Smoothness Adaptive Hypothesis Transfer Learning

Haotian Lin[1] Matthew Reimherr[1]

[1] Pennsylvania State University

ICML 2024

Background

Problem Setting

- Nonparametric regression models:

$$\text{Target: } y_{T,i} = f_T(x_{T,i}) + \epsilon_{T,i}, \quad i = 1, \dots, n_T,$$
$$\text{Source: } y_{S,i} = f_S(x_{S,i}) + \epsilon_{S,i}, \quad i = 1, \dots, n_S.$$

- Model Shift: $P(x_S) = P(x_T)$ but $P(y_T|x_T) \neq P(y_S|x_S)$.
- Source function: $f_S \in H^{m_0}$, a Sobolev space of order $m_0 \geq d/2$.
- Offset function: $f_\delta = f_T - f_S \in H^m$ for some $m \geq m_0$.

Hypothesis Transfer Learning (HTL)

Transferring knowledge from a source domain to a target domain by using the trained source model (hypothesis) while learning the target model.

Background: Learning Framework

Kernel-based HTL

HTL and kernel methods are connected via offset/bias regularization.

- **Input:** Source and target dataset, employ kernel K.
- Phase 1: Source hypothesis training

$$\hat{f}_S = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{n_S} \sum_{i=1}^{n_S} (y_{S,i} - f(x_{S,i}))^2 + \lambda_1 \|f\|_K^2$$

- Phase 2: Transfer via offset regularization

$$\hat{f}_\delta = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{n_T} \sum_{i=1}^{n_T} (y_{T,i} - \hat{f}_S(x_{T,i}) - f(x_{T,i}))^2 + \lambda_2 \|f\|_K^2$$

- **Output:**

$$\hat{f}_T = \hat{f}_S + \hat{f}_\delta$$

Background: Limitation

Limitation in existing works

- **Smoothness-agnostic:** Without knowing the relative smoothness of the f_S and f_δ, using the same kernel regularization in both phases, which against the "simpler" offset principle that leads to the success of HTL.
- **Non-adaptive:** Rate optimality of this two-phase learning framework relies on knowing smoothness m_0 and m and employing the "right" kernels in both phases.

$\Rightarrow$ Question:
How to develop an HTL algorithm so that f_S and f_δ can be learned adaptively and optimally with varying smoothness?

Potential Solution

KRR Revisited

- For a kernel K, and the induced RKHS $\mathcal{H}_K$, KRR estimate is given as

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|^2_{\mathcal{H}_K}.$$

- Assume f_T is in H^{m_0}, the minimax convergence rate is

$$\|\hat{f} - f_T\|^2_{L_2} = \int_{\mathcal{X}} (\hat{f}_0(x) - f_0(x))^2 \, dx = \mathcal{O}_{\mathbb{P}}(n^{-\frac{2m_0}{2m_0+d}}).$$

- If $\mathcal{H}_K$ coincides with H^{m_0} and $\lambda \asymp n^{-\frac{2m_0}{2m_0+d}}$, the minimax rate is attainable.

Problem: How to choose the kernel K to achieve this rate without knowing m_0?

Potential Solution

Robustness of Employed Kernels in KRR

Proposition 1 Let $\hat{f}_T$ be the target-only KRR estimator and K as the imposed kernel,

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

$$\|\hat{f}_T - f_T\|_{L_2}^2 = \mathcal{O}_{\mathbb{P}}\left(n_T^{-2\gamma m_0'/(2\gamma m_0'+d)}\right),$$

which achieves minimax optimal rate $n_T^{-2m_0/(2m_0+d)}$ when $m_0 \leq 2m_0'$.

2. (Saturation Effect) For $m_0' < m_0/2$ and any choice of parameter $\lambda(n_T)$ satisfying that $\lambda(n_T) \to 0$, we have

$$\|\hat{f}_T - f_T\|_{L_2}^2 = \Omega_{\mathbb{P}}\left(n_T^{-4m_0'/(4m_0'+d)}\right).$$

Potential Solution

Solution via Misspecified kernel

- **Possibility**: Imposed misspecified Matérn kernels to achieve rate-optimal HTL.
- **Drawback**: End up choosing a less smooth kernel ($m_0 > 2m_0'$) and never being able to attain the minimax rate because of the saturation effect.
- Demand for a kernel with a more robust misspecified property.

Table of Contents

Target-Only KRR with Gaussian Kernels

Motivation

1. Better to use “over-smooth” misspecified Matérn kernels.
2. The Gaussian kernel is the limit of Matérn kernels.

Theorem (Non-adaptive Rate)

Let the imposed kernel, K, be the Gaussian kernel with fixed bandwidth and $\hat{f}$ be the KRR estimator learned from target dataset $\{(x_i, y_i)\}_{i=1}^{n}$. Under certain standard Assumptions, if $log(1/\lambda) \asymp n^{\frac{2}{2m_0+d}}$, then the following statement holds,

$$\|\hat{f} - f_T\|_{L_2}^2 = \mathcal{O}_{\mathbb{P}}(n^{-\frac{2m_0}{2m_0+d}}).$$

Key takeaway:

- Attain **minimax optimal** convergence rate with fixed bandwidth Gaussian kernels.
- Gaussian kernel smooths a lot, so λ has to decay **exponentially**; Misspecified Matérn kernel requires λ to scale **polynomially**.

Target-Only KRR with Gaussian Kernels

Table: 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^{m_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 $H^{m_0'}$ means the Sobolev space with smoothness order m_0'.

Paper	Imposed RKHS	Rate	λ	γ
[1], [2]	$H^{m_0'}, m_0' > \frac{m_0}{2}$	$n^{-\frac{2m_0}{2m_0+d}}$	$n^{-\frac{2m_0'}{2m_0+d}}$	$-$
[3]	$\mathcal{H}_K$	$n^{-\frac{2m_0}{2m_0+d}+\eta}, \forall \eta > 0$	n^{-1}	$n^{-\frac{1}{2m_0+d}}$
[4]	$\mathcal{H}_K$	$n^{-\frac{2m_0}{2m_0+d}} \log^{d+1}(n)$	n^{-1}	$n^{-\frac{1}{2m_0+d}}$
This work	$\mathcal{H}_K$	$n^{-\frac{2m_0}{2m_0+d}}$	$\exp\{-Cn^{\frac{2}{2m_0+d}}\}$	$-$

Target-Only KRR with Gaussian Kernels

Adaptive process via Training/Validation

Construct a smoothness candidate set $\mathcal{M} = \{m_1, \cdots, m_N\}$ with $m_j - m_{j-1} \asymp 1/\log n_T$ and divide the target dataset into $\mathcal{D}_{T,1}$ and $\mathcal{D}_{T,2}$.

1. For each $m \in \mathcal{M}$, obtain non-adaptive $\hat{f}_{\lambda_m}$ by KRR with $\mathcal{D}_{T,1}$.
2. Obtain the adaptive $\hat{f}_{\lambda_{\hat{m}}}$ by minimizing empirical L_2 error on $\mathcal{D}_{T,2}$, i.e.

$$\hat{f}_{\lambda_{\hat{m}}} = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ \frac{1}{|\mathcal{D}_{T,2}|} \sum_{(y_i, x_i) \in \mathcal{D}_{T,2}} (y_i - \hat{f}_{\lambda_m}(x_i))^2 \right\}.$$

Theorem (Adaptive Rate)

For the adaptive estimator constructed via training/validation method, one has

$$\|\hat{f}_{\lambda_{\hat{m}}} - f_T\|_{L_2}^2 = \mathcal{O}_{\mathbb{P}}\left(\left(\frac{n_T}{\log n_T} \right)^{-\frac{2m_0}{2m_0+d}} \right).$$

Table of Contents

SATL via two-step KRR

Algorithm 1 Smoothess Adaptive Hypothesis Transfer Learning

1. Let the smoothness candidate set for f_S as $\mathcal{M}_S = \{\frac{Q_1}{\log(n_S)}, \cdots, \frac{Q_1 N_1}{\log(n_S)}\}$ and the smoothness candidate set for f_δ 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. Conduct the two-phase KRR-based HTL with each phase follows the training/validation process with $\mathcal{M}_S$ and $\mathcal{M}_\delta$.

Optimality of SATL

Define the parameter space as,

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

Theorem (Optimality of SATL)

Let C_L and C_U be some constants independent of n_S, n_T, R, h, and δ. For $\delta \in (0,1)$, with probability $1-\delta$, we have

1. *(**Lower bound**)*

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

where $\xi_L \propto h^2/R^2$.

2. *(**Upper bound**)*

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

where $\xi_U \propto (h^2 + \sigma_T^2)/(R^2 + \sigma_S^2)$.

Transfer Dynamic and Efficacy

- Upper bound of target-only learning:

$$\left(\frac{n_T}{\log n_T}\right)^{-\frac{2m_0}{2m_0+d}}$$

- Upper bound of SATL:

$$\underbrace{\left(\frac{n_S}{\log n_S}\right)^{-\frac{2m_0}{2m_0+d}}}_{\text{rough estimation error}} + \underbrace{\left(\frac{n_T}{\log n_T}\right)^{-\frac{2m}{2m+d}}}_{\text{offset estimation error}} \xi_U$$

- Jointly determine by source sample size n_S and factor ξ_U.
- Compared to the target-only KRR rate, SATL produces a faster rate with small ξ_U (high similarity) and large n_S.

Comparing to Existing Bounds

- Ours:

$$\left(\frac{n_S}{\log n_S}\right)^{-\frac{2m_0}{2m_0+d}} + \left(\frac{n_T}{\log n_T}\right)^{-\frac{2m}{2m+d}} \xi_U$$

- Existing works via offset TL:

$$(n_S)^{-\frac{2m_0}{2m_0+d}} + (n_T)^{-\frac{2m}{2m+d}} h^2$$

- The logarithmic factor due to adaptivity.
- Our bound indicates the transfer efficacy via the offset TL not singly depends on the margin of dissimilarity measure h, but jointly depends on the ratio of the signal strength between offset and source models (a.k.a. the angle).

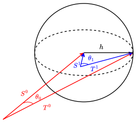

Figure: Geometric illustration for how ξ_U will affect the HTL.

Experiments: Target-only KRR

Construct f_T from Gaussian process s.t. $f_T \in H^{m_0}$.

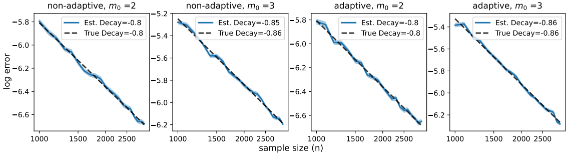

Figure: Emipirical and theoretical error decay curves for different m_0.

Experiments: SATL

$$n_S = n_T^{3/2},\ f_S \in H^1 \text{ and } f_\delta \in H^m \text{ such that } \|\hat{f}_T - f_T\|_{L^2}^2 = O(n_T^{-\frac{2m}{2m+1}}).$$

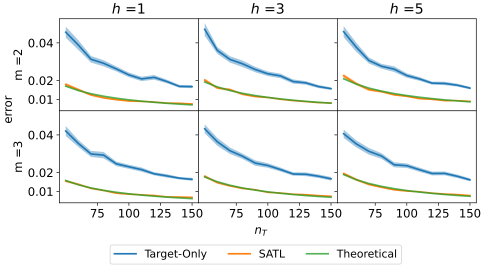

Figure: Generalization error for different m and h.

Summary of Contribution

Kernel Ridge Regression

- When the true function lies in a Sobolev space H^{m_0}, we rigorously prove that employing fixed bandwidth Gaussian kernels in KRR attains the minimax optimal rate.
- The optimal decay rate for λ is $\lambda \asymp \exp\{-Cn^{\frac{2}{2m_0+d}}\}$, which decays exponentially in n.

Transfer Learning

- We present a smoothness-adaptive and rate-optimal hypothesis transfer learning algorithm for nonparametric regression, called SATL.
  - Optimality: Employing Gaussian kernels to avoid saturation and guarantee the possibility of optimality.
  - Adaptivity: Training and validation process to achieve adaptive rate.

Reference I

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

Haobo Zhang, Yicheng Li, Weihao Lu, and Qian Lin.
On the optimality of misspecified kernel ridge regression.
In *International Conference on Machine Learning*, pages 41331–41353. PMLR, 2023.

Mona Eberts and Ingo Steinwart.
Optimal regression rates for SVMs using Gaussian kernels.
Electronic Journal of Statistics, 7(none):1 – 42, 2013.

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