and} \quad tr(V) = \mathcal{N}(\lambda)^2.$$

The proof is finished by applying Lemma C.8 to A_i and V. □

Proposition C.4. *Suppose that Assumptions in the estimation error theorem hold. We have*

$$\left\|(T_K+\lambda I)^{-\frac{1}{2}}\left(g_n-(T_{K,n}+\lambda I)^{-1}f_\lambda\right)\right\|_{\mathcal{H}_K} \leq Cln(\frac{4}{\delta})n^{-\frac{\alpha_0}{2\alpha_0+d}}$$

where C is a constant.

Proof. Denote

$$\begin{aligned}\xi_i &= \xi(x_i,y_i) = (T_K+\lambda I)^{-\frac{1}{2}}(K_{x_i}y_i-T_{K,x_i}f_\lambda)\\ \xi_x &= \xi(x,y) = (T_K+\lambda I)^{-\frac{1}{2}}(K_xy-T_{K,x}f_\lambda),\end{aligned}$$

then it is equivalent to show

$$\left\|\frac{1}{n}\sum_{i=1}^n\xi_i-\mathrm{E}\,\xi_x\right\|_{\mathcal{H}_K} \leq Cln(\frac{4}{\delta})n^{-\frac{\alpha_0}{2\alpha_0+d}}$$

Define $\Omega_1 = \{x\in\mathcal{X} : |f_0| \leq t\}$ and $\Omega_2 = \mathcal{X}\backslash\Omega_1$. We decompose ξ_i and ξ_x over Ω_1 and Ω_2, which leads to

$$\begin{aligned}\left\|(T_K+\lambda I)^{-\frac{1}{2}}\left(g_n-(T_{K,n}+\lambda I)^{-1}f_\lambda\right)\right\|_{\mathcal{H}_K} \leq& \underbrace{\left\|\frac{1}{n}\sum_{i=1}^n\xi_iI_{x_i\in\Omega_1}-\mathrm{E}\,\xi_xI_{x\in\Omega_1}\right\|_{\mathcal{H}_K}}_{I_1}\\ &+\underbrace{\left\|\frac{1}{n}\sum_{i=1}^n\xi_iI_{x_i\in\Omega_2}\right\|_{\mathcal{H}_K}}_{I_2}+\underbrace{\|\mathrm{E}\,\xi_xI_{x\in\Omega_2}\|_{\mathcal{H}_K}}_{I_3}.\end{aligned}$$

For I_1, applying Proposition C.5, for any $\delta\in(0,1)$, with probability $1-\delta$, we have

$$I_1 \leq \log(\frac{2}{\delta})\left(\frac{C_1\sqrt{\mathcal{N}(\lambda)}}{n}\tilde{M}+\frac{C_2\sqrt{\mathcal{N}(\lambda)}}{\sqrt{n}}+\frac{C_1log(\frac{1}{\lambda})^{-\frac{\alpha_0}{2}}\sqrt{\mathcal{N}(\lambda)}}{\sqrt{n}}\right) \tag{3}$$

where $C_1 = 8\sqrt{2}$, $C_2 = 8\sigma$ and $\tilde{M} = L + (\mathcal{N}(\lambda) + 1)t$. By choosing $log(1/\lambda) \asymp \exp\{-Cn^{\frac{2}{2\alpha_0+d}}\}$ and applying Lemma C.6, we have

- for the second term in (3),
$$\frac{C_2\sqrt{\mathcal{N}(\lambda)}}{\sqrt{n}} \asymp n^{-\frac{\alpha_0}{2\alpha_0+d}}.$$
- for the third term,
$$\frac{C_1 log(\frac{1}{\lambda})^{-\frac{\alpha_0}{2}}\sqrt{\mathcal{N}(\lambda)}}{\sqrt{n}} \lesssim \frac{C_2\sqrt{\mathcal{N}(\lambda)}}{\sqrt{n}} \asymp n^{-\frac{\alpha_0}{2\alpha_0+d}}.$$
- for the first term,
$$\frac{C_1\sqrt{\mathcal{N}(\lambda)}}{n}\tilde{M} \leq \frac{C_1 L\sqrt{\mathcal{N}(\lambda)}}{n} + \frac{C_1 t\mathcal{N}(\lambda)^{\frac{3}{2}}}{n} \lesssim n^{-\frac{\alpha_0}{2\alpha_0+d}} \quad \text{given} \quad t \leq n^{\frac{2\alpha_0-d}{2(2\alpha_0+d)}}.$$

Combining all facts, if $t \leq n^{\frac{2\alpha_0-d}{2(2\alpha_0+d)}}$, with probability $1-\delta$ we have
$$I_1 \leq Cln(\frac{2}{\delta})n^{-\frac{\alpha_0}{2\alpha_0+d}}.$$

For I_2, we have
$$\begin{aligned}\tau_n := P\left(I_2 > \frac{\sqrt{\mathcal{N}(\lambda)}}{\sqrt{n}}\right) &\leq P\left(\exists x_i \text{ s.t. } x_i \in \Omega_2\right) \\ &= 1 - P\left(x \notin \Omega_2\right)^n \\ &= 1 - P\left(|f_0(x)| \leq t\right)^n \\ &\leq 1 - \left(1 - \frac{(C_q)^q}{t^q}\right)^n.\end{aligned}$$

Letting $\tau_n \to 0$ leading $t \gg n^{\frac{1}{q}}$. That is to say, if $t \gg n^{\frac{1}{q}}$ holds, we have $\tau_n = P\left(I_2 > I_1\right) \to 0$.

For I_3, we have
$$\begin{aligned}I_3 &\leq \mathrm{E}\left\|\xi_x I_{x\in\Omega_2}\right\|_{\mathcal{H}_K} \\ &\leq \mathrm{E}\left[\left\|(T_K + \lambda I)^{-\frac{1}{2}}K(x,\cdot)\right\|_{\mathcal{H}_K}\left|\left(y - f_0(x)\right)I_{x\in\Omega_2}\right|\right] \\ &\leq E_K^2\mathcal{N}(\lambda)\,\mathrm{E}\left|\left(y - f_0(x)\right)I_{x\in\Omega_2}\right| \\ &\leq E_K^2\mathcal{N}(\lambda)\left(\mathrm{E}\left|\left(f_\lambda - f_0(x)\right)I_{x\in\Omega_2}\right| + \mathrm{E}\left|\epsilon I_{x\in\Omega_2}\right|\right)\end{aligned}$$

Using Cauchy-Schwarz inequality yields
$$\mathrm{E}\left|\left(f_\lambda - f_0(x)\right)I_{x\in\Omega_2}\right| \leq \left\|f_0 - f_\lambda\right\|_{L_2} P(x \in \Omega_2) \leq log(\frac{1}{\lambda})^{-\frac{\alpha_0}{2}}(C_q)^q t^{-q}$$

In addition, we have
$$\mathrm{E}\left|\epsilon I_{x\in\Omega_2}\right| \leq \sigma\,\mathrm{E}\left|I_{x\in\Omega_2}\right| \leq \sigma(C_q)^q t^{-q}.$$

Together, we have
$$I_3 \leq log(\frac{1}{\lambda})^{-\frac{\alpha_0}{2}}(C_q)^q t^{-q} + \sigma(C_q)^q t^{-q}.$$

Notice if we pick $q \geq \frac{2(2\alpha_0+d)}{2\alpha_0-d}$, there exist t such that with probability $1-\delta-\tau_n$, we have
$$I_1 + I_2 + I_3 \leq Cln(\frac{2}{\delta})n^{-\frac{\alpha_0}{2\alpha_0+d}}.$$

For fixed δ, as $n \to \infty$, τ_n is sufficiently small such that $\tau_n = o(\delta)$, therefore without loss of generality, we can say with probability $1-\delta-\tau_n$, we have
$$I_1 + I_2 + I_3 \leq Cln(\frac{2}{\delta})n^{-\frac{\alpha_0}{2\alpha_0+d}}.$$

□

Proposition C.5. *Under the same conditions as the Proposition, we have*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \xi_i I_{x_i\in\Omega_1} - \mathrm{E}\,\xi_x I_{x\in\Omega_1} \right\|_{\mathcal{H}_K} \le \log(\frac{2}{\delta}) \left(\frac{C_1\sqrt{\mathcal{N}(\lambda)}}{n}\tilde{M} + \frac{C_2\sqrt{\mathcal{N}(\lambda)}}{\sqrt{n}} + \frac{C_1 log(\frac{1}{\lambda})^{-\frac{\alpha_0}{2}}\sqrt{\mathcal{N}(\lambda)}}{\sqrt{n}} \right)$$

where $C_1 = 8\sqrt{2}$, $C_2 = 8\sigma$ *and* $\tilde{M} = L + (\mathcal{N}(\lambda)+1)t$.

Proof. In order to use Bernstein inequality (Lemma C.9), we first bound the m-th moment of $\xi_x I_{x\in\Omega_1}$.

$$\begin{aligned}\mathrm{E}\left\|\xi_x I_{x\in\Omega_1}\right\|_{\mathcal{H}_K}^m &= \mathrm{E}\left\|(T_K+\lambda I)^{-\frac{1}{2}} K_x\left(y - f_\lambda(x)\right) I_{x\in\Omega_1}\right\|_{\mathcal{H}_K}^m \\ &\le \mathrm{E}\left(\left\|(T_K+\lambda I)^{-\frac{1}{2}} K(x,\cdot)\right\|_{\mathcal{H}_K}^m \mathrm{E}\left(\left|\left(y-f_\lambda(x)\right) I_{x\in\Omega_1}\right|^m \mid x\right)\right).\end{aligned}$$

Using the inequality $(a+b)^m \le 2^{m-1}\left(a^m + b^m\right)$, we have

$$\begin{aligned}\left|y - f_\lambda(x)\right|^m &\le 2^{m-1}\left(\left|f_\lambda(x) - f_\rho^*(x)\right|^m + \left|f_\rho^*(x) - y\right|^m\right) \\ &= 2^{m-1}\left(\left|f_\lambda(x) - f_\rho^*(x)\right|^m + |\epsilon|^m\right).\end{aligned}$$

Combining the inequalities, we have

$$\begin{aligned}\mathrm{E}\left\|\xi_x I_{x\in\Omega_1}\right\|_{\mathcal{H}_K}^m \le\ & \underbrace{2^{m-1}\,\mathrm{E}\left(\left\|(T_K+\lambda I)^{-\frac{1}{2}} K(x,\cdot)\right\|_{\mathcal{H}_K}^m \left|\left(f_\lambda(x) - f_\rho^*(x)\right) I_{x\in\Omega_1}\right|^m\right)}_{B_1} \\ & \underbrace{+\, 2^{m-1}\,\mathrm{E}\left(\left\|(T_K+\lambda I)^{-\frac{1}{2}} K(x,\cdot)\right\|_{\mathcal{H}_K}^m \mathrm{E}\left(\left|\epsilon I_{x\in\Omega_1}\right|^m \mid x\right)\right)}_{B_2}.\end{aligned}$$

We first focus on B_2, by Lemma C.7, we have

$$\mathrm{E}\left\|(T_K+\lambda I)^{-\frac{1}{2}} K(x,\cdot)\right\|_{\mathcal{H}_K}^m \le \left(E_K^2\mathcal{N}(\lambda)\right)^{\frac{m}{2}}.$$

By the error moment assumption, we have

$$\mathrm{E}\left(\left|\epsilon I_{x\in\Omega_1}\right|^m \mid x\right) \le \mathrm{E}\left(|\epsilon|^m \mid x\right) \le \frac{1}{2}m!\sigma^2 L^{m-2},$$

together, we have

$$B_2 \le \frac{1}{2}m!\left(\sqrt{2}\sigma\sqrt{\mathcal{N}(\lambda)}\right)^2 \left(2L\mathcal{N}(\lambda)\right)^{m-2}. \tag{4}$$

Turning to bounding B_1, we first have

$$\begin{aligned}\left\|(f_\lambda - f_0)I_{x\in\Omega_1}\right\|_{L_\infty} &\le \left\|f_\lambda I_{x\in\Omega_1}\right\|_{L_\infty} + \left\|f_0 I_{x\in\Omega_1}\right\|_{L_\infty} \\ &\le \left\|(T_K+\lambda I)^{-1}T_K(f_0)I_{x\in\Omega_1}\right\|_{L_\infty} + \left\|f_0 I_{x\in\Omega_1}\right\|_{L_\infty} \\ &\le \left(\left\|(T_K+\lambda I)^{-1}T_K\right\|_{op} + 1\right)\left\|f_0 I_{x\in\Omega_1}\right\|_{L_\infty} \\ &\le \left(\mathcal{N}(\lambda)+1\right)t := M.\end{aligned}$$

With bounds on approximation error, we get the upper bound for B_1 as

$$\begin{aligned}B_1 &\le 2^{m-1}\mathcal{N}(\lambda)^{\frac{m}{2}}\left\|(f_\lambda - f_0)I_{x\in\Omega_1}\right\|_{L_\infty}^{m-2}\left\|(f_\lambda - f_0)I_{x\in\Omega_1}\right\|_{L_2}^2 \\ &\le 2^{m-1}\mathcal{N}(\lambda)^{\frac{m}{2}}M^{m-2}log(\frac{1}{\lambda})^{-\alpha_0} \\ &\le \frac{1}{2}m!\left(2log(\frac{1}{\lambda})^{-\frac{\alpha_0}{2}}\sqrt{\mathcal{N}(\lambda)}\right)^2\left(2M\sqrt{\mathcal{N}(\lambda)}\right)^{m-2}.\end{aligned} \tag{5}$$

Denote

$$\tilde{L} = 2(L+M)\sqrt{\mathcal{N}(\lambda)}$$
$$\tilde{\sigma} = \sqrt{2}\sigma\sqrt{\mathcal{N}(\lambda)} + 2log(\frac{1}{\lambda})^{-\frac{\alpha_0}{2}}\sqrt{\mathcal{N}(\lambda)}$$

and combine the upper bounds for B_1 and B_2, i.e. (5) and (4), then we have

$$\mathrm{E}\left\|\xi_x I_{x\in\Omega_1}\right\|_{\mathcal{H}_K}^m \leq \frac{1}{2}m!\tilde{\sigma}^2\tilde{L}^{m-2}.$$

The proof is finished by applying Bernstein inequality i.e. Lemma C.9. □

C.2.5. Auxiliary Lemma

Lemma C.6. *If $s_j = C_1\exp(-C_2 j^2)$, by choosing $\log(1/\lambda) \asymp n^{\frac{2}{2\alpha_0+d}}$, we have*

$$\mathcal{N}(\lambda) = O\left(n^{\frac{d}{2\alpha_0+d}}\right)$$

Proof. For a positive integer $J \geq 1$

$$\begin{aligned}\mathcal{N}(\lambda) &= \sum_{j=1}^{J}\frac{s_j}{s_j+\lambda} + \sum_{j=J+1}^{\infty}\frac{s_j}{s_j+\lambda}\\ &\leq J + \sum_{j=J+1}^{\infty}\frac{s_j}{s_j+\lambda}\\ &\leq J + \frac{C_1}{\lambda}\int_J^\infty exp\{-C_2x^2\}dx\\ &\leq J + \frac{1}{\lambda}\frac{C_1 exp\{-C_2J^2\}}{2C_2J}\end{aligned}$$

where we use the fact that the eigenvalue of the Gaussian kernel decays at an exponential rate, i.e. $s_j \leq C_1\exp\{-C_2j^2\}$ and the inequality

$$\int_x^\infty exp\{\frac{-t^2}{2}\}dt \leq \int_x^\infty \frac{t}{x}exp\{\frac{-t^2}{2}\}dt \leq \frac{exp\{-\frac{x^2}{2}\}}{x}.$$

Then select $J = \lfloor n^{\frac{d}{2\alpha_0+d}}\rfloor$ and $\lambda = exp\{-C^{'}n^{\frac{2}{2\alpha_0+d}}\}$ with $C' \leq C_2$ leads to

$$\mathcal{N}(\lambda) = O\left(n^{\frac{d}{2\alpha_0+d}}\right).$$

□

Lemma C.7. *For μ-almost $x \in \mathcal{X}$, we have*

$$\left\|(T_K+\lambda I)^{-\frac{1}{2}}K(x,\cdot)\right\|_{\mathcal{H}_K}^2 \leq E_K^2\mathcal{N}(\lambda), \quad \textit{and} \quad \mathrm{E}\left\|(T_K+\lambda I)^{-\frac{1}{2}}K(x,\cdot)\right\|_{\mathcal{H}_K}^2 \leq \mathcal{N}(\lambda).$$

For some constant E_K. Consequently, we also have

$$\left\|(T_K+\lambda I)^{-\frac{1}{2}}T_{K,x}(T_K+\lambda I)^{-\frac{1}{2}}\right\|_{op} \leq E_K^2\mathcal{N}(\lambda).$$

Proof. We first state a fact on the Gaussian kernel. If K is a Gaussian kernel function with fixed bandwidth, then there exists a constant E_K such that the eigenfunction of K is uniformly bounded for all $j \geq 1$, i.e. $\sup_{j\geq1}\|e_j\|_{L_\infty} \leq E_K$. This is indeed the so-called "uniformly bounded eigenfunction" assumption that usually appears in nonparametric regression literature, especially for those who consider misspecified kernel in KRR, see Mendelson & Neeman (2010); Wang & Jing (2022). Based on the explicit construction of the RKHS associated with the Gaussian kernel (Steinwart et al., 2006), we know the uniformly bounded eigenfunction holds for the Gaussian kernel.

Based on the fact of uniformly bounded eigenfunction, we know $e_j^2(x) \le E_K^2$ for all $x \in \mathcal{X}$ and $j \ge 1$. Then, we prove the first inequality by the following procedure,

$$\begin{aligned}\left\| (T_K + \lambda I)^{-\frac{1}{2}} K(x,\cdot) \right\|_{\mathcal{H}_K}^2 &= \left\| \sum_{j=1} \frac{1}{\sqrt{s_j + \lambda}} s_j e_j(x) e_j(\cdot) \right\|_{\mathcal{H}_K}^2 \\ &= \sum_{j=1}^{\infty} \frac{s_j}{s_j + \lambda} e_j^2(x) \\ &\le E_K^2 \mathcal{N}(\lambda).\end{aligned}$$

The second inequality follows given the fact that $\operatorname{E} e_j^2(x) = 1$. The third inequality comes from the observation that for any $f \in \mathcal{H}_K$

$$(T_K + \lambda I)^{-\frac{1}{2}} T_{K,x} (T_K + \lambda I)^{-\frac{1}{2}} (f) = \left\langle (T_K + \lambda I)^{-\frac{1}{2}} K(x,\cdot), f \right\rangle_{\mathcal{H}_K} (T_K + \lambda I)^{-\frac{1}{2}} K(x,\cdot)$$

and

$$\begin{aligned}\left\| (T_K + \lambda I)^{-\frac{1}{2}} T_{K,x} (T_K + \lambda I)^{-\frac{1}{2}} \right\|_{op} &= \sup_{\|f\|_{\mathcal{H}_k}=1} \| (T_K + \lambda I)^{-\frac{1}{2}} T_{K,x} (T_K + \lambda I)^{-\frac{1}{2}} (f) \|_{\mathcal{H}_K} \\ &= \left\| (T_K + \lambda I)^{-\frac{1}{2}} K(x,\cdot) \right\|_{\mathcal{H}_K}^2\end{aligned}$$

□

The following lemma provides the concentration inequality about self-adjoint Hilbert-Schmidt operator-valued random variables, which is widely used in related kernel method literature, e.g., Theorem 27 in Fischer & Steinwart (2020), Lemma 26 in Lin & Cevher (2020) and Lemma E.3 in Zhang et al. (2023).

Lemma C.8. *(Lemma E.3 in Zhang et al. (2023)) Let $(\mathcal{X}, \mathcal{B}, \mu)$ be a probability space, and $\mathcal{H}$ be a separable Hilbert space. Suppose $A_1, \cdots, A_n$ are i.i.d. random variables whose values in the set of self-adjoint Hilbert-Schmidt operators. If $\operatorname{E} A_i = 0$ and the operator norm $\|A_i\| \le L$ μ-a.e. $x \in \mathcal{X}$, and there exists a self-adjoint positive semi-definite trace class operator V with $\operatorname{E} A_i^2 \preceq V$. Then for $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} A_i \right\| \le \frac{2L\beta}{3n} + \sqrt{\frac{2\|V\|\beta}{n}},$$

where $\beta = \log(4tr(V)/\delta\|V\|)$.

Lemma C.9. *(Bernstein inequality) Let $(\Omega, \mathcal{B}, P)$ be a probability space, H be a separable Hilbert space, and $\xi : \Omega \to H$ be a random variable with*

$$\operatorname{E} \|\xi\|_H^m \le \frac{1}{2} m! \sigma^2 L^{m-2}$$

for all $m > 2$. Then for $\delta \in (0,1)$, ξ_i are i.i.d. random variables, with probability at least $1-\delta$, we have

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i - \operatorname{E} \xi \right\|_H \le 4\sqrt{2} \log\left(\frac{2}{\delta}\right) \left(\frac{L}{n} + \frac{\sigma}{\sqrt{n}}\right)$$

C.3. Proof of Adaptive Rate (Theorem 3.3)

Proof. To simplify the notation, for a given smoothness α and sample size n, we define

$$\psi_n(\alpha) = \left(\frac{n}{\log n}\right)^{-\frac{2\alpha}{2\alpha+d}}.$$

First, we show that it is sufficient to consider the true Sobolev space α in $\mathcal{A} = \{\alpha_1, \cdots, \alpha_N\}$ with $\alpha_j - \alpha_{j-1} \asymp 1/\log n$. If $\alpha_0 \in (\alpha_{j-1}, \alpha_j)$, then $H^{\alpha_j} \subset H^{\alpha_0} \subset H^{\alpha_{j-1}}$. Therefore, since $\psi_n(\alpha_0)$ is squeezed between $\psi_n(\alpha_{j-1})$ and $\psi_n(\alpha_j)$, we just need to show $\psi_n(\alpha_{j-1}) \asymp \psi_n(\alpha_j)$. By the definition of $\psi_n(\alpha)$, the claim follows since

$$\log \frac{\psi_n(\alpha_{j-1})}{\psi_n(\alpha_j)} = \left(-\frac{2\alpha_{j-1}}{2\alpha_{j-1}+d} + \frac{2\alpha_j}{2\alpha_j+d}\right) \log \frac{n}{\log n} \asymp (\alpha_j - \alpha_{j-1}) \log n \asymp 1.$$

Therefore, we assume $f_0 \in H^{\alpha_i}$ where $i \in \{1, 2, \cdots, N\}$.

Let $m = \lfloor \frac{n}{2} + 1 \rfloor$, i.e. $m \geq \frac{n}{2}$, by Theorem 3.1, for some constants C that doesn't depend on n, we have

$$\mathcal{E}(\hat{f}_{\lambda_\alpha, \mathcal{D}_1}) \leq \left(\log \frac{4}{\delta}\right)^2 (\mathrm{E}(\lambda_\alpha, m) + \mathrm{A}(\lambda_\alpha, m)) \tag{6}$$

for all $\alpha \in \mathcal{A}$ simultaneously with probability at least $1 - N\delta$. Here, $\mathrm{E}(\lambda, n)$ and $\mathrm{A}(\lambda, n)$ denote the estimation and approximation error that depends on the regularization parameter λ and sample size n in non-adaptive rate proof.

Furthermore, by Theorem 7.2 in Steinwart & Christmann (2008) and Assumption 2.5, we have

$$\begin{aligned} \mathcal{E}(\hat{f}_{\lambda_{\hat{\alpha}}}) &< 6 \left(\inf_{\alpha \in \mathcal{A}} \mathcal{E}(\hat{f}_{\lambda_\alpha})\right) + \frac{128\sigma^2 L^2 \left(\log \frac{1}{\delta} + \log(1+N)\right)}{n-m} \\ &< 6 \left(\inf_{\alpha \in \mathcal{A}} \mathcal{E}(\hat{f}_{\lambda_\alpha})\right) + \frac{512\sigma^2 L^2 \left(\log \frac{1}{\delta} + \log(1+N)\right)}{n} \end{aligned} \tag{7}$$

with probability $1 - \delta$, where the last inequality is based on the fact that $n - m \geq \frac{n}{2} - 1 \geq \frac{n}{4}$.

Combining (6) and (7), we have

$$\begin{aligned} \mathcal{E}(\hat{f}_{\lambda_{\hat{\alpha}}}) &< 6 \left(\log \frac{4}{\delta}\right)^2 \left(\inf_{\alpha \in \mathcal{A}} \mathrm{E}(\lambda_\alpha, m) + \mathrm{A}(\lambda_\alpha, m)\right) + \frac{512\sigma^2 L^2 \left(\log \frac{1}{\delta} + \log(1+N)\right)}{n} \\ &\leq 6C \left(\log \frac{4}{\delta}\right)^2 m^{-\frac{2\alpha_0}{2\alpha_0+d}} + \frac{512\sigma^2 L^2 \left(\log \frac{1}{\delta} + \log(1+N)\right)}{n} \\ &\leq 12C \left(\log \frac{4}{\delta}\right)^2 n^{-\frac{2\alpha_0}{2\alpha_0+d}} + \frac{512\sigma^2 L^2 \left(\log \frac{1}{\delta} + \log(1+N)\right)}{n} \end{aligned}$$

with probability at least $1 - (1+N)\delta$. With a variable transformation, we have

$$\mathcal{E}(\hat{f}_{\lambda_{\hat{\alpha}}}) \leq 12C \left(\log \frac{4(1+N)}{\delta}\right)^2 n^{-\frac{2\alpha_0}{2\alpha_0+d}} + \frac{512\sigma^2 L^2 \left(\log \frac{1+N}{\delta} + \log(1+N)\right)}{n} \tag{8}$$

with probability $1 - \delta$. Therefore, for the first term

$$\begin{aligned} 12C \left(\log \frac{4(1+N)}{\delta}\right)^2 n^{-\frac{2\alpha_0}{2\alpha_0+d}} &\leq 24C \left\{\left(\log \frac{4}{\delta}\right)^2 \log^2(1+N) + 1\right\} n^{-\frac{2\alpha_0}{2\alpha_0+d}} \\ &\leq 24C' \left(\log \frac{4}{\delta}\right)^2 \left(\frac{n}{\log n}\right)^{-\frac{2\alpha_0}{2\alpha_0+d}} + 24Cn^{-\frac{2\alpha_0}{2\alpha_0+d}} \end{aligned} \tag{9}$$

where the first inequality is based on the fact that $a + b < ab + 1$ for $a, b > 1$, while the second inequality is based on the fact that $\log(x) \leq x^{\frac{\alpha_0}{2\alpha_0+d}}$ for some n such that $\log(\log n)/\log n < 1/4$. For the second term,

$$\begin{aligned} \frac{512\sigma^2 L^2 \left(\log \frac{1+N}{\delta} + \log(1+N)\right)}{n} &\leq \frac{512\sigma^2 L^2 \left(\log \frac{1}{\delta} + 1 + 2\log(1+N)\right)}{n} \\ &\leq \frac{512\sigma^2 L^2 \left(\log \frac{1}{\delta} + 1 + 2\log n\right)}{n} \end{aligned} \tag{10}$$

The proof is finished by combining (8), (9) and (10). □

C.4. Comparison to Previous Work

In Table 1, we compare our results with some state-of-the-art works (to the best of our knowledge) that consider general/Matérn misspecified kernels and Gaussian kernels in target-only setting KRR. For a detailed review of the optimality of misspecified KRR, we refer readers to Zhang et al. (2023).

Wang & Jing (2022) considered the true function lies in H^{α_0} while the imposed kernels are misspecified Matérn kernels. On the other hand, Zhang et al. (2023) considered the minimax optimality for misspecified KRR in general RKHS, i.e., the imposed kernel is K while the true function $f_0 \in [\mathcal{H}_K]^s$ for $s \in (0, 2]$. However, when the RKHS is specified as the Sobolev space, the results in both papers are equivalent by applying the real interpolation technique in Appendix B. Therefore, we place them in the same row. Unlike the necessary conditions that the imposed RKHS must fulfill $\alpha_0' > \alpha_0/2$ to achieve optimality (Wang & Jing, 2022; Zhang et al., 2023), our results circumvent this requirement, thereby being more robust. Compared to other works on Gaussian kernel-based KRR, our result shows that the optimality can be achieved only via a fixed bandwidth Gaussian kernel.

We would like to note that our statistical rates and Eberts & Steinwart (2013); Hamm & Steinwart (2021) might not be directly comparable. Our results are derived under bounded moment assumption, i.e., Assumption 2.5, while results of Eberts & Steinwart (2013); Hamm & Steinwart (2021) are derived from bounded response assumption, i.e., there exists a constant such that $Y \in [-M, M]$. Moreover, Hamm & Steinwart (2021) considers a broader space (Besov space) and both regression/classification problems, while whether these hold on fixed bandwidth settings is still unknown. Although with slightly different settings, the table highlights the difference in optimal order of λ (and γ) for fixed and variable bandwidth settings.

Table 1. Comparison of generalization error convergence rate (non-adaptive) between our result and the prior literature. Here, we assume the mean function f_0 belongs to Sobolev space H^{α_0}, imposed RKHS means the RKHS that $\hat{f}$ belongs to. "$-$" in column γ means the bandwidth is fixed during training and does not have an optimal order in n. $\mathcal{H}_K$ means the RKHS associated with the Gaussian kernel while $H^{\alpha_0'}$ means the Sobolev space with smoothness order α_0'.

Paper	Imposed RKHS	Rate	λ	γ
Wang & Jing (2022), Zhang et al. (2023)	$H^{\alpha_0'}, \alpha_0' > \frac{\alpha_0}{2}$	$n^{-\frac{2\alpha_0}{2\alpha_0+d}}$	$n^{-\frac{2\alpha_0'}{2\alpha_0+d}}$	$-$
Eberts & Steinwart (2013)	$\mathcal{H}_K$	$n^{-\frac{2\alpha_0}{2\alpha_0+d}+\eta}, \forall \eta > 0$	n^{-1}	$n^{-\frac{1}{2\alpha_0+d}}$
Hamm & Steinwart (2021)	$\mathcal{H}_K$	$n^{-\frac{2\alpha_0}{2\alpha_0+d}} \log^{d+1}(n)$	n^{-1}	$n^{-\frac{1}{2\alpha_0+d}}$
This work	$\mathcal{H}_K$	$n^{-\frac{2\alpha_0}{2\alpha_0+d}}$	$\exp\{-Cn^{\frac{2}{2\alpha_0+d}}\}$	$-$

D. Smoothness Adaptive Transfer Learning Results

D.1. Proof of Lower Bound

In this part, we proof the alternative version for the lower bound, i.e.

$$\inf_{\tilde{f}} \sup_{\Theta(h,m_0,m)} \mathbb{P}\left(\|\tilde{f} - f_T\|_{L_2}^2 \geq C\delta R^2 \left((n_S + n_T)^{-\frac{2m_0}{2m_0+d}} + n_T^{-\frac{m}{2m+d}} \xi_L \right) \right) \geq 1 - \delta$$

for some constant C that are independent of n_S, n_T, R, h and δ, and $\xi_L \propto h^2/R^2$.

This alternative form is also used to prove the lower bound in other transfer learning contexts like high-dimensional linear regression or GLM, see Li et al. (2022); Tian & Feng (2022). However, the upper bound we derive for SATL can still be sharp since in the transfer learning regime, it is always assumed $n_S \gg n_T$, and leads to $(n_S + n_T)^{-\frac{2m_0}{2m_0+d}} \asymp n_S^{-\frac{2m_0}{2m_0+d}}$.

On the other hand, one can modify the first phase in OTL by including the target dataset to obtain $\hat{f}_S$, which produces an alternative upper bound $(n_S + n_T)^{-\frac{2m_0}{2m_0+d}} + n_T^{-\frac{2m}{2m+d}} \xi_L$, and mathematically aligns with the alternative lower bound we mention above. However, we would like to note that such a modified OTL is not computationally efficient for transfer learning since for each new upcoming target task, OTL needs to recalculate a new $\hat{f}_S$ with the combination of the target and source datasets.

Note that any lower bound for a specific case will immediately yield a lower bound for the general case. Therefore, we consider the following two cases.

(1) Consider $h = 0$, i.e. both source and target data are drawn from ρ_T. In this case, the problem can be viewed as obtaining the lower bound for classical nonparametric regression with sample size $n_T + n_S$ and prediction function as $f_T \in H^{m_0}$ with $\|f_T\|_{H^{m_0}} \leq R$. Then using the Proposition D.1, we have

$$\inf_{\tilde{f}} \sup_{\Theta(h,m_0,m)} \mathbb{P}\left(\|\tilde{f} - f_T\|_{L_2}^2 \geq C_1 \delta R^2 (n_T + n_S)^{-\frac{2m_0}{2m_0+d}}\right) \geq 1 - \delta,$$

where C_1 is independent of δ, R, n_S and n_T.

(2) Consider $f_S = 0$, i.e. the source model has no similarity to f_T and all the information about f_T is stored in the target dataset. By the assumptions, we have $f_T \in \{f : f \in H^m, \|f\|_{H^m} \leq h\}$. Again, using the Proposition D.1, we have

$$\inf_{\tilde{f}} \sup_{\Theta(h,m_0,m)} \mathbb{P}\left(\|\tilde{f} - f_T\|_{L_2}^2 \geq C_2 \delta h^2 (n_T)^{-\frac{2m}{2m+d}}\right) \geq 1 - \delta,$$

where C_2 is independent of δ, h, and n_T.

Combining the lower bound in case (1) and case (2), we obtain the desired lower bound.

Discussion. Here, we prove the asymptotic lower bound as

$$\Omega_{\mathbb{P}}\left(R^2 (n_T + n_S)^{-\frac{2m_0}{2m_0+d}} + h^2 (n_T)^{-\frac{2m}{2m+d}}\right). \tag{11}$$

By factoring out the R^2, we obtain the form in Theorem 4.1, i.e.,

$$\Omega_{\mathbb{P}}\left((n_T + n_S)^{-\frac{2m_0}{2m_0+d}} + \xi_L (n_T)^{-\frac{2m}{2m+d}}\right), \tag{12}$$

where the constant should be proportional to R^2. We present the results in form (12) instead of the form (11) as we would like to emphasize how the transfer dynamic and efficacy depend on both h and R, compared to the form (1) in most existing OTL works. The constant ξ_U in the upper bound is designed in the same philosophy.

D.2. Proof of Upper Bound

The final estimator for target regression function is $\hat{f}_T = \hat{f}_S + \hat{f}_\delta$. By triangle inequality

$$\left\|\hat{f}_T - f_T\right\|_{L_2} \leq \left\|\hat{f}_S - f_S\right\|_{L_2} + \left\|\hat{f}_\delta - f_\delta\right\|_{L_2} \tag{13}$$

For the first term in the r.h.s. of (13), since the marginal distribution over $\mathcal{X}$ are equivalent for both target and source, applying Theorem 3.3 directly leads to with probability at least $1 - \delta$

$$\left\|\hat{f}_S - f_S\right\|_{L_2} \leq C \left(\log \frac{4}{\delta}\right) \cdot \left(\frac{n_S}{\log n_S}\right)^{-\frac{m_0}{2m_0+d}},$$

where C is independent of n_S and δ, and proportional to $\sqrt{\sigma_S^2 + \|f_S\|_{H^{m_0}}^2}$ and thus $\sqrt{\sigma_S^2 + R^2}$.

For the second term, we modify the proof of Theorem 1 with the same logic. Note that the estimated offset function has the following expression

$$
\begin{aligned}
\hat{f}_\delta &= (T_{K,n_T} + \lambda_2 \mathbf{I})^{-1} \left(\frac{1}{n_T} \sum_{i=1}^{n_T} K_{x_{T,i}} \left(y_{T,i} - \hat{f}_{\mathcal{S}}(x_{T,i}) \right) \right) \\
&= (T_{K,n_T} + \lambda_2 \mathbf{I})^{-1} \left(\frac{1}{n_T} \sum_{i=1}^{n_T} K_{x_{T,i}} \left(f_{\mathcal{S}}(x_{T,i}) - \hat{f}_{\mathcal{S}}(x_{T,i}) + f_\delta(x_{T,i}) + \epsilon_{T,i} \right) \right) \\
&= \underbrace{(T_{K,n_T} + \lambda_2 \mathbf{I})^{-1} \left(\frac{1}{n_T} \sum_{i=1}^{n_T} K_{x_{T,i}} \left(f_{\mathcal{S}}(x_{T,i}) - \hat{f}_{\mathcal{S}}(x_{T,i}) \right) \right)}_{\hat{f}_{\delta 1}} \\
&\quad + \underbrace{(T_{K,n_T} + \lambda_2 \mathbf{I})^{-1} \left(\frac{1}{n_T} \sum_{i=1}^{n_T} K_{x_{T,i}} \left(f_\delta(x_{T,i}) + \epsilon_{T,i} \right) \right)}_{\hat{f}_{\delta 2}}.
\end{aligned}
$$

We decompose the estimated offset function $\hat{f}_\delta$ into $\hat{f}_{\delta 1}$ and $\hat{f}_{\delta 2}$ since in the second phase, we are using estimated source function $\hat{f}_{\mathcal{S}}$ to generate pseudo label instead of the true source function. Therefore, the term $\hat{f}_{\delta 1}$ counts the introduced bias of using the estimated version of $f_{\mathcal{S}}$. By triangle inequality, one has

$$
\left\| \hat{f}_\delta - f_\delta \right\|_{L_2} \leq \underbrace{\left\| \hat{f}_{\delta 1} \right\|_{L_2}}_{D_1} + \underbrace{\left\| \hat{f}_{\delta 2} - f_\delta \right\|_{L_2}}_{D_2}.
$$

For D_2, since the label is observed from true offset function f_δ with noise, we can directly apply Theorem 3.3. Define the intermediate term $f_{\delta,\lambda}$ as

$$
f_{\delta,\lambda} = (T_K + \lambda_2 \mathbf{I})^{-1} (T_K(f_\delta))
$$

To control the approximation error, using Proposition C.2 with the fact that $\|f_\delta\|_{H^m} \leq h$, we have

$$
\|f_{\delta,\lambda} - f_\delta\|_{L_2}^2 \leq log(\frac{1}{\lambda_2})^{-m} h^2.
$$

For estimation error, the same proof the same proof procedure of Theorem C.1 can be directly applied. Finally, by applying Theorem 3.3, we have

$$
D_2 = \left\| \hat{f}_{\delta 2} - f_\delta \right\|_{L_2} \leq C \left(\log \frac{4}{\delta} \right) \cdot \left(\frac{n_T}{\log n_T} \right)^{-\frac{m}{2m+d}}.
$$

where C is independent of n_T and δ, and proportional to $\sqrt{\sigma_T^2 + \|f_\delta\|_{H^m}^2}$ and thus $\sqrt{\sigma_T^2 + h^2}$.

Turning to D_1,

$$
\begin{aligned}
\left\| \hat{f}_{\delta 1} \right\|_{L_2} &= \left\| (T_{K,n_T} + \lambda_2 \mathbf{I})^{-1} \left(T_{K,n_T} \left(f_{\mathcal{S}} - \hat{f}_{\mathcal{S}} \right) \right) \right\|_{L_2} \\
&\leq \left\| (T_{K,n_T} + \lambda_2 \mathbf{I})^{-1} T_{K,n_T} \right\|_{op} \left\| f_{\mathcal{S}} - \hat{f}_{\mathcal{S}} \right\|_{L_2} \\
&\leq \left\| f_{\mathcal{S}} - \hat{f}_{\mathcal{S}} \right\|_{L_2}.
\end{aligned}
$$

For the second inequality, we used the fact that the upper bound of the largest eigenvalue of $(T_{K,n_T} + \lambda_2 \mathbf{I})^{-1} T_{K,n_T}$ is bounded by 1.

Finally, the proof is finished by combining the result of $\hat{f}_S$ and $\hat{f}_\delta$ and noticing the results hold with probability at least $(1-\delta) \cdot (1-\delta) \geq 1 - 2\delta$.

D.3. Propositions

Proposition D.1 (Lower bound for target-only KRR)**.** *In target-only KRR problem, suppose the observed data are* $\{(x_i, y_i)\}_{i=1}^n$ *and the underlying true function* $f_0 \in \{f \in H^{m_0} : \|f\|_{H^{m_0}} \leq R\} := \mathcal{B}_{m_0}(R)$*. Then, when* n *is sufficiently* *large, one has*

$$\inf_{\tilde{f}} \sup_{f \in \mathcal{B}_{m_0}(R)} \mathbb{P}\left(\|\tilde{f} - f\|_{L_2}^2 \geq C\delta n^{-\frac{2m_0}{2m_0+d}} \right) \geq 1 - \delta,$$

where the constant C is proportional to R^2 and independent of δ and n.

Remark D.2. The proof for the lower bound in target-only KRR is standard and can be found in many works. However, while most of the works omit the property of constant C, we prove it is proportional to R^2.

Proof. For every $f \in \mathcal{B}_{m_0}(R)$, define the probability distribution P_f on $\mathcal{X} \times \mathcal{Y}$ so that $y = f(x) + \epsilon$, where $\epsilon \sim N(0, \bar{\sigma}^2)$ and $\bar{\sigma} = min(\sigma, L)$. Such form of $\bar{\sigma}$ ensures the Assumption 2.5 holds.

We construct a series function, term $f_0, f_1, \cdots, f_N \in \mathcal{B}_{m_0}(R)$, as follows,

$$f_i = \frac{C_0}{\sqrt{M}} \sum_{k=M+1}^{2M} \theta_k^{(i)} T_K^{\frac{1}{2}}(\phi_k).$$

Here, K denotes the reproducing kernel of H^{m_0} and ϕ_k represents the k-th eigenfunction. We set M the smallest integer great than $n^{\frac{d}{2m_0+d}}$, and $\theta^{(0)}, \theta^{(1)}, \cdots, \theta^{(N)} \in \{0,1\}^M$ for some $N \geq 2^{M/8}$ such that the for all $0 \leq i \leq j \leq N$, the Hamming distance between $\theta^{(i)}$ and $\theta^{(j)}$ greater than $M/8$. One can then verify $f_i \in \mathcal{B}_{m_0}(R)$

$$\|f_i\|_{H^{m_0}}^2 = \sum_{i=N+1}^{2N} \frac{(\theta_{k-N}^{(i)} C_0)^2}{N} \left\| T_K^{\frac{1}{2}}(\phi_i) \right\|_{H^{m_0}}^2 \leq R^2,$$

by having $C_0 \leq R$.

Denote s_j as the j-th eigenvalue of K and by the properties of Sobolev space, we have

$$C_1 j^{-\frac{2m_0}{d}} \leq s_j \leq C_2 j^{-\frac{2m_0}{d}}, \quad \forall j \geq 1,$$

where C_1 and C_2 are some constants.

With Lemma D.4,

$$\begin{aligned} KL(P_{f_i}^n, P_{f_j}^n) &= \frac{n}{2\bar{\sigma}^2} \|f_i - f_j\|_{L_2} \\ &\leq \frac{n}{2\bar{\sigma}^2} \frac{C_0^2}{M} s_M H(\theta^{(i)}, \theta^{(j)}) \\ &\leq \frac{n}{2\bar{\sigma}^2} C_0^2 s_M \\ &\leq C_2 C_0 \frac{n}{2\bar{\sigma}^2} M^{-\frac{2m_0}{d}}, \end{aligned}$$

where the first inequality use the fact that $s_M \geq \cdots, s_{2M}$, the second inequality based on the fact that $\theta^{(i)}$ and $\theta^{(j)}$ are elements in $\{0,1\}^M$, and the third inequality based on the property of the eigenvalue of K. Notice we take M the smallest integer great than $n^{\frac{d}{2m_0+d}}$, for a fixed $a \in (0, 1/8)$, letting

$$KL(P_{f_i}^n, P_{f_j}^n) \leq C_2 C_0^2 \frac{n}{2\bar{\sigma}^2} M^{-\frac{2m_0}{d}} \leq C_2 \frac{C_0^2}{2\bar{\sigma}^2} M \leq a \frac{\log 2}{8} M \leq a \log N,$$

leads to

$$\frac{C_2 C_0^2}{\bar{\sigma}^2} \leq a \frac{\log 2}{4}.$$

Besides,

$$
\begin{aligned}
\|f_i - f_j\|_{L_2}^2 &= \frac{C_0^2}{M} \sum_{k=M+1}^{2M} \left(\theta_k^{(i)} - \theta_k^{(j)}\right)^2 s_k \\
&\geq \frac{C_0^2}{M} s_{2M} H\left(\theta_k^{(i)}, \theta_k^{(j)}\right) \\
&\geq \frac{C_0^2}{8} s_{2M} \\
&\geq \frac{C_0^2 C_1}{8} (2M)^{-\frac{2m_0}{d}} \\
&\geq \frac{C_0^2 C_1}{8} 2^{-\frac{2m_0}{d}} n^{-\frac{2m_0}{2m_0+d}},
\end{aligned}
$$

where the second inequality is based on Lemma D.5.

To use Lemma D.3, we set $C_0 = \alpha R\sqrt{a}$ for some positive α, thus for any fixed R and $a \in (0, 1/8)$, we can choose an α such that

$$
C_0 \leq R \quad \text{and} \quad C_0^2 \leq \frac{\bar{\sigma}^2}{4} \frac{\log 2}{C_2} a
$$

are satisfied. Therefore, by applying Lemma D.3,w e have

$$
\inf_{\tilde{f}} \sup_{f \in \mathcal{B}_{m_0}(R)} \mathbb{P}\left(\|\tilde{f} - f\|_{L_2}^2 \geq C' R^2 a n^{-\frac{2m_0}{2m_0+d}}\right) \geq \frac{\sqrt{N}}{1+\sqrt{N}} \left(1 - 2a - \sqrt{\frac{2a}{\log N}}\right),
$$

where $C' = (\alpha^2 2^{-\frac{2m_0}{d}} C_1)/8$. When n is sufficiently large so that N is sufficiently large, the L.H.S. of the above inequality holds a probability greater than $1 - 3a$. For $\delta \in (0, 1)$, choose $a = \delta/3$ completes the proof. □

D.4. Lemma

We provide some Lemma that are important for proving the lower bound. All these lemmas are standard for proving the lower bound and can be found in extensive works.

Lemma D.3. *Suppose that there is a non-parametric class of functions Θ and a (semi-)distance $d(\cdot, \cdot)$ on Θ. $\{P_\theta, \theta \in \Theta\}$ is a family of probability distributions indexed by Θ. Assume that $N \geq 2$ and suppose that Θ contains elements $\theta_0, \theta_1, \cdots, \theta_N$ such that,*

(1) $d(\theta_j, \theta_k) \geq 2s > 0, \quad \forall 0 \leq j < k \leq N$*;*

(2) $P_j \ll P_0, \quad \forall j = 1, \cdots, N$*, and*

$$
\frac{1}{N} \sum_{j=1}^{N} K(P_j, P_0) \leq a \log N,
$$

with $0 < a < 1/8$ and $P_j = P_{\theta_j}, j = 0, 1, \cdots, N$. Then

$$
\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_\theta(d(\hat{\theta}, \theta) \geq s) \geq \frac{\sqrt{N}}{1+\sqrt{N}} \left(1 - 2a - \sqrt{\frac{2a}{\log N}}\right).
$$

Lemma D.4. *Suppose that μ is a distribution on $\mathcal{X}$ and $f_i \in L_2(\mathcal{X}, \mu)$. Suppose that*

$$
y = f_i(x) + \epsilon, \quad i = 1, 2,
$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ are independent Gaussian random error. Denote the two corresponding distributions on $\mathcal{X} \times \mathcal{Y}$ as $\rho_i, i = 1, 2$. The KL divergence of two probability distributions on Ω is

$$
K(P_1, P_2) := \int_\Omega \log\left(\frac{\mathrm{d}P_1}{\mathrm{d}P_2}\right) \mathrm{d}P_1,
$$

if $P_1 \ll P_2$ and otherwise $K(P_1, P_2) \coloneqq \infty$. Then we have

$$\mathrm{KL}\left(\rho_1^n, \rho_2^n\right) = n\mathrm{KL}\left(\rho_1, \rho_2\right) = \frac{n}{2\sigma^2} \left\| f_1 - f_2 \right\|_{L^2(\mathcal{X}, d\mu)}^2,$$

where ρ_i^n denotes the independent product of n distributions $\rho_i, i = 1, 2$.

Lemma D.5 (Varshamov-Gilbert bound (Varshamov, 1957))**.** *Denote $\Omega = \{\omega = (\omega_1, \cdots, \omega_M), \omega_i \in \{0, 1\}\} = \{0, 1\}^M$. Let $m \geq 8$, there exists a subset $\left\{\omega^{(0)}, \cdots, \omega^{(M)}\right\}$ of Ω such that $\omega^{(0)} = (0, \cdots, 0)$,*

$$H\left(\omega^{(i)}, \omega^{(j)}\right) \coloneqq \sum_{k=1}^{M} \left|\omega_k^{(i)} - \omega_k^{(j)}\right| \geq \frac{M}{8}, \quad \forall 0 \leq i < j \leq N,$$

and $N \geq 2^{M/8}$. Here $H\left(\omega^{(i)}, \omega^{(j)}\right)$ is the Hamming distance between $\omega^{(i)}$ and $\omega^{(j)}$.

E. Additional Simulation Results

E.1. Additional Results for Target-Only KRR with Gaussian kernels

In Section 5.1, we only present the best lines with the optimal C. In this part, we report the generalization error decay curve for different C. We report the results in Figure 4. Each subfigure contains 7 different lines that center around the optimal line. One can see even with different C, the empirical error decay curves are still aligned with the theoretical ones.

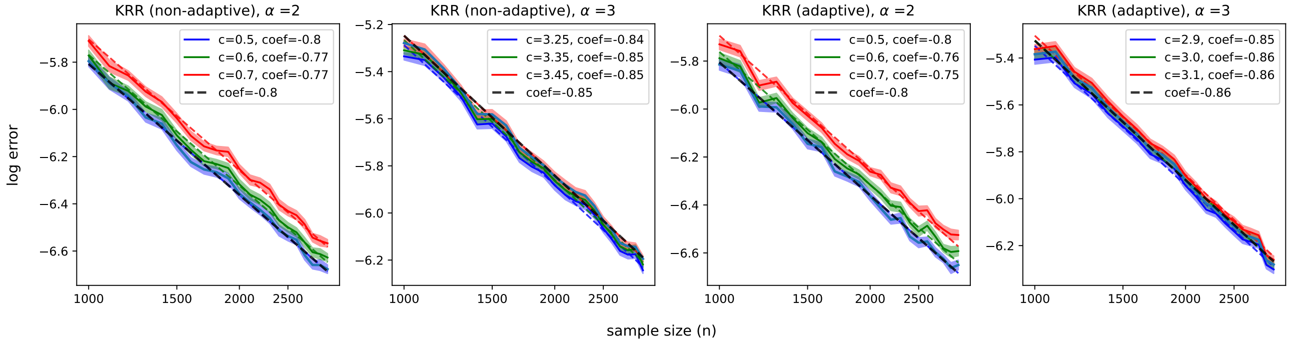


Figure 4. Error decay curves of target-only KRR based on Gaussian kernel, both axes are in log scale. The curves with different colors correspond to different C and denote the average logarithmic generalization errors over 100 trials. The dashed black lines denote the theoretical decay rates.

E.2. Additional Details for TL Algorithm Comparison

Implementation of Finite Basis Expansion: Denote the finite basis estimator (FBE) for a regression function as

$$\hat{f}_M(x) = \sum_{j=1}^{M} \beta_j B_j(x)$$

where B_j are given finite basis or spline functions with a different order, and M denotes the truncation number, which generally controls the variance-bias trade-off of the estimator. Then, the transfer learning procedure proposed in Wang et al. (2016) can be summarized by the following 4 steps.

1. Estimate f_S using the FBE and source data, output $\hat{f}_{S,M_1}$
2. Produce the pseudo label $\hat{y}_{T,i}$ using $\hat{f}_{S,M_1}$ and $x_{T,i}$, obtain the offset estimation as $y_{T,i} - \hat{y}_{T,i}$.
3. Estimate the offset function using the FBE with $\{(y_{T,i} - \hat{y}_{T,i}, x_{T,i})\}_{i=1}^{N_T}$, output $\hat{f}_{\delta,M_2}$.
4. Return $\hat{f}_{S,M_1} + \hat{f}_{\delta,M_2}$ as the estimator for f_T.

Regularization Selection in SATL: In target-only KRR results, for all n, we fixed the constant C and reported the best generalization error decay curves in Figure 2 and other error decay curves for other Cs in Figure 4. In SATL, one can also conduct a similar tuning strategy and select the best performer C. However, this can be computationally insufficient. For example, if one has 40 candidates for C, then there would be a total of 40^2 constants combinations in a two-step transfer learning process. Such a problem also appeared in FBE approaches where one needs to tune the optimal M (number of bases or the degree of B-spline).

Therefore, for each α, we determine the constant C in $\exp\{-Cn^{-\frac{1}{2\alpha+d}}\}$ via following cross-validation (CV) approach. We consider the largest sample size in the current setting, i.e., largest n_S while estimating the source model and largest n_T while estimating the offset, and the estimate C is obtained by the classical K-fold CV, then the estimated $\hat{C}$ is used for all sample size in the experiments.

Additional Results for different basis in FBE: In this part, we provide a detailed description of the implementation for the comparison between SATL and FBE in Figure 3a and 3b. In our implementation, we consider the finite basis as (1) Fourier basis $B_j(x) = \sqrt{2}cos(\pi * k * x)$ (which was used in Wang et al. (2016)) and (2) B_j being the j-th order B-spline. In Wang et al. (2016), the authors use $m_1 = m_2 = 500$, but we notice this will hugely degrade the algorithm performance. Therefore, we use CV to select m_1 and m_2 to optimize the FBE algorithm performance.