

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/283079574>

# NYTRO: When Subsampling Meets Early Stopping

Conference Paper · January 2016

CITATIONS

9

READS

59

4 authors:



**Tomás Angles**

Ecole Normale Supérieure de Paris

4 PUBLICATIONS 9 CITATIONS

SEE PROFILE



**Raffaello Camoriano**

Istituto Italiano di Tecnologia

17 PUBLICATIONS 189 CITATIONS

SEE PROFILE



**Alessandro Rudi**

INRIA, Ecole Normale Supérieure de Paris

34 PUBLICATIONS 256 CITATIONS

SEE PROFILE



**Lorenzo Rosasco**

Massachusetts Institute of Technology

175 PUBLICATIONS 3,266 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Theory of Deep Learning: [View project](#)

# NYTRO: When Subsampling Meets Early Stopping

Tomas Angles\*<sup>1</sup>, Raffaello Camoriano<sup>o\*2,3</sup>, Alessandro Rudi<sup>3</sup>, Lorenzo Rosasco<sup>1,3</sup>

<sup>1</sup> Massachusetts Institute of Technology and Istituto Italiano di Tecnologia  
Laboratory for Computational and Statistical Learning, Cambridge, MA 02139, USA  
{ale\_rudi, lrosasco}@mit.edu

<sup>2</sup> Istituto Italiano di Tecnologia  
iCub facility, Via Morego 30, Genova, Italy  
raffaello.camoriano@iit.it

<sup>3</sup> Università degli Studi di Genova  
DIBRIS, Via Dodecaneso 35, Genova, Italy

October 21, 2015

## Abstract

Early stopping is a well known approach to reduce the time complexity for performing training and model selection of large scale learning machines. On the other hand, memory/space (rather than time) complexity is the main constraint in many applications, and randomized subsampling techniques have been proposed to tackle this issue. In this paper we ask whether early stopping and subsampling ideas can be combined in a fruitful way. We consider the question in a least squares regression setting and propose a form of randomized iterative regularization based on early stopping and subsampling. In this context, we analyze the statistical and computational properties of the proposed method. Theoretical results are complemented and validated by a thorough experimental analysis.

## 1 Introduction

Availability of large scale datasets requires the development of ever more efficient machine learning procedures. A key feature towards scalability is being able to tailor computational requirements to the generalization properties/statistical accuracy allowed by the data. In other words, the precision with which computations need to be performed should be determined not only by the the amount, but also by the quality of the available data.

Early stopping, known as iterative regularization in inverse problem theory (Engl et al., 1996; Zhang and Yu, 2005; Bauer et al., 2007; Yao et al., 2007; Caponnetto and Yao, 2010), provides a simple and sound implementation of this intuition. An empirical objective function is optimized in an iterative way with no explicit constraint or penalization

---

\*The authors have contributed equally.

<sup>o</sup>Corresponding author.

and regularization is achieved by suitably stopping the iteration. Too many iterations might lead to overfitting, while stopping too early might result in oversmoothing (Zhang and Yu, 2005; Bauer et al., 2007; Yao et al., 2007; Caponnetto and Yao, 2010). Then, the best stopping rule arises from a form of bias-variance trade-off (Hastie et al., 2001). Towards the discussion in the paper, the key observation is that the number of iterations controls at the same time the computational complexity as well as the statistical properties of the obtained learning algorithm (Yao et al., 2007). Training and model selection can hence be performed with often considerable gain in time complexity.

Despite these nice properties, early stopping procedures often share the same space complexity requirements, hence bottle necks, of other methods, such as those based on variational regularization *à la Tikhonov* (see Tikhonov, 1963; Hoerl and Kennard, 1970). A natural way to tackle these issues is to consider randomized subsampling/sketching approaches. Roughly speaking, these methods achieve memory and time savings by reducing the size of the problem in a stochastic way (Smola and Schölkopf, 2000; Williams and Seeger, 2000). Subsampling methods are typically used successfully together with penalized regularization. In particular, they are popular in the context of kernel methods, where they are often referred to as Nyström approaches and provide one of the main methods towards large scale extensions (Smola and Schölkopf, 2000; Williams and Seeger, 2000; Zhang et al., 2008; Kumar et al., 2009; Li et al., 2010; Dai et al., 2014; Huang et al., 2014; Si et al., 2014).

In this paper, we ask whether early stopping and subsampling methods can be fruitfully combined. With the context of kernel methods in mind, we propose and study NYTRO (NYström iTerative RegularizatiOn), a simple algorithm combining these two ideas. After recalling the properties and advantages of different regularization approaches in Section 2, in Section 3 we present in detail NYTRO and our main result, the characterization of its generalization properties. In particular, we analyze the conditions under which it attains the same statistical properties of subsampling and early stopping. Indeed, our study shows that while both techniques share similar, optimal, statistical properties, they are computationally advantageous in different regimes and NYTRO outperforms early stopping in the appropriate regime, as discussed in Section 3.3. The theoretical results are validated empirically in Section 4, where NYTRO is shown to provide competitive results even at a fraction of the computational time, on a variety of benchmark datasets.

## 2 Learning and Regularization

In this section we introduce the problem of learning in the fixed design setting and discuss different regularized learning approaches, comparing their statistical and computational properties. This section is a survey that might be interesting in its own right, and reviews several results providing the context for the study in the paper.

### 2.1 The Learning Problem

We introduce the learning setting we consider in the paper. Let  $\mathcal{X} = \mathbb{R}^d$  be the input space and  $\mathcal{Y} \subseteq \mathbb{R}$  the output space. Consider a *fixed design* setting (Bach, 2013) where the input

points  $x_1, \dots, x_n \in \mathcal{X}$  are fixed, while the outputs  $y_1, \dots, y_n \in \mathcal{Y}$  are given by

$$y_i = f_*(x_i) + \epsilon_i, \quad \forall i \in \{1, \dots, n\}$$

where  $f_* : \mathcal{X} \rightarrow \mathcal{Y}$  is a fixed function and  $\epsilon_1, \dots, \epsilon_n$  are random variables. The latter can be seen as noise and are assumed to be independently and identically distributed according to a probability distribution  $\rho$  with zero mean and variance  $\sigma^2$ . In this context, the goal is to minimize the *expected risk*, that is

$$\min_{f \in \mathcal{H}} \mathcal{E}(f), \quad \mathcal{E}(f) = \mathbb{E} \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2, \quad \forall f \in \mathcal{H}, \quad (1)$$

where  $\mathcal{H}$  is a space of functions, called *hypothesis space*. In a real applications,  $\rho$  and  $f_*$  are unknown and accessible only by means of a single realization  $(x_1, y_1), \dots, (x_n, y_n)$  called *training set* and an approximate solution needs to be found. The quality of a solution  $f$  is measured by the *excess risk*, defined as

$$R(f) = \mathcal{E}(f) - \inf_{v \in \mathcal{H}} \mathcal{E}(v), \quad \forall f \in \mathcal{H}.$$

We next discuss estimation schemes to find a solution and compare their computational and statistical properties.

## 2.2 From (Kernel) Ordinary Least Square to Tikhonov Regularization

A classical approach to derive an empirical solution to Problem (1) is the so called *empirical risk minimization*

$$f_{\text{ols}} = \operatorname{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2. \quad (2)$$

In this paper, we are interested in the case where  $\mathcal{H}$  is the reproducing kernel Hilbert space

$$\mathcal{H} = \overline{\operatorname{span}\{k(x, \cdot) \mid x \in \mathcal{X}\}},$$

induced by a positive definite kernel  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  (see [Schölkopf and Smola, 2002](#)). In this case Problem (6) corresponds to the *Kernel Ordinary Least Squares* (KOLS) and has the closed form solution

$$f_{\text{ols}}(x) = \sum_{i=1}^n \alpha_{\text{ols},i} k(x, x_i), \quad \alpha_{\text{ols}} = K^\dagger \mathbf{y}, \quad (3)$$

for all  $x \in \mathcal{X}$ , where  $(K)^\dagger$  denotes the pseudo-inverse of the  $\in \mathbb{R}^{n \times n}$  empirical kernel matrix  $K_{ij} = k(x_i, x_j)$  and  $\mathbf{y} = (y_1, \dots, y_n)$ . The cost for computing the coefficients  $\alpha_{\text{ols}}$  is  $O(n^2)$  in memory and  $O(n^3 + q(\mathcal{X})n^2)$  in time, where  $q(\mathcal{X})n^2$  is the cost for computing  $K$  and  $n^3$  the cost for obtaining its pseudo-inverse. Here  $q(\mathcal{X})$  is the cost of evaluating the kernel function. In the following, we are concerned with the dependence on  $n$  and hence view  $q(\mathcal{X})$  as a constant.

The statistical properties of KOLS, and related methods, can be characterized by suitable notions of *dimension* that we recall next. The simplest is the *full* dimension, that is

$$d^* = \text{rank } K$$

which measures the degrees of freedom of the kernel matrix. This latter quantity might not be stable when  $K$  is ill-conditioned. A more robust notion is provided by the *effective dimension*

$$d_{\text{eff}}(\lambda) = \text{Tr}(K(K + \lambda nI)^{-1}), \quad \lambda > 0.$$

Indeed, the above quantity can be shown to be related to the eigenvalue decay of  $K$  (Bach, 2013; Alaoui and Mahoney, 2014) and can be considerably smaller than  $d^*$ , as discussed in the following. Finally, consider

$$\tilde{d}(\lambda) = n \max_i (K(K + \lambda nI)^{-1})_{ii}, \quad \lambda > 0. \quad (4)$$

It is easy to see that the following inequalities hold,

$$d_{\text{eff}}(\lambda) \leq \tilde{d}(\lambda) \leq 1/\lambda, \quad d_{\text{eff}}(\lambda) \leq d^* \leq n, \quad \forall \lambda > 0.$$

Aside from the above notion of dimensionality, the statistical accuracy of empirical least squares solutions depends on a natural form of signal to noise ratio defined next. Note that the function that minimizes the excess risk in  $\mathcal{H}$  is given by

$$\begin{aligned} f_{\text{opt}} &= \sum_{i=1}^n \alpha_{\text{opt},i} k(x, x_i), \quad \forall x \in \mathcal{X} \\ \alpha_{\text{opt}} &= K^\dagger \mu, \quad \text{with } \mu = \mathbb{E}y. \end{aligned}$$

Then, the signal to noise ratio is defined as

$$\text{SNR} = \frac{\|f_{\text{opt}}\|_{\mathcal{H}}^2}{\sigma^2}. \quad (5)$$

Provided with the above definitions, we can recall a first basic results characterizing the statistical accuracy of KOLS.

**Theorem 1.** *Under the assumptions of Section 2.1, the following equation holds,*

$$\mathbb{E}R(f_{\text{ols}}) = \frac{\sigma^2 d^*}{n}.$$

The above result shows that the excess risk of KOLS can be bounded in terms of the full dimension, the noise level and the number of points. However, in general empirical risk minimization *does not* provide the best results and regularization is needed. We next recall this fact, considering first Tikhonov regularization, that is the *Kernel Regularized Least Squares* (KRLS) algorithm given by,

$$\bar{f}_\lambda = \underset{f \in \mathcal{H}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2. \quad (6)$$

The above algorithm is a penalized empirical risk minimization problem. The *representer theorem* (Schölkopf and Smola, 2002) shows that Problem (6) can be restricted to

$$\mathcal{H}_n = \left\{ \sum_{i=1}^n \alpha_i k(\cdot, x_i) \mid \alpha_1, \dots, \alpha_n \in \mathbb{R} \right\}. \quad (7)$$

Indeed, a direct computation shows that the solution of Problem (6) is

$$\bar{f}_\lambda(x) = \sum_{i=1}^n \bar{\alpha}_{\lambda i} k(x, x_i), \quad \bar{\alpha}_\lambda = (\mathbf{K} + \lambda n \mathbf{I})^{-1} \mathbf{y}, \quad (8)$$

for all  $x \in \mathcal{X}$ . The intuition that regularization can be beneficial is made precise by the following result comparing KOLS and KRLS.

**Theorem 2.** *Let  $\lambda^* = \frac{1}{\text{SNR}}$ . The following inequalities hold,*

$$\mathbb{E}\mathbb{R}(\bar{f}_{\lambda^*}) \leq \frac{\sigma^2 d_{\text{eff}}(\lambda^*)}{n} \leq \frac{\sigma^2 d^*}{n} = \mathbb{E}\mathbb{R}(f_{\text{ols}}).$$

We add a few comments. First, as announced, the above result quantifies the benefits of regularization. Indeed, it shows that there exists a  $\lambda^*$  for which the expected excess risk of KRLS is smaller than the one of KOLS. As discussed in Table 1 of Bach (2013), if  $d^* = n$  and the kernel is sufficiently “rich”, namely universal (Micchelli et al., 2006), then  $d_{\text{eff}}$  can be less than a fractional power of  $d^*$ , so that  $d_{\text{eff}} \ll d^*$  and

$$\mathbb{E}\mathbb{R}(\bar{f}_{\lambda^*}) \ll \mathbb{E}\mathbb{R}(f_{\text{ols}}).$$

Second, note that the choice of the regularization parameter depends on a form of signal to noise ratio, which is usually unknown. In practice, a regularization path<sup>1</sup> is computed and then a model selected or found by aggregation (Hastie et al., 2001). Assuming the selection/aggregation step to have negligible computational cost, the complexity of performing training *and* model selection is then  $O(n^2)$  in memory and  $O(n^3 |\Lambda|)$  in time. These latter requirements can become prohibitive when  $n$  is large and the question is whether the same statistical accuracy of KRLS can be achieved while reducing time/memory requirements.

### 2.3 Early Stopping and Nyström Methods

In this section, we first recall how early stopping regularization allows to achieve the same statistical accuracy of KRLS with potential saving in time complexity. Then, we recall how subsampling ideas can be used in the framework of Tikhonov regularization to reduce the space complexity with no loss of statistical accuracy.

<sup>1</sup>The set of solutions corresponding to regularization parameters in a discrete set  $\Lambda \subset \mathbb{R}$ .

**Iterative Regularization by Early Stopping** The idea is to consider the gradient descent minimization of Problem 3 for a fixed number of steps  $t$ . The corresponding algorithm is then

$$\check{f}_t(x) = \sum_{i=1}^n \check{\alpha}_{t,i} k(x_i, x), \quad (9)$$

$$\check{\alpha}_t = \check{\alpha}_{t-1} - \frac{\gamma}{n} (K \check{\alpha}_{t-1} - y), \quad (10)$$

where  $\gamma < 1/\|K\|$  and  $\check{\alpha}_0 = 0$ . Note that in the above algorithm regularization is not achieved by explicit penalization or imposing constraints, and the only tuning parameter is the number of steps. Indeed, as shown next, the latter controls at the same time the computational complexity and statistical accuracy of the algorithm. The following theorem compares the expected excess error of early stopping with the one of KRLS.

**Theorem 3.** *When  $\gamma < 1/\|K\|$  and  $t \geq 2$  the following holds*

$$\mathbb{E}R(\check{f}_{\gamma,t}) \leq c_t \mathbb{E}R(\bar{f}_{\frac{1}{\gamma t}}).$$

with  $c_t = 4 \left(1 + \frac{1}{t-1}\right)^2 \leq 20$ .

The above theorem follows as a corollary of our main result given in Theorem 5 and recovers results essentially given in [Raskutti et al. \(2014\)](#). Combining the above result with Theorem 2, and setting  $t^* = \frac{1}{\gamma \lambda^*} = \frac{\text{SNR}}{\gamma}$ , we have that

$$\mathbb{E}R(\check{f}_{\gamma,t^*}) \approx \mathbb{E}R(\bar{f}_{\lambda^*}) \leq \mathbb{E}R(f_{\text{ols}}).$$

The statistical accuracy of early stopping is essentially the same as KRLS and can be vastly better than a naïve ERM approach. Note that the cost of computing the best possible solution with early stopping is  $O(n^2 t^*) = O(n^2 \text{SNR})$ . Thus, the computational time of early stopping is proportional to the signal to noise ratio. Hence, it could be much better than KRLS for noisy problems, that is when SNR is small. The main bottle neck of early stopping regularization is that it has the same space requirements of KRLS. Subsampling approaches have been proposed to tackle this issue.

**Subsampling and Regularization** Recall that the solution of the standard KRLS problem belongs to  $\mathcal{H}_n$ . A basic idea (see [Smola and Schölkopf, 2000](#)) is to consider *Nyström KRLS* (NKRLS), restricting Problem (6) to a subspace  $\mathcal{H}_m \subseteq \mathcal{H}_n$  defined as

$$\mathcal{H}_m = \left\{ \sum_{i=1}^m c_i k(\cdot, \tilde{x}_i) \mid c_1, \dots, c_m \in \mathbb{R} \right\}. \quad (11)$$

Here  $M = \{\tilde{x}_1, \dots, \tilde{x}_m\}$  is a subset of the training set and  $m \leq n$ . It is easy to see that the corresponding solution is given by

$$\tilde{f}_{m,\lambda}(x) = \sum_{i=1}^m (\tilde{\alpha}_{m,\lambda})_i k(x, \tilde{x}_i), \quad (12)$$

$$\tilde{\alpha}_{m,\lambda} = (K_{nm}^\top K_{nm} + \lambda n K_{mm})^\dagger K_{nm}^\top y, \quad (13)$$

for all  $x \in \mathcal{X}$ , where  $(\cdot)^\dagger$  is the pseudoinverse,  $\lambda > 0$ ,  $K_{nm} \in \mathbb{R}^{n \times m}$  with  $(K_{nm})_{ij} = k(x_i, \tilde{x}_j)$  and  $K_{mm} \in \mathbb{R}^{m \times m}$  with  $(K_{mm})_{ij} = k(\tilde{x}_i, \tilde{x}_j)$ . A more efficient formulation can also be derived. Indeed, we rewrite Problem (6), restricted to  $\mathcal{H}_m$ , as

$$\tilde{\alpha}_{m,\lambda} = \underset{\alpha \in \mathbb{R}^m}{\operatorname{argmin}} \|K_{nm}\alpha - y\|^2 + \lambda \alpha^\top K_{mm} \alpha \quad (14)$$

$$= \underset{\beta \in \mathbb{R}^k}{\operatorname{R argmin}} \|K_{nm}R\beta - y\|^2 + \lambda \|\beta\|^2 \quad (15)$$

where in the last step we performed the change of variable  $\alpha = R\beta$  where  $R \in \mathbb{R}^{m \times k}$  is a matrix such that  $RR^\top = K_{mm}^\dagger$  and  $k$  is the rank of  $K_{mm}$ . Then, we can obtain the following closed form expression,

$$\tilde{\alpha}_{m,\lambda} = R(A^\top A + \lambda n I)^{-1} A^\top y. \quad (16)$$

(see Prop. 1 in Section B of the appendix for a complete proof). This last formulation is convenient because it is possible to compute  $R$  by  $R = ST^{-1}$  where  $K_{mm} = SD$  is the economic QR decomposition of  $K_{mm}$ , with  $S \in \mathbb{R}^{m \times k}$  such that  $S^\top S = I$ ,  $D \in \mathbb{R}^{k \times m}$  an upper triangular matrix and  $T \in \mathbb{R}^{k \times k}$  an invertible triangular matrix that is the Cholesky decomposition of  $S^\top K_{mm} S$ . Assuming  $k \approx m$ , the complexity of Nyström KRLS is then  $O(nm)$  in space and  $O(nm^2 + m^3|\Lambda|)$  in time. The following known result establishes the statistical accuracy of the solution obtained by suitably choosing the points in  $M$ .

**Theorem 4** (Theorem 1 of Bach (2013)). *Let  $m \leq n$  and  $M = \{\tilde{x}_1, \dots, \tilde{x}_m\}$  be a subset of the training set uniformly chosen at random. Let  $\tilde{f}_{m,\lambda}$  be as in Equation (12) and  $\bar{f}_\lambda$  as in Equation (8) for any  $\lambda > 0$ . Let  $\delta \in (0, 1)$ , when*

$$m \geq \left( \frac{32\tilde{d}(\lambda)}{\delta} + 2 \right) \log \frac{\|K\|n}{\delta\lambda}$$

with  $\tilde{d}(\lambda) = n \sup_{1 \leq i \leq n} (K(K + \lambda n I)^{-1})_{ii}$ , then the following holds

$$\mathbb{E}_M \mathbb{E} \mathbb{R} \left( \tilde{f}_{m,\lambda} \right) \leq (1 + 4\delta) \mathbb{E} \mathbb{R} \left( \bar{f}_\lambda \right).$$

The above result shows that the space/time complexity of NKRLS can be adaptive to the statistical properties of the data while preserving the same statistical accuracy of KRLS. Indeed, using Theorem 2, we have that

$$\mathbb{E}_M \mathbb{E} \mathbb{R} \left( \tilde{f}_{m,\lambda^*} \right) \approx \mathbb{E} \mathbb{R} \left( \bar{f}_{\lambda^*} \right) \leq \mathbb{E} \mathbb{R} \left( f_{\text{ols}} \right),$$

requiring  $O(n\tilde{d}(\lambda^*) \log \frac{n}{\lambda^*})$  in memory and  $O(n\tilde{d}(\lambda^*)^2 (\log \frac{n}{\lambda^*})^2)$  in time. Thus, NKRLS is more efficient with respect to KRLS when  $\tilde{d}(\lambda^*)$  is smaller than  $\frac{n}{\log \frac{n}{\lambda^*}}$ , that is when the problem is mildly complex.

Given the above discussion it is natural to ask whether subsampling and early stopping ideas can be fruitfully combined. Providing a positive answer to this question is the main contribution of this paper that we discuss next.

### 3 Proposed Algorithm and Main Results

We begin by describing the proposed algorithm incorporating the Nyström approach described above in iterative regularization by early stopping. The intuition is that the algorithm thus obtained could have memory and time complexity adapted to the statistical accuracy allowed by the data, while automatically computing the whole regularization path. Indeed, this intuition is then confirmed through a statistical analysis of the corresponding excess risk. Our result indicates in which regimes KRLS, NKRLS, Early Stopping and NYTRO are preferable.

#### 3.1 The Algorithm

NYTRO is obtained considering a finite number of iterations of the gradient descent minimization of the empirical risk in Problem (2) over the space in Equation (11). The algorithm thus obtained is given by,

$$\hat{f}_{m,\gamma,t}(x) = \sum_{i=1}^m (\hat{\alpha}_{m,t})_i k(\tilde{x}_i, x), \quad (17)$$

$$\hat{\beta}_{m,\gamma,t} = \hat{\beta}_{m,t-1} - \frac{\gamma}{n} \mathbf{R}^\top (\mathbf{K}_{nm}^\top (\mathbf{K}_{nm} \hat{\beta}_{m,t-1} - \mathbf{y})), \quad (18)$$

$$\hat{\alpha}_{m,\gamma,t} = \mathbf{R} \beta_{m,t}, \quad (19)$$

for all  $x \in \mathcal{X}$ , where  $\gamma = 1/(\sup_{1 \leq i \leq n} k(x_i, x_i))$  and  $\hat{\beta}_{m,0} = 0$ . Considering that the cost of computing  $\mathbf{R}$  is  $O(m^3)$ , the total cost for the above algorithm is  $O(nm)$  in memory and  $O(nmt + m^3)$  in time.

In the previous section, we have seen that NKRLS has an accuracy comparable to the one of the standard KRLS under a suitable choice of  $m$ . We next show that, under the same conditions, the accuracy of NYTRO is comparable with the ones of KRLS and NKRLS, for suitable choices of  $t$  and  $m$ .

#### 3.2 Error Analysis

We next establish excess risk bounds for NYTRO by providing a direct comparison with NKRLS and KRLS.

**Theorem 5** (NYTRO and NKRLS). *Let  $m \leq n$  and  $M$  be a subset of the training set. Let  $\hat{f}_{m,\gamma,t}$  be the NYTRO solution as in Equation (17),  $\tilde{f}_{m,\frac{1}{\gamma t}}$  the NKRLS solution as in Equation (12). When  $t \geq 2$  and  $\gamma < \|\mathbf{K}_{nm} \mathbf{R}\|^2$  (for example  $\gamma = 1/\max_i k(x_i, x_i)$ ) the following holds*

$$\mathbb{E} \mathbf{R} (\hat{f}_{m,\gamma,t}) \leq c_t \mathbb{E} \mathbf{R} \left( \tilde{f}_{m,\frac{1}{\gamma t}} \right).$$

with  $c_t = 4 \left(1 + \frac{1}{t-1}\right)^2 \leq 16$ .

Note that the above result holds for any  $m \leq n$  and any selection strategy of the Nyström subset  $M$ . The proof of Theorem 5 is different from the one of Theorem 4 and is based only on geometric properties of the estimator and tools from spectral theory and

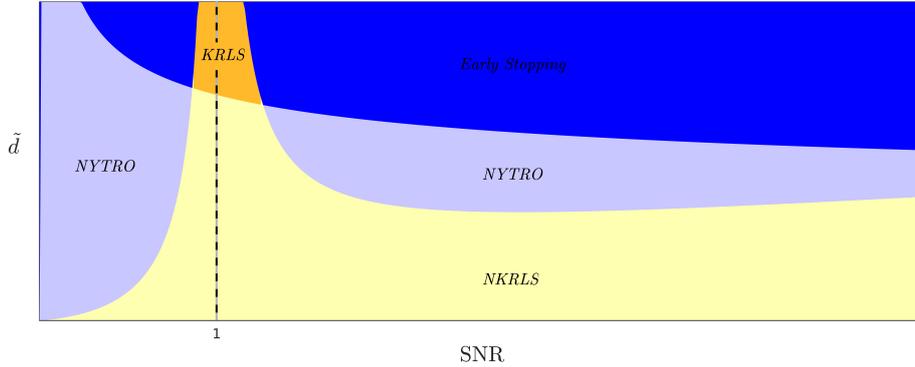


Figure 1: The graph represents the family of learning problems parametrized by the dimensionality  $\tilde{d}$  and the signal-to-noise ratio SNR (see Equations 4, 5). The four regions represent the regimes where different some algorithm are faster than the others. Purple: NYTRO is faster, Blue: Early Stopping is faster, Orange: KRLS is faster, Yellow: NKRLS is faster – see Section 3.3.

inverse problems (see Engl et al., 1996). In the next corollary we compare NYTRO and KRLS, by combining Theorems 4 and 5, hence considering  $M$  to be chosen uniformly at random from the training set.

**Corollary 1.** Let  $t \geq 2$ ,  $\gamma = 1/\|K\|$ ,  $\delta \in (0, 1)$  and  $m$  be chosen as

$$m \geq \left( 32 \frac{\tilde{d}(1/(\gamma t))}{\delta} + 2 \right) \log \frac{n\|K\|\gamma t}{\delta}.$$

Let  $\tilde{f}_{\frac{1}{\gamma t}}$  be the KRLS solution as in Equation 8 and  $\hat{f}_{m,\gamma,t}$  be the NYTRO solution. When the subset  $M$  is chosen uniformly at random from the training set, the following holds

$$\mathbb{E}_M \mathbb{E} \mathbb{R}(\hat{f}_{m,\gamma,t}) \leq c_{t,\delta} \mathbb{E} \mathbb{R}\left(\tilde{f}_{\frac{1}{\gamma t}}\right)$$

where  $c_{t,\delta} = 4 \left(1 + \frac{1}{t-1}\right)^2 (1 + 4\delta) \leq 80$ .

The above result shows that NYTRO can achieve essentially the same results as KRLS. In the next section we compare NYTRO to the other regularization algorithms introduced so far, by discussing how their computational complexity adapts to the statical accuracy in the data. In particular, by parametrizing the learning problems with respect to their dimension and their signal-to-noise ratio, we characterize the regions of the problem space where one algorithm is more efficient than the others.

### 3.3 Discussion

In Section 2 we have compared the expected excess risk of different regularization algorithms. More precisely, we have seen that there exists a suitable choice of  $\lambda$  that is  $\lambda^* = \frac{1}{\text{SNR}}$ , where SNR is the signal-to-noise ratio associated to the learning problem, such

that the expected risk of KRLS is smaller than the one of KOLS, and indeed potentially much smaller. For this reason, in the other result, statistical accuracy of the other methods was directly compared to that of KRLS with  $\lambda = \lambda^*$ .

We exploit these results to analyze the complexity of the algorithms with respect to the statistical accuracy allowed by the data. If we choose  $m \approx \tilde{d}(\lambda^*) \log(n/\lambda^*)$  and  $t = \frac{1}{\gamma\lambda^*}$ , then combining Theorem 2 with Corollary 1 and with Theorem 4, respectively, we see that the expected excess risk of both NYTRO and NKRLS is in the same order of the one of KRLS. Both algorithms have a memory requirement of  $O(nm)$  (compared to  $O(n^2)$  for KRLS), but they differ in their time requirement. For NYTRO we have  $O(n \frac{\tilde{d}(\lambda^*)}{\lambda^*} \log \frac{n}{\lambda^*})$ , while for NKRLS it is  $O(n \tilde{d}(\lambda^*)^2 (\log \frac{n}{\lambda^*})^2)$ . Now note that  $\tilde{d}(\lambda^*)$  by definition is bounded by

$$d_{\text{eff}}(\lambda) \leq \tilde{d}(\lambda) \leq \frac{1}{\lambda}, \quad \forall \lambda > 0,$$

thus, by comparing the two computational times, we can identify two regimes

$$\begin{cases} d_{\text{eff}}(\lambda^*) \leq \tilde{d}(\lambda^*) \leq \frac{1}{\lambda^* \log \frac{n}{\lambda^*}} & \implies \text{NKRLS faster} \\ \frac{1}{\lambda^* \log \frac{n}{\lambda^*}} \leq \tilde{d}(\lambda^*) \leq \frac{1}{\lambda^*} & \implies \text{NYTRO faster} \end{cases}$$

To illustrate the regimes in which different algorithms can be preferable from a computational point of view while achieving the same error as KRLS with  $\lambda^*$  (see Figure 1), it is useful to parametrize the family of learning problems with respect to the signal-to-noise ratio defined in Equation (5) and to the dimensionality of the problem  $\tilde{d} := \tilde{d}(\lambda^*)$  defined in Equation (4). We choose  $\tilde{d}$  as a measure of dimensionality with respect to  $d_{\text{eff}}$ , because  $\tilde{d}$  directly affects the computational properties of the analyzed algorithms. In Figure 1, the parameter space describing the learning problems is partitioned in regions given by the curve

$$c_1(\text{SNR}) = \frac{n}{|\log(n\text{SNR})|},$$

that separates the subsampling methods from the standard methods and

$$c_2(\text{SNR}) = \frac{\text{SNR}}{|\log(\text{SNR})|},$$

that separates the iterative from Tikhonov methods.

As illustrated in Figure 1, NYTRO is preferable when  $\text{SNR} \leq 1$ , that is when the problem is quite noisy. When  $\text{SNR} > 1$ , then NYTRO is faster when the dimension of the problem is sufficiently large. Note that, in particular, the area of the NYTRO region on  $\text{SNR} > 1$  increases with  $n$ , and the curve  $c_1$  is quite flat when  $n$  is very large. On the opposite extremes we have early stopping and NKRLS. Indeed, one is effective when the dimensionality is very large, while the second when it is very small. There is a peak around  $\text{SNR} \approx 1$  for which it seems that the only useful algorithm is NKRLS when the dimensionality is sufficiently large. The only region where KRLS is more effective is when  $\text{SNR} \approx 1$  and the dimensionality is close to  $n$ .

In the next section, the theoretical results are validated by an experimental analysis on benchmark datasets. We add one remark first.

Table 1: Specifications of the Datasets Used in Time-accuracy Comparison Experiments.  $\sigma$  is the Bandwidth of the Gaussian Kernel.

<i>Dataset</i>	<i>n</i>	<i>n</i> <sub>test</sub>	<i>d</i>	$\sigma$
<i>InsuranceCompany</i>	5822	4000	85	3
<i>Adult</i>	32562	16282	123	6.6
<i>Ijcn</i>	49990	91701	22	1
<i>YearPrediction</i>	463715	51630	90	1
<i>CoverttypeBinary</i>	522910	58102	54	1

Table 2: Time-accuracy Comparison on Benchmark Datasets.

<i>Dataset</i>		<i>KOLS</i>	<i>KRLS</i>	<i>Early Stopping</i>	<i>NKRLS</i>	<i>NYTRO</i>
<i>InsuranceCompany</i> n = 5822 m = 2000	Time (s)	<b>1.04</b>	97.48 $\pm$ 0.77	2.92 $\pm$ 0.04	20.32 $\pm$ 0.50	5.49 $\pm$ 0.12
	RMSE	5.179	<b>0.4651 <math>\pm</math> 0.0001</b>	<b>0.4650 <math>\pm</math> 0.0002</b>	<b>0.4651 <math>\pm</math> 0.0003</b>	<b>0.4651 <math>\pm</math> 0.0003</b>
	Par.	NA	3.27e-04	494 $\pm$ 1.7	5.14e-04 $\pm$ 1.42e-04	491 $\pm$ 3
<i>Adult</i> n = 32562 m = 1000	Time (s)	112	4360 $\pm$ 9.29	5.52 $\pm$ 0.23	5.95 $\pm$ 0.31	<b>0.85 <math>\pm</math> 0.05</b>
	RMSE	1765	<b>0.645 <math>\pm</math> 0.001</b>	0.685 $\pm$ 0.002	0.6462 $\pm$ 0.003	0.6873 $\pm$ 0.003
	Par.	NA	4.04e-05 $\pm$ 1.04e-05	39.2 $\pm$ 1.1	4.04e-05 $\pm$ 1.83e-05	44.9 $\pm$ 0.3
<i>Ijcn</i> n = 49990 m = 5000	Time (s)	271	825.01 $\pm$ 6.81	154.82 $\pm$ 1.24	160.28 $\pm$ 1.54	<b>80.9 <math>\pm</math> 0.4</b>
	RMSE	730.62	0.615 $\pm$ 0.002	<b>0.457 <math>\pm</math> 0.001</b>	0.469 $\pm$ 0.003	<b>0.457 <math>\pm</math> 0.001</b>
	Par.	NA	1.07e-08 $\pm$ 1.47e-08	489 $\pm$ 7.2	1.07e-07 $\pm$ 1.15e-07	328.7 $\pm$ 2.6
<i>YearPrediction</i> n = 463715 m = 10000	Time (s)				1188.47 $\pm$ 36.7	<b>887 <math>\pm</math> 6</b>
	RMSE	NA	NA	NA	<b>0.1015 <math>\pm</math> 0.0002</b>	0.1149 $\pm$ 0.0002
	Par.				3.05e-07 $\pm$ 1.05e-07	481 $\pm$ 6.1
<i>CoverttypeBinary</i> n = 522910 m = 10000	Time (s)				1235.21 $\pm$ 42.1	<b>92.69 <math>\pm</math> 2.35</b>
	RMSE	NA	NA	NA	1.204 $\pm$ 0.008	<b>0.918 <math>\pm</math> 0.006</b>
	Par.				9.33e-09 $\pm$ 1.12e-09	39.2 $\pm$ 2.3

**Remark 1** (Empirical parameter choices and regularization path). *Note that an important aspect that is not covered by Figure 1 is that iterative algorithms have the further desirable property of computing the regularization path. In fact, for KRLS and NKRLS computations are slowed by a factor of  $|\Lambda|$ , where  $\Lambda$  is the discrete set of cross-validated  $\lambda$ s. This last aspect is very relevant in practice, because the optimal regularization parameter values are not known and need to be found via model selection/aggregation.*

## 4 Experiments

In this section we present an empirical evaluation of the NYTRO algorithm, showing regimes in which it provides a significant model selection speedup with respect to NKRLS and the other exact kernelized learning algorithms mentioned above (KOLS, KRLS and Early Stopping). We consider the Gaussian kernel and the subsampling of the training set points for kernel matrix approximation is performed uniformly at random. All experiments have been carried out on a server with  $12 \times 2.10\text{GHz}$  Intel<sup>®</sup> Xeon<sup>®</sup> E5-2620 v2 CPUs and 132 GB of RAM.

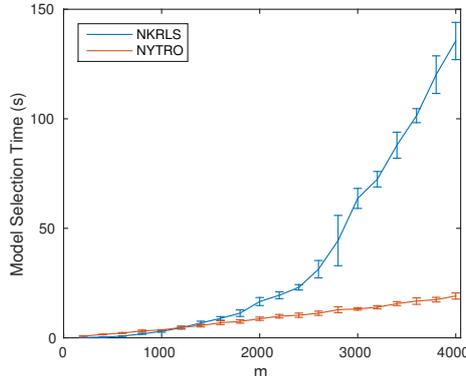


Figure 2: Training Time of NKRLS and NYTRO on the `cpuSmall` Dataset as the Subsampling Level  $m$  Varies Linearly Between 100 and 4000. Experiment With 5 Repetitions. Mean and Standard Deviation Reported.

We compare the algorithms on the benchmark datasets reported in Table 1<sup>2</sup>. In the table we also report the bandwidth parameter  $\sigma$  adopted for the Gaussian kernel computation. Following (Si et al., 2014), we measure performance by the root mean squared error (RMSE). Note that for the `YearPredictionMSD` dataset outputs are scaled in  $[0, 1]$ .

For all the algorithms, model selection is performed via hold-out cross validation, where the validation set is composed of 20% of the training points chosen uniformly at random at each trial. We select the regularization parameter  $\lambda$  for NKRLS between 100 guesses logarithmically spaced in  $[10^{-15}, 1]$ , by computing the validation error for each model and choosing the  $\lambda^*$  associated with the lowest error. NYTRO’s regularization parameter is the number of iterations  $t$ . We select the optimal  $t^*$  by considering the evolution of the validation error. As an early stopping rule, we choose an iteration such that the validation error ceases to be decreasing up to a given threshold chosen to be the 5% of the relative RMSE. After model selection, we evaluate the performance on the test set. We report the results in Table 2 and discuss them further below.

**Time Complexity Comparison** We start by showing how the time complexity changes with the subsampling level  $m$ , making NYTRO more convenient than NKRLS if  $m$  is large enough. For example, consider Figure 2. We performed training on the `cpuSmall`<sup>3</sup> dataset ( $n = 6554$ ,  $d = 12$ ), with  $m$  spanning between 100 and 4000 at 100-points linear intervals. The experiment is repeated 5 times, and we report the mean and standard deviation of the NYTRO and NKRLS model selection times. We consider 100 guesses for  $\lambda$ , while the NYTRO iterations are fixed to a maximum of 500. As revealed by the plot, the time complexity grows linearly with  $m$  for NYTRO and quadratically for NKRLS. This is consistent with the time complexities outlined in Sections 2 and 3 ( $O(nm^2 + m^3)$  for NKRLS and  $O(nmt + m^3)$  for NYTRO).

<sup>2</sup>All the datasets are available at <http://archive.ics.uci.edu/ml> or <https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>

<sup>3</sup><http://www.cs.toronto.edu/~delve/data/datasets.html>

**Time-accuracy Benchmarking** We also compared the training time and accuracy performances for KRLS, KOLS, Early Stopping (ES), NKRLS and NYTRO, reporting the selected hyperparameter ( $\lambda^*$  for KRLS and NKRLS,  $t^*$  for ES and NYTRO), the model selection time and the test error in Table 2. All the experiments are repeated 5 times. The standard deviation of the results is negligible. Notably, NYTRO achieves comparable or superior predictive performances with respect to its counterparts in a fraction of the model selection time. In particular, the absolute time gains are most evident on large scale datasets such as *Covertime* and *YearPredictionMSD*, for which a reduction of an order of magnitude in cross-validation time corresponds to saving tens of minutes. Note that exact methods such as KOLS, KRLS and ES cannot be applied to such large scale datasets due to their prohibitive memory requirements. Remarkably, NYTRO’s predictive performance is not significantly penalized in these regimes and can even be improved with respect to other methods, as in the *Covertime* case, where it requires 90% less time for model selection.

## 5 Acknowledgements

The work described in this paper is supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216, and FIRB project RBFR12M3AC, funded by the Italian Ministry of Education, University and Research.

## References

- Ahmed Alaoui and Michael W Mahoney. Fast Randomized Kernel Methods With Statistical Guarantees. *arXiv*, 2014.
- Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *COLT*, volume 30 of *JMLR Proceedings*, pages 185–209. JMLR.org, 2013.
- F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.
- A. Caponnetto and Yuan Yao. Adaptive rates for regularization operators in learning theory. *Analysis and Applications*, 08, 2010.
- Bo Dai, Bo Xie 0002, Niao He, Yingyu Liang, Anant Raj, Maria-Florina Balcan, and Le Song. Scalable Kernel Methods via Doubly Stochastic Gradients. In *NIPS*, pages 3041–3049, 2014.
- Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2001.
- A. E. Hoerl and R. W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12:55–67, 1970.

- Po-Sen Huang, Haim Avron, Tara N. Sainath, Vikas Sindhwani, and Bhuvana Ramabhadran. Kernel methods match Deep Neural Networks on TIMIT. In *ICASSP*, 2014.
- Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Ensemble Nystrom Method. In *NIPS*, pages 1060–1068. Curran Associates, Inc., 2009.
- Mu Li, James T. Kwok, and Bao-Liang Lu. Making Large-Scale Nyström Approximation Possible. In *ICML*, pages 631–638. Omnipress, 2010.
- Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *The Journal of Machine Learning Research*, 7:2651–2667, 2006.
- Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early Stopping and Non-parametric Regression: An Optimal Data-dependent Stopping Rule. *J. Mach. Learn. Res.*, 15(1): 335–366, January 2014. ISSN 1532-4435.
- Ryan Rifkin, Gene Yeo, and Tomaso Poggio. Regularized least-squares classification. *Nato Science Series Sub Series III Computer and Systems Sciences*, 190:131–154, 2003.
- Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. MIT Press, 2002.
- Si Si, Cho-Jui Hsieh, and Inderjit S. Dhillon. Memory Efficient Kernel Approximation. In *ICML*, volume 32 of *JMLR Proceedings*, pages 701–709. JMLR.org, 2014.
- Alex J. Smola and Bernhard Schölkopf. Sparse Greedy Matrix Approximation for Machine Learning. In *ICML*, pages 911–918. Morgan Kaufmann, 2000. ISBN 1-55860-707-2.
- A. N. Tikhonov. On the solution of ill-posed problems and the method of regularization. *Dokl. Akad. Nauk SSSR*, 151:501–504, 1963.
- Christopher Williams and Matthias Seeger. Using the Nyström Method to Speed Up Kernel Machines. In *NIPS*, pages 682–688. MIT Press, 2000.
- Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On Early Stopping in Gradient Descent Learning. *Constructive Approximation*, 26(2):289–315, 2007. ISSN 0176-4276. doi: 10.1007/s00365-006-0663-2.
- Kai Zhang, Ivor W. Tsang, and James T. Kwok. Improved Nyström Low-rank Approximation and Error Analysis. *ICML*, pages 1232–1239. ACM, 2008. doi: 10.1145/1390156.1390311.
- Tong Zhang and Bin Yu. Boosting with early stopping: convergence and consistency. *Annals of Statistics*, pages 1538–1579, 2005.

## A Proofs

*Proof of Theorem 2.* By applying Prop. 2 to the estimator of Equation 3 we have  $Q_{\text{ols}} = K^\dagger K = P$ . Now note that  $P^2 = P$  by definition,  $\text{Tr}(P) = d^*$  and that  $P(I - P) = 0$ , therefore

$$\mathbb{E}R(f_{\text{ols}}) = \frac{\sigma^2}{n} \text{Tr}(P^2) + \frac{1}{n} \|P(I - P)\mu\| = \frac{\sigma^2 d^*}{n}.$$

Now let  $K = U\Sigma U^\top$  be the eigen-decomposition of  $K$ , with  $U$  an orthonormal matrix and  $\Sigma$  a diagonal matrix with  $\sigma_1 \geq \dots \geq \sigma_n \geq 0$ . Let  $\bar{Q}_\lambda = (K + \lambda n I)^{-1} K$ ,  $\beta = U^\top P \mu$  with  $P = K^\dagger K$  the projection operator on the range of  $K$ . By applying Prop. 2 to the estimator of Equation (3) and considering that  $P(I - \bar{Q}_\lambda) = (I - \bar{Q}_\lambda)P$  and  $I - \bar{Q}_\lambda = \lambda n (K + \lambda n I)^{-1}$ , we have

$$\begin{aligned} \mathbb{E}R(\bar{f}_\lambda) &= \frac{\sigma^2}{n} \text{Tr}(\bar{Q}_\lambda^2) + \frac{1}{n} \|P(I - \bar{Q}_\lambda)\mu\|^2 \\ &= \frac{\sigma^2}{n} \text{Tr}(\bar{Q}_\lambda^2) + \frac{1}{n} \|(I - \bar{Q}_\lambda)P\mu\|^2 \\ &= \frac{1}{n} \sum_{i=1}^n \frac{\sigma^2 \sigma_i^2 + \lambda^2 n^2 \beta_i^2}{(\sigma_i + \lambda n)^2} \\ &= \frac{1}{n} \sum_{i=1}^{d^*} \frac{\sigma^2 \sigma_i^2 + \lambda^2 n^2 \beta_i^2}{(\sigma_i + \lambda n)^2} \end{aligned}$$

where the last step is due to the fact that  $\sigma_i = \beta_i = 0$  for  $i > d^*$ . Now we study  $\mathbb{E}R(\bar{f}_{\lambda^*})$  when  $\lambda^* = \sigma^2 / \|K^{\dagger 1/2} P \mu\|^2 = \sigma^2 / T$  with  $T = \sum_{i=1}^{d^*} \tau_i^2$ . Let  $\bar{\sigma}_i = \sigma_i / n$  and  $\tau_i = \sigma_i^{-1/2} \beta_i$  for  $1 \leq i \leq d^*$ . We have

$$\begin{aligned} \mathbb{E}R(\bar{f}_{\lambda^*}) &= \frac{\sigma^2}{n} \sum_{i=1}^{d^*} \frac{\bar{\sigma}_i}{\bar{\sigma}_i + \lambda^*} \frac{\bar{\sigma}_i + \sigma^2 \tau_i^2 / T^2}{\bar{\sigma}_i + \sigma^2 / T} \\ &= \frac{\sigma^2}{n} \sum_{i=1}^{d^*} \frac{\bar{\sigma}_i}{\bar{\sigma}_i + \lambda^*} \frac{(\bar{\sigma}_i + \sigma^2 / T) - (\sigma^2 / T)(1 - \tau_i^2 / T)}{\bar{\sigma}_i + \sigma^2 / T} \\ &= \frac{\sigma^2}{n} \sum_{i=1}^{d^*} \frac{\bar{\sigma}_i}{\bar{\sigma}_i + \lambda^*} \left( 1 - \frac{1 - \tau_i^2 / T}{1 + T \bar{\sigma}_i / \sigma^2} \right) \\ &\leq \frac{\sigma^2}{n} \sum_{i=1}^{d^*} \frac{\bar{\sigma}_i}{\bar{\sigma}_i + \lambda^*} = \frac{\sigma^2}{n} \sum_{i=1}^{d^*} \frac{\sigma_i}{\sigma_i + \lambda^* n} \\ &= \frac{\sigma^2}{n} \text{Tr}(\Sigma(\Sigma + \lambda^* n I)^{-1}) = \frac{\sigma^2}{n} d_{\text{eff}}(\lambda^*). \end{aligned}$$

□

*Proof of Theorem 3.* It is an application of Theorem 5 when we select the whole training set ( $m = n$ ) for the Nyström approximation. In that case, the expected excess risks of Nyström KRLS and NYTRO are just equal to the ones of KRLS and Early Stopping, indeed

when  $m = n$  we have that  $K_{mm} = K_{nm} = K$ . If we call  $\bar{Q}_\lambda$  and  $\tilde{Q}_{n,\lambda}$  the  $Q$ -matrices for the two algorithms (see Prop. 2) and  $R$  such that  $RR^\top = K_{mm}^\dagger$ , for any  $\lambda > 0$  we have

$$\begin{aligned}\bar{Q}_\lambda &= (K + \lambda nI)^{-1}K = (KK^\dagger K + \lambda nI)^{-1}KK^\dagger K \\ &= (KRR^\top K + \lambda nI)^{-1}KRR^\top K \\ &= KR(R^\top K^2 R + \lambda nI)^{-1}R^\top K = \tilde{Q}_{n,\lambda}.\end{aligned}$$

□

*Proof of Theorem 5.* In the following we assume without loss of generality that the selected points  $\tilde{x}_1, \dots, \tilde{x}_m$  are the first  $m$  points in the dataset. In Prop. 2 we have seen that the behavior of an algorithm in a fixed design setting is completely described by a matrix  $Q = KC$  when the coefficients of the estimator are of the form  $Cy$ . We now find the associated  $Q$  for NYTRO, that is  $\hat{Q}_{m,\gamma,t}$ . By solving the recursion of Equation (17), we have for any  $i \in \{1, \dots, n\}$

$$\begin{aligned}\hat{f}_{m,\gamma,t}(x_i) &= k_i^\top Cy, \text{ with } C = \begin{pmatrix} C_{m,\gamma,t} \\ 0_{(n-m) \times n} \end{pmatrix}, \\ C_{m,\gamma,t} &= \gamma \sum_{p=0}^{t-1} R(I - \gamma A^\top A)^p A^\top\end{aligned}$$

with  $A = K_{nm}R$  and  $k_i = (k(x_i, x_1), \dots, k(x_i, x_n))$ . Therefore, we have

$$\begin{aligned}\hat{Q}_{m,\gamma,t} &= KC = \gamma \sum_{p=0}^{t-1} K_{nm}R(I - \gamma A^\top A)^p A^\top \\ &= \gamma \sum_{p=0}^{t-1} A(I - \gamma A^\top A)^p A^\top.\end{aligned}$$

**Rewriting of  $\hat{Q}_{m,\gamma,t}$ .** Now we rewrite  $\hat{Q}_{m,\gamma,t}$  in a suitable form to bound the bias and variance errors. First of all, we apply Prop. 3 to  $\hat{Q}_{m,\gamma,t}$ . Let  $f(\sigma) = \gamma \sum_{i=0}^{t-1} (1 - \gamma/n\sigma)^p$  with  $\sigma \in [0, n/\gamma]$ , we have that

$$\hat{Q}_{m,\gamma,t} = Af(A^\top A)A^\top = f(AA^\top)AA^\top = g(AA^\top),$$

where  $g(\sigma) = f(\sigma)\sigma$ . Now note that

$$g(\sigma) = \gamma\sigma \sum_{i=0}^{t-1} (1 - \gamma/n\sigma)^p = 1 - (1 - \gamma/n\sigma)^t,$$

therefore we have

$$\hat{Q}_{m,\gamma,t} = g(AA^\top) = I - (I - \gamma/nAA^\top)^t.$$

**Bound of the Bias** Now we are going to bound the bias for NYTRO. Let  $\lambda = 1/(\gamma t)$  and  $Z = AA^\top$ , then

$$\begin{aligned} B(\hat{Q}_{m,\gamma,t}) &= \frac{1}{n} \|\mathbb{P}(I - \hat{Q}_{m,\gamma,t})\mu\|^2 \\ &= \frac{1}{n} \|\mathbb{P}(I - \frac{\gamma}{n}Z)^t \mu\|^2 = \frac{1}{n} \|(I - \frac{\gamma}{n}Z)^t \mathbb{P}\mu\|^2 \\ &= \frac{1}{n} \|(I - \frac{\gamma}{n}Z)^t (Z + \lambda n I)(Z + \lambda n I)^{-1} \mathbb{P}\mu\|^2 \\ &\leq \frac{1}{n} q(A, \lambda n) \|(Z + \lambda n I)^{-1} \mathbb{P}\mu\|^2 \end{aligned}$$

and  $q(A, \lambda n) = \|(I - \gamma/n AA^\top)^t (AA^\top + \lambda n I)\|^2$ . Note that the third step is due to the fact that  $\text{range}(Z) \subseteq \text{range}(K) = \text{range}(P)$  and  $Z$  is symmetric. Therefore,  $\text{Ph}(Z) = h(Z)P$  as a consequence of Prop. 3 for any spectral function  $h$ . Let  $\sigma_1, \dots, \sigma_n$  be the singular values of  $Z$ , we have

$$\begin{aligned} q\left(A, \frac{n}{\gamma t}\right) &= \sup_{i \in \{1, \dots, n\}} (1 - \gamma/n \sigma_i)^{2t} \left(\sigma_i + \frac{n}{\gamma t}\right)^2 \\ &\leq \sup_{0 \leq \sigma \leq n/\gamma} (1 - \gamma/n \sigma)^{2t} \left(\sigma + \frac{n}{\gamma t}\right)^2 \leq \frac{n^2}{\gamma^2 t^2} \end{aligned}$$

Therefore we have

$$B(\hat{Q}_{m,\gamma,t}) \leq \lambda^2 n \|(Z + \lambda n I)^{-1} \mathbb{P}\mu\|^2.$$

**Bound of the Variance** Let  $t \geq 2$ ,  $\lambda = \frac{1}{\gamma t}$ ,  $r(\sigma) = (1 - \gamma/n \sigma)^t$  and

$$v(\sigma) = \sigma/(t-1) + \sigma(1+r(\sigma)) - \lambda n(1-r(\sigma)).$$

We have  $v(\sigma) \geq 0$  for  $0 \leq \sigma \leq n/\gamma$ . Indeed, for  $\lambda n < \sigma \leq n/\gamma$  we have  $v(\sigma) \geq 0$ , since  $0 \leq r(\sigma) \leq 1$ , while for  $0 \leq \sigma \leq \lambda n$  we have

$$\begin{aligned} \lambda n(1-r(\sigma)) &= \lambda n \left(1 - e^{-t \log \frac{1-\gamma\sigma}{1-\gamma\lambda n}}\right) \leq \frac{n}{\gamma t} t \log \frac{1}{1-\frac{\gamma\sigma}{n}} \\ &\leq \frac{n}{\gamma} \frac{\gamma/n \sigma}{1-\gamma/n \sigma} \leq \frac{\sigma}{1-\frac{1}{t}} = \frac{\sigma}{t-1} + \sigma \\ &\leq \frac{\sigma}{t-1} + \sigma(1+r(\sigma)), \end{aligned}$$

therefore  $v(\sigma) \geq 0$ . Now let  $0 \leq \sigma \leq n/\gamma$ . Since  $v(\sigma) \geq 0$ , the function  $w(\sigma) = v(\sigma)/(\sigma + \lambda n)$  is  $w(\sigma) \geq 0$ . Now we rewrite  $w$  a bit. First of all, note that

$$w(\sigma) = (2t-1)/(t-1)w_1(\sigma) - g(\sigma),$$

with  $w_1(\sigma) = \sigma/(\sigma + \lambda n)$ . The fact that  $w(\sigma) \geq 0$  and that  $g(\sigma) \geq 0$  implies that

$$\left(\frac{2t-1}{t-1}\right)^2 w_1(\sigma)^2 \geq g(\sigma)^2. \quad \forall 0 \leq \sigma \leq \frac{n}{\gamma}, t \geq 2$$

Let us now focus on  $\text{Tr}(\hat{Q}_{\gamma t}^2)$ . Let  $Z = U\Sigma U^\top$  be its eigenvalue decomposition with  $U$  an orthonormal matrix and  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$  with  $\sigma_1 \geq \dots \geq \sigma_n \geq 0$ ,

$$\begin{aligned} \text{Tr}(\hat{Q}_{m,\gamma,t}^2) &= \text{Tr}(g^2(Z)) = \text{Tr}(Ug^2(\Sigma)U^\top) = \text{Tr}(g^2(\Sigma)) \\ &= \sum_{i=1}^n g(\sigma_i)^2 \leq c_t \sum_{i=1}^n w_1(\sigma_i)^2 = c_t \text{Tr}(w_1(\Sigma)^2) \\ &= c_t \text{Tr}(Uw_1(\Sigma)^2U^\top) = c_t \text{Tr}(w_1(Z)^2) \\ &= c_t \text{Tr}(Z^2(Z + \lambda n I)^{-2}) \end{aligned}$$

where we applied many times Prop. 3 and the fact that the trace is invariant to unitary transforms. Thus

$$V(\hat{Q}_{m,\gamma,t}, n) \leq \frac{\sigma^2}{n} \left( \frac{2t-1}{t-1} \right)^2 \text{Tr} \left( Z(Z + n/(\gamma t)I)^{-1} \right)^2.$$

**The Expected Excess Risk for Nyström KRLS** The Nyström KRLS estimator with linear kernel is a function of the form

$$\begin{aligned} \tilde{f}(x_i) &= k_i^\top C y, \quad \text{with } C = \begin{pmatrix} \tilde{C}_{m,\lambda} \\ 0_{(n-m) \times n} \end{pmatrix}, \\ \tilde{C}_{m,\lambda} &= R(A^\top A + \lambda n I)^\dagger A^\top, \end{aligned}$$

with  $k_i = (k(x_i, x_1), \dots, k(x_i, x_n))$  for any  $i \in \{1, \dots, n\}$ . Now, by applying Prop. 3 we have

$$\begin{aligned} \tilde{Q}_{m,\lambda} &= KC = K_{nm} \tilde{C}_{m,\lambda} \\ &= A(A^\top A + \lambda n I)^{-1} A = AA^\top (AA^\top + \lambda I)^{-1} \\ &= Z(Z + \lambda n I)^{-1} \end{aligned}$$

Thus, we have

$$\begin{aligned} V(\tilde{Q}_{m,\lambda}) &= \frac{\sigma^2}{n} \text{Tr}(\tilde{Q}_{m,\lambda})^2 = \frac{\sigma^2}{n} \text{Tr} \left( Z(Z + \lambda n I)^{-1} \right)^2 \\ B(\tilde{Q}_{m,\lambda}) &= \frac{1}{n} \|\mathbb{P}(I - Z(Z + \lambda n I)^{-1})\mu\|^2 \\ &= \lambda^2 n \|\mathbb{P}(Z + \lambda n I)^{-1}\mu\|^2 \\ &= \lambda^2 n \|(Z + \lambda n I)^{-1}\mathbb{P}\mu\|^2. \end{aligned}$$

where the last step is due to the same reasoning as in the bound for the bias of NYTRO. Finally, by applying twice Prop. 2 and calling  $c_t = \left( \frac{2t-1}{t-1} \right)^2$ , we have that

$$\begin{aligned} R(\hat{f}_{m,\gamma,t}) &= V(\hat{Q}_{m,\gamma,t}, n) + B(\hat{Q}_{m,\gamma,t}) \\ &\leq c_t V(\tilde{Q}_{m,\frac{1}{\gamma t}}, n) + B(\tilde{Q}_{m,\frac{1}{\gamma t}}) \\ &\leq c_t \left( V(\tilde{Q}_{m,\frac{1}{\gamma t}}, n) + B(\tilde{Q}_{m,\frac{1}{\gamma t}}) \right) \\ &= c_t R(\tilde{f}_{m,\frac{1}{\gamma t}}) \end{aligned}$$

for  $\|Z\| \leq n/\gamma$  and  $t \geq 2$ . Now the choice  $\gamma = 1/(\max_{1 \leq i \leq n} k(x_i, x_i))$  is valid, indeed

$$\begin{aligned} \gamma \|Z\|^2 &= \gamma \|K_{nm} R R^\top K_{nm}^\top\| = \gamma \|K_{nm} K_{mm}^\dagger K_{nm}^\top\| \\ &\leq \gamma \|K\| \leq \gamma n \max_{1 \leq i \leq n} (K)_{ii} = \gamma n \max_{1 \leq i \leq n} k(x_i, x_i), \end{aligned}$$

where  $\|K_{nm} K_{mm}^\dagger K_{nm}^\top\| \leq \|K\|$  can be found in [Bach \(2013\)](#); [Alaoui and Mahoney \(2014\)](#).  $\square$

*Proof of Corollary 1.* Theorem 5 combined with Theorem 1 of [Bach \(2013\)](#).  $\square$

## B Some Useful Results

**Proposition 1.** *With the notation of Section 2.3, let  $R \in \mathbb{R}^{m \times k}$  such that  $K_{mm}^\dagger = R R^\top$  and  $A = K_{nm} R$ . Then, for any  $\lambda, m > 0$ ,  $\tilde{\alpha}_{m,\lambda}$  is characterized by Equation 16.*

*Proof.* By Equation 7.7 of [Rifkin et al. \(2003\)](#) we have that

$$\begin{aligned} \tilde{\alpha}_{m,\lambda} &= K_{mm}^\dagger K_{nm}^\top (K_{nm} K_{mm}^\dagger K_{nm}^\top + \lambda n I)^{-1} \mathbf{y} \\ &= R R^\top K_{nm}^\top (K_{nm} R R^\top K_{nm}^\top + \lambda n I)^{-1} \mathbf{y} \\ &= R A^\top (A A^\top + \lambda n I)^{-1} \mathbf{y} \\ &= R (A^\top A + \lambda n I)^{-1} A^\top \mathbf{y}, \end{aligned}$$

where the last step is due to Prop. 3.  $\square$

**Proposition 2.** *Let  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  be a kernel function on  $\mathcal{X}$ ,  $x_1, \dots, x_n$  be the given points and  $\mathbf{y} = (y_1, \dots, y_n)$  be the labels of the dataset. For any function of the form  $f(x) = \sum_{i=1}^n w_i k(x, x_i)$  with  $w = C \mathbf{y}$  for any  $x \in \mathcal{X}$ , with  $C \in \mathbb{R}^{n \times n}$  independent from  $\mathbf{y}$ , the following holds*

$$\mathbb{E}_{\mathbf{y}} \mathcal{R}(f) = \underbrace{\frac{\sigma^2}{n} \text{Tr}(Q^2)}_{\text{Variance } V(Q)} + \underbrace{\frac{1}{n} \|P(I - Q)\mu\|^2}_{\text{Bias } B(Q)},$$

with  $Q = KC \in \mathbb{R}^{n \times n}$ ,  $K$  the kernel matrix,  $\mu = \mathbb{E} \mathbf{y} \in \mathbb{R}^n$  and  $P = K^\dagger K$  the projection operator on the range of  $K$ .

*Proof.* A function  $f \in \mathcal{H}$  is of the form  $f(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$  for any  $x \in \mathcal{X}$ . If we compute it on a point of the dataset  $x_i$ , with  $i \in \{1, \dots, n\}$ , we have  $f(x_i) = \sum_{j=1}^n \alpha_j k(x_i, x_j) = k_i^\top w$  with  $w = C \mathbf{y}$  and  $k_i = (k(x_i, x_1), \dots, k(x_i, x_n))$ . Note that  $K = (k_1, \dots, k_n)$ .

**Rewriting of  $\mathcal{E}, \mathcal{R}$  for Fixed Design** We have

$$\begin{aligned} \mathcal{E}(w) &= \frac{1}{n} \sum_{i=1}^n \mathbb{E} (k_i^\top w - y_i)^2 = \frac{1}{n} \sum_{i=1}^n (\mathbb{E} (k_i^\top w - \mu_i)^2 \\ &\quad - 2 (k_i^\top w - \mu_i) (y_i - \mu_i) + (y_i - \mu_i)^2) \\ &= \frac{1}{n} \sum_{i=1}^n (k_i^\top w - \mu_i)^2 + \frac{\sigma^2}{n} = \frac{\sigma^2}{n} + \frac{1}{n} \|Kw - \mu\|^2, \end{aligned}$$

Now note that  $\mathbf{PK} = \mathbf{K}$  and  $(\mathbf{I} - \mathbf{P})\mathbf{K} = \mathbf{0}$ , that  $\|\mathbf{q}\|^2 = \|\mathbf{Pq}\|^2 + \|(\mathbf{I} - \mathbf{P})\mathbf{q}\|^2$  for any  $\mathbf{q} \in \mathcal{H}$  and that  $\inf_{\mathbf{v} \in \mathcal{X}} \mathcal{E}(\mathbf{v}) = \sigma^2 + \|(\mathbf{I} - \mathbf{P})\boldsymbol{\mu}\|^2$ , then the excess risk can be rewritten as

$$\begin{aligned} \mathbf{R}(\mathbf{w}) &= \frac{1}{n} \|\mathbf{Kw} - \boldsymbol{\mu}\|^2 - \frac{1}{n} \|(\mathbf{I} - \mathbf{P})\boldsymbol{\mu}\|^2 \\ &= \frac{1}{n} \|\mathbf{P}(\mathbf{Kw} - \boldsymbol{\mu})\|^2 + \frac{1}{n} \|(\mathbf{I} - \mathbf{P})(\mathbf{Kw} - \boldsymbol{\mu})\|^2 \\ &\quad - \frac{1}{n} \|(\mathbf{I} - \mathbf{P})\boldsymbol{\mu}\|^2 = \frac{1}{n} \|\mathbf{P}(\mathbf{Kw} - \boldsymbol{\mu})\|^2. \end{aligned}$$

**Expected Excess Risk** We focus on the expectation of  $\mathbf{R}$  with respect to the dataset for linear functions that depend linearly on the observed labels  $\mathbf{y}$ . Indeed we have

$$\begin{aligned} \mathbb{E}\mathbf{R}(\mathbf{w}) &= \frac{1}{n} \mathbb{E} \|\mathbf{P}(\mathbf{K}\mathbf{C}\mathbf{y} - \mathbf{P}\boldsymbol{\mu})\|^2 \\ &= \frac{1}{n} \mathbb{E} \|\mathbf{P}\mathbf{Q}(\mathbf{y} - \boldsymbol{\mu}) + \mathbf{P}(\mathbf{I} - \mathbf{Q})\boldsymbol{\mu}\|^2 \\ &= \frac{1}{n} \mathbb{E} \text{Tr}(\mathbf{Q}(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{Q}) + \frac{1}{n} \|\mathbf{P}(\mathbf{I} - \mathbf{Q})\boldsymbol{\mu}\|^2 \\ &\quad - \frac{2}{n} \mathbb{E}(\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{Q}\mathbf{P}(\mathbf{I} - \mathbf{Q})\boldsymbol{\mu} \\ &= \frac{1}{n} \text{Tr}(\mathbf{Q}\mathbb{E}(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{Q}) + \frac{1}{n} \|\mathbf{P}(\mathbf{I} - \mathbf{Q})\boldsymbol{\mu}\|^2 \\ &= \frac{\sigma^2}{n} \text{Tr}(\mathbf{Q}^2) + \frac{1}{n} \|\mathbf{P}(\mathbf{I} - \mathbf{Q})\boldsymbol{\mu}\|^2. \end{aligned}$$

Here the third step is due to  $\|\mathbf{a} - \mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\mathbf{a}^\top \mathbf{b}$  and that  $\|\mathbf{a}\|^2 = \text{Tr}(\mathbf{a}\mathbf{a}^\top)$ , for any vector  $\mathbf{a}, \mathbf{b}$ . The last term in the third step vanishes due to the fact that  $\mathbf{y} - \boldsymbol{\mu}$  is a zero mean random variable, moreover note that  $(\mathbb{E}(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^\top)_{ij} = \mathbb{E}(y_i - \mu_i)(y_j - \mu_j) = \sigma^2 \delta_{ij}$ , therefore  $\mathbb{E}(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^\top = \sigma^2 \mathbf{I}$ .  $\square$

**Proposition 3** (Spectral functions). *Let  $f, g : [0, T] \rightarrow \mathbb{R}$  be a continuous function and  $\mathbf{A} \in \mathbb{R}^{n \times n}$  symmetric with  $\|\mathbf{A}\| \leq T$ , for a  $T > 0$ ,  $n \geq 1$ . Let  $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^\top$  be its eigenvalue decomposition with  $\mathbf{U} \in \mathbb{R}^{n \times n}$  an orthonormal matrix,  $\mathbf{U}^\top \mathbf{U} = \mathbf{U}\mathbf{U}^\top = \mathbf{I}$  and  $\boldsymbol{\Sigma}$  a diagonal matrix, then*

$$\begin{aligned} f(\mathbf{A}) &= \mathbf{U}f(\boldsymbol{\Sigma})\mathbf{U}^\top, \\ f(\mathbf{A}) + g(\mathbf{A}) &= (f + g)(\mathbf{A}), \quad f(\mathbf{A})g(\mathbf{A}) = (fg)(\mathbf{A}) \end{aligned}$$

where  $f(\boldsymbol{\Sigma}) = \text{diag}(f(\sigma_1), \dots, f(\sigma_n))$ . Moreover, let  $\mathbf{B} \in \mathbb{R}^{n \times m}$  with  $n, m \geq 1$ , then

$$f(\mathbf{B}^\top \mathbf{B})\mathbf{B}^\top = \mathbf{B}^\top f(\mathbf{B}\mathbf{B}^\top).$$