Università degli Studi di Genova
XXXV Cycle of PhD Programme in Mathematics and Applications

Optimal and Efficient Learning In Classification

Andrea Della Vecchia

A thesis submitted for the degree of
Doctor of Philosophy
under the advice and supervision of

Professor Ernesto De Vito
&
Professor Lorenzo Rosasco

Contents

Introduction

In last decades what is commonly called *learning from data* has become more and more important in science. The huge amount of information we have available nowadays, combined with the enormous growth in power and capacity of modern computers, has completely modified many fields in science: from bioinformatics to medicine, from finance to astronomy and many others. The use of machine learning is increasingly common and the applications are entering in people's everyday life: for fraud detection systems that can identify patterns of fraudulent behavior in financial transactions, for recommender systems that can suggest products, services, or content based on a user's preferences and behavior, for natural language processing systems that can process and understand human language, allowing for applications such as chatbots and language translation.

This has created opportunities but also challenges. The need for processing huge quantities of data has made, inevitably, statistical and computational problems to dramatically grow both in terms of size and complexity. In this thesis our focus will be on developing large scale models and algorithms towards efficient, and hence sustainable machine learning. In particular, the goal will be to study both the statistical and the computational aspects of these algorithms.

The thesis will be divided in two parts. In the first part we will investigate how projections, and in particular random projections (sketches), can allow learning with optimal prediction properties and minimal memory footprint. In this context, projections are a useful mathematical and algorithmic tool to reduce data size (volume and input/output dimensionality). Finding projections with minimal size is the key to efficiently solve many problems in machine learning since only compressed data need be processed. With this scope in mind, our contribution will consist in developing a theory for sketched algorithms working with loss functions as general as possible, extending the known results limited to square loss and smooth losses (logistic) [Rudi et al., 2015, Bach, 2013, Marteau-Ferey et al., 2019] (see Chapters 4-5-6). This will be done in the context of classification with *surrogate* losses, i.e. convex proxy of the 0-1 loss, and we will investigate the main properties of learning when considering convex, possibly non differentiable, loss functions. Our focus will be on the Empirical Risk Minimization algorithm and we will exploit random projections techniques, in particular Nyström method, to make it efficient and able to deal with large scale datasets. Statistical guarantees will be given for this simplified model and we will show that, in some regimes, decreasing the computational cost will come together with no drop in accuracy, allowing no tradeoff between the two. Fast rates for the convergence of the excess risk will be proven. 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 empirical process theory [Boucheron et al., 2013] with results for random projections. Nevertheless, known results for smooth loss functions will be recovered in a common regime. We will also show how to pass from the obtained statistical bounds for surrogate losses to excess risk bounds for the classification risk, i.e. the risk associated with the non convex 0-1 loss. Finally, these theoretical results will be tested with some numerical experiments over real data.

In the second part of the thesis we will focus on the plug-in approach when considering generalized performance metrics for classification tasks. In particular, we will study the so called linear fractional performance measures, that includes as particular cases many well known metrics such as accuracy, F-score, Jaccard and AM-measure among others [Koyejo et al., 2014]. The reason is that, in different applications we may want to penalise more some types of "mistakes" with respect to others (penalise more false negative with respect to false positive for example). Our original work will consist in showing that, with this kind of error measures, the optimal classifier is nothing else than a plug-in rule, i.e. a step function depending on the regression function and on an optimal threshold. Given the optimal shape of the classifier, we will assume to have an estimator for the unknown regression function and we will give an algorithm to estimate the unknown threshold. In fact, differently from the usual way of estimating the threshold via grid search, for which no statistical guarantees have been proven, we will derive our estimated threshold through efficiently solving a simple fixed point equation only depending on the inputs distribution and not on the labels. Using only an unlabelled dataset to derive the threshold is a key advantage in practice in all the many cases where the number of labelled example is limited. We will give post-processing bounds for our estimator, i.e. excess risk bounds depending on the norm difference between the true regression function and the given estimator of it. These post-processing bounds will be derived both without assumptions on the probability distribution of the data and, a refined version of them, under a margin assumption.

As regards the first part the structure is the following.
In Chapter 1 we will give an introduction to Statistical Learning Theory and present the mathematical framework and the statistical model that will be used in the following sections. We will define the main quantities in this setting, such as loss functions, Bayes predictor,

excess risk etc. We will introduce the notion of consistency of an estimator. We will recall the reader the well-known Empirical Risk Minimization (ERM) algorithm and its regularized version, needed to avoid the overfitting phenomenon. To nicely introduce the reader to excess risk bounds, we will show a simple way of splitting the error and we will mention the well known bias-variance tradeoff phenomenon. We will suggest some ways of controlling the various pieces appearing in the bound after the splitting and particular emphasis will be given on how to control the statistical *estimation* error. At this scope, Rademacher averages will be presented, together with some of their properties that we will need later on.

In Chapter 2 we will introduce kernel methods and define the notion of Reproducing Kernel Hilbert Spaces (RKHS). We will state a fundamental theorem in our analysis: the representer theorem and we will show how it deeply connects to our problem.

In Chapter 3 we explain what is learning on random subspaces. We will start introducing the main ideas in the simplified setting of linear kernel and square loss. We will briefly discuss some well-known approaches from the deterministic and expensive Principal Component Analysis (PCA), to cheaper random methods such as Sketching, Random Features and finally, the Nyström method that we will exploit in our results. We will conclude this section with a small summary of the main result contained in [Rudi et al., 2015], that, limited to square loss, will be the starting point for our work and extensions.

In Chapter 4 we present our original work. In Section 4.1 we will explain our setting and the basic assumptions in our framework. In Section 4.2 we will discuss ERM in this setting, considering also the computational aspect to motivate the need of random projections in the following sections. We will provide via a simple proof some new excess risk bounds for sub-gaussian random variables, equivalent to the known ones in the bounded case. In Section 4.3 we exploit Nyström method to obtain cheaper estimators. We analyse their computational complexity and provide bounds for the associated excess risk under some assumption on the eigenvalues decay of the covariance operator. In Section 4.4 we will finally present our main results: fast rates of convergence for the excess risk of these projected predictors. Under an additional assumption of a Bernstein condition on the loss function, we will show that fast rates up to $(1/n)$ can be achieved and that they match the state-of-art bounds in [Steinwart and Christmann, 2008], but with possibly high computational savings. We will compare these results with the known ones for Random Feature, showing some advantages in our refined analysis. We conclude the section specifying our results for differentiable loss (square loss and logistic in particular), recovering, in the common regime, the results presented in [Rudi et al., 2015].

In Chapter 5 we study how it is possible to relate our theory for convex surrogate losses to non-convex 0-1 loss and its classical misclassification risk. In particular, we will show

how to easily pass from our just derived risk bounds for surrogate losses to corresponding upper bounds controlling the 0-1 risk. While doing so, we will also introduce a margin condition (or low noise condition), that consists in assuming that the regression function is unlikely to be very close to 1/2. We finally compare the derived 0-1 risk bounds obtained starting from hinge loss or square loss.

In Chapter 6 we present some numerical experiments to show the computational benefit of our approach with random projections when dealing also with real data. We will focus on binary classification with SVMs. Our sketched estimator will be shown to match the performance of standard SVMs while using only a fraction of the available training set, leading to notable savings both in time and memory.

For the second part the organization is the following.
In Chapter 7 we will present the plug-in approach and its development in recent years. We will present the setting of our classification problem and we will introduce the generalized performance metrics that will be employed in the rest of the chapter. Given a linear fractional performance measure, we will derive the corresponding optimal thresholding, with the optimal threshold that will satisfy a fixed point equation. We will show that this optimal thresholding function is indeed optimal among all possible classifiers.

In Chapter 8 we will propose our estimator, given the shape of the derived optimal classifier. We will show a post-processing bound when no assumption is made on the probability distribution of the data. Next, we will introduce a margin assumption and we will give a refined version of the previous bound.

Part I

Empirical Risk Minimization with Surrogate Losses

0.1 Main Notation

For the reader's convenience we collect the main notation we will use in this thesis. 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 $\mathcal{X}$ is separable and $\mathcal{Y}$ is a Polish space [Steinwart and Christmann, 2008]. Table 0.1 summarizes the main notations.

Table 1: Definition of the main quantities used

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

Chapter 1

Statistical Learning Theory

In this section we give an introduction about Statistical Learning Theory. We will present some of the main challenges when designing a learning algorithm, together with the principal assumptions we based our statistical model on. We will briefly recall some of the technical tools we will need in the next chapters in order to develop a sound mathematical theory and solid statistical guarantees on our results. The rest of the chapter is based on the following references: [Bousquet et al., 2003, Bach, 2021, Shalev-Shwartz and Ben-David, 2014, Boucheron et al., 2005, Boucheron et al., 2013]

The goal of statistical learning theory is to build a mathematical framework and provide theoretical guarantees in the context of statistical inference. Among all, this involves building models and making predictions. The problem is studied in a statistical setting, which means there are assumptions of statistical nature about the underlying phenomena and in particular about how the data are generated.
Inspired by the classical scientific method, we aim to study the process of inductive inference that we can roughly summarize as the following process:

1. we observe a particular phenomenon,

2. we construct a mathematical model of it as accurate as possible,

3. we use this model to make predictions in the future given some new observations.

This is no different from the approach to every other natural science. The peculiarity of Machine Learning, and the main reason of its success, lies on the fact that this process is completely automated and the goal of Learning Theory is actually to give a mathematical formalization of it.

In the following we will focus into classical supervised learning where the data consists of instance-label pairs. The product of our learning algorithm is a function mapping instances to labels, with the objective of making as few as possible mistakes when predicting the

labels of new unseen instances.
The naive intuition of building this function simply fitting exactly the training data is often not working. In fact, in presence of noise, this approach can lead to poor *generalization* performance on new data. This phenomenon is known as *overfitting* and avoiding it is one of the key points to keep in mind when designing our learning procedure. Intuitively, what we expect from our algorithm is to find regularities in the observed phenomenon, i.e. patterns and structures in the data, to extract knowledge from them and to exploit it when predicting unseen inputs. Instead, what we really want to avoid is *learning by memorization* that, despite perfect performances on the training set, would likely be unuseful when labelling new instances. We will talk in details about this in the following.

1.1 Statistical Learning Theory Framework

In classical *supervised* learning the data, or *training set*, is a set of n examples represented as pairs

$$(x_1, y_1), \dots, (x_n, y_n)$$

with x_i the input and y_i the output (or *label* in classification). We will assume that the inputs belong to an input space $\mathcal{X}$, often chosen as a subset of $\mathbb{R}^d$, while, similarly, $\mathcal{Y}$ will be the output space. The dimension d of the input space is referred to as the number of features of the problem. As regards some well-known choices for the output space, $\mathcal{Y} \subseteq \mathbb{R}$ is commonly called regression, $\mathcal{Y} = \{-1, 1\}$ is classification and $\mathcal{Y} = \{1, 2, \dots, K\}$ is multiclass (or multi-category) classification. We indicate $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ the *data space.*

The goal of the *learner* is to understand how x's and y's are linked together. Mathematically, this means that, given a so called *training set* $(x_1, y_1, \dots, x_n, y_n)$, we want to find a function f that controls the input-output relation, i.e.

$$\widehat{f} : \mathcal{X} \to \mathcal{Y}$$

such that, given a new input x_{new} not present in our training set, then

$$\widehat{f}(x_{\text{new}}) \sim y_{\text{new}}.$$

The prediction rule $\widehat{f}$ is usually called *hypothesis* or, precisely, *predictor.*
In the following, all the quantities that depend on the training set, as $\widehat{f}$, will be referred as *empirical quantities* and, for sake of clarity, will be denoted with an "hat".
To conclude, the map taking the training data as input and returning a predictor $\widehat{f}$

$$(x_1, y_1, \dots, x_n, y_n) \longmapsto \widehat{f}_{(x_1,y_1),\dots,(x_n,y_n)} = \widehat{f}$$

is called a *learning algorithm.* A learning algorithm is good as far as its provided solution is able to predict or classify well new, previously unseen data. In this case, the algorithm is said to *generalize well.*

1.1.1 The Statistical Model

The statistical model comes from the fact that learning is studied from a random sample and it should take into account possible uncertainties coming from the task and the data. For this reason, consider (X, Y) as a pair of random variables taking values in $\mathcal{X}$ and $\mathcal{Y}$ and denote with P their joint distribution. Training set $(x_1, y_1, \ldots, x_n, y_n)$ is simply considered as a random realization of random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$, i.e. n i.i.d. copies of (X, Y), with joint distribution P.

To measure performance we define the so called *loss function* $\ell : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$. The purpose of introducing such a function is to quantitatively evaluate our algorithm, penalizing mistakes and inaccuracy in its predictions. In fact, $\ell(y, f(x))$ is a point-wise measure of the error we incur in when we predict $f(x)$ in place of y. Some classical examples are:

(a) hinge loss:

$$\ell(y, a) = |1 - ya|_+ = \max\{0, 1 - ya\} \tag{1.1}$$

(b) logistic loss:

$$\ell(y, a) = \log(1 + e^{-ya}) \tag{1.2}$$

(c) square loss:

$$\ell(y, a) = (y - a)^2 . \tag{1.3}$$

The choice of the loss is deeply related with the problem at hand: hinge and logistic losses are chosen in *classification* tasks, when the goal is to separate inputs into two (or multiple) different classes; on the contrary, the natural application for square loss is in *regression* problems, where the goal is to predict a real number given an input (but can be easily adapted to classification).

As already mentioned above, we are interested in the error we expect to make when predicting Y given X. For this purpose, given a loss function, we define the *expected risk* (or *expected error*) as

$$L(f) := \mathbb{E}_{(X,Y)\sim P}[\ell(Y, f(X))]. \tag{1.4}$$

With the above choice of error measure, the best input-output relation is the minimizer of the expected risk $L(f)$ over the possible functions $f : \mathcal{X} \to \mathcal{Y}$. This minimizer is usually referred to as the *target*, or *Bayes, function* f^*, i.e.

$$L(f^*) = \min_{f:\mathcal{X}\to\mathcal{Y}} L(f). \tag{1.5}$$

Clearly, the target function cannot be computed since the probability distribution P is unknown. Note also that in general $L(f^*)$ is not zero and, in case of complex problems with high uncertainty, it can be big. For this reason, what we are really interested in is the relative performance of our algorithm compared with the one of f^*. Consequently, we define the *excess risk* of an estimator f as the difference between its expected risk with the best possible one given by the target function:

$$\mathcal{E}(f) := L(f) - L(f^*). \tag{1.6}$$

1.1.2 Measures of performance and consistency

Our goal is monitor the performance of our algorithm $\widehat{f}$ via controlling its excess risk. Given its dependence though the data, $\widehat{f}$ is a random quantity and, then, $L(\widehat{f})$ will be a (positive) random variable as well. Hence we cannot expect the excess risk to be small for all possible training sets. Anyway, we can require for example our estimator to perform well in *expectation*, meaning that the *expected error*

$$\mathbb{E}\left[L(\widehat{f}) - L\left(f^*\right)\right]$$

is small, with the expectation taken over all possible training sets

$$(x_1, y_1, \ldots, x_n, y_n) \in (\mathcal{X} \times \mathcal{Y})^n.$$

One other classical option leads to *probably approximately correct* (PAC) learning, where we look for estimators with low excess risk *in probability.* This simply means that we require $\widehat{f}$ to perform well on "almost" all the possible training sets, mathematically

$$\mathbb{P}\left[L(\widehat{f}) - L\left(f^*\right) \leqslant \epsilon\right] \geqslant 1 - \eta$$

where $\varepsilon > 0$ is a given bound on the error and $0 < 1 - \eta < 1$ is the confidence level.

An important property we want to enforce on our algorithm is that, in the limit of our training set being infinite, we are able to recover the Bayes solution f^* both in probability and expectation.

Definition 1 (Consistency)**.** *A learning algorithm $\widehat{f}$ is called consistent in probability if*

$$\lim_{n\to+\infty} \mathbb{P}\left[L(\widehat{f}) - L\left(f^*\right) > \epsilon\right] = 0 \quad \forall \epsilon > 0.$$

Similarly, $\widehat{f}$ is called consistent in expectation if

$$\lim_{n\to+\infty} \mathbb{E}\left[L(\widehat{f}) - L\left(f^*\right)\right] = 0.$$

Applying Markov inequality we can directly show that consistency in expectation implies consistency in probability:

$$\mathbb{P}\left[L(\widehat{f}) - L\left(f^*\right) > \epsilon\right] \leqslant \frac{\mathbb{E}\left[L(\widehat{f}) - L\left(f^*\right)\right]}{\epsilon}.$$

The converse does not hold in general.

An algorithm is called *universally consistent* if the above conditions hold for all possible probability distributions P.

Nevertheless, the convergence speed will depend on P. In particular, over all possible distributions, this rate can be arbitrary slow and there is no hope of having a uniform rate in general. This result is well-known and formulated in different versions in the so called *no free lunch theorems*. We report here the one contained in [Devroye et al., 1996].

Theorem 1 (No free lunch - sequence of errors)**.** *Consider a binary classification problem with the 0-1 loss, with X infinite. Let $\mathcal{P}$ denote the set of all probability distributions on $X \times \{0,1\}$. There exists $p \in \mathcal{P}$, such that, for any decreasing sequence a_n tending to zero and such that $a_1 \leqslant 1/16$, for any learning algorithm $\widehat{f} = \widehat{f}(\mathcal{D}_n(p))$, with $\mathcal{D}_n(p)$ the training data drawn from p, for all $n \geqslant 1$, then:*

$$\mathbb{E}\left[L_p(\widehat{f})\right] - L_p^* \geqslant a_n.$$

This means that no method can be universal and achieve a good convergence rate on all problems. However, such negative results consider classes of problems which are arbitrarily large.

1.1.3 Empirical Risk Minimization

We introduce here one of the most famous and successful algorithms to solve our the learning problem. This will be the a key ingredient for the rest of the thesis.
As mentioned above, the problem of learning is to solve

$$\min_{f:\mathcal{X}\to\mathcal{Y}} L(f)$$

with f a measurable function and for a fixed yet unknown distribution P. Still, we can address P trough the training set $\mathcal{D}_n = (x_1, y_1), \dots, (x_n, y_n) \sim P^n$, i.e. a collection of n samples identically and independently distributed (i.i.d.) with respect to P. Nevertheless, finding an exact solution is in general not possible since only finite data are given. In fact, the expected risk is defined through an expectation that might be hard or impossible to compute, as well as the minimization over the space of all measurable functions can be unfeasible. Then, it can be reasonable to substitute the expected risk with the empirical risk $\widehat{L}$ defined as

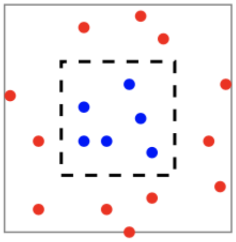
$$\widehat{L}(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$$

and to look for a solution over a restricted set, i.e. the so called *hypothesis space* $\mathcal{H}$.
Then, Empirical Risk Minimization (ERM) is nothing else than the minimization problem

$$\min_{f\in\mathcal{H}} \widehat{L}(f), \quad \text{with} \quad \widehat{f} = \arg\min_{f\in\mathcal{H}} \widehat{L}(f)$$

1.2 Overfitting and Regularization

Despite ERM algorithm seems very reasonable, without being careful, it can also fail miserably. To explain this, we recur to a simple example: imagine we want to solve a classification problem with a sample as depicted in the following:



We assume that the underlying probability distribution P is such that instances are distributed uniformly inside the grey square, while the label is 1 if the instance is within the blue square (with area that is half of the grey square's one) and 0 otherwise. Now, let's pick as our predictor

$$\widehat{f}(x) = \begin{cases} y_i & \text{if } \exists i \in [n] \text{ s.t. } x_i = x \\ 0 & \text{otherwise.} \end{cases}$$

It's clear that, by construction, no matter what the sample is,

$$\widehat{L}(\widehat{f}) = 0,$$

i.e. we are fitting the data, and therefore this is a possible choice for our ERM algorithm since no other predictor can have a smaller empirical error. It is intuitive to realise that, despite the perfect performance on the training set, $\widehat{f}$ will be terrible when generalizing to new examples. In fact, it's easy to see that the true error of a classifier predicting 1 only on a finite number of instances is, given this simple model, exactly $1/2$, i.e. no better than coin flip:

$$L(\widehat{f}) = 1/2.$$

This phenomenon, where our performance is perfect on the training set yet very poor on new data, is called *overfitting*. Clearly, this comes from the fact that the algorithm is focusing too much on the specific piece of data we have at hand, perfectly fitting the given examples, without learning the actual rule to separate the two classes (in this case we are actually only memorizing the training set).

A classical way to avoid this problem is to introduce a bias towards *simple* solutions. Typically, one can choose a large hypothesis space $\mathcal{H}$ and define on $\mathcal{H}$ a *regularizer*, for example a norm $\|f\|$. Then the goal is to minimize a slightly different problem, the regularized empirical risk

$$\widehat{L}_\lambda(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda\|f\|^2.$$

The free parameter λ is called the *regularization parameter* and it allows to choose the right tradeoff between fit and complexity. The common way to tune λ is using a *cross-validation* strategy on an extra set of data.

Given the above definitions, in next section we show a standard way of controlling the excess risk of such estimators.

1.3 Error-splitting and bias-variance tradeoff

We start this section with some classical definitions. Let's analyse the excess risk $L(f)-L(f^*)$ of some function $f \in \mathcal{H}$. We add and subtract the risk of a *best-in-class* hypothesis $f_\mathcal{H}$, i.e. a hypothesis in $\mathcal{H}$ with minimal error

$$f_\mathcal{H} = \arg\min_{f\in\mathcal{H}} L(f),$$

obtaining

$$L(f) - L(f^*) = \underbrace{L(f) - L(f_\mathcal{H})}_{\text{estimation}} + \underbrace{L(f_\mathcal{H}) - L(f^*)}_{\text{approximation}}.$$

The above splitting is the simplest one and we will refine it later. The first difference in right hand side is usually called *estimation error* and it measures the quality of the hypothesis f with respect to the best hypothesis in the class. Similarly, the second difference is commonly referred to as the *approximation error* and it gives instead a measure of how well the Bayes risk can be approximated using $\mathcal{H}$; it's deterministic and depends on the underlying distribution P and the class $\mathcal{H}$. It's worth noticing that the choice of $\mathcal{H}$ is in fact crucial. Increasing $\mathcal{H}$ would clearly lead to an improvement in the approximation error since this will lead to a better approximation of f^*. On the contrary, the estimation error would be increased. This is known as the *bias-variance tradeoff*, showing the need of a compromise between the two sources of error.

In the following we analyse a further refined splitting of the excess risk. We take $\widehat{f}$ as our chosen hypothesis, i.e. the output of the ERM algorithm. We add and subtract $\widehat{L}(\widehat{f})$ and $\widehat{L}(f_\mathcal{H})$, i.e. the empirical risk of $\widehat{f}$ and $f_\mathcal{H}$, to the excess risk of $\widehat{f}$

$$\begin{aligned} L(\widehat{f}) - L(f^*) &= L(\widehat{f}) - \widehat{L}(\widehat{f}) + \widehat{L}(\widehat{f}) - \widehat{L}(f_\mathcal{H}) + \widehat{L}(f_\mathcal{H}) - L(f_\mathcal{H}) + L(f_\mathcal{H}) - L(f^*) \\ &\leqslant L(\widehat{f}) - \widehat{L}(\widehat{f}) + \widehat{L}(f_\mathcal{H}) - L(f_\mathcal{H}) + L(f_\mathcal{H}) - L(f^*) \end{aligned}$$

where the inequality comes from the fact that $\widehat{L}(\widehat{f}) - \widehat{L}(f_{\mathcal{H}}) \leqslant 0$ since $\widehat{f}$ is exactly the empirical risk minimizer. Now we can study how to bound the following three different terms:

- $L(f_{\mathcal{H}}) - L(f^*)$ is only the approximation error we already discussed
- $\widehat{L}(f_{\mathcal{H}}) - L(f_{\mathcal{H}})$ is difference between the expectation and the empirical average of $\ell \circ f_{\mathcal{H}}$. By the law of large numbers, we immediately obtain that

  $$\mathbb{P}\left[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i, f_{\mathcal{H}}(X_i)\right) - \mathbb{E}[\ell\left(Y_i, f_{\mathcal{H}}(X_i)\right)] = 0\right] = 1.$$

  So, as expected, with enough samples, the empirical risk of a function is a good approximation to its true risk. For a quantitative version of the above law of large numbers, Hoeffding's inequality can be used when the variables are bounded. More refined bounds can be obtained using additional informations, Bernstein's inequality for example allows to use the variance to get a finer non-asymptotic result. Any of these concentration inequalities will give a bound in $O(1/\sqrt{n})$.

- $L(\widehat{f}) - \widehat{L}(\widehat{f})$ is the more problematic term: both the empirical average $\widehat{L}$ and $\widehat{f}$ depend statistically on the data. This dependence does not allow us to directly follow the approach at the previous point. The standard way of dealing with this problem is by controlling uniformly this term over all $f \in \mathcal{H}$:

  $$L(\widehat{f}) - \widehat{L}(\widehat{f}) \leqslant \sup_{f\in\mathcal{H}} |L(f) - \widehat{L}(f)|. \tag{1.7}$$

  However, when controlling the maximal deviations over many functions f, there is always a small chance that one of these deviations get large. One very powerful tool that allows to sharply bound this term is Rademacher complexities [Boucheron et al., 2005] or Gaussian complexities [Bartlett and Mendelson, 2002]. We will focus on Rademacher complexity in the following.

Since in Eq. (1.7) we want an uniform bound on the difference of expected and empirical risk over a set of function, we need a method to compute some measure of ”complexity” of this hypothesis space. For simplicity let's introduce $Z \in \mathcal{Z}$, $\mathcal{F}$ the space of measurable function from $\mathcal{Z}$ to $\mathbb{R}$ and $\mathcal{D} = \{z_1, \ldots, z_n\}$ the data (we will recover problem (1.7) defining $Z = (X, Y) \in \mathcal{Z}$, and $\mathcal{F} = \{(X, Y) \mapsto \ell(X, h(X)), h \in \mathcal{H}\}$).

Definition 2 (Rademacher Complexity)**.** *We define the Rademacher complexity of the class of functions $\mathcal{F}$ from $\mathcal{Z}$ to $\mathbb{R}$:*

$$\widehat{R}(\mathcal{F}) = \mathbb{E}\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f\left(z_i\right)\right), \tag{1.8}$$

where $\varepsilon \in \mathbb{R}^n$ is a vector of independent Rademacher random variables, i.e. random variables taking values -1 or 1 with equal probabilities, that are also independent of $\mathcal{D}$.

Note that is a deterministic quantity that only depends on n and $\mathcal{F}$. In words, the Rademacher complexity is nothing else than the expected value of the maximal dot-product between values of a function f at the observations z_i and random labels. This means that the idea is somehow to measure the complexities of the space $\mathcal{F}$ through its ability of fitting random noise.

Theorem 2. *Given $(X, Y) \sim P$ as a pair of random variables taking values in $\mathcal{X}$ and $\mathcal{Y}$, training data $(x_1, y_1), \ldots, (x_n, y_n)$ as realizations of i.i.d. random variables $(X_1, Y_1) \ldots, (X_n, Y_n) \sim P$, i.e. n i.i.d. copies of (X, Y), a loss function $\ell : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ G-Lipschitz in its second argument, then*

$$\mathbb{E}\left[\sup_{f \in \mathcal{H}} \left| L(f) - \widehat{L}(f) \right| \right] \leqslant 2G\widehat{R}(\mathcal{H}).$$

Moreover, suppose $\mathcal{H}$ is a class of bounded functions from $\mathcal{X}$ to $[0, 1]$ then, with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{H}} \left| L(f) - \widehat{L}(f) \right| \leqslant 2\widehat{R}(\mathcal{H}) + \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}.$$

The proof of this theorem exploits the well known *symmetrization* argument.

Proposition 1 (Symmetrization)**.** *Given the Rademacher complexity of $\mathcal{F}$ defined in Eq. (1.8), we have:*

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f\left(z_i\right) - \mathbb{E}[f(z)] \right| \right] \leqslant 2\widehat{R}(\mathcal{F}).$$

Proof. Let $\mathcal{D}' = \{z_1', \ldots, z_n'\}$ be an independent copy of the data $\mathcal{D} = \{z_1, \ldots, z_n\}$ (often known as *ghost samples*). Let $(\varepsilon_i)_{i \in \{1, \ldots, n\}}$ be i.i.d. Rademacher random variables, which are also independent of $\mathcal{D}$ and $\mathcal{D}'$. Using that for all i in $\{1, \ldots, n\}$, $\mathbb{E}\left[f\left(z_i'\right) \mid \mathcal{D}\right] = \mathbb{E}[f(z)]$, we have:

$$\begin{aligned} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(\mathbb{E}[f(z)] - \frac{1}{n} \sum_{i=1}^{n} f\left(z_i\right)\right)\right] &= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[f\left(z_i'\right) \mid \mathcal{D}\right] - \frac{1}{n} \sum_{i=1}^{n} f\left(z_i\right)\right)\right] \\ &= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[f\left(z_i'\right) - f\left(z_i\right) \mid \mathcal{D}\right]\right)\right] \end{aligned}$$

by definition of the independent copy $\mathcal{D}'$. Then

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(\mathbb{E}[f(z)] - \frac{1}{n} \sum_{i=1}^{n} f\left(z_i\right)\right)\right] \leqslant \mathbb{E}\left[\mathbb{E}\left(\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \left[f\left(z_i'\right) - f\left(z_i\right)\right]\right) \mid \mathcal{D}\right)\right]$$

using that the supremum of the expectation is less than expectation of the supremum. Thus, by the towering law of expectation, we get

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}[f(z)]-\frac{1}{n}\sum_{i=1}^{n} f\left(z_i\right)\right)\right] \leqslant \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\left[f\left(z_i'\right)-f\left(z_i\right)\right]\right)\right].$$

We can now use the symmetry of the laws of ε_i and $f\left(z_i'\right)-f\left(z_i\right)$, to get:

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}[f(z)]-\frac{1}{n}\sum_{i=1}^{n} f\left(z_i\right)\right)\right]\\
\leqslant&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left(f\left(z_i'\right)-f\left(z_i\right)\right)\right)\right]\\
\leqslant&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left(f\left(z_i\right)\right)\right)\right]+\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left(-f\left(z_i\right)\right)\right)\right]\\
=&2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f\left(z_i\right)\right)\right]=2\widehat{R}(\mathcal{F}).
\end{aligned}$$

The reasoning is essentially identical for $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n} f\left(z_i\right)-\mathbb{E}[f(z)]\right)\right] \leqslant 2\widehat{R}(\mathcal{F})$. □

We will also need this other known result that we report here with a simple proof taken from [Meir and Zhang, 2003].

Proposition 2 (Contraction principle - Lipschitz-continuous functions)**.** *Given any functions $b, a_i : \Theta \rightarrow \mathbb{R}$ (no assumption) and $\varphi_i : \mathbb{R} \rightarrow \mathbb{R}$ any 1-Lipschitz-functions, for $i=1,\ldots,n$, we have, for $\varepsilon \in \mathbb{R}^n$ a vector of independent Rademacher random variables:*

$$\mathbb{E}_{\varepsilon}\left[\sup_{\theta\in\Theta}\left\{b(\theta)+\sum_{i=1}^{n}\varepsilon_i\varphi_i\left(a_i(\theta)\right)\right\}\right] \leqslant \mathbb{E}_{\varepsilon}\left[\sup_{\theta\in\Theta}\left\{b(\theta)+\sum_{i=1}^{n}\varepsilon_i a_i(\theta)\right\}\right].$$

Proof. We consider a proof by induction on n. The case $n=0$ is trivial, and we show how to go from $n \geqslant 0$ to $n+1$. We thus consider $\mathbb{E}_{\varepsilon_1,\ldots,\varepsilon_{n+1}}\left[\sup_{\theta\in\Theta}\left\{b(\theta)+\sum_{i=1}^{n+1}\varepsilon_i\varphi_i\left(a_i(\theta)\right)\right\}\right]$ and compute the expectation with respect to ε_{n+1} explicitly, by considering the two potential values with probability 1/2 :

$$\begin{aligned}
&\mathbb{E}_{\varepsilon_1,\ldots,\varepsilon_{n+1}}\left[\sup_{\theta\in\Theta}\left\{b(\theta)+\sum_{i=1}^{n+1}\varepsilon_i\varphi_i\left(a_i(\theta)\right)\right\}\right]\\
&\quad=\frac{1}{2}\mathbb{E}_{\varepsilon_1,\ldots,\varepsilon_n}\left[\sup_{\theta\in\Theta}\left\{b(\theta)+\sum_{i=1}^{n}\varepsilon_i\varphi_i\left(a_i(\theta)\right)+\varphi_{n+1}\left(a_{n+1}(\theta)\right)\right\}\right]\\
&\quad+\frac{1}{2}\mathbb{E}_{\varepsilon_1,\ldots,\varepsilon_n}\left[\sup_{\theta\in\Theta}\left\{b(\theta)+\sum_{i=1}^{n}\varepsilon_i\varphi_i\left(a_i(\theta)\right)-\varphi_{n+1}\left(a_{n+1}(\theta)\right)\right\}\right]\\
&\quad=\mathbb{E}_{\varepsilon_1,\ldots,\varepsilon_n}\left[\sup_{\theta,\theta'\in\Theta}\left\{\frac{b(\theta)+b\left(\theta'\right)}{2}+\sum_{i=1}^{n}\varepsilon_i\frac{\varphi_i\left(a_i(\theta)\right)+\varphi_i\left(a_i\left(\theta'\right)\right)}{2}+\frac{\varphi_{n+1}\left(a_{n+1}(\theta)\right)-\varphi_{n+1}\left(a_{n+1}\left(\theta'\right)\right)}{2}\right\}\right],
\end{aligned}\tag{1.9}$$

by assembling the terms. By taking the supremum over (θ,θ') and (θ',θ), we get

$$\begin{aligned}
(1.9)\leqslant&\mathbb{E}_{\varepsilon_1,\ldots,\varepsilon_n}\left[\sup_{\theta,\theta'\in\Theta}\left\{\frac{b(\theta)+b\left(\theta'\right)}{2}+\sum_{i=1}^{n}\varepsilon_i\frac{\varphi_i\left(a_i(\theta)\right)+\varphi_i\left(a_i\left(\theta'\right)\right)}{2}+\frac{\left|\varphi_{n+1}\left(a_{n+1}(\theta)\right)-\varphi_{n+1}\left(a_{n+1}\left(\theta'\right)\right)\right|}{2}\right\}\right]\\
\leqslant&\mathbb{E}_{\varepsilon_1,\ldots,\varepsilon_n}\left[\sup_{\theta,\theta'\in\Theta}\left\{\frac{b(\theta)+b\left(\theta'\right)}{2}+\sum_{i=1}^{n}\varepsilon_i\frac{\varphi_i\left(a_i(\theta)\right)+\varphi_i\left(a_i\left(\theta'\right)\right)}{2}+\frac{\left|a_{n+1}(\theta)-a_{n+1}\left(\theta'\right)\right|}{2}\right\}\right],
\end{aligned}$$

using Lipschitz-continuity. Exploiting again the fact that ε_{n+1} is a Rademacher random variable that takes values ±1 with equal probability, we can write

$$\begin{aligned}
&\mathbb{E}_{\varepsilon_1,\ldots,\varepsilon_n}\mathbb{E}_{\varepsilon_{n+1}}\left[\sup_{\theta\in\Theta}\left\{b(\theta)+\varepsilon_{n+1}a_{n+1}(\theta)+\sum_{i=1}^{n}\varepsilon_i\varphi_i\left(a_i(\theta)\right)\right\}\right]\\
\leqslant&\mathbb{E}_{\varepsilon_1,\ldots,\varepsilon_n,\varepsilon_{n+1}}\left[\sup_{\theta\in\Theta}\left\{b(\theta)+\varepsilon_{n+1}a_{n+1}(\theta)+\sum_{i=1}^{n}\varepsilon_i a_i(\theta)\right\}\right]
\end{aligned}$$

by the induction hypothesis, which leads to the desired result. □

An analogous result is valid also with absolute values [Ledoux and Talagrand, 1991].

Proposition 3 (Contraction principle - absolute values)**.** *Given any functions $a_i:\Theta\to\mathbb{R}$ (no assumption) and $\varphi_i:\mathbb{R}\to\mathbb{R}$ any 1-Lipschitz-functions such that $\varphi_i(0)=0$, for $i=1,\ldots,n$, we have, for $\varepsilon\in\mathbb{R}^n$ a vector of independent Rademacher random variables:*

$$\mathbb{E}_{\varepsilon}\left[\sup_{\theta\in\Theta}\left|\sum_{i=1}^{n}\varepsilon_i\varphi_i\left(a_i(\theta)\right)\right|\right]\leqslant 2\mathbb{E}_{\varepsilon}\left[\sup_{\theta\in\Theta}\left|\sum_{i=1}^{n}\varepsilon_i a_i(\theta)\right|\right].$$

Finally, to prove the second part of Theorem 2 and pass from a bound in expectation to a high probability control over the uniform deviations in Eq. (1.7), we will make use of a famous exponential concentration inequality, i.e. *McDiarmid's inequality*. Given n independent random variables, this inequality is useful when we want to concentrate other quantities than their average. What we need is the function of these random variables to have *bounded variation*.

Proposition 4 (McDiarmid's inequality)**.** *Let $Z_1, \ldots, Z_n$ be independent random variables (in any measurable space $\mathcal{Z}$), and $f : Z^n \to \mathbb{R}$ a function of "bounded variation", that is, such that for all i, and all $z_1, \ldots, z_n, z_i' \in \mathcal{Z}$, we have*

$$\left|f\left(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_n\right) - f\left(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n\right)\right| \leqslant c.$$

Then the random variable $W = f(Z_1, \ldots, Z_n)$ satisfies

$$\mathbb{P}\left(|W - \mathbb{E}W| \geqslant t\right) \leqslant 2 \exp\left(-2t^2 / \left(nc^2\right)\right),$$

or equivalently, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$|W - \mathbb{E}W| \leqslant \sqrt{\frac{nc^2}{2} \log \frac{1}{\delta}}.$$

The bounded differences assumption simply means that if the i-th variable of function f varies while keeping all the others fixed, the value of the function cannot change by more than c.

We can finally give the proof of the above theorem.

Proof of Theorem 2. The first part of the theorem is simply the application of the symmetrization argument and contraction principle to the class $\mathcal{F} = \{\ell \circ f : f \in \mathcal{H}\}$, with loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0, +\infty)$ G-Lipschitz in its second argument, i.e. $a \mapsto \ell(y,a)$ is G-Lipschitz.
Then

$$\mathbb{E}_{\varepsilon}\left(\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \ell\left(y_i, f\left(x_i\right)\right) \mid \mathcal{D}\right) \leqslant G \cdot \mathbb{E}_{\varepsilon}\left(\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f\left(x_i\right) \mid \mathcal{D}\right)$$

which leads to

$$\widehat{R}(\mathcal{F}) \leqslant G \cdot \widehat{R}(\mathcal{H}).$$

As regards the second part of the theorem, if we define

$$W := \sup_{f \in \mathcal{H}} \left|L(f) - \widehat{L}(f)\right|,$$

then, with $\mathcal{H}$ a class of bounded functions from $\mathcal{X}$ to $[0,1]$, W satisfies the bounded differences assumption with $c = 2/n$ and we get immediately the result.

□

Chapter 2

Kernel Methods and Reproducing Kernel Hilbert Spaces

In this chapter we introduce the so called *kernel methods*. These methods are widely used in machine learning and they are a powerful tool to extend linear models to non-linear ones. The basic idea is that a *kernel function* implicitly define an inner product in a high-dimensional space, given some technical conditions of symmetry and *positive-definiteness*. For this reason, suppose we are solving a classification problem: when the inner product in the input space is replaced with positive definite kernels we immediately extend our algorithms to a linear separation in a high-dimensional space, or, equivalently, to a non-linear separation in the initial input space.
We will proceed staring from the definition of a kernel and of a reproducing kernel Hilbert space. Afterwards, we will connect this mathematical definition with the more intuitive idea of embedding the data into a high dimensional feature space. We will recall the well-known representer theorem and it's application to regularized ERM in supervised learning. We will finally introduce the so called *kernel trick*, i.e. for many algorithms the solution can be carried out just on the basis of the values of the kernel function over pairs of domain points, without ever explicitly expressing the embedding given by the feature map.
This chapter is based on the following references [Mohri et al., 2018, Shalev-Shwartz and Ben-David, 2014, Bach, 2021]

2.1 Kernels and RKHS

We start with the mathematical definition of a kernel.

Definition 3. *Given a set $\mathcal{X}$, we call semi-positive definite kernel a map $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that*

$$\begin{array}{ll} K\left(x, x'\right) = K\left(x', x\right) & x, x' \in \mathcal{X} \\ \sum_{i,j=1}^{n} c_i c_j K\left(x_i, x_j\right) \geqslant 0 & x_1, \ldots, x_n \in \mathcal{X}, \quad c_1, \ldots, c_n \in \mathbb{R}. \end{array}$$

In this thesis, what we especially care about is how it is possible to univocally define a Hilbert space of functions associated with a kernel K. We recall here this known result.

Theorem 3 (Kolmogorov-Moore theorem)**.** *Take a semi-positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, then there exists a unique space $\mathcal{H}$ with the following properties:*

(a) the elements of $\mathcal{H}$ are functions $f : \mathcal{X} \to \mathbb{R}$

(b) $\mathcal{H}$ is a vector space with respect to the usual pointwise operations of sum and product by scalar

(c) $\mathcal{H}$ is a Hilbert space with scalar product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\| \cdot \|_{\mathcal{H}}$

(d) for all $x \in \mathcal{X}$ there is $K_x \in \mathcal{H}$ such that

$$f(x) = \langle f, K_x \rangle_{\mathcal{H}}, \quad f \in \mathcal{H} \quad \textit{(reproducing property)}$$

such that

$$K\left(x, x'\right) = \langle K_x, K_{x'} \rangle_{\mathcal{H}} \quad x, x' \in \mathcal{X}. \tag{2.1}$$

Conversely, take a space $\mathcal{H}$ satisfying (a)-(d) and define K by means of (2.1), then K is a semi-positive definite kernel.

A space with the above described properties it's called *Reproducing Kernel Hilbert Space* (RKHS) on $\mathcal{X}$ with reproducing kernel K

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad K\left(x, x'\right) = \langle K_x, K_{x'} \rangle_{\mathcal{H}}.$$

The *reproducing property* is fundamental since, exploiting it, we can evaluate functions through scalar products in $\mathcal{H}$.

2.1.1 Feature maps

It is possible to equivalently define an RKHS through a so called *feature map*. Let's call $\mathcal{W}$ a Hilbert space with scalar product $\langle \cdot, \cdot \rangle_{\mathcal{W}}$, we define a map

$$\Phi : \mathcal{X} \to \mathcal{W} \quad x \mapsto \Phi(x),$$

that will be our *feature map*. For all $w \in \mathcal{W}$ we define also

$$f_w : \mathcal{X} \to \mathbb{R} \quad f_w(x) = \langle w, \Phi(x) \rangle_{\mathcal{W}}$$

satisfying the following property

$$\mathcal{W} = \overline{\operatorname{span}}\{\Phi(x) \mid x \in \mathcal{X}\},$$

which is equivalent to assume that the map $w \to f_w$ is injective. If the above property is not satisfied it is always possible to replace $\mathcal{W}$ with $\overline{\operatorname{span}}\{\Phi(x) \mid x \in \mathcal{X}\}$. Denote

$$\mathcal{H} = \{f_w \mid w \in \mathcal{W}_\Phi\}.$$

It can be shown that $\mathcal{H}$ is a RKHS with respect to the scalar product

$$\langle f_w, f_{w'}\rangle_{\mathcal{H}} = \langle w, w'\rangle_{\mathcal{W}}.$$

More than that, $\mathcal{H}$ is the unique reproducing kernel Hilbert space with reproducing kernel

$$K\left(x, x'\right) = \left\langle \Phi(x), \Phi\left(x'\right)\right\rangle_{\mathcal{W}}.$$

The reproducing property, combined with the idea of feature maps, justify why in Chapter 4 we will solve the ERM problem hin the simplified case of linear functions in infinite dimensional spaces, i.e. $f(x) = \langle w, x\rangle$ with $w \in \mathcal{H}$ and $\mathcal{H}$ being a reproducing kernel Hilbert space. In fact, as just seen, to study this model is exactly equivalent to work with kernels and learn non-linear functions in the input space, since for every $f \in \mathcal{H}$ we can write $f(x) = \langle w, \Phi(x)\rangle$ for some $w \in \mathcal{H}$ and a feature map Φ, recovering the infinite dimensional linear case.

Example 1 (Gaussian kernel)**.** *Following the notation we used above let $\mathcal{W} = \ell_2$ be the Hilbert space of square-summable sequences with the scalar product*

$$\left\langle (a_\ell)_\ell, (b_n)_\ell \right\rangle_{\ell_2} = \sum_{\ell=0}^{+\infty} a_\ell b_\ell.$$

Define the feature map

$$\Phi : \mathbb{R} \to \ell_2, \qquad \Phi(x)_\ell = e^{-\frac{x^2}{2}} \frac{x^\ell}{\sqrt{\ell!}} = \varphi_\ell(x), \qquad \ell \in \mathbb{N},$$

which is well-defined since

$$\sum_{\ell=0}^{+\infty} \varphi_\ell(x)^2 = e^{-x^2} \sum_{\ell=0}^{+\infty} \frac{x^{2\ell}}{\ell!} = e^{-x^2} e^{x^2} = 1.$$

Hence, according to the theory above

$$\mathcal{H} = \left\{ f : \mathbb{R} \to \mathbb{R} \mid f(x) = \sum_{\ell=1}^{+\infty} a_\ell \varphi_\ell(x), \quad \sum_{\ell=1}^{+\infty} a_\ell^2 < +\infty \right\}$$

is a reproducing kernel Hilbert space with kernel

$$K\left(x, x'\right) = e^{-\frac{x^2}{2} - \frac{x'^2}{2}} \sum_{\ell=0}^{+\infty} \frac{\left(xx'\right)^\ell}{\ell!} = e^{-\frac{x^2}{2} - \frac{x'^2}{2} + xx'} = \exp\left(-\frac{\left(x - x'\right)^2}{2}\right),$$

which is the Gaussian kernel with $\sigma = \sqrt{2}$. It's easy to check also injectivity:

$$\sum_{\ell=1}^{+\infty} a_\ell e^{-\frac{x^2}{2}} \frac{x^\ell}{\sqrt{\ell!}} = 0, \quad \forall x \in \mathbb{R} \quad \Rightarrow \quad a_\ell = 0, \quad \forall \ell \in \mathbb{N}$$

2.2 Representer Theorem

Infinite dimensional models introduced above seem unpratical at first sight since ML algorithms cannot be run in infinite dimensions. We present here some results in kernel theory that will help overcoming this issue.
Our starting point will be the optimization problem coming from machine learning with linear models, with data $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \ldots, n$:

$$\min_{w \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, \langle \varphi\left(x_i\right), w \rangle\right) + \frac{\lambda}{2} \|w\|^2. \tag{2.2}$$

It's important to notice that the objective function in Eq. (2.2) accesses the input observations $x_1, \ldots, x_n \in \mathcal{X}$, only through dot-products $\langle w, \varphi\left(x_i\right)\rangle, i = 1, \ldots, n$, and that we penalize using the Hilbert norm $\|w\|$. The following theorem is crucial and it will be exploited several times in the following of this thesis.

Theorem 4 (Representer theorem [Kimeldorf and Wahba, 1971, Schölkopf et al., 2001])**.** *Let $\varphi : \mathcal{X} \to \mathcal{H}$. Let $(x_1, \ldots, x_n) \in \mathcal{X}^n$, and assume that the functional $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ is strictly increasing with respect to the last variable.*
Then the infimum of $\Psi\left(\langle w, \varphi\left(x_1\right)\rangle, \cdots, \langle w, \varphi\left(x_n\right)\rangle, \|w\|^2\right)$ can be obtained by restricting to a vector w of the form

$$w = \sum_{i=1}^{n} \alpha_i \varphi\left(x_i\right),$$

with $\alpha \in \mathbb{R}^n$.

Proof. Proof Let $w \in \mathcal{H}$, and $\mathcal{H}_{\mathcal{D}} = \left\{\sum_{i=1}^{n} \alpha_i \varphi\left(x_i\right), \alpha \in \mathbb{R}^n\right\} \subset \mathcal{H}$, the linear span of the feature vectors. Let $w_{\mathcal{D}} \in \mathcal{H}_{\mathcal{D}}$ and $w_\perp \in \mathcal{H}_{\mathcal{D}}^\perp$ be such that $w = w_{\mathcal{D}} + w_\perp$, using the Hilbertian structure of $\mathcal{H}$. Then $\forall i \in \{1, \ldots, n\}, \langle w, \varphi\left(x_i\right)\rangle = \langle w_{\mathcal{D}}, \varphi\left(x_i\right)\rangle + \langle w_\perp, \varphi\left(x_i\right)\rangle$ with $\langle w_\perp, \varphi\left(x_i\right)\rangle = 0$ From Pythagorean theorem, we get: $\|w\|^2 = \|w_{\mathcal{D}}\|^2 + \|w_\perp\|^2$. Therefore we have:

$$\begin{aligned}\Psi\left(\langle w, \varphi\left(x_1\right)\rangle, \ldots, \langle w, \varphi\left(x_n\right)\rangle, \|w\|^2\right) &= \Psi\left(\langle w_{\mathcal{D}}, \varphi\left(x_1\right)\rangle, \ldots, \langle w_{\mathcal{D}}, \varphi\left(x_n\right)\rangle, \|w_{\mathcal{D}}\|^2 + \|w_\perp\|^2\right) \\ &\geqslant \Psi\left(\langle w_{\mathcal{D}}, \varphi\left(x_1\right)\rangle, \ldots, \langle w_{\mathcal{D}}, \varphi\left(x_n\right)\rangle, \|w_{\mathcal{D}}\|^2\right)\end{aligned}$$

Thus

$$\inf_{w \in \mathcal{H}} \Psi\left(\langle w, \varphi\left(x_1\right)\rangle, \cdots, \langle w, \varphi\left(x_n\right)\rangle, \|w\|^2\right) = \inf_{w \in \mathcal{H}_{\mathcal{D}}} \Psi\left(\langle w, \varphi\left(x_1\right)\rangle, \ldots, \langle w, \varphi\left(x_n\right)\rangle, \|w\|^2\right),$$

obtaining the stated result. □

What is fundamental to notice is that the problem of finding the solution of Eq. (2.2) in the infinite-dimensional space $\mathcal{H}$ is now equivalent to find a vector $\alpha \in R^n$, given $w = \sum_{i=1}^{n} \alpha_i \varphi(x_i)$.

Corollary 1 (Representer theorem for supervised learning)**.** *For $\lambda > 0$,*

$$\inf_{w \in \mathcal{H}} \frac{1}{n} \sum \ell\left(y_i, \langle w, \varphi(x_i) \rangle\right) + \frac{\lambda}{2} \|w\|^2 = \inf_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum \ell\left(y_i, \langle w, \varphi(x_i) \rangle\right) + \frac{\lambda}{2} \|w\|^2 \ \ s.t. \ \ w = \sum_{i=1}^{n} \alpha_i \varphi(x_i) .$$

No assumption on the loss function ℓ is needed.

Thanks to the above Corollary, we can rewrite the learning problem with kernels notation. We will need the kernel function k which is the dot product between feature vectors:

$$k\left(x, x'\right) = \left\langle \varphi(x), \varphi\left(x'\right) \right\rangle .$$

We have:

$$\forall j \in \{1, \ldots, n\}, \langle w, \varphi(x_j) \rangle = \sum_{i=1}^{n} \alpha_i k(x_i, x_j) = (K\alpha)_j$$

where $K \in \mathbb{R}^{n \times n}$ is the kernel matrix, such that $K_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle = k(x_i, x_j)$, and

$$\|w\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \langle \varphi(x_i), \varphi(x_j) \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K_{ij} = \alpha^\top K \alpha$$

We can then write:

$$\inf_{w \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, \langle w, \varphi(x_i) \rangle\right) + \frac{\lambda}{2} \|w\|^2 = \inf_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, (K\alpha)_i\right) + \frac{\lambda}{2} \alpha^\top K \alpha.$$

For a test point $x \in X$, we have $f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$.
Thus, the input observations are contained in the kernel matrix and the kernel function, regardless of the dimension of $\mathcal{H}$. Remarkably, we never need to explicitly compute the feature vector $\varphi(x)$, that would be often unfeasible. This is commonly called the *kernel trick*, which is one of the keys of kernels' success. Note again that with this new formulation we replaced $\mathcal{H}$ by $\mathbb{R}^n$; this is clearly interesting computationally when the dimension of $\mathcal{H}$ is very large.

Chapter 3

Learning on Random Subspaces

As mentioned above, the goal in supervised learning is to learn from examples a function able to predict well on new data. Often this prediction function can be highly non-linear and non-parametric learning methods must be taken into account. The just described kernel methods are among the most popular non-parametric learning tools in machine learning and this is thanks to their excellent theoretical properties, widely studied in literature. Anyway, despite sound statistical guarantees, kernel methods have limited applications in large scale learning because of time and memory requirements, typically at least quadratic in the number of data points. Overcoming these scaling issues has motivated a lot of work in the direction of efficiency and computational savings. Then, a variety of practical approaches, with the goal of improving time complexity, have been studied. These include gradient methods, as well accelerated, stochastic and preconditioned extensions [Caponnetto and Yao, 2010, Avron et al., 2017, Gonen et al., 2016]. At the same, random projections approaches have been developed to reduce also memory requirements. Popular methods of this kind include Nyström subsampling [Williams and Seeger, 2001, Smola and Schölkopf, 2000], random features [Rahimi and Recht, 2007], and their numerous variations.
From a theoretical point of view, the main challenge is to characterize the statistical and computational tradeoffs derived by these approximation tools. In particular, this means to understand if, or under which conditions, computational gains come at the expense of statistical accuracy.

In this chapter we will briefly present some examples of random projections techniques when solving

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, f\left(x_i\right)\right) + \lambda \|f\|_{\mathcal{H}}^2, \tag{3.1}$$

for ℓ being convex with respect to its second variable and $\mathcal{H}$ a RKHS. For simplicity we study the special case of the square loss (ridge regression), where we have an closed-form solution, and linear kernel $k(x, x') = x^\top x'$ (feature map φ is the identity), corresponding to learn linear functions $f_w(x) = w^\top x$ and with $K = \widehat{X}\widehat{X}^\top$, where $\widehat{X} \in \mathbb{R}^{n\times d}$ is the matrix of the data. Then, the problem takes the form

$$\min_{\alpha\in\mathbb{R}^n} \frac{1}{n}\|y - K\alpha\|_2^2 + \lambda\alpha^\top K\alpha$$

and setting the gradient to 0 we obtain

$$\left(K^2 + n\lambda K\right)\alpha = Ky,$$

with a solution

$$\alpha = (K + n\lambda I)^{-1}y.$$

The main issue with this approach is that computing α requires $O(n^3)$ complexity in time and $O(n^2)$ in memory. What we need is some techniques to reduce the dimensionality of the problem.

At this scope, define $S \in \mathbb{R}^{d\times m}$, with $m \ll n, d$, and

$$\underbrace{\widehat{X}_m}_{n\times m} = \underbrace{\widehat{X}}_{n\times d}\underbrace{S}_{d\times m}.$$

This is nothing else than a transformation of the d-dimensional inputs to a lower dimensionality m

$$x \in \mathbb{R}^d \quad \to \quad \tilde{x} = (s_j^\top x)_{j=1}^m \in \mathbb{R}^m, \tag{3.2}$$

with $s_1, \dots, s_m$ columns of S.

We can reformulate problem (3.1) as

$$\min_{c\in\mathbb{R}^m} \frac{1}{n}\left\|c^\top \widehat{X}_m - \widehat{Y}\right\|^2 + \lambda\|c\|^2$$

where the complexity is reduced from d to m. Clearly the reader may wonder which kind of statistical guarantees this new formulation has. In particular, how much have we lost in terms of accuracy of our estimator when transforming the inputs through the embedding in (3.2)? This question will the main topic of next chapters. Another key ingredient is how to choose the embedding of the input point: we present in the following some of the most common choices.

3.1 PCA

The most straightforward choice for approximating K and reducing the $O(n^3)$ time complexity when inverting it is using its Singular Value Decomposition (SVD). This approach consists in selecting only the m *principal components* of K given by its SVD and discarding the remaining ones. Calculating the SVD of $K = \widehat{X}\widehat{X}^\top$ (again we work with the linear kernel for simplicity) we have

$$K = \widehat{X}\widehat{X}^\top = U\Sigma^2 U^\top.$$

Then

$$\widehat{X} = U\Sigma V^T \Rightarrow V = \widehat{X}^\top U \Sigma^{-1}$$

and matrix $S \in \mathbb{R}^{d\times m}$, in the above notation, can be taken as

$$S = V_m = \widehat{X}^\top U_m \Sigma_m^{-1}$$

where we called $U_m \in \mathbb{R}^{n\times m}$ the matrix of the first m columns of U (the order is given by the magnitude of the corresponding eigenvalues) and $\Sigma_m \in \mathbb{R}^{m\times m}$ the matrix of the biggest m eigenvalues.
With v_j the j-th column of V_m, we can rewrite the embedding of any x as

$$(x_m)_j = x^\top v_j = \sum_{i=1}^n \underbrace{x^\top x_i}_{k(x,x_i)} \frac{u_j^i}{\sigma_j},$$

with $(u_j, \sigma_j^2)_j$ the couples of eigenvectors and eigenvalues of K and $x_1, \dots, x_n$ the data.
With $\widehat{X}_m \in \mathbb{R}^{n\times m}$ the matrix of the embedded data, we consider now the simplified problem

$$\min_{c\in\mathbb{R}^m} \frac{1}{n} \left\| \widehat{X}_m c - \widehat{Y} \right\|^2,$$

the solution is

$$\widehat{c}_m = \widehat{X}_m^\top \left(\widehat{X}_m \widehat{X}_m^\top \right)^{-1} \widehat{Y}$$

with $\widehat{c}_m \in \mathbb{R}^m$. Our predictor becomes

$$\widehat{f}_{\lambda,m}(x) := \widehat{c}_m^\top x.$$

This is usually called *principal component regression* in statistics. Still, computing the SVD for K requires $O(n^3)$ in time and $O(n^2)$ in memory.

3.2 Random Projection

A possible solution to avoid the complexity of computing the SVD is to work with random projections methods. We briefly present here some of the most popular approaches.

Sketching Let take S as a random $d \times m$ matrix such that $S_{ij} \sim \mathcal{N}(0,1)$ and the linear embedding $\widehat{X}_m \in \mathbb{R}^{n \times m}$ as

$$\widehat{X}_m = \widehat{X} S.$$

It's important to note that if $\tilde{x} = S^\top x$ and $\tilde{x}' = S^\top x'$, then

$$\frac{1}{m}\mathbb{E}[\tilde{x}^\top \tilde{x}'] = \frac{1}{m}\mathbb{E}[x^\top S S^\top x'] = x^\top \mathbb{E}[S S^\top] x' = \frac{1}{m} x^\top \sum_{j=1}^{m} \mathbb{E}[s_j s_j^\top] x' = x^\top x'.$$

This means that inner products and norms distances are preserved in expectation. Consider again the regularized ERM problem

$$\min_{c \in \mathbb{R}^m} \frac{1}{n} \left\| \widehat{X}_m c - \widehat{Y} \right\|^2 + \lambda \|c\|^2.$$

The solution of the problem can be written as

$$\widehat{c}_m = \left(\widehat{X}_m^\top \widehat{X}_m + \lambda n I \right)^{-1} \widehat{X}_m^\top \widehat{Y},$$

and our predictor becomes

$$\widehat{f}_{\lambda,m}(x) = x^\top S \widehat{c}_m.$$

Computing $\widehat{c}_m$ is time $\mathrm{O}\left(nm^2 + ndm\right)$ and memory $\mathrm{O}(nm)$, that can lead to huge improvements with respect to previous $O(n^3)$ and $O(n^2)$ when m is small.

Random Features The idea behind random feature is again to approximate K and deal with a lower dimensional object. Consider kernels having the particular form

$$K(x, x') = \int_v \varphi(x, v) \varphi\left(x', v\right) d\mu(v),$$

where $d\mu$ is a probability distribution on some space $\mathcal{V}$ and $\varphi(x, v) \in \mathbb{R}$. We can then approximate the expectation by an empirical average

$$\tilde{K}\left(x, x'\right) = \frac{1}{m} \sum_{i=1}^{m} \varphi\left(x, v_i\right) \varphi\left(x', v_i\right),$$

where the v_i's are sampled i.i.d. from $d\mu$. We can thus use an explicit feature representation $\widehat{\varphi}(x) = \left(\frac{1}{\sqrt{m}} \varphi\left(x, v_i\right) \right)_{i \in \{1, \ldots, m\}}$, and defining $\widehat{X}_m^\top := (\widehat{\varphi}(x_1), \ldots, \widehat{\varphi}(x_n))$ we have

$$\widehat{f}_{\lambda,m}(x) := \widehat{\varphi}(x)^\top \widehat{c}_m, \quad \text{with} \quad \widehat{c}_m := \left(\widehat{X}_m^\top \widehat{X}_m + \lambda n I \right)^{-1} \widehat{X}_m^\top \widehat{Y},$$

For this scheme to makes sense, the number m of random features has to be significantly smaller than n, which is often sufficient in practice (see [Rudi and Rosasco, 2017]).
A simple example of this technique is given by Random Fourier features [Rahimi and Recht, 2007] and Gaussian kernel.

Example 2 (Random Fourier features)**.** *If we write the Gaussian kernel as $K(x,x') = G(x-x')$, with $G(z) = e^{-\frac{1}{2\sigma^2}\|z\|^2}$, for a $\sigma > 0$, then since the inverse Fourier transform of G is a Gaussian, and using a basic symmetry argument, it is easy to show that*

$$G\left(x-x'\right) = \frac{1}{Z}\int\int_0^{2\pi} \sqrt{2}\cos\left(w^\top x + b\right)\sqrt{2}\cos\left(w^\top x' + b\right) e^{-\frac{\sigma^2}{2}\|w\|^2} dwdb$$

where Z is a normalizing factor. Then, the Gaussian kernel has an approximation of the form (5) with $\widehat{\varphi}(x) = m^{-1/2}\left(\sqrt{2}\cos\left(w_1^\top x + b_1\right), \ldots, \sqrt{2}\cos\left(w_m^\top x + b_m\right)\right)$, and $w_1, \ldots, w_m$ and $b_1, \ldots, b_m$ sampled independently from $\frac{1}{Z}e^{-\sigma^2\|w\|^2/2}$ and uniformly in $[0, 2\pi]$, respectively.

Note that dimension reduction is performed independently of the input data, that is the random feature functions $\varphi(\cdot, v_i)$ are selected before the data are observed. This is opposed to the column sampling scheme that we will study next which is a data-dependent dimension reduction scheme.

Nyström Our last example of random projection is Nyström subsampling. As seen in the previous chapter, applying the representer theorem, the solution $\widehat{f}_\lambda$ of Problem (3.1) can be written as

$$\widehat{f}_\lambda(x) = \sum_{i=1}^n \widehat{\alpha}_i K(x_i, x) \qquad \text{with} \qquad \widehat{\alpha} = \left(K_n + \lambda n I\right)^{-1} y$$

with K_n the $n \times n$ kernel matrix, i.e. $K_{ij} = K(x_i, x_j)\ \forall i, j \in [n]$. Note that this means that we can restrict the minimization in (3.1) to the smaller space,

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

The simple idea of Nyström method is to further reduce this space minimizing the objective of (3.1) over

$$\mathcal{H}_m = \left\{ f \mid f = \sum_{i=1}^m c_i K(\tilde{x}_i, \cdot), \quad c_i, \ldots, c_m \in \mathbb{R} \right\}$$

where $m \leqslant n$ and $\{\tilde{x}_1, \ldots, \tilde{x}_m\}$ is a random subset of the n input points in the training set. The solution $\widehat{f}_{\lambda,m}$ of the corresponding minimization problem can now be written as,

$$\widehat{f}_{\lambda,m}(x) = \sum_{i=1}^m \widehat{c}_i K(\tilde{x}_i, x) \qquad \text{with} \qquad \widehat{c} = \left(K_{nm}^\top K_{nm} + \lambda n K_{mm}\right)^\dagger K_{nm}^\top \widehat{Y}, \tag{3.3}$$

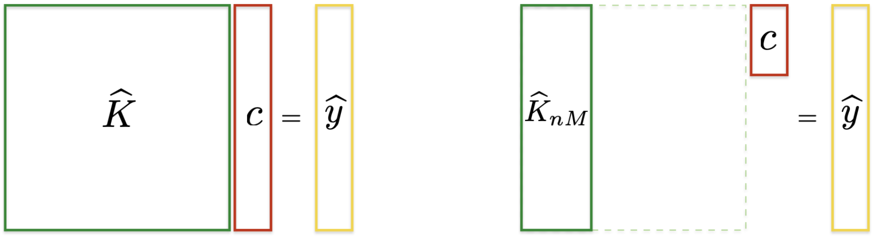
where $A^\dagger$ denotes the Moore-Penrose pseudoinverse of a matrix A, and $(K_{nm})_{ij} = K(x_i, \tilde{x}_j)$, $(K_{mm})_{kj} = K(\tilde{x}_k, \tilde{x}_j)$ with $i \in \{1, \ldots, n\}$ and $j, k \in [m]$.
This approach can be also seen as a way of approximating the kernel matrix K with

$$K \approx K_{nm} K_{mm}^{-1} K_{mn}.$$

Eq. (3.3) shows why this method is often referred to as column subsampling, where random columns of the kernel matrix K_n are sampled to reduce the dimensionality of the problem such that

$$\underbrace{f(x) = \sum_{i=1}^{n} k(x, x_i)c_i}_{\text{ERM predictor}} \to Kc = \widehat{Y} \qquad \Rightarrow \qquad \underbrace{f(x) = \sum_{i=1}^{m} k(x, \tilde{x}_i)c_i}_{\text{Nyström predictor}} \to K_{nm}c = \widehat{Y}$$

which graphically can be seen in the following picture:



This approximation improves the complexity from $O\left(n^3\right)$ in time and $O\left(n^2\right)$ in space to $O\left(nm^2 + m^3\right)$ and $O(nm)$, respectively.

In the next chapter we will focus on Nyström method and we will show our original results when applying it to generic Lipschitz convex losses. Before that, we briefly summarize here some known results about ERM with Nyström subsampling when the square loss is considered.

3.3 Empirical Risk Minimization on Random Subspaces with Square Loss

Before presenting our results for convex loss functions in the next chapter, we summarize here the known results for square loss proved in [Rudi et al., 2015]. As shown before, such a simplified estimator is cheaper to compute with respect to the original ERM solver and this allows to deal with kernel methods and large scale dataset. The main question addressed by this work is if this simplification leads to a drop in the predictive ability of the algorithm. The success of the paper lies in the fact the answer is actually negative and, up a certain point, there is no tradeoff between computational savings and accuracy.
We report here a simplified version of their main assumptions and results. The hypothesis space is assumed to be an RKHS $\mathcal{H}$ with bounded kernel, P the joint distribution of X, Y and P_X the marginal probability distribution of X.
The first basic assumption is that the excess risk $\mathcal{E}$ admits at least a minimizer.

Assumption 1. *There exists an $f_{\mathcal{H}} \in \mathcal{H}$ such that*

$$\mathcal{E}\left(f_{\mathcal{H}}\right) = \min_{f \in \mathcal{H}} \mathcal{E}(f)$$

The second assumption is known as *capacity condition.*

Assumption 2 (Capacity condition)**.** *Defined $d_\alpha := \operatorname{Tr}((\Sigma + \alpha \mathrm{I})^{-1}\Sigma)$ the effective dimension of $\mathcal{H}$, assume*

$$d_\alpha = O(\alpha^{-p}), \qquad 0 < p \leqslant 1 \tag{3.4}$$

where Σ is the covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$, $\langle f, \Sigma g\rangle_{\mathcal{H}} = \int_X f(x)g(x)dP_X(x)$, $\forall f, g \in \mathcal{H}$.

Condition (3.4) quantifies the capacity assumption and is related to covering/entropy number conditions [Steinwart and Christmann, 2008]. Roughly speaking, the effective dimension d_α controls the complexity of the hypothesis space $\mathcal{H}$ according to the marginal measure P_X [Caponnetto and De Vito, 2007]. In particular, Condition (3.4) is ensured if the eigenvalues $(\sigma_i)_i$ of Σ satisfy a polynomial decaying condition $\sigma_i \sim i^{-\frac{1}{p}}$ (see Proposition 8 in Appendix F). Note that, when the kernel is bounded, the operator Σ is trace class and then Condition (3.4) always holds for $p = 1$.

This following assumption is usually called *source condition.*

Assumption 3 (Source condition)**.** *There exists $r \geqslant 0, 1 \leqslant R < \infty$, such that*

$$\left\|\Sigma^{-r} f_{\mathcal{H}}\right\|_{\mathcal{H}} < R. \tag{3.5}$$

Intuitively, it quantifies the degree to which $f_{\mathcal{H}}$ can be well approximated by functions in the RKHS $\mathcal{H}$ and allows to control the bias/approximation error of a learning solution. For $s = 0$, it is always satisfied. For larger s, we are assuming $f_{\mathcal{H}}$ to belong to subspaces of $\mathcal{H}$ that are the images of the fractional compact operators Σ^s [Rudi et al., 2015, Engl et al., 1996].

Theorem 5 (Theorem 1 in [Rudi et al., 2015])**.** *Given a bounded kernel, i.e. $\sup_{x \in \mathcal{X}} k(x, x) = \kappa^2 < +\infty$, n large enough, and under the above capacity and source conditions with $v = \min(r, 1/2)$, then, let $\delta > 0$, with probability at least $1 - \delta$:*

$$L\left(\widehat{f}_{\lambda,m}\right) - L\left(f_{\mathcal{H}}\right) \lesssim n^{-\frac{2v+1}{2v+p+1}}$$

with $\widehat{f}_{\lambda,m}$ as in Eq. (3.3), $\lambda \sim n^{-\frac{1}{2v+p+1}}$ and

$$m \gtrsim \frac{1}{\lambda} \log \frac{1}{\delta\lambda}.$$

It can be shown that this rate is optimal in minimax sense [Caponnetto and De Vito, 2007, Steinwart et al., 2009]. What is worth to notice is the admissible choices for the number of Nyström centers m: take for example $r = 0$ (no source condition) and $p = 1$ (always granted), then the optimal rate (under these assumptions)

$$L\left(\widehat{f}_{\lambda,m}\right) - L\left(f_{\mathcal{H}}\right) \lesssim \sqrt{\frac{1}{n}}$$

can be reached with only

$$m \sim \sqrt{n}$$

neglecting log terms. Therefore, going back to the complexity analysis of Nyström algorithm in the previous section, this approach leads to complexity $O(nm^2 + m^3) = O(n^2)$ in time and $O(nm) = O(n\sqrt{n})$ in memory, in contrast with respectively $O(n^3)$ and $O(n^2)$ complexities of the standard not projected ERM approach.

In the following chapter we will extend this kind of results to generic convex losses, beyond square loss. The case of square loss and the result in Theorem 5 for a specific regime will be recovered as a particular example.

Chapter 4

ERM on Random Subspaces with General Convex Losses

In this chapter we will present our original work that can be found in [Della Vecchia et al., 2021], together with some our extensions not published yet.

4.1 Setting

We start introducing the learning setting and the assumptions we consider. Let $\mathcal{H}$ be a real separable Hilbert space with scalar product $\langle \cdot, \cdot \rangle$ and $\mathcal{Y}$ a Polish space, *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 denote by P their joint distribution defined on the Borel σ-algebra of $\mathcal{H} \times \mathcal{Y}$.
Let $\ell : \mathcal{Y} \times \mathbb{R} \to [0, \infty]$ be a loss function and

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

the corresponding expected risk already defined in Section 1.
As already described there, what we are interested in is to solve the problem

$$\inf_{w \in \mathcal{H}} L(w), \tag{4.1}$$

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 data are finite, we cannot expect to solve the problem exactly. 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 $(x_i, y_i)_{i=1}^n$. We make the following assumptions on the data distributions and the loss.

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

We recall that a random vector X taking value in a Hilbert space $\mathcal{H}$ is called C-sub-gaussian if

$$\|\langle X, u\rangle\|_p \leqslant C\sqrt{p}\|\langle X, u\rangle\|_2 \qquad \forall u \in \mathcal{H}, p \geqslant 2, \tag{4.2}$$

where $\|\langle X, u\rangle\|_p^p = \mathbb{E}\left[\|\langle X, u\rangle\|^p\right]$ [Koltchinskii and Lounici, 2014]. Note that (4.2) 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

$$\| \langle X, u\rangle \|_{\psi_2} = \sup_{p \geqslant 2} \frac{\|\langle X, u\rangle\|_p}{\sqrt{p}} \tag{4.3}$$

should be bounded from above by the L_2-norm $\|\langle X, u\rangle\|_2$. In particular, we note 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

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

is a trace-class positive operator. We define the effective rank of Σ as

$$r_\Sigma = \frac{\mathrm{Tr}\Sigma}{\|\Sigma\|} \tag{4.4}$$

where $\mathrm{Tr}\,\Sigma = \mathbb{E}[\|X\|^2]$ is the trace of Σ. We recall the already mentioned definition of the effective dimension [Zhang, 2005, Caponnetto and De Vito, 2007], for $\alpha > 0$, as

$$d_\alpha = \mathrm{Tr}((\Sigma + \alpha I)^{-1}\Sigma) = \sum_j \frac{\sigma_j}{\sigma_j + \alpha} \tag{4.5}$$

where $(\sigma_j)_j$ are the strictly positive eigenvalues of Σ, 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_α is always finite since Σ is trace class.

The next assumption is on the loss function.

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*

$$|\ell(y, a) - \ell(y, a')| \leq G|a - a'| \quad \forall y \in \mathcal{Y} \quad \textit{and} \quad a, a' \in \mathbb{R}. \tag{4.6}$$

We also assume $\ell_0 = \sup_{y \in \mathcal{Y}} \ell(y, 0) < +\infty$ for all $y \in \mathcal{Y}$.

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

Example 3. *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 examples of loss functions satisfying the above assumption.

Example 4. *The main examples are*

(a) hinge loss:

$$\ell(y, a) = |1 - ya|_+ = \max\{0, 1 - ya\} \qquad \mathcal{Y} = \{-1, 1\} \tag{4.7}$$

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

(b) logistic loss

$$\ell(y, a) = \log(1 + e^{-ya}) \qquad \mathcal{Y} = \{-1, 1\} \tag{4.8}$$

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

(c) square loss

$$\ell(y, a) = (y - a)^2 \qquad \mathcal{Y} \subseteq [-M, M], \tag{4.9}$$

which is convex and differentiable with $G_{loc} = 2M$ (locally Lipschitz with $a \in [-M, M]$) and $\ell_0 = M^2$.

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

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

which is not convex and cannot be directly studied using the theory we will present in this chapter. A way of dealing with the 0-1 loss to obtain statistical bounds for it will be described in Chapter 5.

4.2 Empirical risk minimization

In Section 1.1.3 we presented a classical approach to derive approximate solutions 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

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

and then considering the (regularized) empirical risk minimization (ERM) based on the solution of the problem,

$$\min_{w \in \mathcal{H}} \widehat{L}_\lambda(w), \qquad \widehat{L}_\lambda(w) = \widehat{L}(w) + \lambda \|w\|^2, \tag{4.10}$$

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] introduced in Section 2.2, there exists $c = \widehat{c}_\lambda \in \mathbb{R}^n$ such that

$$\widehat{w}_\lambda = \widehat{X}^\top c \in \operatorname{span}\{x_1, \ldots, x_n\}, \tag{4.11}$$

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

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

The explicit form of the coefficient vector c depends on the considered loss function. In Section 4.2.1 we briefly recall some possible approaches to compute c, whereas in Section 4.2.2 we analyze the statistical properties of the above estimator.

Example 5 (Representer theorem for kernel machines)**.** *As seen in Chapter 2, in the context of kernel methods (see also Example 3), the above discussion, and in particular* (4.11)*, can be easily adapted. Indeed, the parameter w corresponds to a function $f \in \mathcal{H}$ in the RKHS, while the norm $\| \cdot \|$ is the RKHS norm $\| \cdot \|_{\mathcal{H}}$. Eq.* (4.11) *simply states that there exist constants c_i such that the solution of the regularized ERM can be written as $\widehat{f}_\lambda(x) = \sum_{i=1}^{n} K(x, x_i) c_i \in span\{K_{x_1}, \ldots, K_{x_n}\}$.*

4.2.1 Computational aspects

Problem (4.10) can be solved in many ways and we provide below some basic considerations. If $\mathcal{H}$ is finite dimensional, gradient methods can be used. For example, the subgradient method [Boyd and Vandenberghe, 2004] applied to (4.10) gives, for some suitable w_0 and step-size sequence $(\eta_t)_t$,

$$w_{t+1} = w_t - \eta_t \left(\frac{1}{n} \sum_{i=1}^{n} y_i x_i g_i(w_t) + 2\lambda w_t \right), \tag{4.12}$$

where for all $(y_i, x_i)_{i=1,\dots,n}$, $g_i(w) \in \partial\ell(y_i, \langle w, x_i\rangle)$ 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 per iteration cost is $O(nd)$ in time and memory. Clearly, other variants can be considered, for example adding a momentum term [Nesterov, 2018], using 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 (4.12) satisfies $w_t = \widehat{X}^\top c_{t+1}$, with

$$c_{t+1} = c_t - \eta_t \left(\frac{1}{n} \sum_{i=1}^{n} y_i e_i g_i(\widehat{X}^\top c_t) + 2\lambda c_t \right), \tag{4.13}$$

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 the inner product. Also in this case, a number of 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 6 (Hinge loss & SVM). *Considering problem (4.10) 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 (4.13) we can take $g_i(w) = -\mathbb{1}_{[y_i\langle w, x_i\rangle \leq 1]}$.*

4.2.2 Statistical analysis

In this section, we summarize the main statistical properties of the regularized ERM under the sub-gaussian hypothesis in Assumption 1. In the following Theorem 6 we provide a finite sample bound on the excess risk of $\widehat{w}_\lambda$ without assuming the existence of w^* (that instead will be assumed in Theorem 7 via Assumption 3). Towards this end, we introduce the approximation error,

$$\mathcal{A}(\lambda) = \inf_{w \in \mathcal{H}} [L(w) + \lambda \|w\|^2] - \inf_{w \in \mathcal{H}} L(w)\,. \tag{4.14}$$

Note that, if w_* exists, then $\mathcal{A}(\lambda) \leqslant \lambda \|w_*\|^2$. More generally, the approximation error decreases with λ and learning rates can be derived assuming a suitable decay.

Theorem 6. *Under Assumptions 1 and 2, fix $\lambda > 0$ and $0 < \delta < 1$. Then, with probability at least $1 - \delta$,*

$$\begin{aligned} L(\widehat{w}_\lambda) - \inf_{w \in \mathcal{H}} L(w) <& 2\mathcal{A}(\lambda) + \frac{D^2 G^2 C^2 \|\Sigma\| ((\sqrt{r_\Sigma} + K)^2 + (\sqrt{r_\Sigma} + \sqrt{\log(1/\delta)})^2)}{4\lambda n} + \\ &+ \frac{DGC(\sqrt{r_\Sigma} + K)\|\Sigma\|^{\frac{1}{2}} + D\ell_0(K + \sqrt{\log(1/\delta)})}{\sqrt{n}}. \end{aligned} \tag{4.15}$$

where C and G are the constants defined respectively in (4.2) and (4.6), D is a numerical constant and

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

The theorem can be easily extended to non-centered sub-gaussian variables.
Notice that the same result is well known for bounded random variables, see for example [Steinwart and Christmann, 2008, Shalev-Shwartz et al., 2010]. We are not aware of a reference for the sub-gaussian case. 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 would yield a suboptimal dependence in λ. Finally, a similar rate, though only in expectation, can be derived through a stability argument [Bousquet and Elisseeff, 2002, Shalev-Shwartz et al., 2010].

The bound (A.3) shows that the learning rate depends on some a-priori assumption on the distribution that allows to control the approximation error $\mathcal{A}(\lambda)$. The simplest assumption is that the best in the model exists.

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

Under the above condition, we have the following result.

Theorem 7. *Under Assumption 1, 2, and 3, take $\lambda > 0$ and $0 < \delta < 1$, then with probability at least $1-\delta$:*

$$\begin{aligned} L(\widehat{w}_\lambda) - L(w_*) <& \lambda\|w_*\|^2 + \frac{D^2G^2C^2(\sqrt{r_\Sigma}+K)^2\|\Sigma\|}{4\lambda n} + \frac{DGC(\sqrt{r_\Sigma}+K)\|\Sigma\|^{\frac{1}{2}}+}{\sqrt{n}} + \\ &+ \frac{D\ell_0(K+\sqrt{\log(8/\delta)})}{\sqrt{n}} + \frac{DGC\|\Sigma\|^{\frac{1}{2}}\|w_*\|\left(\sqrt{r_\Sigma}+\sqrt{\log(8/\delta)}\right)}{\sqrt{n}}. \end{aligned} \tag{4.16}$$

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

$$L(\widehat{w}_{\lambda_n}) - L(w_*) = O(\|w_*\|\sqrt{\log(1/\delta)/n}), \tag{4.17}$$

up to a $\log\log n$ term.

As above, the proof is given in Appendix A. In a nutshell, what Thm. 7 shows is 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 $w_* \in \mathcal{X}$ exists. With the choice $\lambda \asymp \sqrt{1/n}$ it holds that

$$L(\widehat{w}_\lambda) - \inf_{w\in\mathcal{H}} L(w) = O(\sqrt{1/n}), \tag{4.18}$$

which provides a benchmark for the results in the next sections.

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

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

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.

4.3 ERM on random subspaces

As explained in the Chapter 3, though the ERM estimator $\widehat{w}_\lambda$ achieves sharp rates, from a computational point of view it can be very expensive for large datasets. To overcome this issue, we apply here the same idea presented in Chapter 3 and we study a variant of ERM based on considering a subspace $\mathcal{B} \subset \mathcal{H}$ and the corresponding regularized ERM problem,

$$\min_{\beta\in\mathcal{B}} \widehat{L}_\lambda(\beta), \tag{4.19}$$

with $\widehat{\beta}_\lambda$ the unique minimizer. As clear from (4.11), choosing $\mathcal{B} = \mathcal{H}_n = \operatorname{span}\{x_1, \ldots, x_n\}$ is not a restriction and yields the same solution as considering (4.10). From this observation a natural choice is to consider for $m \leq n$, recovering the Nyström approach introduced in Section 3.2:

$$\mathcal{B}_m = \operatorname{span}\{\widetilde{x}_1, \ldots, \widetilde{x}_m\} \tag{4.20}$$

with $\{\widetilde{x}_1, \ldots, \widetilde{x}_m\} \subset \{x_1, \ldots, x_n\}$ 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.*

$$\widehat{\beta}_{\lambda,m} = \operatorname*{argmin}_{\beta\in\mathcal{B}_m} \widehat{L}_\lambda(\beta). \tag{4.21}$$

In the rest of the thesis all the results are valid when the Nyström points are selected using approximate leverage scores (ALS) sampling. Recall that leverages scores are defined as [Drineas et al., 2012]:

$$l_i(\alpha) = \left\langle x_i, (\widehat{X}\widehat{X}^\top x + \alpha In)^{-1} x_i \right\rangle \qquad i = 1, \dots, n \tag{4.22}$$

where $\alpha > 0$. Since in practice the leverage scores $l_i(\alpha)$ are expensive to compute, approximations have been considered [Drineas et al., 2012, Cohen et al., 2015, Alaoui and Mahoney, 2015, Rudi et al., 2018]. In particular, we consider approximations of the form described in the following definition.

Definition 4 (Approximate leverage scores sampling (ALS))**.** *Let $(l_i(\alpha))_{i=1}^n$ be the leverage scores (4.22). Given $\alpha_0 > 0$ and $T \geqslant 1$, we say that a family $(\widehat{l}_i(\alpha))_{i=1}^n$ is (T, α_0)-approximate leverage scores with confidence $\delta \in (0,1)$ if*

$$\frac{1}{T} l_i(\alpha) \leqslant \widehat{l}_i(\alpha) \leqslant T l_i(\alpha), \qquad \forall i \in \{1, \dots, n\}, \;\; \alpha \geqslant \alpha_0, \tag{4.23}$$

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 $Q_\alpha(i) = \widehat{l}_i(\alpha) / \sum_j \widehat{l}_j(\alpha)$.

We now focus on the computational benefits of considering ERM on random subspaces and we analyze the corresponding statistical properties in Section 4.3.2.

4.3.1 Computational aspects

The choice of $\mathcal{B}_m$ as in (4.20) allows to improve computations with respect to (4.11). Indeed, $\beta \in \mathcal{B}_m$ if and only if $\exists b \in \mathbb{R}^m$ s.t. $\beta = \widetilde{X}^\top b$, with $\widetilde{X} : \mathcal{H} \to \mathbb{R}^m$ the matrix with rows the chosen Nyström points. Then, we can replace (4.19) problem with

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

where $\langle \cdot, \cdot \rangle_m$ is the 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 (4.10), considering the reparameterization $a = (\widetilde{X}\widetilde{X}^\top)^{1/2} b$ which leads to

$$\min_{a \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^n \ell\left(y_i, \langle a, \mathbbm{x}_i \rangle_m\right) + \lambda \left\| a \right\|_m^2, \tag{4.24}$$

where for all $i = 1, \dots, n$, we defined the embedding $x_i \mapsto \mathbbm{x}_i = ((\widetilde{X}\widetilde{X}^\top)^{1/2})^\dagger \widetilde{X} x_i$ and with $\| \cdot \|_m$ we denoted the norm in $\mathbb{R}^m$. Note that the computation of the embedding $x_i \to \mathbbm{x}_i$

only involves the inner product in $\mathcal{H}$ and can be computed in $O(m^3 + nm^2 C_K)$ time. The subgradient method for (4.24) 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 (4.13) 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. We add one example.

Example 7 (Kernel methods and Nyström approximations)**.** *Again, following Example 3 and Example 5, our setting can be easily specialized to kernel methods, where* $\beta \in \mathcal{B}_m = \text{span}\{\widetilde{x}_1, \dots, \widetilde{x}_m\}$ *is replaced by* $\widetilde{f}(x) = \sum_{i=1}^m K(x, \widetilde{x}_i)\widetilde{c}_i \in \text{span}\{K_{\widetilde{x}_1}, \dots, K_{\widetilde{x}_m}\}$*, while the embedding* $x_i \mapsto \mathbb{x}_i = ((\widetilde{X}\widetilde{X}^\top)^{1/2})^\dagger \widetilde{X} x_i$ *becomes* $x_i \mapsto \mathbb{x}_i = (\widetilde{K}^{1/2})^\dagger (K(\widetilde{x}_1, x_i), \dots, K(\widetilde{x}_m, x_i))^\top$*, with* $\widetilde{K}_{i,j} = K(\widetilde{x}_i, \widetilde{x}_j)$*.*

4.3.2 Statistical analysis

In this section we will show, under a suitable polynomial (or exponential) decay condition on the spectrum of Σ (see (4.28)) that,

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

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

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

where the exponent p controls how strong the polynomial decay condition is (see (4.28)). Compared to the results for exact ERM (4.18), we get the same convergence rate up to a log factor, but the computational 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 is obtained for exponential decay in which case we can take $m \simeq \log^2 n$ Nyström points. We observe that under the above decay conditions on the spectrum of Σ classical ERM algorithm achieves fast rates. In Section 4.4, we will show that also randomized ERM can achieve fast rates, but this will require a more refined analysis.

We now state the detailed results. We recall that the Nyström points are sampled according to ALS, see Definition 4.

Theorem 8. *Under Assumption 1, 2 and 3, fix $\alpha, \lambda, \delta > 0$. Then, with probability at least $1-\delta$:*

$$L(\widehat{\beta}_{\lambda,m}) - L(w_*) \lesssim \frac{\log(1/\delta)}{\lambda n} + \frac{\|w_*\|\sqrt{\log(1/\delta)}}{\sqrt{n}} + \sqrt{\alpha}\|w_*\| + \lambda\|w_*\|^2 \tag{4.25}$$

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 8 with explicit constants is given in Appendix B, here we only add some comments. Note that

$$d_\alpha = \int \left\langle w, (\Sigma + \alpha I)^{-1} w \right\rangle dP_X(w) \leqslant \int \|w\|^2 \left\|(\Sigma + \alpha I)^{-1}\right\| dP_X(w) \leqslant \alpha^{-1}\mathbb{E}[\|X\|^2] \lesssim \alpha^{-1}, \tag{4.26}$$

using 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

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

we get that with high probability

$$L(\widehat{\beta}_{\lambda_n,m}) - L(w_*) \lesssim \frac{\|w_*\|\sqrt{\log(1/\delta)}}{\sqrt{n}} \tag{4.27}$$

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 decay conditions on the spectrum of the covariance operator Σ, 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

$$\sigma_j \lesssim j^{-\frac{1}{p}} \tag{4.28}$$

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

$$\sigma_j \lesssim e^{-\beta j}. \tag{4.29}$$

Under the above conditions, we have the following result.

Theorem 9. *Under the assumptions of Theorem 8, fix $\delta > 0$, then, with probability at least $1-\delta$:*

$$L(\widehat{\beta}_{\lambda,m}) - L(w_*) \lesssim \frac{\log(1/\delta)}{\lambda n} + \frac{\|w_*\|\sqrt{\log(1/\delta)}}{\sqrt{n}} + \sqrt{\alpha}\|w_*\| + \lambda\|w_*\|^2 \tag{4.30}$$

and, with the choice

(a) for the polynomial decay (4.28)

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

(b) for the exponential decay (4.29)

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

then, it holds that

$$L(\widehat{\beta}_{\lambda_n,m}) - L(w_*) \lesssim \frac{\|w_*\|\sqrt{\log(1/\delta)}}{\sqrt{n}} \tag{4.31}$$

The proof of the above result is given in Appendix B. Theorem 9 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. However, note that our bound on the number of Nyström points is, in the case of square loss, worse than the bound in [Rudi et al., 2015]. In Section 4.5, by specializing the analysis for smooth losses and exploiting the special structure of the quadratic loss, we obtain the right estimate of Nyström points matching the result in [Rudi et al., 2015].

Theorem 9 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. 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 random subspaces. Compared to random features, the proof technique 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 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, next, we focus on random subspaces defined by leverage scores sampling and show that the assumption on the eigendecay not only allows for smaller subspace dimensions, but can also lead to faster learning rates.

Remark 2. *Following [Rudi et al., 2015], other choices of $\mathcal{B} \subseteq \mathcal{H}$ are possible. Indeed for any $q \in \mathbb{N}$ and $z_1, \ldots, z_q \in \mathcal{H}$ we could consider $\mathcal{B} = span\{z_1, \ldots, z_q\}$ and derive a formulation as in (4.24) replacing $\widetilde{X}$ with the matrix Z with rows $z_1, \ldots, z_q$. We leave this discussion for future work. We simply state the following result where*

$$\mu_{\mathcal{B}} = \left\| \Sigma^{1/2}(I - \mathcal{P}) \right\|, \tag{4.32}$$

and $\mathcal{P}$ is the projection onto $\mathcal{B}$. Then, for a generic subset $\mathcal{B} \subseteq \mathcal{H}$ we have the following theorem.

Theorem 10. *Choose $\mathcal{B} \subseteq \mathcal{H}$. Under Assumptions 1, 2, 3, fix $\lambda > 0$ and $0 < \delta < 1$, with probability at least $1 - \delta$:*

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

Compared to Theorem 7, 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.

4.4 Fast rates

In this section, we prove that 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 the results, we recall some definitions and basic facts, see [Steinwart and Christmann, 2008, Chapter 6].

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 ℓ can be *clipped* at $M > 0$ if for all $y \in \mathcal{Y}, a \in \mathbb{R}$,

$$\ell(y, a^{cl}) \leqslant \ell(y, a), \tag{4.33}$$

For convex loss functions, as considered in this thesis, 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 can set

$$f_*(x) = \int_{\mathcal{Y}} \ell(y, \varphi(x)) dP(y|x), \tag{4.34}$$

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 is easy to check that hinge loss and square loss with bounded outputs can be clipped. Even if the logistic loss can not be clipped, we will show in Section 4.5.2 how we can easily bypass this issue with an ad hoc fix. 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.

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

Assumption 5 (Universality)**.**

$$\inf_{w \in \mathcal{H}} L(w) = L(f_*). \tag{4.35}$$

Recalling that the target function f_* is the minimizer of the expected error over all possible functions f, condition (4.35) means that f_* can be arbitrarily well approximated by a linear function $\langle w, x \rangle$ for some $w \in \mathcal{H}$. When considering 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 3 it is satisfied by universal kernels [Steinwart and Christmann, 2008]. Condition (4.35) may be relaxed at the cost of an additional approximation term, but the analysis is just lengthier and it won't be discussed in here. A sufficient stronger condition is provided by assuming the target function to be linear (well specified model).

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.

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:*

$$\ell(Y, \langle w, X\rangle^{cl}) \leqslant B, \tag{4.36}$$

$$\mathbb{E}\big[\big\{\ell(Y, \langle w, X\rangle^{cl}) - \ell(Y, f_*(X))\big\}^2\big] \leqslant V(\mathbb{E}[\ell(Y, \langle w, X\rangle^{cl}) - \ell(Y, f_*(X))])^{\theta} \tag{4.37}$$

$$\mathbb{E}\big[\big\{\ell(Y, \langle w, X\rangle) - \ell(Y, f_*(X))\big\}^2\big] \leqslant V(\mathbb{E}[\ell(Y, \langle w, X\rangle) - \ell(Y, f_*(X))])^{\theta} \tag{4.38}$$

Condition (4.36) is called supremum bound [Steinwart and Christmann, 2008] and, thanks to the clipping, it is satisfied by Lipschitz loss functions. Condition (4.37) is called variance bound [Steinwart and Christmann, 2008] 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 classification, the variance bound is implied by so called margin conditions (see Section 5 and Theorem 8.24 in [Steinwart and Christmann, 2008]), and the parameter θ characterizes how easy or hard the classification problem is [Steinwart and Christmann, 2008]. With respect to [Steinwart and Christmann, 2008], condition (4.38) is a technical one that we need in the proof.
To state our result, we will make use again of the approximation error $\mathcal{A}(\lambda)$ defined in (4.14). The following theorem provides fast rates for Nyström algorithm, where we recall the Nyström points are sampled according to ALS, see Definition 4.

Theorem 11. *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 (4.28)

$$L(\widehat{\beta}^{cl}_{\lambda,m}) - L(f_*) \lesssim \Big(\frac{1}{\lambda^p n}\Big)^{\frac{1}{2-p-\theta+\theta p}} + \sqrt{\frac{\alpha\mathcal{A}(\lambda)}{\lambda}} + \Big(\frac{\log(3/\delta)}{n}\Big)^{\frac{1}{2-\theta}} + \frac{\log(3/\delta)}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} + \mathcal{A}(\lambda) \tag{4.39}$$

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 condition (4.29)

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

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 11 is given in Appendix C. Notice that a faster decay condition on the spectrum of Σ leads to improvements in both the excess risk bound and the parameters' choices. As regards the learning rate, under exponential decay in *(b)*, first term of (4.39) improves to $1/n$ up to logarithmic factors. At the same time, the range of admissible α gets larger while the control on the effective dimension gets tighter. Let us comment these results more precisely in the following.

4.4.1 Polynomial decay of Σ

In this section we assume the polynomial decay (4.28) condition on the spectrum of Σ. By omitting numerical constants, logarithmic and higher order terms, Theorem 11 (a) 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

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

then with high probability

$$L(\widehat{\beta}^{cl}_{\lambda_n,m}) - L(f_*) \lesssim n^{-\min\{\frac{2r}{r+1},\frac{r}{r(2-p-\theta+\theta p)+p}\}}. \tag{4.40}$$

The above bound further simplifies when the variance bound (4.37) 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

$$L(\widehat{\beta}^{cl}_{\lambda_n,m}) - L(w_*) \lesssim n^{-\frac{1}{1+p}} \tag{4.41}$$

with the choice

$$\lambda_n \asymp n^{-\frac{1}{1+p}}, \quad \alpha_n \asymp n^{-\frac{2}{1+p}}, \quad m \gtrsim n^{\frac{2p}{1+p}} \log n, \tag{4.42}$$

see Appendix C for the proof.
By comparing bound (4.41) with (4.31), 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* error estimates since the decay of the eigenvalues is used both for the subspace approximation error and the estimation error. As it is clear from (4.41), 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 (4.31). In this case, the subspace defined by leverage scores needs to be chosen of dimension at least $O(n^{2/3})$.
For arbitrary θ and r, bound (4.40) 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 λ 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 θ and r leads to worse bounds than (4.41), which is the best rate in this context. Since, given any acceptable choice of p, r and θ, the quantity $\min\{2p, \frac{p(r+1)}{r(2-p-\theta+\theta p)+p}\}$ takes values in $(0,1)$, the best rate, that differently from before can also be slower than $\sqrt{1/n}$, can always be achieved choosing $m < n$ (up to logarithmic terms).

4.4.2 Exponential decay of Σ

We can further improve the bounds above assuming an exponential decay (4.28) condition on the spectrum of Σ. By omitting numerical constants, logarithmic and higher order terms, from Theorem 11 *(b)* we have that with high probability

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

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

$$\lambda_n \asymp \log n/n^2, \quad \alpha_n \asymp 1/n^2, \quad m \gtrsim \log^2 n, \tag{4.43}$$

with high probability,

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

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

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

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

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

see Appendix C for the proof.

Remark 3. *Whereas the results of Section 4.3.2 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 8 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 (4.42) and (4.43), which provide the optimal convergence rates. Under the assumption that X is bounded, Lemma 8 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 (4.42) and (4.43).*

4.4.3 Comparison with Random Features

We start comparing our results with the ones obtained using random features in [Sun et al., 2018]. Specifically, their Theorem 1 is based on similar assumptions as our result in Eq. (4.41), 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 α in (4.42). 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.
Table 4.1 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 4.1: Comparison among the different regimes using hinge loss.

	Assumptions	Eigen-decay	Rate	m
Theorem 7	1,2,3	/	$n^{-1/2}$	/
Eq. (4.31)	1,2,3	$\sigma_j \lesssim j^{-\frac{1}{p}}$	$n^{-1/2}$	n^p
Eq. (4.31)	1,2,3	$\sigma_j \lesssim e^{-\beta j}$	$n^{-1/2}$	$\log^2 n$
Eq: (4.41)	1,2,6,7*	$\sigma_j \lesssim j^{-\frac{1}{p}}$	$n^{-\frac{1}{1+p}}$	$n^{\frac{2p}{1+p}}$
Eq: (4.40)	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*	$\sigma_j \lesssim j^{-\frac{1}{p}}$	$n^{-\frac{1}{2p+1}}$	$n^{\frac{2p}{2p+1}}$

4.5 Differentiable loss functions

In this section we specify the above results to differentiable losses and, in particular, to quadratic and logistic losses. In both cases, we will provide for this setting equivalent bounds of the ones presented in Theorem 11.

4.5.1 Square loss

Next, we specialized the analysis to square loss defined by (4.9) 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 ℓ can be clipped at 1. A well known variance bound for least squares loss gives that

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

so that variance bound (4.37) holds for $V=16$ and $\theta=1$.
Finally, the least squares loss restricted to $[-1,1]$ is Lipschitz continuous, that is

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

for all $y \in [-1,1]$ and $t,t' \in [-1,1]$.
The following theorem specializes to the square loss the previous states, see Appendix D.0.1 for the proof. As usual the Nyström points are sampled according to ALS, see Definition 4.

Theorem 12. *Under Assumption 1 and the polynomial decay condition (4.28), 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

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

Comparing the above bound and the one in (4.40) with $\theta = 1$, we get the same 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 3 we are able to prove the above results only under the assumption that X is 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 α. We state the result without reporting the proof, which is a variant of the proof of Theorem 12 taking into account the above remark.

Corollary 2. *Assume that X is bounded almost surely, under Assumption 6 and polynomial decay of the spectrum (4.28), fix $\lambda > 0$, $\alpha \gtrsim 1/n$, and $0 < \delta < 1$. Then, with probability at least $1 - 2\delta$:*

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

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

$$\lambda \asymp n^{-\frac{1}{1+p}}, \quad \alpha \asymp n^{-\frac{1}{1+p}}, \quad m \gtrsim n^{\frac{p}{1+p}} \log n, \tag{4.44}$$

with high probability,

$$L(\widehat{\beta}^{cl}_{\lambda,m}) - L(w_*) \lesssim n^{-\frac{1}{1+p}}. \tag{4.45}$$

Remark 4 (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 Σ^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 4.2 the two works return exactly the same rate and the same requirement for m.*

Table 4.2: 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 12	1	$\sigma_j \lesssim j^{-\frac{1}{p}}$	$n^{-\min\{\frac{2r}{r+1},\frac{r}{r+p}\}}$	$n^{\min\{\frac{2p}{r+1},\frac{p}{r+p}\}}$

4.5.2 Logistic loss

As already mentioned, let's start noticing that logistic loss defined by (4.8) cannot be clipped according to (4.33) [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

$$\ell(y, a^{cl}) \leqslant \ell(y,a) + \frac{1}{n}, \tag{4.46}$$

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

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

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. $\ell(y,a^{cl}) \nleqslant \ell(y,a)$, 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 the same excess risk bounds of the square loss in Theorem 12 and Corollary 2 for the logistic loss. The simple adaptation of proofs is given in Appendix D.0.2.

Chapter 5

From Surrogates to Classification Error

In this chapter, we consider 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. $\ell_{0-1}(y, a) := \mathbb{1}_{(-\infty,0]}(y \text{ sign}(a))$. Through out this section, we study how the previous bounds for surrogate losses relate to the 0-1 classification risk. In the following, 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]. The following definition is taken from [Tsybakov, 2004]:

Definition 5. *Distribution P has noise exponent $0 \leqslant \gamma < 1$ if it satisfies one of the following conditions:*

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

$$\Pr[f(X)(2\eta(X) - 1) < 0] \leqslant c \left(L_{0-1}(f) - L^*_{0-1}\right)^\gamma ; \tag{5.1}$$

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

$$\Pr\left[0 < |2\eta(X) - 1| \leqslant \epsilon\right] \leqslant c\epsilon^{\frac{\gamma}{1-\gamma}} ; \tag{5.2}$$

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

We will assume the following low-noise condition:

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

Using Lemma 14 in Appendix F, when dealing with the square loss, there is a standard way of transforming its excess risk bound into the following bound on the classification risk:

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:*

$$L_{0-1}(f) - L^*_{0-1} \lesssim \left(L_{square}(f) - L^*_{square}\right)^{\frac{1}{2-\gamma}} . \tag{5.3}$$

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

For the hinge loss, the bound given by Lemma 13 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 (4.37) through Theorem 8.24 in [Steinwart and Christmann, 2008] (see Lemma 15 in Appendix F). In particular, if we assume a low noise condition with parameter γ, then the variance bound in Assumption 8 is always satisfied for the hinge loss with $\theta = \gamma$.

5.0.1 From square and logistic losses to classification loss

Starting from Theorem 12, 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 (4.14), with respect the square loss and the Nyström points are sampled, as always, accordingly to ALS, see Definition 4.

Theorem 13. *Under Assumptions 1 and 8 and the polynomial decay condition (4.28), 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 analogous bounds hold for logistic loss, up to constant or negligible terms.

5.0.2 From hinge loss to classification loss

Starting from Theorem 11, 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 15 we obtain the following theorem, where $\mathcal{A}_{\text{hinge}}(\lambda)$ is the approximation error, see (4.14), with respect to the hinge loss.

Theorem 14. *Under Assumptions 1, 8 and under polynomial decay condition (4.28), 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 5.1: Comparison between the $0-1$ classification risk upper bounds derived from square, logistic and hinge loss under low noise condition

	Assump	Eigen-decay	Rate	m
Square Loss: Theorem 13	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 14	1,8	$\sigma_j \lesssim j^{-\frac{1}{p}}$	$n^{-\min\{\frac{2r}{r+1}, \frac{r}{r(2-p-\gamma+\gamma p)+p}\}}$	$n^{\min\{2p, \frac{p(r+1)}{r(2-p-\gamma+\gamma p)+p}\}}$

Next, we will discuss the results obtained in Table 4.

5.0.3 Discussion of the results

We want now to compare the two upper bounds we obtained in Theorem 13 and Theorem 14. 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, γ and r the bound for the classification error derived using the hinge loss can always achieve a better rate than the one derived from the square loss. On the other hand, since $\min\{\frac{2p}{r+1}, \frac{p}{r+p}\} \leqslant \min\{2p, \frac{p(r+1)}{r(2-p-\gamma+\gamma p)+p}\}$, the choice of the hinge loss results, according to our upper bounds, 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 γ 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 $m_{\mathrm{s}} = n^{R(2-\gamma)p/r}$ for square loss and $m_{\mathrm{h}} = n^{R(1+r)p/r}$ for hinge loss. Since when $\gamma + r < 1$ then $m_h \geqslant m_s$, we have that, given a fixed rate for the 0-1 loss, using hinge is, according to our upper bounds, computationally *cheaper* than using square loss (see Figure 5.1). This suggests that when the problem is hard, hinge loss seems to be even *less expensive* than the square loss (and viceversa).

Similarly, imagine now to have some budget constraint on m so that we are maybe not allowed to choose its optimal value: which loss will show a faster rate? Again the condition on $\gamma + r$ is the key, with the upper bound for hinge loss going to 0 faster as n increases than the one for square loss, when $\gamma + r < 1$ (see Figure 5.2, where also the saturation effect can be seen).

In summary, when studying the misclassification error using surrogates, the comparison between our two upper bounds obtained from hinge and square loss does not suggest an univocal better choice for all regimes. When the problem is *hard*, i.e. slow decay of the approximation error ($\lambda \ll 1$) and/or strong noise ($\gamma \ll 1$), the upper bound for the hinge loss behaves better than the one for the square loss; the opposite when the problem is *easy*.

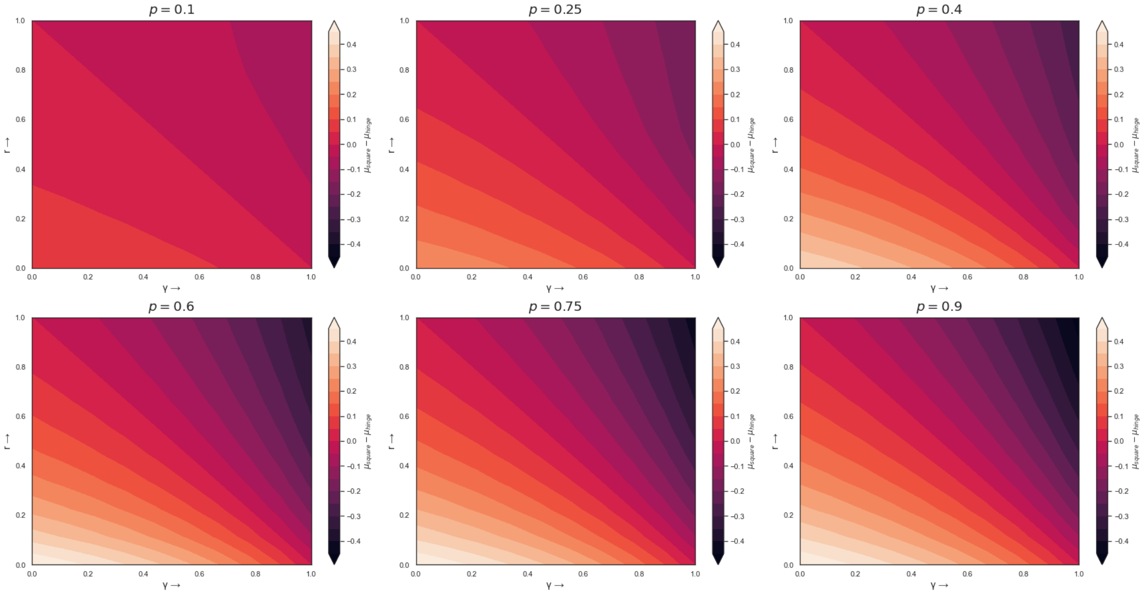


Figure 5.1: Comparison between the number of Nyström points needed according to our bounds for square and hinge loss to get a fixed common rate: the plots above show $\mu_{\text{square}} - \mu_{\text{hinge}}$, where $0 \leqslant \mu \leqslant 1$ is the exponent controlling m, i.e. $m \asymp n^{\mu}$. Light colors represent then the regimes where hinge loss is *cheaper* than square loss.

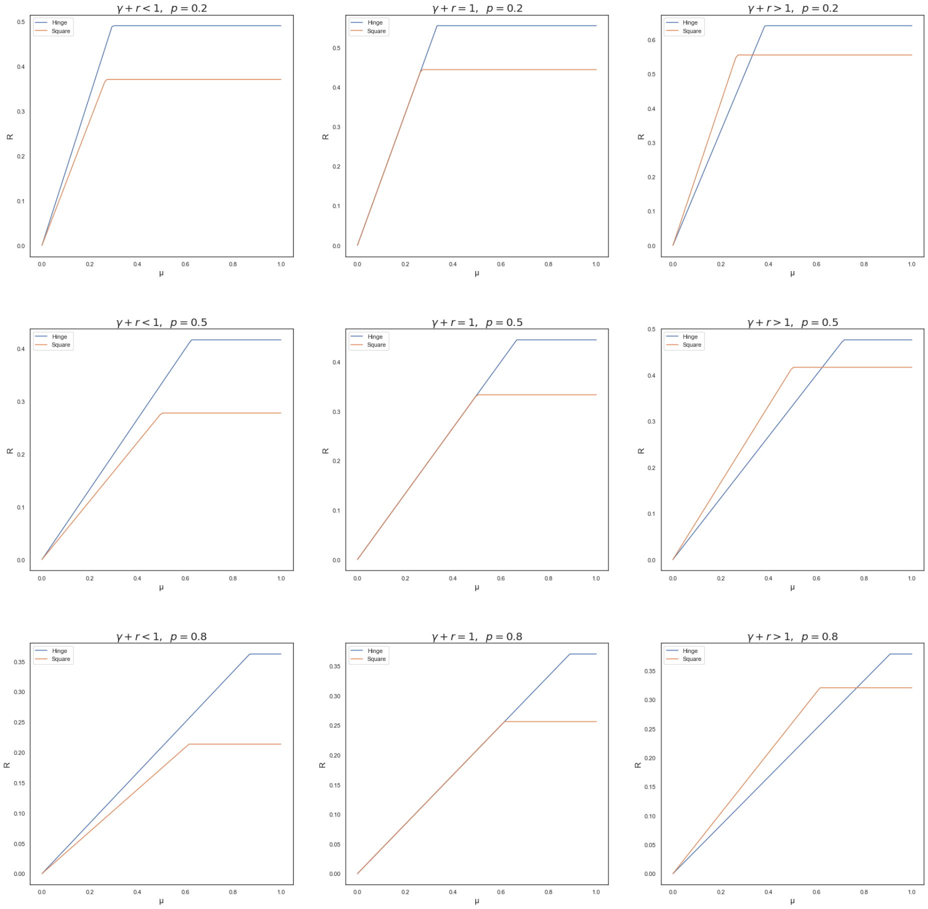


Figure 5.2: Comparison between the rate achieved by the bounds for square and hinge loss varying m: the plots above have R in the y-axis, where $0 \leqslant R \leqslant 1$ is the exponent of the resulting rate, i.e. rate $= n^{-R}$; in the x-axis we have μ, with $m = n^{\mu}$ and $0 \leqslant \mu \leqslant 1$ ($\mu = 1$ 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.

Chapter 6

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 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 11 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 (4.12) (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 4.3. 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 4.3, 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).

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

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.

Results. We compare with linear (used only as baseline) and K-SVM see Table 6.1. 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 6.1 we illustrate the interplay between λ 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.

Table 6.1: 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).

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

[†] 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

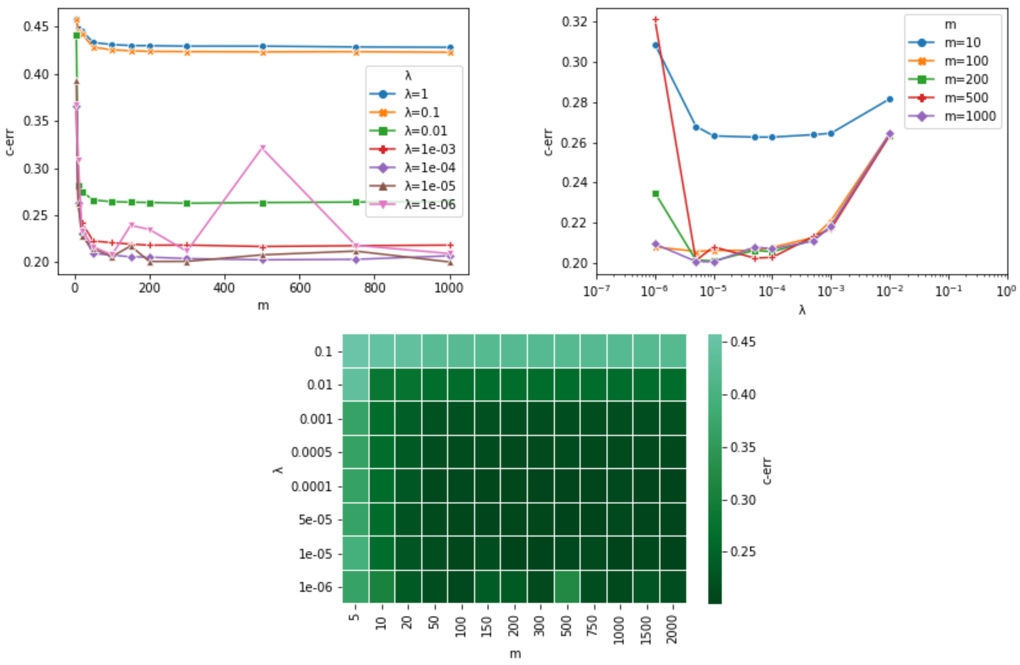


Figure 6.1: The graphs above are obtained from SUSY data set: on the top left we show how c-err measure changes for different choices of λ parameter; in top right figure the focus is on the stability of the algorithm varying λ; on the bottom the combined behavior is presented with a heatmap.

Part II

Plug-in Classifiers for Generalized Performance Metrics

Chapter 7

The Plug-in Approach

In the first part of the thesis we focused on one popular approach for classification based on ERM and surrogate losses. In this second part, we study the so-called plug-in approach for classification. The plug-in approach consists in transforming an estimator of the regression function into a classifier. The way of proceeding is the following: first the optimal classifier is derived at the population level, second all the unknown quantities are replaced with their empirical estimators. We give a simple example of this procedure studying the misclassification error. Let (X, Y) be a random couple taking values in $\mathbb{R}^d \times \{0, 1\}$ with joint distribution P and denote with P_X the marginal distribution of X. Let's define $g^* : \mathbb{R}^d \to \{0, 1\}$ the optimal classifier among all measurable functions such that

$$g^* = \underset{g:\mathbb{R}\to\{0,1\}}{\arg\min}\ \mathbb{P}(Y \neq g(X)).$$

The shape of g^* can be easily derived:

$$\begin{aligned}\mathbb{P}(Y \neq g(X)) &= \mathbb{E}[\mathbb{1}(Y \neq g(X))] = \mathbb{E}[(1-2Y)g(X) + Y] = \\ &= \mathbb{E}[\mathbb{E}[(1-2Y)g(X)|X]] + \mathbb{E}[Y] = \mathbb{E}[g(X)\mathbb{E}[(1-2Y)|X]] + \mathbb{E}[Y] = \\ &= \mathbb{E}[g(X)(1-2\eta(X))] + \mathbb{E}[Y]\end{aligned}$$

where $\eta(X) = \mathbb{E}[Y|X]$ is the regression function and we used some basic properties of expectations and the fact that Y and g takes only values 0 and 1. It's straightforward to notice that, to minimize this quantity, we want our predictor to predict 1 when $1-2\eta(X) < 0$ and 0 viceversa. We recover the classic Bayes predictor in misclassification

$$g^*(\cdot) = \mathbb{1}\{\eta(\cdot) > 1/2\},$$

i.e. a step function depending on the unknown regression function η and with threshold 1/2.

At this point we want to build our plug-in estimator $\widehat{g}^{\mathrm{PI}}$ with the same structure of the optimal g^* but replacing unknown quantities, i.e. quantities depending on the unknown distribution of data, with their empirical estimators derived from some given samples:

$$\widehat{g}^{\mathrm{PI}}(\cdot) = \mathbb{1}\{\widehat{\eta}(\cdot) > 1/2\},$$

where indeed $\widehat{\eta}$ is some estimator of the regression function.

Historically, both ERM based methods and plug-in methods have been studied and strong theoretical guarantees have been proven. Nevertheless, for a long time, ERM approaches have been considered superior. In the context of binary classification, where more theoretical advances have been developed, this belief was initially supported by the results in [Yang, 1999]. In this work they showed that, under general assumptions, rates of convergence faster than $O(1/\sqrt{n})$ cannot not be achieved by plug-in estimators. More than that, this kind of classifiers suffer also the curse of dimensionality. In contrast, as seen in the first part of the thesis, ERM methods, under slightly different hypothesis, can achieve this $O(1/\sqrt{n})$ with the dimensionality playing no role. In addition, under Tsybakov's *margin assumption* [Tsybakov, 2004], also fast convergence rates up to $O(1/n)$ can be reached. These results raised some scepticism around plug-in rules in favour of ERM-based ones. This idea was proved to be wrong in general around a decade after by [Audibert and Tsybakov, 2007] and [Rigollet and Vert, 2009]. The main criticism to the above reasoning is that the two approaches follow different sets of assumptions, with none of them including the other. For this reason, naively comparing the two methods can be misleading and far from legitimate.
In details, let's consider plug-in rules, i.e. classifiers of the form

$$\widehat{g}^{\mathrm{PI}}(X) = \mathbb{1}\left\{\widehat{\eta}(X) \geqslant \widehat{\theta}\right\}$$

with $\widehat{\eta}$ an estimator of the regression function η and $\widehat{\theta} \in \mathbb{R}$ the threshold. For standard binary classification with 0-1 loss (accuracy) the best choice for $\widehat{\theta}$ is exactly $1/2$, as shown above. A typical well adapted assumption for this kind of classifiers can be made on the smoothness of η. In fact, it is possible to show that closeness of $\widehat{\eta}$ to η implies closeness of $\widehat{g}^{\mathrm{PI}}$ to g^*.

Lemma 2 (Theorem 2.2 in [Devroye et al., 2013])**.** *For the error probability of the plug-in decision g^{PI} defined above, we have*

$$\mathbb{P}\{\widehat{g}^{\mathrm{PI}}(X) \neq Y\} - \mathbb{P}\{g^*(X) \neq Y\} \leqslant 2\int_{\mathbb{R}^d} |\eta(x) - \widehat{\eta}(x)| P_X(dx) = 2\mathbb{E}|\eta(X) - \widehat{\eta}(X)|.$$

Proof. If for some $x \in \mathbb{R}^d$, $\widehat{g}^{\mathrm{PI}}(x) = g^*(x)$, then clearly the difference between the conditional error probabilities of $\widehat{g}^{\mathrm{PI}}$ and f^* is zero:

$$\mathbb{P}\{\widehat{g}^{\mathrm{PI}}(X) \neq Y \mid X = x\} - \mathbb{P}\left\{g^*(X) \neq Y \mid X = x\right\} = 0.$$

Otherwise, if $\widehat{g}^{\mathrm{PI}}(x) \neq g^*(x)$ the difference may be written as

$$\begin{aligned}\mathbb{P}\{\widehat{g}^{\mathrm{PI}}(X) \neq Y \mid X = x\} - \mathbb{P}\left\{g^*(X) \neq Y \mid X = x\right\} \\ &= (2\eta(x) - 1)\left(\mathbb{1}\{g^*(x) = 1\} - \mathbb{1}\{\widehat{g}^{\mathrm{PI}}(x) = 1\}\right) \\ &= |2\eta(x) - 1|\mathbb{1}\left\{\widehat{g}^{\mathrm{PI}}(x) \neq g^*(x)\right\}.\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{P}\{\widehat{g}^{\mathrm{PI}}(X) \neq Y\} - \mathbb{P}\{g^*(X) \neq Y\} &= \int_{\mathbb{R}^d} 2|\eta(x) - 1/2|\mathbb{1}\left\{\widehat{g}^{\mathrm{PI}}(x) \neq g^*(x)\right\}P_X(dx) \\ &\leqslant \int_{\mathbb{R}^d} 2|\eta(x) - \widehat{\eta}(x)|P_X(dx)\end{aligned}$$

□

Then, smoothness assumption on η, referred to as *complexity assumption on the regression function* (CAR) [Audibert and Tsybakov, 2007], implies that a good nonparametric estimator (kernel, local polynomial, orthogonal series or other) $\widehat{\eta}$ converges with some rate to the regression function η, as $n \to \infty$. Typically, this approach leads to bounds of the kind

$$L\left(\widehat{g}^{\mathrm{PI}}\right) - L(g^*) = O\left(n^{-\beta/(2\beta+d)}\right) \tag{7.1}$$

with β smoothness parameter. This rate is always slower than $n^{-1/2}$ and it deteriorates dramatically as the dimension d increases. Moreover, under general assumptions, this bound can be proved to be optimal in minimax sense, see [Yang, 1999].
On the contrary, results exploiting ERM techniques are known to reach

$$L\left(\widehat{g}^{\mathrm{ERM}}\right) - L(g^*) = O\left(n^{-1/2}\right)$$

rates under slightly different assumptions, with none of them including the others. This makes hard to compare the two approaches fairly.
Nevertheless, the limitation about the slow rate for plug-in classifiers in Eq. (7.1) has been overcome afterwards in [Audibert and Tsybakov, 2007] under Tsybakov's *margin assumption* [Tsybakov, 2004]. The authors proved that, in this framework, plug-in classifiers can also reach fast rates up to $O(1/n)$ and even super-fast rates, revealing that plug-in methods should not be considered inferior to ERM methods and, more importantly, that this new type of assumption on the regression function is a critical point in the general analysis of classification procedures. This chapter is based on the following references: [Audibert and Tsybakov, 2007, Rigollet and Vert, 2009, Chzhen, 2019b, Gaucher et al., 2022, Chzhen, 2020, Devroye et al., 2013, Koyejo et al., 2014]

7.1 Settings and Main Notation

In this chapter we will still consider the setting of supervised learning. We already extensively discussed it in Chapter 1 and we briefly recall it here. The statistical model comes from the fact that the data are probabilistically generated, i.e. there exists a couple of input-label random variables (X, Y) taking values in $\mathbb{R}^d \times \{0, 1\}$ with joint distribution P. We further indicate with P_X the marginal distribution of the feature vector $X \in \mathbb{R}^d$ and with $P_{Y|X}$ the conditional distribution of $Y|X \in \{0, 1\}$. Differently from before we introduce now two different datasets: a labelled one $\mathcal{D}_n^{\mathrm{L}} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ consisting of n i.i.d. copies of $(X, Y) \sim P$ and an unlabelled one $\mathcal{D}_N^{U} = \{X_{n+1}, \ldots, X_{n+N}\}$ consisting of N independent copies of $X \sim P_X$, with $N \geqslant n$. The first one will be used to estimate the regression function $\eta(X) := \mathbb{E}[Y|X]$, that will be a fundamental element in our analysis. As regards the second unlabelled one, it will be used to approximate the marginal distribution P_X by its empirical counterpart $\widehat{P}_X$, with $\widehat{P}_X = \frac{1}{N}\sum_{X \in \mathcal{D}_N^{\mathrm{U}}} \delta_X$. A good estimate of both these quantities will be necessary when designing our predictor, as we will explain in the next chapter. Finally, we call $\mathcal{G}$ the set of all possible classifiers, i.e. any measurable function $g : \mathbb{R}^d \to \{0, 1\}$.

7.2 Generalized Performance Metrics

Given a classifier g, there exist many different measures on how g is expected to perform in predicting Y given X. The most natural way to measure its risk is via its misclassification error $\mathbb{P}(Y \neq g(X))$. Nevertheless, in many applications, practitioners are more interested in optimizing different kind of metrics, for example balancing differently precision and recall of the constructed classifier. To give an example, it is well known that classification accuracy is probably not the best metric when considering rare event classification problems such as medical diagnosis, fraud detection, click rate prediction and text retrieval applications [Gu et al., 2009, He and Garcia, 2009]. Instead, alternative metrics better tuned to imbalanced classification (such as the F-score) are more suitable. Similarly, cost-sensitive metrics may be useful for addressing asymmetry in real-world costs associated with specific classes. In this thesis we study an utility metric that is general enough to recover as particular cases popular measures such as accuracy, F-score, Jaccard and AM-measure among others. In particular, we consider a family of performance metrics given by ratios of linear combinations of the four population quantities associated with the confusion matrix: true positive, true negative, false positive and false negative. Therefore, we introduce the so-called linear fractional performance measures. The performance of a classifier $g \in \mathcal{G}$ is measured by its utility (note that it's called utility in this case since we want to maximize it):

$$\mathrm{U}_{(\boldsymbol{a},\boldsymbol{b})}(g) := \frac{a_0\mathrm{TP} + a_1\mathrm{TN} + a_2\mathrm{FP} + a_3\mathrm{FN}}{b_0\mathrm{TP} + b_1\mathrm{TN} + b_2\mathrm{FP} + b_3\mathrm{FN}}$$

	Expression	$(\mathrm{c}_0, \mathrm{c}_1, \mathrm{c}_2)$	$(\mathrm{d}_0,\ \mathrm{d}_1,\ \mathrm{d}_2)$
Accuracy	$\mathbb{P}(Y = g(X, S))$	$(1 - \mathbb{P}(Y = 1), 2, -1)$	$(1, 0, 0)$
$\mathbf{F_b}$**-score**	$\frac{(1+b^2)\mathbb{P}(Y=1, g(X,S)=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g(X,S)=1)}$	$(0, 1 + b^2, 0)$	$(b^2\mathbb{P}(Y = 1), 0, 1)$
Jaccard	$\mathbb{P}(Y = 1, g(X, S) = 1)$	$(0, 1, 0)$	$(\mathbb{P}(Y = 1), 0, 0)$
AM-measure	$\frac{1}{2}\sum_{i=\{0,1\}} \mathbb{P}(g(X, S) = i \vert Y = i)$	$\left(\frac{1}{2}, \frac{1}{2\mathbb{P}(Y=1)} + \frac{1}{2\mathbb{P}(Y=0)}, -\frac{1}{2\mathbb{P}(Y=0)}\right)$	$(1, 0, 0)$
Recall	$\mathbb{P}(g(X, S) = 1 \mid Y = 1)$	$(0, 1, 0)$	$(\mathbb{P}(Y = 1), -1, 1)$

Table 7.1: Choices of parameters $\boldsymbol{c}$ and $\boldsymbol{d}$ for some known performance metrics.

with $\boldsymbol{a} = (a_0, a_1, a_2, a_3) \in \mathbb{R}^4$, $\boldsymbol{b} = (b_0, b_1, b_2, b_3) \in \mathbb{R}^4$ and $\mathrm{TP} = \mathbb{P}[Y = 1, g(X) = 1]$, $\mathrm{TN} = \mathbb{P}[Y = 0, g(X) = 0]$, $\mathrm{FP} = \mathbb{P}[Y = 0, g(X) = 1]$, $\mathrm{FN} = \mathbb{P}[Y = 1, g(X) = 0]$. Given the redundancy in these definitions, noticing for example that $\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN} = 1$ or $\mathrm{TP} + \mathrm{FN} = \mathbb{P}[Y = 1]$, with some algebraic manipulation we can rewrite equivalently the utility function as

$$\mathrm{U}_{(\boldsymbol{c},\boldsymbol{d})}(g) = \mathrm{U}(g) := \frac{c_0 + c_1\mathbb{P}[Y = 1, g(X) = 1] + c_2\mathbb{P}[g(X) = 1]}{d_0 + d_1\mathbb{P}[Y = 1, g(X) = 1] + d_2\mathbb{P}[g(X) = 1]} \tag{7.2}$$

with $\boldsymbol{c} = (c_0, c_1, c_2) \in \mathbb{R}^3$ and $\boldsymbol{d} = (d_0, d_1, d_2) \in \mathbb{R}^3$, with $\boldsymbol{c}$, $\boldsymbol{d}$ not depending on g (but possibly on $\mathbb{P}_Y$).

It can be easily shown the above mentioned performance measures (accuracy, F-score, AM, Jaccard and others) corresponds to some particular choices of $\boldsymbol{c}$ and $\boldsymbol{d}$ (see Table 7.2). We start our analysis deriving the optimal $g^* \in \mathcal{G}$ that maximizes the utility function.

7.3 Optimality of thresholding functions

To derive the Bayes function, we need to maximize the utility function $U(g)$ in Eq.(7.2) with $g \in \mathcal{G}$. First step we simplify the problem restricting the optimization to predictors of the form $g_\theta(\cdot) = \mathbb{1}\{\eta(\cdot) > \theta\}$: we call this parametric class $\mathcal{G}_\theta = \{g_\theta : g_\theta(\cdot) = \mathbb{1}\{\eta(\cdot) > \theta\}\}$ and we want to find

$$g_{\theta^*} \in \arg\max_{\mathcal{G}_\theta} \mathrm{U}(g).$$

Before proceeding we state here some conditions on coefficients $\boldsymbol{c}$ and $\boldsymbol{d}$ in Eq. (7.2) we will need in the rest of the thesis.

Assumption 4. *Coefficients c_0, c_1, c_2, d_0, d_1, d_2 in Eq. (7.2), with $d_2c_1 \neq c_2d_1$ (see Remark 5 below for the equality), must satisfy the following conditions:*

1.

$$c_1 \geqslant 0,$$

2.

$$d_0 + \min\{\min\{d_1, 0\} + d_2, 0\} > 0,$$

3.

$$\begin{cases} d_2 c_1 > c_2 d_1 \\ \frac{c_0 d_2 - d_0 c_2}{c_2 d_1 - d_2 c_1} \leqslant \mathbb{P}(Y=1) \\ d_0 c_1 - c_0 d_1 \geqslant (c_0 d_2 - d_0 c_2)_+ \,. \end{cases}$$

Condition 1 will be needed in Lemma 4 and makes the maximization of U reasonable: the more the classifier align with Y the better. Condition 2 ensures the denominator is positive (and $d_0 > 0$). Finally, condition 3 will ensure that the fixed point equation (7.3) is satisfied for exactly one $\theta \in [0,1]$. It can be verified that all the measures presented in Table 7.2 do indeed satisfy these conditions.

Remark 5. *If $d_2 c_1 = c_2 d_1$ we can substitute condition 3) in Assumption 4 with*

$$\begin{cases} d_2 c_1 = c_2 d_1 \\ c_1 d_0 > d_1 c_0 \\ \frac{d_0 c_2 - c_0 d_2}{c_0 d_1 - d_0 c_1} \in [0,1] \end{cases}$$

We can proceed now with deriving the best step function considering its corresponding utility. The following Lemma gives the optimal classifier in $\mathcal{G}_\theta$, i.e. the optimal θ^* such that g_{θ^*} maximizes $\mathrm{U}(g)$ in $\mathcal{G}_\theta$.

Lemma 3. *Let's consider the utility function given in (7.2) under conditions in Assumption 4. Consider only classifiers g belonging to the parameter class $\mathcal{G}_\theta = \{g_\theta : g_\theta(\cdot) = \mathbb{1}\{\eta(\cdot) > \theta\}\}$ for $\theta \in [0,1]$ and $\eta(X) = \mathbb{P}[Y=1|X]$, then*

$$\mathbb{E}[(\eta(X) - \theta^*)_+] = \frac{c_0 d_1 - c_1 d_0}{c_2 d_1 - c_1 d_2}\theta^* + \frac{c_0 d_2 - c_2 d_0}{c_2 d_1 - c_1 d_2} \tag{7.3}$$

where θ^* is the optimal parameter such that $g_{\theta^*}(X) = \mathbb{1}\{\eta(X) > \theta^*\}$ maximizes (7.2) in $\mathcal{G}_\theta$. If $d_2 c_1 = c_2 d_1$ we have a simple explicit expression for the threshold

$$\theta^* = \frac{d_0 c_2 - c_0 d_2}{c_0 d_1 - d_0 c_1}.$$

Proof of Lemma 3. The idea is to apply the first-order optimality condition to $\mathrm{U}(g_\theta) = \mathrm{U}(\theta)$. We start noting that

$$\begin{aligned} \mathbb{P}[Y=1, g(X)=1] = \mathbb{E}[Y g(X)] &= \int_{\mathcal{X}} \eta(x) \mathbb{1}\{\eta(x) > \theta\} \mathrm{d}\mathbb{P}_X(x) \\ &= \int_\theta^1 z \mathrm{d}F_Z(z) = \int_{F_Z(\theta)}^{F_Z(1)} F_Z^{-1}(u)\mathrm{d}u \end{aligned} \tag{7.4}$$

where we took $Z = \eta(X)$ and we used the fact that $\mathbb{E}_{X\sim F_X}[X] = \mathbb{E}_{U\sim \text{Unif}[0,1]}[F^{-1}(U)]$. Differentiating (7.4) wrt θ we obtain

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\int_{F_{\eta(X)}(\theta)}^{F_{\eta(X)}(1)} F^{-1}_{\eta(X)}(u)\mathrm{d}u\right) = -F^{-1}_{\eta(X)}(F_{\eta(X)}(\theta))F'_{\eta(X)}(\theta) = -\theta f_{\eta(X)}(\theta)$$

with $f_{\eta(X)}(\cdot)$ the density function.
Similarly, we have

$$\mathbb{P}[g(X)=1] = \mathbb{P}[\eta(X) > \theta] = 1 - F_{\eta(X)}(\theta)$$

and

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left(1 - F_{\eta(X)}(\theta)\right) = -f_{\eta(X)}(\theta).$$

Then, applying the first order optimality condition to $\mathrm{U}(\theta)$ we obtain the following equation

$$\begin{aligned}-f_{\eta(X)}(c_1\theta + c_2)(d_0 + d_1\mathbb{E}[(\eta(X)-\theta)_+] + d_1\theta(1-F_{\eta(X)}(\theta)) + d_2(1-F_{\eta(X)}(\theta)))+\\ +f_{\eta(X)}(d_1\theta + d_2)(c_0 + c_1\mathbb{E}[(\eta(X)-\theta)_+] + c_1\theta(1-F_{\eta(X)}(\theta)) + c_2(1-F_{\eta(X)}(\theta))) = 0\end{aligned}$$

Rearranging and simplifying the various terms we get the desired condition on θ^*:

$$\mathbb{E}[(\eta(X)-\theta^*)_+] = \frac{c_0d_1 - c_1d_0}{c_2d_1 - c_1d_2}\theta^* + \frac{c_0d_2 - c_2d_0}{c_2d_1 - c_1d_2} \tag{7.5}$$

where we used $\mathbb{E}[Yg(X)] = \mathbb{E}[\eta(X)\mathbb{1}\{\eta(X) > \theta\}] = \mathbb{E}[(\eta(X)-\theta)_+] + \theta\mathbb{P}[\eta(X) > \theta]$. □

The above Lemma 3 gives the form of the best step-classifier, with threshold θ^* the solution of the fixed point equation in (7.5). Anyway, we still don't know if the performance of this step-classifier is competitive against general classifiers.
Therefore, we want to show now that the step-classifier $g_{\theta^*}(\cdot) = \mathbb{1}\{\eta(\cdot) > \theta^*\}$ found above is indeed the optimal choice even in $\mathcal{G}$, i.e. among all possible classifiers.

Lemma 4. *Let's consider the utility function given in (7.2) under conditions in Assumption 4. Let $g \in \mathcal{G}$ be any classifier, then*

$$\mathrm{U}(g_{\theta^*}) - \mathrm{U}(g) = \frac{c_1\mathbb{E}[|\eta(X)-\theta^*|\mathbb{1}\{g_{\theta^*}(X) \neq g(X)\}]}{d_0 + d_1\mathbb{E}[\eta(X)\mathbb{1}\{g(X)=1\}] + d_2\mathbb{P}[g(X)=1]} \geqslant 0.$$

This means that no classifier can have higher utility than our step-classifier. The proof of this result can be found in Appendix H.
In summary, this section proves that, when considering linear fractional performance measures for binary classification, the shape of the optimal Bayes classifier assumes the simple shape of a step function depending on the regression function η and on a threshold that, in the case $d_2c_1 \neq c_2d_1$, can be derived by the fixed point equation (7.5) and depending only on $\boldsymbol{c}$, $\boldsymbol{d}$, η and the distribution of X, i.e. P_X, but not on P_Y. The case with $d_2c_1 = c_2d_1$ is even simpler and we have an explicit expression for the threshold θ^* depending on coefficients $\boldsymbol{c}$ and $\boldsymbol{d}$.

Chapter 8

Post-processing Bounds

In this section our goal is to give post-processing bounds of any plug-in estimator. The idea, already introduced above, of post-processing bounds is that we have access to two different dataset: a labelled one $\mathcal{D}_n^{\mathrm{L}} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ consisting of n i.i.d. copies of $(X, Y) \sim P$ and an unlabelled one $\mathcal{D}_N^{U} = \{X_{n+1}, \ldots, X_{n+N}\}$ consisting of N independent copies of $X \sim P_X$, with $N \geqslant n$. For clarity, when an expectation will be taken with respect to $\mathcal{D}_N^{\mathrm{U}} \sim P_X^{\otimes N}$, we will indicate it with $\mathbb{E}_{\mathcal{D}_N^{\mathrm{U}}}$. We will keep the simple $\mathbb{E}$ symbol for generic expectations.

Now, suppose to have any consistent estimator $\widehat{\eta}$ of the regression function η derived using the labelled data $\mathcal{D}_n^{\mathrm{L}}$. Exploiting $\widehat{\eta}$ and the unlabelled data $\mathcal{D}_N^{\mathrm{U}}$ we can compute the estimator $\widehat{\theta}$ of θ^*, for example applying a bisection algorithm (see Algorithm 1) to the empirical version of (7.3), i.e.

$$\widehat{R}(\theta) = \frac{c_0 d_1 - c_1 d_0}{c_2 d_1 - c_1 d_2}\theta + \frac{c_0 d_2 - c_2 d_0}{c_2 d_1 - c_1 d_2} - \widehat{\mathbb{E}}_N(\widehat{\eta}(X) - \theta)_+, \tag{8.1}$$

where $\widehat{\mathbb{E}}_N$ is an expectation taken with respect to the empirical measure $\frac{1}{N}\sum_{i=n+1}^{n+N} \delta_{X_i}$ evaluated on unlabelled data. Notice that to have one and only one solution $\bar{\theta}$ satisfying $\widehat{R}(\bar{\theta}) = 0$ we need the additional assumption:

$$\frac{c_0 d_2 - d_0 c_2}{c_2 d_1 - d_2 c_1} \leqslant \widehat{\mathbb{E}}_N(\widehat{\eta}(X)) \tag{8.2}$$

so that $\widehat{R}(0) < 0$, $\widehat{R}(1) > 0$ (i.e. empirical counterpart of condition 3. in Assumption 4). Let's define our estimator $\widehat{g}$ (depending on both datasets) as

$$\widehat{g}(\cdot) := \mathbb{1}\{\widehat{\eta}(\cdot) > \widehat{\theta}\}.$$

The goal of this section is to control the "excess risk" of $\widehat{g}$:

$$\mathrm{U}(g_{\theta^*}) - \mathrm{U}(\widehat{g}) = \frac{c_1 \mathbb{E}[|\eta(X) - \theta^*| \mathbb{1}\{g_{\theta^*}(X) \neq \widehat{g}(X)\}]}{d_0 + d_1 \mathbb{E}[\eta(X)\mathbb{1}\{\widehat{g}(X) = 1\}] + d_2 \mathbb{P}[\widehat{g}(X) = 1]}.$$

Algorithm 1 Threshold estimation

```
Input: unlabeled data $\mathcal{D}_N^{\mathrm{U}}$; estimator $\hat{\eta}$; parameters $\boldsymbol{c}, \boldsymbol{d}$; # of iterations $K_{\max}$
Output: threshold estimator $\hat{\theta}$
procedure BISECTION ESTIMATOR
    $\hat{R}(\theta) \leftarrow \frac{c_0 d_1 - c_1 d_0}{c_2 d_1 - c_1 d_2}\theta + \frac{c_0 d_2 - c_2 d_0}{c_2 d_1 - c_1 d_2} - \widehat{\mathbb{E}}_N(\widehat{\eta}(X) - \theta)_+$
    $\theta_{\min} \leftarrow 0, \theta_{\max} \leftarrow 1, K \leftarrow 1$
while $K \leqslant K_{\max}$ :
    if $\hat{R}\left(\frac{\theta_{\min}+\theta_{\max}}{2}\right) = 0$ then return $\frac{\theta_{\min}+\theta_{\max}}{2}$
    if $\hat{R}\left(\frac{\theta_{\min}+\theta_{\max}}{2}\right) < 0$ then $\theta_{\min} \leftarrow \frac{\theta_{\min}+\theta_{\max}}{2}$ else $\theta_{\max} \leftarrow \frac{\theta_{\min}+\theta_{\max}}{2}$
    $K \leftarrow K + 1$
endwhile
    return $\hat{\theta} = \frac{\theta_{\min}+\theta_{\max}}{2}$
```

8.0.1 Assumption-free post-processing bound

We start presenting our first post-processing bound, where no assumptions are made on the probability distribution P.

Proposition 5. *Let's consider the utility function given in* (7.2) *under conditions in Assumption 4 and the additional assumption in Eq.* (8.2). *Let $\widehat{\eta}$ be any estimator of η such that $\widehat{\eta}(x) \in (0, 1]$ almost surely. Consider $\widehat{g}(\cdot) = \mathbb{1}\{\widehat{\eta}(\cdot) \geqslant \widehat{\theta}\}$ where $\widehat{\theta}$ is the output of the bisection algorithm described in Algorithm 1 with some $K_{\max} \in \mathbb{N}$, then it holds that*

$$\mathrm{U}(g_{\theta^*}) - \mathbb{E}\mathrm{U}(\widehat{g}) \leqslant c_1 K_0 \left(\mathbb{E}\|\eta - \widehat{\eta}\|_1 + K_1 \left(\sqrt{\frac{\pi \mathbb{P}(Y=1)}{N}} + \frac{4}{N} + \mathbb{E}\|\eta - \widehat{\eta}\|_1 \right) + 2^{-K_{\max}} \right)$$

with $K_0 = 1/(d_0 + \min\{d_1 + d_2, 0\}\mathbb{P}[Y = 1])$ and $K_1 = \frac{c_2 d_1 - c_1 d_2}{c_0 d_1 - c_1 d_0}$.

The proof of the proposition can be found in Appendix H.
The interpretation of the bound in Propositions 5 is straightforward. We can distinguish the different terms: $\mathbb{E}\|\eta - \widehat{\eta}\|_1$ is the estimation error of the regression function; the $N^{-1/2}$ term is the price we pay for not knowing the marginal distribution of the features; the $2^{-K_{\max}}$ term is the error coming from Algorithm 1 since we are not solving $\widehat{R}(\theta) = 0$ exactly.

In the case of F-score measure (see Table 7.2 for the choices of $\boldsymbol{c}$ and $\boldsymbol{d}$) the result is coherent with [Chzhen, 2019a].

Remark 6 (F-score)**.** *For F-score measure we have $c_0 = 0$, $c_1 = 1 + b^2$, $c_2 = 0$, $d_0 =$*

$b^2\mathbb{P}(Y=1)$, $d_1 = 0$ and $d_2 = 1$ so that we obtain

$$\begin{aligned}\mathrm{F}(g_{\theta^*}) - \mathbb{E}\mathrm{F}(\widehat{g}) &\leqslant \frac{1+b^2}{b^2\mathbb{P}(Y=1)}\left(\mathbb{E}\|\eta-\widehat{\eta}\|_1 + \frac{1}{b^2\mathbb{P}(Y=1)}\left(\sqrt{\frac{\pi\mathbb{P}(Y=1)}{N}} + \frac{4}{N} + \mathbb{E}\|\eta-\widehat{\eta}\|_1\right) + 2^{-K_{\max}}\right)\\ &\leqslant (1+b^2)\Bigg(\frac{1}{b^2\mathbb{P}(Y=1)} \vee 2\left(\frac{1}{b^2\mathbb{P}(Y=1)}\right)^2 \mathbb{E}\|\eta-\widehat{\eta}\|_1 + \\ &\quad + \left(\frac{1}{b^2\mathbb{P}(Y=1)}\right)^2\left(\sqrt{\frac{\pi\mathbb{P}(Y=1)}{N}} + \frac{4}{N}\right)\Bigg) + \frac{1+b^2}{b^2\mathbb{P}(Y=1)}2^{-K_{\max}}\end{aligned}$$

that is exactly the result in [Chzhen, 2019a].

8.0.2 Post-processing bound under margin assumption

In this section we want to improve the above result and obtain fast rates in case of favourable assumptions on the probability distribution of the data. In particular, we will focus on a well-known margin-like assumption.
We start with a lemma we will need in the following.

Lemma 5. *On the set $\{X : \widehat{g}(X) \neq g_{\theta^*}(X)\}$ we have*

$$|\eta(X) - \theta^*| \leqslant 2\max\{|\eta(X) - \widehat{\eta}(X)|, |\widehat{\theta} - \theta^*|\}.$$

Proof. We analyse the 2 possible cases:

- $\widehat{\theta} < \theta^*$, $\eta(X) \geqslant \theta^*$, $\widehat{\eta}(X) < \widehat{\theta}$ (analogous to the symmetric case $\widehat{\theta} \geqslant \theta^*$, $\eta(X) < \theta^*$, $\widehat{\eta}(X) \geqslant \widehat{\theta}$): then it's easy to see that $|\eta(X) - \theta^*| \leqslant |\eta(X) - \widehat{\eta}(X)|$
- $\widehat{\theta} < \theta^*$, $\eta(X) < \theta^*$, $\widehat{\eta}(X) \geqslant \widehat{\theta}$ (analogous to the symmetric case $\widehat{\theta} \geqslant \theta^*$, $\eta(X) \geqslant \theta^*$, $\widehat{\eta}(X) < \widehat{\theta}$): we have $|\eta(X) - \theta^*| \leqslant |\eta(X) - \widehat{\eta}(X)| + |\widehat{\theta} - \theta^*|$.

Putting all together we obtain the result. □

We now want to introduce a slightly modified version of the low noise condition (already presented in Definition 5) and adapted to our case of generalized metrics.

Assumption 5 (Margin Assumption / Low Noise Condition)**.** *Distribution P of the pair $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$ is said to satisfy the α-margin assumption if there exist constants $M > 0$ and $\alpha \geqslant 0$ such that*

$$P(0 < |\eta(X) - \theta^*| \leqslant t) \leqslant Mt^{\alpha} \quad \forall t > 0.$$

Intuitively, the margin condition specifies the behavior of the regression function around the decision threshold θ^*.

We can show now an improved version of Proposition 5 under the above margin assumption, with generic p-norm controlling the difference between true and estimated regression function.

Proposition 6. *Let's consider the utility function given in* (7.2) *under conditions in Assumption 4 and the additional assumption in Eq.* (8.2)*. Assume also the margin assumption in 5. Let $\widehat{\eta}$ be any estimator of η such that $\widehat{\eta}(x) \in (0,1]$ almost surely. Consider $\widehat{g}(\cdot) = \mathbb{1}\{\widehat{\eta}(\cdot) \geqslant \widehat{\theta}\}$ where $\widehat{\theta}$ is the output of the bisection algorithm with some $K_{\max} \in \mathbb{N}$, then it holds that:*

$$\mathrm{U}(g_{\theta^*}) - \mathrm{U}(\widehat{g}) \leqslant C_0(p, \alpha, M, P, \boldsymbol{c}, \boldsymbol{d}) \|\eta - \widehat{\eta}\|_p{}^{\frac{p(1+\alpha)}{p+\alpha}} + \frac{C_1(\alpha, M, P, \boldsymbol{c}, \boldsymbol{d})}{\sqrt{N^{1+\alpha}}} + \frac{C_2(\alpha, M, P, \boldsymbol{c}, \boldsymbol{d})}{N^{1+\alpha}} + \\ C_3(\alpha, M, P, \boldsymbol{c}, \boldsymbol{d}) \|\eta - \widehat{\eta}\|_{1+\alpha}^{1+\alpha} + \frac{C_4(\alpha, M, P, \boldsymbol{c}, \boldsymbol{d})}{2^{(1+\alpha)K_{max}}}$$

where $C_0 = 2c_1 K_0 \frac{p+\alpha}{p} \left(\frac{p}{\alpha}\right)^{\frac{\alpha}{p+\alpha}} M^{\frac{p-1}{p+\alpha}}$, $C_1 = 2^{5+5\alpha} c_1 K_0 M K_1^{1+\alpha} \sqrt{P(Y=1)}(1+\alpha)\Gamma\left(\frac{1+\alpha}{2}\right)$, $C_2 = 2^{5+5\alpha} c_1 K_0 M K_1^{1+\alpha} P(Y=1)(1+\alpha)\Gamma(1+\alpha)$, $C_3 = 2^{1+3\alpha} c_1 K_0 M^{1+\alpha}$, $C_4 = c_1 K_0 M 2^{1+2\alpha}$, $K_0 = 1/(d_0 + \min\{d_1 + d_2, 0\} P[Y=1])$, $K_1 = (c_2 d_1 - c_1 d_2)/(c_0 d_1 - c_1 d_0)$.

The proof of the proposition can be found in Appendix H.
Note that this result matches the one in Proposition 5 for $p = 1$ and $\alpha = 0$, i.e. when we have no margin assumption. Nevertheless, when $\alpha > 0$ we have the desired improvement in the rate of convergence of the bound.

Bibliography

[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.

[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.

[Audibert and Tsybakov, 2007] Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers.

[Avron et al., 2017] Avron, H., Clarkson, K. L., and Woodruff, D. P. (2017). Faster kernel ridge regression using sketching and preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1116–1138.

[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.

[Bach, 2021] Bach, F. (2021). Learning theory from first principles. *Draft of a book, version of Sept*, 6:2021.

[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.

[Boucheron et al., 2005] Boucheron, S., Bousquet, O., and Lugosi, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375.

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

[Bousquet et al., 2003] Bousquet, O., Boucheron, S., and Lugosi, G. (2003). Introduction to statistical learning theory. In *Summer school on machine learning*, pages 169–207. Springer.

[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.

[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.

[Caponnetto and Yao, 2010] Caponnetto, A. and Yao, Y. (2010). Adaptive rates for regularization operators in learning theory. *Analysis and Applications*, 8.

[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.

[Chzhen, 2019a] Chzhen, E. (2019a). Optimal rates for f-score binary classification. *arXiv preprint arXiv:1905.04039*.

[Chzhen, 2019b] Chzhen, E. (2019b). *Plug-in methods in classification*. PhD thesis, Paris Est.

[Chzhen, 2020] Chzhen, E. (2020). Optimal rates for nonparametric f-score binary classification via post-processing. *Mathematical Methods of Statistics*, 29:87–105.

[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.

[Devroye et al., 1996] Devroye, L., Györfi, L., Lugosi, G., Devroye, L., Györfi, L., and Lugosi, G. (1996). Vapnik-chervonenkis theory. *A probabilistic theory of pattern recognition*, pages 187–213.

[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.

[Engl et al., 1996] Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of inverse problems*, volume 375. Springer Science & Business Media.

[Gaucher et al., 2022] Gaucher, S., Schreuder, N., and Chzhen, E. (2022). Fair learning with wasserstein barycenters for non-decomposable performance measures. *arXiv preprint arXiv:2209.00427*.

[Gonen et al., 2016] Gonen, A., Orabona, F., and Shalev-Shwartz, S. (2016). Solving ridge regression using sketched preconditioned svrg. In *International conference on machine learning*, pages 1397–1405. PMLR.

[Gu et al., 2009] Gu, Q., Zhu, L., and Cai, Z. (2009). Evaluation measures of the classification performance of imbalanced data sets. In *Computational Intelligence and Intelligent Systems: 4th International Symposium, ISICA 2009, Huangshi, China, October 23-25, 2009. Proceedings 4*, pages 461–471. Springer.

[He and Garcia, 2009] He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

[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.

[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.

[Kimeldorf and Wahba, 1971] Kimeldorf, G. and Wahba, G. (1971). Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95.

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

[Koyejo et al., 2014] Koyejo, O. O., Natarajan, N., Ravikumar, P. K., and Dhillon, I. S. (2014). Consistent binary classification with generalized performance metrics. *Advances in neural information processing systems*, 27.

[Ledoux and Talagrand, 1991] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media.

[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*.

[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.

[Mohri et al., 2018] Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

[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.

[Rahimi and Recht, 2007] Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20.

[Rigollet and Vert, 2009] Rigollet, P. and Vert, R. (2009). Optimal rates for plug-in estimators of density level sets.

[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.

[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 and Ben-David, 2014] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms.* Cambridge university press.

[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.

[Yang, 1999] Yang, Y. (1999). Minimax nonparametric classification. i. rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284.

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

Appendix A

Proof of Section 4.2

This section is devoted to the proof of Theorems 6 and 7. With slight abuse of notation we set

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

With this notation $L(w) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(w,z)dP(z)$.

The following result is known, [Alquier et al., 2019, Lemma 8.1]. We provide an alternative proof tailored to the Hilbert setting.

Lemma 6. *Under Assumptions 1 and 2, fix $R>0$ and $\tau>0$, with probability at least $1-\delta$,*

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

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

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

Proof. In the proof D denotes an absolute numerical constant, whose value can change from line to line. Fix $w\in\mathcal{X}$ 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 $w,w'\in\mathcal{X}$

$$\|Z_w - Z_{w'}\|_{\psi_2} \leqslant 2CG\|\left\langle X, w-w'\right\rangle\|_2, \tag{A.3}$$

where $\|Z_w - Z_{w'}\|_{\psi_2}$ is defined by (4.28). 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

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

where the last two inequalities are consequence of (4.6) and (4.2), respectively. Hence

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

so that (A.3) is clear. Furthermore, since

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

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]

$$\|(\widehat{L}(w)-L(w))-(\widehat{L}(w')-L(w'))\|_{\psi_2} \leqslant \frac{D}{\sqrt{n}}\|Z_w - Z_{w'}\|_{\psi_2} \leqslant \frac{D}{\sqrt{n}}CG\| \left\langle X, w-w'\right\rangle)\|_2, \tag{A.4}$$

where the last inequality is a consequence of (A.3) and D is an absolute constant. Consider $\mathcal{X}$ as a metric space with respect to the metric

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

where without loss of generality we assume that Σ is injective, then (A.4) states that the centered random process $\big(\widehat{L}(w) - L(w)\big)_{w\in\mathcal{X}}$ has sub-gaussian increments and the generic chaining tail bound [Vershynin, 2010, Theorem 8.5.5] implies that, with probability at least $1-2e^{-\tau}$,

$$\sup_{w,w'\in B_R} \left|(\widehat{L}(w) - L(w)) - (\widehat{L}(w') - L(w'))\right| \leqslant \frac{D}{\sqrt{n}}CG\big(\sqrt{\tau}\,\mathrm{diam}(B_R) + \gamma_2(B_R)\big), \tag{A.5}$$

where $B_R = \{w \in \mathcal{X} : \|w\| \leqslant R\}$, $\mathrm{diam}(B_R)$ and $\gamma_2(B_R)$ are the diamater with respect to the metric d and the Talagrand's γ_2 functional of B_R, [Vershynin, 2010, Definition 8.5.1].

Let G be the Gaussian random vector in $\mathcal{X}$ with covariance Σ, which always exists since Σ 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

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

Plugin these last two bounds in (A.5), it holds that

$$\sup_{w,w' \in B_R} \left|(\widehat{L}(w) - L(w)) - (\widehat{L}(w') - L(w'))\right| \leqslant \frac{D}{\sqrt{n}} CGR\left(\sqrt{\tau}\|\Sigma\|^{\frac{1}{2}} + \operatorname{Tr}(\Sigma)^{\frac{1}{2}}\right). \tag{A.6}$$

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 (4.6), 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 $1-2e^{-\tau}$,

$$|\widehat{L}(0) - L(0)| \leqslant 2\ell_0 \sqrt{\frac{2\tau}{n}}. \tag{A.7}$$

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 (A.6) and (A.7) give (A.1) with $4\exp(-\tau) = \delta$. Bound (A.4) with $w' = 0$ implies (A.2). □

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 λ and thus a suboptimal rate.

The next step in the proof is to make the bound of Lemma 6 valid for all norms R, so that it can be applied to the random quantity $R = \|\widehat{w}_\lambda\|$. This is done in Lemma 7 below though a union bound.

Lemma 7. *Under Assumptions 1 and 2, $\forall w \in \mathcal{H}$, with probability $1-\delta$:*

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

Proof. Fix $\delta \in (0,1)$. For $p \geqslant 1$, let $R_p := 2^p$ and $\delta_p = \delta/(p(p+1))$. By Lemma 6, one has for every $p \geqslant 1$,

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

Collecting the terms containing δ_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:

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

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 $1 - \delta$:

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

This is precisely the desired bound. □

We are now able to prove the two theorems.

Proof of Theorem 6. Since the bound of Lemma 7 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 $1 - \delta$,

$$\begin{aligned} L(\widehat{w}_\lambda) - \widehat{L}(\widehat{w}_\lambda) \leqslant & \frac{DGC\|\Sigma\|^{\frac{1}{2}}(1 + \|\widehat{w}_\lambda\|)\sqrt{r_\Sigma}}{\sqrt{n}} + \\ & + \frac{D}{\sqrt{n}} \Big(GC\|\Sigma\|^{\frac{1}{2}}(1 + \|\widehat{w}_\lambda\|) + \ell_0 \Big) \sqrt{\log(1 + \log_2(3 + \ell_0/\lambda)) + \log(1/\delta)}. \end{aligned} \tag{A.8}$$

Now, let $K = K_{\lambda,\delta} = \sqrt{\log(1 + \log_2(3 + \ell_0/\lambda)) + \log(1/\delta)}$. Eq (A.8) writes

$$L(\widehat{w}_\lambda) - \widehat{L}(\widehat{w}_\lambda) \leqslant \frac{DGC\|\Sigma\|^{\frac{1}{2}}(1 + \|\widehat{w}_\lambda\|)\sqrt{r_\Sigma}}{\sqrt{n}} + \frac{DK}{\sqrt{n}} \Big(GC\|\Sigma\|^{\frac{1}{2}}(1 + \|\widehat{w}_\lambda\|) + \ell_0 \Big) \tag{A.9}$$

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

$$\begin{aligned}
L(\widehat{w}_\lambda) &\leqslant \widehat{L}(\widehat{w}_\lambda) + \sqrt{r_\Sigma}\frac{DGC\|\Sigma\|^{\frac{1}{2}}(1+\|\widehat{w}_\lambda\|)}{\sqrt{n}} + K\frac{DGC\|\Sigma\|^{\frac{1}{2}}(1+\|\widehat{w}_\lambda\|)}{\sqrt{n}} + \frac{DK\ell_0}{\sqrt{n}} \\
&\leqslant \widehat{L}(\widehat{w}_\lambda) + (\sqrt{r_\Sigma}+K)\frac{DGC\|\Sigma\|^{\frac{1}{2}}\,\|\widehat{w}_\lambda\|}{\sqrt{n}} + (\sqrt{r_\Sigma}+K)\frac{DGC\|\Sigma\|^{\frac{1}{2}}}{\sqrt{n}} + \frac{DK\ell_0}{\sqrt{n}} \\
&\leqslant \widehat{L}(\widehat{w}_\lambda) + \lambda\|\widehat{w}_\lambda\|^2 + \frac{D^2G^2C^2(\sqrt{r_\Sigma}+K)^2\|\Sigma\|}{4\lambda n} + (\sqrt{r_\Sigma}+K)\frac{DGC\|\Sigma\|^{\frac{1}{2}}}{\sqrt{n}} + \frac{DK\ell_0}{\sqrt{n}} \\
&\leqslant \widehat{L}(w_\lambda) + \lambda\|w_\lambda\|^2 + \frac{D^2G^2C^2(\sqrt{r_\Sigma}+K)^2\|\Sigma\|}{4\lambda n} + (\sqrt{r_\Sigma}+K)\frac{DGC\|\Sigma\|^{\frac{1}{2}}}{\sqrt{n}} + \frac{DK\ell_0}{\sqrt{n}}
\end{aligned} \tag{A.10}$$

where (A.10) holds by definition of $\widehat{w}_\lambda$. Now, using again Lemma 6 for $\|w_\lambda\|$ we have that, with probability $1-\delta$:

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

Combining this inequality with (A.10) with a union bound, with probability $1-2\delta$:

$$\begin{aligned}
L(\widehat{w}_\lambda) <& L(w_\lambda) + \lambda\|w_\lambda\|^2 + \frac{D^2G^2C^2(\sqrt{r_\Sigma}+K)^2\|\Sigma\|}{4\lambda n} + (\sqrt{r_\Sigma}+K)\frac{DGC\|\Sigma\|^{\frac{1}{2}}}{\sqrt{n}} + \frac{DK\ell_0}{\sqrt{n}} + \\
&+ \frac{DGC\|\Sigma\|^{\frac{1}{2}}\,\|w_\lambda\|\,\big(\sqrt{r_\Sigma}+\sqrt{\log(4/\delta)}\big)}{\sqrt{n}} + \frac{D\ell_0\sqrt{\log(4/\delta)}}{\sqrt{n}}.
\end{aligned} \tag{A.11}$$

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

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

so that (A.11) implies, with probability $1-2\delta$:

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

After replacing δ by $\delta/2$, we get bound (4.15). □

Proof of Theorem 7. Assume that $w_* = \arg\min_{w \in \mathcal{H}} L(w)$ exists. Then, by definition of w_λ,

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

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 (A.11) that, with probability $1 - 2\delta$,

$$\begin{aligned} L(\widehat{w}_\lambda) <& L(w_*) + \lambda\|w_*\|^2 + \frac{D^2G^2C^2(\sqrt{r_\Sigma} + K)^2\|\Sigma\|}{4\lambda n} + \frac{DGC(\sqrt{r_\Sigma}+K)\|\Sigma\|^{\frac{1}{2}} + DK\ell_0}{\sqrt{n}} + \\ &+ \frac{DGC\|\Sigma\|^{\frac{1}{2}} \|w_*\| \left(\sqrt{r_\Sigma} + \sqrt{\log(4/\delta)}\right)}{\sqrt{n}} + \frac{D\ell_0\sqrt{\log(4/\delta)}}{\sqrt{n}} \end{aligned} \tag{A.12}$$

The bound (A.12) precisely corresponds to the desired bound (4.16) after replacing δ 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 $\widetilde{O}(DGC \|\Sigma\|^{1/2} \|w_*\| \sqrt{\log(1/\delta)/n})$. □

Appendix B

Proof of Section 4.3

In order to prove Theorem 8, we need to previously extend Lemma 7 in [Rudi et al., 2015] to sub-gaussian random variables.

Lemma 8. *Fix $\delta > 0$ and a (T, α_0)-approximate leverage scores $(\hat{l}_i(\alpha))_{i=1}^n$ with confidence $\delta > 0$. Given $\alpha > \alpha_0$, let $\{\widetilde{x}_1, \ldots, \widetilde{x}_m\}$ be the Nyström points selected according to Definition 4 and set $\mathcal{B}_m = span\{\widetilde{x}_1, \ldots, \widetilde{x}_m\}$. Under Assumption 1, with probability at least $1-\delta$:*

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

provided that

$$n \gtrsim d_\alpha \vee \log(5/\delta) \tag{B.2}$$

$$m \gtrsim d_\alpha \log(\frac{10n}{\delta}). \tag{B.3}$$

Furthermore, if the spectrum of Σ satisfies the decay conditions (4.28) (polynomial decay) or (4.29) (exponential decay), it is enough to assume that

$$n \gtrsim \log(5/\delta) \qquad \alpha \gtrsim n^{-1/p} \qquad m \gtrsim \alpha^{-p}\log(\frac{10n}{\delta}) \qquad \textit{polynomial decay} \tag{B.4}$$

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

Proof. Exploiting sub-gaussianity anyway the various terms are bounded differently. In particular, to bound β_1 we refer to Theorem 9 in [Koltchinskii and Lounici, 2014], obtaining with probability at least $1-\delta$

$$\beta_1(\alpha) \lesssim \max\left\{\sqrt{\frac{d_\alpha}{n}}, \sqrt{\frac{\log(1/\delta)}{n}}\right\}. \tag{B.6}$$

As regards β_3 term we apply Proposition 7 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 $1-5\delta$

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

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 (4.28) or (4.29) . The second claim is consequence of Proposition 8 or Proposition 9. □

We can proceed now with the proof of Theorem 8:

Proof of Theorem 8. We recall the notation.

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

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:

$$L(\widehat{\beta}_\lambda) - L(w_*) \leqslant \left|L(\widehat{\beta}_\lambda) - \widehat{L}(\widehat{\beta}_\lambda) - \lambda\|\widehat{\beta}_\lambda\|^2_{\mathcal{H}}\right| + \left|\widehat{L}(\widehat{\beta}_\lambda) + \lambda\|\widehat{\beta}_\lambda\|^2_{\mathcal{H}} - \widehat{L}(\mathcal{P}_m w_*) - \lambda\|\mathcal{P}_m w_*\|^2_{\mathcal{H}}\right| + \\ + \left|\widehat{L}(\mathcal{P}_m w_*) - L(\mathcal{P}_m w_*)\right| + |L(\mathcal{P}_m w_*) - L(w_*)| + \lambda\|\mathcal{P}_m w_*\|^2_{\mathcal{H}} \tag{B.7}$$

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 7 for $\widehat{\beta}_\lambda$ and we get

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

with $K = K_{\lambda,\delta} = \sqrt{\log(1 + \log_2(3 + \ell_0/\lambda)) + \log(1/\delta)}$ as in (A.9).

Now since $xy \leqslant \lambda x^2 + y^2/(4\lambda)$, we can write

$$\frac{DGC(\sqrt{r_\Sigma} + K)\|\widehat{\beta}_\lambda\|\|\Sigma\|^{\frac{1}{2}}}{\sqrt{n}} \leqslant \lambda\|\widehat{\beta}_\lambda\|^2 + \frac{D^2G^2C^2(\sqrt{r_\Sigma} + K)^2\|\Sigma\|}{\lambda n} \tag{B.8}$$

hence,

$$\left|L(\widehat{\beta}_\lambda)-\widehat{L}(\widehat{\beta}_\lambda)-\lambda\|\widehat{\beta}_\lambda\|^2\right| \leqslant \frac{D^2G^2C^2(\sqrt{r_\Sigma}+K)^2\|\Sigma\|}{\lambda n}+\frac{DGC(\sqrt{r_\Sigma}+K)\|\Sigma\|^{\frac{1}{2}}}{\sqrt{n}}+\frac{DK\ell_0}{\sqrt{n}}, \tag{B.9}$$

Term $\left|\widehat{L}(\widehat{\beta}_\lambda)+\lambda\|\widehat{\beta}_\lambda\|^2_{\mathcal{H}}-\widehat{L}(\mathcal{P}_m w_*)-\lambda\|\mathcal{P}_m w_*\|^2_{\mathcal{H}}\right|$ 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 $1-\delta$:

$$\begin{aligned}\left|\widehat{L}(\mathcal{P}_m w_*)-L(\mathcal{P}_m w_*)\right| &\leqslant \sup_{\|w\|\leqslant\|w_*\|}\left(\left|\widehat{L}(w)-L(w)\right|\right)\\ &< \frac{D}{\sqrt{n}}\Big(GC\|w_*\|\|\Sigma\|^{\frac{1}{2}}\big(\sqrt{r_\Sigma}+\sqrt{\log(4/\delta)}\big)+\ell_0\sqrt{\log(4/\delta)}\Big).\end{aligned} \tag{B.10}$$

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 6 where the role of L and $\widehat{L}$ is interchanged.

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

$$\begin{aligned}|L(\mathcal{P}_m w_*)-L(w_*)| &\leqslant G\int|\langle w,\mathcal{P}_m w_*\rangle-\langle w,w_*\rangle|dP_X(w)\\ &\leqslant G\left(\int|\langle w,(I-\mathcal{P}_m)w_*\rangle|^2 dP_X(w)\right)^{\frac{1}{2}}\\ &= G\,\langle\Sigma(I-\mathcal{P}_m)w_*,(I-\mathcal{P}_m)w_*\rangle^{\frac{1}{2}} \qquad \text{(B.11)}\\ &= G\|\Sigma^{1/2}(I-\mathcal{P}_m)w_*\|_{\mathcal{H}}\\ &\leqslant G\|\Sigma^{1/2}(I-\mathcal{P}_m)\|\|w_*\|_{\mathcal{H}}\\ &= G\|(I-\mathcal{P}_m)\Sigma^{1/2}\|\|w_*\|_{\mathcal{H}} \leqslant G\sqrt{3\alpha}\|w_*\|, \qquad \text{(B.12)}\end{aligned}$$

where the last bound is a consequence of Lemma 8 and it holds true with probability at least $1-\delta$.

Putting the pieces together we finally get the result in Theorem 8 by replacing δ with $\delta/3$. □

Proof of Theorem. 9. Under polynomial decay assumption (4.28), the claim is a consequence of Theorem 8 with Proposition 8 with $\beta = 1/p$ so that

$$m \gtrsim d_\alpha \log n, \qquad d_\alpha \lesssim \alpha^{-p}, \qquad m \asymp n^p(\log n)^{1-p} \tag{B.13}$$

Under exponential decay assumption (4.29), the claim is a consequence of Theorem 8 with Proposition 9 so that

$$m \gtrsim d_\alpha \log n, \qquad d_\alpha \lesssim \log(1/\alpha), \qquad m \asymp \log^2 n \tag{B.14}$$

□

Proof of Theorem 10. The proof is given by decomposing the excess risk as in (B.7) where $\mathcal{P}_m$ is replaced by $\mathcal{P}_\mathcal{B}$, (B.9) bounds term A, (B.10) bounds term B and (B.11) and 4.32 bound term C. □

Appendix C

Proofs of Section 4.4

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 $d_\alpha = \mathrm{Tr}((\Sigma_\alpha + \alpha\,\mathrm{I})^{-1}\Sigma)$.

Proposition 7. *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 $1-\delta$*

$$d_\alpha(\widehat{\Sigma}) \leqslant 16 d_\alpha \tag{C.1}$$

Proof. Let V_α be the space spanned by eigenvectors of Σ with corresponding eigenvalues $\alpha_j \geqslant \alpha$, and call D_α 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_α 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_α, we have

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

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 $d_\alpha(\Sigma) = \mathrm{Tr}\, g(\Sigma) = \mathrm{Tr}((\Sigma_\alpha + \alpha\,\mathrm{I})^{-1}\Sigma)$,

$$d_\alpha(\widehat{\Sigma}) \leqslant 2(d_\alpha(\widehat{\Sigma}_1) + d_\alpha(\widehat{\Sigma}_2)) \tag{C.3}$$

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

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

Now,

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

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

$$\|\langle v, X\rangle\|_{\psi_2} \leqslant C\|\langle v, X\rangle\|_{L_2} \qquad \forall v \in \mathcal{H} \tag{C.5}$$

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

$$\left\|\|X_2\|^2\right\|_{\psi_1} = \left\|\sum_i \langle X_2, e_i\rangle^2\right\|_{\psi_1} \leqslant \sum_i \left\|\langle X_2, e_i\rangle^2\right\|_{\psi_1} \tag{C.6}$$

$$= \sum_i \|\langle X_2, e_i\rangle\|_{\psi_2}^2 \leqslant C^2 \|\langle X_2, e_i\rangle\|_{L_2}^2 \tag{C.7}$$

$$= C^2 \sum_i \alpha_i = C^2 \mathrm{Tr}\left[\Sigma_2\right] = C^2 \mathbb{E}\left[\|X_2\|^2\right] \tag{C.8}$$

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

$$d_\alpha = \sum_{i=1}^\infty \frac{\alpha_i}{\alpha_i + \alpha} \geqslant \sum_{i:\alpha_i<\alpha} \frac{\alpha_i}{\alpha_i + \alpha} \geqslant \sum_{i:\alpha_i<\alpha} \frac{\alpha_i}{2\alpha} = \frac{\mathrm{Tr}(\Sigma_2)}{2\alpha} \tag{C.9}$$

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

$$n\mathbb{E}\left[\|X_2\|^4\right] \leqslant \underbrace{4nC^2\alpha^2 d_\alpha^2(\Sigma)}_{\nu} \tag{C.10}$$

$$c = C\alpha d_\alpha \tag{C.11}$$

where we used the bound on the moments of a sub-exponential variable (see [Vershynin, 2010]). With high probability (C.4) becomes

$$d_\alpha(\widehat{\Sigma}) \leqslant 8d_\alpha + \frac{4Cd_\alpha\sqrt{2\log(1/\delta)}}{\sqrt{n}} + \frac{2Cd_\alpha \log(1/\delta)}{n} \leqslant 16d_\alpha \tag{C.12}$$

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

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:

Theorem 15 (Theorem 4 in [Adamczak, 2008])**.** *Let $X, X_1, \ldots, X_n$ 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

$$\mathbb{P}\left(Z \geqslant (1+\eta)\mathbb{E}Z + \frac{K_1 \left\|\sup_{f\in\mathcal{F}} |f(X)|\right\|_{\psi_1} (2+\tau)}{n} + \sqrt{\frac{3(1+\tau)\sigma^2}{n}} \right) \leqslant e^{-\tau} \tag{C.13}$$

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

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

$$g_{w,r} := \frac{h_w - \mathbb{E}h_w}{\lambda \|w\|^2 + \mathbb{E}h_w + r}, \quad w \in \mathcal{H}, \quad r > 0 \tag{C.14}$$

(notice that in [Steinwart and Christmann, 2008] they define $-g_{w,r}$).
Our plan is to apply Theorem 15 to $g_{\widehat{w}_0,r}$, with $\widehat{w}_0 \in \mathcal{B}_m \subseteq \mathcal{H}$ and $\|\widehat{w}_0\| \leqslant \|w_*\|$.

Corollary 3. *Under the hypothesis of Theorem 15, for all $\tau > 0$ we have*

$$\sup_{w\in\mathcal{H},\|w\|\leqslant\|w_*\|} \frac{\widehat{\mathbb{E}}h_w - \mathbb{E}h_w}{\lambda\|w\|^2 + \mathbb{E}h_w + r} < 2\mathbb{E}_{D\sim \mathrm{P}^n} \sup_{w\in\mathcal{H},\|w\|\leqslant\|w_*\|} \frac{\widehat{\mathbb{E}}h_w - \mathbb{E}h_w}{\lambda\|w\|^2 + \mathbb{E}h_w + r} + \sqrt{\frac{3V(1+\tau)}{nr^{2-\theta}}} + 2GK_1\|w_*\| \frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\|X\|)(2+\tau)}{nr} \tag{C.15}$$

Proof. In Theorem 15, we take

$$Z = \sup_{w\in\mathcal{H},\|w\|\leqslant R} \left| \frac{1}{n}\sum_{i=1}^{n} g_{w,r}(X_i) \right| . \tag{C.16}$$

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,

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

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

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

Applying Theorem 15 with $\eta = 1$ we get the result. □

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

Theorem 16. *Under assumptions 1, 2, 4 and 3, the covariance matrix satisfies the polynomial decay condition (4.28), and the Bernstein conditions (4.36)–(4.37) hold true. Fix a closed subspace $\widehat{\mathcal{F}}$ of $\mathcal{H}$ and set*

$$
w_{\widehat{\mathcal{F}},\lambda} = \operatorname*{argmin}_{w\in\widehat{\mathcal{F}}} \left(\widehat{L}(w) + \lambda\|w\|^2 \right) \qquad \lambda > 0. \tag{C.17}
$$

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

$$\lambda \|\widehat{w}_{\mathcal{F},\lambda}\|^2 + L(\widehat{w}^{cl}_{\mathcal{F},\lambda}) - L(f_*) \leqslant 7\left(\lambda \left\|\widehat{w}_0\right\|^2 + L(\widehat{w}_0) - L(f_*)\right) + K_3 \left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + \\ + 2\left(\frac{72V\log(3/\delta)}{n}\right)^{\frac{1}{2-\vartheta}} + 16GK_1 \left\|w_*\right\| \frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2+\log(3/\delta))}{n} \tag{C.18}$$

where the constant a only depends on (4.28) *and $K_3 \geqslant 1$ only depends on p, M, B, ϑ, 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 (4.28) is equivalent to condition (7.48) of Theorem 7.23, which is given in terms of entropy numbers e_j, see Lemma 12. 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 15 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)],

$$\lambda \left\|\widehat{w}_\lambda\right\|^2 + \mathbb{E}h_{\widehat{w}^{cl}_\lambda} \leqslant (\lambda \left\|\widehat{w}_0\right\|^2 + \mathbb{E}h_{\widehat{w}_0}) + (\widehat{\mathbb{E}}h_{\widehat{w}_0} - \mathbb{E}h_{\widehat{w}_0}) + (\mathbb{E}h_{\widehat{w}^{cl}_\lambda} - \widehat{\mathbb{E}}h_{\widehat{w}^{cl}_\lambda}) \tag{C.19}$$

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 (C.14), 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

$$\frac{\widehat{\mathbb{E}}h_{\widehat{w}_0} - \mathbb{E}h_{\widehat{w}_0}}{\lambda \left\|\widehat{w}_0\right\|^2 + \mathbb{E}h_{\widehat{w}_0} + r} < 2\mathbb{E}_{D\sim \mathrm{P}^n} \frac{\widehat{\mathbb{E}}h_{\widehat{w}_0} - \mathbb{E}h_{\widehat{w}_0}}{\lambda \left\|\widehat{w}_0\right\|^2 + \mathbb{E}h_{\widehat{w}_0} + r} \\ + \sqrt{\frac{3V(1+\tau)}{nr^{2-\theta}}} + 2GK_1 \left\|w_*\right\| \frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\left\|X\right\|)(2+\tau)}{nr}. \tag{C.20}$$

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

$$\begin{aligned}\mathbb{E}_{D\sim \mathrm{P}^n} \sup_{w\in\mathcal{B}_{m,r},\|w\|\leqslant\|w_*\|} \left|\widehat{\mathbb{E}}h_w - \mathbb{E}h_w\right| &\leqslant \mathbb{E}_{D\sim \mathrm{P}^n} \sup_{w\in\mathcal{H}_r,\|w\|\leqslant\|w_*\|} \left|\widehat{\mathbb{E}}h_w - \mathbb{E}h_w\right| \\ &\leqslant 2\mathbb{E}_{D\sim \mathrm{P}^n}\widehat{\mathrm{Rad}}(\mathcal{H}_r, n) \leqslant 2\varphi_n(r).\end{aligned} \tag{C.21}$$

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

$$\mathbb{E}_{D\sim \mathrm{P}^n} \sup_{w\in\mathcal{B}_{m},\|w\|\leqslant\|w_*\|} \left|\widehat{\mathbb{E}}g_{w,r}\right| \leqslant \mathbb{E}_{D\sim \mathrm{P}^n} \sup_{w\in\mathcal{H},\|w\|\leqslant\|w_*\|} \left|\widehat{\mathbb{E}}g_{w,r}\right| \leqslant \frac{8\varphi_n(r)}{r} \tag{C.22}$$

Putting all together we get w.h.p.

$$\begin{aligned}\widehat{\mathbb{E}}h_{\widehat{w}_0} - \mathbb{E}h_{\widehat{w}_0} < (\lambda \|\widehat{w}_0\|^2 + \mathbb{E}h_{\widehat{w}_0}) &\left(\frac{10\varphi_n(r)}{r} + \sqrt{\frac{3V(1+\tau)}{nr^{2-\theta}}} + 2GK_1\|w_*\| \frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\|X\|)(2+\tau)}{nr}\right) \\ &+ 10\varphi_n(r) + \sqrt{\frac{3V(1+\tau)r^\theta}{n}} + 2GK_1\|w_*\| \frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\|X\|)(2+\tau)}{n}\end{aligned} \tag{C.23}$$

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

$$\begin{aligned}&\lambda \|\widehat{w}_\lambda\|^2 + \mathbb{E}h_{\widehat{w}^{cl}_\lambda} < \left(\lambda \|\widehat{w}_0\|^2 + \mathbb{E}h_{\widehat{w}_0}\right) + \\ &\quad + (\lambda \|\widehat{w}_0\|^2 + \mathbb{E}h_{\widehat{w}_0}) \left(\frac{10\varphi_n(r)}{r} + \sqrt{\frac{3V(1+\tau)}{nr^{2-\theta}}} + 2GK_1\|w_*\| \frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\|X\|)(2+\tau)}{nr}\right) + \\ &\quad + 10\varphi_n(r) + \sqrt{\frac{3V(1+\tau)r^\theta}{n}} + 2GK_1\|w_*\| \frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\|X\|)(2+\tau)}{n} + \\ &\quad + \left(\lambda \|\widehat{w}_\lambda\|^2 + \mathbb{E}h_{\widehat{w}^{cl}_\lambda}\right) \left(\frac{10\varphi_n(r)}{r} + \sqrt{\frac{2V\tau}{nr^{2-\theta}}} + \frac{28B\tau}{3nr}\right) \\ &\quad + 10\varphi_n(r) + \sqrt{\frac{2V\tau r^\theta}{n}} + \frac{28B\tau}{3n}\end{aligned} \tag{C.24}$$

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 $r \geqslant 8GK_1 \left\|w_*\right\| \frac{(C\sqrt{2\mathrm{Tr}\Sigma}+\mathbb{E}\|X\|)(2+\tau)}{n}$ gives

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

We finally obtain

$$\begin{aligned} \lambda\left\|\widehat{w}_\lambda\right\|^2 + \mathbb{E}h_{\widehat{w}^{cl}_\lambda} &< \frac{11}{6}\left(\lambda\left\|\widehat{w}_0\right\|^2 + \mathbb{E}h_{\widehat{w}_0}\right) + \frac{79}{54}r + \epsilon + \frac{17}{27}\left(\lambda\left\|\widehat{w}_\lambda\right\|^2 + \mathbb{E}h_{\widehat{w}^{cl}_\lambda}\right) \\ &\leqslant 5\left(\lambda\left\|\widehat{w}_0\right\|^2 + \mathbb{E}h_{\widehat{w}_0}\right) + 2r \end{aligned} \tag{C.25}$$

with

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

□

Remark 7. *Notice that the same reasoning can be applied in Section 4.4 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 11, polynomial decay. Applying Theorem 16 in the general case of Remark 7, with the choice $\widehat{\mathcal{F}} = \mathcal{B}_m$ and $\widehat{w}_0 = \mathcal{P}_{\mathcal{B}_m} w_\lambda$, we rewrite (C.18) as:

$$\begin{aligned}
\lambda\|\widehat{\beta}_{\lambda,m}\|^2 + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(f_*) &\leqslant 7(\lambda\|\mathcal{P}_{\mathcal{B}_m} w_\lambda\|^2 + L(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L(f_*)) + K_3\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\theta+\theta p}} + \\
&\quad + 2\Big(\frac{72V\log(3/\delta)}{n}\Big)^{\frac{1}{2-\theta}} + 16GK_1\|w_\lambda\|\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\|X\|)(2+\log(3/\delta))}{n} \\
&= 7(\lambda\|\mathcal{P}_{\mathcal{B}_m} w_\lambda\|^2 + L(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L(w_\lambda) + L(w_\lambda) - L(f_*)) + K_3\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\theta+\theta p}} + \\
&\quad + 2\Big(\frac{72V\log(3/\delta)}{n}\Big)^{\frac{1}{2-\theta}} + 16GK_1\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\|X\|)(2+\log(3/\delta))}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} \\
&\leqslant 7(L(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L(w_\lambda) + \lambda\|w_\lambda\|^2 + L(w_\lambda) - L(f_*)) + K_3\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\theta+\theta p}} + \\
&\quad + 2\Big(\frac{72V\log(3/\delta)}{n}\Big)^{\frac{1}{2-\theta}} + 16GK_1\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\|X\|)(2+\log(3/\delta))}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} \\
&= 7\mathcal{A}(\lambda) + 7(L(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L(w_\lambda)) + K_3\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\theta+\theta p}} + 2\Big(\frac{72V\log(3/\delta)}{n}\Big)^{\frac{1}{2-\theta}} + \\
&\quad + 16GK_1\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\|X\|)(2+\log(3/\delta))}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} \qquad \text{(C.26)}
\end{aligned}$$

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 (B.11) (but where we use Lemma 8 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 $1-\delta$

$$L(\mathcal{P}_{\mathcal{B}_m} w_\lambda) - L(w_\lambda) \leqslant K_2 G\sqrt{\alpha}\,\|w_\lambda\| \leqslant K_2 G\sqrt{\alpha}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} \qquad \text{(C.27)}$$

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

$$\begin{aligned}
\lambda\|\widehat{\beta}_{\lambda,m}\|^2_{\mathcal{H}} + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(f_*) &\leqslant 7\mathcal{A}(\lambda) + 7K_2G\sqrt{\frac{\alpha\mathcal{A}(\lambda)}{\lambda}} + K_3\Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\theta+\theta p}} + 2\Big(\frac{72V\log(3/\delta)}{n}\Big)^{\frac{1}{2-\theta}} + \\
&\quad + 16GK_1\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\|X\|)(2+\log(3/\delta))}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}} \qquad \text{(C.28)}
\end{aligned}$$

which proves the first claim. □

The following corollary provides the optimal rates.

Corollary 4. *Fix $\delta > 0$. Under the Theorem 11 and the source condition*

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

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

$$\lambda \asymp n^{-\min\{\frac{2}{r+1}, \frac{1}{r(2-p-\theta+\theta p)+p}\}} \tag{C.29}$$

$$\alpha \asymp n^{-\min\{2, \frac{r+1}{r(2-p-\theta+\theta p)+p}\}} \tag{C.30}$$

$$m \gtrsim n^{\min\{2p, \frac{p(r+1)}{r(2-p-\theta+\theta p)+p}\}} \tag{C.31}$$

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

$$\lambda \|\widehat{\beta}_{\lambda,m}\|^2 + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(f_*) \lesssim n^{-\min\{\frac{2r}{r+1}, \frac{r}{r(2-p-\theta+\theta p)+p}\}} \tag{C.32}$$

Proof. Lemma 8 with Proposition 8 gives

$$m \gtrsim d_\alpha \log(n/\delta), \qquad d_\alpha \lesssim \alpha^{-p} \qquad \alpha \asymp \frac{\log^{1/p}(n/\delta)}{m^{1/p}} \tag{C.33}$$

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 λ, α and m given by (C.29)–(C.31) 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 8.

When we are in the well specified case (see Section 4.4.1), i.e. w_* exists, we have the following results.

Corollary 5. *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 (4.28), then, with probability at least $1 - 2\delta$:*

$$L(\widehat{\beta}^{cl}_{\lambda,m}) - L(w_*) \lesssim \frac{1}{\lambda^p n} + \lambda \|w_*\|^2 + \sqrt{\alpha}\, \|w_*\| \tag{C.34}$$

provided that n and m are large enough.

Proof. The proof mimics the proof of Theorem 11 *(a)* where in (C.18) we choose $\widehat{w}_0 = \mathcal{P}_{\mathcal{B}_m} w_*$ Hence (C.18) with $\theta = 1$ reads

$$\begin{aligned}
\lambda \|\widehat{\beta}_{\lambda,m}\|^2 + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(w_*) &\leqslant 7(\lambda \|\mathcal{P}_{\mathcal{B}_m} w_*\|^2 + L(\mathcal{P}_{\mathcal{B}_m} w_*) - L(w_*)) + K_3 \frac{a^{2p}}{\lambda^p n} + 144V \frac{\log(3/\delta)}{n} + \\
&\quad + 16GK_1 \|w_*\| \frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\,\|X\|)(2 + \log(3/\delta))}{n} \\
&\leqslant 7\lambda \|w_*\|^2 + 7(L(\mathcal{P}_{\mathcal{B}_m} w_*) - L(w_*)) + K_3 \frac{a^{2p}}{\lambda^p n} + 144V \frac{\log(3/\delta)}{n} + \\
&\quad + 16GK_1 \|w_*\| \frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\,\|X\|)(2 + \log(3/\delta))}{n}
\end{aligned} \tag{C.35}$$

We can deal wit h the term $L(\mathcal{P}_{\mathcal{B}_m} w_*) - L(w_*)$ as in (B.11), 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 $1-2\delta$:

$$\lambda\|\widehat{\beta}_{\lambda,m}\|^2 + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(w_*) \leqslant 7\lambda\|w_*\|^2 + 7K_2G\sqrt{\alpha}\|w_*\| + K_3\frac{a^{2p}}{\lambda^p n} + 144V\frac{\log(3/\delta)}{n} + \\ 16GK_1\,\|w_*\|\,\frac{(C\sqrt{2\mathrm{Tr}\Sigma} + \mathbb{E}\,\|X\|)(2+\log(3/\delta))}{n} \tag{C.36}$$

which proves the claim. □

And, similarly to Corollary 4, we obtain the optimal rate presented in Eq. 4.41.

Corollary 6. *Fix $\delta > 0$. Under the assumptions of Theorem 11 (a), when the variance bound (4.37) holds true with the optimal paratemer $\theta = 1$ and the model is well specified, i.e. $r = 1$, set*

$$\lambda \asymp n^{-\frac{1}{1+p}} \tag{C.37}$$

$$\alpha \asymp n^{-\frac{2}{1+p}} \tag{C.38}$$

$$m \gtrsim n^{\frac{2p}{1+p}}\log n \tag{C.39}$$

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

$$\lambda\|\widehat{\beta}_{\lambda,m}\|^2_{\mathcal{H}} + L(\widehat{\beta}^{cl}_{\lambda,m}) - L(w_*) \lesssim \|w_*\|\left(\frac{1}{n}\right)^{\frac{1}{1+p}}. \tag{C.40}$$

Notice that $\alpha \asymp n^{-\frac{2}{1+p}}$ is compatible with condition $\alpha \gtrsim d_\alpha \asymp n^{-1/p}$ in Lemma 8.

C.0.1 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 16 in the case of exponential decay using the following Lemma:

Lemma 9. *When*

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

we have

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

We can finally prove the second part of Theorem 11 under exponential decay:

Proof of Theorem 11, exponential decay. We follow exactly the proof of Theorem 16 for polynomial decay presented above in the previous subsection, but using the estimate in Lemma 9 for r:

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

□

Appendix D

Proofs of Section 4.5

D.0.1 Square loss

We report in this section the proofs of Theorem 12.
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 12. The proof follows the one of Theorem 11 in Appendix C with some differences coming from the fact that we are working now with the square loss. Since Theorem 16 works also with locally Lipschitz loss functions we have:

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

Using the fact that L_λ is quadratic and expanding around the the minimum w_λ we have

$$L_\lambda(\mathcal{P}_m w_\lambda) - L_\lambda(w_\lambda) = \|(\Sigma+\alpha)^{1/2}(I-\mathcal{P}_m)w_\lambda\|^2 \qquad \text{(D.2)}$$

Using Lemma 8 we get the result

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

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.0.2 Logistic Loss

Since logistic loss is not clippable, we prove how the modification of the definition of the clipping in (4.46) 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 16.

As explained in subSection 4.5.2, let's note that we have $h_f(X) - h^{cl}_f(X) + \frac{1}{n} \geqslant 0$. Therefore we can simply rewrite the splitting of the error (C.19) as

$$\lambda\,\|\widehat{w}_\lambda\|^2 + \mathbb{E}h_{\widehat{w}^{cl}_\lambda} \leqslant (\lambda\,\|\widehat{w}_0\|^2 + \mathbb{E}h_{\widehat{w}_0}) + (\widehat{\mathbb{E}}h_{\widehat{w}_0} - \mathbb{E}h_{\widehat{w}_0}) + (\mathbb{E}h_{\widehat{w}^{cl}_\lambda} - \widehat{\mathbb{E}}h_{\widehat{w}^{cl}_\lambda}) + \frac{1}{n}. \qquad \text{(D.4)}$$

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

$$\begin{aligned}\lambda\|\widehat{\beta}_{\lambda,m}\|^2+L(\widehat{\beta}^{cl}_{\lambda,m})-L(f_*)&\leqslant 7(\lambda\|\mathcal{P}_{\mathcal{B}_m}w_\lambda\|^2+L(\mathcal{P}_{\mathcal{B}_m}w_\lambda)-L(f_*))+K_3\frac{a^{2p}}{\lambda^p n}+\\&+\frac{144V\log(3/\delta)}{n}+16GK_1\left\|w_\lambda\right\|\frac{(C\sqrt{2\mathrm{Tr}\Sigma}+\mathbb{E}\left\|X\right\|)(2+\log(3/\delta))}{n}+\frac{1}{n}\\&=7(L_\lambda(\mathcal{P}_{\mathcal{B}_m}w_\lambda)-L_\lambda(w_\lambda)+L_\lambda(w_\lambda)-L(f_*))+K_3\frac{a^{2p}}{\lambda^p n}+\frac{144V\log(3/\delta)}{n}+\\&+16GK_1\frac{(C\sqrt{2\mathrm{Tr}\Sigma}+\mathbb{E}\left\|X\right\|)(2+\log(3/\delta))}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}}+\frac{1}{n}\\&=7\mathcal{A}(\lambda)+7(L_\lambda(\mathcal{P}_{\mathcal{B}_m}w_\lambda)-L_\lambda(w_\lambda))+K_3\frac{a^{2p}}{\lambda^p n}+\frac{144V\log(3/\delta)}{n}+\\&+16GK_1\frac{(C\sqrt{2\mathrm{Tr}\Sigma}+\mathbb{E}\left\|X\right\|)(2+\log(3/\delta))}{n}\sqrt{\frac{\mathcal{A}(\lambda)}{\lambda}}+\frac{1}{n}\end{aligned}\tag{D.5}$$

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

$$L_\lambda(\mathcal{P}_{\mathcal{B}_m}w_\lambda)-L_\lambda(w_\lambda)=\frac{1}{2}\langle(HL)(w')(\mathcal{P}_{\mathcal{B}_m}w_\lambda-w_\lambda),(\mathcal{P}_{\mathcal{B}_m}w_\lambda-w_\lambda)\rangle\tag{D.6}$$

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

$$\begin{aligned}\langle(HL)(w)v,v\rangle&=\frac{d^2}{dt^2}L(w+tv)|_{t=0}=\frac{d}{dt}\mathbb{E}\left[\ell'(\langle w+tv,X\rangle,Y)\langle v,X\rangle\right]|_{t=0}\\&=\mathbb{E}\left[\ell''(\langle w+tv,X\rangle,Y)(\langle v,X\rangle)^2\right]|_{t=0}\leqslant M\mathbb{E}\left[\langle v,X\rangle^2\right]\end{aligned}\tag{D.7}$$

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 (D.7) and considering $L_\lambda(\cdot)=L(\cdot)+\lambda\left\|\cdot\right\|^2$ we get

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

Finally we can rewrite (D.6) as

$$L_\lambda(\mathcal{P}_{\mathcal{B}_m}w_\lambda)-L_\lambda(w_\lambda)\leqslant\frac{1}{2}\left\|\Sigma_\lambda^{1/2}(\mathcal{P}_{\mathcal{B}_m}w_\lambda-w_\lambda)\right\|^2\tag{D.8}$$

and proceed exactly as in the case of the square loss (see appendix D.0.1).

Appendix E

Entropy Numbers and Exponential Decay

We analyse here the main steps needed to obtain the results for exponential decay in Theorem 8 and Theorem 11.

E.0.1 Entropy numbers in Hilbert spaces

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

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

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 Σ 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 Σ is trace-class, S is compact, so that by (E.1)

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

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.0.2 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.0.3 Bound the Rademacher Complexity of $\mathcal{F}_r$

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

Lemma 10.

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

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

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

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

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

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

$$\int_T^\infty 2^{x/2} 2^{-c\sqrt{2^i - 1}}\, \mathrm{d}x \leqslant \int_T^\infty 2^{x/2} 2^{-c_2\sqrt{2^i}}\, \mathrm{d}x \leqslant \frac{2^{-c_2\sqrt{2^T} + 1}}{c_2 \log^2(2)} \tag{E.3}$$

$$\leqslant c_3 2^{-c_2\sqrt{2^T}} \tag{E.4}$$

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 ρ. 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

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

□

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

Lemma 11.

$$R\left(\mathcal{F}_r\right) \leqslant C_1 \sqrt{\frac{V}{n}} \log_2\left(\frac{1}{\lambda}\right) \sqrt{r} + C_2 \frac{\log_2^2(1/\lambda)}{n} \tag{E.5}$$

where C_1 and C_2 are two absolute constants.

Proof.

$$R\left(\mathcal{F}_r\right)=\mathbb{E}[\widehat{R}\left(\mathcal{F}_r\right)] \leqslant \sqrt{\frac{(\log 16)}{n}} \log_2\left(\frac{1}{\lambda}\right)\left(3\mathbb{E} \sup_{f\in\mathcal{F}_r}\|f\|_{L_2(D)}+2c_3\sqrt{r}\right) \tag{E.6}$$

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

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

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

$$\begin{aligned}R\left(\mathcal{F}_r\right) &\leqslant \sqrt{\frac{\log 16}{n}}\log_2\left(\frac{1}{\lambda}\right)\left(9\sigma+2c_3\sqrt{r}\right) \qquad \text{(E.7)}\\ &\leqslant \sqrt{\frac{\log 16}{n}}\log_2\left(\frac{1}{\lambda}\right)\left(9\sqrt{Vr^{\theta}}+2c_3\sqrt{r}\right) \qquad \text{(E.8)}\\ &\leqslant c_5\sqrt{\frac{V}{n}}\log_2\left(\frac{1}{\lambda}\right)\sqrt{r} \qquad \text{(E.9)}\end{aligned}$$

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 $\sigma^2\leqslant R\left(\mathcal{F}_r\right)$, we have

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

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

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

where C_1 and C_2 are two absolute constants. □

Appendix F

Known results

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

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 σ_j of Σ from [Caponnetto and De Vito, 2007]. We report the proofs for sake of completeness.

Proposition 8 (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

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

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

$$d_\alpha = \mathrm{Tr}(\Sigma(\Sigma + \alpha I)^{-1}) = \sum_{i=1}^{\infty} \frac{\sigma_i}{\sigma_i + \alpha} \leqslant \sum_{i=1}^{\infty} \frac{\gamma}{\gamma + i^\beta \alpha} \tag{F.2}$$

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

$$\begin{aligned} d_\alpha &\leqslant \int_0^\infty \frac{\gamma}{\gamma + x^\beta \alpha} dx \\ &= \alpha^{-1/\beta} \int_0^\infty \frac{\gamma}{\gamma + \tau^\beta} d\tau \\ &\leqslant \gamma \frac{\beta}{\beta - 1} \alpha^{-1/\beta} \end{aligned} \tag{F.3}$$

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

Proposition 9 (Exponential eigenvalues decay).
If for some $\gamma, \beta \in \mathbb{R}^+ \sigma_i \leqslant \gamma e^{-\beta i}$ *then*

$$d_\alpha \leqslant \frac{\log(1+\gamma/\alpha)}{\beta} \tag{F.4}$$

Proof.

$$d_\alpha = \sum_{i=1}^{\infty} \frac{\sigma_i}{\sigma_i + \alpha} = \sum_{i=1}^{\infty} \frac{1}{1+\alpha/\sigma_i} \leqslant \sum_{i=1}^{\infty} \frac{1}{1+\alpha' e^{\beta i}} \leqslant \int_0^{+\infty} \frac{1}{1+\alpha' e^{\beta x}} dx \tag{F.5}$$

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

$$\begin{aligned}(F.5) &= \frac{1}{\beta}\int_1^{+\infty} \frac{1}{1+\alpha' t}\frac{1}{t} dt = \frac{1}{\beta}\int_1^{+\infty} \Big[\frac{1}{t} - \frac{\alpha'}{1+\alpha' t}\Big] dt = \frac{1}{\beta}\Big[\log t - \log(1+\alpha' t)\Big]_1^{+\infty} \\ &= \frac{1}{\beta}\Big[\log\Big(\frac{t}{1+\alpha' t}\Big)\Big]_1^{+\infty} = \frac{1}{\beta}\Big[\log(1/\alpha') + \log(1+\alpha')\Big] \end{aligned} \tag{F.6}$$

So we finally obtain

$$d_\alpha \leqslant \frac{1}{\beta}\Big[\log(\gamma/\alpha) + \log(1+\alpha/\gamma)\Big] = \frac{\log(1+\gamma/\alpha)}{\beta} \tag{F.7}$$

□

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 $x_i, \dots, x_n$. Regard the data matrix $\widehat{X}$ as the inclusion operator $\mathrm{id} : \mathcal{H} \to L_2(\widehat{P})$

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

Lemma 12. *Let* $p \in (0,1)$*. Then*

$$\mathbb{E}_{\widehat{P}}[e_j(\mathrm{id} : \mathcal{H} \to L_2(\widehat{P}))] \sim j^{-\frac{1}{2p}} \tag{F.8}$$

if and only if

$$\sigma_j \sim j^{-\frac{1}{p}} \tag{F.9}$$

As regard results in Section 5, from [Bartlett et al., 2006] we report the following lemma:

Lemma 13. *For any nonnegative loss function* ϕ, *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 13:

Lemma 14 (Theorem 3 in [Bartlett et al., 2006])**.** *Suppose that P has noise exponent $0 \leqslant \gamma \leqslant 1$, and that ϕ 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 ϕ is the square loss, $\psi(x) = x$ when ϕ is the hinge loss and $\psi(x) \geqslant \frac{x}{2}$ when ϕ 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:

Lemma 15. *[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 ℓ. 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 (5.1).

Appendix G

Experiments: datasets and tuning

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 σ and λ 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 γ. The details of the considered data sets and the chosen parameters for our algorithm in Table 6.1 and G.1 are the following:

SUSY (Table 6.1 and G.1, $n = 5 \times 10^6$, $d = 18$): we used a Gaussian kernel with $\sigma = 4$, $\lambda = 3 \times 10^{-6}$ and $m_{ALS} = 2500$, $m_{uniform} = 2500$.

Mnist binary (Table 6.1 and G.1, $n = 7 \times 10^4$, $d = 784$): we used a Gaussian kernel with $\sigma = 10$, $\lambda = 3 \times 10^{-6}$ and $m_{ALS} = 15000$, $m_{uniform} = 20000$.

Usps (Table 6.1 and G.1, $n = 9298$, $d = 256$): we used a Gaussian kernel with $\sigma = 10$, $\lambda = 5 \times 10^{-6}$ and $m_{ALS} = 2500$, $m_{uniform} = 4000$.

Webspam (Table 6.1 and G.1, $n = 3.5 \times 10^5$, $d = 254$): we used a Gaussian kernel with $\sigma = 0.25$, $\lambda = 8 \times 10^{-7}$ and $m_{ALS} = 11500$, $m_{uniform} = 20000$.

a9a (Table 6.1 and G.1, $n = 48842$, $d = 123$): we used a Gaussian kernel with $\sigma = 10$, $\lambda = 1 \times 10^{-5}$ and $m_{ALS} = 800$, $m_{uniform} = 1500$.

CIFAR (Table 6.1 and G.1, $n = 6 \times 10^4$, $d = 400$): we used a Gaussian kernel with $\sigma = 10$, $\lambda = 2 \times 10^{-6}$ and $m_{ALS} = 20500$, $m_{uniform} = 20000$.

Table G.1: 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).

	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 ± 2	134 ± 4	$20.1\% \pm 0.2\%$	592 ± 2	129 ± 1
Mnist bin	$2.2\% \pm 0.1\%$	1342 ± 5	491 ± 32	$2.3\% \pm 0.1\%$	1814 ± 8	954 ± 21
Usps	$3.0\% \pm 0.1\%$	19.8 ± 0.1	7.3 ± 0.3	$3.0\% \pm 0.2\%$	66.1 ± 0.1	48 ± 8
Webspam	$1.3\% \pm 0.1\%$	2440 ± 5	376 ± 18	$1.3\% \pm 0.1\%$	4198 ± 40	1455 ± 180
a9a	$15.1\% \pm 0.2\%$	29.3 ± 0.2	1.5 ± 0.1	$15.1\% \pm 0.2\%$	30.9 ± 0.2	3.2 ± 0.1
CIFAR	$19.2\% \pm 0.1\%$	2408 ± 14	820 ± 47	$19.0\% \pm 0.1\%$	2168 ± 19	709 ± 13

Appendix H

Plug-in Classifiers: Main Proofs

H.1 Proofs of Section 7

Proof of Lemma 4. Let's call $D(g_{\theta^*}) := d_0 + d_1\mathbb{E}[\eta(X)\mathbb{1}\{\eta(x) > \theta^*\}] + d_2\mathbb{P}[\eta(x) > \theta^*]$ and $D(g) := d_0 + d_1\mathbb{E}[\eta(X)\mathbb{1}\{g(X) = 1\}] + d_2\mathbb{P}[g(X) = 1]$, we have

$$
\begin{aligned}
&\mathrm{U}_{(c,d)}(g_{\theta^*}) - \mathrm{U}_{(c,d)}(g) = \\
&= \frac{c_0 + c_1\mathbb{E}[\eta(X)\mathbb{1}\{\eta(x) > \theta^*\}] + c_2\mathbb{P}[\eta(x) > \theta^*]}{D(g_{\theta^*})} - \frac{c_0 + c_1\mathbb{E}[\eta(X)\mathbb{1}\{g(X) = 1\}] + c_2\mathbb{P}[g(X) = 1]}{D(g)} \\
&= \frac{c_1\mathbb{E}[(\eta(X) - \theta^*)(\mathbb{1}\{\eta(x) > \theta^*\} - \mathbb{1}\{g(X) = 1\})] + (c_1\theta^* + c_2)(\mathbb{P}[\eta(x) > \theta^*] - \mathbb{P}[g(X) = 1])}{D(g_{\theta^*})} + \\
&\quad + \frac{(c_0 + c_1\mathbb{E}[\eta(X)\mathbb{1}\{g(X) = 1\}] + c_2\mathbb{P}[g(X) = 1])(d_0 + d_1\mathbb{E}[\eta(X)\mathbb{1}\{g(X) = 1\}] + d_2\mathbb{P}[g(X) = 1])}{D(g)D(g_{\theta^*})} \\
&\quad - \frac{(c_0 + c_1\mathbb{E}[\eta(X)\mathbb{1}\{g(X) = 1\}] + c_2\mathbb{P}[g(X) = 1])(d_0 + d_1\mathbb{E}[\eta(X)\mathbb{1}\{\eta(x) > \theta^*\}] + d_2\mathbb{P}[\eta(x) > \theta^*])}{D(g)D(g_{\theta^*})} \\
&= \frac{c_1\mathbb{E}[(\eta(X) - \theta^*)(\mathbb{1}\{\eta(x) > \theta^*\} - \mathbb{1}\{g(X) = 1\})] + (c_1\theta^* + c_2)(\mathbb{P}[\eta(X) > \theta^*] - \mathbb{P}[g(X) = 1])}{D(g_{\theta^*})} + \\
&\quad + \frac{(c_0 + c_1\mathbb{E}[\eta(X)\mathbb{1}\{g(X) = 1\}] + c_2\mathbb{P}[g(X) = 1])(d_1\mathbb{E}[\eta(X)(\mathbb{1}\{g(X) = 1\} - \mathbb{1}\{\eta(X) > \theta^*\})] +}{D(g)D(g_{\theta^*})} \\
&\quad \underline{d_2(\mathbb{P}[g(X) = 1] - \mathbb{P}[\eta(X) > \theta^*]))} \\
&= \frac{c_1\mathbb{E}[(\eta(X) - \theta^*)(\mathbb{1}\{\eta(X) > \theta^*\} - \mathbb{1}\{g(X) = 1\})] + (c_1\theta^* + c_2)(\mathbb{P}[\eta(X) > \theta^*] - \mathbb{P}[g(X) = 1])}{D(g_{\theta^*})} + \\
&\quad + \mathrm{U}_{(c,d)}(g)\frac{d_1\mathbb{E}[\eta(X)(\mathbb{1}\{g(X) = 1\} - \mathbb{1}\{\eta(X) > \theta^*\})] + d_2(\mathbb{P}[g(X) = 1] - \mathbb{P}[\eta(X) > \theta^*])}{D(g_{\theta^*})} \\
&= \frac{c_1\mathbb{E}[|\eta(X) - \theta^*|\mathbb{1}\{g_{\theta^*}(X) \neq g(X)\}]}{D(g_{\theta^*})} + \mathrm{U}_{(c,d)}(g)\frac{d_1\mathbb{E}[\eta(X)(\mathbb{1}\{g(X) = 1\} - \mathbb{1}\{\eta(X) > \theta^*\})]}{D(g_{\theta^*})} + \\
&\quad + (c_1\theta^* + c_2 - d_2\mathrm{U}_{(c,d)}(g))\frac{\mathbb{P}[\eta(X) > \theta^*] - \mathbb{P}[g(X) = 1]}{D(g_{\theta^*})} \\
&= \frac{c_1\mathbb{E}[|\eta(X) - \theta^*|\mathbb{1}\{g_{\theta^*}(X) \neq g(X)\}]}{D(g_{\theta^*})} + \\
&\quad + d_1(\mathrm{U}_{(c,d)}(g_{\theta^*}) - \mathrm{U}_{(c,d)}(g))\frac{\mathbb{E}[\eta(X)\mathbb{1}\{\eta(X) > \theta^*\} - \mathbb{1}\{g(X) = 1\}]}{D(g_{\theta^*})} + \\
&\quad + d_2(\mathrm{U}_{(c,d)}(g_{\theta^*}) - \mathrm{U}_{(c,d)}(g))\frac{\mathbb{P}[\eta(x) > \theta^*] - \mathbb{P}[g(X) = 1]}{D(g_{\theta^*})}
\end{aligned}
$$

where we added and subtracted $\frac{c_0+c_1\mathbb{E}[\eta(X)\mathbb{1}\{g(X)=1\}]+c_2\mathbb{P}[g(X)=1]}{D(g_{\theta^*})}$ and we used the fact that $\mathrm{U}_{(c,d)}\left(g_{\theta^*}\right)=\frac{c_2+\theta^* c_1}{d_2+\theta^* d_1}$ (see Lemma 16). Solving for $\mathrm{U}_{(c,d)}$ we get

$$\mathrm{U}_{(c,d)}(g_{\theta^*})-\mathrm{U}_{(c,d)}(g)=\frac{c_1\mathbb{E}[|\eta(X)-\theta^*|\mathbb{1}\{g_{\theta^*}(X)\neq g(X)\}]}{d_0+d_1\mathbb{E}[\eta(X)\mathbb{1}\{g(X)=1\}]+d_2\mathbb{P}[g(X)=1]}$$

that is non-negative under Assumption 4. □

Lemma 16 (Lemma 4.1 in [Gaucher et al., 2022])**.** *The utility function given in* (7.2) *and evaluated for g_{θ^*} takes the form*

$$\mathrm{U}\left(g_{\theta^*}\right)=\frac{c_2+\theta^* c_1}{d_2+\theta^* d_1}.$$

Proof. Using the optimality condition (7.3) for θ^* we have

$$\mathbb{E}[\eta(X)\mathbb{1}\{\eta(X)>\theta^*\}]=\mathbb{E}[(\eta(X)-\theta^*)_+]+\theta^*\mathbb{P}[\eta(X)>\theta^*]=\frac{c_0d_1-c_1d_0}{c_2d_1-c_1d_2}\theta^*+\frac{c_0d_2-c_2d_0}{c_2d_1-c_1d_2}+\theta^*\mathbb{P}[\eta(X)>\theta^*].$$

Then, we can rewrite $\mathrm{U}_{(c,d)}(g_{\theta^*})$ as

$$\mathrm{U}_{(c,d)}(g_{\theta^*})=\frac{c_0(c_2d_1-d_2c_1)+c_1(\theta^*(c_0d_1-d_0c_1)+(c_0d_2-d_0c_2))+(c_2+\theta^*c_1)(c_2d_1-d_2c_1)\mathbb{P}[\eta(X)>\theta^*]}{d_0(c_2d_1-d_2c_1)+d_1(\theta^*(c_0d_1-d_0c_1)+(c_0d_2-d_0c_2))+(d_2+\theta^*d_1)(c_2d_1-d_2c_1)\mathbb{P}[\eta(X)>\theta^*]}$$

Simplifying some terms and factorizing the numerator and the denominator by $(c_2+\theta^*c_1)$ and $(d_2+\theta^*d_1)$ respectively, we obtain

$$\mathrm{U}_{(c,d)}\left(g_{\theta^*}\right)=\frac{c_2+\theta^*c_1}{d_2+\theta^*d_1}\cdot\frac{(c_0d_1-d_0c_1)+(c_2d_1-d_2c_1)\mathbb{P}[\eta(X)>\theta^*]}{(c_0d_1-d_0c_1)+(c_2d_1-d_2c_1)\mathbb{P}[\eta(X)>\theta^*]}=\frac{c_2+\theta^*c_1}{d_2+\theta^*d_1}.$$

□

H.2 Proofs of Section 8

Proof of Proposition 5. We already know we can write the "excess risk" as

$$\mathrm{U}(g_{\theta^*})-\mathrm{U}(\widehat{g})=\frac{c_1\mathbb{E}[|\eta(X)-\theta^*|\mathbb{1}\{g_{\theta^*}(X)\neq\widehat{g}(X)\}]}{d_0+d_1\mathbb{E}[\eta(X)\mathbb{1}\{\widehat{g}(X)=1\}]+d_2\mathbb{P}[\widehat{g}(X)=1]}.$$

We can rewrite the numerator as:

$$\begin{aligned}\mathbb{E}[|\eta(X)-\theta^*|\mathbb{1}\{g_{\theta^*}(X)\neq\widehat{g}(X)\}]&=\\&=\mathbb{E}[(\eta(X)-\theta^*)\mathbb{1}\{\eta(X)>\theta^*,\widehat{\eta}(X)<\widehat{\theta}\}]+\mathbb{E}[(\theta^*-\eta(X))\mathbb{1}\{\eta(X)<\theta^*,\widehat{\eta}(X)>\widehat{\theta}\}]\\&\leqslant\mathbb{E}[(\eta(X)-\theta^*-\widehat{\eta}(X)+\widehat{\theta})\mathbb{1}\{\eta(X)>\theta^*,\widehat{\eta}(X)<\widehat{\theta}\}]+\\&\quad+\mathbb{E}[(\theta^*-\eta(X)-\widehat{\theta}+\widehat{\eta}(X))\mathbb{1}\{\eta(X)<\theta^*,\widehat{\eta}(X)>\widehat{\theta}\}]\\&\leqslant\|\eta-\widehat{\eta}\|_1+|\theta^*-\widehat{\theta}|.\end{aligned}$$

Then we obtain

$$\mathrm{U}(g_{\theta^*}) - \mathrm{U}(\widehat{g}) \leqslant c_1 \frac{\|\eta - \widehat{\eta}\|_1 + |\theta^* - \widehat{\theta}|}{d_0 + d_1 \mathbb{E}[\eta(X)\mathbb{1}\{\widehat{g}(X) = 1\}] + d_2 \mathbb{P}[\widehat{g}(X) = 1]}.$$

We start by controlling $|\theta^* - \widehat{\theta}|$. Define $\bar{\theta}$ such that

$$\widehat{R}(\bar{\theta}) = 0, \qquad \text{with} \quad \widehat{R}(\theta) = \frac{c_0 d_1 - c_1 d_0}{c_2 d_1 - c_1 d_2}\theta + \frac{c_0 d_2 - c_2 d_0}{c_2 d_1 - c_1 d_2} - \widehat{\mathbb{E}}_N(\widehat{\eta}(X) - \theta)_+ \tag{H.1}$$

with $\widehat{\theta}$ an estimator of $\bar{\theta}$ (for instance the output of a bisection algorithm applied to $\widehat{R}(\theta)$) and $\widehat{\mathbb{E}}_N$ is an expectation taken with respect to the empirical measure $\frac{1}{N}\sum_{i=n+1}^{n+N} \delta_{X_i}$ evaluated on unlabelled data.

We can decompose

$$\mathbb{E}_{\mathcal{D}_N^U}|\theta^* - \widehat{\theta}| \leqslant \mathbb{E}_{\mathcal{D}_N^U}|\bar{\theta} - \widehat{\theta}| + \mathbb{E}_{\mathcal{D}_N^U}|\bar{\theta} - \theta^*|$$

The first term is simply an optimization error related with the chosen algorithm to solve $\widehat{R}(\theta) = 0$. Notice that $\widehat{R}(\theta)$ is continuous on $[0, 1]$ and $\widehat{R}(0) < 0$, $\widehat{R}(1) > 0$ thanks to the assumption in Eq. (8.2), then classical analysis of the bisection algorithm implies that $|\bar{\theta} - \widehat{\theta}| \leqslant 2^{-K_{max}}$ almost surely.

To bound $|\bar{\theta} - \theta^*|$ we mimic the proof of Proposition 3 in [Chzhen, 2019a].

$$|\bar{\theta} - \theta^*| = \frac{c_2 d_1 - c_1 d_2}{c_0 d_1 - c_1 d_0}\left|\int_{\theta^*}^1 (1 - F_\eta(t))dt - \int_{\bar{\theta}}^1 (1 - \widehat{F}_{\widehat{\eta}}(t))dt\right| \tag{H.2}$$

Consider $\theta^* \geqslant \bar{\theta}$:

$$\begin{aligned}&\left|\int_{\theta^*}^1 (1 - F_\eta(t))dt - \int_{\bar{\theta}}^1 (1 - \widehat{F}_{\widehat{\eta}}(t))dt\right| \leqslant \left|\int_{\theta^*}^1 (1 - F_\eta(t))dt - \int_{\theta^*}^1 (1 - \widehat{F}_{\widehat{\eta}}(t))dt\right| \\ &\leqslant \left|\int_{\theta^*}^1 (\widehat{F}_{\widehat{\eta}}(t) - F_\eta(t))dt\right| \leqslant \int_{\theta^*}^1 \left|\widehat{F}_{\widehat{\eta}}(t) - F_\eta(t)\right| dt \leqslant \int_0^1 \left|\widehat{F}_{\widehat{\eta}}(t) - F_\eta(t)\right| dt \leqslant \|F_{\widehat{\eta}} - F_\eta\|_1\end{aligned}$$

Consider $\theta^* \leqslant \bar{\theta}$:

$$\begin{aligned}&\left|\int_{\theta^*}^1 (1 - F_\eta(t))dt - \int_{\bar{\theta}}^1 (1 - \widehat{F}_{\widehat{\eta}}(t))dt\right| \leqslant \left|\int_{\bar{\theta}}^1 (1 - F_\eta(t))dt - \int_{\bar{\theta}}^1 (1 - \widehat{F}_{\widehat{\eta}}(t))dt\right| \\ &\leqslant \left|\int_{\bar{\theta}}^1 (\widehat{F}_{\widehat{\eta}}(t) - F_\eta(t))dt\right| \leqslant \int_{\bar{\theta}}^1 \left|\widehat{F}_{\widehat{\eta}}(t) - F_\eta(t)\right| dt \leqslant \int_0^1 \left|\widehat{F}_{\widehat{\eta}}(t) - F_\eta(t)\right| dt \leqslant \|\widehat{F}_{\widehat{\eta}} - F_\eta\|_1\end{aligned}$$

So we have

$$|\bar{\theta} - \theta^*| = \frac{c_2 d_1 - c_1 d_2}{c_0 d_1 - c_1 d_0}\left|\int_{\theta^*}^1 (1 - F_\eta(t))dt - \int_{\bar{\theta}}^1 (1 - \widehat{F}_{\widehat{\eta}}(t))dt\right| \leqslant \frac{c_2 d_1 - c_1 d_2}{c_0 d_1 - c_1 d_0}\|\widehat{F}_{\widehat{\eta}} - F_\eta\|_1$$

Now we can introduce also $\widehat{F}_\eta$ that is the empirical cumulative distribution of $\eta(X)$ based on $\mathcal{D}_N^U$. Using the triangular inequality again we get

$$\|\widehat{F}_{\widehat{\eta}} - F_\eta\|_1 \leqslant \|F_\eta - \widehat{F}_\eta\|_1 + \|\widehat{F}_{\widehat{\eta}} - \widehat{F}_\eta\|_1$$

Following the reasoning in [Chzhen, 2019a], we can concentrate the first term using Bernstein's inequality. Let $p(t) = P(\eta(X) \geqslant t)$, we have:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_N^U}\left|P(\eta(X) \geqslant t) - \widehat{P}_N(\eta(X) \geqslant t)\right| &= \int_0^\infty P_{\mathcal{D}_N^U}\left(\left|P(\eta(X) \geqslant t) - \widehat{P}_N(\eta(X) \geqslant t)\right| \geqslant x\right) dx \\ &\leqslant 2\int_0^\infty \exp\left(-\frac{Nx^2}{2\left(p(t)+\frac{1}{3}x\right)}\right) dx\end{aligned}$$

with $\widehat{P}_N$ the empirical measure $\frac{1}{N}\sum_{i=n+1}^{n+N}\delta_{X_i}$ evaluated on the unlabelled data. Furthermore, we can write for the inner integral

$$\begin{aligned}\int_0^\infty \exp\left(-\frac{Nx^2}{2\left(p(t)+\frac{1}{3}x\right)}\right) dx &= \left(\int_0^{3p(t)} + \int_{3p(t)}^\infty\right) \exp\left(-\frac{Nx^2}{2\left(p(t)+\frac{1}{3}x\right)}\right) dx \\ &\leqslant \int_0^{3p(t)} \exp\left(-\frac{N^2x^2}{4p(t)}\right) dx + \int_{3p(t)}^\infty \exp\left(-\frac{Nx}{4}\right) dx \\ &\leqslant \sqrt{\frac{\pi p(t)}{N}} + \frac{4}{N}.\end{aligned}$$

Therefore, using Fubini's theorem, we obtain

$$\mathbb{E}_{\mathcal{D}_N^U}\left\|F_\eta - \widehat{F}_\eta\right\|_1 = \int_0^1 \mathbb{E}_{\mathcal{D}_N^U}\left|P(\eta(X) \geqslant t) - \widehat{P}_N(\eta(X) \geqslant t)\right| dt \leqslant \sqrt{\frac{\pi}{N}}\int_0^1 \sqrt{P(\eta(X) \geqslant t)}dt + \frac{4}{N}.$$

Applying Cauchy-Schwarz inequality to the first term on the r.h.s. of the above inequality we derive the following bound

$$\mathbb{E}_{\mathcal{D}_N^U}\left\|F_\eta - \widehat{F}_\eta\right\|_1 \leqslant \sqrt{\frac{\pi\mathbb{P}(Y=1)}{N}} + \frac{4}{N}.$$

It remains to bound $\left\|\widehat{F}_{\widehat{\eta}} - \widehat{F}_\eta\right\|_1$. Let $Z_i = \eta(X_i)$ and $\widehat{Z}_i = \widehat{\eta}(X_i)$ for all $i = n+1, \ldots, N$, then for the second term on the r.h.s. of Eq. (10) we can write

$$\left\|\widehat{F}_{\widehat{\eta}} - \widehat{F}_\eta\right\|_1 = \frac{1}{N}\int_0^1 \left|\sum_{i=n+1}^{n+N}\left(\mathbb{1}\{Z_i \leqslant t\} - \mathbb{1}\left\{\widehat{Z}_i \leqslant t\right\}\right)\right| dt$$

This expression corresponds to the Wasserstein-1 distance between empirical measures of $\{Z_{n+1}, \ldots, Z_{n+N}\}$ and $\left\{\widehat{Z}_{n+1}, \ldots, \widehat{Z}_{n+N}\right\}$. Hence, using its alternative definition we get

$$\left\|\widehat{F}_{\widehat{\eta}} - \widehat{F}_{\eta}\right\|_1 = \inf_{\omega \in \mathbb{S}_N} \frac{1}{N} \sum_{i=n+1}^{n+N} \left|Z_i - \widehat{Z}_{\omega(i)}\right| \leqslant \frac{1}{N} \sum_{i=n+1}^{n+N} \left|\eta\left(X_i\right) - \widehat{\eta}\left(X_i\right)\right|$$

where the infimum is taken over all permutations $\mathbb{S}_N$ of $\{n+1, \ldots, n+N\}$. Finally, since conditionally on the labelled data $\mathcal{D}_n^{\mathrm{L}}$ the random variables $\left|\eta\left(X_i\right) - \widehat{\eta}\left(X_i\right)\right|$ with $i = n+1, \ldots, n+N$ are i.i.d., then $\mathbb{E}_{\mathcal{D}_N^U}\left\|\widehat{F}_{\widehat{\eta}} - \widehat{F}_{\eta}\right\|_1 \leqslant \mathbb{E}_{\mathcal{D}_N^U}\|\eta - \widehat{\eta}\|_1$. Hence,

$$\mathbb{E}_{\mathcal{D}_N^U}\|\widehat{F}_{\widehat{\eta}} - F_{\eta}\|_1 \leqslant \sqrt{\frac{\pi\mathbb{P}(Y=1)}{N}} + \frac{4}{N} + \|\eta - \widehat{\eta}\|_1.$$

Putting all together

$$\mathbb{E}_{\mathcal{D}_N^U}|\theta^* - \widehat{\theta}| \leqslant \frac{c_2 d_1 - c_1 d_2}{c_0 d_1 - c_1 d_0}\left(\sqrt{\frac{\pi\mathbb{P}(Y=1)}{N}} + \frac{4}{N} + \|\eta - \widehat{\eta}\|_1\right) + 2^{-K_{\max}}$$

As regards the denominator, conditions in 4 assure that $d_0 \geqslant 0$. For all the most known measures that can be represented by (7.2) we have also $d_2 \geqslant 0$. Then we get

$$\begin{aligned} d_0 + d_1\mathbb{P}[Y=1, \widehat{g}(X)=1] + d_2\mathbb{P}[\widehat{g}(X)=1] &\geqslant d_0 + (d_1+d_2)\mathbb{P}[Y=1, \widehat{g}(X)=1] \\ &\geqslant d_0 + \min\{d_1+d_2, 0\}\mathbb{P}[Y=1]. \end{aligned}$$

So, with $K_0(\mathbb{P}, d_0, d_1) = 1/(d_0 + \min\{d_1+d_2, 0\}\mathbb{P}[Y=1])$, we obtain the final bound

$$\begin{aligned} \mathrm{U}(g_{\theta^*}) - \mathbb{E}_{\mathcal{D}_N^U}\mathrm{U}(\widehat{g}) &= c_1\frac{\|\eta - \widehat{\eta}\|_1 + \mathbb{E}_{\mathcal{D}_N^U}|\theta^* - \widehat{\theta}|}{d_0 + d_1\mathbb{P}[Y=1, \widehat{g}(X)=1] + d_2\mathbb{P}[\widehat{g}(X)=1]} \\ &\leqslant c_1 K_0\left(\|\eta - \widehat{\eta}\|_1 + \frac{c_2 d_1 - c_1 d_2}{c_0 d_1 - c_1 d_0}\left(\sqrt{\frac{\pi\mathbb{P}(Y=1)}{N}} + \frac{4}{N} + \|\eta - \widehat{\eta}\|_1\right) + 2^{-K_{\max}}\right) \end{aligned}$$

Finally, taking the expectation $\mathbb{E}_{\mathcal{D}_n^L}$ over the labelled dataset we obtain the results. □

Proof of Lemma 6.

$$\begin{aligned}
\frac{\mathrm{U}(g_{\theta^*}) - \mathrm{U}(\widehat{g})}{c_1 K_0} &\leqslant \mathbb{E}[|\eta(X) - \theta^*|\mathbb{1}\{g_{\theta^*}(X) \neq \widehat{g}(X)\}] \\
&\leqslant \mathbb{E}[|\eta(X) - \theta^*|\mathbb{1}\{|\eta(X) - \theta^*| \leqslant 2|\eta(X) - \widehat{\eta}(X)|\}]+ \\
&\quad + \mathbb{E}[|\eta(X) - \theta^*|\mathbb{1}\{|\eta(X) - \theta^*| \leqslant 2|\widehat{\theta} - \theta^*|\}] \\
&\leqslant \mathbb{E}[|\eta(X) - \theta^*|\mathbb{1}\{|\eta(X) - \theta^*| \leqslant 2|\eta(X) - \widehat{\eta}(X)|\}\mathbb{1}\{|\eta(X) - \theta^*| \leqslant t\}]+ \\
&\quad + \mathbb{E}[|\eta(X) - \theta^*|\mathbb{1}\{|\eta(X) - \theta^*| \leqslant 2|\eta(X) - \widehat{\eta}(X)|\}\mathbb{1}\{|\eta(X) - \theta^*| > t\}]+ \\
&\quad + 2\mathbb{E}[|\widehat{\theta} - \theta^*|\mathbb{1}\{|\eta(X) - \theta^*| \leqslant 2|\widehat{\theta} - \theta^*|\}] \\
&\leqslant 2\mathbb{E}[|\eta(X) - \widehat{\eta}(X)|\mathbb{1}\{|\eta(X) - \theta^*| \leqslant t\}]+ \\
&\quad + 2\mathbb{E}[|\eta(X) - \widehat{\eta}(X)|\frac{|\eta(X) - \widehat{\eta}(X)|^{p-1}}{t^{p-1}}\mathbb{1}\{|\eta(X) - \theta^*| > t\}]+ \\
&\quad + 2|\widehat{\theta} - \theta^*|\mathbb{P}(|\eta(X) - \theta^*| \leqslant 2|\widehat{\theta} - \theta^*|) \\
&\leqslant 2\|\eta - \widehat{\eta}\|_p \, \mathbb{P}\left(|\eta(X) - \theta^*| \leqslant t\right)^{1-\frac{1}{p}} + 2\frac{\|\eta - \widehat{\eta}\|_p^p}{t^{p-1}} + 2^{1+\alpha}M|\widehat{\theta} - \theta^*|^{1+\alpha} \\
&\leqslant 2M^{1-\frac{1}{p}}\|\eta - \widehat{\eta}\|_p t^{\alpha(1-\frac{1}{p})} + 2\frac{\|\eta - \widehat{\eta}\|_p^p}{t^{p-1}} + 2^{1+\alpha}M|\widehat{\theta} - \theta^*|^{1+\alpha}
\end{aligned}$$

Optimizing for t we get (from [Audibert and Tsybakov, 2007], to check...)

$$\frac{\mathrm{U}(g_{\theta^*}) - \mathrm{U}(\widehat{g})}{c_1 K_0} \leqslant 2\frac{p+\alpha}{p}\left(\frac{p}{\alpha}\right)^{\frac{\alpha}{p+\alpha}} M^{\frac{p-1}{p+\alpha}}\|\eta - \widehat{\eta}\|_p^{\frac{p(1+\alpha)}{p+\alpha}} + 2^{1+\alpha}M|\widehat{\theta} - \theta^*|^{1+\alpha}$$

To control the last term notice that

$$|\widehat{\theta} - \theta^*|^{1+\alpha} \leqslant 2^{\alpha}(|\widehat{\theta} - \bar{\theta}|^{1+\alpha} + |\bar{\theta} - \theta^*|^{1+\alpha})$$

As regards the first term, by using the bisection algorithm, we know $|\widehat{\theta} - \bar{\theta}| \leqslant 2^{-K_{\max}}$ almost surely, where $K_{\max}$ is the number of iterations. Then we simply have:

$$|\widehat{\theta} - \bar{\theta}|^{1+\alpha} \leqslant 2^{-(1+\alpha)K_{\max}} \quad \text{a.s.}$$

Let's focus now on the second term and, similarly to the proof of Proposition 5, we have:

$$\frac{1}{\mathcal{B}^{1+\alpha}}|\bar{\theta} - \theta^*|^{1+\alpha} \leqslant \|\widehat{F}_{\widehat{\eta}} - F_\eta\|_1^{1+\alpha} \leqslant 2^{\alpha}\left(\|F_\eta - \widehat{F}_\eta\|_1^{1+\alpha} + \|\widehat{F}_{\widehat{\eta}} - \widehat{F}_\eta\|_1^{1+\alpha}\right). \tag{H.3}$$

For the first term we use:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_N^U}\|F_\eta - \widehat{F}_\eta\|_1^{1+\alpha} &= \mathbb{E}_{\mathcal{D}_N^U}\left(\int_0^1 |P(\eta(X) \geqslant t) - \widehat{P}_N(\eta(X) \geqslant t)|dt\right)^{1+\alpha} \\
&\leqslant \int_0^1 \mathbb{E}_{\mathcal{D}_N^U}|P(\eta(X) \geqslant t) - \widehat{P}_N(\eta(X) \geqslant t)|^{1+\alpha}dt.
\end{aligned}$$

and, with $p(t) = P(\eta(X) \geqslant t)$, we get:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_N^U} |P(\eta(X) \geqslant t) - \widehat{P}_N(\eta(X) \geqslant t)|^{1+\alpha} \\
&\qquad = \int_0^{+\infty} P_{\mathcal{D}_N^U} \left(|P(\eta(X) \geqslant t) - \widehat{P}_N(\eta(X) \geqslant t)|^{1+\alpha} \geqslant x \right) dx \\
&\qquad = \int_0^{+\infty} P_{\mathcal{D}_N^U} \left(|P(\eta(X) \geqslant t) - \widehat{P}_N(\eta(X) \geqslant t)| \geqslant x^{\frac{1}{1+\alpha}} \right) dx \\
&\qquad = \int_0^{+\infty} P_{\mathcal{D}_N^U} \left(|P(\eta(X) \geqslant t) - \widehat{P}_N(\eta(X) \geqslant t)| \geqslant u \right) (1+\alpha) u^\alpha du \\
&\qquad \leqslant 2(1+\alpha) \int_0^{+\infty} \exp\left(- \frac{N u^2}{2(p(t) + \frac{1}{3}u)} \right) u^\alpha du \\
&\qquad = 2(1+\alpha) \left(\int_0^{3p(t)} u^\alpha \exp\left(- \frac{N u^2}{4p(t)} \right) du + \int_{3p(t)}^{+\infty} u^\alpha \exp\left(- \frac{Nu}{4} \right) du \right) \\
&\qquad \leqslant (1+\alpha) \left(\left(\frac{4p(t)}{N} \right)^{\frac{1+\alpha}{2}} \Gamma\left(\frac{1+\alpha}{2} \right) + 2 \left(\frac{4p(t)}{N} \right)^{1+\alpha} \Gamma\left(1+\alpha\right) \right)
\end{aligned}
$$

Finally

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}_N^U} \|F_\eta - \widehat{F}_\eta\|_1^{1+\alpha} &\leqslant (1+\alpha) \int_0^1 \left(\frac{4p(t)}{N} \right)^{\frac{1+\alpha}{2}} \Gamma\left(\frac{1+\alpha}{2} \right) + 2 \left(\frac{4p(t)}{N} \right)^{1+\alpha} \Gamma\left(1+\alpha\right) dt \\
&\leqslant 4^{\alpha+2}(1+\alpha) \left(\frac{\sqrt{P(Y=1)}}{N^{\frac{1+\alpha}{2}}} \Gamma\left(\frac{1+\alpha}{2} \right) + \frac{P(Y=1)}{N^{1+\alpha}} \Gamma\left(1+\alpha\right) \right)
\end{aligned}
$$

For the second term in Eq. (H.3), similarly to the proof of Proposition 5, we have:

$$
\begin{aligned}
\left\| \widehat{F}_{\widehat{\eta}} - \widehat{F}_\eta \right\|_1^{1+\alpha} &= \left(\inf_{\omega \in \mathbb{S}_N} \frac{1}{N} \sum_{i=n+1}^{n+N} \left| Z_i - \widehat{Z}_{\omega(i)} \right| \right)^{1+\alpha} \leqslant \left(\frac{\sum_{i=n+1}^{n+N} |\eta(X_i) - \widehat{\eta}(X_i)|}{N} \right)^{1+\alpha} \\
&\leqslant \frac{\sum_{i=n+1}^{n+N} |\eta(X_i) - \widehat{\eta}(X_i)|^{1+\alpha}}{N}.
\end{aligned}
$$

Taking the expectation both sides we obtain:

$$
\mathbb{E}_{\mathcal{D}_N^U} \left\| \widehat{F}_{\widehat{\eta}} - \widehat{F}_\eta \right\|_1^{1+\alpha} \leqslant \|\eta - \widehat{\eta}\|_{1+\alpha}^{1+\alpha}.
$$

Putting things together we recover the final bound:

$$\begin{aligned}\frac{\mathrm{U}(g_{\theta^*}) - \mathrm{U}(\widehat{g})}{c_1 K_0} \leqslant & 2\frac{p+\alpha}{p}\left(\frac{p}{\alpha}\right)^{\frac{\alpha}{p+\alpha}} M^{\frac{p-1}{p+\alpha}} \|\eta - \widehat{\eta}\|_p^{\frac{p(1+\alpha)}{p+\alpha}} + 2^{1+3\alpha} M K_1^{1+\alpha} \|\eta - \widehat{\eta}\|_{1+\alpha}^{1+\alpha} \\ & + 2^{5+5\alpha} M K_1^{1+\alpha}(1+\alpha)\left(\frac{\sqrt{P(Y=1)}}{N^{\frac{1+\alpha}{2}}}\Gamma\left(\frac{1+\alpha}{2}\right) + \frac{P(Y=1)}{N^{1+\alpha}}\Gamma\left(1+\alpha\right)\right) \\ & + M 2^{1+2\alpha} 2^{-(1+\alpha)K_{\max}}.\end{aligned}$$

□