Kernel Truncated Randomized Ridge Regression: Optimal Rates and Low Noise Acceleration

Kwang-Sung Jun (U of Arizona), Ashok Cutkosky (Google Research), and Francesco Orabona (Boston University)

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 to recent interest on the generalization ability of the interpolator.
- **Algorithm**: A randomized variant of the kernel ridge regression.

Background

Let $\mathbb{X}$ and $\mathbb{Y}$ be the feature and label space respectively. Task: Given an i.i.d. training set $S = \{\boldsymbol{x}_t \in \mathbb{X}, y_t \in \mathbb{Y}\}_{t=1}^n$ from an unknown distribution ρ, find $\hat{f}$ whose risk

$$\mathcal{R}(\hat{f}) := \int_{\mathbb{X}\times\mathbb{Y}} \left(\hat{f}(\boldsymbol{x}) - y\right)^2 d\rho$$

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

[Definitions]

- $f_\rho(\boldsymbol{x}) := \int_{\mathbb{Y}} y d\rho(y|\boldsymbol{x})$: the regression function $\implies$ achieves the optimal risk $\mathcal{R}^\star$.
- $\rho_{\mathbb{X}}$: the marginal distribution. $\mathcal{L}^2_{\rho_{\mathbb{X}}}$: the space of square integrable functions w.r.t. $\rho_{\mathbb{X}}$.
- $L_K : \mathcal{L}^2_{\rho_{\mathbb{X}}} \to \mathcal{L}^2_{\rho_{\mathbb{X}}}$: the integral operator defined by $(L_K f)(\boldsymbol{x}) = \int_{\mathbb{X}} K(\boldsymbol{x},\boldsymbol{x}')f(\boldsymbol{x}')d\rho_{\mathbb{X}}(\boldsymbol{x}')$.
- $\exists$ an orthonormal basis $\{\Phi_1, \Phi_2, \cdots\}$ of $\mathcal{L}^2_{\rho_{\mathbb{X}}}$ consisting of eigenfunctions of L_K with corresponding non-negative eigenvalues $\{\lambda_1, \lambda_2, \cdots\}$. Fact: the set $\{\lambda_i\}$ is finite or $\lambda_k \to 0$ when $k \to \infty$.

[Assumptions]

(i) Regularity: Separable RKHS $\mathcal{H}_K$ associated to a Mercer kernel $K : \mathbb{X}\times\mathbb{X} \to \mathbb{R}$.

(ii) Boundedness: $\sup_{\boldsymbol{x}\in\mathbb{X}} K(\boldsymbol{x},\boldsymbol{x}) = R^2 < \infty$. (set $R = 1$ for simplicity).
$\mathbb{Y} \in [-Y, Y]$ with $Y < \infty$.

(iii) **Source condition**: Define

$$L_K^\beta(\mathcal{L}^2_{\rho_{\mathbb{X}}}) := \left\{ f = \sum_{i=1}^\infty \lambda_i^\beta a_i \Phi_i \, : \|L_K^{-\beta} f\|^2_\rho := \sum_{i=1}^\infty a_i^2 < \infty \right\}.$$

We assume that

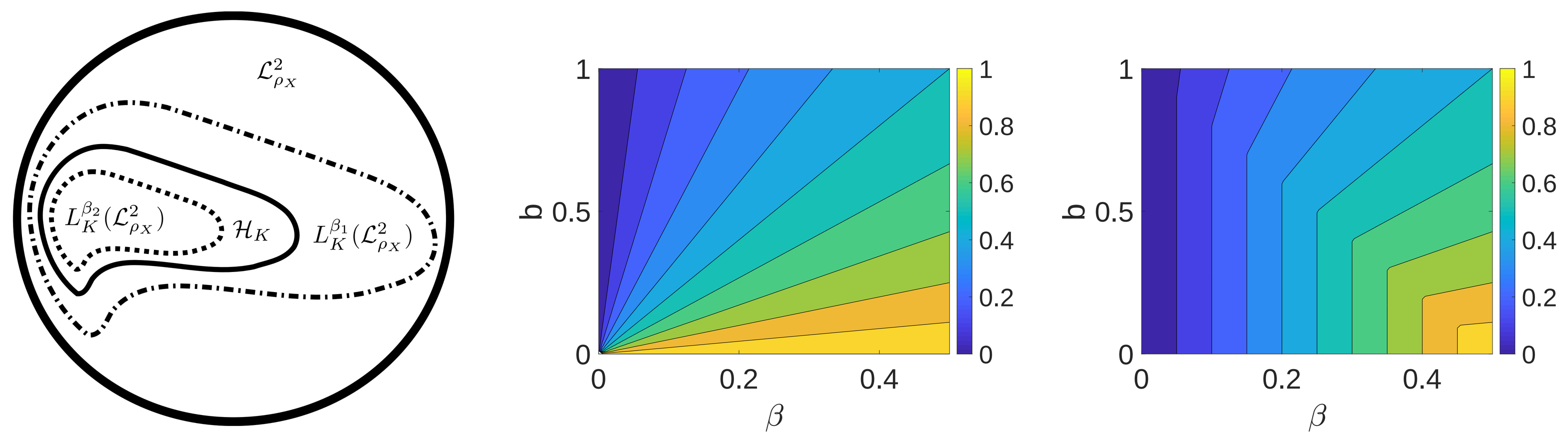
$$f_\rho \in L_K^\beta(\mathcal{L}^2_{\rho_{\mathbb{X}}}) \text{ for } 0 < \beta \le 1/2 \qquad (\text{i.e., } \exists g \in \mathcal{L}^2_{\rho_{\mathbb{X}}} : f_\rho = L_K^\beta(g)).$$

$\implies$ characterizes the "complexity" of the function, answering "how much infinite" the function f's norm is. Smaller β means that f is more complex.

(iv) **Eigenvalue decay**: $\exists b \in [0,1]$ such that $\mathrm{Tr}[L_K^b] < \infty$.

[Facts]

- $\beta = 1/2$ means $f_\rho \in \mathcal{H}_K$.
- $b = 0$ means that the kernel induces finite dimensions.
- Sum of the eigenvalues of L_K is at most R^2.



(a) $0 < \beta_1 < \frac{1}{2} < \beta_2$ (b) known lower bound (exponent of $\frac{1}{n}$) (c) known upper bound

Kernel Truncated Randomized Ridget Regression (KTR^3)

Algorithm 1 KTR^3: Kernel Truncated Randomized Ridge Regression

Input: A training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, a regularization parameter $\lambda \ge 0$
Randomly permute the training set S
for $t = 0, 1, \ldots, n-1$ **do**
  Set $f_t = \operatorname{argmin}_{f\in\mathcal{H}_K} \lambda\|f\|^2 + \frac{1}{n}\sum_{i=1}^t (f(\boldsymbol{x}_i) - y_i)^2$
  (break ties by the minimum norm)
end for
Return $f_{S,\lambda} := T^Y \circ f_k$, where k is uniformly at random between 0 and $n-1$

Theorem 1 *(simplified)*. *There exists a setting of $\lambda \ge 0$ such that:*
(i) When $b \ne 0$,

$$\mathbb{E}\left[\mathcal{R}(f_{S,\lambda})\right] - \mathcal{R}(f_\rho) \le O\left(\min\left((n/\mathcal{R}(f_\rho))^{-\frac{2\beta}{2\beta+1}} + n^{-2\beta}, n^{-\frac{2\beta}{2\beta+b}}\right)\right).$$

(ii) In the case $b = 0$ and $\beta = \frac{1}{2}$,

$$\mathbb{E}\left[\mathcal{R}(f_{S,\lambda})\right] - \mathcal{R}(f_\rho) \le O\left(n^{-1}\mathrm{Tr}[L_K^0]\log\left(1 + n/\mathrm{Tr}[L_K^0]\right)\right).$$

(When $b = 0$ the space is finite dimensional, hence β can only have value 0 or $1/2$ and there is no convergence to the Bayes risk when $\beta = 0$.)

[Remarks]

- **Optimal rate**: Our rate $n^{-\frac{2\beta}{2\beta+b}}$ matches the worst-case lower bound (Fischer and Steinwart, 2017) without additional assumptions for the first time in the literature.
  $\implies$ In the regime $2\beta + b < 1$, prior works have a slower rate of $n^{-2\beta}$.
- **Low-noise acceleration**: When $\mathcal{R}(f_\rho) = 0$, we obtain a faster rate of $n^{-\frac{2\beta}{\min\{2\beta+b,1\}}}$.
  $\implies$ The first of its kind; no known lower bounds.
- **Interpolation** (almost): When $\mathcal{R}(f_\rho) = 0$, the optimal λ that minimizes the generalization upper bound in Theorem 1 goes to zero when β goes to $1/2$ and becomes exactly 0 when β is exactly $1/2$.

Technical ingredients

- Online-to-batch conversion (Cesa-Bianchi et al., 2004)
  $\implies$ Allows us to leverage strong inequalities from online learning.
- "The identity" for online Kernel ridge regression (Zhdanov and Kalnishkan, 2013).
  $\implies$ A rather obscure result that says: The online error of KRR, adjusted by some weights, is *exactly* the minimum of the batch training error objective.

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

$$\frac{1}{n}\sum_{t=1}^n \frac{(f_{t-1}(\boldsymbol{x}_t) - y_t)^2}{1 + \frac{d_t}{\lambda n}} = \min_{f\in\mathcal{H}_K} \lambda\|f\|^2 + \frac{1}{n}\sum_{t=1}^n \left(f(\boldsymbol{x}_t) - y_t\right)^2,$$

where $d_t := K(\boldsymbol{x}_t,\boldsymbol{x}_t) - \boldsymbol{k}_{t-1}(\boldsymbol{x}_t)^\top(K_{t-1} + \lambda n I)^{-1}\boldsymbol{k}_{t-1}(\boldsymbol{x}_t) \ge 0$, $\boldsymbol{k}_{t-1}(\boldsymbol{x}_t) := [K(\boldsymbol{x}_t,\boldsymbol{x}_1), \ldots, K(\boldsymbol{x}_t,\boldsymbol{x}_{t-1})]^\top$, and K_{t-1} is the Gram matrix of the samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}$.

Lemma 1. *(Classic result; e.g., (Cesa-Bianchi et al., 2005))*

$$\mathbb{E}\left[\sum_{t=1}^n \frac{d_t}{d_t + \lambda n}\right] \le \sum_{i=1}^\infty \log\left(1 + \frac{\lambda_i}{\lambda}\right)$$

[E.g., Linear version]

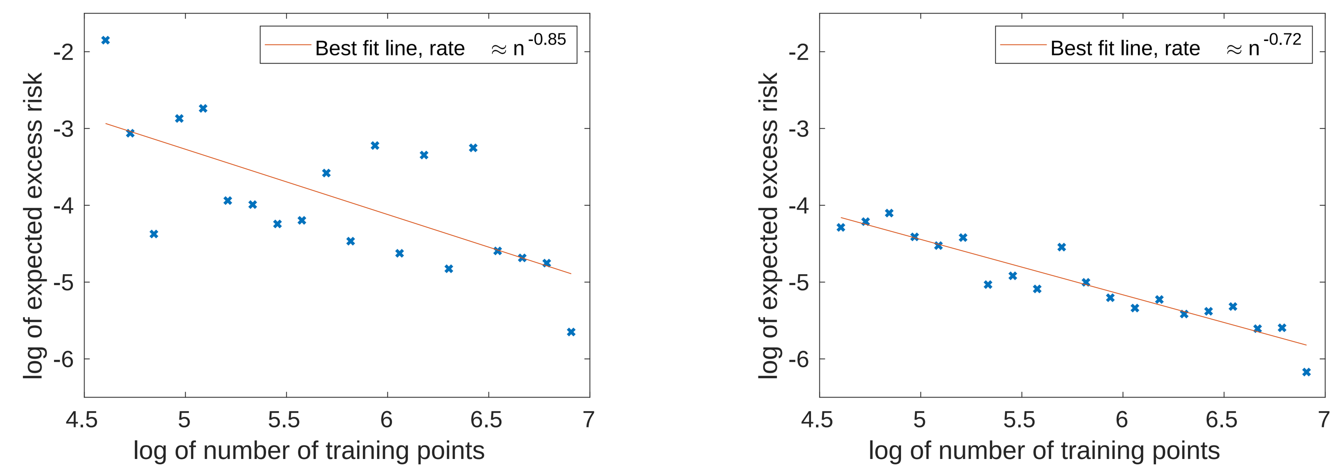
- Define $\boldsymbol{V}_t = \lambda n \boldsymbol{I} + \sum_{s=1}^t \boldsymbol{x}_s\boldsymbol{x}_s^\top$. Then, $d_t = (\lambda n)\cdot\|\boldsymbol{x}_t\|^2_{\boldsymbol{V}_{t-1}^{-1}}$.
  $\implies$ The test error at time t is weighted by $\frac{1}{1+\|\boldsymbol{x}_t\|^2_{\boldsymbol{V}_{t-1}^{-1}}}$.
- Also, $\mathbb{E}\left[\sum_{t=1}^n \frac{d_t}{d_t+\lambda n}\right] = \mathbb{E}[\sum_{t=1}^n \|\boldsymbol{x}_t\|^2_{\boldsymbol{V}_t^{-1}}] \le \sum_{i=1}^d \log\left(1 + \frac{\lambda_i}{\lambda}\right)$

Related work

Notable ones:

- (Lin et al., 2018) and (Dieuleveut and Bach, 2016): suboptimal rate of $O\left(n^{-2\beta}\right)$ for regime $2\beta + b < 1$.
- With an additional assumption, (Pillaud-Vivien et al., 2018) achieve the optimal rate in a subregime of $2\beta + b < 1$.
- Low-noise acceleration: (Orabona, 2014) achieve $O(n^{-\frac{2\beta}{\beta+1}})$ when $\mathcal{R}(f_\rho) = 0$, for smooth and Lipschitz losses.
- Asymptotic result on finite dimensional case: (Hastie et al., 2019) show that when $\mathcal{R}(f_\rho) = 0$ the best ridge regression parameter λ is 0.

Experiments



(a) $b = \frac{1}{8}$, $\beta = \frac{7}{16}$, theoretical rate: $n^{-\frac{7}{8}}$ (b) $\beta = \frac{1}{4}$, $b = \frac{1}{6}$, theoretical rate: $n^{-\frac{3}{4}}$

- A spline regression task.
- For each training set size n we choose the λ that minimizes the average excess risk.
- For (b), previously-known bounds predicts a slower rate of $n^{-\frac{1}{2}}$.

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 $R(f_\rho)$, which allows accelerated rates.

- We conjecture the standard KRR would enjoy a similar upper bound; we believe the randomization of KTR^3 just provided an easy pathway to the proof.
- What about the regime $\beta > 1/2$? Our method suffers from 'saturation' effect due to the regularizer.
- What would be the lower bound for the case $R(f_\rho) = 0$? Note this is not unrealistic, e.g., in vision tasks where human can do a near-perfect classification of images.

References

N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Trans. Inf. Theory*, 50(9):2050–2057, 2004.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. A second-order Perceptron algorithm. *SIAM Journal on Computing*, 34(3):640–668, 2005.

A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. *arXiv preprint arXiv:1702.07254*, 2017.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560v2*, 2019.

J. Lin, A. Rudi, L. Rosasco, and V. Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Applied and Computational Harmonic Analysis*, 2018.

F. Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Advances in Neural Information Processing Systems 27*, 2014.

L. Pillaud-Vivien, A. Rudi, and F. Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems 31*, pages 8114–8124. Curran Associates, Inc., 2018.

F. Zhdanov and Y. Kalnishkan. An identity for kernel ridge regression. *Theor. Comput. Sci.*, 473:157–178, February 2013.