NYTRO: When Subsampling Meets Early Stopping

Raffaello Camoriano[1]
iCub Facility - IIT[2]
LCSL - IIT & MIT[3]
raffaello.camoriano@iit.it

Tomás Angles[1]
École Polytechnique
LCSL - IIT & MIT[3]
tomas.angles@gmail.com

Alessandro Rudi
DIBRIS - UniGe[4]
LCSL - IIT & MIT[3]
ale_rudi@mit.edu

Lorenzo Rosasco
DIBRIS - UniGe[4]
LCSL - IIT & MIT[3]
lrosasco@mit.edu

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 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 (Zhang et al., 2008; Kumar et al., 2009; Li et al., 2010; Dai et al., 2014; Huang et al., 2014; Si et al., 2014; Rudi et al., 2015).

[1]The authors contributed equally.
[2]Istituto Italiano di Tecnologia, Genoa, Italy.
[3]Laboratory for Computational and Statistical Learning, Istituto Italiano di Tecnologia and Massachusetts Institute of Technology.
[4]Università degli Studi di Genova, Italy.

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51.

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

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

where $\mathcal{H}$ is a space of functions, called *hypothesis space*. In real applications, ρ 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*

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

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

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

induced by a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (see Schölkopf and Smola, 2002). In this case Problem (8) corresponds to the *Kernel Ordinary Least Squares* (KOLS) and has the closed form solution

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

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 $y = (y_1, \dots, y_n)$. The cost for computing the coefficients α_{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

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

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*

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

Indeed, the above quantity can be shown to be related to the eigenvalue decay of K (Bach, 2013; Alaoui and

Mahoney, 2014; Rudi et al., 2015) and can considerably smaller than d^*, as discussed in the following. Finally, consider

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

It is easy to see that the following inequalities hold,

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

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

$$f_{\text{opt}} = \sum_{i=1}^{n} \alpha_{\text{opt},i} k(x, x_i), \quad \forall x \in \mathcal{X} \tag{5}$$

$$\alpha_{\text{opt}} = K^{\dagger}\mu, \quad \text{with } \mu = \mathbb{E}y. \tag{6}$$

Then, the signal to noise ratio is defined as

$$\text{SNR} = \frac{\|f_{\text{opt}}\|_{\mathcal{H}}^2}{\sigma^2}. \tag{7}$$

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

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

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

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* yield 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,

$$\bar{f}_\lambda = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2. \tag{8}$$

The above algorithm is a penalized empirical risk minimization problem. The *representer theorem* (Schölkopf and Smola, 2002) shows that Problem (8) can be restricted to

$$\mathcal{H}_n = \{\sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \mid \alpha_1, \ldots, \alpha_n \in \mathbb{R}\}. \tag{9}$$

Indeed, a direct computation shows that the solution of Problem (8) is

$$\bar{f}_\lambda(x) = \sum_{i=1}^{n} \bar{\alpha}_{\lambda i} k(x, x_i), \quad \bar{\alpha}_\lambda = (K + \lambda nI)^{-1} y, \tag{10}$$

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 $\lambda^* = \frac{1}{SNR}$, the following inequalities hold,*

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

We add a few comments. First, as announced, the above result quantifies the benefits of regularization. Indeed, it shows that there exists a λ^* 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_{eff} can be less than a fractional power of d^*, so that $d_{\text{eff}} \ll d^*$ and

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

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 is computed and then a model selected or found by aggregation (Hastie et al., 2001). Let $\Lambda \subset \mathbb{R}$ be the discrete set of regularization parameter guesses. 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 $O\left(n^3|\Lambda|\right)$ 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.

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

$$\breve{f}_t(x) = \sum_{i=1}^{n} \breve{\alpha}_{t,i} k(x_i, x), \tag{11}$$

$$\breve{\alpha}_t = \breve{\alpha}_{t-1} - \frac{\gamma}{n}(K\breve{\alpha}_{t-1} - y), \tag{12}$$

where $\gamma < 1/\|K\|$ and $\breve{\alpha}_0 = 0$. 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 that of KRLS.

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

$$\mathbb{E}R\left(\breve{f}_{\gamma,t}\right) \leq c_t \, \mathbb{E}R\left(\bar{f}_{\frac{1}{\gamma t}}\right),$$

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 $t^* = \frac{1}{\gamma\lambda^*} = \frac{\mathrm{SNR}}{\gamma}$, we have that

$$\mathbb{E}R\left(\breve{f}_{\gamma,t^*}\right) \approx \mathbb{E}R\left(\bar{f}_{\lambda^*}\right) \leq \mathbb{E}R\left(f_{\mathrm{ols}}\right).$$

The statistical accuracy of early stopping is essentially the same as KRLS and can be vastly better than a naive ERM approach. Note that the cost of computing the best possible solution with early stopping is $O(n^2 t^*) = O(n^2 \mathrm{SNR})$. Thus the computational time of early stopping is proportional to the signal to noise ratio, hence 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 (8) to a subspace $\mathcal{H}_m \subseteq \mathcal{H}_n$ defined as

$$\mathcal{H}_m = \{\sum_{i=1}^{m} c_i k(\cdot, \tilde{x}_i) | c_1, \ldots, c_m \in \mathbb{R}\}. \tag{13}$$

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

$$\tilde{f}_{m,\lambda}(x) = \sum_{i=1}^{m} (\tilde{\alpha}_{m,\lambda})_i k(x, \tilde{x}_i), \tag{14}$$

$$\tilde{\alpha}_{m,\lambda} = (K_{nm}^\top K_{nm} + \lambda n K_{mm})^\dagger K_{nm}^\top y, \tag{15}$$

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 $(K_{mm})_{i,j} = k(\tilde{x}_i, \tilde{x}_j)$. A more efficient formulation can also be derived. Indeed, we rewrite Problem (8), restricted to $\mathcal{H}_m$, as

$$\tilde{\alpha}_{m,\lambda} = \operatorname*{argmin}_{\alpha \in \mathbb{R}^m} \|K_{nm}\alpha - y\|^2 + \lambda \alpha^\top K_{mm} \alpha \tag{16}$$

$$= R \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|K_{nm} R\beta - y\|^2 + \lambda \|\beta\|^2, \tag{17}$$

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

$$\tilde{\alpha}_{m,\lambda} = R(A^\top A + \lambda n I)^{-1} A^\top y, \tag{18}$$

with $A = K_{nm} R$ (see Prop. 2 in Section A 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 $S \in \mathbb{R}^{m \times p}$ such that $S^\top S = I$, $D \in \mathbb{R}^{p \times m}$ an upper triangular matrix and $T \in \mathbb{R}^{p \times p}$ an invertible triangular matrix that is the Cholesky decomposition of $S^\top K_{mm} S$. Assuming $p \approx m$, 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 thus obtained suitably choosing the points in M.

Theorem 4 (Theorem 1 of Bach (2013))**.** *Let $m \leq n$ and $M = \{\tilde{x}_1, \ldots, \tilde{x}_m\}$ be a subset of the training set uniformly chosen at random. Let $\tilde{f}_{m,\lambda}$ be as in Equation* (14) *and $\bar{f}_\lambda$ as in Equation* (10) *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

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

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

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

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 results indicate 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 (13). The algorithm thus obtained is given by,

$$\hat{f}_{m,\gamma,t}(x) = \sum_{i=1}^{m} (\hat{\alpha}_{m,t})_i k(\tilde{x}_i, x), \tag{19}$$

$$\hat{\beta}_{m,\gamma,t} = \hat{\beta}_{m,t-1} - \frac{\gamma}{n} R^\top (K_{nm}^\top (K_{nm} \hat{\beta}_{m,t-1} - y)), \tag{20}$$

$$\hat{\alpha}_{m,\gamma,t} = R \beta_{m,t}, \tag{21}$$

for all $x \in \mathcal{X}$, where $\gamma = 1/(\sup_{1 \le i \le n} k(x_i, x_i))$ and $\hat{\beta}_{m,0} = 0$. Considering that the cost of computing 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 one 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 $m \le n$ and M be a subset of the training set. Let $\hat{f}_{m,\gamma,t}$ be the NYTRO solution as in Equation* (19)*, $\tilde{f}_{m,\frac{1}{\gamma t}}$ the NKRLS solution as in Equation* (14)*. When $t \ge 2$ and $\gamma < \|K_{nm} R\|^2$ (for example $\gamma = 1/\max_i k(x_i, x_i)$) the following holds*

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

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

Note that the above result holds for any $m \le n$ 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 inverse problems (see Engl et al., 1996). In the next corollary we compare NYTRO and KRLS, by combining Theorem 5 and 4, hence considering M to be chosen uniformly at random from the training set.

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

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

Let $\bar{f}_{\frac{1}{\gamma t}}$ be the KRLS solution as in Equation 10 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

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

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

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 statistical 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 λ, that is $\lambda^* = \frac{1}{\mathrm{SNR}}$, 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

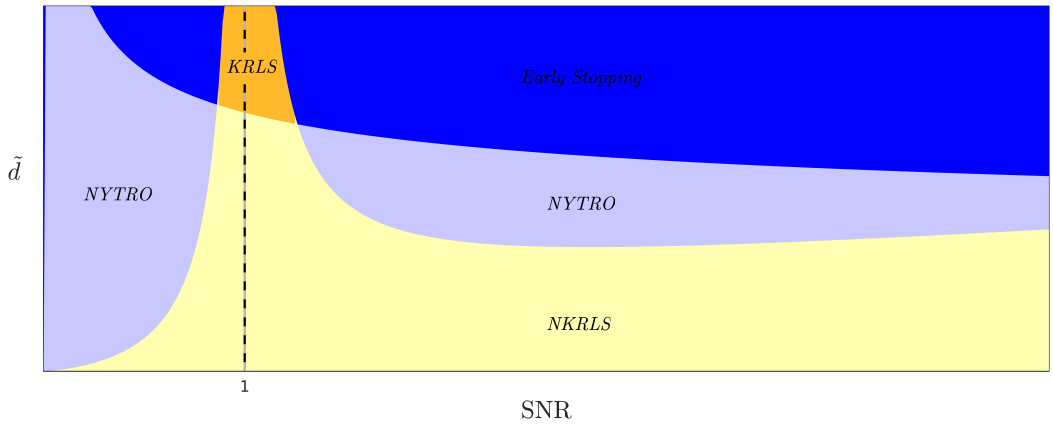

Figure 1: The graph represents the family of learning problems parametrized by the dimensionality $\tilde{d}$ and the signal to noise ratio SNR (see Equations 4, 7). The four regions represent the regimes where different some algorithm are faster than the others. Purple: NYTRO is faster, Blue: Early Stopping is faster, Orange: KRLS is faster, Yellow: NKRLS is faster – see Section 3.3.

requirements. 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 λ^* (see Figure 1), it is useful to parametrize the family of learning problems with respect to the signal to noise ratio defined in Equation (7) 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_{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.

Table 1: Specifications of the datasets used in time-accuracy comparison experiments. σ is the bandwidth of the Gaussian kernel.

Dataset	n	n_{test}	d	σ
InsuranceCompany	5822	4000	85	3
Adult	32562	16282	123	6.6
Ijcnn	49990	91701	22	1
YearPrediction	463715	51630	90	1
CovertypeBinary	522910	58102	54	1

As illustrated in Figure 1, NYTRO is preferable when SNR ≤ 1, that is when the problem is quite noisy. When SNR > 1, then NYTRO is faster when the dimensionality of the problem is sufficiently big. Note that, in particular, the area of the NYTRO region when 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 big, while the second when it is very small. There is a peak around SNR ≈ 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 SNR ≈ 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.

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 Λ is the discrete set of cross-validated λ guesses. This last aspect is practically very relevant 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[1], 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 × 2.10GHz Intel® Xeon® E5-2620 v2 CPUs and 132 GB of RAM.

We compare the algorithms on the benchmark datasets reported in Table 1[2]. In the table we also report the bandwidth parameter σ 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 normalized to be between 0 and 1.

Model selection. 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. The Tikhonov regularization parameter λ for KRLS and NKRLS is chosen among 100 guesses logarithmically spaced in $\left[10^{-15}, 1\right]$, by computing the validation error for each model and selecting the λ^* associated with the lowest one. The regularization parameter for ES and NYTRO is the number of iterations t. The optimal t^* is selected by considering the evolution of the validation error. As an early stopping rule, for each dataset an iteration t^* is chosen such that the relative decrease of the validation error at t^* is smaller than a threshold. After model selection, we evaluate the predictive 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

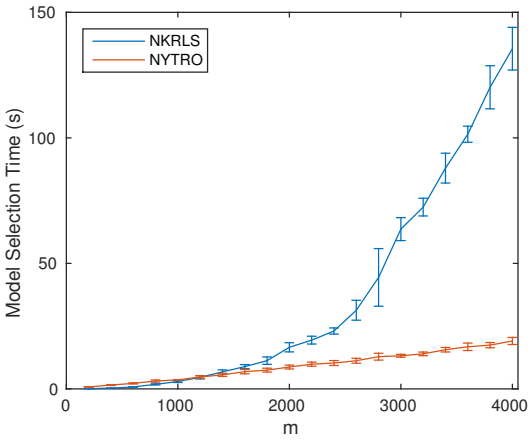

Figure 2: Training time of NKRLS and NYTRO on the `cpuSmall` dataset as the subsampling level m varies linearly between 100 and 4000. Experiment with 5 repetitions. Mean and standard deviation reported.

Figure 2. We performed training on the `cpuSmall`[3] 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 λ, 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).

Time-accuracy benchmarking. We also compare the training time and accuracy performances for KRLS, KOLS, Early Stopping (ES), NKRLS and NYTRO, reporting the selected hyperparameter, the model selection time and the test error in Table 2. The model selection time refers to the choice of λ^* for KRLS and NKRLS and of the iteration t^* for ES and NYTRO, realized via hold-out cross validation as previously specified. All the experiments are repeated 5 times. On the considered datasets, 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 `Covertype` 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

[1]Code available at: `http://lcsl.github.io/NYTRO/`

[2]Datasets available at `http://archive.ics.uci.edu/ml` or `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`

[3]Available at `http://www.cs.toronto.edu/~delve/data/datasets.html`

Table 2: Model selection time and test accuracy comparison on benchmark datasets.

Dataset		*KOLS*	*KRLS*	*Early Stopping*	*NKRLS*	*NYTRO*
InsuranceCompany	*Time* (s)	**1.04**	97.48 ± 0.77	2.92 ± 0.04	20.32 ± 0.50	5.49 ± 0.12
$n = 5822$	*RMSE*	5.179	**0.4651 ± 0.0001**	**0.4650 ± 0.0002**	**0.4651 ± 0.0003**	**0.4651 ± 0.0003**
$m = 2000$	*Par.*	NA	3.27e-04	494 ± 1.7	5.14e-04 ± 1.42e-04	491 ± 3
Adult	*Time* (s)	112	4360 ± 9.29	5.52 ± 0.23	5.95 ± 0.31	**0.85 ± 0.05**
$n = 32562$	*RMSE*	1765	**0.645 ± 0.001**	0.685 ± 0.002	0.6462 ± 0.003	0.6873 ± 0.003
$m = 1000$	*Par.*	NA	4.04e-05 ± 1.04e-05	39.2 ± 1.1	4.04e-05 ± 1.83e-05	4.9 ± 0.3
Ijcnn	*Time* (s)	271	825.01 ± 6.81	154.82 ± 1.24	160.28 ± 1.54	**80.9 ± 0.4**
$n = 49990$	*RMSE*	730.62	0.615 ± 0.002	**0.457 ± 0.001**	0.469 ± 0.003	**0.457 ± 0.001**
$m = 5000$	*Par.*	NA	1.07e-08 ± 1.47e-08	489 ± 7.2	1.07e-07 ± 1.15e-07	328.7 ± 2.6
YearPrediction	*Time* (s)				1188.47 ± 36.7	**887 ± 6**
$n = 463715$	*RMSE*	NA	NA	NA	**0.1015 ± 0.0002**	0.1149 ± 0.0002
$m = 10000$	*Par.*				3.05e-07 ± 1.05e-07	481 ± 6.1
CovertypeBinary	*Time* (s)				1235.21 ± 42.1	**92.69 ± 2.35**
$n = 522910$	*RMSE*	NA	NA	NA	1.204 ± 0.008	**0.918 ± 0.006**
$m = 10000$	*Par.*				9.33e-09 ± 1.12e-09	39.2 ± 2.3

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 `Covertype` case, where it requires 90% less time for model selection.

Acknowledgments

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 by 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.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1657–1665. Curran Associates, Inc., 2015.

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*, 151:501504, 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.

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.