# Regularized ERM on random subspaces

Andrea Della Vecchia [*], Ernesto De Vito [†], Jaouad Mourtada [‡], Lorenzo Rosasco [§]

December 6, 2022

**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 unified analysis requires developing new proofs, that use different technical tools, such as sub-gaussian inputs, to achieve fast rates. 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.

## 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 or 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, this includes Nyström approaches used for kernel methods [Smola and Schölkopf, 2000] and Gaussian processes [Williams and Seeger, 2001]. Recent works in statistical learning has focused on smooth loss functions [Rudi et al., 2015, Bach, 2013, Marteau-Ferey et al., 2019], whereas here we want to extend those analysis also to convex, Lipschitz but possibly non smooth losses.

In particular, if compared with previous 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 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

---

[*] andrea.dellavecchia@edu.unige.it

[†] ernesto.devito@unige.it

[‡] jaouad.mourtada@ensae.fr

[§] lorenzo.rosasco@unige.it

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] and sub-gaussian inputs [Koltchinskii and Lounici, 2014, Vershynin, 2010]. 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, while giving a unified theory including smooth and non-smooth losses. We do so studying the interplay between regularization, subspace size and the different parameters describing the hardness of the problem. 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. More than that, differently from quadratic loss, also a fast eigenvalues decay of the covariance operator is fundamental to have some computational savings. Related with this, local Rademacher complexities prove to be the right tool to fully exploit this latter eigen-decay assumption modifying also in the structure of the risk bound and leading to fast rates of convergence.
Once derived guarantees for both square and hinge losses, an interesting question is which one is better when solving a classification task. To have a fair comparison we will convert the previous bounds into the standard *classification risk*, i.e. the one derived by using the $0-1$ loss. Here we introduce the low noise condition that will play a key role. The final result shows that hinge loss can have always a better rate than square loss, no matter the choice of the parameter of eigen-decay, noise condition or approximation error that depend on the specific data. Nevertheless, if we match the rate achieved by the two to a fixed, reachable one, then hinge loss is *cheaper* than square loss, in terms of needed Nyström points, only when the problem is *hard* in some sense that will be clear in the following.
The rest of the paper is organized as follow. In Section 2, we introduce the setting and the main notation. In Section 3, we review the ERM approach and in Section 4 we introduce ERM on random subspaces and our setting. In Section 5, we present and discuss the main results and defer the proofs to the appendix. In Section 6, we extend our previous results to smooth losses. In Section 7 we pass to *classification risk* with $0-1$ loss and discuss the comparison between hinge and square losses. In Section 8, we collect some simple numerical results.

## 2 Setting and notations

ec: setting

We fix a real separable Hilbert space $\mathcal{H}$ with scalar product $\langle \cdot, \cdot \rangle$ and a Polish space $\mathcal{Y}$, *i.e* a separable complete metrizable topological space. Let $(X, Y)$ be a pair of random variables taking value in $\mathcal{H}$ and $\mathcal{Y}$, respectively and describing the input-output sampling procedure. We denote by $P$ the joint distribution of $(X, Y)$ defined on the Borel $\sigma$-algebra of $\mathcal{H} \times \mathcal{Y}$ and we choose a loss function $\ell : \mathcal{Y} \times \mathbb{R} \to [0, \infty]$ and set

$$L : \mathcal{H} \to [0, \infty) \qquad L(w) = \int_{\mathcal{H}\times\mathcal{Y}} \ell(y, \langle w, x \rangle) dP(x, y) = \mathbb{E}[\ell(Y, \langle w, X \rangle)]$$

to be the corresponding expected risk. Given $w \in \mathcal{H}$, $\ell(y, \langle w, x \rangle)$ can be viewed as the error made predicting $y$ with the linear function $f(x) = \langle w, x \rangle$, while $L(w)$ can be interpreted as the mean prediction error.
In this setting, we are interested in solving the problem

$$\inf_{w \in \mathcal{H}} L(w), \tag{1}$$

def of the p

when the distribution $P$ is only known through a training set $(x_i, y_i)_{i=1}^n$, which is a realization of $(X_1, Y_1)$, $\ldots$, $(X_n, Y_n)$, i.e. $n$ i.i.d. copies of $(X, Y)$. Since we only have some data, 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 $(x_i, y_i)_{i=1}^n$. In the following we are interested in characterizing its distribution for finite sample sizes and afterwards in discussing how approximate solutions can be obtained from data. To this aim we need to make some mathematical assumptions.

`sub-Gaussian` **Assumption 1.** *There exists $C > 0$ such that $X$ is a $C$-sub-Gaussian centered random vector.*

We recall that, according to [Koltchinskii and Lounici, 2014] a random vector $X$ taking value in a Hilbert space $\mathcal{H}$ is called $C$-sub-Gaussian if

$$\|\langle X, u\rangle\|_p \leqslant C\sqrt{p}\|\langle X, u\rangle\|_2 \qquad \forall u \in \mathcal{H}, p \geqslant 2 \tag{2}$$ `def: subgau`

where $\|\langle X, u\rangle\|_p^p = \mathbb{E}\left[\|\langle X, u\rangle\|^p\right]$. Note that (2) implies that for any vector $u \in \mathcal{H}$, the projection $\langle X, u\rangle$ is a real sub-Gaussian random variable [Vershynin, 2010], but this latter condition is not sufficient since the sub-Gaussian norm

$$\| \langle X, u\rangle \|_{\psi_2} = \sup_{p\geqslant 2} \frac{\|\langle X, u\rangle\|_p}{\sqrt{p}} \tag{3}$$ `eq:13`

should be bounded from above by the $L_2$-norm $\|\langle X, u\rangle\|_2$. In particular, we stress that, in general, bounded random vectors in $\mathcal{H}$ are not sub-Gaussian.
Under the above conditions, $\mathbb{E}[\|X\|^2]$ is finite, so that the (non-centered) covariance operator

$$\Sigma : \mathcal{H} \to \mathcal{H} \qquad \Sigma = \mathbb{E}[X \otimes X]$$

is a trace-class positive operator, and

$$r_\Sigma = \sqrt{\frac{\mathrm{Tr}\Sigma}{\|\Sigma\|}} \tag{4}$$ `eq:1`

denotes the effective rank of $\Sigma$, where $\mathrm{Tr}\,\Sigma = \mathbb{E}[\|X\|^2]$ is the trace of $\Sigma$. Further, it is also useful to introduce the so-called effective dimension [Zhang, 2005, Caponnetto and De Vito, 2007, Rudi et al., 2015]. We define for $\alpha > 0$

$$d_\alpha = \mathrm{Tr}((\Sigma + \alpha I)^{-1}\Sigma) = \sum_j \frac{\sigma_j}{\sigma_j + \alpha} \tag{5}$$

where $(\sigma_j)_j$ are the strictly positive eigenvalues of $\Sigma$, with eigenvalues counted with respect to their multiplicity and ordered in a non-increasing way, and $(u_j)$ is the corresponding family of eigenvectors. Note that $d_\alpha$ is finite since $\Sigma$ is trace class.

Next assumption is on the loss function.

`ass:loss` **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*

$$|\ell(y, a) - \ell(y, a')| \leq G|a - a'| \quad \textit{and} \quad \ell_0 = \sup_{y\in\mathcal{Y}} \ell(y, 0). \tag{6}$$ `eq:5`

*for all $y \in \mathcal{Y}$ and $a, a' \in \mathbb{R}$.*

Under the above condition, the expected risk $L(w)$ is finite, convex and Lipschitz. We now provide some relevant examples. The classical linear regression problem corresponds to the choice $\mathcal{H} = \mathbb{R}^d$ and $\mathcal{Y}$. Another example is provided by kernel methods [Steinwart and Christmann, 2008].

`ex kernel` **Example 1.** The input variable $X$ takes value in an abstract measurable set $\mathcal{X}$. We fix a reproducing kernel Hilbert space on $\mathcal{X}$ with (measurable) reproducing kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. By mapping the inputs from $\mathcal{X}$ to $\mathcal{H}$ through the feature map

$$\mathcal{H} \ni x \mapsto K(\cdot, x) = K_x \in \mathcal{H},$$

we can always identify $X$ with $K_X$, which is a random variable taking value in $\mathcal{H}$.

We now provide some example of loss functions.

**Example 2.** The main examples are

(a) hinge loss:

$$\ell(y,a) = |1-ya|_+ = \max\{0, 1-ya\} \qquad \mathcal{Y} = \{-1,1\} \tag{7}$$

eq:9

which is convex, but non-differentiable with $G = 1$ and $\ell_0 = 1$;

(b) logistic loss

$$\ell(y,a) = \log(1+e^{-ya}) \qquad \mathcal{Y} = \{-1,1\} \tag{8}$$

eq:10

which is convex and differentiable with $G = 1$ and $\ell_0 = \log 2$;

(c) square loss

$$\ell(y,a) = (y-a)^2 \qquad \mathcal{Y} \subseteq [-M, M], \tag{9}$$

eq:11

which is convex and differentiable with $G = 2M$ and $\ell_0 = 4M^2$.

For classification task $\mathcal{Y} = \{-1,1\}$, a natural loss function is given by the $0-1$ loss

$$\ell_{0-1}(y,a) := \mathbb{1}_{(-\infty,0]}(y \ \mathrm{sign}\, a),$$

which is not convex. However, it is known that a bound for the excess risk leads directly to a bound on the classification risk [Bartlett et al., 2006]. In Section 7 we recall some standard result and we apply them to our setting.

## 2.1 Notations

For reader's convenience we collect the main notation we introduced in the paper. We denote with the "hat", e.g. $\widehat{\cdot}$, random quantities depending on the data. Given a linear operator $A$ we denote by $A^\top$ its adjoint (transpose for matrices). For any $n \in \mathbb{N}$, we denote by $\langle \cdot, \cdot \rangle_n$ , $\|\cdot\|_n$ the inner product and norm in $\mathbb{R}^n$. Given two quantities $a, b$ (depending on some parameters), the notation $a \lesssim b$, or $a = O(b)$ means that there exists a constant $C$ such that $a \leqslant Cb$. We denote by $P_X$ the marginal distribution of $X$ and by $P(\cdot|x)$ is the conditional distribution of $Y$ given $X = x$. The conditional probability is well-defined since $\mathcal{H}$ is separable and $\mathcal{Y}$ is a Polish space [Steinwart and Christmann, 2008]. Table 2.1 summarizes main notation.

tab:1

Table 1: Definition of the main quantities used in the paper

| | Definition |
|---|---|
| $L(w)$ | $\int_{\mathcal{H}\times\mathcal{Y}} \ell(y, \langle w, x\rangle) dP(x,y)$ |
| $L_\lambda(w)$ | $L(w) + \lambda\|w\|^2$ |
| $\widehat{L}(w)$ | $n^{-1}\sum_{i=1}^n \ell(y_i, \langle w, x_i\rangle)$ |
| $\widehat{L}_\lambda(w)$ | $\widehat{L}(w) + \lambda\|w\|^2$ |
| $w_*$ | $\arg\min_{w\in\mathcal{H}} L(w)$ |
| $w_\lambda$ | $\arg\min_{w\in\mathcal{H}} L_\lambda(w)$ |
| $\widehat{w}_\lambda$ | $\arg\min_{w\in\mathcal{H}} \widehat{L}_\lambda(w)$ |
| $\beta_{\lambda,\mathcal{B}}$ | $\arg\min_{\beta\in\mathcal{B}} L_\lambda(\beta)$ |
| $\widehat{\beta}_{\lambda,\mathcal{B}}$ | $\arg\min_{\beta\in\mathcal{B}} \widehat{L}_\lambda(\beta)$ |
| $f_*(x)$ | $\arg\min_{a\in\mathbb{R}} \int_{\mathcal{Y}} \ell(y,a) dP(y\|x)$ |
| $\mathcal{B}_m$ | $\mathcal{B}_m = \mathrm{span}\{\widetilde{x}_1, \dots, \widetilde{x}_m\}$ |
| $\mathcal{P}_\mathcal{B}$ | projection operator onto $\mathcal{B}$ |
| $\mathcal{P}_m$ | projection operator onto $\mathcal{B}_m$ |
| $\mathfrak{R}(\cdot)$ | population Rademacher Complexity |
| $\widehat{\mathfrak{R}}(\cdot)$ | empirical Rademacher Complexity |
| $e_n$ | (dyadic) entropy numbers $e_n = \varepsilon_{2^{n-1}}$ |

# 3 Empirical risk minimization: a review

`sec:ERM`

A classical 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 the (regularized) empirical risk minimization (ERM) based on the solution of the problem,

$$\min_{w \in \mathcal{H}} \widehat{L}_\lambda(w), \qquad \widehat{L}_\lambda(w) = \widehat{L}(w) + \lambda \|w\|^2, \tag{10}$$

`perm`

where $\lambda > 0$ is a positive regularization parameter. Since $\widehat{L}_\lambda : \mathcal{H} \to \mathbb{R}$ is continuous and strongly convex, there exists a unique minimizer $\widehat{w}_\lambda$ and, by the representer theorem [Wahba, 1990, Schölkopf et al., 2001], there exists $c = \widehat{c}_\lambda \in \mathbb{R}^n$ such that

$$\widehat{w}_\lambda = \widehat{X}^\top c \in \operatorname{span}\{x_1, \dots, x_n\}, \tag{11}$$

`repre`

where $\widehat{X} : \mathcal{H} \to \mathbb{R}^n$ denotes the *data matrix*

$$(\widehat{X} w)_i = \langle w, x_i \rangle \qquad i = 1, \dots, n, \quad w \in \mathcal{H}.$$

The explicit form of the coefficient vector $c$ depends on the considered loss function. In Section 3.1 we briefly recall some possible approach to compute $c$, whereas in Section 3.2 we analyze its statistical properties.

`kernel span`

**Example 3** (Representer theorem for kernel machines)**.** In the context of kernel methods, see Example 1, the above discussion, and in particular (11) 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 (11) then simply states that there exists constants $c_i$ such that the solution of the regularized ERM can be written as $\widehat{f}_\lambda(x) = \sum_{i=1}^n K(x, x_i) c_i \in \operatorname{span}\{K_{x_1}, \dots, K_{x_n}\}$.

## 3.1 Computational aspects

`sec:comp`

Minimizing (10) can be solved in many ways and we provide below some basic considerations. If $\mathcal{H}$ is finite dimensional, iterative via gradient methods can be used. For example, the subgradient method [Boyd and Vandenberghe, 2004] applied to (10) gives, for some suitable $w_0$ and step-size sequence $(\eta_t)_t$,

$$w_{t+1} = w_t - \eta_t \left( \frac{1}{n} \sum_{i=1}^{n} y_i x_i g_i(w_t) + 2\lambda w_t \right), \tag{12}$$

`subgsvm`

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 minibatching 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 (12) satisfies $w_t = \widehat{X}^\top c_{t+1}$, where

$$c_{t+1} = c_t - \eta_t \left( \frac{1}{n} \sum_{i=1}^{n} y_i e_i g_i(\widehat{X}^\top c_t) + 2\lambda c_t \right), \tag{13}$$

`subgsvm1`

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 (10) 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 (13) we can take $g_i(w) = -\mathbb{1}_{[y_i \langle w, x_i\rangle \leq 1]}$.

`ec:statistic`

## 3.2 Statistical analysis

In this section we summarizes the main statistical properties of the regularized ERM under the sub-Gaussian hypothesis. Thm. 2 will show that with high probability

$$L(\widehat{w}_\lambda) - \inf_{w\in\mathcal{H}} L(w) \lesssim \frac{1}{\lambda n} + \lambda \left\|w_*\right\|^2 ,$$

provided that the best in model $w_* \in \mathcal{H}$ exists, see Assumption 3. With the choice $\lambda \asymp \sqrt{1/n}$ it holds that

$$L(\widehat{w}_\lambda) - \inf_{w\in\mathcal{H}} L(w) = O(\sqrt{1/n}), \tag{14}$$

`eq:38`

which provides a benchmark for our results. More generally, the following theorem provides a finite sample bound on the excess risk of $\widehat{w}_\lambda$ without assuming the existence of $w^*$. To this aim, we introduce the approximation error

$$\begin{aligned} \mathcal{A}(\lambda) &= \inf_{w\in\mathcal{H}} [L(w) + \lambda\|w\|^2] - \inf_{w\in\mathcal{H}} L(w) \\ &= L(w_\lambda) + \lambda\|w_\lambda\|^2 - \inf_{w\in\mathcal{H}} L(w) \,. \end{aligned} \tag{15}$$

`eq:24`

`erm-standard`

**Theorem 1.** *Under Assumptions 1 and 2, fix $\lambda > 0$ and $0 < \delta < 1$. Then, with probability at least $1 - \delta$,*

$$\begin{aligned} L(\widehat{w}_\lambda) - \inf_{w\in\mathcal{H}} L(w) <&2\mathcal{A}(\lambda) + \frac{D^2G^2C^2\|\Sigma\|(K^2 + (r_\Sigma + \sqrt{\log(1/\delta)})^2)}{4\lambda n} + \\ &+ \frac{DGCK\|\Sigma\|^{\frac{1}{2}} + D\ell_0(K - r_\Sigma + \sqrt{\log(1/\delta)})}{\sqrt{n}}. \end{aligned} \tag{16}$$

`eq:excessris`

*where $C$ and $G$ are the constants defined respectively in (2) and (6), $D$ is an absolute numerical constant and*

$$K = K_{\lambda,\delta} = r_\Sigma + \sqrt{\log(1 + \log_2(3 + \ell_0/\lambda)) + \log(1/\delta)} = O(\sqrt{\log\log(3 + \ell_0/\lambda) + \log(1/\delta)}).$$

The theorem can be easily extended to non-centered sub-Gaussian variables. Notice that the same result is well known for bounded random variables, by specializing more refined analysis, see for example [Steinwart and Christmann, 2008, Shalev-Shwartz et al., 2010]. As regards the sub-Gaussian case, we are not aware of a previous reference. In Appendix A we provide a simple self-contained proof, which holds true also for the bounded case [Della Vecchia et al., 2021]. It is based on the fact that the excess risk bound for regularized ERM arises from a trade-off between an estimation and an approximation term. 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 $\lambda$. 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].

Bound (55) shows that the learning rate depends on some a-priori assumption of the distribution, allowing to controll the approximation error $\mathcal{A}(\lambda)$. The simplest assumption is that the best in the model exists.

`ass:best`

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

Under this condition, we have the following result, claimed at the beginning of the section:

ed-full-space **Theorem 2.** *Under Assumption 1, 2, and 3, take* $\lambda > 0$ *and* $0 < \delta < 1$*, then with probability at least* $1 - \delta$*:*

$$L(\widehat{w}_\lambda) - L(w_*) < \lambda\|w_*\|^2 + \frac{D^2G^2C^2K^2\|\Sigma\|}{4\lambda n} + \frac{DGCK\|\Sigma\|^{\frac{1}{2}} + D\ell_0(K - r_\Sigma + \sqrt{\log(8/\delta)})}{\sqrt{n}} + \\ + \frac{DGC\|\Sigma\|^{\frac{1}{2}}\,\|w_*\|\left(r_\Sigma + \sqrt{\log(8/\delta)}\right)}{\sqrt{n}}. \tag{17}$$

eq:excessris

*Hence, let* $\lambda = \lambda_n \asymp (DGC\,\|\Sigma\|^{1/2}/\|w_*\|)\sqrt{\log(1/\delta)/n}$ *with high probability:*

$$L(\widehat{w}_{\lambda_n}) - L(w_*) = O(\|w_*\|\sqrt{\log(1/\delta)/n}), \tag{18}$$

eq:17

*up to a* $\log\log n$ *terms.*

As above the proof is given in Appendix A.

*Remark* 1. Note that for all $w \in \mathcal{H}$ with $\|w\| \leqslant R$,

$$\mathcal{A}(\lambda) \leqslant L(w) + \lambda\|w\|^2 - \inf_{\mathcal{H}} L \leqslant L(w) - \inf_{\mathcal{H}} L + \lambda R^2$$

hence $\mathcal{A}(\lambda) \leqslant \inf_{\|w\|\leqslant R} L(w) - \inf_{\mathcal{H}} L + \lambda R^2$ and

$$L(\widehat{w}_\lambda) - \inf_{w\in\mathcal{H}} L(w) < 2\Big(\inf_{\|w\|\leqslant R} L(w) - \inf_{\mathcal{H}} L\Big) + 2\lambda R^2 + \frac{D^2G^2C^2\|\Sigma\|(K^2 + (r_\Sigma + \sqrt{\log(8/\delta)})^2)}{4\lambda n} \\ + \frac{DGCK\|\Sigma\|^{\frac{1}{2}} + D\ell_0(K - r_\Sigma) + D\ell_0\sqrt{\log(8/\delta)}}{\sqrt{n}},$$

Letting $\lambda \asymp 1/(R\sqrt{n})$, this gives $L(\widehat{w}_\lambda) - \inf_{w\in\mathcal{H}} L(w) \leqslant 2(\inf_{\|w\|\leqslant R} L(w) - \inf_{\mathcal{H}} L) + O(R/\sqrt{n})$ with high probability.

dom subspace

## 4 ERM on random subspaces

As explained in the introduction, though the ERM algoritm $\widehat{w}_{\lambda_n}$ achieves optimal rates, from a computational point of view is very expensive with large datasets. To overcome this issue, we consider a variant of ERM based on considering a subspace $\mathcal{B} \subset \mathcal{H}$ and the corresponding regularized ERM problem,

$$\min_{\beta\in\mathcal{B}} \widehat{L}_\lambda(\beta) \tag{19}$$

sperm

with $\widehat{\beta}_\lambda$ the unique minimizer. As clear from (11), choosing $\mathcal{B} = \mathcal{H}_n = \operatorname{span}\{x_1, \dots, x_n\}$ is not a restriction and yields the same solution as considering (10). From this observation a natural choice is to consider for $m \leq n$,

$$\mathcal{B}_m = \operatorname{span}\{\widetilde{x}_1, \dots, \widetilde{x}_m\} \tag{20}$$

randspace

with $\{\widetilde{x}_1, \dots, \widetilde{x}_m\} \subset \{x_1, \dots, x_n\}$ a subset of the input points, called the Nyström points. We denote by $\mathcal{P}_m = \mathcal{P}_{\mathcal{B}_m}$ the corresponding projection and by $\widehat{\beta}_{\lambda,m}$ the unique minimizer of $\widehat{L}_\lambda$ on $\mathcal{B}_m$, *i.e.*

$$\widehat{\beta}_{\lambda,m} = \operatorname*{argmin}_{\beta\in\mathcal{B}_m} \widehat{L}_\lambda(\beta). \tag{21}$$

eq:37

We now focus on the computational benefits of considering ERM on random subspaces and we analyze the corresponding statistical properties, whereas the statistical analysis is given in Section 4.2. A more advanced analysis is provided in Section 5 to obtain fast rates.

### 4.1 Computational aspects

The choice of $\mathcal{B}_m$ as in (20) allows to improve computations with respect to (11). 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 (19) 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 (10), considering the reparameterization $a = (\widetilde{X}\widetilde{X}^\top)^{1/2} b$ which leads to,

$$\min_{a \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^n \ell\left(y_i, \langle a, \mathbb{x}_i \rangle_m\right) + \lambda \|a\|_m^2, \tag{22}$$ embedsvm

where for all $i = 1, \ldots, n$, we defined the embedding $x_i \mapsto \mathbb{x}_i = ((\widetilde{X}\widetilde{X}^\top)^{1/2})^\dagger \widetilde{X} x_i$ 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 (22) 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 (13) 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.

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

### 4.2 Statistical analysis

tics-Nystrom

In this paper, we consider approximate leverage scores sampling procedures of the Nyström points. The reason is that, we wil show that under a suitable $p$-polynomial (or exponential) decay condition on the spectrum, see (28),

$$L(\widehat{\beta}_{\lambda,m}) - L(w_*) \lesssim \frac{\sqrt{\log(1/\delta)}}{\sqrt{n}}.$$

provided that the best in model $w_* \in \mathcal{H}$ exists, see Assumption 3, and, up to log terms,

$$\lambda \asymp \frac{1}{\sqrt{n}}, \qquad m \gtrsim n^p.$$

By comparing with the benchmark (14), we get the same convergence rate up to a log factor, but the complexity of the algorithm is dramatically reduced, for example if $p = 1/2$ we only need $m \simeq \sqrt{n}$ Nyström points. A similar result can be obtained for exponential decay with $m \simeq \log^2 n$ Nyström points. Finally, we observe that under the above decay conditions on the spectrum of $\Sigma$ classical ERM algorithm achieves fast rates. In Section 5 we will show that also Nyström sub-sampling has fast rates, but this requires a more advanced analysis.

We now describe ALS sampling procedure. With this method we sample according to the leverages scores [Drineas et al., 2012]:

$$l_i(\alpha) = \left\langle x_i, (\widehat{X}\widehat{X}^\top x + \alpha I n)^{-1} x_i \right\rangle \qquad i = 1, \ldots, n \tag{23}$$ eq:3

where $\alpha > 0$. Since in practice the leverage scores $l_i(\alpha)$ defined by (23) are onerous to compute, approximations $(\hat{l}_i(\alpha))_{i=1}^n$ have been considered [Drineas et al., 2012, Cohen et al., 2015, Alaoui and Mahoney, 2015]. In particular, in the following we are interested in suitable approximations defined as follows, see [Rudi et al., 2018] and references therein.

x_lev_scores

**Definition 1** (Approximate leverage scores sampling (ALS))**.** Let $(l_i(\alpha))_{i=1}^n$ be the leverage scores given by (23). Given $\alpha_0 > 0$ and $T \geqslant 1$, we say that a family $(\hat{l}_i(\alpha))_{i=1}^n$ is $(T, \alpha_0)$-approximate leverage scores with confidence $\delta \in (0,1)$ if

$$\frac{1}{T} l_i(\alpha) \leqslant \hat{l}_i(\alpha) \leqslant T l_i(\alpha), \qquad \forall i \in \{1, \dots, n\}, \;\; \alpha \geqslant \alpha_0 \tag{24}$$

with probability at least $1 - \delta$. Under this condition, the approximate leverage scores (ALS) sampling selects the Nyström points $\{\tilde{x}_1, \dots, \tilde{x}_m\}$ from the training set $\{x_1, \dots, x_n\}$ independently with replacement and with probability $Q_\alpha(i) = \hat{l}_i(\alpha) / \sum_j \hat{l}_j(\alpha)$.

We can state our first result. We recall the Nyström points are sampled according to ALS, see Definition 1.

thm:1

**Theorem 3.** *Under Assumption 1, 2 and 3, fix $\alpha, \lambda, \delta > 0$. With probability at least $1 - \delta$:*

$$L(\widehat{\beta}_{\lambda,m}) - L(w_*) \lesssim \frac{\log(1/\delta)}{\lambda n} + \frac{\|w_*\| \sqrt{\log(1/\delta)}}{\sqrt{n}} + \sqrt{\alpha} \|w_*\| + \lambda \|w_*\|^2 \tag{25}$$

reg pb f_H m

*up to $\log(\log(1/\lambda))$ terms and provided that $n \gtrsim d_\alpha \vee \log(1/\delta)$ and $m \gtrsim d_\alpha \log(\frac{2n}{\delta})$.*

The proof of Theorem 3 with explicit constants is given in Appendix B, whereas here we add some comments. Notice that

$$d_\alpha = \int \langle w, (\Sigma + \alpha I)^{-1} w \rangle \, dP_X(w) \leqslant \int \|w\|^2 \left\| (\Sigma + \alpha I)^{-1} \right\| dP_X(w) \leqslant \alpha^{-1} \mathbb{E}[\|X\|^2] \lesssim \alpha^{-1}, \tag{26}$$

eq:6bis

where we used the fact that the second moment of a sub-Gaussian variable is finite.
Using the above bound, we get that, up to log terms, with high probability

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

provided that $m \gtrsim \alpha^{-1}$. With the choice

$$\lambda_n \asymp \frac{1}{\|w_*\| \sqrt{n}}, \qquad \alpha \asymp 1/n$$

we get that with high probability

$$L(\widehat{\beta}_{\lambda_n,m}) - L(w_*) \lesssim \frac{\|w_*\| \sqrt{\log(1/\delta)}}{\sqrt{n}} \tag{27}$$

up to log factors in $n$ and with $m \gtrsim n$.
Despite of the fact that the rate is optimal (up to the logarithmic term), the required number of subsampled points is $m \gtrsim n$, so that the procedure is not effective. However, the following proposition shows that under a fast decay of the spectrum of the covariance operator $\Sigma$, the ALS method becomes computationally efficient. We assume one of the following two conditions:

a) polinomial decay: there exists $p \in (0,1)$ such that

$$\sigma_j \lesssim j^{-\frac{1}{p}} \tag{28}$$

eq:2

b) exponential decay: there exists $\beta > 0$ such that

$$\sigma_j \lesssim e^{-\beta j}. \tag{29}$$

eq:19

Under the above condition, we have the following result.

: constr pb

**Theorem 4.** *Under the assumptions of Theorem 3, fix $\delta > 0$, with probability at least $1 - \delta$:*

*(a) for the polynomial decay* (28)

$$L(\widehat{\beta}_{\lambda,m}) - L(w_*) \lesssim \frac{\log(1/\delta)}{\lambda n} + \frac{\|w_*\|\sqrt{\log(1/\delta)}}{\sqrt{n}} + \sqrt{\alpha}\|w_*\| + \lambda\|w_*\|^2 \tag{30}$$

*and, with the choice*

$$\lambda_n \asymp \frac{\|w_*\|\sqrt{\log(1/\delta)}}{\sqrt{n}}, \qquad \alpha_n \asymp \frac{\log(1/\delta)}{n}, \qquad m \gtrsim n^p,$$

*it holds that*

$$L(\widehat{\beta}_{\lambda_n,m}) - L(w_*) \lesssim \frac{\|w_*\|\sqrt{\log(1/\delta)}}{\sqrt{n}} \tag{31}$$

`eq:12`

*(b) for exponential decay* (29)

$$L(\widehat{\beta}_{\lambda,m}) - L(w_*) \lesssim \frac{\log(1/\delta)}{\lambda n} + \frac{\|w_*\|\sqrt{\log(1/\delta)}}{\sqrt{n}} + \sqrt{\alpha}\|w_*\| + \lambda\|w_*\|^2 \tag{32}$$

`eq: thm 3 e`

*and with the choice*

$$\lambda_n \asymp \frac{\|w_*\|\sqrt{\log(1/\delta)}}{\sqrt{n}}, \qquad \alpha_n \asymp \frac{\log(1/\delta)}{n}, \qquad m \gtrsim \log^2 n,$$

*it holds that*

$$L(\widehat{\beta}_{\lambda_n,m}) - L(w_*) \lesssim \frac{\|w_*\|\sqrt{\log(1/\delta)}}{\sqrt{n}}.$$

The proof of the above result is given in Appendix B. Theorem 4 is already known for square loss [Rudi et al., 2015] and for smooth loss functions [Marteau-Ferey et al., 2019] under the assumption that the input $X$ is bounded. The cited references deal also the case of uniform sampling of Nyström points. Notice that, compared with the analogous theorem for square loss, our bound on the number of Nyström points is worse than the bound given in [Rudi et al., 2015]. In Section 6, by exploiting in the projection term the square in the definition of the quadratic loss, we obtain the right estimate of Nyström points matching the result in [Rudi et al., 2015].
Theorem 4 shows that for an arbitrary convex, possibly non-smooth, loss function, 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 of 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, considering $p = 1/2$, then the choice $m \simeq \sqrt{n}$ is enough. Notice also that other decays, e.g. exponential, can 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. Compared to random features, the proof techniques presents 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 $d_\alpha \asymp d_{\alpha,\infty}$, where $d_{\alpha,\infty} := \sup_{w \in \operatorname{supp}(P_X)} \langle w, (\Sigma + \alpha I)^{-1} w \rangle$, see [Rudi et al., 2015]. 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.

## 4.3 Further choices

Following [Rudi et al., 2015], other choices of $\mathcal{B} \subseteq \mathcal{H}$ 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 (22) replacing $\widetilde{X}$ with the matrix $Z$ with rows $z_1, \dots, z_q$. We leave this discussion for future work. We simply state the following result where

$$\mu_{\mathcal{B}} = \left\| \Sigma^{1/2}(I - \mathcal{P}) \right\|, \tag{33}$$

`proj`

and $\mathcal{P}$ is the projection onto $\mathcal{B}$.

`firstmain` **Theorem 5.** *Choose $\mathcal{B} \subseteq \mathcal{H}$. Under Assumptions 1, 2, 3, fix $\lambda > 0$ and $0 < \delta < 1$, with probability at least $1 - \delta$:*

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

Compared to Theorem 2, 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., 2019]. Roughly speaking, it captures how well the subspace $\mathcal{B}$ is adapted to the problem.

# 5 Fast rates

`:statistical`

In this section we prove the Nyström algorithm achieves fast rates under a Bernstein condition on the loss function, see Assumption 7, which is quite standard in order to have fast rates for regularized ERM [Steinwart and Christmann, 2008, Bartlett et al., 2005]. To state our results, we recall some definitions and basic facts, see [Steinwart and Christmann, 2008, Chapter 6] for a full account.
Given a threshold parameter $M > 0$, for any $a \in \mathbb{R}$, $a^{cl}$ denotes the clipped value of $a$ at $\pm M$

$$a^{cl} = -M \quad \text{if } \ a \leqslant -M, \qquad a^{cl} = a \quad \text{if } \ a \in [-M,M], \qquad a^{cl} = M \quad \text{if } \ a \geqslant M.$$

We say that the loss function $\ell$ can be *clipped* at $M > 0$ if for all $y \in \mathcal{Y}, a \in \mathbb{R}$,

$$\ell(y, a^{cl}) \leqslant \ell(y, a), \tag{34}$$ `eq: clippin`

For convex loss functions, as we consider in this paper, the above definition is equivalent to the fact that for all $y \in \mathcal{Y}$, there exists $a_y \in [-M, M]$ such that

$$\ell(y, a_y) = \min_{a \in \mathbb{R}} \ell(y, a),$$

see [Steinwart and Christmann, 2008, Lemma 2.23]. Furthermore, Aumann's measurable selection principle [Steinwart and Christmann, 2008, Lemma A.3.18] implies that there exists a measurable map $\varphi : \mathcal{Y} \to \mathbb{R}$ such that

$$\ell(y, \varphi(y)) = \min_{a \in \mathbb{R}} \ell(y, a), \qquad |\varphi(y)| \leqslant M.$$

and we set

$$f_*(x) = \int_{\mathcal{Y}} \ell(y, \varphi(x)) dP(y|x) \tag{35}$$ `target_def`

for $P_X$-almost all $x \in \mathcal{H}$. The function $f^*$ is the target function since

$$L(f_*) = \inf_f L(f),$$

where the infimum is taken over all the measurable functions $f : \mathcal{H} \to \mathbb{R}$. It easy to check that hinge loss and square loss with bounded outputs can be clipped, whereas the logistic loss can not be clipped, however we are able to include also this latter case in Section 6.2. We also introduce the following notation, for all $w \in \mathcal{H}$, we set

$$w^{cl} : \mathcal{H} \to \mathbb{R} \qquad w^{cl}(x) = \langle w, x \rangle^{cl} .$$

In the following we assume the conditions below.

`ass:clipped` **Assumption 4** (Clippability)**.** *There exists $M > 0$ such that the loss function can be clipped at $M$.*

`ss:universal` **Assumption 5** (Universality)**.**

$$\inf_{w \in \mathcal{H}} L(w) = L(f_*). \tag{36}$$ `universal`

Recalling that the target function $f_*$ is the minimizer of the expected error over all possible functions $f$, condition (36) means that $f_*$ can be arbitrarily well approximated by a linear function $\langle w, x\rangle$ for some $w \in \mathcal{H}$. For the square loss, this condition is equivalent to the fact that $\mathcal{H}$ is dense in $L^2(\mathcal{H}, P_X)$ and, in the context of kernel methods, see Example 1 it is satisfied by universal kernels [Steinwart and Christmann, 2008]. Condition (36) may be relaxed at the cost of an additional approximation term, but the analysis is lengthier and it won't be discussed in here. A sufficient stronger condition is provided by the well specified case.

`target`

**Assumption 6** (Well specified model)**.** *There exists $w_* \in \mathcal{H}$ such that*

$$f_*(x) = \langle w_*, x\rangle$$

*for $P_X$-almost $x \in \mathcal{H}$.*

We further assume the following condition.

`ass:berstein`

**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:*

$$\ell(Y, \langle w, X\rangle^{cl}) \leqslant B, \tag{37}$$

`supremum bou`

$$\mathbb{E}[\ell(Y, \langle w, X\rangle^{cl}) - \ell(Y, f_*(X))]^2 \leqslant V(\mathbb{E}[\ell(Y, \langle w, X\rangle^{cl}) - \ell(Y, f_*(X))])^\theta \tag{38}$$

`variance bou`

$$\mathbb{E}[\ell(Y, \langle w, X\rangle) - \ell(Y, f_*(X))]^2 \leqslant V(\mathbb{E}[\ell(Y, \langle w, X\rangle) - \ell(Y, f_*(X))])^\theta \tag{39}$$

`variance bou`

Condition (37) is called supremum bound and it is satisfied by almost all loss functions. Condition (38) is called the variance bound and the optimal exponent corresponds to the choice $\theta = 1$. For the square loss with bounded output, the variance bound always holds true with $\theta = 1$, see [Steinwart and Christmann, 2008, Example 7.3] . For other loss functions the above condition is hard to verify for all distributions. For example for classification, the variance bound is implied by margin conditions, and the parameter $\theta$ characterizes how easy or hard the classification problem is [Steinwart and Christmann, 2008]. With respect to [Steinwart and Christmann, 2008], condition (39) is a technical one that we need in the proof.
To state our result, we introduce the approximation error,

$$\mathcal{A}(\lambda) = \min_{w \in \mathcal{H}} \left(L(w) + \lambda \|w\|^2\right) - \inf_{w \in \mathcal{H}} L(w). \tag{40}$$

`ass:apprx`

Note that, if $w_*$ exists, then $\mathcal{A}(\lambda) \leqslant \lambda \|w_*\|^2$. More generally, the approximation error decreases with $\lambda$ and learning rates can be derived assuming a suitable decay. The following theorem provides fast rates for Nyström algorithm, where we recall the Nyström points are sampled according to ALS, see Definition 1.

`te A(lambda)`

**Theorem 6.** *Under Assumptions 1, 2, 4, 7, let fix $0 < \delta < 1$, then, with probability at least $1 - 2\delta$:*

*(a) for the polynomial decay condition* (28)

$$L(\widehat{\beta}^{cl}_{\lambda,m}) - L(f_*) \lesssim \left(\frac{1}{\lambda^p n}\right)^{\frac{1}{2-p-\theta+\theta p}} + \sqrt{\frac{\alpha \mathcal{A}(\lambda)}{\lambda}} + \left(\frac{\log(3/\delta)}{n}\right)^{\frac{1}{2-\theta}} + \frac{\log(3/\delta)}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} + \mathcal{A}(\lambda) \tag{41}$$

`fast rate A(`

*provided that*

$$\alpha \gtrsim n^{-1/p}, \qquad n \gtrsim d_\alpha \vee \log(1/\delta), \qquad m \gtrsim d_\alpha \log(\frac{2n}{\delta}),$$

*(b) for the exponential decay* (29)

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

*provided that*

$$\alpha \gtrsim e^{-n}, \qquad n \gtrsim d_\alpha \vee \log(1/\delta), \qquad m \gtrsim d_\alpha \log(\frac{2n}{\delta}).$$

The proof of Theorem 6 is given in Appendix C. Let us comment the above result in different settings.

### 5.1 Polynomial decay of $\Sigma$

In this section we assume the polynomial decay (28) of the spectrum of $\Sigma$. By omitting numerical constants, logarithmic and higher order terms, Theorem 6 implies that with high probability

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

To have an explicit rate, we further assume that there exists $r \in (0,1]$ such that

$$\mathcal{A}(\lambda) \lesssim \lambda^r.$$

Under this condition, with the choice

$$\begin{aligned}
\lambda_n &\asymp n^{-\min\{\frac{2}{r+1}, \frac{1}{r(2-p-\theta+\theta p)+p}\}} \\
\alpha_n &\asymp n^{-\min\{2, \frac{r+1}{r(2-p-\theta+\theta p)+p}\}} \\
m &\gtrsim n^{\min\{2p, \frac{p(r+1)}{r(2-p-\theta+\theta p)+p}\}} \log n
\end{aligned}$$

then with high probability

$$L(\widehat{\beta}^{cl}_{\lambda_n,m}) - L(f_*) \lesssim n^{-\min\{\frac{2r}{r+1}, \frac{r}{r(2-p-\theta+\theta p)+p}\}}. \tag{42}$$

eq:25

The above bound further simplifies when the variance bound (38) holds true with the optimal paratemer $\theta = 1$ and the model is well specified as in (6) since we can set $r = 1$. Under these conditions, we get that

$$L(\widehat{\beta}^{cl}_{\lambda_n,m}) - L(w_*) \lesssim n^{-\frac{1}{1+p}}. \tag{43}$$

rate2

with the choice

$$\lambda_n \asymp n^{-\frac{1}{1+p}}, \quad \alpha_n \asymp n^{-\frac{2}{1+p}}, \quad m \gtrsim n^{\frac{2p}{1+p}} \log n. \tag{44}$$

params_choic

By comparing bound (43) with (31), 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 it is clear from (43), 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 (31). In this case, the subspace defined by leverage scores needs to be chosen of dimension at least $O(n^{2/3})$.
Coming back to arbitrary $\theta$ and $r$, bound (42) 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]. Note that large values of $\lambda$ are prevented, indicating a saturation effect (see [Vito et al., 2005, Mücke et al., 2019]). As discussed before, the bound improves when there is a fast eigendecay. Smaller values of $\theta$ and $r$ leads to worse bounds than (43), which is the best rate in this context. Since, given any acceptable choice of $p, r$ and $\theta$, 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).

We conclude this section stating the result just discussed for the *well specified case.*

ast rate f_H

**Corollary 1.** *Fix $\lambda > 0$, $\alpha \gtrsim n^{-1/p}$ and $0 < \delta < 1$. Under Assumptions 1, 2, 6, 7 (with $\theta = 1$) and polynomial decay condition (28), then, with probability at least $1 - 2\delta$:*

$$L(\widehat{\beta}^{cl}_{\lambda,m}) - L(w_*) \lesssim \frac{1}{\lambda^p n} + \lambda \|w_*\|^2 + \sqrt{\alpha}\,\|w_*\| \tag{45}$$

*provided that $n$ and $m$ are large enough.*

## 5.2 Exponential decay of $\Sigma$

exponential

We can further improve the bounds above assuming an exponential decay (28) of the spectrum of $\Sigma$. By omitting numerical constants, logarithmic and higher order terms, Theorem 6 implies that with high probability

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

Under an exponential decay condition, it is reasonable to modify the source condition controlling the behaviour of the approximation error $\mathcal{A}(\lambda)$ from polynomial to logarithmic. We therefore assume that

$$\mathcal{A}(\lambda) \lesssim \log^{-1}(1/\lambda)$$

and, with the choice

$$\lambda_n \asymp \log n / n^2, \quad \alpha_n \asymp 1/n^2, \quad m \gtrsim \log^2 n, \tag{46}$$

params choic

with high probability,

$$L(\widehat{\beta}^{cl}_{\lambda_n,m}) - L(w_*) \lesssim 1/\log n.$$

If the model is well-specified as in (6) and $\theta = 1$, we get

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

provided that $n$ and $m$ are large enough, and $\alpha \gtrsim e^{-n}$. With the choice

$$\lambda_n \asymp 1/n, \quad \alpha_n \asymp 1/n^2, \quad m \gtrsim \log^2 n,$$

with high probability

$$L(\widehat{\beta}^{cl}_{\lambda_n,m}) - L(w_*) \lesssim 1/n.$$

why

*Remark* 2. Whereas the results of Section 4.2 also hold true for bounded inputs $X$, to have fast rates we are forced to assume the sub-Gaussianity of $X$. Under this latter condition in fact, Lemma 4 requires *only* that $\alpha \gtrsim n^{-1/p}$ for polynomial decay and $\alpha \gtrsim e^{-n}$ for exponential decay. These ranges are compatible with the choices (44) and (46), which provide the optimal convergence rates. Under the assumption that $X$ is bounded, Lemma 4 is replaced by Lemma 7 in [Rudi et al., 2015], which requires instead that $\alpha \gtrsim n^{-1}$ both for polynomial and exponential decay, which is not compatible with (44) and (46).

## 5.3 Comparison with Random Features

We start comparing our results with the work [Sun et al., 2018] on random features. Specifically, their Theorem 1 is based on similar assumptions as our Corollary 1, i.e. the Bayes predictor belongs to the RKHS (realizable case), Massart's low-noise condition (implying our variance condition), and the spectrum of the covariance operator decays polynomially: $\sigma_i \asymp i^{-1/p}$, $0 < p < 1$. They obtain a rate of $n^{-1/(2p+1)}$ using $n^{2p/(2p+1)}$ random features. We can obtain the same rate with the same number of Nyström points, but our analysis also provides an improved rate of $n^{-1/(p+1)}$ using $n^{2p/(p+1)}$ Nyström points; this improvement is due to our refined analysis, allowing to consider smaller values of $\alpha$ in Corollary 1. We do not know whether this improvement comes from a better adaptivity of Nyström sampling, or it's a byproduct of our analysis. Regarding [Li et al., 2019], comparison with their fast rates is more difficult, as they assume that the Bayes predictor belongs to the random space spanned by random features. We do not make this strong assumption, and indeed controlling the approximation error of the random subspace is one of the key challenges in our work.

The following table provides a comparison (up to logarithmic factors) among the various rates for the hinge loss discussed above.

---

[*] $\theta = 1$

[†] Here $m$ is number of random features

[‡] $X$ bounded

Table 2: Comparison among the different regimes using hinge loss.

| | Assumptions | Eigen-decay | Rate | m |
|---|---|---|---|---|
| Theorem 2 | 1,2,3 | / | $n^{-1/2}$ | / |
| Eq. (31) | 1,2,3 | $\sigma_j \lesssim j^{-\frac{1}{p}}$ | $n^{-1/2}$ | $n^p$ |
| Eq. (32) | 1,2,3 | $\sigma_j \lesssim e^{-\beta j}$ | $n^{-1/2}$ | $\log^2 n$ |
| Eq: (43) | 1,2,6,7* | $\sigma_j \lesssim j^{-\frac{1}{p}}$ | $n^{-\frac{1}{1+p}}$ | $n^{\frac{2p}{1+p}}$ |
| Eq: (42) | 1,2,7 | $\sigma_j \lesssim j^{-\frac{1}{p}}$ | $n^{-\min\{\frac{2r}{r+1},\frac{r}{r(2-p-\theta+\theta p)+p}\}}$ | $n^{\min\{2p,\frac{p(r+1)}{r(2-p-\theta+\theta p)+p}\}}$ |
| RF† [Sun et al., 2018] | ·‡,2,6,7* | $\sigma_j \lesssim j^{-\frac{1}{p}}$ | $n^{-\frac{1}{2p+1}}$ | $n^{\frac{2p}{2p+1}}$ |

# 6 Differentiable loss functions

sec:other

## 6.1 Square loss

In this section we specialized the analysis to square loss defined by (9) under the assumption that $\mathcal{Y} \subset [-1,1]$. The interval $[-1,1]$ can be replaced by $[-M,M]$, but we take $M=1$ since, in the following section, we will consider binary classification. It is easy to see that

$$\ell(y,t) \leqslant 4, \qquad y,t \in [-1,1],$$

and $\ell$ can be clipped at 1 and a well known variance bound for the least squares shows

$$\left(\ell(y,f^{cl}(x)) - \ell\left(y,f^*(x)\right)\right)^2 = \left(\left(f^{cl}(x)+f^*(x)-2y\right)\left(f^{cl}(x)-f^*(x)\right)\right)^2 \leqslant 16\left(f^{cl}(x)-f^*(x)\right)^2,$$

so that variance bound (38) holds for $V := 16$ and $\theta = 1$.
Finally, the least squares loss restricted to $[-1,1]$ is Lipschitz continuous, that is

$$|L(y,t) - L\left(y,t'\right)| \leqslant 4\left|t-t'\right|$$

for all $y \in [-1,1]$ and $t,t' \in [-1,1]$.
The result for the square loss, whose proof is given in Appendix D.1, reads as follows.

n:square loss

**Theorem 7.** *Under Assumption 1 and polynomial decay condition (28), fix $\lambda > 0$, $\alpha \gtrsim n^{-1/p}$ and $0<\delta<1$. then with probability at least $1-2\delta$:*

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

*Furthermore, if there exists $r \in (0,1]$ such that $\mathcal{A}(\lambda) \lesssim \lambda^r$, then*

$$\lambda_n \asymp n^{-\min\{\frac{2}{r+1},\frac{1}{r+p}\}}, \qquad \alpha_n \asymp n^{-\min\{\frac{2}{r+1},\frac{1}{r+p}\}}, \qquad m \gtrsim n^{\min\{\frac{2p}{r+1},\frac{p}{r+p}\}}\log n$$

*with high probability*

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

As usual the Nyström points are sampled according to ALS, see Definition 1.
Comparing the above bound and (42) with $\theta = 1$, we get the same optimal convergence rates, but the number $m$ of Nyström points reduces from $n^{\min\{2p,\frac{p(r+1)}{r+p}\}}\log n$ to $n^{\min\{\frac{2p}{r+1},\frac{p}{r+p}\}}\log n$, matching the bound in [Rudi et al., 2015].
As already observed in Remark 2 we are able to prove the above results only under the assumption that $X$ sub-Gaussian. However, it is possible to show that in the *well specified case*, see Assumption 6, corresponding to the choice $r=1$, the above result holds true also for bounded inputs $X$. This is due to the additional square we get in the projection term thanks to the quadratic properties of the loss, namely

$$L(\mathcal{P}_m w_*) - L(w_*) = \left\|\Sigma^{1/2}(I-\mathcal{P}_m)w_*\right\|^2$$

so that condition $\alpha \gtrsim n^{-1}$ in Lemma 7 in [Rudi et al., 2015] can still be fulfilled for our choice of the parameter $\alpha$. We state the result without reporting the proof, which is a variant of the proof of Theorem 7 taking into account the above remark.

`e loss basic`

**Corollary 2.** *Assume that $X$ is bounded almost surely, under Assumption 6 and polynomial decay of the spectrum (28), fix $\lambda > 0$, $\alpha \gtrsim 1/n$, and $0 < \delta < 1$. Then, with probability at least $1 - 2\delta$:*

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

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

$$\lambda \asymp n^{-\frac{1}{1+p}}, \quad \alpha \asymp n^{-\frac{1}{1+p}}, \quad m \gtrsim n^{\frac{p}{1+p}} \log n, \tag{47}$$

*with high probability,*

$$L(\widehat{\beta}^{cl}_{\lambda,m}) - L(w_*) \lesssim n^{-\frac{1}{1+p}}. \tag{48}$$

Table 3: Comparison among the different regimes with square loss

| | Assumptions | Eigen-decay | Rate | m |
|---|---|---|---|---|
| Theorem 2 | 1,6 | $\sigma_j \lesssim j^{-\frac{1}{p}}$ | $n^{-\frac{1}{1+p}}$ | $n^{\frac{p}{1+p}}$ |
| [Rudi et al., 2015] | $X$ bounded, 6 | $\sigma_j \lesssim j^{-\frac{1}{p}}$ | $n^{-\frac{1}{1+p}}$ | $n^{\frac{p}{1+p}}$ |
| Theorem 7 | 1 | $\sigma_j \lesssim j^{-\frac{1}{p}}$ | $n^{-\min\{\frac{2r}{r+1},\frac{r}{r+p}\}}$ | $n^{\min\{\frac{2p}{r+1},\frac{p}{r+p}\}}$ |

`tab: square`

*Remark* 3 (Comparison with [Rudi et al., 2015]). The comparison makes sense only when choosing $s = 0$ in the source condition $\|\Sigma^{-s} w_*\|_{\mathcal{H}} < R$ in [Rudi et al., 2015]. The reason is that while in [Rudi et al., 2015] they study the problem in the well specified case –improving the result when $w_*$ belongs to subspaces of $\mathcal{H}$ that are the images of the fractional compact operators $\Sigma^s$– here instead we go in the opposite direction studying the case where $w_*$ does not exists and the approximation error must be introduced. The only intersection is for $s = 0$ where it's reasonable to compare their bound with our Theorem 2. As detailed in Table 3 the two works return exactly the same rate and the same requirement for $m$.

## 6.2 Logistic loss

`c: logistic`

As already mentioned, let's start noticing that logistic loss defined by (8) cannot be clipped according to (34) [Steinwart and Christmann, 2008]. Nevertheless, we can still clip our loss $\ell(y, a)$ at $M = \log n$ so that for all $y \in \mathcal{Y}$, $a \in \mathbb{R}$ it's easy to verify that

$$\ell(y, a^{cl}) \leqslant \ell(y, a) + \frac{1}{n}, \tag{49}$$

`eq: clippin`

where $a^{cl}$ denotes the clipped value of $a$ at $\pm \log(n)$, that is

$$\begin{aligned} a^{cl} &= -\log(n) && \text{if } a \leqslant -\log(n), \\ a^{cl} &= y && \text{if } a \in [-\log(n), \log(n)], \\ a^{cl} &= \log(n) && \text{if } a \geqslant \log(n). \end{aligned}$$

The key point here is that, even though the loss is not always reduced by clipping, i.e. $\exists\, y \in \mathcal{Y}$, $a \in \mathbb{R}$ s.t. $\ell(y, a^{cl}) \nleq \ell(y, a)$, it can only increase at most of $1/n$. This is important since it does not affect the resulting bounds on the excess risk. In particular, we recover Theorem 7 and Corollary 2 for the logistic loss. The proof is given in Appendix D.2.

# 7 From surrogates to classification loss

ec:se0-01loss

In this section we deal with a classification task, so that $\mathcal{Y} = \{\pm 1\}$ and the natural way of measuring performances is by using the 0-1 loss, i.e. $\ell_{0-1}(y, a) := \mathbb{1}_{(-\infty,0]}(y \operatorname{sign}(a))$. Through this section we study how the previous bounds for surrogate losses relate with the 0-1 classification risk. In the following, to make it more explicit, we will indicate with $L_{0-1}$, $L_{hinge}$, $L_{square}$ the risks associated respectively with 0-1, hinge and square losses.
A key role will be played by the well-known low noise condition [Mammen and Tsybakov, 1999, Tsybakov, 2004, Massart et al., 2006]. In the following, the definition is taken from [Tsybakov, 2004]:

**Definition 2.** Distribution $P$ has noise exponent $0 \leqslant \gamma < 1$ if it satisfies equivalently one of the following:

- $N_\gamma$: for some $c > 0$ and all measurable $f : \mathcal{H} \to \{\pm 1\}$

$$\Pr[f(X)(2\eta(X) - 1) < 0] \leqslant c\left(L_{0-1}(f) - L^*_{0-1}\right)^\gamma \tag{50}$$

ass:noise co

- $M_{\frac{\gamma}{1-\gamma}}$: for some $c > 0$ and all $\epsilon > 0$

$$\Pr\left[0 < |2\eta(X) - 1| \leqslant \epsilon\right] \leqslant c\epsilon^{\frac{\gamma}{1-\gamma}} \tag{51}$$

where $\eta(X) = \Pr(Y = 1|X)$ and for $\gamma = 1$ we have $M_\infty$ equivalent to $N_1$.

In the following we will assume this low-noise condition:

: low-noise

**Assumption 8** (Low-noise condition)**.** *The distribution $P$ has noise exponent $\gamma \in [0, 1]$.*

Using the Lemma 10 in Appendix F, when dealing with the square loss, we have the following bound on the classification risk:

quare to 0-1

**Lemma 1** (Square loss)**.** *Under Assumption 8, there is a $c > 0$ such that for any measurable $f : \mathcal{X} \to \mathbb{R}$ we have:*

$$L_{0-1}(f) - L^*_{0-1} \lesssim \left(L_{square}(f) - L^*_{square}\right)^{\frac{1}{2-\gamma}} \tag{52}$$

It's easy to see that an analogous bound can be obtained for logistic loss.

As regards hinge loss, the bound given by Lemma 9 in Appendix F can not be improved even under low noise in Assumption 8. Anyway, it is worth noticing that an assumption of low noise is directly connected with the variance bound (38) through Theorem 8.24 in [Steinwart and Christmann, 2008] (see Lemma 11 in Appendix F). In particular, if we assume a low noise condition with parameter $\gamma$, then the variance bound in Assumption 8 is always satisfied for the hinge loss with $\theta = \gamma$.

## 7.1 From square and logistic losses to classification loss

Starting from Theorem 7, we can now derive an upper bound for the classification risk using the results obtained for the surrogate square loss. We assume low-noise condition and exploit Lemma 1 to obtain the following theorem, where $\mathcal{A}_{\text{square}}(\lambda)$ is the approximation error, see (15), with respect the square loss and the Nyström points are sampled, as always, accordingly to ALS, see Definition 1.

uare loss 01

**Theorem 8.** *Under Assumptions 1 and 8 and the polynomial decay condition (28), fix $\lambda > 0$, $\alpha \gtrsim n^{-1/p}$ and $0 < \delta < 1$, then with probability at least $1 - 2\delta$:*

$$L_{0-1}(\widehat{\beta}^{cl}_{\lambda,m}) - L_{0-1}(f_*) \lesssim \left(\frac{1}{\lambda^p n} + \frac{\alpha \mathcal{A}_{square}(\lambda)}{\lambda} + \frac{\log(3/\delta)}{n}\sqrt{\frac{\mathcal{A}_{square}(\lambda)}{\lambda}} + \mathcal{A}_{square}(\lambda)\right)^{\frac{1}{2-\gamma}}.$$

*Furthermore, if there exists $r \in (0, 1]$ such that $\mathcal{A}_{square}(\lambda) \lesssim \lambda^r$ and choosing*

$$\lambda \asymp n^{-\min\{\frac{2}{r+1}, \frac{1}{r+p}\}}, \qquad \alpha \asymp n^{-\min\{\frac{2}{r+1}, \frac{1}{r+p}\}}, \qquad m \gtrsim n^{\min\{\frac{2p}{r+1}, \frac{p}{r+p}\}} \log n,$$

*then, with high probability*

$$L_{0-1}(\widehat{\beta}^{cl}_{\lambda,m}) - L_{0-1}(f_*) \lesssim n^{-\min\{\frac{2r}{(2-\gamma)(r+1)}, \frac{r}{(2-\gamma)(r+p)}\}}.$$

Once again we have the analogous bounds, up to constant or negligible terms, for logistic loss.

### 7.2 From hinge loss to classification loss

Similarly, starting from Theorem 6, we can derive another upper bound for the classification risk but using as surrogate the hinge loss. Under the low noise assumption and exploiting Lemma 11 as described above, we obtain the following theorem, where $\mathcal{A}_{\text{hinge}}(\lambda)$ is the approximation error, see (15), with respect the hinge loss:

nge loss 0-1

**Theorem 9.** *Under Assumptions 1, 8 and under polynomial decay condition* (28)*, fix* $\lambda > 0$*,* $\alpha \gtrsim n^{-1/p}$ *and* $0 < \delta < 1$*, then with probability at least* $1 - 2\delta$*:*

$$L_{0-1}(\widehat{\beta}^{cl}_{\lambda,m}) - L_{0-1}(f_*) \lesssim \left(\frac{1}{\lambda^p n}\right)^{\frac{1}{2-p-\gamma+\gamma p}} + \sqrt{\frac{\alpha \mathcal{A}_{hinge}(\lambda)}{\lambda}} + \frac{\log(3/\delta)}{n}\sqrt{\frac{\mathcal{A}_{hinge}(\lambda)}{\lambda}} + \mathcal{A}_{hinge}(\lambda).$$

*Furthermore, if there exists* $r \in (0,1]$ *such that* $\mathcal{A}_{hinge}(\lambda) \lesssim \lambda^r$ *and choosing*

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

*then, with high probability*

$$L_{0-1}(\widehat{\beta}^{cl}_{\lambda,m}) - L_{0-1}(f_*) \lesssim n^{-\min\{\frac{2r}{r+1}, \frac{r}{r(2-p-\gamma+\gamma p)+p}\}}.$$

Table 4: Comparison between the $0-1$ classification risk derived from square, logistic and hinge loss under low noise condition

| | Assump | Eigen-decay | Rate | m |
|---|---|---|---|---|
| *Square Loss:* Theorem 8 | 1,8 | $\sigma_j \lesssim j^{-\frac{1}{p}}$ | $n^{-\min\{\frac{2r}{(2-\gamma)(r+1)}, \frac{r}{(2-\gamma)(r+p)}\}}$ | $n^{\min\{\frac{2p}{r+1}, \frac{p}{r+p}\}}$ |
| *Logistic Loss* | 1,8 | $\sigma_j \lesssim j^{-\frac{1}{p}}$ | $n^{-\min\{\frac{2r}{(2-\gamma)(r+1)}, \frac{r}{(2-\gamma)(r+p)}\}}$ | $n^{\min\{\frac{2p}{r+1}, \frac{p}{r+p}\}}$ |
| *Hinge Loss:* Theorem 9 | 1,8 | $\sigma_j \lesssim j^{-\frac{1}{p}}$ | $n^{-\min\{\frac{2r}{r+1}, \frac{r}{r(2-p-\gamma+\gamma p)+p}\}}$ | $n^{\min\{2p, \frac{p(r+1)}{r(2-p-\gamma+\gamma p)+p}\}}$ |

### 7.3 Discussion of the results

Since $\min\{\frac{2r}{(2-\gamma)(r+1)}, \frac{r}{(2-\gamma)(r+p)}\} \leqslant \min\{\frac{2r}{r+1}, \frac{r}{r(2-p-\gamma+\gamma p)+p}\}$ for all the choices of $p$, $\gamma$ and $r$ the bound for the classification error derived using the hinge loss has always a better rate than the one derived from the square loss. On the other hand, since $\min\{\frac{2p}{r+1}, \frac{p}{r+p}\} \leqslant \min\{\frac{2p}{r+1}, \frac{p}{r+p}\}$, the choice of the hinge loss results to be more expensive in term of $m$ (while achieving a better rate). Therefore, we can try to compare the two rates while fixing the number of number of Nyström points selected, or, viceversa, we can fix the rate and compare the number of Nyström points needed to achieve it. The results here are less obvious and we do not have a clear winner. What appears from the analysis is that the discriminant is the choice of the low noise condition parameter $\gamma$ and the $r$ parameter, which controls the approximation error decay.
Let's imagine to fix a realizable convergence rate $O\left(n^{-R}\right)$ for the classification excess risk. To achieve this rate we need $\alpha_{square} = n^{-R(2-\gamma)/r}$ for square loss and $\alpha_{hinge} = n^{-R(1+r)/r}$ for hinge loss. Since having $\alpha_{hinge} \leqslant \alpha_{square}$ means $m_h \geqslant m_s$, we have that, given a fixed rate for the $0-1$ loss, using hinge is *cheaper* than using square loss, when condition $\gamma + r < 1$ is fulfilled (see Figure 1). This suggests that when the problem is hard, hinge loss seems to be even *less expensive* than square loss.
Similarly, imagine now to have some budget constraint on $m$ so that we are not allowed to choose the optimal value: which loss will show a faster non-optimal rate? Again the condition above is key, with hinge loss performing better than square loss when $\gamma + r < 1$ (see Figure 2, where also the saturation effect can be seen).

## 8 Experiments

:experiments

As mentioned in the introduction, a main of motivation for our study is showing that the computational savings can be achieved without incurring in any loss of accuracy. In this section, we complement our theoretical

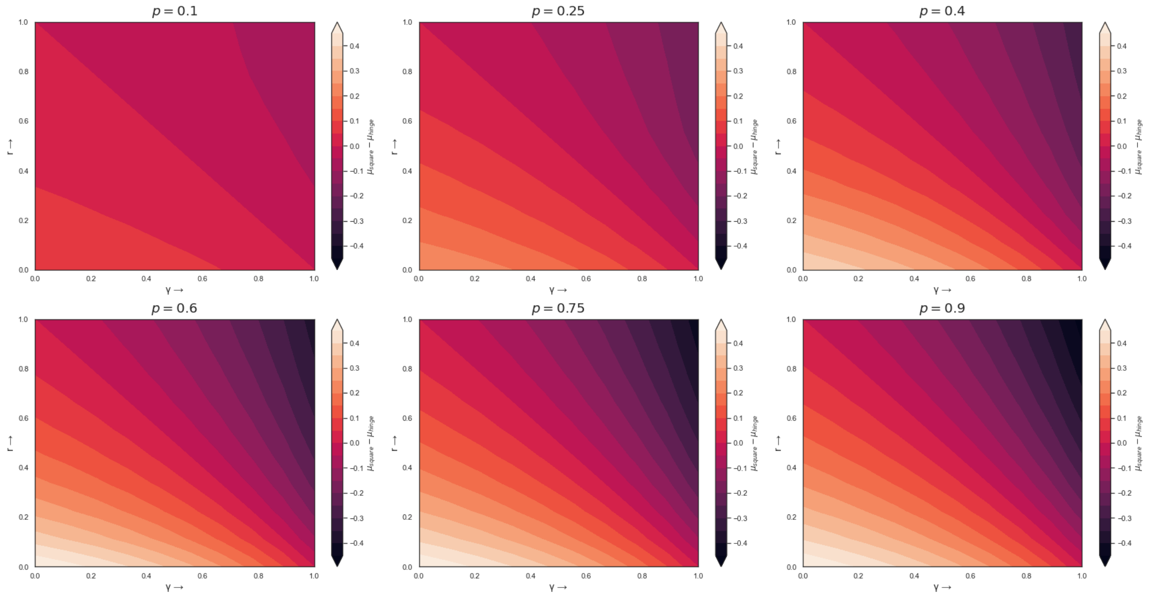

Figure 1: Comparison between the number of Nyström points needed by square and hinge loss to get a fixed common rate: the plots above show $\mu_{square} - \mu_{hinge}$, where $0 \leqslant \mu \leqslant 1$ is the exponent controlling $m$, i.e. $m \asymp n^{\mu}$. Light colours represent then the regimes where hinge loss is *cheaper* than square loss.

fig: m

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 6 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 (12) (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[§]. We consider a procedure in two steps. First, we compute the embedding discussed in Section 4. With kernels it takes the form $\mathbb{x}_i = (K_m^{\dagger})^{1/2}(K(x_i, \tilde{x}_1), \ldots, K(x_i, \tilde{x}_m))^T$, 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 4, the total cost is $O(nm^2 C_K + nm \cdot \#iter)$ 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 & setup (see Appendix G).** We consider five datasets[¶] 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, which is our target, we increase the number of sampled Nyström points $m < n$ as long as also Nyström-Pegasos matches that result.

---

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

[¶] 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

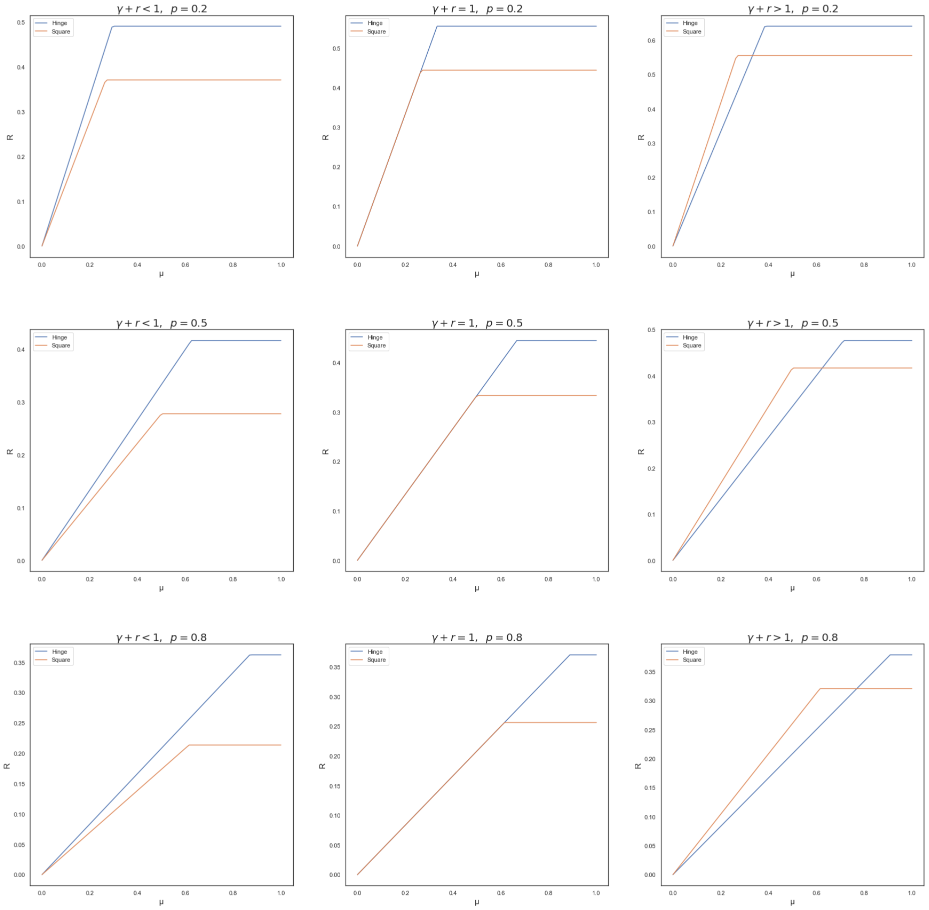

Figure 2: Comparison between the rate achieved by square and hinge loss varying $m$: the plots above have $R$ in the $y$-axis, where $0 \leqslant R \leqslant 1$ is the exponent of the resulting rate, i.e. rate $= n^{-R}$; in the $x$-axis we have $\mu$, with $m = e^{\mu}$ and $0 \leqslant \mu \leqslant 1$ ($\mu = 1$ is equivalent to sample the entire dataset). Every row shows the different behaviours when $\gamma + r$ is respectively less, equal or greater than 1, with $p$ fixed. Note also the saturation effects for hinge and square once we achieve the optimal values for $m$, with hinge loss always reaching a better rate at the end.

`fig: rate`

**Results.** We compare with linear (used only as baseline) and K-SVM see Table 5. 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 3 we illustrate the interplay between $\lambda$ and $m$ for the Nyström-Pegasos considering SUSY data set. In Appendix G we compare also with results obtained using the simpler uniform sampling of the points.

# References

[Adamczak, 2008] Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability*, 13:1000–1034.

[Alaoui and Mahoney, 2015] 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.

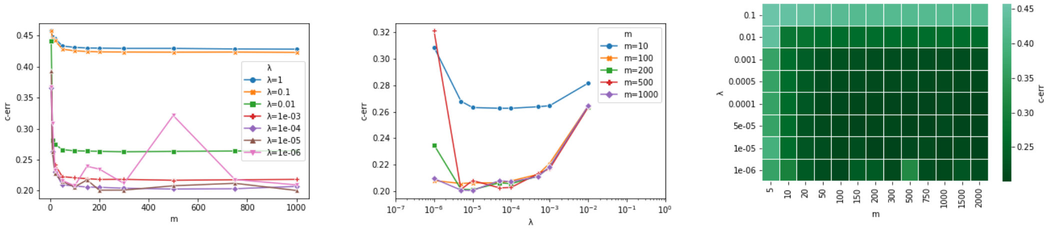

Figure 3: The graphs above are obtained from SUSY data set: on the left we show how c-err measure changes for different choices of $\lambda$ parameter; in the central figure the focus is on the stability of the algorithm varying $\lambda$; on the right the combined behavior is presented with a heatmap.

`fig`

Table 5: Architecture: single machine with AMD EPYC 7301 16-Core Processor and 256GB of RAM. For Nyström-Pegaos, ALS sampling has been used (see [Rudi et al., 2018]) and the results are presented 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).

`tab:results`

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

[Alquier et al., 2019] Alquier, P., Cottet, V., and Lecué, G. (2019). Estimation bounds and sharp oracle inequalities of regularized procedures with lipschitz loss functions. *The Annals of Statistics*, 47(4):2117–2144.

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

[Bach, 2017] 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 et al., 2005] Bartlett, P. L., Bousquet, O., Mendelson, S., et al. (2005). Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

[Bartlett et al., 2006] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

[Bartlett and Mendelson, 2002] 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 and Bousquet, 2008] Bottou, L. and Bousquet, O. (2008). The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168.

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

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

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

[Calandriello et al., 2017] 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 and Rosasco, 2018] 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 and De Vito, 2007] Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368.

[Carl and Stephani, 1990] Carl, B. and Stephani, I. (1990). *Entropy, compactness and the approximation of operators*. Number 98. Cambridge University Press.

[Chang and Lin, 2011] 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 et al., 2015] 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.

[Della Vecchia et al., 2021] Della Vecchia, A., Mourtada, J., De Vito, E., and Rosasco, L. (2021). Regularized erm on random subspaces. In *International Conference on Artificial Intelligence and Statistics*, pages 4006–4014. PMLR.

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

[Drineas et al., 2012] 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 and Mahoney, 2005] 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.

[Hsieh et al., 2014] 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, 1998] Joachims, T. (1998). Making large-scale svm learning practical. Technical report, Technical Report.

[Johnson and Zhang, 2013] 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 et al., 2013] 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 et al., 2009] 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 et al., 2006] Koltchinskii, V. et al. (2006). Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656.

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

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

[Li et al., 2019] 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 et al., 2016] 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, 2011] Mahoney, M. W. (2011). Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224.

[Mammen and Tsybakov, 1999] Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829.

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

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

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

[Mücke et al., 2019] 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, 2018] Nesterov, Y. (2018). *Lectures on convex optimization*, volume 137. Springer.

[Pedregosa et al., 2011] 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, 1970] Rockafellar, R. T. (1970). *Convex analysis*. Number 28. Princeton university press.

[Rudi et al., 2018] 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 et al., 2015] 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 and Rosasco, 2017] 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 et al., 2017] 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 et al., 2001] 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 et al., 2010] 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 et al., 2011] 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 and Zhang, 2013] 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 and Schölkopf, 2000] Smola, A. J. and Schölkopf, B. (2000). Sparse greedy matrix approximation for machine learning.

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

[Steinwart et al., 2009] 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 et al., 2018] 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, 2012] Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434.

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

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

[Vito et al., 2005] 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, 1990] Wahba, G. (1990). *Spline models for observational data*, volume 59. Siam.

[Williams and Seeger, 2001] 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, 2014] Woodruff, D. P. (2014). Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*.

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

# A Proof of Section 3

:radem-compl

This section is devoted to the proof of Theorems 1 and 2. With slight abuse of notation we set

$$\ell(w,z)=\ell(y,\langle w,x\rangle), \qquad z=(x,y)\in\mathcal{H}\times\mathcal{Y},\ w\in\mathcal{H}.$$

With this notation $L(w)=\int_{\mathcal{H}\times\mathcal{Y}}\ell(w,z)dP(z)$.
The following result is known, [Alquier et al., 2019, Lemma 8.1]. We provide an alternative proof tailored to the Hilbert setting.

p-ball-gauss

**Lemma 2.** *Under Assumptions 1 and 2, fix $R>0$ and $\tau>0$, with probability at least $1-\delta$,*

$$\sup_{\|w\|\leqslant R}\left|\widehat{L}(w)-L(w)\right|<\frac{D}{\sqrt{n}}\Big(GRC\|\Sigma\|^{\frac{1}{2}}\big(r_\Sigma+\sqrt{\log(4/\delta)}\big)+\ell_0\sqrt{\log(4/\delta)}\Big), \tag{53}$$

eq:gen-gap-b

*where $D>0$ is an absolute numerical constant and $r_\Sigma^2=\mathrm{Tr}\Sigma/\|\Sigma\|$ is the effective rank of $\Sigma$. Furthermore, for each $w\in\mathcal{H}$, $\widehat{L}(w)-L(w)$ is a sub-Gaussian centered real random variable and*

$$\|\widehat{L}(w)-L(w)\|_{\psi_2}\leqslant\frac{2}{\sqrt{n}}(\ell_0+CG\|\langle X,w\rangle\|_2). \tag{54}$$

eq:23

*Proof.* In the proof $D$ denotes an absolute numerical constant, whose value can change from line to line. Fix $w\in\mathcal{H}$ and define the centered real random variable

$$Z_w=\ell(Y,\langle X,w\rangle)-\mathbb{E}[\ell(Y,\langle X,w\rangle)].$$

We claim that, for any pair $w,w'\in\mathcal{H}$

$$\|Z_w-Z_{w'}\|_{\psi_2}\leqslant 2CG\|\langle X,w-w'\rangle\|_2, \tag{55}$$

eq:14

where $\|Z_w-Z_{w'}\|_{\psi_2}$ is defined by (28). Indeed, for all $p\geqslant 1$, recalling that $\|\xi\|_p=\mathbb{E}[|\xi|^p]^{\frac{1}{p}}$, then triangular inequality and continuity of expectation give

$$\begin{aligned}\|Z_w-Z_{w'}\|_p&\leqslant\|\ell(Y,\langle X,w\rangle)-\ell(Y,\langle X,w'\rangle)\|_p+\|\ell(Y,\langle X,w\rangle)-\ell(Y,\langle X,w'\rangle)\|_1\\&\leqslant 2\|\ell(Y,\langle X,w\rangle)-\ell(Y,\langle X,w'\rangle)\|_p\\&\leqslant 2G\|\langle X,w-w'\rangle)\|_p\leqslant 2GC\sqrt{p}\|\langle X,w-w'\rangle)\|_2\end{aligned}$$

where the last two inequalities are consequence of (6) and (2), respectively. Hence

$$\sup_{p\geqslant 2}\frac{\|Z_w-Z_{w'}\|_p}{\sqrt{p}}\leqslant 2GC\|\langle X,w-w'\rangle\|_2,$$

so that (55) is clear. Furthermore, since

$$\begin{aligned}\big(\widehat{L}(w)-L(w)\big)-\big(\widehat{L}(w')-L(w')\big)=\frac{1}{n}\sum_{i=1}^{n}&((\ell(Y_i,\langle X_i,w\rangle)-\mathbb{E}[\ell(Y_i,\langle X_i,w\rangle)])\\&-(\ell(Y_i,\langle X_i,w'\rangle)-\mathbb{E}[\ell(Y_i,\langle X_i,w'\rangle)]))\end{aligned}$$

is a sum of independent sub-Gaussian random variables distributed as $(Z_w-Z'_w)/n$, then by rotational invariance theorem [Vershynin, 2010, Proposition 2.6.1]

$$\|(\widehat{L}(w)-L(w))-(\widehat{L}(w')-L(w'))\|_{\psi_2}\leqslant\frac{D}{\sqrt{n}}\|Z_w-Z_{w'}\|_{\psi_2}\leqslant\frac{D}{\sqrt{n}}CG\|\langle X,w-w'\rangle)\|_2, \tag{56}$$

eq:18

where the last inequality is a consequence of (55) and $D$ is an absolute constant. Consider $\mathcal{H}$ as a metric space with respect to the metric

$$d(w,w')=\|\langle X,w-w'\rangle\|_2$$

where without loss of generality we assume that $\Sigma$ is injective, then (56) states that the centered random process $\big(\widehat{L}(w)-L(w)\big)_{w\in\mathcal{H}}$ has sub-Gaussian increments and the generic chaining tail bound [Vershynin, 2010, Theorem 8.5.5] implies that, with probability at least $1-2e^{-\tau}$,

$$\sup_{w,w'\in B_R} \big|(\widehat{L}(w)-L(w))-(\widehat{L}(w')-L(w'))\big| \leqslant \frac{D}{\sqrt{n}}CG\big(\sqrt{\tau}\,\mathrm{diam}(B_R)+\gamma_2(B_R)\big), \tag{57}$$

`eq:20`

where $B_R=\{w\in\mathcal{H} : \|w\|\leqslant R\}$, $\mathrm{diam}(B_R)$ and $\gamma_2(B_R)$ are the diamater with respect to the metric $d$ and the Talagrand's $\gamma_2$ functional of $B_R$, [Vershynin, 2010, Definition 8.5.1].
Let $G$ be the Gaussian random vector in $\mathcal{H}$ with covariance $\Sigma$, which always exists since $\Sigma$ is a trace class operator. Talagrand's majorizing measure theorem [Vershynin, 2010, Theorem 8.6.1] implies that

$$\gamma_2(B_R)\leqslant D\mathbb{E}[\sup_{w\in B_R}\langle G,w\rangle]=\mathbb{E}[\sup_{w\in B_R}|\langle G,w\rangle|]=R\,\mathbb{E}[\|G\|]\leqslant R\,\mathbb{E}[\|G\|^2]^{\frac{1}{2}}=R\,\mathrm{Tr}(\Sigma)^{\frac{1}{2}},$$

where the first equality is due to the fact that $B_R$ is symmetric, the second inequality is a consequence of Jansen inequality and the last equality by definition of $G$. Furthermore, the definition of $d$ gives that

$$\mathrm{diam}(B_R)\leqslant 2R\|\Sigma\|^{\frac{1}{2}}.$$

Plugin these last two bounds in (57), it holds that

$$\sup_{w,w'\in B_R} \big|(\widehat{L}(w)-L(w))-(\widehat{L}(w')-L(w'))\big| \leqslant \frac{D}{\sqrt{n}}CGR\big(\sqrt{\tau}\|\Sigma\|^{\frac{1}{2}}+\mathrm{Tr}(\Sigma)^{\frac{1}{2}}\big). \tag{58}$$

`eq:21`

with high probability. Finally, observe that

$$|\ell(Y,0)-\mathbb{E}[\ell(Y,0)]|\leqslant 2\sup_{y\in Y}\ell(y,0)=2\ell_0,$$

by (6), and

$$\widehat{L}(0)-L(0)=\frac{1}{n}\sum_{i=1}(\ell(Y_i,0)-\mathbb{E}[\ell(Y_i,0)])$$

so that Hoeffding's inequality [Boucheron et al., 2013] implies that, with probability $1-2e^{-\tau}$,

$$|\widehat{L}(0)-L(0)|\leqslant 2\ell_0\sqrt{\frac{2\tau}{n}}. \tag{59}$$

`eq:22`

Finally, since

$$\sup_{w\in B_R}|\widehat{L}(w)-L(w)|\leqslant \sup_{w\in B_R}|\widehat{L}(w)-L(w)-(\widehat{L}(0)-L(0))|+|\widehat{L}(0)-L(0)|$$

bounds (58) and (59) give (53) with $4\exp(-\tau)=\delta$. Bound (56) with $w'=0$ implies (54). ☐

This result cannot be readily applied to $\widehat{w}_\lambda$, since its norm $\|\widehat{w}_\lambda\|$ is itself random. Observe that, by definition and by Assumption 2,

$$\lambda\|\widehat{w}_\lambda\|^2\leqslant \widehat{L}_\lambda(\widehat{w}_\lambda)\leqslant \widehat{L}_\lambda(0)=\widehat{L}(0)\leqslant\sup_{y\in\mathcal{Y}}\ell(y,0)=\ell_0,$$

so that $\|\widehat{w}_\lambda\|\leqslant\sqrt{\ell_0/\lambda}$. One could in principle apply this bound on $\widehat{w}_\lambda$, but this would yield a suboptimal dependence on $\lambda$ and thus a suboptimal rate.
The next step in the proof is to make the bound of Lemma 2 valid for all norms $R$, so that it can be applied to the random quantity $R=\|\widehat{w}_\lambda\|$. This is done in Lemma 3 below though a union bound.

`en-gap-union` **Lemma 3.** *Under Assumptions 1 and 2, $\forall w\in\mathcal{H}$, with probability $1-\delta$:*

$$L(w)-\widehat{L}(w)\leqslant\frac{DGC\|\Sigma\|^{\frac{1}{2}}(1+\|w\|)r_\Sigma}{\sqrt{n}}+\frac{D}{\sqrt{n}}\Big(GC\|\Sigma\|^{\frac{1}{2}}(1+\|w\|)+\ell_0\Big)\sqrt{\log(2+\log_2(1+\|w\|))+\log(1/\delta)}.$$

*Proof.* Fix $\delta \in (0,1)$. For $p \geqslant 1$, let $R_p := 2^p$ and $\delta_p = \delta/(p(p+1))$. By Lemma 2, one has for every $p \geqslant 1$,

$$\mathbb{P}\left(\sup_{\|w\|\leqslant R_p}\Big[L(w)-\widehat{L}(w)\Big] \geqslant \frac{D}{\sqrt{n}}\Big(GR_pC\|\Sigma\|^{\frac{1}{2}}(r_\Sigma+\sqrt{\log(1/\delta_p)})+\ell_0\sqrt{\log(1/\delta_p)}\Big)\right) \leqslant \delta_p.$$

Collecting the terms containing $\delta_p$ and taking a union bound over $p \geqslant 1$ while using that $\sum_{p\geqslant 1}\delta_p = \delta$ and $\delta_p \geqslant \delta^2/(p+1)^2$, we get:

$$\mathbb{P}\left(\exists p \geqslant 1,\quad \sup_{\|w\|\leqslant R_p}\Big[L(w)-\widehat{L}(w)\Big] \geqslant \frac{D}{\sqrt{n}}\Big(GR_pC\|\Sigma\|^{\frac{1}{2}}(r_\Sigma+\sqrt{\log\frac{p+1}{\delta}})+\ell_0\sqrt{\log\frac{p+1}{\delta}}\Big)\right) \leqslant \delta.$$

Now, for $w \in \mathcal{H}$, let $p = \lceil \log_2(1+\|w\|)\rceil$; then, $1+\|w\| \leqslant R_p = 2^p \leqslant 2(1+\|w\|)$, so $\|w\| \leqslant R_p$. Hence, $\forall w \in \mathcal{H}$, with probability $1-\delta$:

$$\begin{aligned}
L(w)-\widehat{L}(w) &\leqslant \frac{DGC\|\Sigma\|^{\frac{1}{2}}(1+\|w\|)r_\Sigma}{\sqrt{n}} + \frac{D}{\sqrt{n}}\sqrt{\log\frac{p+1}{\delta}}\Big(GC\|\Sigma\|^{\frac{1}{2}}(1+\|w\|)+\ell_0\Big)\\
&\leqslant \frac{DGC\|\Sigma\|^{\frac{1}{2}}(1+\|w\|)r_\Sigma}{\sqrt{n}} + \frac{D}{\sqrt{n}}\Big(GC\|\Sigma\|^{\frac{1}{2}}(1+\|w\|)+\ell_0\Big)\sqrt{\log(2+\log_2(1+\|w\|))+\log(1/\delta)}\\
&\leqslant \delta.
\end{aligned}$$

This is precisely the desired bound. □

We are now able to prove the two theorems.

*Proof of Theorem 1.* Since the bound of Lemma 3 holds simultaneously for all $w \in \mathcal{H}$, one can apply it to $\widehat{w}_\lambda$; using the inequality $\|\widehat{w}_\lambda\| \leqslant \sqrt{\ell_0/\lambda} \leqslant (1+\ell_0/\lambda)/2$ to bound the $\log\log$ term, this gives with probability $1-\delta$,

$$L(\widehat{w}_\lambda)-\widehat{L}(\widehat{w}_\lambda) \leqslant \frac{DGC\|\Sigma\|^{\frac{1}{2}}(1+\|\widehat{w}_\lambda\|)r_\Sigma}{\sqrt{n}}+\frac{D}{\sqrt{n}}\Big(GC\|\Sigma\|^{\frac{1}{2}}(1+\|\widehat{w}_\lambda\|)+\ell_0\Big)\sqrt{\log(1+\log_2(3+\ell_0/\lambda))+\log(1/\delta)}. \tag{60}$$

`eq:gen-gap-w`

Now, let $K = K_{\lambda,\delta} = r_\Sigma + \sqrt{\log(1+\log_2(3+\ell_0/\lambda))+\log(1/\delta)}$. Eq (60) writes

$$L(\widehat{w}_\lambda)-\widehat{L}(\widehat{w}_\lambda) \leqslant \frac{DGCK\|\Sigma\|^{\frac{1}{2}}(1+\|\widehat{w}_\lambda\|)}{\sqrt{n}}+\frac{D\ell_0(K-r_\Sigma)}{\sqrt{n}} \tag{61}$$

`eq:lemma3 si`

Using that $ab \leqslant \lambda a^2 + b^2/(4\lambda)$ for $a,b \geqslant 0$, one can then write

$$\begin{aligned}
L(\widehat{w}_\lambda) &\leqslant \widehat{L}(\widehat{w}_\lambda)+\frac{DGCK\|\Sigma\|^{\frac{1}{2}}\,\|\widehat{w}_\lambda\|}{\sqrt{n}}+\frac{DGCK\|\Sigma\|^{\frac{1}{2}}}{\sqrt{n}}+\frac{D\ell_0(K-r_\Sigma)}{\sqrt{n}}\\
&\leqslant \widehat{L}(\widehat{w}_\lambda)+\lambda\|\widehat{w}_\lambda\|^2+\frac{D^2G^2C^2K^2\|\Sigma\|}{4\lambda n}+\frac{DGCK\|\Sigma\|^{\frac{1}{2}}+D\ell_0(K-r_\Sigma)}{\sqrt{n}}\\
&\leqslant \widehat{L}(w_\lambda)+\lambda\|w_\lambda\|^2+\frac{D^2G^2C^2K^2\|\Sigma\|}{4\lambda n}+\frac{DGCK\|\Sigma\|^{\frac{1}{2}}+D\ell_0(K-r_\Sigma)}{\sqrt{n}} \qquad (62)
\end{aligned}$$

`eq:proof-wla`

where (62) holds by definition of $\widehat{w}_\lambda$. Now, using again Lemma 2 for $\|w_\lambda\|$ we have that, with probability $1-\delta$:

$$\widehat{L}(w_\lambda)-L(w_\lambda) < \frac{D}{\sqrt{n}}\Big(GC\|\Sigma\|^{\frac{1}{2}}\,\|w_\lambda\|\,\big(r_\Sigma+\sqrt{\log(4/\delta)}\big)+\ell_0\sqrt{\log(4/\delta)}\Big).$$

Combining this inequality with (62) with a union bound, with probability $1-2\delta$:

$$\begin{aligned}
L(\widehat{w}_\lambda) <\,& L(w_\lambda)+\lambda\|w_\lambda\|^2+\frac{D^2G^2C^2K^2\|\Sigma\|}{4\lambda n}+\frac{DGCK\|\Sigma\|^{\frac{1}{2}}+D\ell_0(K-r_\Sigma)}{\sqrt{n}}+\\
&+\frac{DGC\|\Sigma\|^{\frac{1}{2}}\,\|w_\lambda\|\,\big(r_\Sigma+\sqrt{\log(4/\delta)}\big)}{\sqrt{n}}+\frac{D\ell_0\sqrt{\log(4/\delta)}}{\sqrt{n}}. \qquad (63)
\end{aligned}$$

`eq:proof-wla`

Since $ab \leqslant \lambda a^2 + b^2/(4\lambda)$, then

$$\begin{aligned}\frac{DGC\|\Sigma\|^{\frac{1}{2}}\,\|w_\lambda\|\left(r_\Sigma+\sqrt{\log(1/\delta)}\right)}{\sqrt{n}} &\leqslant \lambda\|w_\lambda\|^2 + \frac{D^2G^2C^2\|\Sigma\|\left(r_\Sigma+\sqrt{\log(4/\delta)}\right)^2}{4\lambda n}\\ &\leqslant \mathcal{A}(\lambda) + \frac{D^2G^2C^2\|\Sigma\|\left(r_\Sigma+\sqrt{\log(4/\delta)}\right)^2}{4\lambda n}\end{aligned}$$

so that (63) implies, with probability $1-2\delta$:

$$L(\widehat{w}_\lambda) - \inf_{w\in\mathcal{H}} L(w) < 2\mathcal{A}(\lambda) + \frac{D^2G^2C^2\|\Sigma\|(K^2+(r_\Sigma+\sqrt{\log(4/\delta)})^2)}{4\lambda n} + \frac{DGCK\|\Sigma\|^{\frac{1}{2}} + D\ell_0(K-r_\Sigma+\sqrt{\log(4/\delta)})}{\sqrt{n}}.$$

After replacing $\delta$ by $\delta/2$, we get bound (16). □

*Proof of Theorem 2.* Assume that $w_* = \arg\min_{w\in\mathcal{H}} L(w)$ exists. Then, by definition of $w_\lambda$,

$$L(w_\lambda) + \lambda\|w_\lambda\|^2 \leqslant L(w_*) + \lambda\|w_*\|^2.$$

In addition, $\|w_\lambda\| \leqslant \|w_*\|$, since otherwise having $\|w_*\| < \|w_\lambda\|$ and $L(w_*) \leqslant L(w_\lambda)$ would imply $L(w_*) + \lambda\|w_*\|^2 < L(w_\lambda) + \lambda\|w_\lambda\|^2$, contradicting the above inequality. Since $L(w_*) = \inf_{\mathcal{H}} L$, it follows from (63) that, with probability $1-2\delta$,

$$\begin{aligned}L(\widehat{w}_\lambda) <& L(w_*) + \lambda\|w_*\|^2 + \frac{D^2G^2C^2K^2\|\Sigma\|}{4\lambda n} + \frac{DGCK\|\Sigma\|^{\frac{1}{2}} + D\ell_0(K-r_\Sigma)}{\sqrt{n}} +\\ &+ \frac{DGC\|\Sigma\|^{\frac{1}{2}}\,\|w_*\|\left(r_\Sigma+\sqrt{\log(4/\delta)}\right)}{\sqrt{n}} + \frac{D\ell_0\sqrt{\log(4/\delta)}}{\sqrt{n}}\end{aligned} \tag{64}$$

eq:proof-mai

The bound (64) precisely corresponds to the desired bound (17) after replacing $\delta$ by $\delta/2$. In particular, tuning $\lambda \asymp (DGCK\,\|\Sigma\|^{1/2}/\|w_*\|)\sqrt{\log(1/\delta)/n}$ yields

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

Omitting the $\log\log n$ term, this bound essentially scales as $\widetilde{O}(DGC\,\|\Sigma\|^{1/2}\,\|w_*\|\sqrt{\log(1/\delta)/n})$. □

roofthmbasic

## B Proof of Section 4

In order to prove Theorem 3, we need to previously extend Lemma 7 in [Rudi et al., 2015] to sub-Gaussian random variables.

lev_subgauss

**Lemma 4.** *Fix $\delta > 0$ and a $(T,\alpha_0)$-approximate leverage scores $(\hat{l}_i(\alpha))_{i=1}^n$ with confidence $\delta > 0$. Given $\alpha > \alpha_0$, let $\{\widetilde{x}_1,\dots,\widetilde{x}_m\}$ be the Nyström points selected according to Definition 1 and set $\mathcal{B}_m = span\{\widetilde{x}_1,\dots,\widetilde{x}_m\}$. Under Assumption 1, with probability at least $1-\delta$:*

$$\left\|(I-\mathcal{P}_{\mathcal{B}_m})\Sigma^{1/2}\right\|^2 \leqslant \left\|(I-\mathcal{P}_{\mathcal{B}_m})(\Sigma+\alpha\,\mathrm{I})^{1/2}\right\|^2 \leqslant 3\alpha, \tag{65}$$

*provided that*

$$n \gtrsim d_\alpha \vee \log(5/\delta) \tag{66}$$

eq: subgaus

$$m \gtrsim d_\alpha \log(\frac{10n}{\delta}). \tag{67}$$

eq: subgaus

*Furthermore, if the spectrum of $\Sigma$ satisfies the decay conditions (28) (polynomial decay) or (29) (exponential decay), it is enough to assume that*

$$n \gtrsim \log(5/\delta) \qquad \alpha \gtrsim n^{-1/p} \qquad m \gtrsim \alpha^{-p}\log(\frac{10n}{\delta}) \qquad \textit{polynomial decay} \tag{68}$$

eq:31a

$$n \gtrsim \log(5/\delta) \qquad \alpha \gtrsim e^{-n} \qquad m \gtrsim \log(1/\alpha)\log(\frac{10n}{\delta}) \qquad \textit{exponential decay} \tag{69}$$

eq:32a

*Proof.* Exploiting sub-Gaussianity anyway the various terms are bounded differently. In particular, to bound $\beta_1$ we refer to Theorem 9 in [Koltchinskii and Lounici, 2014], obtaining with probability at least $1-\delta$

$$\beta_1(\alpha) \lesssim \max\left\{\sqrt{\frac{d_\alpha}{n}}, \sqrt{\frac{\log(1/\delta)}{n}}\right\}. \tag{70}$$

As regards $\beta_3$ term we apply Proposition 1 below to get with probability greater than $1-3\delta$

$$\beta_3(\alpha) \leqslant \frac{2\log\frac{2n}{\delta}}{3m} + \sqrt{\frac{32T^2 d_\alpha \log\frac{2n}{\delta}}{m}}$$

for $n \geqslant 2C^2\log(1/\delta)$.
Finally, taking a union bound we have with probability at least $1-5\delta$

$$\begin{aligned}\beta(\alpha) \lesssim &\max\left\{\sqrt{\frac{d_\alpha}{n}}, \sqrt{\frac{\log(\frac{1}{\delta})}{n}}\right\} + \\ &+ \left(1+\max\left\{\sqrt{\frac{d_\alpha}{n}}, \sqrt{\frac{\log(\frac{1}{\delta})}{n}}\right\}\right)\left(\frac{2\log\frac{2n}{\delta}}{3m} + \sqrt{\frac{32T^2 d_\alpha \log\frac{2n}{\delta}}{m}}\right) \lesssim 1\end{aligned}$$

when $n \gtrsim d_\alpha \vee \log(1/\delta)$ and $m \gtrsim d_\alpha \log\frac{2n}{\delta}$. See [Rudi et al., 2015] to conclude the proof of the first claim. Assume now (28) or (29) . The second claim is consequence of Proposition 2 or Proposition 3. ☐

We can proceed now with the proof of Theorem 3:

*Proof of Theorem 3.* We recall the notation.

$$\mathcal{B}_m = \operatorname{span}\{\tilde{x}_1, \dots, \tilde{x}_m\}, \qquad \widehat{\beta}_\lambda = \operatorname*{arg\,min}_{w\in\mathcal{B}_m} \widehat{L}(w), \qquad w_* = \operatorname*{arg\,min}_{w\in\mathcal{H}} L(w)$$

and $\mathcal{P}_m = \mathcal{P}_{\mathcal{B}_m}$ the orthogonal projector operator onto $\mathcal{B}_m$.
In order to bound the excess risk of $\widehat{\beta}_\lambda$, we decompose the error as follows:

$$\begin{aligned}L(\widehat{\beta}_\lambda) - L(w_*) \leqslant &\left|L(\widehat{\beta}_\lambda) - \widehat{L}(\widehat{\beta}_\lambda) - \lambda\|\widehat{\beta}_\lambda\|^2_{\mathcal{H}}\right| + \left|\widehat{L}(\widehat{\beta}_\lambda) + \lambda\|\widehat{\beta}_\lambda\|^2_{\mathcal{H}} - \widehat{L}(\mathcal{P}_m w_*) - \lambda\|\mathcal{P}_m w_*\|^2_{\mathcal{H}}\right| + \\ &+ \left|\widehat{L}(\mathcal{P}_m w_*) - L(\mathcal{P}_m w_*)\right| + |L(\mathcal{P}_m w_*) - L(w_*)| + \lambda\|\mathcal{P}_m w_*\|^2_{\mathcal{H}}\end{aligned} \tag{71}$$

1st_split_de

To bound the first term $\left|L(\widehat{\beta}_\lambda) - \widehat{L}(\widehat{\beta}_\lambda) - \lambda\|\widehat{\beta}_\lambda\|^2_{\mathcal{H}}\right|$ we apply Lemma 3 for $\widehat{\beta}_\lambda$ and we get

$$L(\widehat{\beta}_\lambda) - \widehat{L}(\widehat{\beta}_\lambda) \leqslant \frac{DGCK\|\Sigma\|^{\frac{1}{2}}(1+\|\widehat{\beta}_\lambda\|)}{\sqrt{n}} + \frac{D\ell_0(K-r_\Sigma)}{\sqrt{n}}$$

with $K = K_{\lambda,\delta} = r_\Sigma + \sqrt{\log(1+\log_2(3+\ell_0/\lambda)) + \log(1/\delta)}$ as in (61).
Now since $xy \leqslant \lambda x^2 + y^2/(4\lambda)$, we can write

$$\frac{DGCK\|\widehat{\beta}_\lambda\|\|\Sigma\|^{\frac{1}{2}}}{\sqrt{n}} \leqslant \lambda\|\widehat{\beta}_\lambda\|^2 + \frac{D^2G^2C^2K^2\|\Sigma\|}{\lambda n} \tag{72}$$

hence,

$$\left|L(\widehat{\beta}_\lambda) - \widehat{L}(\widehat{\beta}_\lambda) - \lambda\|\widehat{\beta}_\lambda\|^2\right| \leqslant \frac{D^2G^2C^2K^2\|\Sigma\|}{\lambda n} + \frac{DGCK\|\Sigma\|^{\frac{1}{2}}}{\sqrt{n}} + \frac{D\ell_0(K-r_\Sigma)}{\sqrt{n}}, \tag{73}$$

bound_A_unio

Term $\left|\widehat{L}(\widehat{\beta}_\lambda) + \lambda\|\widehat{\beta}_\lambda\|_{\mathcal{H}}^2 - \widehat{L}(\mathcal{P}_m w_*) - \lambda\|\mathcal{P}_m w_*\|_{\mathcal{H}}^2\right|$ is less or equal than 0.

As regards term $\left|\widehat{L}(\mathcal{P}_m w_*) - L(\mathcal{P}_m w_*)\right|$, since $\mathcal{P}_m$ is a projection $\|\mathcal{P}_m w_*\| \leqslant \|w_*\|$, so that with probability at least $1-\delta$:

$$\begin{aligned} \left|\widehat{L}(\mathcal{P}_m w_*) - L(\mathcal{P}_m w_*)\right| &\leqslant \sup_{\|w\|\leqslant\|w_*\|} \left(\left|\widehat{L}(w) - L(w)\right|\right) \\ &< \frac{D}{\sqrt{n}}\left(GC\|w_*\|\|\Sigma\|^{\frac{1}{2}}\left(r_\Sigma + \sqrt{\log(4/\delta)}\right) + \ell_0\sqrt{\log(4/\delta)}\right). \qquad (74) \end{aligned}$$

eq:6

where in the sup in the left hand side is taken over all possible Nyström points and the second inequality is the content of Lemma 2 where the role of $L$ and $\widehat{L}$ is interchanged.

Finally, term $|L(\mathcal{P}_m w_*) - L(w_*)|$ can be rewritten as

$$\begin{aligned} |L(\mathcal{P}_m w_*) - L(w_*)| &\leqslant G\int |\langle w, \mathcal{P}_m w_*\rangle - \langle w, w_*\rangle| dP_X(w) \\ &\leqslant G\left(\int |\langle w, (I-\mathcal{P}_m)w_*\rangle|^2 dP_X(w)\right)^{\frac{1}{2}} \\ &= G\left\langle \Sigma(I-\mathcal{P}_m)w_*, (I-\mathcal{P}_m)w_*\right\rangle^{\frac{1}{2}} \qquad (75) \\ &= G\|\Sigma^{1/2}(I-\mathcal{P}_m)w_*\|_{\mathcal{H}} \\ &\leqslant G\|\Sigma^{1/2}(I-\mathcal{P}_m)\|\|w_*\|_{\mathcal{H}} \\ &= G\|(I-\mathcal{P}_m)\Sigma^{1/2}\|\|w_*\|_{\mathcal{H}} \leqslant G\sqrt{3\alpha}\|w_*\|, \qquad (76) \end{aligned}$$

reg C term

where the last bound is a consequence of Lemma 4 and it holds true with probability at least $1-\delta$. Putting the pieces together we finally get the result in Theorem 3 by replacing $\delta$ with $\delta/3$. ☐

*Proof of Theorem. 4.* Under polynomial decay assumption (28), the claim is a consequence of Theorem 3 with Proposition 2 with $\beta = 1/p$ so that

$$m \gtrsim d_\alpha \log n, \qquad d_\alpha \lesssim \alpha^{-p}, \qquad m \asymp n^p(\log n)^{1-p} \qquad (77)$$

Under exponential decay assumption (29), the claim is a consequence of Theorem 3 with Proposition 3 so that

$$m \gtrsim d_\alpha \log n, \qquad d_\alpha \lesssim \log(1/\alpha), \qquad m \asymp \log^2 n \qquad (78)$$

☐

*Proof of Theorem 5.* The proof is given by decomposing the excess risk as in (71) where $\mathcal{P}_m$ is replaced by $\mathcal{P}_{\mathcal{B}}$, (73) bounds term A, (74) bounds term B and (75) and 33 bound term C. ☐

## C Proofs of Section 5

pp:theorem 4

The following proposition provides a bound on the empirical effective dimension $d_\alpha(\widehat{\Sigma}) = \mathrm{Tr}(\widehat{\Sigma}_\alpha^{-1}\widehat{\Sigma})$ in terms of the correspondent population quantity $d_\alpha = \mathrm{Tr}((\Sigma_\alpha + \alpha\,\mathrm{I})^{-1}\Sigma)$.

und_emp_deff

**Proposition 1.** *Let $X, X_1, \ldots, X_n$ be iid $C$-sub-Gaussian random variables in $\mathcal{H}$. For any $\delta > 0$ and $n \geqslant 2C^2\log(1/\delta)$, then the following hold with probability $1-\delta$*

$$d_\alpha(\widehat{\Sigma}) \leqslant 16 d_\alpha \qquad (79)$$

*Proof.* Let $V_\alpha$ be the space spanned by eigenvectors of $\Sigma$ with corresponding eigenvalues $\alpha_j \geqslant \alpha$, and call $D_\alpha$ its dimension. Notice that $D_\alpha \leqslant 2d_\alpha$ since $d_\alpha = \mathrm{Tr}((\Sigma_\alpha + \alpha\,\mathrm{I})^{-1}\Sigma) = \sum \frac{\alpha_i}{\alpha_i+\alpha}$, where in the sum we have

$D_\alpha$ terms greater or equal than $1/2$.
Let $X = X_1 + X_2$, where $X_1$ is the orthogonal projection of $X$ on the space $V_\alpha$, we have

$$\widehat{\Sigma} = \widehat{\Sigma}_1 + \widehat{\Sigma}_2 + \frac{1}{n}\sum_{i=1}^{n}(X_{1,i}X_{2,i}^\top + X_{2,i}X_{1,i}^\top) \preccurlyeq 2(\widehat{\Sigma}_1 + \widehat{\Sigma}_2) \tag{80}$$

Now, since the function $g : t \mapsto \frac{t}{t+\alpha}$ is sub-additive (meaning that $g(t+t') \leqslant g(t) + g(t')$), denoting $d_\alpha(\Sigma) = \operatorname{Tr} g(\Sigma) = \operatorname{Tr}((\Sigma_\alpha + \alpha\,\mathrm{I})^{-1}\Sigma)$,

$$d_\alpha(\widehat{\Sigma}) \leqslant 2(d_\alpha(\widehat{\Sigma}_1) + d_\alpha(\widehat{\Sigma}_2)) \tag{81}$$

and, since $(\widehat{\Sigma}_1 + \alpha)^{-1}\widehat{\Sigma}_1 \preccurlyeq I_{V_\alpha}$,

$$\operatorname{Tr}((\widehat{\Sigma}_\alpha + \alpha\,\mathrm{I})^{-1}\widehat{\Sigma}) \leqslant 2D_\alpha + \frac{2\operatorname{Tr}(\widehat{\Sigma}_2)}{\alpha} = 4d_\alpha + \frac{2\operatorname{Tr}(\widehat{\Sigma}_2)}{\alpha} \tag{82}$$

`upper_bound_`

Now,

$$\operatorname{Tr}(\widehat{\Sigma}_2) = \frac{1}{n}\sum_{i=1}^{n}\|X_{2,i}\|^2$$

It thus suffices establish concentration for averages of the random variable $\|X_2\|^2$.
Since $X$ is sub-Gaussian then $\|X_2\|^2$ is sub-exponential. In fact, since $X$ is $C$-sub-Gaussian then

$$\|\langle v, X\rangle\|_{\psi_2} \leqslant C\|\langle v, X\rangle\|_{L_2} \qquad \forall v \in \mathcal{H} \tag{83}$$

and given that $\langle v, \mathcal{P}X\rangle = \langle \mathcal{P}v, X\rangle$ with $\mathcal{P}$ an orthogonal projection, then also $X_2$ is $C$-sub-Gaussian. Now take $e_i$ the orthonormal basis of $V$ composed by the eigenvectors of $\Sigma_2 = \mathbb{E}[X_2X_2^T]$, then

$$\begin{aligned}\left\|\|X_2\|^2\right\|_{\psi_1} &= \left\|\sum_i \langle X_2, e_i\rangle^2\right\|_{\psi_1} \leqslant \sum_i \left\|\langle X_2, e_i\rangle^2\right\|_{\psi_1} &(84)\\ &= \sum_i \|\langle X_2, e_i\rangle\|_{\psi_2}^2 \leqslant C^2\,\|\langle X_2, e_i\rangle\|_{L_2}^2 &(85)\\ &= C^2\sum_i \alpha_i = C^2\operatorname{Tr}[\Sigma_2] = C^2\mathbb{E}\left[\|X_2\|^2\right] &(86)\end{aligned}$$

so $\|X_2\|^2$ is $C^2\mathbb{E}\left[\|X_2\|^2\right]$-sub-exponential. Note that $\mathbb{E}\|X_2\|^2 = \mathbb{E}[\operatorname{Tr}(X_2X_2^\top)] = \operatorname{Tr}(\Sigma_2) \leqslant 2\alpha d_\alpha(\Sigma)$, in fact

$$d_\alpha = \sum_{i=1}^{\infty}\frac{\alpha_i}{\alpha_i+\alpha} \geqslant \sum_{i:\alpha_i<\alpha}\frac{\alpha_i}{\alpha_i+\alpha} \geqslant \sum_{i:\alpha_i<\alpha}\frac{\alpha_i}{2\alpha} = \frac{\operatorname{Tr}(\Sigma_2)}{2\alpha} \tag{87}$$

Hence, we can apply then Bernstein inequality for sub-exponential scalar variables (see Theorem 2.10 in [Boucheron et al., 2013]), with parameters $\nu$ and $c$ given by

$$\begin{aligned}&n\mathbb{E}\left[\|X_2\|^4\right] \leqslant \underbrace{4nC^2\alpha^2d_\alpha^2(\Sigma)}_{\nu} &(88)\\ &c = C\alpha d_\alpha &(89)\end{aligned}$$

where we used the bound on the moments of a sub-exponential variable (see [Vershynin, 2010]).
With high probability (82) becomes

$$d_\alpha(\widehat{\Sigma}) \leqslant 8d_\alpha + \frac{4Cd_\alpha\sqrt{2\log(1/\delta)}}{\sqrt{n}} + \frac{2Cd_\alpha\log(1/\delta)}{n} \leqslant 16d_\alpha \tag{90}$$

for $n \geqslant 2C^2\log(1/\delta)$. $\square$

From [Adamczak, 2008] Theorem 4 we write a concentration inequality we will use in the following, corresponding to the simplified Talagrand's inequality in Theorem 7.5 of [Steinwart and Christmann, 2008] but for sub-exponential random variables:

`grand subexp`

**Theorem 10** (Theorem 4 in [Adamczak, 2008])**.** *Let $X, X_1, \dots, X_n$ be i.i.d. random variables with values in a measurable space $(\mathcal{S}, \mathcal{B})$ and let $\mathcal{F}$ be a countable class of measurable functions $f : \mathcal{S} \to \mathbb{R}$. Assume that $\mathbb{E}f(X) = 0$ and $\left\|\sup_f |f(X)|\right\|_{\psi_1} < \infty$ for every $f \in \mathcal{F}$. Let*

$$Z = \sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right|$$

*and define*

$$\sigma^2 = \sup_{f\in\mathcal{F}} \mathbb{E}f(X)^2 .$$

*Then, for all $\tau > 0$ and $\eta > 0$, we have*

$$\mathbb{P}\left( Z \geqslant (1+\eta)\mathbb{E}Z + \frac{K_1 \left\|\sup_{f\in\mathcal{F}} |f(X)|\right\|_{\psi_1} (2+\tau)}{n} + \sqrt{\frac{3(1+\tau)\sigma^2}{n}} \right) \leqslant e^{-\tau} \tag{91}$$

*where $K_1 = K_1(\delta, \eta)$.*

Similarly to [Steinwart and Christmann, 2008], we define the quantity

$$g_{w,r} := \frac{h_w - \mathbb{E}h_w}{\lambda \|w\|^2 + \mathbb{E}h_w + r}, \quad w \in \mathcal{H}, \quad r > 0 \tag{92}$$

`def: g_w,r`

(notice that in [Steinwart and Christmann, 2008] they define $-g_{w,r}$).
Our plan is to apply Theorem 10 to $g_{\widehat{w}_0,r}$, with $\widehat{w}_0 \in \mathcal{B}_m \subseteq \mathcal{H}$ and $\|\widehat{w}_0\| \leqslant \|w_*\|$.

`: Talagrand`

**Corollary 3.** *Under the hypothesis of Theorem 10, for all $\tau > 0$ we have*

$$\begin{aligned} \sup_{w\in\mathcal{H}, \|w\|\leqslant\|w_*\|} \frac{\widehat{\mathbb{E}}h_w - \mathbb{E}h_w}{\lambda\|w\|^2 + \mathbb{E}h_w + r} <& 2\mathbb{E}_{D\sim\mathrm{P}^n} \sup_{w\in\mathcal{H}, \|w\|\leqslant\|w_*\|} \frac{\widehat{\mathbb{E}}h_w - \mathbb{E}h_w}{\lambda\|w\|^2 + \mathbb{E}h_w + r} \\ &+ \sqrt{\frac{3V(1+\tau)}{nr^{2-\theta}}} + 2GK_1\|w_*\| \frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\|X\|)(2+\tau)}{nr} \end{aligned} \tag{93}$$

*Proof.* In Theorem 10, we take

$$Z = \sup_{w\in\mathcal{H}, \|w\|\leqslant R} \left| \frac{1}{n} \sum_{i=1}^{n} g_{w,r}(X_i) \right| . \tag{94}$$

We have also that, using the second inequality of Lemma 7.1 in [Steinwart and Christmann, 2008] and taking $\theta > 0$, $q := \frac{2}{2-\theta}$, $q' := \frac{2}{\theta}$, $a := r$, and $b := \mathbb{E}h_w \neq 0$:

$$\mathbb{E}g_{w,r}^2 \leqslant \frac{\mathbb{E}h_w^2}{\left(\lambda\|w\|^2 + \mathbb{E}h_w + r\right)^2} \leqslant \frac{(2-\theta)^{2-\theta}\theta^\theta \mathbb{E}h_w^2}{4r^{2-\theta}\left(\mathbb{E}h_w\right)^\theta} \leqslant Vr^{\theta-2} = \sigma^2$$

Moreover,

$$
\begin{aligned}
&\left\| \sup_{w\in\mathcal{H},\|w\|\leqslant\|w_*\|} |g_{w,r}(X)| \right\|_{\psi_1} = \left\| \sup_{w\in\mathcal{H},\|w\|\leqslant\|w_*\|} \left| \frac{h_w(X)-\mathbb{E}h_w}{\lambda\|w\|^2+\mathbb{E}h_w+r} \right| \right\|_{\psi_1} \leqslant \frac{1}{r}\left\| \sup_{w\in\mathcal{H},\|w\|\leqslant\|w_*\|} |h_w-\mathbb{E}h_w(X)| \right\|_{\psi_1}\\
&= \frac{1}{r}\left\| \sup_{w\in\mathcal{H},\|w\|\leqslant\|w_*\|} |\ell(\langle w,X\rangle,Y)-\ell(\langle w_*,X\rangle,Y)-\mathbb{E}[\ell(\langle w,X\rangle,Y)-\ell(\langle w_*,X\rangle,Y)]| \right\|_{\psi_1}\\
&\leqslant \frac{1}{r}\left\| \sup_{w\in\mathcal{H},\|w\|\leqslant\|w_*\|} |\ell(\langle w,X\rangle,Y)-\ell(\langle w_*,X\rangle,Y)| + \sup_{w\in\mathcal{H},\|w\|\leqslant\|w_*\|} |\mathbb{E}[\ell(\langle w,X\rangle,Y)-\ell(\langle w_*,X\rangle,Y)]| \right\|_{\psi_1}\\
&\leqslant \frac{1}{r}\left\| G \sup_{w\in\mathcal{H},\|w\|\leqslant\|w_*\|} |\langle w-w_*,X\rangle| + G \sup_{w\in\mathcal{H},\|w\|\leqslant\|w_*\|} \mathbb{E}\,|\langle w-w_*,X\rangle| \right\|_{\psi_1}\\
&\leqslant \frac{1}{r}\Big\| 2G\|w_*\|\,\|X\| + 2G\|w_*\|\,\mathbb{E}\,\|X\| \Big\|_{\psi_1} = \frac{2G\|w_*\|}{r}\Big\| \|X\| + \mathbb{E}\,\|X\| \Big\|_{\psi_1} \leqslant \frac{2G\|w_*\|}{r}\Big\| \|X\| + \mathbb{E}\,\|X\| \Big\|_{\psi_2}\\
&\leqslant \frac{2G\|w_*\|\,(C\sqrt{2\mathrm{Tr}\Sigma}+\mathbb{E}\,\|X\|)}{r}
\end{aligned}
$$

where last inequality derives from the fact that $\|X\|$ is sub-Gaussian since, given an orthonormal basis $e_i$,

$$
\begin{aligned}
\big\| \|X\| \big\|_{\psi_2}^2 &\leqslant \big\| \|X\|^2 \big\|_{\psi_1} = \Big\| \sum_i \langle X,e_i\rangle^2 \Big\|_{\psi_1} \leqslant \sum_i \Big\| \langle X,e_i\rangle^2 \Big\|_{\psi_1}\\
&\leqslant 2\sum_i \|\langle X,e_i\rangle\|_{\psi_2}^2 \leqslant 2C^2\,\|\langle X,e_i\rangle\|_{L_2}^2 = 2C^2\,\mathrm{Tr}\,[\Sigma]
\end{aligned}
$$

Applying Theorem 10 with $\eta = 1$ we get the result. □

We now adapt Theorem 7.23 in [Steinwart and Christmann, 2008] to our setting:

ion thm 7.23

**Theorem 11.** *Under assumptions 1, 2, 4 and 3, the covariance matrix satisfies the polynomial decay condition (28), and the Bernstein conditions (37)–(38) hold true. Fix a closed subspace $\widehat{\mathcal{F}}$ of $\mathcal{H}$ and set*

$$
w_{\widehat{\mathcal{F}},\lambda} = \operatorname*{argmin}_{w\in\widehat{\mathcal{F}}} \left(\widehat{L}(w)+\lambda\|w\|^2\right) \qquad \lambda > 0. \tag{95}
$$

eq:34

*Choose $\widehat{w}_0 \in \widehat{\mathcal{F}}$, fix $\delta > 0$, then with probability at least $1-\delta$*

$$
\begin{aligned}
\lambda\|\widehat{w}_{\mathcal{F},\lambda}\|^2 + L(\widehat{w}^{cl}_{\mathcal{F},\lambda}) - L(f_*) \leqslant 7\left(\lambda\,\|\widehat{w}_0\|^2 + L(\widehat{w}_0) - L(f_*)\right) + K_3\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} +\\
+ 2\left(\frac{72V\log(3/\delta)}{n}\right)^{\frac{1}{2-\vartheta}} + 16GK_1\,\|w_*\|\,\frac{(C\sqrt{2\mathrm{Tr}\Sigma}+\mathbb{E}\,\|X\|)(2+\log(3/\delta))}{n}
\end{aligned} \tag{96}
$$

fast rate th

*where the constant $a$ only depends on (28) and $K_3 \geqslant 1$ only depends on $p, M, B, \vartheta$, and $V$.*

*Proof.* The proof mimics the one of Theorem 7.23 [Steinwart and Christmann, 2008], with some major differences.
We start recalling that Theorem 15 in [Steinwart et al., 2009] shows that that the decay condition (28) is equivalent to condition (7.48) of Theorem 7.23, which is given in terms of entropy numbers $e_j$, see Lemma 8. Note that the constant $a$ is defined by the bound (7.48). Using this remark, the above assumptions let us upper bound the empirical Rademacher complexity of $\mathcal{H}_r$ in term of a function $\varphi_n(r)$ defined as in [Steinwart and Christmann, 2008] (see pag. 267). Thus, the result comes from the application of Steinwart's Theorem 7.20, with the key difference that our $X$ is not bounded but sub-Gaussian and that $\widehat{w}_0$ here is not deterministic but depends on the data.

As a consequence, in order to control the quantity $\widehat{\mathbb{E}}h_{\widehat{w}_0} - \mathbb{E}h_{\widehat{w}_0}$ we cannot simply apply a Bernstein's inequality for sub-Gaussian but we need to use the more refined Corollary 3. In particular, we mimic the reasoning to derive [Steinwart and Christmann, 2008, eq. (7.44)], but where Talagrand's inequality for bounded random variables is replaced by our Theorem 10 for sub-exponential ones and in the specific case of Corollary 3. We split the error as in [Steinwart and Christmann, 2008, eq. (7.39)],

$$\lambda \left\|\widehat{w}_\lambda\right\|^2 + \mathbb{E}h_{\widehat{w}_\lambda^{cl}} \leqslant (\lambda \left\|\widehat{w}_0\right\|^2 + \mathbb{E}h_{\widehat{w}_0}) + (\widehat{\mathbb{E}}h_{\widehat{w}_0} - \mathbb{E}h_{\widehat{w}_0}) + (\mathbb{E}h_{\widehat{w}_\lambda^{cl}} - \widehat{\mathbb{E}}h_{\widehat{w}_\lambda^{cl}}) \tag{97}$$

eq: split 7

and we start with controlling the term $\widehat{\mathbb{E}}h_{\widehat{w}_0} - \mathbb{E}h_{\widehat{w}_0}$.
Exploiting the definition of $g_{w,r}$ in (92), we know that for all the $w \in \mathcal{H}$ with $\|w\| \leqslant \|w_*\|$ and $r > 0$ we can apply Corollary 3. In particular, since $\widehat{w}_0 \in \mathcal{B}_m \subseteq \mathcal{H}$, the bound in the Corollary is valid also for $\widehat{w}_0$, i.e

$$\begin{aligned}\frac{\widehat{\mathbb{E}}h_{\widehat{w}_0} - \mathbb{E}h_{\widehat{w}_0}}{\lambda \left\|\widehat{w}_0\right\|^2 + \mathbb{E}h_{\widehat{w}_0} + r} <& 2\mathbb{E}_{D\sim \mathrm{P}^n} \frac{\widehat{\mathbb{E}}h_{\widehat{w}_0} - \mathbb{E}h_{\widehat{w}_0}}{\lambda \left\|\widehat{w}_0\right\|^2 + \mathbb{E}h_{\widehat{w}_0} + r} \\ &+ \sqrt{\frac{3V(1+\tau)}{nr^{2-\theta}}} + 2GK_1 \left\|w_*\right\| \frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2+\tau)}{nr}.\end{aligned} \tag{98}$$

Using symmetrization (see Prop. 7.10 in [Steinwart and Christmann, 2008]) we have

$$\begin{aligned}\mathbb{E}_{D\sim \mathrm{P}^n} \sup_{w\in\mathcal{B}_{m,r}, \|w\|\leqslant\|w_*\|} \left|\widehat{\mathbb{E}}h_w - \mathbb{E}h_w\right| &\leqslant \mathbb{E}_{D\sim \mathrm{P}^n} \sup_{w\in\mathcal{H}_r, \|w\|\leqslant\|w_*\|} \left|\widehat{\mathbb{E}}h_w - \mathbb{E}h_w\right| \\ &\leqslant 2\mathbb{E}_{D\sim \mathrm{P}^n} \widehat{\mathrm{Rad}}(\mathcal{H}_r, n) \leqslant 2\varphi_n(r).\end{aligned} \tag{99}$$

Peeling by Steinwart's Theorem 7.7 together with $\mathcal{H}_r = \{w \in \mathcal{H} : \lambda \left\|w\right\|^2 + \mathbb{E}h_w \leqslant r\}$ hence gives

$$\mathbb{E}_{D\sim \mathrm{P}^n} \sup_{w\in\mathcal{B}_m, \|w\|\leqslant\|w_*\|} \left|\widehat{\mathbb{E}}g_{w,r}\right| \leqslant \mathbb{E}_{D\sim \mathrm{P}^n} \sup_{w\in\mathcal{H}, \|w\|\leqslant\|w_*\|} \left|\widehat{\mathbb{E}}g_{w,r}\right| \leqslant \frac{8\varphi_n(r)}{r} \tag{100}$$

Putting all together we get w.h.p.

$$\begin{aligned}\widehat{\mathbb{E}}h_{\widehat{w}_0} - \mathbb{E}h_{\widehat{w}_0} < (\lambda \left\|\widehat{w}_0\right\|^2 + \mathbb{E}h_{\widehat{w}_0}) &\left(\frac{10\varphi_n(r)}{r} + \sqrt{\frac{3V(1+\tau)}{nr^{2-\theta}}} + 2GK_1 \left\|w_*\right\| \frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2+\tau)}{nr}\right) \\ &+ 10\varphi_n(r) + \sqrt{\frac{3V(1+\tau)r^\theta}{n}} + 2GK_1 \left\|w_*\right\| \frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2+\tau)}{n}\end{aligned} \tag{101}$$

As regards the term $\mathbb{E}h_{w_\lambda^{cl}} - \widehat{\mathbb{E}}h_{w_\lambda^{cl}}$ we proceed as in [Steinwart and Christmann, 2008]. We finally obtain, for $\widehat{w}_0 \in \mathcal{B}_m$ with $\left\|\widehat{w}_0\right\| \leqslant \left\|w_*\right\|$ and with $r \geqslant r^*_{\mathcal{B}_m} \geqslant r^*_{\mathcal{H}}$, w.h.p.

$$\begin{aligned}\lambda \left\|\widehat{w}_\lambda\right\|^2 + \mathbb{E}h_{\widehat{w}_\lambda^{cl}} <& \left(\lambda \left\|\widehat{w}_0\right\|^2 + \mathbb{E}h_{\widehat{w}_0}\right) + \\ &+ (\lambda \left\|\widehat{w}_0\right\|^2 + \mathbb{E}h_{\widehat{w}_0}) \left(\frac{10\varphi_n(r)}{r} + \sqrt{\frac{3V(1+\tau)}{nr^{2-\theta}}} + 2GK_1 \left\|w_*\right\| \frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2+\tau)}{nr}\right) + \\ &+ 10\varphi_n(r) + \sqrt{\frac{3V(1+\tau)r^\theta}{n}} + 2GK_1 \left\|w_*\right\| \frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2+\tau)}{n} + \\ &+ \left(\lambda \left\|\widehat{w}_\lambda\right\|^2 + \mathbb{E}h_{\widehat{w}_\lambda^{cl}}\right) \left(\frac{10\varphi_n(r)}{r} + \sqrt{\frac{2V\tau}{nr^{2-\theta}}} + \frac{28B\tau}{3nr}\right) \\ &+ 10\varphi_n(r) + \sqrt{\frac{2V\tau r^\theta}{n}} + \frac{28B\tau}{3n}\end{aligned} \tag{102}$$

which replaces (7.44) in [Steinwart and Christmann, 2008].

Observe now that $r \geqslant 30\varphi_n(r)$ implies $10\varphi_n(r)r^{-1} \leqslant 1/3$ and $10\varphi_n(r) \leqslant r/3$. Moreover, $r \geqslant \left(\frac{72V(1+\tau)}{n}\right)^{1/(2-\theta)}$ yields

$$\left(\frac{2V\tau}{nr^{2-\theta}}\right)^{1/2} \leqslant \frac{1}{6} \quad \text{and} \quad \left(\frac{2V\tau r^\theta}{n}\right)^{1/2} \leqslant \frac{r}{6}$$

and

$$\left(\frac{3V(1+\tau)}{nr^{2-\theta}}\right)^{1/2} \leqslant \frac{1}{4} \quad \text{and} \quad \left(\frac{2V(1+\tau)r^{\theta}}{n}\right)^{1/2} \leqslant \frac{r}{4}$$

In addition $n \geqslant 72(1+\tau)$, $V \geqslant B^{2-\theta}$, and $r \geqslant \left(\frac{72V(1+\tau)}{n}\right)^{1/(2-\theta)}$ imply

$$\frac{28B\tau}{3nr} = \frac{7}{54}\cdot\frac{72\tau}{n}\cdot\frac{B}{r} \leqslant \frac{7}{54}\cdot\left(\frac{72\tau}{n}\right)^{\frac{1}{2-\theta}}\cdot\frac{V^{\frac{1}{2-\theta}}}{r} \leqslant \frac{7}{54}$$

and $\frac{28B\tau}{3n} \leqslant \frac{7r}{54}$. Finally $r \geqslant 8GK_1\|w_*\|\frac{(C\sqrt{2\mathrm{Tr}\Sigma}+\mathbb{E}\|X\|)(2+\tau)}{n}$ gives

$$2GK_1\|w_*\|\frac{(C\sqrt{2\mathrm{Tr}\Sigma}+\mathbb{E}\|X\|)(2+\tau)}{nr} \leqslant \frac{1}{4} \quad \text{and} \quad 2GK_1\|w_*\|\frac{(C\sqrt{2\mathrm{Tr}\Sigma}+\mathbb{E}\|X\|)(2+\tau)}{n} \leqslant \frac{r}{4}$$

We finally obtain

$$\begin{aligned}\lambda\|\widehat{w}_\lambda\|^2+\mathbb{E}h_{\widehat{w}_\lambda^{cl}} &< \frac{11}{6}\left(\lambda\|\widehat{w}_0\|^2+\mathbb{E}h_{\widehat{w}_0}\right)+\frac{79}{54}r+\epsilon+\frac{17}{27}\left(\lambda\|\widehat{w}_\lambda\|^2+\mathbb{E}h_{\widehat{w}_\lambda^{cl}}\right)\\ &\leqslant 5\left(\lambda\|\widehat{w}_0\|^2+\mathbb{E}h_{\widehat{w}_0}\right)+2r \qquad (103)\end{aligned}$$

with

$$r > \max\left\{30\varphi_n(r), \left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}}, 8GK_1\|w_*\|\frac{(C\sqrt{2\mathrm{Tr}\Sigma}+\mathbb{E}\|X\|)(2+\tau)}{n}, r^*_{\mathcal{H}}\right\}$$

□

remark: w_la *Remark* 4. Notice that the same reasoning can be applied in Section 5 in the more general framework where $w_*$ does not exist. In that case $w_*$ will be replaced by $w_\lambda := \arg\min_{w\in\mathcal{H}} L(w)+\lambda\|w\|^2$, with $\|w_\lambda\| \leqslant \sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}}$.

We are now ready to prove our main result:

*Proof of Theorem 6, polynomial decay.* Applying Theorem 11 in the general case of Remark 4, with the choice $\widehat{\mathcal{F}} = \mathcal{B}_m$ and $\widehat{w}_0 = \mathcal{P}_{\mathcal{B}_m}w_\lambda$, we rewrite (96) as:

$$\begin{aligned}\lambda\|\widehat{\beta}_{\lambda,m}\|^2+L(\widehat{\beta}^{cl}_{\lambda,m})-L(f_*) &\leqslant 7(\lambda\|\mathcal{P}_{\mathcal{B}_m}w_\lambda\|^2+L(\mathcal{P}_{\mathcal{B}_m}w_\lambda)-L(f_*))+K_3\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\theta+\theta p}}+2\Big(\frac{72V\log(3/\delta)}{n}\Big)^{\frac{1}{2-\theta}}+\\ &\quad+16GK_1\|w_\lambda\|\frac{(C\sqrt{2\mathrm{Tr}\Sigma}+\mathbb{E}\|X\|)(2+\log(3/\delta))}{n}\\ &=7(\lambda\|\mathcal{P}_{\mathcal{B}_m}w_\lambda\|^2+L(\mathcal{P}_{\mathcal{B}_m}w_\lambda)-L(w_\lambda)+L(w_\lambda)-L(f_*))+K_3\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\theta+\theta p}}+\\ &\quad+2\Big(\frac{72V\log(3/\delta)}{n}\Big)^{\frac{1}{2-\theta}}+16GK_1\frac{(C\sqrt{2\mathrm{Tr}\Sigma}+\mathbb{E}\|X\|)(2+\log(3/\delta))}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}}\\ &\leqslant 7(L(\mathcal{P}_{\mathcal{B}_m}w_\lambda)-L(w_\lambda)+\lambda\|w_\lambda\|^2+L(w_\lambda)-L(f_*))+K_3\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\theta+\theta p}}+\\ &\quad+2\Big(\frac{72V\log(3/\delta)}{n}\Big)^{\frac{1}{2-\theta}}+16GK_1\frac{(C\sqrt{2\mathrm{Tr}\Sigma}+\mathbb{E}\|X\|)(2+\log(3/\delta))}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}}\\ &=7\mathcal{A}(\lambda)+7(L(\mathcal{P}_{\mathcal{B}_m}w_\lambda)-L(w_\lambda))+K_3\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\theta+\theta p}}+2\Big(\frac{72V\log(3/\delta)}{n}\Big)^{\frac{1}{2-\theta}}+\\ &\quad+16GK_1\frac{(C\sqrt{2\mathrm{Tr}\Sigma}+\mathbb{E}\|X\|)(2+\log(3/\delta))}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} \qquad (104)\end{aligned}$$

eq:39

where we used the fact that $\|w_\lambda\| \leqslant \sqrt{\mathcal{A}(\lambda)/\lambda}$.
We can deal with the term $L(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L(w_\lambda)$ as in (75) (but where we use Lemma 4 instead of Lemma 7 in [Rudi et al., 2015] to exploit sub-Gaussianity), so that for $\alpha \gtrsim n^{-1/p}$ with probability greater than $1-\delta$

$$L(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L(w_\lambda) \leqslant K_2 G\sqrt{\alpha}\,\|w_\lambda\| \leqslant K_2 G\sqrt{\alpha}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} \tag{105}$$

`eq:40`

for some universal constant $K_2 > 0$. We finally obtain with probability greater than $1-2\delta$:

$$\begin{aligned}\lambda\|\widehat{\beta}_{\lambda,m}\|^2_{\mathcal{H}} + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(f_*) \leqslant 7\mathcal{A}(\lambda) + 7K_2 G\sqrt{\frac{\alpha\mathcal{A}(\lambda)}{\lambda}} + K_3\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\theta+\theta p}} + 2\Big(\frac{72V\log(3/\delta)}{n}\Big)^{\frac{1}{2-\theta}} + \\ + 16GK_1\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\,\|X\|)(2+\log(3/\delta))}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} \qquad (106)\end{aligned}$$

which proves the first claim. ☐

The following corollary provides the optimal rates.

`ptimal_rates`

**Corollary 4.** *Fix $\delta > 0$. Under the Theorem 6 and the source condition*

$$\mathcal{A}(\lambda) \leqslant A_0\lambda^r$$

*for some $r \in (0,1]$, set*

$$\lambda \asymp n^{-\min\{\frac{2}{r+1},\frac{1}{r(2-p-\theta+\theta p)+p}\}} \tag{107}$$

`eq:3a`

$$\alpha \asymp n^{-\min\{2,\frac{r+1}{r(2-p-\theta+\theta p)+p}\}} \tag{108}$$

`eq:3b`

$$m \gtrsim n^{\min\{2p,\frac{p(r+1)}{r(2-p-\theta+\theta p)+p}\}} \tag{109}$$

`eq:3c`

*with probability at least $1-2\delta$:*

$$\lambda\|\widehat{\beta}_{\lambda,m}\|^2 + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(f_*) \lesssim n^{-\min\{\frac{2r}{r+1},\frac{r}{r(2-p-\theta+\theta p)+p}\}} \tag{110}$$

*Proof.* Lemma 4 with Proposition 2 gives

$$m \gtrsim d_\alpha \log(n/\delta), \qquad d_\alpha \lesssim \alpha^{-p} \qquad \alpha \asymp \frac{\log^{1/p}(n/\delta)}{m^{1/p}} \tag{111}$$

Lemma A.1.7 in [Steinwart and Christmann, 2008] with $r=2$, $1/\gamma = (2-p-\theta+\theta p)$, $\alpha = p$, $\beta = r$ shows that the choice of $\lambda$, $\alpha$ and $m$ given by (107)–(109) provides the optimal rate. ☐

Notice that $\alpha \asymp n^{-\min\{2,\frac{r+1}{r(2-p-\theta+\theta p)+p}\}}$ is compatible with condition $\alpha \gtrsim d_\alpha \asymp n^{-1/p}$ in Lemma 4.

*Proof of Corollary 1.* The proof mimics the proof of Theorem 6 where in (96) we choose $\widehat{w}_0 = \mathcal{P}_{\mathcal{B}_m} w_*$ Hence (96) with $\theta = 1$ reads

$$\begin{aligned}\lambda\|\widehat{\beta}_{\lambda,m}\|^2 + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(w_*) &\leqslant 7(\lambda\|\mathcal{P}_{\mathcal{B}_m} w_*\|^2 + L(\mathcal{P}_{\mathcal{B}_m} w_*) - L(w_*)) + K_3\frac{a^{2p}}{\lambda^p n} + 144V\frac{\log(3/\delta)}{n} + \\ &\quad + 16GK_1\,\|w_*\|\,\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\,\|X\|)(2+\log(3/\delta))}{n} \\ &\leqslant 7\lambda\|w_*\|^2 + 7(L(\mathcal{P}_{\mathcal{B}_m} w_*) - L(w_*)) + K_3\frac{a^{2p}}{\lambda^p n} + 144V\frac{\log(3/\delta)}{n} + \\ &\quad + 16GK_1\,\|w_*\|\,\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\,\|X\|)(2+\log(3/\delta))}{n} \qquad (112)\end{aligned}$$

We can deal wit h the term $L(\mathcal{P}_{\mathcal{B}_m} w_*) - L(w_*)$ as in (75), so that for $\alpha \gtrsim n^{-1/p}$ with probability greater than $1-\delta$

$$L(\mathcal{P}_{\mathcal{B}_m} w_*) - L(w_*) \leqslant K_2 G\sqrt{\alpha}\,\|w_*\|$$

for some $K_2 > 0$. Hence, with probability at least $1 - 2\delta$:

$$\lambda\|\widehat{\beta}_{\lambda,m}\|^2 + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(w_*) \leqslant 7\lambda\|w_*\|^2 + 7K_2G\sqrt{\alpha}\|w_*\| + K_3\frac{a^{2p}}{\lambda^p n} + 144V\frac{\log(3/\delta)}{n} + \\ 16GK_1\left\|w_*\right\|\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2 + \log(3/\delta))}{n} \tag{113}$$

which proves the claim. ☐

The following corollary provides the optimal rates, whose proof is the same as for Corollary 4

`cor:optimal` **Corollary 5.** *Fix $\delta > 0$. Under the Theorem 1 set*

$$\lambda \asymp n^{-\frac{1}{1+p}} \tag{114}$$
`eq:4a`

$$\alpha \asymp n^{-\frac{2}{1+p}} \tag{115}$$
`eq:4b`

$$m \gtrsim n^{\frac{2p}{1+p}}\log n \tag{116}$$
`eq:4c`

*then, for ALS sampling, with probability at least $1 - 2\delta$:*

$$\lambda\|\widehat{\beta}_{\lambda,m}\|^2_{\mathcal{H}} + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(w_*) \lesssim \|w_*\|\left(\frac{1}{n}\right)^{\frac{1}{1+p}} \tag{117}$$

Notice that $\alpha \asymp n^{-\frac{2}{1+p}}$ is compatible with condition $\alpha \gtrsim d_\alpha \asymp n^{-1/p}$ in Lemma 4.

### C.1 Excess risk under exponential decay

As regards exponential decay, given the discussion in Appendix E, we have a different bound on the empirical Rademacher complexity of $\mathcal{H}_r$. In particular, we obtain $\varphi_n(r) := C_1\sqrt{\frac{V}{n}}\log_2\left(\frac{1}{\lambda}\right)\sqrt{r} + C_2\frac{\log_2^2(1/\lambda)}{n}$ and we modify Theorem 11 in the case of exponential decay using the following Lemma:

`n:  exp decay` **Lemma 5.** *When*

$$r = C_3\frac{\log_2^2(1/\lambda)}{n} + \left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + 8GK_1\left\|w_*\right\|\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2 + \tau)}{n}$$

*we have*

$$r \geqslant \max\left\{30\varphi_n(r), \left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}}, 8GK_1\left\|w_*\right\|\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2 + \tau)}{n}\right\}$$

We can finally prove the second part of Theorem 6 under exponential decay:

*Proof of Theorem 6, exponential decay.* We follow exactly the proof of Theorem 11 for polynomial decay presented above in the previous subsection, but using the estimate in Lemma 5 for $r$:

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

☐

## D Proofs of Section 6

### D.1 Square loss

`app:  square`

We report in this section the proofs of Theorem 7.
As mentioned above, in the case where $w_*$ does not exists, the assumption of sub-Gaussianity is necessary to get fast rates:

*Proof of Theorem 7.* The proof follows the one of Theorem 6 in Appendix C with some differences coming from the fact that we are working now with the square loss. Since Theorem 11 works also with locally Lipschitz loss functions we have:

$$\begin{aligned}\lambda\|\widehat{\beta}_{\lambda,m}\|^2 + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(f_*) &\leqslant 7(\lambda\|\mathcal{P}_{\mathcal{B}_m} w_\lambda\|^2 + L(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L(f_*)) + K_3\frac{a^{2p}}{\lambda^p n} + 2\frac{72V\log(3/\delta)}{n} + \\ &+ 16GK_1\left\|w_\lambda\right\| \frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2+\log(3/\delta))}{n} \\ &= 7(L_\lambda(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L_\lambda(w_\lambda) + L_\lambda(w_\lambda) - L(f_*)) + K_3\frac{a^{2p}}{\lambda^p n} + \\ &+ 2\frac{72V\log(3/\delta)}{n} + 16GK_1\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2+\log(3/\delta))}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} \\ &= 7\mathcal{A}(\lambda) + 7(L_\lambda(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L_\lambda(w_\lambda)) + K_3\frac{a^{2p}}{\lambda^p n} + 2\frac{72V\log(3/\delta)}{n} + \\ &+ 16GK_1\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2+\log(3/\delta))}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}}\end{aligned} \tag{118}$$

Using the fact that $L_\lambda$ is quadratic and expanding around the the minimum $w_\lambda$ we have

$$L_\lambda(\mathcal{P}_m w_\lambda) - L_\lambda(w_\lambda) = \|(\Sigma+\alpha)^{1/2}(I-\mathcal{P}_m)w_\lambda\|^2 \tag{119}$$

Using Lemma 4 we get the result

$$\begin{aligned}\lambda\|\widehat{\beta}_{\lambda,m}\|^2 + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(f_*) &\leqslant 7\mathcal{A}(\lambda) + 7\|(\Sigma+\alpha)^{1/2}(I-\mathcal{P}_m)w_\lambda\|^2 + K_3\frac{a^{2p}}{\lambda^p n} + 2\frac{72V\log(3/\delta)}{n} + \\ &+ 16GK_1\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2+\log(3/\delta))}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} \\ &\lesssim 7\mathcal{A}(\lambda) + 7\alpha\frac{\mathcal{A}(\lambda)}{\lambda} + K_3\frac{a^{2p}}{\lambda^p n} + 2\frac{72V\log(3/\delta)}{n} + \\ &+ 16GK_1\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2+\log(3/\delta))}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}}\end{aligned} \tag{120}$$

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

$$\begin{aligned}&\lambda \asymp n^{-\min\{\frac{2}{r+1},\frac{1}{r+p}\}} \\ &\alpha \asymp n^{-\min\{\frac{2}{r+1},\frac{1}{r+p}\}} \\ &m \gtrsim n^{\min\{\frac{2p}{r+1},\frac{p}{r+p}\}}\log n\end{aligned}$$

with high probability

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

□

## D.2 Logistic Loss

`p:  logistic`

Since logistic loss is not clippable, we prove how the modification of the definition of the clipping in (49) and the similar treatment of the projection term, up to constants, between square and logistic losses asymptotically lead to the same excess risk bounds. We start adjusting the proof of Theorem 11.
As explained in subsection 6.2, let's note that we have $h_f(X) - h_f^{cl}(X) + \frac{1}{n} \geqslant 0$. Therefore we can simply rewrite the splitting of the error (97) as

$$\lambda\left\|\widehat{w}_\lambda\right\|^2 + \mathbb{E}h_{\widehat{w}^{cl}_\lambda} \leqslant (\lambda\left\|\widehat{w}_0\right\|^2 + \mathbb{E}h_{\widehat{w}_0}) + (\widehat{\mathbb{E}}h_{\widehat{w}_0} - \mathbb{E}h_{\widehat{w}_0}) + (\mathbb{E}h_{\widehat{w}^{cl}_\lambda} - \widehat{\mathbb{E}}h_{\widehat{w}^{cl}_\lambda}) + \frac{1}{n}. \tag{121}$$

Clearly last term $1/n$ does not spoil the rate and we can proceed as for square loss:

$$\begin{aligned}
\lambda\|\widehat{\beta}_{\lambda,m}\|^2 + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(f_*) &\leqslant 7(\lambda\|\mathcal{P}_{\mathcal{B}_m} w_\lambda\|^2 + L(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L(f_*)) + K_3\frac{a^{2p}}{\lambda^p n} + \frac{144V\log(3/\delta)}{n} + \\
&\quad + 16GK_1\left\|w_\lambda\right\|\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2+\log(3/\delta))}{n} + \frac{1}{n} \\
&= 7(L_\lambda(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L_\lambda(w_\lambda) + L_\lambda(w_\lambda) - L(f_*)) + K_3\frac{a^{2p}}{\lambda^p n} + \frac{144V\log(3/\delta)}{n} + \\
&\quad + 16GK_1\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2+\log(3/\delta))}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} + \frac{1}{n} \\
&= 7\mathcal{A}(\lambda) + 7(L_\lambda(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L_\lambda(w_\lambda)) + K_3\frac{a^{2p}}{\lambda^p n} + \frac{144V\log(3/\delta)}{n} + \\
&\quad + 16GK_1\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2+\log(3/\delta))}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} + \frac{1}{n} \qquad (122)
\end{aligned}$$

To deal with the projection term $L_\lambda(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L_\lambda(w_\lambda)$ we do a Taylor expansion

$$L_\lambda(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L_\lambda(w_\lambda) = \frac{1}{2}\langle (HL)(w')(\mathcal{P}_{\mathcal{B}_m} w_\lambda - w_\lambda), (\mathcal{P}_{\mathcal{B}_m} w_\lambda - w_\lambda)\rangle \qquad (123)$$

eq: taylor

where $w' = w_\lambda + t(\mathcal{P}_{\mathcal{B}_m} w_\lambda - w_\lambda)$ with $t \in [0,1]$ and using the fact that $\nabla L_\lambda(w_\lambda) = 0$. We can find the expression of the Hessian $H$ of $L$ in $w \in \mathcal{H}$ exploiting its definition

$$\begin{aligned}
\langle (HL)(w)v, v\rangle &= \frac{d^2}{dt^2}L(w+tv)|_{t=0} = \frac{d}{dt}\mathbb{E}\left[\ell'(\langle w+tv, X\rangle, Y)\langle v, X\rangle\right]|_{t=0} \\
&= \mathbb{E}\left[\ell''(\langle w+tv, X\rangle, Y)(\langle v, X\rangle)^2\right]|_{t=0} \leqslant M\mathbb{E}\left[\langle v, X\rangle^2\right] \qquad (124)
\end{aligned}$$

eq: hessian

where $M = \sup_{\tau\in\mathbb{R}, y\in\mathcal{Y}}\ell''(\tau, y)$ and $v \in \mathcal{H}$. For the logistic loss we have

$$\ell''(\tau, y) = \sigma(y\tau)(1-\sigma(y\tau)) \leqslant \frac{1}{4}, \qquad \forall\tau\in\mathbb{R}, y\in\mathcal{Y}$$

where $\sigma(\cdot)$ is the sigmoid which is upper bounded by 1. So combining this result with (124) and considering $L_\lambda(\cdot) = L(\cdot) + \lambda\left\|\cdot\right\|^2$ we get

$$(HL_\lambda)(w) \leqslant \Sigma_\lambda.$$

Finally we can rewrite (123) as

$$L_\lambda(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L_\lambda(w_\lambda) \leqslant \frac{1}{2}\left\|\Sigma_\lambda^{1/2}(\mathcal{P}_{\mathcal{B}_m} w_\lambda - w_\lambda)\right\|^2 \qquad (125)$$

and proceed exactly as in the case of the square loss (see appendix D.1).

# E Entropy Numbers and Exponential Decay

pp: entropy

We analyse here the main steps needed to obtain the results for exponential decay in Theorem 3 and Theorem 6.

## E.1 Entropy numbers in Hilbert spaces

Let $\mathcal{H}$ and $\mathcal{K}$ be real Hilbert spaces. For all $n \in \mathbb{N}, n \geqslant 1$

$$\sup_{1\leqslant k<\infty}\left(n^{-1/k}\left(\Pi_{\ell=1}^k a_\ell(T)\right)^{1/k}\right) \leqslant \varepsilon_n(T) \leqslant 14\sup_{1\leqslant k<\infty}\left(n^{-1/k}\left(\Pi_{\ell=1}^k a_\ell(T)\right)^{1/k}\right) \qquad (126)$$

eq: entropy

where $\varepsilon_n(T)$ are the entropy numbers, see (3.4.15) of [Carl and Stephani, 1990].
Let $X$ be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking value in a real Hilbert space $\mathcal{H}$ such that $\mathbb{E}\left[|\langle X, v\rangle|^2\right]$ is finite for all $v \in \mathcal{H}$. Define

$$T : \mathcal{H} \to L_2(\Omega, \mathbb{P}) \quad T(v)(\omega) = \langle X(\omega), v\rangle$$

so that $\Sigma = T^*T$ is (non-centered) covariance matrix. We assume that $\Sigma$ is a trace-class operator and the corresponding eigenvalues have an exponential decay

$$\Sigma = \sum_{n=1}^{+\infty} \lambda_n(\Sigma) v_n \otimes v_n \quad \lambda_n(\Sigma) \simeq 2^{-2an}$$

where $(v_n)_n$ is a base of $\mathcal{H}$. Since $\Sigma$ is trace-class, $S$ is compact, so that by (126)

$$e_n(T) \simeq \sup_{1 \leqslant k < \infty} 2^{-(n-1)/k} \left(\Pi_{\ell=1}^{k} a_n(T)\right)^{1/k}$$

with $e_n(T) = \varepsilon_{2^{n-1}}(T)$ the (dyadic) entropy numbers and where by [Carl and Stephani, 1990]

$$a_n(T) = a_n(|T|) = \lambda_n(|T|) = \lambda_n(\Sigma)^{1/2} \simeq 2^{-an}.$$

We have

$$2^{-(n-1)/k} \left(\Pi_{\ell=1}^{k} 2^{-a\ell}\right)^{1/k} = 2^{-\left(\frac{n-1}{k} + \frac{a(k+1)}{2}\right)}.$$

Observe that the minimum on $(0, +\infty)$ of the function

$$f(x) = \left(\frac{n-1}{x} + \frac{ax}{2}\right)$$

is $f(\sqrt{2(n-1)/a}) = \sqrt{2a(n-1)}$, then

$$e_n(T) \simeq 2^{-\sqrt{an}}.$$

## E.2 Entropy numbers of $\mathcal{F}_r$

Given the above calculation we want to upper bound the entropy number of $\mathcal{F}_r$, we recall here some definitions:

$$\mathcal{H}_r := \left\{f \in \mathcal{H} : \Upsilon(f) + L(f^{cl}) - L(f_*) \leqslant r\right\} \qquad r > r^*$$

$$\mathcal{F}_r := \left\{\ell \circ f^{cl} - \ell \circ f_* : f \in \mathcal{H}_r\right\} \qquad r > r^*$$

Using the above discussion we obtain

$$e_i(\mathcal{F}_r) \leqslant G e_i(\mathcal{H}_r) \leqslant G\sqrt{\frac{r}{\lambda}} e_i(\mathcal{B}_{\mathcal{H}}) = G\sqrt{\frac{r}{\lambda}} 2^{-c\sqrt{i}}$$

## E.3 Bound the Rademacher Complexity of $\mathcal{F}_r$

Now we are ready to upper bound the empirical Rademacher Complexity $\widehat{\mathfrak{R}}$ of $\mathcal{F}_r$:

**Lemma 6.**

$$\widehat{\mathfrak{R}}\left(\mathcal{F}_r\right) \leqslant \sqrt{\frac{\log 16}{n}} \log\left(\frac{1}{\lambda}\right)(3\rho + 2c_3\sqrt{r}) \tag{127}$$

*where* $\rho = \sup_{f \in \mathcal{F}_r} \|f\|_{L_2(D)}$ *and* $\|f\|_{L_2(D)} := \left(\frac{1}{m}\sum_i f^2\left(x_i\right)\right)^{1/2}$.

*Proof.* Using Theorem 7.13 in [Steinwart and Christmann, 2008], we have

$$\widehat{\mathfrak{R}}\left(\mathcal{F}_r\right) \leqslant \sqrt{\frac{\log 16}{n}}\left(\sum_{i=1}^{\infty} 2^{i/2} e_{2^i}\left(\mathcal{F}_r \cup \{0\}, \|\cdot\|_{L_2(D)}\right) + \sup_{f \in \mathcal{F}_r} \|f\|_{L_2(D)}\right)$$

It is easy to see that $e_i\left(\mathcal{F}_r \cup \{0\}\right) \leqslant e_{i-1}\left(\mathcal{F}_r\right)$ and $e_0\left(\mathcal{F}_r\right) \leqslant \sup_{f \in \mathcal{F}_r} \|f\|_{L_2(D)}$. Since $e_i\left(\mathcal{F}_r\right)$ is a decreasing sequence with respect to $i$, together with the lemma above, we know that

$$e_i\left(\mathcal{F}_r\right) \leqslant \min\left\{\sup_{f \in \mathcal{F}_r} \|f\|_{L_2(D)}, \sqrt{\frac{2r}{\lambda}} 2^{-c\sqrt{i}}\right\}$$

Even though the second one decays exponentially, it may be much greater than the first term when $2r/\lambda$ is huge for small $i$ s. To achieve the balance between these two bounds, we use the first one for first $T$ terms in the sum and the second one for the tail. So

$$\widehat{\mathfrak{R}}\left(\mathcal{F}_r\right) \leqslant \sqrt{\frac{\log 16}{n}}\left(\sup_{f \in \mathcal{F}_r} \|f\|_{L_2(D)} \sum_{i=0}^{T-1} 2^{i/2} + \sqrt{\frac{2r}{\lambda}} \sum_{i=T}^{\infty} 2^{i/2} 2^{-c\sqrt{2^i-1}}\right)$$

The first sum is $\frac{\sqrt{2}^T-1}{\sqrt{2}-1}$. When $T$ is large enough, the second sum is upper bounded by the integral

$$\begin{aligned} \int_T^{\infty} 2^{x/2} 2^{-c\sqrt{2^i-1}} \,\mathrm{d}x \leqslant \int_T^{\infty} 2^{x/2} 2^{-c_2\sqrt{2^i}} \,\mathrm{d}x &\leqslant \frac{2^{-c_2\sqrt{2^T}+1}}{c_2 \log^2(2)} && (128) \\ &\leqslant c_3 2^{-c_2\sqrt{2^T}} && (129) \end{aligned}$$

To make the form simpler, we bound $\frac{\sqrt{2}^T-1}{\sqrt{2}-1}$ by $3 \cdot 2^{T/2}$, and denote $\sup_{h \in \mathcal{F}_r} \|h\|_{L_2(D)}$ by $\rho$. Taking $T$ to be

$$\log_2\left(c_4^2 \log_2^2\left(\frac{1}{\lambda}\right)\right),$$

with $c_4$ such that $c_2 c_4 > 1/2$, we get the upper bound of the form

$$\widehat{\mathfrak{R}}\left(\mathcal{F}_r\right) \leqslant \sqrt{\frac{\log 16}{n}}\left(3\rho \log\left(\frac{1}{\lambda}\right) + c_3\sqrt{\frac{2r}{\lambda}} \lambda^{c_2 c_4}\right) \leqslant \sqrt{\frac{\log 16}{n}} \log\left(\frac{1}{\lambda}\right)\left(3\rho + 2c_3\sqrt{r}\right)$$

□

Now we can directly compute the upper bound for the population Rademacher Complexity $\mathfrak{R}\left(\mathcal{F}_r\right)$ by taking expectation over $D \sim P^m$:

**Lemma 7.**

$$\mathfrak{R}\left(\mathcal{F}_r\right) \leqslant C_1\sqrt{\frac{V}{n}} \log_2\left(\frac{1}{\lambda}\right)\sqrt{r} + C_2 \frac{\log_2^2(1/\lambda)}{n} \tag{130}$$

*where $C_1$ and $C_2$ are two absolute constants.*

*Proof.*

$$\mathfrak{R}\left(\mathcal{F}_r\right) = \mathbb{E}[\widehat{\mathfrak{R}}\left(\mathcal{F}_r\right)] \leqslant \sqrt{\frac{(\log 16)}{n}} \log_2\left(\frac{1}{\lambda}\right)\left(3\mathbb{E} \sup_{f \in \mathcal{F}_r} \|f\|_{L_2(D)} + 2c_3\sqrt{r}\right) \tag{131}$$

By Jensen's inequality and Corollary A.8.5 in [Steinwart and Christmann, 2008], we have

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}_r} \|f\|_{L_2(D)} \leqslant \left(\mathbb{E} \sup_{f \in \mathcal{F}_r} \|f\|_{L_2(D)}^2\right)^{1/2} &\leqslant \left(\mathbb{E} \sup_{f \in \mathcal{F}_r} \frac{1}{m} \sum_{i=1}^{m} f^2\left(x_i, y_i\right)\right)^{1/2} \\ &\leqslant \left(\sigma^2 + 8\mathfrak{R}\left(\mathcal{F}_r\right)\right)^{1/2} \end{aligned}$$

where $\sigma^2 := \mathbb{E}f^2$. When $\sigma^2 > \mathfrak{R}(\mathcal{F}_r)$, we have

$$\begin{aligned}
\mathfrak{R}(\mathcal{F}_r) &\leqslant \sqrt{\frac{\log 16}{n}} \log_2\left(\frac{1}{\lambda}\right)(9\sigma + 2c_3\sqrt{r}) && (132)\\
&\leqslant \sqrt{\frac{\log 16}{n}} \log_2\left(\frac{1}{\lambda}\right)(9\sqrt{Vr^\theta} + 2c_3\sqrt{r}) && (133)\\
&\leqslant c_5\sqrt{\frac{V}{n}} \log_2\left(\frac{1}{\lambda}\right)\sqrt{r} && (134)
\end{aligned}$$

The second inequality is because $\mathbb{E}f^2 \leqslant V(\mathbb{E}f)^\theta$ and $\mathbb{E}f \leqslant r$ for $f \in \mathcal{F}_r$. When $\sigma^2 \leqslant \mathfrak{R}(\mathcal{F}_r)$, we have

$$\begin{aligned}
\mathfrak{R}(\mathcal{F}_r) &\leqslant \sqrt{\frac{\log 16}{n}} \log_2\left(\frac{1}{\lambda}\right)\left(9\sqrt{\mathfrak{R}(\mathcal{F}_r)} + 2c_3\sqrt{r}\right)\\
&\leqslant (9+2c_3)c_3\sqrt{\frac{\log 16}{n}} \log_2\left(\frac{1}{\lambda}\right)\sqrt{r} + (9+2c_3)^2\frac{(\log 16)\log_2^2(1/\lambda)}{n}
\end{aligned}$$

The last inequality can be obtained by dividing the formula into two cases, either $\mathfrak{R}(\mathcal{F}_r) < r$ or $\mathfrak{R}(\mathcal{F}_r) \geqslant r$ and then take the sum of the upper bounds of two cases. Combining all these inequalities, we finally obtain an upper bound

$$\mathfrak{R}(\mathcal{F}_r) \leqslant C_1\sqrt{\frac{V}{n}} \log_2\left(\frac{1}{\lambda}\right)\sqrt{r} + C_2\frac{\log_2^2(1/\lambda)}{n}$$

where $C_1$ and $C_2$ are two absolute constants. □

# F Known results

nown results

For sake of completeness we recall the following known results, we freely use in the paper.

The following two results provide a tight bound on the effecticbe dimension under the assumption of a polynomial decay or an exponential decay of the eigenvalues $\sigma_j$ of $\Sigma$ from [Caponnetto and De Vito, 2007]. We report the proofs for sake of completeness.

olynom decay

**Proposition 2** (Proposition 3 in [Caponnetto and De Vito, 2007])**.** *If for some* $\gamma \in \mathbb{R}^+$ *and* $1 < \beta < +\infty$

$$\sigma_i \leqslant \gamma i^{-\beta}$$

*then*

$$d_\alpha \leqslant \gamma\frac{\beta}{\beta - 1}\alpha^{-1/\beta} \tag{135}$$

*Proof.* Since the function $\sigma/(\sigma+\alpha)$ is increasing in $\sigma$ and using the spectral theorem $\Sigma = UDU^*$ combined with the fact that $\mathrm{Tr}(UDU^*) = \mathrm{Tr}(U(U^*D)) = \mathrm{Tr}D$

$$d_\alpha = \mathrm{Tr}(\Sigma(\Sigma+\alpha I)^{-1}) = \sum_{i=1}^{\infty}\frac{\sigma_i}{\sigma_i+\alpha} \leqslant \sum_{i=1}^{\infty}\frac{\gamma}{\gamma + i^\beta\alpha} \tag{136}$$

The function $\gamma/(\gamma + x^\beta\alpha)$ is positive and decreasing, so

$$\begin{aligned}
d_\alpha &\leqslant \int_0^\infty \frac{\gamma}{\gamma + x^\beta\alpha}dx\\
&= \alpha^{-1/\beta}\int_0^\infty \frac{\gamma}{\gamma+\tau^\beta}d\tau\\
&\leqslant \gamma\frac{\beta}{\beta-1}\alpha^{-1/\beta} && (137)
\end{aligned}$$

since $\int_0^\infty(\gamma+\tau^\beta)^{-1} \leqslant \beta/(\beta-1)$. □

exp decay **Proposition 3** (Exponential eigenvalues decay)**.**
*If for some* $\gamma, \beta \in \mathbb{R}^+ \sigma_i \leqslant \gamma e^{-\beta i}$ *then*

$$d_\alpha \leqslant \frac{\log(1+\gamma/\alpha)}{\beta} \tag{138}$$

*Proof.*

$$d_\alpha = \sum_{i=1}^{\infty} \frac{\sigma_i}{\sigma_i + \alpha} = \sum_{i=1}^{\infty} \frac{1}{1+\alpha/\sigma_i} \leqslant \sum_{i=1}^{\infty} \frac{1}{1+\alpha' e^{\beta i}} \leqslant \int_0^{+\infty} \frac{1}{1+\alpha' e^{\beta x}} dx \tag{139}$$

exp decay

where $\alpha' = \alpha/\gamma$. Using the change of variables $t = e^{\beta x}$ we get

$$\begin{aligned}(139) &= \frac{1}{\beta}\int_1^{+\infty} \frac{1}{1+\alpha' t}\frac{1}{t} dt = \frac{1}{\beta}\int_1^{+\infty}\left[\frac{1}{t} - \frac{\alpha'}{1+\alpha' t}\right] dt = \frac{1}{\beta}\Big[\log t - \log(1+\alpha' t)\Big]_1^{+\infty} \\ &= \frac{1}{\beta}\left[\log\left(\frac{t}{1+\alpha' t}\right)\right]_1^{+\infty} = \frac{1}{\beta}\Big[\log(1/\alpha') + \log(1+\alpha')\Big] \end{aligned} \tag{140}$$

So we finally obtain

$$d_\alpha \leqslant \frac{1}{\beta}\Big[\log(\gamma/\alpha) + \log(1+\alpha/\gamma)\Big] = \frac{\log(1+\gamma/\alpha)}{\beta} \tag{141}$$

□

The following result provides a bound on the entropy number and it is the content of Theorem 15 in [Steinwart et al., 2009]. We recall that, given a bounded operator $A$ between two Hilbert spaces $\mathcal{H}_1$ and $H_2$, we denote by $e_j(A)$ the (dyadic) entropy numbers of $A$ and by $\widehat{P}_{\mathcal{H}} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ the empirical (marginal) measure associated with the input data $x_i, \ldots, x_n$. Regard the data matrix $\widehat{X}$ as the inclusion operator $\mathrm{id} : \mathcal{H} \to L_2(\widehat{P})$

$$(\mathrm{id}\, w)(x_i) = \langle w, x_i \rangle \qquad i = 1, \ldots, n$$

entropy **Lemma 8.** *Let* $p \in (0,1)$*. Then*

$$\mathbb{E}_{\widehat{P}}[e_j(\mathrm{id} : \mathcal{H} \to L_2(\widehat{P}))] \sim j^{-\frac{1}{2p}} \tag{142}$$

*if and only if*

$$\sigma_j \sim j^{-\frac{1}{p}} \tag{143}$$

decad lemma

As regard results in Section 7, from [Bartlett et al., 2006] we report the following lemma:

m0-1_to_surr **Lemma 9.** *For any nonnegative loss function* $\phi$*, any measurable* $f : \mathcal{H} \to \mathbb{R}$*, and any probability distribution on* $\mathcal{H} \times \{\pm 1\}$

$$\psi\left(L_{0-1}(f) - L_{0-1}^*\right) \leqslant L_\phi(f) - L_\phi^*.$$

*In particular, for square, hinge and logistic losses we can write*

- *square loss:* $L_{0-1}(f) - L_{0-1}^* \leqslant \sqrt{L_{square}(f) - L_{square}^*}$*,*
- *hinge loss:* $L_{0-1}(f) - L_{0-1}^* \leqslant L_{hinge}(f) - L_{hinge}^*$*,*
- *logistic loss:* $L_{0-1}(f) - L_{0-1}^* \leqslant 2\sqrt{L_{logistic}(f) - L_{logistic}^*}$*.*

Under the assumption of low noise we can improve the above bounds in Lemma 9:

k wirh gamma **Lemma 10** (Theorem 3 in [Bartlett et al., 2006])**.** *Suppose that* $P$ *has noise exponent* $0 \leqslant \gamma \leqslant 1$*, and that* $\phi$ *is classification-calibrated (which is the case for square, hinge and logistic losses). Then there is a* $c > 0$ *such that for any* $f : \mathcal{X} \to \mathbb{R}$

$$c\left(L_{0-1}(f) - L_{0-1}^*\right)^\gamma \psi\left(\frac{\left(L_{0-1}(f) - L_{0-1}^*\right)^{1-\gamma}}{2c}\right) \leqslant L_\phi(f) - L_\phi^*$$

*where* $\psi(x) = x^2$ *when* $\phi$ *is the square loss,* $\psi(x) = x$ *when* $\phi$ *is the hinge loss and* $\psi(x) \geqslant \frac{x}{2}$ *when* $\phi$ *is the logistic loss.*

We copy also this results from [Steinwart and Christmann, 2008], linking the variance bound in Assumption 7 with low noise condition in Assumption 8 for hinge loss:

noise_bernst

**Lemma 11.** *[Theorem 8.24 [Steinwart and Christmann, 2008]] (Variance bound for the hinge loss). Let* P *be a distribution on* $X \times Y$ *that has noise exponent* $\gamma \in [0,1]$. *Moreover, let* $f_* : X \to [-1,1]$ *be a fixed Bayes decision function for the hinge loss* $\ell$. *Then, for all measurable* $f : X \to \mathbb{R}$, *we have*

$$\mathbb{E}\left(\ell \circ f^{cl} - \ell \circ f_*\right)^2 \leqslant 6c\left(\mathbb{E}\left(\ell \circ f^{cl} - \ell \circ f_*\right)\right)^\gamma$$

*where* $c$ *is the constant appearing in* (50).

## G Experiments: datasets and tuning

appexp

Here we report further information on the used data sets and the set up used for parameter tuning.
For Nyström SVM with Pegaos we tuned the kernel parameter $\sigma$ and $\lambda$ regularizer with a simple grid search ($\sigma \in [0.1, 20]$, $\lambda \in [10^{-8}, 10^{-1}]$, initially with a coarse grid and then more refined around the best candidates). An analogous procedure has been used for K-SVM with its parameters $C$ and $\gamma$. The details of the considered data sets and the chosen parameters for our algorithm in Table 5 and 6 are the following:
**SUSY** (Table 5 and 6, $n = 5 \times 10^6$, $d = 18$): we used a Gaussian kernel with $\sigma = 4$, $\lambda = 3 \times 10^{-6}$ and $m_{ALS} = 2500$, $m_{uniform} = 2500$.
**Mnist binary** (Table 5 and 6, $n = 7 \times 10^4$, $d = 784$): we used a Gaussian kernel with $\sigma = 10$, $\lambda = 3 \times 10^{-6}$ and $m_{ALS} = 15000$, $m_{uniform} = 20000$.
**Usps** (Table 5 and 6, $n = 9298$, $d = 256$): we used a Gaussian kernel with $\sigma = 10$, $\lambda = 5 \times 10^{-6}$ and $m_{ALS} = 2500$, $m_{uniform} = 4000$.
**Webspam** (Table 5 and 6, $n = 3.5 \times 10^5$, $d = 254$): we used a Gaussian kernel with $\sigma = 0.25$, $\lambda = 8 \times 10^{-7}$ and $m_{ALS} = 11500$, $m_{uniform} = 20000$.
**a9a** (Table 5 and 6, $n = 48842$, $d = 123$): we used a Gaussian kernel with $\sigma = 10$, $\lambda = 1 \times 10^{-5}$ and $m_{ALS} = 800$, $m_{uniform} = 1500$.
**CIFAR** (Table 5 and 6, $n = 6 \times 10^4$, $d = 400$): we used a Gaussian kernel with $\sigma = 10$, $\lambda = 2 \times 10^{-6}$ and $m_{ALS} = 20500$, $m_{uniform} = 20000$.

Table 6: Comparison between ALS and uniform sampling. To achieve similar accuracy, uniform sampling usually requires larger $m$ than ALS sampling. Therefore, even if it does not need leverage scores computations, Nyström-Pegasos with uniform sampling can be more expensive both in terms of memory and time (in seconds).

tab:comparis

| | Nyström-Pegasos (ALS) | | | Nyström-Pegasos (Uniform) | | |
|---|---|---|---|---|---|---|
| Datasets | c-err | t train | t pred | c-err | t train | t pred |
| SUSY | $20.0\% \pm 0.2\%$ | $608 \pm 2$ | $134 \pm 4$ | $20.1\% \pm 0.2\%$ | $592 \pm 2$ | $129 \pm 1$ |
| Mnist bin | $2.2\% \pm 0.1\%$ | $1342 \pm 5$ | $491 \pm 32$ | $2.3\% \pm 0.1\%$ | $1814 \pm 8$ | $954 \pm 21$ |
| Usps | $3.0\% \pm 0.1\%$ | $19.8 \pm 0.1$ | $7.3 \pm 0.3$ | $3.0\% \pm 0.2\%$ | $66.1 \pm 0.1$ | $48 \pm 8$ |
| Webspam | $1.3\% \pm 0.1\%$ | $2440 \pm 5$ | $376 \pm 18$ | $1.3\% \pm 0.1\%$ | $4198 \pm 40$ | $1455 \pm 180$ |
| a9a | $15.1\% \pm 0.2\%$ | $29.3 \pm 0.2$ | $1.5 \pm 0.1$ | $15.1\% \pm 0.2\%$ | $30.9 \pm 0.2$ | $3.2 \pm 0.1$ |
| CIFAR | $19.2\% \pm 0.1\%$ | $2408 \pm 14$ | $820 \pm 47$ | $19.0\% \pm 0.1\%$ | $2168 \pm 19$ | $709 \pm 13$ |