KERNEL TRUNCATED RANDOMIZED RIDGE REGRESSION

Kwang-Sung Jun, Ashok Cutkosky, Francesco Orabona Published in **NeurIPS 2019**

Gaurav Gijare

Kernel Methods EE5605

Intro.

Prediction with Regression

Introduction

- Our Training Set $S := \{x_t, y_t\}_{t=1}^{n}$ with n identically and independently distributed from a fixed but unknown distribution ρ on $\mathbb{X}x\mathbb{Y}$.
- The main goal of a non parametric regression task is to find a function $\hat{f}$ whose risk(Error) is as close to the optimal risk(Noise).
- In other words, Risk Function

$$R(\hat{f}) = \int_{\mathbb{X}x\mathbb{Y}} (\hat{f}(x) - y)^2 d\rho$$

is as close as possible to Optimal Risk

$$R^* =_f^{inf} \mathcal{R}(f)$$

Prediction with Regression

Introduction

- Our analysis is with Kernel methods, and we thus define our classic Kernel Ridge Regression prediction function as

$$\hat{f} = \underset{f \in \mathcal{H}_k}{argmin}\ \lambda \mid\mid f \mid\mid^2 + \frac{1}{n}\sum_{t=1}^{n}(f(x_t - y_t)^2)$$

  where $\mathcal{H}_k$ is associated with a kernel K and λ is the hyper-parameter controlling the amount of regularization.
- Further, we have referred to this regularization parameter λ

Prediction with Regression

Findings

- Empirically it has been shown that over a particular dataset, KRR performs better without a regularization parameter, which contradicts initial theoretical findings which tells us that a non-zero regularization is needed in all cases for learning in infinite dimensional RKHSs.

Motivation

- It is unclear if this mismatch between theory and practice is due to
  - Sub optimal analyses that lead to sub-optimal choices of the amount of regularization
  - Not taking into account crucial data-dependent quantities (e.g., capturing "easiness' of the problem) that allow fast rates and minimal regularization.
- We explore an algorithm that answers the above, as well as gives us a better convergence rate.

Problem

Problem Statement

Problem

- We want to explore an algorithm that tightens and gives optimal error bounds.

Use of Kernels in this paper

- The main advantage of kernel ridge regression over a normal non parametric least squares regression problem is that it automatically eliminates the domain boundaries bias, associated with a locally weighted approach.
- We use kernels because we can transform the data to a higher dimensional space to make the data linearly separable, and thus operations are easier to handle.
- The algorithm discussed is an improvement to kernel ridge regression as stated in the introduction.

Setting

Notation

- Let $\mathbb{X} \subset \mathbb{R}^d$ be a compact set
- $\mathcal{H}_K$ is a separable RKHS
- $K_{i,j} = K(x_i, x_j)$ is our Kernel function where x_i, x_j belong to $S_t \subseteq S$ containing the first t elements of the set S
- $\rho_{\mathbb{X}}$ is the marginal probability measure on $\mathbb{X}$

Notation

- $\mathcal{L}^2$ is the space of square integrable functions
- $f_\rho(x) = \int_{\mathbb{Y}} y d\rho(y|x)$
- $\mathcal{R}^* = \mathcal{R}(f_\rho) = \inf_{f \in \mathcal{L}^2_{\rho_{\mathbb{X}}}} \mathcal{R}(f)$
- $(L_K f)x = \int_{\mathbb{X}} K(x, x') f(x') d\rho_{\mathbb{X}}(x')$
-

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

Assumptions

Assumption 1 - Boundedness

- We assume our $\mathcal{K}$ to be bounded ie: $\sup_{x \in \mathbb{X}} \mathcal{K}(x, x) = R^2 < \infty$, further, without loss of generality we take $R = 1$.
- We assume our training set target values(given) to be bounded, ie: $\mathbb{Y} \in$[-Y, Y];Y$<\infty$

Assumptions

Assumption 2 - Source Condition

- Assume that $f_\rho \in L_K^\beta(\mathcal{L}^2_{\rho_\mathbb{X}})$ where $0 < \beta \leq \frac{1}{2}$
- ie:$g \in \mathrm{L}^2_{\rho_\mathbb{X}}$: $f_\rho = L_K^\beta(g)$

Assumption 3 - Eigen Value Decay

- Assume that there exists $b \in [0, 1]$ such that $\mathrm{Tr}[L_K^b] < \infty$.

Literature Review

Literature Review

Theis paper has cited many other papers which form a part of the *RKHS* and *Kernel Ridge Regression* literature.

- ***(J. Lin et. al 2018)***:
  In this paper regression problems over a separable Hilbert space with square loss is studied and non parametric regression over a RKHS is also studied. Many algorithms like ridge regression, pca and gradient methods are used. Almost sure convergence is obtained with optimal rates for all the algorithms. They show that KRR converges the fastest to Bayes risk out of all the Kernel based algorithms. This paper also has some of the best results for non-parameterized ridge regression using Assumptions 1-3 given in our paper. These results are the rates which are suboptimal in the regime $2\beta + b < 1$. In contrast to this, the rates derived in our paper are sub optimal in all regimes. Also, these rates do not depend in any way on the risk of the optimal function f. Hence, they never support the choice of a regularization parameter being zero.

Literature Review

- ***(Steinwart, et. al 2000)***:
  In this paper, new oracle inequality for kernel based ridge regression is established, which is in turn used to derive learning rates for these methods. The rates are found to be independent of the exponent in the regularization term. It is also shown that the rates are asymptotically optimal.
- ***(C. Zhang, et. al 2008, M. Belkin et. al)***:
  These papers try out KRR without the regularization term on real datasets using deep learning comparisons which works very well. The datasets are mainly for classifiers.

Literature Review - Comparisions

- ***(Steinwart, et. al 2008,2009 and A. Dieuleveut et. al. 2016)***:
  Papers from which the notation has been derived to maintain consistency. All of these papers are on least-squares ridge regression using RKHSs' and Hilbert Spaces.
- ***(A. Caponnetto and E. De Vito 2007)***:
  In the case for $R(f_\rho) \neq 0$, the author's rate matches the worst case lower bound as derived in the above paper. The rate is $n^{\frac{-2\beta}{(2\beta+b)}}$.
- ***(T. Hastie et al 2019)***
  This paper states an asymptotic bound, asymptotic result (as $n \to \infty$) that the best regularization parameter of ridge regression is 0 when there is no label noise (i.e., $R(f_\rho) = 0$) and $\beta = \frac{1}{2}$. Their result aligns well with given paper, but given paper is not limited to asymptotic regimes nor finite dimensional spaces. On the other hand, given paper's guarantee is an upper bound on the risk rather than an equality.

Literature Review - Comparisions

- ***(F. Zhdanov and Y. Kalnishkan. 2013)***:
  Proof of theorem 1 in this paper, authors of the paper have used a recent result on online kernel ridge regression as seen in Theorem 3 included in given paper. This result has been used to prove the Theorem 1 given in the paper. Lemma 3 in this paper in also used in support to derive the bound in the paper.
- ***(J. Lin and V. Cevher 2018)***:
  The authors take a different approach than the one taken in the J. Lin and V. Cevher's paper. The main difference in the approaches is that "*the study of the convergence of empirical covariance operator to the population one, which seems to deteriorate when the regularization parameter becomes too small, which is precisely needed in the regime* $2\beta + b < 1$." The authors use the result in the Zhdanov paper(Theorem 3) as it "*is the key to obtain the improved rates in the regime* $2\beta + b < 1$. *In particular, it allows us to analyze the effect of the eigenvalues using only the expectation of the Gram matrix Kn and nothing else.*"

Literature Review - Comparisions

- ***(F. Cucker and D. X. Zhou 2007)***:
  Theorem 4 in the given paper is based on Proposition 8.5 in this paper.
- ***(N. Cesa-Bianchi et al 2004)***:
  Given paper uses a result from the this paper known as *online-to-batch conversion* to arrive at the bound stated in Theorem 1.
- ***(S. Shalev-Shwartz and S. Ben-David. 2014)***:
  The stability bound for KRR from this paper is compared to the one obtained in the given paper.
- ***(T. Zhang 2003)***:
  Leave-One-Out analysis, in this paper states another bound.

Solution

Solution

Kernel Truncated Randomized Ridge Regression : Algorithm

- **Input:** A training set $S = \{(x_i, y_i)\}_{i=1}^{n}$ and a regularization parameter λ. Permute the training set S
- **Algorithm:**
  for t = 0, 1,...., n-1 :
  $$f_t = \underset{f \in \mathcal{H}_k}{argmin}\ \lambda ||f||^2 + \frac{1}{n}\sum_{i=1}^{t}(f(x_i) - y_i)^2$$
  (take minimum norm solution when unique solution is not available)
- **Output:** Return $T^Y \circ f_k$ where k is uniform random between 0 and n-1 and T^Y is the truncation function

Solution

The Truncation function

- As the name suggests, the truncation function cuts all values below and above a predefined constant.
- ie; $T^{Y}(z) = min(Y, |z|).sign(z)$

Kernel Truncated Randomized Ridge Regression

- So what the algorithm does is, it fixes one parameter λ, our regularization parameter.
- Next, it randomly permutes our training set and prepares n models, each computed with 't' terms as shown in the above algorithm(t = 0, 1, ..., n-1).
- Then, it selects one of these n functions at random, and applies the *truncation function* to it, which is the final returned prediction model that we use.

Calculating Excess Risk

We calculate bounds on excess risk in 2 cases, the first being a generic $\lambda > 0$ amd the second one being the 'easy' $\lambda = 0$.

Theorem 1

- Let $\mathbb{X} \subset \mathbb{R}^d$ be a compact domain and K be a Mercer kernel such that assumptions 1, 2, and 3 are verified. Define by $f_{S,\lambda}$ the function returned by the KTR^3 algorithm on a training set S with regularisation parameter $\lambda > 0$

Bound on our excess risk is defined as follows:

$$\mathrm{E}[\mathrm{R}(\mathrm{f}_{S,\lambda})] - \mathcal{R}(f_\rho) \leq \lambda^{2\beta} ||L_K^{-\beta} f_\rho||_\rho^2 + min\left[\frac{4Y^2 Tr[L_k^b]}{\lambda^b n} min\left(ln^{1-b}\left(1+\frac{1}{\lambda}\right), \frac{1}{b}\right), \frac{\lambda^{2\beta-1}||L_k^{-\beta} f_\rho||_\rho^2}{n} + \frac{\mathcal{R}(f_\rho)}{\lambda n}\right]$$

Calculating Excess Risk

Theorem 2

- As mentioned, we take $\lambda = 0$, and similar conditions to theorem 1. Assume $\beta = \frac{1}{2}$ and $\mathcal{R}(f_\rho) = 0$, along with the distribution ρ satisfying $\mathcal{K}_n$ is invertible with probability 1.
- Then, $\mathbb{E}[\mathcal{R}(f_{S,\lambda})] - \mathcal{R}(f_\rho) = \mathcal{O}(n^{-1})$

Calculating Excess Risk

Remarks

- From the first theorem, we see that our optimal λ (Required minimize generalization upper bound) goes to 0 when β goes to $\frac{1}{2}$ (and $\lambda = 0$ when $\beta = \frac{1}{2}$) which answers our question from the Motivation section.
- From the second theorem above, we see that because of the $\frac{1}{n}$ dependence of our excess risk, bounds can be can thus be easily predicted for a subset of our original set.
- If we sample from $\lfloor (1-\alpha)n \rfloor$ to $n-1$, we see that our algorithm gives worse bounds by a factor of $\frac{1}{\alpha}$ ($\alpha \in (0, 1]$).

Corollary

Corollary 1

Under the assumptions in Theorem 1, there exists a setting $\lambda \geq 0$ such that:

1. when $b \neq 0$,

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

2. when $b = 0$ and $\beta = \frac{1}{2}$*,

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

*When $b = 0$, our space is finite dimensional which implies β must be either 0 or $\frac{1}{2}$ and clearly with $\beta = 0$, there is no convergence to bayes risk.

Results

Understanding the significance of the theorems/corollary

Remarks

- Previous research says that $\mathcal{O}(n^{-2\beta})$ was the best known bound without any assumptions.
- From the Corollary 1([1]]), we see that if we put $2\beta + b < 1$, then our rate is in the order $\mathcal{O}(n^{-\frac{2\beta}{2\beta+b}})$ which is a **strict improvement** (as $2\beta < \frac{2\beta}{2\beta+b}$), and hence this region can be considered as optimal.

Understanding the significance of the theorems/corollary

Remarks

- Again, in Corollary 1([1]]), if we take our $\mathcal{R}(f_\rho) = 0$(Best attainable bayes risk = 0), we see that our convergence happens at a rate of $\mathcal{O}(n^{-\frac{2\beta}{min(1,2\beta+b)}})$.
- This is clearly a faster rate than without the condition.
- This means that our algorithm **accelerates** if bayes risk is close to 0.
- Another interesting point to note, is that this holds without any conditions on β, ie: Our f_ρ need not even belong to a hilbert space (β can be less than $\frac{1}{2}$).

Comparing with previous results

Novelty of this solution

- The Theorem in *Shalev-Shwartz and Ben-David [2014]* holds only for $\lambda \geq \frac{4}{m}$. This does not allow setting $\lambda = 0$ which prevents from achieving the rate of $\mathcal{O}(n^{-1})$ for the case of $\beta = 0$ and Bayes risk 0.
- *Lin.et al [2018] got the optimal rate as -*

$$\mathbb{E}[\mathcal{R}(f_{S,\lambda})] - \mathcal{R}(f_\rho) \leq \begin{cases} \mathcal{O}(n^{-\frac{2\beta}{2\beta+b}}), & \text{if } 2\beta + b \geq 1 \\ \mathcal{O}(n^{-2\beta}), & \text{otherwise} \end{cases}$$

These rates are sub-optimal in the regime $2\beta + b < 1$ whereas our rates are optimal in all the regimes. Since these rates do not depend on $\mathcal{R}(f_\rho)$, they don't support the regularisation parameter being set to 0.

Thoughts

Limitations and Suggestions

Loosening our assumptions

- Considerable limitations to this approach are the conditions from Assumptions 1, 2, and 3.
- Due to Assumption 2, a constraint is imposed such that $\beta \leq \frac{1}{2}$ and hence our rates are applicable only for this bound (Although this has been a consequence of phenomena discussed in *Yao et al [2007]*).
- Also the bounds on K being finite, if loosened would be interesting to look at.

Final Thoughts

Discussions in Class

- We discussed the general problem of Kernel Ridge Regression in class, but did not talk about empirical worst case bounds.
- This paper spoke about a modification of this algorithm while introducing a certain degree of randomness.
- Through this paper, we learnt that such analyses are necessary since error bounds can be quite deviant in terms of larger problems with complex features and data.

Conclusions

- The use of Kernels and in particular the randomness introduced helps with such complexity and deviation.
- Since in most classification tasks, the ground truth is set by humans, there is a close to no chance in error, so researchers should also focus on the case with $\mathcal{R}(f_\rho) = 0$ instead of focusing only on cases that make the analysis mathematically complete.

*****