Recent work

and my reading on deep learning theory

Yukun Dong

dyk2021@mail.ustc.edu.cn

May 30, 2024

Section 1

Results on kernel machines/interpolators

Problem settings

We observe i.i.d pairs of data (x_i, y_i), where x_i are the covariates in the compact domain $\Omega \subset \mathbb{R}^d$ and $y_i \in \mathbb{R}$ are the corresponding labels. Suppose the n pairs are drawn from an unknown probability distribution $\mu(x, y)$.
We are interested in estimating the conditional expectation function $f_*(x) = \mathbb{E}(y|X = x)$, which is assumed to lie in a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$.

Conventional wisdom: kernel ridge regression

According to traditional statistical wisdom, explicit regularization should be added to the least-squares objective

$$\widehat{f}_{n,\lambda}(x) = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - y_i\right)^2 + \lambda \|f\|_{\mathcal{H}}^2.$$

The solution can be explicitly written as

$$\widehat{f}_{n,\lambda}(x) = \mathbb{K}(x, X) \left\{\mathbb{K}(X, X) + n\lambda\right\}^{-1} Y.$$

Min-norm interpolate solution

C. Zhang et al. (2017) simply fitted a model using linear kernel that interpolates the MNIST dataset, and observed that the model achieved 1.2% testing error. Furthermore, adding regularization does not improve model's performance.
This phenomenon lead more research in ridgeless interpolators. It is well known that the solution can be written as:

$$\widehat{f}_{\text{inter}}(x) = \mathbb{K}(x, X)\mathbb{K}(X, X)^{-1}Y$$

Bounds for kernel interpolators

- Rudi, Camoriano, and Rosasco (2016) provides bound on the excess risk of KRR estimator $\widehat{f}_{n,\lambda}$.

Theorem 1 (Rudi, Camoriano, and Rosasco 2016)

With high probability,

$$L(\widehat{f}_{n,\lambda}) - L(f_*) \lesssim \frac{\|\widehat{f}_{n,\lambda}\|_{\mathcal{H}}^{\alpha}}{n^{\beta}}$$

for some constants α, β.

Bounds for kernel interpolators(Continued)

- Belkin, Ma, and Mandal (2018) illustrates the strong generalization performance of interpolated classifiers. But norm-based concentration bounds fails to explain this phenomenon and a new theory of kernel methods is needed to understand this behavior.

Theorem 1 (Belkin, Ma, and Mandal 2018)

With high probability, any h that t-overfits the data, satisfies

$$\|f\|_{\mathcal{H}} > Ae^{Bn^{1/d}}$$

for some constants $A, B > 0$ depending on t.

Bounds by Liang and Rakhlin (2020)

- Liang and Rakhlin (2020) showed that kernel interpolation can generalize when $d \asymp n$. Note that the bound is data-dependent.

Assumptions

- $\|\Sigma_d\|_{op} \leq 1$, where $\Sigma_d = \mathrm{Var}(X)$.
- X has $8 + m$ moments for some $m > 0$.
- $\mathbb{E}[X_i] = 0$.
- $\sup_{x \in \mathbb{R}^d} \mathrm{Var}(Y \mid X = x) \leq \sigma^2$ for some $\sigma > 0$.
- $c < \frac{d}{n} < C \quad \text{for} \quad c, C \in (0, \infty)$.

Bounds by Liang and Rakhlin (2020)(Continued)

Theorem 1 (Liang and Rakhlin 2020)

Under the above assumptions,

$$L(\hat{f}_n) - L(f_*) \leq \phi_{n,d}(X, f_*) + \epsilon_{n,d},$$

where

$$\phi_{n,d}(X, f_*) = \frac{8\sigma^2 \|\Sigma_d\|_{op}}{d} \sum_j \frac{\lambda_j\left(\frac{XX^\top}{d} + \frac{\alpha}{\beta} 11^\top\right)}{\left[\frac{\gamma}{\beta} + \lambda_j\left(\frac{XX^\top}{d} + \frac{\alpha}{\beta} 11^\top\right)\right]^2} +$$

$$\|f_*\|_{\mathcal{H}}^2 \inf_{0 \leq k \leq n} \left\{ \frac{1}{n} \sum_{j=k}^{n} \lambda_j(K(X, X)) + 2M\sqrt{\frac{k}{n}} \right\},$$

Bounds by Liang and Rakhlin (2020)(Continued)

Theorem 1 (Liang and Rakhlin 2020)(Continued)

Here the error term is

$$\epsilon_{n,d} = O\left(d^{-m/(m+8)} \log^{4.1} d\right) + O\left(n^{-1/2} \log^{0.5} n\right).$$

The curvature-related parameters are

$$\alpha := g(0) + g''(0)\frac{\mathrm{Tr}(\Sigma_d^2)}{d^2}, \quad \beta := g'(0),$$

$$\gamma := g\left(\frac{\mathrm{Tr}(\Sigma_d)}{d}\right) - g(0) - g'(0)\frac{\mathrm{Tr}(\Sigma_d)}{d}.$$

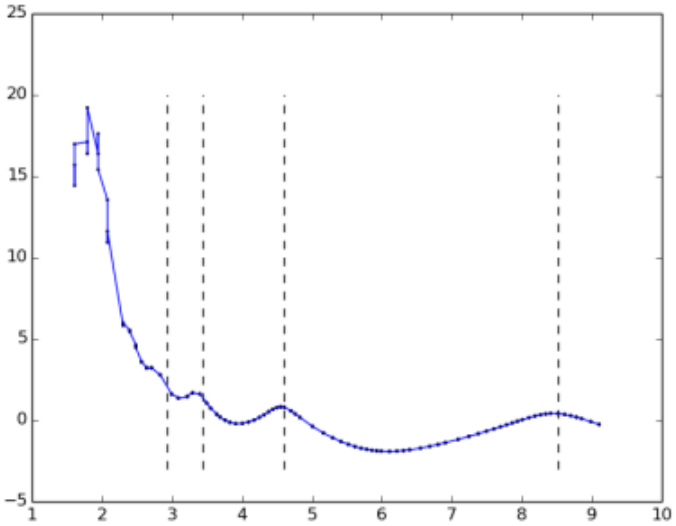
Few remarks on Liang and Rakhlin (2020)

- The bounds only work under regimes where $d \asymp n$. Its mechanism of implicit regularization relies on high dimensionality d of the input space.
- Non-linearity of g is crucial. Results of Theorem 1 still holds when g is the RBF kernel.
- Bias-variance trade-off: Fast eigenvalue decay leads to insufficient regularization, indicating large variance; Slow eigenvalue decay brings about too much regularization, inducing large bias.

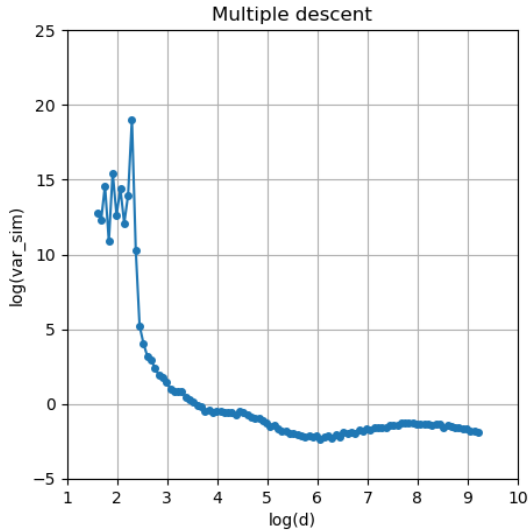
Other bounds

- (fixed input dimension case) Rakhlin and Zhai (2018) suggest that minimum-norm interpolant does not appear to perform well in low dimensions, by studying the case with Laplacian kernels. Li, H. Zhang, and Lin (2023) proved more general results in fixed dimension settings.
- $(d \asymp n^{\alpha}, \alpha \in (0,1))$ upper bounds on the risk are of a multiple-descent shape. (Liang, Rakhlin, and Zhai 2020)

Multiple descent



(a) Origin



(b) My

Figure: Multiple descent with inner product kernel

Kernel machines: computational perspective

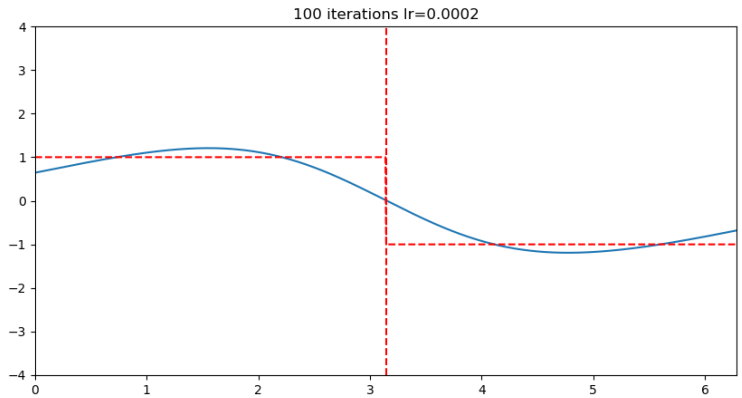
I refer to M.Belkin's series of EigenPro here.
Ma and Belkin (2017) show extremely slow convergence of gradient descent on kernel interpolation setting.

Corollary 1 (Ma and Belkin 2017)

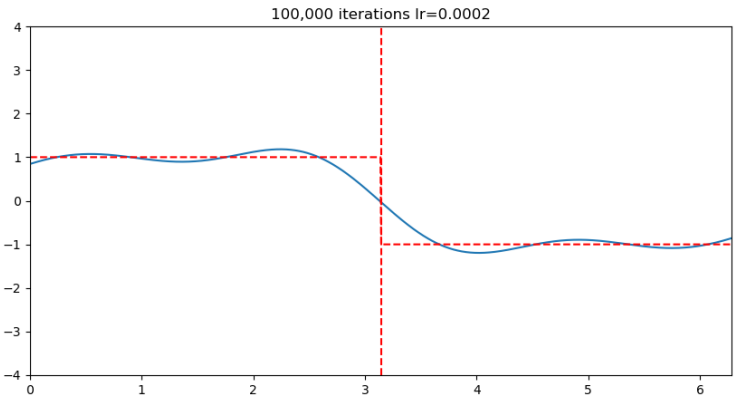
Any $f \in L^2(\Omega)$ that for any $\epsilon > 0$ can be ϵ-approximated with polynomial in $1/\epsilon$ number of steps of gradient descent is infinitely differentiable. Thus, if f is not infinitely differentiable it cannot be ϵ-approximated in $L^2(\Omega)$ using a polynomial number of gradient descent steps.

Limits of computational reach of GD methods

An example considers the Heaviside step function $f(x)$, taking 1 and -1 values for $x \in (0, \pi]$ and $x \in (\pi, 2\pi]$, respectively.



(a) 100 iters



(b) 100,000 iters

The approximation for the Heaviside function is only marginally improved by going from 100 to 10^6 iterations of gradient descent.

EigenPro iterations

- Ma and Belkin (2017) proposed EigenPro, using a left preconditioner to reduce the top k eigenvalues of covariance matrix $H = \sum_{i=1}^{n} x_i x_i^T$.
- Ma and Belkin (2019) designed EigenPro2.0 for a class of classical kernel machines. This work extends linear scaling to match the parallel computing capacity of a resource.
- Abedsoltan, Belkin, and Pandit (2023) introduced EigenPro 3.0, an algorithm based on projected dual preconditioned SGD. This enables kernel machines to scale to large datasets.

Something interesting but not yet mentioned

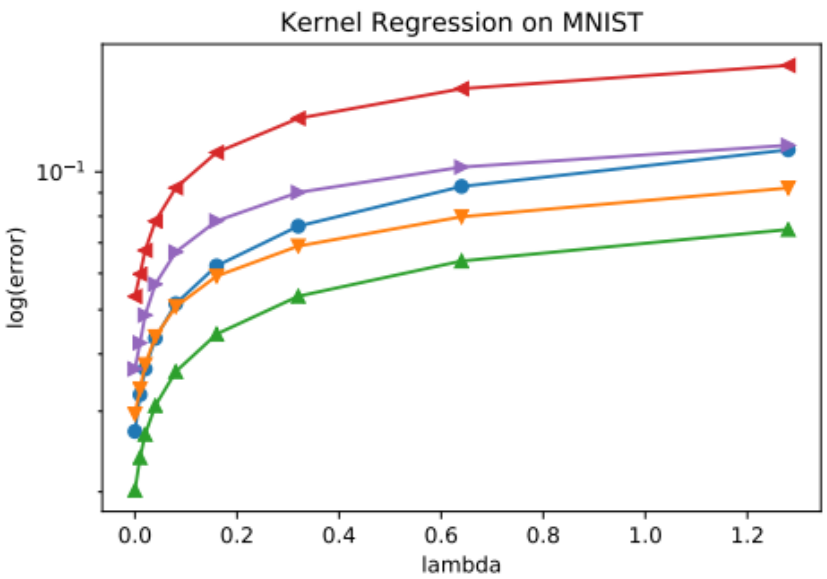
- NTK, RMT, empirical process
- Mean-field theory
- DL-GP
- Spectral complexity of deep neural networks. (Lillo et al. 2024)
- Covariate shift. (Ge et al. 2023)
- Inductive bias.
- Double & multiple descent.
- Moreau Envelope generalization theory.
- Optimistic rates. (Zhou et al. 2021)

Section 2

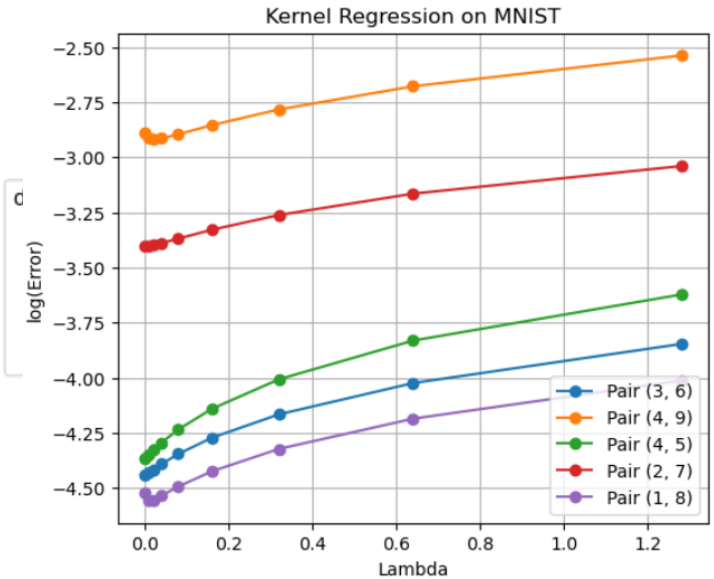
My recent work

Recent work

- Pytorch realization of paper Liang and Rakhlin (2020).



(a) Origin

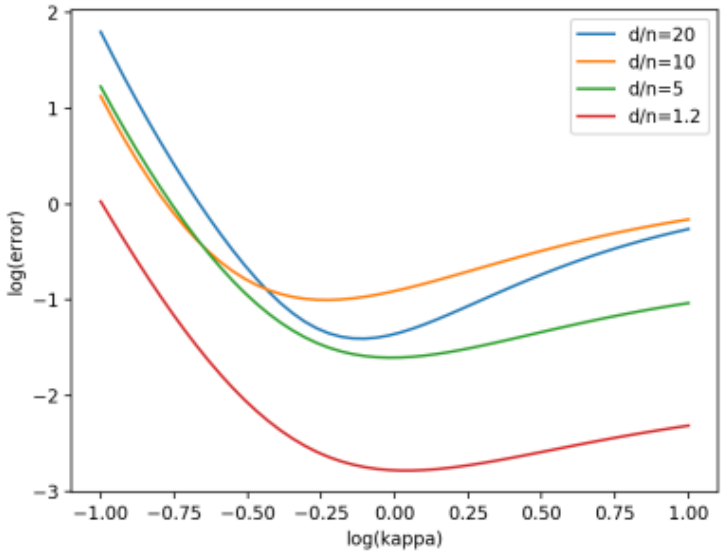


(b) My

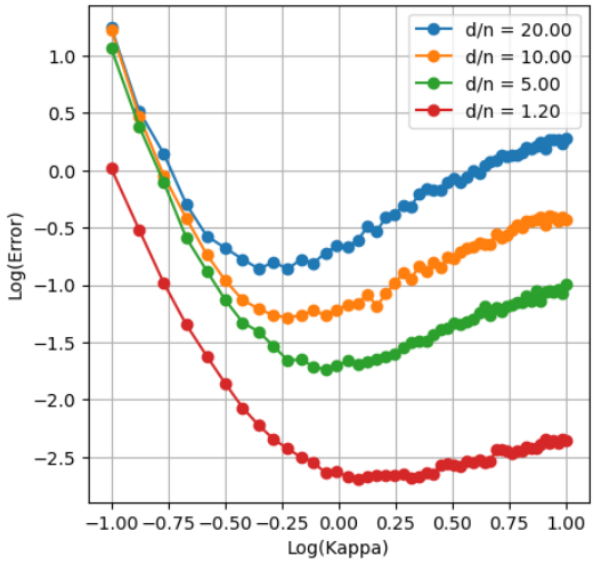
Figure: MNIST experiment

Recent work

- Pytorch realization of paper Liang and Rakhlin (2020).



(a) Origin



(b) My

Figure: Synthetic dataset

References I


Abedsoltan, Amirhesam, Mikhail Belkin, and Parthe Pandit (23–29 Jul 2023). "Toward Large Kernel Models". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 61–78. URL: https://proceedings.mlr.press/v202/abedsoltan23a.html.


Belkin, Mikhail, Siyuan Ma, and Soumik Mandal (Oct. 2018). "To Understand Deep Learning We Need to Understand Kernel Learning". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 541–549. URL: https://proceedings.mlr.press/v80/belkin18a.html.

References II

- Ge, Jiawei et al. (2023). *Maximum Likelihood Estimation is All You Need for Well-Specified Covariate Shift*. arXiv: 2311.15961 [stat.ML].
- Li, Yicheng, Haobo Zhang, and Qian Lin (Aug. 2023). "Kernel interpolation generalizes poorly". In: *Biometrika* 111.2, pp. 715–722. ISSN: 1464-3510. DOI: 10.1093/biomet/asad048. URL: http://dx.doi.org/10.1093/biomet/asad048.
- Liang, Tengyuan and Alexander Rakhlin (June 2020). "Just interpolate: Kernel "Ridgeless" regression can generalize". In: *The Annals of Statistics* 48.3. ISSN: 0090-5364. DOI: 10.1214/19-aos1849. URL: http://dx.doi.org/10.1214/19-AOS1849.

References III

- Liang, Tengyuan, Alexander Rakhlin, and Xiyu Zhai (2020). *On the Multiple Descent of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels*. arXiv: 1908.10292 [math.ST].
- Lillo, Simmaco Di et al. (2024). *Spectral complexity of deep neural networks*. arXiv: 2405.09541 [stat.ML].
- Ma, Siyuan and Mikhail Belkin (2017). *Diving into the shallows: a computational perspective on large-scale shallow learning*. arXiv: 1703.10622 [stat.ML].
- — (2019). "Kernel machines that adapt to GPUs for effective large batch training". In: *Proceedings of Machine Learning and Systems* 1, pp. 360–373.

References IV

- Rakhlin, Alexander and Xiyu Zhai (2018). *Consistency of Interpolation with Laplace Kernels is a High-Dimensional Phenomenon*. arXiv: 1812.11167 [stat.ML].
- Rudi, Alessandro, Raffaello Camoriano, and Lorenzo Rosasco (2016). *Less is More: Nyström Computational Regularization*. arXiv: 1507.04717 [stat.ML].
- Zhang, Chiyuan et al. (2017). *Understanding deep learning requires rethinking generalization*. arXiv: 1611.03530 [cs.LG].
- Zhou, Lijia et al. (2021). *Optimistic Rates: A Unifying Theory for Interpolation Learning and Regularization in Linear Regression*. arXiv: 2112.04470 [stat.ML].