Regularized ERM on random subspaces

Andrea Della Vecchia
MaLGa, DIMA
University of Genova
dellavecchia@dima.unige.it

Jaouad Mourtada
CREST, ENSAE
IP Paris
jaouad.mourtada@ensae.fr

Ernesto De Vito
MaLGa, DIMA
University of Genova
ernesto.devito@unige.it

Lorenzo Rosasco
MaLGa, DIBRIS
University of Genova
IIT
CBMM, MIT
lorenzo.rosasco@unige.it

Abstract

We study a natural extension of classical empirical risk minimization, where the hypothesis space is a random subspace of a given space. In particular, we consider possibly data dependent subspaces spanned by a random subset of the data, recovering as a special case Nyström approaches for kernel methods. Considering random subspaces naturally leads to computational savings, but the question is whether the corresponding learning accuracy is degraded. These statistical-computational tradeoffs have been recently explored for the least squares loss and self-concordant loss functions, such as the logistic loss. Here, we work to extend these results to convex Lipschitz loss functions, that might not be smooth, such as the hinge loss used in support vector machines. This extension requires developing new proofs, that use different technical tools. Our main results show the existence of different settings, depending on how hard the learning problem is, for which computational efficiency can be improved with no loss in performance. Theoretical results are illustrated with simple numerical experiments.

1 Introduction

Despite excellent practical performances, state of the art machine learning (ML) methods often require huge computational resources, motivating the search for more efficient solutions. This has led to a number of new results in optimization [Johnson and Zhang (2013); Schmidt et al. (2017)], as well as the development of approaches mixing linear algebra and randomized algorithms [Mahoney (2011); Drineas and Mahoney (2005); Woodruff (2014); Calandriello et al. (2017)]. While these techniques are applied to empirical objectives, in the context of learning it is natural to study how different numerical solutions affect statistical accuracy. Interestingly, it is now clear that there is a whole set of problems and approaches where computational savings do not lead to any degradation in terms of learning performance [Rudi et al. (2015); Bach (2017); Bottou and Bousquet (2008); Sun et al. (2018); Li et al. (2019); Rudi and Rosasco (2017); Calandriello and Rosasco (2018)].

Here, we follow this line of research and study an instance of regularized empirical risk minimization where, given a fixed high– possibly infinite– dimensional hypothesis space, the search for a solution is restricted to a smaller– possibly random– subspace. This is equivalent to considering sketching operators [Kpotufe and Sriperumbudur (2019)], or equivalently regularization with random projections [Woodruff (2014)]. For infinite dimensional hypothesis spaces, it includes Nyström methods used for kernel methods [Smola and Schölkopf (2000)] and Gaussian processes [Williams and Seeger (2001)]. Recent works in supervised statistical learning has focused on smooth loss functions [Rudi et al. (2015); Bach (2013); Marteau-Ferey et al. (2019b)], whereas here we consider convex, Lipschitz but possibly non smooth losses.

In particular, if compared with results for quadratic and logistic loss, our proof follows a different path. For square loss, all relevant quantities (i.e. loss function, excess risk) are quadratic, while the regularized estimator has an explicit expression, allowing for an explicit analysis based on linear algebra and matrix concentration [Tropp (2012)]. Similarly, the study for logistic loss can be reduced to the quadratic case through a local quadratic approximation based on the

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

self-concordance property. Instead here convex Lipschitz but non-smooth losses such as the hinge loss do not allow for such a quadratic approximation and we need to combine empirical process theory [Boucheron et al. (2013)] with results for random projections. In particular, fast rates require considering localized complexity measures [Steinwart and Christmann (2008); Bartlett et al. (2005); Koltchinskii et al. (2006)]. Related ideas have been used to extend results for random features from the square loss [Rudi and Rosasco (2017)] to general loss functions [Li et al. (2019); Sun et al. (2018)].

Our main interest is characterizing the relation between computational efficiency and statistical accuracy. We do so studying the interplay between regularization, subspace size and different parameters describing how are hard or easy is the considered problem. Indeed, our analysis starts from basic assumption, that eventually we first strengthen to get faster rates, and then weaken to consider more general scenarios. Our results show that also for convex, Lipschitz losses there are settings in which the best known statistical bounds can be obtained while substantially reducing computational requirements. Interestingly, these effects are relevant but also less marked than for smooth losses. In particular, some form of adaptive sampling seems needed to ensure no loss of accuracy and achieve sharp learning bounds. In contrast, uniform sampling suffices to achieve similar results for smooth loss functions. It is an open question whether this is a byproduct of our analysis, or a fundamental limitation. Some preliminary numerical results complemented with numerical experiments are given considering benchmark datasets.

The rest of the paper is organized as follow. In Section 2, we introduce the setting. In Section 3, we introduce the ERM approach we consider. In Section 4, we present and discuss the main results and defer the proofs to the appendix. In Section 5, we collect some numerical results.

2 Statistical learning with ERM

Let (X, Y) be random variables in $\mathcal{H} \times \mathcal{Y}$, with distribution P satisfying the following conditions.

Assumption 1. *The space $\mathcal{H}$ is a real separable Hilbert space with scalar product $\langle \cdot, \cdot \rangle$, $\mathcal{Y}$ is a Polish space, and there exists $\kappa > 0$ such that $\|X\| \leq \kappa$ almost surely.*

Since X is bounded, the covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$ given by $\Sigma = \mathbb{E}[X \otimes X]$ can be shown to be self-adjoint, positive and trace class with $\mathrm{Tr}(\Sigma) \leq \kappa$. We can think of $\mathcal{H}$ and $\mathcal{Y}$ as input and output spaces, respectively, and some examples are relevant.

Example 1. An example is linear estimation, where $\mathcal{H}$ is $\mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$. Another example is kernel methods, where $\mathcal{H}$ is a separable reproducing kernel Hilbert space on a measurable space $\mathcal{X}$. The data are then mapped from $\mathcal{X}$ to $\mathcal{H}$ through the feature map $x \mapsto K(\cdot, x) = K_x$ where $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the (measurable) reproducing kernel of $\mathcal{H}$ [Steinwart and Christmann (2008)].

We denote by $\ell : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ the loss function. Given a function f on $\mathcal{H}$ with values in $\mathbb{R}$, we view $\ell(y, f(x))$ as the error made predicting y by $f(x)$. We make the following assumption.

Assumption 2 (Lipschitz loss)**.** *The loss function $\ell : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is convex and Lipschitz in its second argument, namely there exists $G > 0$ such that for all $y \in \mathcal{Y}$ and $a, a' \in \mathbb{R}$,*

$$|\ell(y, a) - \ell(y, a')| \leq G|a - a'| \quad \text{and} \quad \ell_0 = \sup_{y \in \mathcal{Y}} \ell(y, 0). \tag{1}$$

Example 2 (Hinge loss & other loss functions)**.** The main example we have in mind is the hinge loss $\ell(y, a) = |1 - ya|_+ = \max\{0, 1 - ya\}$, with $\mathcal{Y} = \{-1, 1\}$, which is convex but not differentiable, and for which $G = 1$ and $\ell_0 = 1$. Another example is the logistic loss $\ell(y, a) = \log(1 + e^{-ya})$, for which $G = 1$ and $\ell_0 = \log 2$.

Given a loss, the corresponding expected risk $L : \mathcal{H} \to [0, \infty)$ is for all $w \in \mathcal{H}$

$$L(w) = \mathbb{E}\Big[\ell(Y, \langle w, X \rangle)\Big] = \int_{\mathcal{H} \times \mathcal{Y}} \ell(y, \langle w, x \rangle) dP(x, y),$$

and can be easily shown to be convex and Lipschitz continuous.

In this setting, we are interested in the problem of solving

$$\min_{w \in \mathcal{H}} L(w), \tag{2}$$

when the distribution P is known only through a training set of independent samples $D = (x_i, y_i)_{i=1}^n \sim P^n$. Since we only have the data D, we cannot solve the problem exactly and given an empirical approximate solution $\widehat{w}$, a natural error measure is the the excess risk

$$L(\widehat{w}) - \inf_{w \in \mathcal{H}} L(w),$$

which is a random variable through its dependence on $\widehat{w}$, and hence on the data. Notice also that, in the case of hinge loss, an upper bound on the excess risk is also an upper bound on the classification risk, i.e. the risk associated with the $0-1$ loss $\ell_{0-1}(y, a) := \mathbb{1}_{(-\infty, 0]}(y \,\mathrm{sign}(a))$ (see Zhang's inequality in Steinwart and Christmann (2008), Theorem 2.31).
In the following we are interested in characterizing its distribution for finite sample sizes. Next we discuss how approximate solutions can be obtained from data.

2.1 Empirical risk minimization (ERM)

A natural approach to derive approximate solutions is based on replacing the expected risk with the empirical risk $\widehat{L} : \mathcal{H} \to [0, \infty)$ defined for all $w \in \mathcal{H}$ as

$$\widehat{L}(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle w, x_i \rangle).$$

We consider regularized empirical risk minimization (ERM) based on the solution of the problem,

$$\min_{w \in \mathcal{H}} \widehat{L}_\lambda(w), \qquad \widehat{L}_\lambda(w) = \widehat{L}(w) + \lambda \|w\|^2. \tag{3}$$

Note that $\widehat{L}_\lambda : \mathcal{H} \to \mathbb{R}$ is continuous and strongly convex, hence there exists a unique minimizer $\widehat{w}_\lambda$. If we let $\widehat{X}$ denote the data matrix, by the representer theorem [Wahba (1990); Schölkopf et al. (2001)] there exists $c \in \mathbb{R}^n$ such that

$$\widehat{w}_\lambda = \widehat{X}^\top c \in \operatorname{span}\{x_1, \dots, x_n\}. \tag{4}$$

The expression of the coefficient c depends on the considered loss function. Next, we comment on different approaches to obtain a solution when ℓ is the hinge loss. We add one remark first.

Remark 1 (Constrained ERM). A related approach is based on considering the problem

$$\min_{\|w\| \le R} \widehat{L}(w). \tag{5}$$

Minimizing (3) can be seen as a Lagrange multiplier formulation of the above problem. While these problems are equivalent (see Boyd and Vandenberghe (2004), Section 5.5.3), the exact correspondence is implicit. As a consequence their statistical analysis differ. We primarily discuss Problem (3), but also analyze Problem (5) in Appendix I.

Example 3 (Representer theorem for kernel machines)**.** In the context of kernel methods, see Example 1, the above discussion, and in particular (4) are related to the well known representer theorem. Indeed, the linear parameter w corresponds to a function $f \in \mathcal{H}$ in the RKHS, while the norm $\|\cdot\|$ is the RKHS norm $\|\cdot\|_{\mathcal{H}}$. The representer theorem (4) then simply states that there exists constants c_i such that the solution of the regularized ERM can be written as $\widehat{f}_\lambda(x) = \sum_{i=1}^n K(x, x_i) c_i \in \operatorname{span}\{K_{x_1}, \dots, K_{x_n}\}$.

2.2 Computations with the hinge loss

Minimizing (3) can be solved in many ways and we provide some basic considerations. If $\mathcal{H}$ is finite dimensional, iteratively via gradient methods can be used. For example, the subgradient method [Boyd and Vandenberghe (2004)] applied to (3) is given, for some suitable w_0 and step-size sequence $(\eta_t)_t$, by

$$w_{t+1} = w_t - \eta_t \left(\frac{1}{n} \sum_{i=1}^{n} y_i x_i g_i(w_t) + 2\lambda w_t \right), \tag{6}$$

where $g_i(w) \in \partial \ell(y_i, \langle w, x_i \rangle)$ is the subgradient of the map $a \mapsto \ell(y_i, a)$ evaluated at $a = \langle w, x_i \rangle$, see also Rockafellar (1970). The corresponding iteration cost is $O(nd)$ in time and memory. Clearly, other variants can be considered, for example adding a momentum term [Nesterov (2018)], stochastic gradients and mini-batching or considering other approaches for example based on coordinate descent [Shalev-Shwartz and Zhang (2013)]. When $\mathcal{H}$ is infinite dimensional a different approach is possible, provided $\langle x, x' \rangle$ can be computed for all $x, x' \in \mathcal{H}$. For example, it is easy to prove by induction that the iteration in (6) satisfies $w_t = \widehat{X}^\top c_{t+1}$, where

$$c_{t+1} = c_t - \eta_t \left(\frac{1}{n} \sum_{i=1}^{n} y_i e_i g_i(\widehat{X}^\top c_t) + 2\lambda c_t \right), \tag{7}$$

and where $e_1, \dots, e_n$ is the canonical basis in $\mathbb{R}^n$. The cost of the above iteration is $O(n^2 C_K)$ for computing $g_i(w) \in \partial \ell \left(y_i, \left\langle \widehat{X}^\top c_t, x_i \right\rangle \right) = \partial \ell \left(y_i, \sum_{j=1}^n \langle x_j, x_i \rangle (c_t)_i \right)$, where C_K is the cost of evaluating one inner product. Also in this case, a number of other approaches can be considered, see e.g. (Steinwart and Christmann, 2008, Chap. 11) and references therein. We illustrate the above ideas for the hinge loss.

Example 4 (Hinge loss & SVM)**.** Considering problem (3) with the hinge loss corresponds to support vector machines for classification. With this choice $\partial \ell(y_i, \langle w, x_i \rangle) = 0$ if $y_i \langle w, x_i \rangle > 1$, $\partial \ell(y_i, \langle w, x_i \rangle) = [-1, 0]$ if $y_i \langle w, x_i \rangle = 1$ and $\partial \ell(y_i, \langle w, x_i \rangle) = -1$ if $y_i \langle w, x_i \rangle < 1$. In particular, in (7) we can take $g_i(w) = -\mathbb{1}_{[y_i \langle w, x_i \rangle \le 1]}$.

3 ERM on random subspaces

In this paper, we consider a variant of ERM based on considering a subspace $\mathcal{B} \subset \mathcal{H}$ and the corresponding regularized ERM problem,

$$\min_{\beta \in \mathcal{B}} \widehat{L}_\lambda(\beta) \tag{8}$$

with $\widehat{\beta}_\lambda$ the unique minimizer. As clear from (4), choosing $\mathcal{B} = \mathcal{H}_n = \operatorname{span}\{x_1, \dots, x_n\}$ is not a restriction and yields the same solution as considering (3). From this observation a natural choice is to consider for $m \le n$,

$$\mathcal{B}_m = \operatorname{span}\{\widetilde{x}_1, \dots, \widetilde{x}_m\} \tag{9}$$

with $\{\widetilde{x}_1, \dots, \widetilde{x}_m\} \subset \{x_1, \dots, x_n\}$ a subset of the input points. A basic idea we consider is to sample the points uniformly at random. Another more refined choice we consider is sampling exactly or approximately (see Definition 2 in the Appendix) according to the leverages scores [Drineas et al. (2012)]

$$l_i(\alpha) = \left\langle x_i, (\widehat{X}\widehat{X}^\top x + \alpha In)^{-1} x_i \right\rangle \qquad i = 1, \dots, n. \tag{10}$$

While leverage scores computation is costly, approximate leverage scores (ALS) computation can be done efficiently, see Rudi et al. (2018) and references therein. Following Rudi et al. (2015), other choices are possible. Indeed for any $q \in \mathbb{N}$ and $z_1, \dots, z_q \in \mathcal{H}$ we could consider $\mathcal{B} = \operatorname{span}\{z_1, \dots, z_q\}$ and derive a formulation as in (11) replacing $\widetilde{X}$ with the matrix Z with rows $z_1, \dots, z_q$. We leave this discussion for future work. Here, we focus on the computational benefits of considering ERM on random subspaces and analyze the corresponding statistical properties.

The choice of $\mathcal{B}_m$ as in (9) allows to improve computations with respect to (4). Indeed, $\beta \in \mathcal{B}_m$ is equivalent to the existence of $b \in \mathbb{R}^m$ s.t. $\beta = \widetilde{X}^\top b$, so that we can replace (8) with the problem

$$\min_{b \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^n \ell\left(y_i, \left\langle \widetilde{X}^\top b, x_i \right\rangle\right) + \lambda \left\langle b, \widetilde{X}\widetilde{X}^\top b \right\rangle_m$$

where $\langle \cdot, \cdot \rangle_m$ is the usual scalar product in $\mathbb{R}^m$. Further, since $\widetilde{X}\widetilde{X}^\top \in \mathbb{R}^{m \times m}$ is symmetric and positive semi-definite, we can derive a formulation close to that in (3), considering the reparameterization $a = (\widetilde{X}\widetilde{X}^\top)^{1/2} b$ which leads to,

$$\min_{a \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^n \ell\left(y_i, \langle a, \aleph_i \rangle_m\right) + \lambda \|a\|_m^2, \tag{11}$$

where for all $i = 1, \dots, n$, we defined the embedding $x_i \mapsto \aleph_i = ((\widetilde{X}\widetilde{X}^\top)^{1/2})^\dagger \widetilde{X} x_i$ and with $\|\cdot\|_m$ we refer to the 2-norm in $\mathbb{R}^m$. Note that this latter operation only involves the inner product in $\mathcal{H}$ and hence can be computed in $O(m^3 + nm^2 C_K)$ time. The subgradient method for (11) has a cost $O(nm)$ per iteration. In summary, we obtained that the cost for the ERM on subspaces is $O(nm^2 C_K + nm \cdot \#\text{iter})$ and should be compared with the cost of solving (7) which is $O(n^2 C_K + n^2 \cdot \#\text{iter})$. The corresponding costs to predict new points are $O(mC_K)$ and $O(nC_K)$, while the memory requirements are $O(mn)$ and $O(n^2)$, respectively. Clearly, memory requirements can be reduced recomputing things on the fly. As clear from the above discussion, computational savings can be drastic, as long as $m < n$, and the question arises of how this affect the corresponding statistical accuracy. Next section is devoted to this question.

Example 5. [Kernel methods and Nyström approximations] Again, following Example 1 and 3, we can specialize our setting to kernel methods where $\beta \in \operatorname{span}\{\widetilde{x}_1, \dots, \widetilde{x}_m\}$ is replaced by $\widetilde{f}(x) = \sum_{i=1}^m K(x, \widetilde{x}_i)\widetilde{c}_i \in \operatorname{span}\{K_{\widetilde{x}_1}, \dots, K_{\widetilde{x}_m}\}$ while the embedding $x_i \mapsto \aleph_i = ((\widetilde{X}\widetilde{X}^\top)^{1/2})^\dagger \widetilde{X} x_i$ becomes $x_i \mapsto \aleph_i = (\widetilde{K}^{1/2})^\dagger (K(\widetilde{x}_1, x_i), \dots, K(\widetilde{x}_m, x_i))^\top$, with $\widetilde{K}_{i,j} = K(\widetilde{x}_i, \widetilde{x}_j)$.

4 Statistical analysis of ERM on random subspaces

We divide the presentation of the results in three parts. First, we consider a setting where we make basic assumptions. Then, we discuss improved results considering more benign assumptions. Finally, we describe general results covering also less favorable conditions. In all cases, we provide simplified statements for the results, omitting numerical constants, logarithmic and higher order terms, for ease of presentation. The complete statements and the proofs are provided in the appendices.

4.1 Basic setting

In this section, we only assume the best in the model to exist.

Assumption 3. *There exists $w_* \in \mathcal{H}$ such that $L(w_*) = \min_{w \in \mathcal{H}} L(w)$.*

We first provide some benchmark results for regularized ERM under this assumption.

Theorem 1 (Regularized ERM)**.** *Under Assumption 1, 2, 3, the following inequality holds, for all $\lambda > 0$ and $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$L(\widehat{w}_\lambda) - L(w_*) \lesssim \frac{G^2 \kappa^2 \log(1/\delta)}{\lambda n} + \lambda \|w_*\|^2.$$

Hence letting $\lambda \asymp (G\kappa/\|w_*\|)\sqrt{\log(1/\delta)/n}$ leads to a rate of $O(\|w_*\|\sqrt{\log(1/\delta)/n})$.

The proof of Theorem 1 is given in Appendix B, where a more general result is stated. It shows the excess risk bound for regularized ERM arises from a trade-off between an estimation and an approximation term. While this result can be derived specializing more refined analysis, see e.g. Steinwart and Christmann (2008) or later sections, as well as Shalev-Shwartz et al. (2010), we provide a simple self-contained proof which is of interest in its own right. Similar bounds in high-probability for ERM constrained to the ball of radius $R \geqslant \|w_*\|$ can be obtained through a uniform convergence argument over such balls, see Bartlett and Mendelson (2002); Meir and Zhang (2003); Kakade et al.

(2009). In order to apply this to regularized ERM, one could in principle use the fact that by Assumption 2, $\|\widehat{w}_\lambda\| \leqslant \sqrt{\ell_0/\lambda}$ (see Appendix) [Steinwart and Christmann (2008)], but this yields a suboptimal dependence in λ. Finally, a similar rate for $\widehat{w}_\lambda$, though only in expectation, can be derived through a stability argument [Bousquet and Elisseeff (2002); Shalev-Shwartz et al. (2010)]. Our proof proceeds as follows. First, by uniform convergence over balls and a union bound, one has $L(\widehat{w}_\lambda) - \widehat{L}(\widehat{w}_\lambda) \leqslant c\kappa\|\widehat{w}_\lambda\|/\sqrt{n}$ with high probability for some c.
Noting that $c\kappa\|\widehat{w}_\lambda\|/\sqrt{n} \leqslant \lambda\|\widehat{w}_\lambda\|^2 + c^2\kappa^2/(\lambda n)$, we obtain

$$L(\widehat{w}_\lambda) \leqslant \widehat{L}_\lambda(\widehat{w}_\lambda) + \frac{c^2\kappa^2}{\lambda n} \leqslant L_\lambda(w_\lambda) + \frac{c^2\kappa^2}{\lambda n} + \frac{c\kappa\|w_\lambda\|}{\sqrt{n}}$$

using the definition of $\widehat{w}_\lambda$ and a Hoeffding bound. One can conclude by noting that $L(w_\lambda)+\lambda\|w_\lambda\|^2 \leqslant L(w_*)+\lambda\|w_*\|^2$ (by definition of w_λ) and $\|w_\lambda\| \leqslant \|w_*\|$.

Theorem 2 (Regularized ERM on subspaces)**.** *Fix $\mathcal{B} \subseteq \mathcal{H}$, $\lambda > 0$ and $0 < \delta < 1$. Under Assumptions 1, 2, 3, with probability at least $1-\delta$,*

$$L(\widehat{\beta}_\lambda) - L(w_*) \lesssim \frac{G^2\kappa^2\log(1/\delta)}{\lambda n} + \lambda\,\|w_*\|^2 + \sqrt{\mu_{\mathcal{B}}}G\,\|w_*\|\,.$$

Compared to Theorem 1, the above result shows that there is an extra approximation error term due to considering a subspace. The coefficient $\mu_{\mathcal{B}}$ appears in the analysis also for other loss functions, see e.g. Rudi et al. (2015); Marteau-Ferey et al. (2019b). Roughly speaking, it captures how well the subspace $\mathcal{B}$ is adapted to the problem. We next develop this reasoning, specializing the above result to a random subspace $\mathcal{B} = \mathcal{B}_m$ as in (9). Note that, if $\mathcal{B}$ is random then $\mu_{\mathcal{B}}$ is a random variable through its dependence on $\mathcal{P}_{\mathcal{B}}$ and on $\mathcal{B}$. We denote by $\widehat{\beta}_{\lambda,m}$ the unique minimizer of $\widehat{L}_\lambda$ on $\mathcal{B}_m$ and by $\mathcal{P}_m = \mathcal{P}_{\mathcal{B}_m}$ the corresponding projection. Further, it is also useful to introduce the so-called effective dimensions [Zhang (2005); Caponnetto and De Vito (2007); Rudi et al. (2015)]. We denote by P_X the distribution of X, with $\mathrm{supp}(P_X) \subseteq \mathcal{H}$ its support[1], and define for $\alpha > 0$

$$d_{\alpha,2} = \mathrm{Tr}((\Sigma + \alpha I)^{-1}\Sigma), \tag{12}$$

$$d_{\alpha,\infty} = \sup_{x \in \mathrm{supp}(P_X)} \langle x, (\Sigma + \alpha I)^{-1} x\rangle\,. \tag{13}$$

Then, $d_{\alpha,2}$ is finite since Σ is trace class, and $d_{\alpha,\infty}$ is finite since $\mathrm{supp}(P_X)$ is bounded. Further, we denote by $(\sigma_j(\Sigma))_j$ the strictly positive eigenvalues of Σ, with eigenvalues counted with respect to their multiplicity and ordered in a non-increasing way. We borrow the following results from Rudi et al. (2015).

[1]Namely, the smallest closed subset of $\mathcal{H}$ with P_X-measure 1, well-defined since $\mathcal{H}$ is a Polish space [Steinwart and Christmann (2008)].

Proposition 1 (Uniform and leverage scores sampling)**.** *Fix $\alpha > 0$ and $0 < \delta < 1$. With probability at least $1-\delta$*

$$\mu_{\mathcal{B}_m}{}^2 = \left\|\Sigma^{1/2}(I - \mathcal{P}_m)\right\|^2 \leq 3\alpha. \tag{14}$$

provided that $m \gtrsim d_{\alpha,\infty}\log\frac{1}{\alpha\delta}$ for uniform sampling or $m \gtrsim d_{\alpha,2}\log\frac{n}{\delta}$ and $\alpha \gtrsim \frac{1}{n}\log\frac{n}{\delta}$ for ALS sampling.

Moreover, if the spectrum of Σ has a polynomial decay, i.e. for some $p \in (0,1)$

$$\sigma_j(\Sigma) \lesssim j^{-\frac{1}{p}} \tag{15}$$

then (14) holds if $m \gtrsim \frac{1}{\alpha}\log\frac{1}{\alpha\delta}$ for uniform sampling or $m \gtrsim \frac{1}{\alpha^p}\log\frac{n}{\delta}$ and $\alpha \gtrsim \frac{1}{n}\log\frac{n}{\delta}$ for ALS sampling.

Combining the above proposition with Theorem 2 we have the following.

Theorem 3 (Uniform and leverage scores sampling under eigen-decay)**.** *Under Assumption 1, 2, 3 and condition (15), for all $\lambda > 0$, $\alpha > 0$ and $0 < \delta < 1$, with probability $1-\delta$,*

$$L(\widehat{\beta}_{\lambda,m}) - L(w_*) \lesssim \frac{G^2\kappa^2\log(3/\delta)}{\lambda n} + \lambda\,\|w_*\|^2 + \sqrt{\alpha}G\,\|w_*\|\,.$$

Taking $\lambda \asymp \sqrt{\frac{1}{n}\log(n/\delta)}$, $\alpha \asymp \lambda^2 \asymp \frac{1}{n}\log(\frac{n}{\delta})$ and choosing $m \gtrsim n\log n$ points by uniform sampling or $m \gtrsim n^p \log n$ by leverage score sampling, leads to a rate of $O(\sqrt{\frac{\log(n/\delta)}{n}})$.

The above results show that it is possible to achieve the same rate of standard regularized ERM (up to a logarithmic factor), but to do so uniform sampling does not seem to provide a computational benefit. As clear from the proof, computational benefits for smaller subspace dimension would lead to worse rates. This behavior is worse than that allowed by smooth loss functions [Rudi et al. (2015); Marteau-Ferey et al. (2019b)]. These results can be recovered with our approach. Indeed, for both least squares and self-concordant losses, the bound in Theorem (2) can be easily improved to have a linear dependence on $\mu_{\mathcal{B}_m}$, leading to straightforward improvements. We will detail this derivation in a longer version of the paper. Due to space constraints, here we focus on non-smooth losses, since these results, and not only their proof, are new. For this class of loss functions, Theorem 3 shows that leverage scores sampling can lead to better results depending on the spectral properties of the covariance operator. Indeed, if there is a fast eigendecay, then using leverage scores and a subspace dimension $m < n$ one can achieve the same rates as exact ERM. For fast eigendecay (p small), the subspace dimension can decrease dramatically. For example, as a reference for $p = 1/2$ then $m = \sqrt{n}$ suffices. Note that, other decays, e.g. exponential, could

also be considered. These observations are consistent with recent results for random features [Bach (2017); Li et al. (2019); Sun et al. (2018)], while they seem new for ERM on subspaces. Compare to random features the proof techniques have similarities but also differences due to the fact that in general random features do not define subspaces. Finding a unifying analysis would be interesting, but it is left for future work. Also, we note that uniform sampling can have the same properties of leverage scores sampling, if $d_{\alpha,2} \asymp d_{\alpha,\infty}$. This happens under the strong assumptions on the eigenvectors of the covariance operator, but can also happen in kernel methods with kernels corresponding to Sobolev spaces [Steinwart et al. (2009)]. With these comments in mind, here, we focus on subspace defined through leverage scores noting that the assumption on the eigendecay not only allows for smaller subspace dimensions, but can also lead to faster learning rates. Indeed, we study this next.

4.2 Fast rates

In this section we obtain fast rates assuming X to be a sub-gaussian random variable. According to Koltchinskii and Lounici (2014) we have the following definition:

Definition 1 (Sub-gaussian random variable)**.** A centered random variable X in $\mathcal{H}$ will be called C-sub-gaussian iff $\forall p \geqslant 2$

$$\|\langle X, u\rangle\|_{L_p(P)} \leqslant C\sqrt{p}\|\langle X, u\rangle\|_{L_2(P)} \qquad \forall u \in \mathcal{H} \quad (16)$$

Note that (16) implies that all the projections $\langle X, u\rangle$ are real sub-gaussian random variables [Vershynin (2010)] but this is not sufficient since the sub-gaussian norm

$$\| \langle X, u\rangle \|_{\psi_2} = \sup_{p\geqslant 2} \frac{\|\langle X, u\rangle\|_{L_p(P)}}{\sqrt{p}}$$

should be bounded from above by the L_2-norm $\|\langle X, u\rangle\|_{L_2(P)}$. In particular, we stress that, in general, bounded random vectors in $\mathcal{H}$ are not sub-gaussian. The following condition replaces Assumption 1:

Assumption 4. *There exists $C > 0$ such that X is a C-sub-gaussian random variable.*

To exploit the eigendecay assumption and derive fast rates, we begin considering further conditions on the problem. We relax these assumptions in the next section. First, we let for P_X-almost all $x \in \mathcal{H}$

$$f_*(x) = \arg\min_{a\in\mathbb{R}} \int_{\mathcal{Y}} \ell(y, a) dP(y|x) \qquad (17)$$

where $P(y|x)$ is the conditional distribution [2] of y given $x \in \mathcal{H}$ and make the following assumption.

[2] The conditional distribution always exists since $\mathcal{H}$ is separable and $\mathcal{Y}$ is a Polish space [Steinwart and Christmann (2008)],

Assumption 5. *There exists $w_* \in \mathcal{H}$ such that, almost surely, $f_*(X) = \langle w_*, X\rangle$.*

In our context, this is the same as requiring the model to be well specified. Second, following Steinwart and Christmann (2008), we consider a loss that can be *clipped* at $M > 0$ that is such that for all $y' \in \mathcal{Y}, y \in \mathbb{R}$,

$$\ell(y', y^{cl}) \leqslant \ell(y', y), \qquad (18)$$

where y^{cl} denotes the clipped value of y at $\pm M$, i.e.

$$\begin{aligned} y^{cl} &= -M \quad \text{if } y \leqslant -M, \\ y^{cl} &= y \quad \text{if } y \in [-M, M], \\ y^{cl} &= M \quad \text{if } y \geqslant M. \end{aligned}$$

If $w \in \mathcal{H}$, w^{cl} denotes the non-linear function $f(x) = \langle w, x\rangle^{cl}$. This assumption holds for hinge loss with $M = 1$, and for bounded regression. Finally, we make the following assumption on the loss.

Assumption 6 (Simplified Bernstein condition)**.** *There are constants $B, V > 0$, such that for all $w \in \mathcal{H}$,*

$$\ell(Y, \langle w, X\rangle^{cl})) \leqslant B \qquad (19)$$

$$\mathbb{E}[\ell(Y, \langle w, X\rangle^{cl}) - \ell(Y, f_*(X))]^2 \leqslant V\mathbb{E}[\ell(Y, \langle w, X\rangle^{cl}) - \ell(Y, f_*(X)]. \qquad (20)$$

This is a standard assumption to derive fast rates for ERM [Steinwart and Christmann (2008); Bartlett et al. (2005)]. In classification with the hinge loss, it is implied by standard margin conditions characterizing classification noise, and in particular by hard margin assumptions on the data distribution [Audibert and Tsybakov (2007); Tsybakov (2004); Massart et al. (2006); Steinwart and Christmann (2008)]. As discussed before, we next focus on subspaces defined by leverage scores and derive fast rates under the above assumptions.

Theorem 4. *Fix $\lambda > 0$, $\alpha \gtrsim n^{-1/p}$, and $0 < \delta < 1$. Under Assumptions 2, 4, 5, 6 and a polynomial decay of the spectrum of Σ with rate $1/p \in (1, \infty)$, as in (15), and including also the additional hypothesis $\mathbb{E}(\ell(Y, \langle \mathcal{P}_m w_*, X\rangle) - \ell(Y, \langle \mathcal{P}_m w_*, X\rangle^{cl})^2 \lesssim \mathbb{E}(\ell(Y, \langle \mathcal{P}_m w_*, X\rangle) - \ell(Y, \langle \mathcal{P}_m w_*, X\rangle^{cl}))$ then, with probability at least $1 - 2\delta$*

$$L(\widehat{\beta}^{cl}_{\lambda,m}) - L(w_*) \lesssim \frac{1}{\lambda^p n} + \lambda \|w_*\|^2 + \sqrt{\alpha}\|w_*\|$$

provided that n and m are large enough. Further, for ALS sampling with the choice

$$\lambda \asymp n^{-\frac{1}{1+p}}, \quad \alpha \asymp n^{-\frac{2}{1+p}}, \quad m \gtrsim n^{\frac{2p}{1+p}} \log n, \qquad (21)$$

with high probability,

$$L(\widehat{\beta}^{cl}_{\lambda,m}) - L(w_*) \lesssim n^{-\frac{1}{1+p}}. \qquad (22)$$

The above result is a special case of the analysis in the next section, but it is easier to interpret. Compared to Theorem 3 the assumption on the spectrum also leads to an improved estimation error bound and hence improved learning rates. In this sense, these are the *correct* estimates since the decay of eigenvalues is used both for the subspace approximation error and the estimation error. As is clear from (22), for fast eigendecay, the obtained rate goes from $O(1/\sqrt{n})$ to $O(1/n)$. Taking again, $p = 1/2$ leads to a rate $O(1/n^{2/3})$ which is better than the one in Theorem 3. In this case, the subspace defined by leverage scores needs to be chosen of dimension at least $O(n^{2/3})$.

We can now clarify also the need of replacing Assumption 1 with 4. Note that, the choice of α in (21) is not admissible when dealing with bounded variables (see conditions in Lemma 4 in the Appendix). Assuming X sub-gaussian solves the problem allowing to enlarge the admissible range of α to $\alpha \gtrsim n^{-1/p}$, which is always compatible with (21) (see Lemma 5 and Corollary 3 in the Appendix).

Note that again, the subspace dimension is even smaller for faster eigendecay. Next, we extend these results considering weaker, more general assumptions.

4.3 General analysis

Last, we give a general analysis relaxing the above assumptions. We replace Assumption 5 by

$$\inf_{w\in\mathcal{H}} L(w) = \mathbb{E}[\ell(Y, f_*(X))], \tag{23}$$

and introduce the approximation error,

$$\mathcal{A}(\lambda) = \min_{w\in\mathcal{H}} L(w) + \lambda \|w\|^2 - \inf_{w\in\mathcal{H}} L(w). \tag{24}$$

Condition (23) may be relaxed at the cost of an additional approximation term, but the analysis is lengthier and is postponed. It has a natural interpretation in the context of kernel methods, see Example 1, where it is satisfied by universal kernels [Steinwart and Christmann (2008)]. Regarding the approximation error, note that, if w_* exists then $\mathcal{A}(\lambda) \leqslant \lambda \|w_*\|^2$, and we can recover the results in Section 4.1. More generally, the approximation error decreases with λ and learning rates can be derived assuming a suitable decay. Further, we consider a more general form of the Bernstein condition.

Assumption 7 (Bernstein condition)**.** *There exist constants $B > 0$, $\theta \in [0, 1]$ and $V \geqslant B^{2-\theta}$, such that for all $w \in \mathcal{H}$, the following inequalities hold almost surely:*

$$\ell(Y, \langle w, X\rangle^{cl}) \leqslant B, \tag{25}$$

$$\mathbb{E}[\ell(Y, \langle w, X\rangle^{cl}) - \ell(Y, f_*(X))]^2 \leqslant V(\mathbb{E}[\ell(Y, \langle w, X\rangle^{cl}) - \ell(Y, f_*(X))])^{\theta}. \tag{26}$$

Again in classification, the above condition is implied by margin conditions, and the parameter θ characterizes how easy or hard the classification problem is. The strongest assumption is choosing $\theta = 1$, with which we recover the result in the previous section. Then, we have the following result.

Theorem 5. *Fix $\lambda > 0$, $\alpha \gtrsim n^{-1/p}$ and $0 < \delta < 1$. Under Assumptions 2, 4, 7, and a polynomial decay $1/p \in (1, \infty)$ of the spectrum of Σ, as in (15), and including also the additional hypothesis $\mathbb{E}(\ell(Y, \langle \mathcal{P}_m w_\lambda, X\rangle) - \ell(Y, \langle \mathcal{P}_m w_\lambda, X\rangle^{cl})^2 \lesssim \mathbb{E}(\ell(Y, \langle \mathcal{P}_m w_\lambda, X\rangle) - \ell(Y, \langle \mathcal{P}_m w_\lambda, X\rangle^{cl}))$, then with probability at least $1 - 2\delta$*

$$L(\widehat{\beta}^{cl}_{\lambda,m}) - L(f_*) \lesssim \left(\frac{1}{\lambda^p n}\right)^{\frac{1}{2-p-\theta+\theta p}} + \sqrt{\frac{\alpha\mathcal{A}(\lambda)}{\lambda}} + \frac{\log(3/\delta)}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} + \mathcal{A}(\lambda).$$

Furthermore, if there exists $r \in (0, 1]$ such that $\mathcal{A}(\lambda) \lesssim \lambda^r$, then with the choice for ALS sampling

$$\lambda \asymp n^{-\min\{\frac{2}{r+1}, \frac{1}{r(2-p-\theta+\theta p)+p}\}}$$

$$\alpha \asymp n^{-\min\{2, \frac{r+1}{r(2-p-\theta+\theta p)+p}\}}$$

$$m \gtrsim n^{\min\{2p, \frac{p(r+1)}{r(2-p-\theta+\theta p)+p}\}} \log n$$

with high probability

$$L(\widehat{\beta}^{cl}_{\lambda,m}) - L(f_*) \lesssim n^{-\min\{\frac{2r}{r+1}, \frac{r}{r(2-p-\theta+\theta p)+p}\}}.$$

The proof of the above bound follows combining Lemma 5 (see Appendix) with results to analyze the learning properties of regularized ERM with kernels [Steinwart and Christmann (2008)]. While general, the obtained bound is harder to parse. For $r \to 0$ the bound become vacuous and there are not enough assumptions to derive a bound [Devroye et al. (2013)]. Taking $r = 1$ gives the best bound, recovering the result in the previous section when $\theta = 1$. Note that large values of λ are prevented, indicating a saturation effect (see Vito et al. (2005); Mücke et al. (2019)). As before the bound improves when there is a fast eigendecay. Taking $\theta = 1$ we recover the previous bounds, whereas smaller θ lead to worse bounds. Since, given any acceptable choice of p, r and θ, the quantity $\min\{2p, \frac{p(r+1)}{r(2-p-\theta+\theta p)+p}\}$ takes values in $(0, 1)$, the best rate, that differently from before can also be slower than $\sqrt{1/n}$, can always be achieved choosing $m < n$ (up to logarithmic terms). Again the assumption of sub-gaussianity it's necessary to make the choice of α admissible.

5 Experiments

As mentioned in the introduction, a main of motivation for our study is showing that the computational sav-

Table 1: Comparison among the different regimes (up to logarithmic factors) under ALS sampling

	Assumptions	Eigen-decay	Rate	m
Theorem 1	1,2,3	/	$n^{-1/2}$	/
Theorem 3	1,2,3	$\sigma_j(\Sigma) \lesssim j^{-\frac{1}{p}}$	$n^{-1/2}$	n^p
Eq. (52)	1,2,3	$\sigma_j(\Sigma) \lesssim e^{-\beta j}$	$n^{-1/2}$	$\log^2 n$
Theorem 4	2,4,5,6	$\sigma_j(\Sigma) \lesssim j^{-\frac{1}{p}}$	$n^{-\frac{1}{1+p}}$	$n^{\frac{2p}{1+p}}$
Theorem 5	2,4,7	$\sigma_j(\Sigma) \lesssim j^{-\frac{1}{p}}$	$n^{-\min\{\frac{2r}{r+1}, \frac{r}{r(2-p-\theta+\theta p)+p}\}}$	$n^{\min\{2p, \frac{p(r+1)}{r(2-p-\theta+\theta p)+p}\}}$
RF Sun et al. (2018)	1,2,5,6	$\sigma_j(\Sigma) \lesssim j^{-\frac{1}{p}}$	$n^{-\frac{1}{2p+1}}$	$n^{\frac{2p}{2p+1}}$

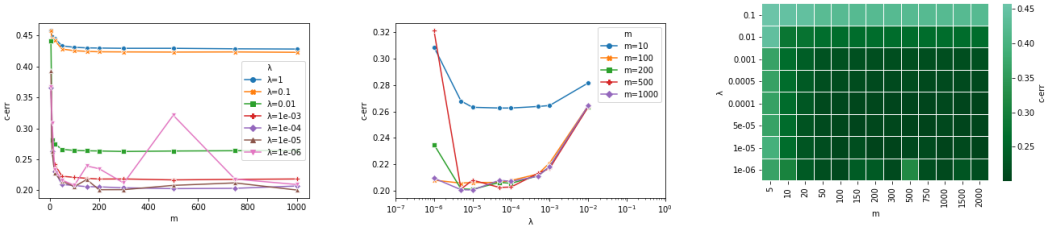

Figure 1: The graphs above are obtained from SUSY data set: on the left we show how c-err measure changes for different choices of λ parameter; in the central figure the focus is on the stability of the algorithm varying λ; on the right the combined behavior is presented with a heatmap.

Table 2: Architecture: single machine with AMD EPYC 7301 16-Core Processor and 256GB of RAM. For Nyström-Pegaos, ALS sampling has been used [Rudi et al. (2018)] and the results are reported as mean and standard deviation deriving from 5 independent runs of the algorithm. The columns of the table report classification error, training time and prediction time (in seconds).

	LinSVM	KSVM			Nyström-Pegasos			
Datasets	c-err	c-err	t train	t pred	c-err	t train	t pred	m
SUSY	28.1%	-	-	-	$20.0\% \pm 0.2\%$	608 ± 2	134 ± 4	2500
Mnist bin	12.4%	2.2%	1601	87	$2.2\% \pm 0.1\%$	1342 ± 5	491 ± 32	15000
Usps	16.5%	3.1%	4.4	1.0	$3.0\% \pm 0.1\%$	19.8 ± 0.1	7.3 ± 0.3	2500
Webspam	8.8%	1.1%	6044	473	$1.3\% \pm 0.1\%$	2440 ± 5	376 ± 18	11500
a9a	16.5%	15.0%	114	31	$15.1\% \pm 0.2\%$	29.3 ± 0.2	1.5 ± 0.1	800
CIFAR	31.5%	19.1%	6339	213	$19.2\% \pm 0.1\%$	2408 ± 14	820 ± 47	20500

ings can be achieved without incurring in any loss of accuracy. In this section, we complement our theoretical results investigating numerically the statistical and computational trade-offs in a relevant setting. More precisely, we report simple experiments in the context of kernel methods, considering Nyström techniques. In particular, we choose the hinge loss, hence SVM for classification. Keeping in mind Theorem 3 we expect we can match the performances of kernel-SVM using a Nyström approximation with only $m \ll n$ centers. The exact number depends on assumptions, such as the eigen-decay of the covariance operator, that might be hard to know in practice, so here we explore this empirically.

Nyström-Pegasos. Classic SVM implementations with hinge loss are based on considering a dual formulation and a quadratic programming problem [Joachims (1998)]. This is the case for example, for the LibSVM library [Chang and Lin (2011)] available on Scikit-learn [Pedregosa et al. (2011)]. We use this implementation for comparison, but find it convenient to combine the Nyström method to a primal solver akin to (6) (see Li et al. (2016); Hsieh et al. (2014) for the dual formulation). More precisely, we use Pegasos [Shalev-Shwartz et al. (2011)] which is based on a simple and easy to use stochastic subgradient iteration[3]. We consider a procedure in two steps. First, we compute the embed-

[3]Python implementation from https://github.com/ejlb/pegasos

ding discussed in Section 3. With kernels it takes the form $\tilde{x}_i = (K_m^\dagger)^{1/2}(K(x_i, \tilde{x}_1), \dots, K(x_i, \tilde{x}_m))^T$, where $K_m \in \mathbb{R}^{m\times m}$ with $(K_m)_{ij} = K(\tilde{x}_i, \tilde{x}_j)$. Second, we use Pegasos on the embedded data. As discussed in Section 3, the total cost is $O(nm^2C_K + nm \cdot \#iter)$ in time (here iter = epoch, i.e. one epoch equals n steps of stochastic subgradient) and $O(m^2)$ in memory (needed to compute the pseudo-inverse and embedding the data in batches of size m).

Datasets & set up (see Appendix J). We consider five datasets[4] of size $10^4 - 10^6$, challenging for standard SVMs. We use a Gaussian kernel, tuning width and regularization parameter as explained in appendix. We report classification error and for data sets with no fixed test set, we set apart 20% of the data.

Procedure Given the accuracy achieved by K-SVM algorithm, we increase the number of sampled Nyström points $m < n$ as long as also Nyström-Pegasos matches that result.

Results We compare with linear (used only as baseline) and K-SVM see Table 2. For all the datasets, the Nyström-Pegasos approach achieves comparable performances of K-SVM with much better time requirements (except for the small-size Usps). Moreover, note that K-SVM cannot be run on millions of points (SUSY), whereas Nyström-Pegasos is still fast and provides much better results than linear SVM. Further comparisons with state-of-art algorithms for SVM are left for a future work. Finally, in Figure 1 we illustrate the interplay between λ and m for the Nyström-Pegasos considering SUSY data set.

6 Conclusions

In this paper, we extended results for square loss [Rudi et al. (2015)] and self-concordant loss functions such as logistic loss [Marteau-Ferey et al. (2019b),Marteau-Ferey et al. (2019a)] to convex Lipschitz non-smooth loss functions such as hinge loss. The main idea is to save computations by solving the regularized ERM problem in a random subspace of the hypothesis space. We analysed the specific case of Nyström, where a data dependent subspace spanned by a random subset of the data is considered. In this setting we proved that under proper assumptions there is no statistical-computational tradeoff and our excess risk bounds can still match state-of-art results for SVM's [Steinwart and Christmann (2008)]. In particular, to achieve this behaviour we need sub-gaussianity of the input variables, a polynomial decay of the spectrum of the covariance operator and leverage scores sampling of the data. Theoretical guarantees have been proven both in the *realizable* case and, introducing the approximation error $\mathcal{A}(\lambda)$, when w_* does not exists. Numerical simulations using real data seem to support our theoretical findings while providing the desired computational savings. The obtained results can match the ones for random features [Sun et al. (2018)], but also allow to reach faster rates with more Nyström points while the others saturate. We leave for a longer version of the paper a unified analysis which includes square and logistic losses as special cases, and the consequences for classification.

[4]Datasets available from LIBSVM website http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/ and from Jose et al. (2013) http://manikvarma.org/code/LDKL/download.html#Jose13

Acknowledgments

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216, and the Italian Institute of Technology. We gratefully acknowledge the support of NVIDIA Corporation for the donation of the Titan Xp GPUs and the Tesla k40 GPU used for this research. Part of this work has been carried out at the Machine Learning Genoa (MaLGa) center, Università di Genova (IT).

L. R. acknowledges the financial support of the European Research Council (grant SLING 819789), the AFOSR projects FA9550-17-1-0390 and BAA-AFRL-AFOSR-2016-0007 (European Office of Aerospace Research and Development), and the EU H2020-MSCA-RISE project NoMADS - DLV-777826.

E. De Vito is a member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

References

Alaoui, A. and Mahoney, M. W. (2015). Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 775–783.

Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633.

Bach, F. (2013). Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209.

Bach, F. (2017). On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751.

Bartlett, P. L., Bousquet, O., Mendelson, S., et al. (2005). Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Bottou, L. and Bousquet, O. (2008). The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Calandriello, D., Lazaric, A., and Valko, M. (2017). Distributed adaptive sampling for kernel matrix approximation. In *Artificial Intelligence and Statistics*, pages 1421–1429. PMLR.

Calandriello, D. and Rosasco, L. (2018). Statistical and computational trade-offs in kernel k-means. In *Advances in Neural Information Processing Systems*, pages 9357–9367.

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

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27.

Cohen, M. B., Lee, Y. T., Musco, C., Musco, C., Peng, R., and Sidford, A. (2015). Uniform sampling for matrix approximation. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 181–190.

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

Drineas, P., Magdon-Ismail, M., Mahoney, M. W., and Woodruff, D. P. (2012). Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(Dec):3475–3506.

Drineas, P. and Mahoney, M. W. (2005). On the nyström method for approximating a gram matrix for improved kernel-based learning. *journal of machine learning research*, 6(Dec):2153–2175.

Giné, E. and Zinn, J. (1984). Some limit theorems for empirical processes. *The Annals of Probability*, pages 929–989.

Hsieh, C.-J., Si, S., and Dhillon, I. S. (2014). Fast prediction for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 3689–3697.

Joachims, T. (1998). Making large-scale svm learning practical. Technical report, Technical Report.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323.

Jose, C., Goyal, P., Aggrwal, P., and Varma, M. (2013). Local deep kernel learning for efficient non-linear svm prediction. In *International conference on machine learning*, pages 486–494.

Kakade, S. M., Sridharan, K., and Tewari, A. (2009). On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems 21*, pages 793–800.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033 of *École d'Été de Probabilités de Saint-Flour*. Springer-Verlag Berlin Heidelberg.

Koltchinskii, V. et al. (2006). Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656.

Koltchinskii, V. and Lounici, K. (2014). Concentration inequalities and moment bounds for sample covariance operators. *arXiv preprint arXiv:1405.2468*.

Kpotufe, S. and Sriperumbudur, B. K. (2019). Kernel sketching yields kernel jl. *arXiv preprint arXiv:1908.05818*.

Li, Z., Ton, J.-F., Oglic, D., and Sejdinovic, D. (2019). Towards a unified analysis of random fourier features. In *International Conference on Machine Learning*, pages 3905–3914. PMLR.

Li, Z., Yang, T., Zhang, L., and Jin, R. (2016). Fast and accurate refined nyström-based kernel svm. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Mahoney, M. W. (2011). Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224.

Marteau-Ferey, U., Bach, F., and Rudi, A. (2019a). Globally convergent newton methods for ill-conditioned generalized self-concordant losses. *arXiv preprint arXiv:1907.01771*.

Marteau-Ferey, U., Ostrovskii, D., Bach, F., and Rudi, A. (2019b). Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In *Conference on Learning Theory*, pages 2294–2340. PMLR.

Massart, P., Nédélec, É., et al. (2006). Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366.

Meir, R. and Zhang, T. (2003). Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4(Oct):839–860.

Mücke, N., Neu, G., and Rosasco, L. (2019). Beating sgd saturation with tail-averaging and minibatching. In *Advances in Neural Information Processing Systems*, pages 12568–12577.

Nesterov, Y. (2018). *Lectures on convex optimization*, volume 137. Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Rockafellar, R. T. (1970). *Convex analysis*. Number 28. Princeton university press.

Rudi, A., Calandriello, D., Carratino, L., and Rosasco, L. (2018). On fast leverage score sampling and optimal learning. In *Advances in Neural Information Processing Systems*, pages 5672–5682.

Rudi, A., Camoriano, R., and Rosasco, L. (2015). Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pages 1657–1665.

Rudi, A. and Rosasco, L. (2017). Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems 30*, pages 3215–3225.

Schmidt, M., Le Roux, N., and Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2010). Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670.

Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. (2011). Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30.

Shalev-Shwartz, S. and Zhang, T. (2013). Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599.

Smola, A. J. and Schölkopf, B. (2000). Sparse greedy matrix approximation for machine learning.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Steinwart, I., Hush, D., and Scovel, C. (2009). Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, pages 79–93.

Sun, Y., Gilbert, A., and Tewari, A. (2018). But how does it work in theory? linear svm with random features. In *Advances in Neural Information Processing Systems*, pages 3379–3388.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vito, E. D., Rosasco, L., Caponnetto, A., Giovannini, U. D., and Odone, F. (2005). Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6(May):883–904.

Wahba, G. (1990). *Spline models for observational data*, volume 59. Siam.

Williams, C. K. and Seeger, M. (2001). Using the nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688.

Woodruff, D. P. (2014). Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*.

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