Smoothed discrepancy principle as an early stopping rule in RKHS

Yaroslav Averyanov [1] & Alain Celisse [2]

[1] *Inria Lille-Nord Europe, yaroslav.averyanov@inria.fr*
[2] *Inria Lille-Nord Europe, alain.celisse@inria.fr*

Abstract. In this paper we work on the estimation of a regression function that belongs to a polynomial decay reproducing kernel Hilbert space (RKHS). We describe spectral filter framework for our estimator that allows us to deal with several iterative algorithms: gradient descent, Tikhonov regularization, etc. The main goal of the paper is to propose a new early stopping rule by introducing smoothing parameter for empirical risk of the estimator in order to improve the previous results [1] on discrepancy principle. Theoretical justifications as well as simulations experiments for the proposed rule are provided.

Keywords. Non-parametric regression, regularization, kernels, stopping rules.

Résumé. Dans ce travail, nous présentons, dans un cadre général, l'estimation de la fonction de régression lorsqu'elle appartient à un RKHS. Les propriétés de plusieurs estimateurs sont analysées à travers des algorithmes itératifs comme la descente de gradient et la régularisation de type Tikhonov. L'objectif principal de notre analyse est de proposer une nouvelle règle d'arrêt prématuré des algorithmes basée sur l'introduction d'un paramètre de lissage dans la définition du risque empirique. Nous illustrons l'efficacité de notre approche et présentons les résultats d'une étude de simulation.

Mots-clés. Régression non-paramétrique, régularisation, noyaux, règles d'arrêt

1 Introduction

In supervised learning, given a sample of pairs of inputs and outputs, the goal is to estimate a regression function in the framework of empirical risk minimization or Tikhonov regularization. Usually properties of the regression function is not known, therefore one can apply different nonparametric techniques to relax this difficulty. Kernel methods [2] are one of the most widely used approaches to learning.

Early stopping rule (ESR) is an algorithmic approach to the regularization of an iterative algorithm such as (stochastic) gradient descent [3], boosting algorithms [4] or EM algorithm [5]. It is based on an idea of stopping iterative process according to a special criterion in order to reach the best statistical precision. ESR has a fairly long history and was first introduced for neural networks [6].

There have been three principal strategies for designing an ESR for a regression function learning. The first one is based on expanding the value of the risk error into Taylor series and optimizing each term of the series. The second one is decomposing the risk error into bias and variance parts and to obtain their high probability upper bounds. At the end, the stopping rule will be defined according to a criterion of the intersection of these two bounds. Several results have been derived regarding this strategy to quantify the ESR performance in the reproducing kernel Hilbert space (RKHS). For example, [7] derived a stopping rule t_w when the regression function belongs to RKHS $\mathcal{H}$. If one stops the learning process at this iteration, the minimax optimal rate for the risk error is achieved for a wide class of functions. The main deficiency of this method is that it requires an accurate upper bound of the regression function in $\mathcal{H}$. The third and the recent one strategy consists of designing an ESR by observing empirical risk and building a threshold for stopping appropriately an iterative process (so-called discrepancy principle). This approach was developed initially by [1] where authors analyzed the behaviour of discrepancy principle for spectral filter algorithms in linear regression model that was further expanded to kernel framework.

In the present work we keep the same spectral framework as [1] by considering gradient descent and kernel ridge regression (or Tikhonov regularization) algorithms. More precisely, we focus on the nonlinear regression function estimation using polynomial decay reproducing decay kernels. We introduce a smoothing parameter for empirical risk, modify previously designed discrepancy principle rule and prove optimality results in terms of L_2-error and as an oracle inequality for a regression function estimator stopped at new rule.

The organization of the paper is as follows. Section 2 introduces a statistical framework and the spectral filter estimator. Section 3 describes the main theoretical results achieved. Section 4 shows the behaviour of the derived ESR in simulations.

2 Statistical framework

Let us assume we have a sample $z_i = (x_i, y_i) \sim \mathbb{P},\ i = 1, ..., n$ with $x_i \in \mathcal{X}$ and $y_i \in \mathbb{R}$ and we consider the usual regression model:

$$Y_i = f^*(x_i) + \sigma\epsilon_i, \quad i = 1, ..., n,$$

where ϵ_i are i.i.d. $\mathcal{N}(0, 1)$ random variables and $\sigma = \text{const}$ is known. There is a large body of work on estimating the noise variance σ in non-parametric regression. In other words

$$Y = F^* + \sigma\epsilon \in \mathbb{R}^n$$

Introducing now $(\hat{\mu}_1, \ldots, \hat{\mu}_n)$ and $(u_1, \ldots, u_n)$ as the eigenvalues and eigenvectors of normalized Gram matrix $K = \mathbb{K}(x_i, x_j)/n$ respectively, where $\mathbb{K}(\cdot, \cdot)$ denotes the reproducing kernel associated with the reproducing kernel Hilbert space $\mathcal{H}$ [8]. Let us further assume

$$\hat{\mu}_1 \geq \hat{\mu}_2 \geq \ldots \geq \hat{\mu}_r > 0 = \hat{\mu}_{r+1} = \hat{\mu}_{r+2} = \ldots = \hat{\mu}_n.$$

We assume that $f^* \in \mathcal{H}$ therefore we would like to use an iterative algorithm to solve

$$\inf_{f \in \mathcal{H}} \{\frac{1}{n} \sum_{i=1}^{n} (y_i - f(X_i))^2\} = \min_{\theta \in \mathbb{R}^n} \|Y - K\theta\|_n^2, \tag{1}$$

by the representer theorem. Then projecting $F^t = K\theta^t$, F^* , Y and ϵ onto the space spanned by $(u_1, \ldots, u_r)$, the r first eigenvectors of K, gives us

$$Z_i = G_i^* + \sigma\tilde{\epsilon}_i, \quad i = 1, \ldots, r$$

Here we used the fact that $G_i^* = \langle F^*, u_i \rangle = 0$ when $i > r$ since $f^* \in \mathcal{H}$.

A non-negative function $\gamma^{(t)} \in \mathbb{R}^r$ is called a spectral filter if it is a non-decreasing function of t (in each of its coordinates), $\gamma_i^{(0)} = 0$ and $\lim_{t \to \infty} \gamma_i^{(t)} = 1$. Several iterative algorithms could be expressed in terms of spectral filter as

$$(G^t)_i = \begin{cases} \gamma_i^{(t)} Z_i, & \text{if } i = 1, \ldots, r \\ 0, & \text{if } i = r+1, \ldots, n \end{cases}.$$

Two examples that we study in this paper:

- Gradient descent (GD) with a constant step-size α: $\gamma_i^{(t)} = 1 - (1 - \alpha\hat{\mu}_i)^t$
- (Iterative) kernel ridge regression (KRR) with a parameter α: $\gamma_i^{(t)} = \frac{\hat{\mu}_i}{\hat{\mu}_i + \lambda_t}$, where $\lambda_t = 1/(\alpha t)$ for linear parameterization case or $\lambda_t = 1/(e^{\alpha t} - 1)$ for exponential parameterization case.

3 Main results

3.1 Previous stopping rule definition

Considering the risk of F^t we can define its bias and variance:

$$\mathbb{E}_\epsilon||F^t - F^*||_n^2 = ||\mathbb{E}_\epsilon F^t - F^*||_n^2 + \mathbb{E}_\epsilon||F^t - \mathbb{E}_\epsilon F_t||_n^2 = B_t^2 + \mathbb{E}_\epsilon V_t,$$
$$B_t^2 = \frac{1}{n}\sum_{i=1}^{r}(1-\gamma_i^{(t)})^2(G_i^*)^2, \quad \mathbb{E}_\epsilon V_t = \frac{\sigma^2}{n}\sum_{i=1}^{r}(\gamma_i^{(t)})^2$$

Bias is a non-increasing convex functions converging to zero and variance is a non-decreasing function converging to $\frac{r\sigma^2}{n}$. Ideally we would like to be able to minimize the risk as a function of t. Actually, this is not possible because it depends on the unknown distribution. Therefore we define empirical risk, a non-increasing convex function converging to zero.

$$R_t = \frac{1}{n}||F^t - Y||_2^2 = \frac{1}{n}\sum_{i=1}^{r}(1-\gamma_i^{(t)})^2 Z_i^2$$

A stopping rule that was designed in [1] consists of properly setting a threshold for empirical risk:

$$\tau = \inf\{t \geq 0 : R_t \leq \frac{r\sigma^2}{n}\} \tag{2}$$

The following theorem describes the performance of τ compared to the global optimum performance:

Theorem 3.1 *For gradient descent and kernel ridge regression filters there exist constants $C_1 \geq 2$ and $C_2 > 0$:*

$$\mathbb{E}_\epsilon||F^\tau - F^*||_n^2 \leq C_1 \inf_{t\geq 0}\Big[\mathbb{E}_\epsilon||F^t - F^*||_n^2\Big] + C_2\frac{\sqrt{r}}{n}$$

Here constants C_1 and C_2 do not depend on the number of samples n. This theorem shows that if our kernel is a finite-rank one then the remainder term is of order $\mathcal{O}(\frac{1}{n})$ and it is converging faster that the optimal value of the risk error. However if we assume that rank of the kernel depends on the number of samples, e.g. for Sobolev kernel $r = n$, then the remainder term of the oracle-type inequality has a slow convergence rate (for Sobolev kernel $\mathcal{O}(\frac{1}{\sqrt{n}})$). Moreover, it appeared that τ, since it is a random quantity itself, has a large variance. Therefore we suggest to use a smoothed version of bias/variance and empirical risk by means of the eigenvalues of Gram matrix and a smoothing parameter $\theta \in [0, 1]$:

$$B_{\theta,t} = \frac{1}{n}\sum_{i=1}^{r}\hat{\mu}_i^\theta(1-\gamma_i^{(t)})^2(G_i^*)^2, \quad \mathbb{E}_\epsilon V_{\theta,t} = \frac{\sigma^2}{n}\sum_{i=1}^{r}\hat{\mu}_i^\theta(\gamma_i^{(t)})^2,$$
$$R_{\theta,t} = \frac{1}{n}\sum_{i=1}^{r}\hat{\mu}_i^\theta(1-\gamma_i^{(t)})^2 Z_i^2, \quad \theta \in [0,1]$$

3.2 Polynomial decay kernels

Let us at the beginning derive a result that introduces minimax optimal rate for a stopping rule that has $\mathcal{O}(\frac{1}{n})$ threshold for smoothed empirical risk with polynomial decay kernels.

Theorem 3.2 *For any $\gamma > 0$ and $n \geq 16$ let us make the following assumptions:*

- $\sup_{x\in\mathcal{X}} \mathbb{K}(x,x) \le M_K$
- $|Y| \le M$ *a.s.*
- *Let us define the kernel integral operator*

$$B : L_2(\mathbb{P}) \to L_2(\mathbb{P}),\ g \mapsto \int \mathbb{K}(\cdot,x)g(x)dP(x)$$

There exists $w \in L_2(\mathbb{P})$ *such that* $f^* = B^{\nu-\frac{1}{2}}w$ *with* $\|w\|_{L_2(\mathbb{P})} \le M_K^{-\nu}\rho$ *and* $\nu \ge \frac{1}{2}$

- *Given two parameters* $s \in (0,1)$ *and* $D \ge 1$ *we consider the polynomial effective dimensionality*

$$\mathcal{N}(\lambda) = Tr(B(B+\lambda I)^{-1}) \le D^2(M_K^{-1}\lambda)^{-s}$$

This notion was first introduced by [9] in a learning context, and used in a number of works since. This assumption is tightly connected with the decay of the eigenvalues of the kernel integral operator B*: if the eigenvalues of the kernel integral operator has a decay* $\mu_i \asymp i^{-\frac{1}{s}}$ *than the condition on the effective dimensionality holds true with a parameter* s*.*

- *If we define an ESR* $\hat{t}_o = \inf\{t > 0 : R_{\theta,t} \le C(\rho, D, M, M_K, \nu, \theta, \gamma)\ n^{\frac{2\nu+\theta}{2\nu+s}}\}$ *for gradient descent and kernel ridge regression filters and we take the smoothing parameter* $\theta = s$ *then*

$$\|f^{\hat{t}_o} - f^*\|_{L_2(\mathbb{P})} \lesssim n^{-\frac{\nu}{2\nu+s}} \text{ with probability } 1-\gamma$$

The rate achieved for the L_2*-error in the theorem is proved to be minimax-optimal(see e.g. [10])*

The theorem shows that, if we choose the smoothing parameter $\theta = s$, our strategy $\hat{t}_o$ will be optimal in minimax sense. Since in practice we do not have an access to the eigenvalues of the kernel integral operator, we propose to use the inverse decay of the eigenvalues of Gram matrix as an estimation of the optimal parameter θ. Since in the definition of $\hat{t}_o$ the constant $C(\rho, D, M, M_K, \nu, \theta, \gamma)$ is non-computable in practice we propose to consider the following stopping rule $\hat{t}$ where the threshold for smoothed empirical risk is of order $\mathcal{O}(\frac{\mathrm{tr}(K_n^\theta)}{n}) = \mathcal{O}(\frac{\log n}{n})$. This stopping rule aims to estimate an iteration of the intersection of smoothed bias and smoothed variance.

$$\hat{t} = \inf\{t > 0 : R_{\theta,t} \le \frac{\sigma^2}{n}\sum_{i=1}^{r} \hat{\mu_i}^\theta \left[(\gamma_i^{(t)})^2 + (1-\gamma_i^{(t)})^2\right]\} \tag{3}$$

Theorem 3.3 *For gradient descent and kernel ridge regression filters there exist constants* $\hat{C}_1 \ge 4$ *and* $\hat{C}_2 > 0$ *such that*

$$\mathbb{E}_\epsilon\|F^{\hat{t}} - F^*\|_n^2 \le \hat{C}_1 \inf_{t>0}\left[\mathbb{E}_\epsilon\|F^t - F^*\|_n^2\right] + \frac{\hat{C}_2}{n}$$

Here constants $\hat{C}_1$ and $\hat{C}_2$ do not depend on the number of samples n so, for this oracle-type inequality, we achieve $\mathcal{O}(\frac{1}{n})$ rate for the remainder term in the right hand side of the inequality.

4 Simulations

We perform simulations experiments on a simple problem: for the regression model $y_i = f^*(x_i) + \sigma\epsilon_i,\ i = 1, ..., n$ where $\epsilon_i \sim \mathcal{N}(0,1)$ and $\sigma = 0.2$ we use fixed design setting $x_i = i/n$. We implement gradient descent and kernel ridge regression algorithms for fixed step-size and fixed parameter α. We choose the

regression function to be either a smooth function $f^*(x) = -0.5 \sin\left[3(x-2)\right]$ (SF) or a piecewise linear function $f^*(x) = |x - 0.5| - 0.5$ (WF) and polynomial decay Sobolev kernel $\mathbb{K}(x_1, x_2) = \min(x_1, x_2)$. We would like to compare our stopping rule to the previous discrepancy principle stopping rule τ described in (2), to another stopping rule t_w [7], that provides state-of-the-art results for gradient descent and kernel ridge regression algorithms and is based on upper bounding bias and variance with high probabilities, as well as to the oracle method, that requires knowledge of f^*, therefore non-computable in practice.

$$t_{\text{or}} = \inf_{t \geq 0} \left[\mathbb{E}_\epsilon \|F^t - F^*\|_n^2 \right] \tag{4}$$

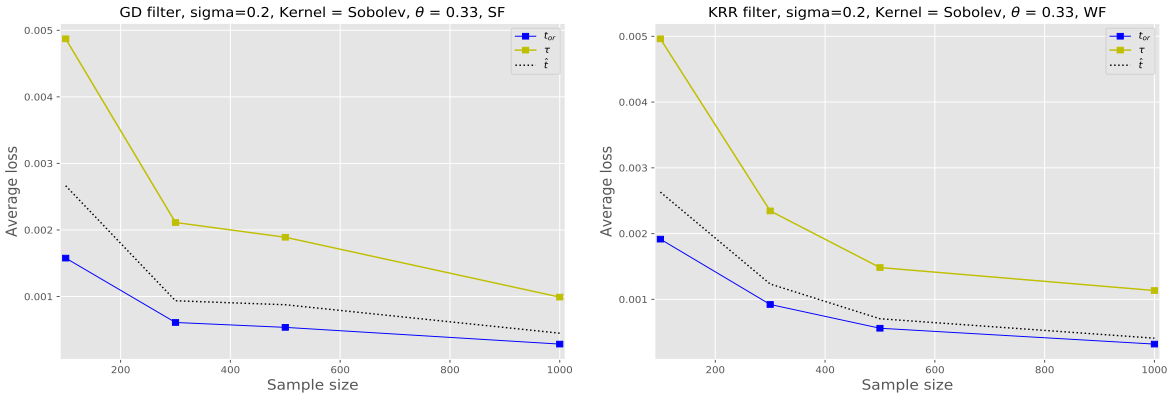

Figure 1: We choose Sobolev kernel, noise level $\sigma = 0.2$ and apply GD and KRR (linear parameterization) filters with $\alpha = 0.5$ for (SF) and (WF) regression functions. Smoothing parameter θ is chosen to be equal to inverse estimation of the decay of the eigenvalues of Gram matrix. Each curve for both of two graphs corresponds to the mean-squared error of a spectral filter estimator, stopped at t_{or}, τ and $\hat{t}$, averaging over 100 independent trials.

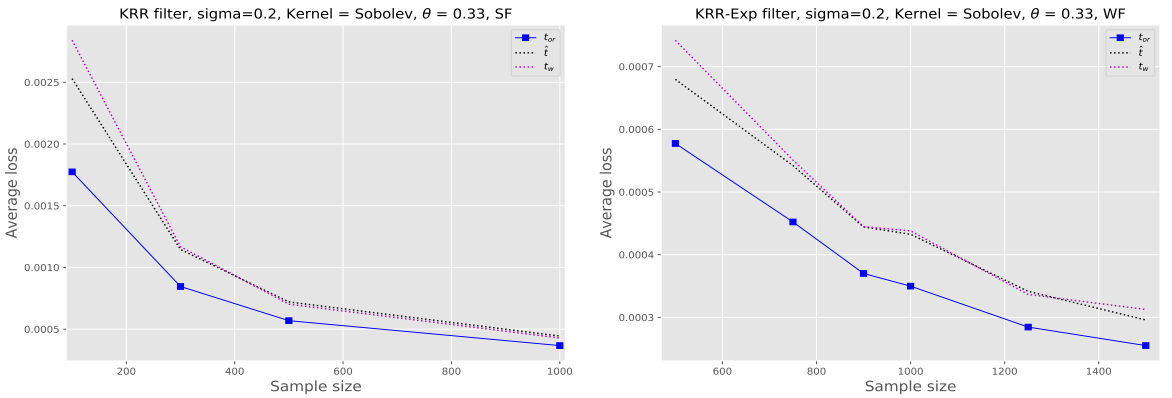

Figure 2: We choose Sobolev kernel, noise level $\sigma = 0.2$ and apply KRR (linear parameterization) and KRR (exponential parameterization) filters with $\alpha = 0.5$ for (SF) and (WF) regression functions. Smoothing parameter θ is chosen to be equal to inverse estimation of the decay of the eigenvalues of Gram matrix. Each curve for both of two graphs corresponds to the mean-squared error of a spectral filter estimator, stopped at t_{or}, t_w and $\hat{t}$, averaging over 100 independent trials.

Figure 1 compares the resulting mean-squared errors of our stopping rule (3), the previous discrepancy principle rule (2) and the oracle stopping rule (4). The new proposed rule exhibits better performance than (2) for all sample sizes. Figure 2 compares the resulting mean-squared errors of our stopping rule (3), the state-of-the-art stopping rule t_w [7] and the oracle stopping rule (4). The new proposed rule exhibits better performance than t_w for a sample size $n < 300$ for KRR (linear parameterization) filter and for a sample size $n < 800$ for KRR (exponential parameterization) filter. Nevertheless, we observe the same asymptotic behaviour of t_w and $\hat{t}$. Since the rule t_w is proved to be minimax optimal in a functional space generated by Sobolev kernels, we can conclude that $\hat{t}$ recovers the same rate in simulations.

5 Conclusion

In this paper we described spectral filter algorithms (gradient descent, Tikhonov regularization) for non-parametric regression function estimation in RKHS. We proposed a new early stopping rule $\hat{t}$ for these algorithms. After that we proved an optimal result in L_2-error and an oracle-type inequality for the developed rule. At the end of the paper we showed the performance of $\hat{t}$ in simulations. The main deficiency of our strategy is that the construction of $\hat{t}$ is based on the assumption that the regression function belongs to a known RKHS and that the results were derived only for polynomial decay kernels.

Bibliography

[1] Gilles Blanchard, Marc Hoffmann, and Markus Reiß. Optimal adaptation for early stopping in statistical inverse problems. *arXiv preprint arXiv:1606.07702*, 2016.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[3] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[4] Tong Zhang, Bin Yu, et al. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.

[5] Gilles Celeux, Didier Chauveau, and Jean Diebolt. *On stochastic versions of the EM algorithm*. PhD thesis, INRIA, 1995.

[6] Lutz Prechelt. Early stopping - but when? In *Neural Networks: Tricks of the Trade, volume 1524 of LNCS, chapter 2*, pages 55–69. Springer-Verlag, 1997.

[7] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(1):335–366, 2014.

[8] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[9] Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.

[10] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.