Overview

- The problem: Nonparametric regression in reproducing Kernel Hilbert Space (RKHS).
- The goal: Close the gap between the known lower and upper bound on the prediction error.
- Contributions:
  - Our proposed algorithm achieves the optimal rate in so-called “hard regime”, resolving a long-standing open problem.
  - We achieve even faster convergence when the Bayes error is 0.
- When the Bayes error is 0, we show that the best regularization is 0, which connects recent interest on the generalization ability of the interpolator.
- Algorithm: A randomized variant of the kernel ridge regression.

Background

Let $X$ and $Y$ be the feature and label space respectively. Task: Given an i.i.d. training set $S = \{(x_i, y_i) \in X \times Y \}_{i=1}^n$ from an unknown distribution $\rho$, find $f$ whose risk

$$\mathcal{R}(f) = \int_{x \in X} (f(x) - y)^2 \, d\rho$$

is close to the optimal risk $\mathcal{R}^* = \inf_f \mathcal{R}(f)$. We consider functions from a Reproducing Kernel Hilbert Space (RKHS).

<table>
<thead>
<tr>
<th>Definitions</th>
</tr>
</thead>
</table>
| $f^\rho(x) = \int x \phi(y|x) \, d\rho$: the regression function, achieves the optimal risk $\mathcal{R}^*$.
| $\mathcal{P}_C$: the marginal distribution, $L_C^\rho \rightarrow L^\rho$, the integral operator defined by $(L_C^\rho f)(x) = \int K(x, x') f(x') \, d\rho(x')$.
| $\lambda$: an orthonormal basis, $\{\phi_i\}_{i=1}^\infty$ of $L^\rho$, consisting of eigenfunctions of $L^\rho$ with corresponding non-negative eigenvalues $\lambda_i$.

Fact: the set $\{\lambda_i\}$ is finite or $\lambda_i \rightarrow 0$ when $k \rightarrow \infty$.

<table>
<thead>
<tr>
<th>Assumptions</th>
</tr>
</thead>
</table>
| (i) Regularity: Separable RKHS $H_K$ associated to a Mercer kernel $K$, $X \times X \rightarrow \mathbb{R}$.
| (ii) Boundness: $\sup_{x, y} |K(x, x')| = R^2 < \infty$. (Set $R = 1$ for simplicity).
| (iii) Source condition: Define $L^\rho_q(L_C^\rho f)(x) = \sum_{i=1}^{\infty} \lambda_i^q f(x') \, d\rho(x')$: $L^\rho_q \rightarrow L^\rho_q$, for $0 < q < 1$.

We assume that $f \in L^\rho_q(L_C^\rho f)$ for $0 < b \leq 1/2$ (i.e., $\exists q \geq 2$, $f \in L^\rho_q(\cdot)$).

- Theorem 1 (simplified): There exists a setting of $\lambda \geq 0$ such that:
  1. When $b \neq 0$,
     $$\mathbb{E} [R(f_{\lambda}) - R(f)] \leq \left( \min \left( \mathbb{E} [R(f)] \right) \right)^{\frac{1}{2}} + n^{-\beta} \lambda^{-\beta},$$
  2. When $b = 0$, we obtain a faster rate of $n^{-\beta'}$.

<table>
<thead>
<tr>
<th>Remarks</th>
</tr>
</thead>
</table>
| Optimal rate: Our rate $n^{-\beta}$ matches the worst-case lower bound (Fischer and Steinwart, 2017) without additional assumptions for the first time in the literature.
| Low-noise acceleration: When $R(f_{\lambda}) = 0$, we obtain a faster rate of $n^{-\beta'}$.

Technical Details

- Online-to-batch conversion (Cesa-Bianchi et al., 2004): Allows us to leverage strong inequalities from online learning.
- “The identity” for online Kernel ridge regression (Zhdanov and Kalnishkan, 2013).

Theorem 2 (Zhdanov and Kalnishkan, 2013, Theorem 1). Take a kernel $K$ on a domain $X$ and a parameter $\lambda > 0$. Then, with the notation of Algorithm 1, we have

$$\frac{1}{n} \sum_{i=1}^n (f_{\lambda}(x_i) - y_i)^2 = \min_{\lambda > 0} \mathbb{E} [R(f_{\lambda})] = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 - \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2,$$

where $d_{\lambda} = K(x_i, x_i) - k_{\lambda}(x_i, x_i)$, $k_{\lambda}(x_i, x_i) = K(x_i, x_i) - K(x_s, x_i)$, $x_s \in X$ such that $k_{\lambda}(x_s, x_i) \geq 0$, $k_{\lambda}(x_i, x_j) = K(x_i, x_j)$, and $K_{\lambda}$ is the Gram matrix of the samples $x_1, \ldots, x_n$.

Lemma 1 (Classic result, e.g., Zhdanov et al., 2005). Let $d_{\lambda} = \frac{1}{\lambda} \sum_{i=1}^n (f(x_i) - y_i)^2 - \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2$. Then $d_{\lambda}$ tends to zero as $\lambda \rightarrow 0$.

KTR Algorithm:

1. Input: A training set $S = \{(x_i, y_i)\}_{i=1}^n$, a regularization parameter $\lambda \geq 0$.
2. Randomly permute the training set $S$ for $t = 0, 1, \ldots, n-1$.
3. Set $f_t = \text{argmin}_{f \in \mathcal{K}} \mathbb{E} [R(f)] + \frac{1}{\lambda} \sum_{i=1}^n (f(x_i) - y_i)^2$ (break ties by the minimum norm).
4. Return $f_S = T^* \circ f_t$, where $T$ is uniformly at random between 0 and $n-1$.

Theorem 1 (simplified): There exists a setting of $\lambda \geq 0$ such that:

(a) When $b \neq 0$,
$$\mathbb{E} [R(f_{\lambda}) - R(f)] \leq \left( \min \left( \mathbb{E} [R(f)] \right) \right)^{\frac{1}{2}} + n^{-\beta} \lambda^{-\beta},$$
(b) When $b = 0$, we obtain a faster rate of $n^{-\beta'}$.

<table>
<thead>
<tr>
<th>Related work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notable ones:</td>
</tr>
</tbody>
</table>
| • (Lin et al., 2018) and (Diederik and Bach, 2016): optimal rate of $O(n^{-\beta})$ for regime $3/2 + b < 1$.
| • With an additional assumption, (Pilland-Vivien et al., 2018) achieve the optimal rate in a subregime of $3/2 + b < 1$.
| • Low-noise acceleration: (Orabona, 2014) achieve $O(n^{-1})$ when $R(f_\lambda) = 0$, for smooth and Lipschitz losses.
| Asymptotic result on finite dimensional case: (Hastie et al., 2019) show that when $R(f_\lambda) = 0$, the best ridge parameter is $\lambda = 0$.

Conclusion

Our work verifies that the previously-known lower bound is indeed optimal by showing a matching upper bound. Furthermore, we open up a new parametrization of the risk bound via the Bayes risk $R(f_\lambda)$, which allows accelerated rates.

- We conjecture the standard KRR would enjoy a similar upper bound; we believe the randomization of KTR just provided an easy path to the proof.
- What about the regime $\beta > 1/2$? Our method suffers from ‘saturation’ effect due to the regularization.
- What would be the lower bound for the case $R(f_\lambda) = 0$? Note this is not unrealistic, e.g., in vision tasks where human can do a near-classification of images.

References