Proceedings of Machine Learning Research vol 134:1–43, 2021

Fast Rates for Structured Prediction

Vivien Cabannes VIVIEN.CABANNES@GMAIL.COM
Francis Bach FRANCIS.BACH@GMAIL.COM
Alessandro Rudi ALESSANDRO.RUDI@GMAIL.COM
INRIA - Département d'Informatique de l'École Normale Supérieure - PSL Research University, Paris, France

Abstract

Discrete supervised learning problems such as classification are often tackled by introducing a continuous surrogate problem akin to regression. Bounding the original error, between estimate and solution, by the surrogate error endows discrete problems with convergence rates already shown for continuous instances. Yet, current approaches do not leverage the fact that discrete problems are essentially predicting a discrete output when continuous problems are predicting a continuous value. In this paper, we tackle this issue for general structured prediction problems, opening the way to "super fast" rates, that is, convergence rates for the excess risk faster than n^{-1}, where n is the number of observations, with even exponential rates with the strongest assumptions. We first illustrate it for predictors based on nearest neighbors, generalizing rates known for binary classification to any discrete problem within the framework of structured prediction. We then consider kernel ridge regression where we improve known rates in $n^{-1/4}$ to arbitrarily fast rates, depending on a parameter characterizing the hardness of the problem, thus allowing, under smoothness assumptions, to bypass the curse of dimensionality.

Keywords: Structured prediction, fast convergence rates, generalization bounds, low-density separation, margin condition, local averaging method, nearest neighbors, kernel methods, kernel ridge regression.

1. Introduction

Machine learning is raising high hopes to tackle a wide variety of prediction problems, such as language translation, fraud detection, traffic routing, speech recognition, self-driving cars, DNA-binding proteins, *etc.*. Its framework is appreciated as it removes humans from the burden to come up with a set of precise rules to accomplish a complex task, such as recognizing a cat on an array of pixels. Yet, it comes at a price, which is of forgetting about algorithm correctness, meaning that machine learning algorithms can make mistakes, *i.e.*, wrong predictions, which can have dramatic implications, *e.g.*, in medical applications. This motivates work on generalization error bounds, quantifying how often one should expect errors.

Many of the problems discussed above are of discrete nature, in the sense that the number of potential outputs is finite, or infinite countable. To learn such problems, a classical technique consists in defining a continuous surrogate problem, which is easier to solve, and such that:

(1) an algorithm on the surrogate problem translates into an algorithm on the original problem;
(2) errors on the original problem are bounded by errors on the surrogate problem.

The first point refers to the concept of plug-in algorithms, while the second point to the notion of calibration inequalities. For example, binary classification can be approached through regression by estimating the conditional expectation of the output Y given an input X (Bartlett et al., 2006).

On the one hand, continuous surrogates for discrete problems are interesting, as they benefit from functional analysis knowledge, when discrete problems are more combinatorial in nature. On the other hand, continuous surrogate can be deceptive, as they are asking to solve for more than needed. Considering the example of binary classification, where $Y \in \{-1, 1\}$, one only has to predict the sign of the conditional expectation, rather than its precise value. Interestingly, without modifying the continuous surrogate approach, this last remark can be leveraged in order to tighten generalization bounds derived through calibration inequalities (Audibert and Tsybakov, 2007). In this work, we extend those considerations, known in binary classification (*e.g.,* Koltchinskii and Beznosova, 2005; Chaudhuri and Dasgupta, 2014), to generic discrete supervised learning problems, and show how it can be applied to the kernel ridge regression algorithm introduced by Ciliberto et al. (2016).

1.1. Contributions

Our contributions are organized in the following order.

- In Section 2, we consider the general structured prediction from Ciliberto et al. (2020) and derive refined calibration inequalities to leverage the fact that learning a mapping into a discrete output space is easier than learning a mapping into a continuous space.
- In Section 3, we show how to exploit exponential concentration inequalities to turn them into fast rates under a condition generalizing the Tsybakov margin condition.
- In Section 4, we apply Section 3 to local averaging methods with the particular example of nearest neighbors. This leads to extending the rates known for regression and classification to a wide variety of structured prediction problems, with rates that match minimax rates known in binary classification.
- In Section 5, we show how Section 3 can be applied to kernel ridge regression. This allows us to improve rates known in $n^{-1/4}$ to arbitrarily fast rates depending on the hardness of the associated discrete problem.

1.2. Related work

Surrogate framework. The surrogate problem we will consider to tackle structured prediction finds its roots in the approximate Bayes rule proposed by Stone (1977), analyzed through the prism of mean estimation as suggested by Friedman (1994) for classification, and analyzed by Ciliberto et al. (2020) in the wide context of structured prediction. In particular, we will specify results on two classes of surrogate estimators: local averaging methods, or kernel ridge regression.

Local averaging methods. Neighborhood methods were first studied by Fix and Hodges (1951) for statistical testing through density estimation. Similarly Parzen–Rosenblatt window methods (Parzen, 1962; Rosenblatt, 1956) were developed. Those methods were cast in the context of regression as nearest neighbors (Cover and Hart, 1967) and Nadayara-Watson estimators (Watson, 1962; Nadaraya, 1964). Stone (1977) was the first to derive consistency results for a large class of localized methods, among which are nearest neighbors and some window estimators (Spiegelman and Sacks, 1980; Devroye and Wagner, 1980). Rates were then derived, with minimax optimality (Stone, 1980; Yang, 1999). Several reviews can be found in the literature, such as Györfi et al. (2002); Tsybakov (2009); Biau and Devroye (2015); Chen and Shah (2018).

Reproducing kernel ridge regression. The theory of real-valued reproducing kernel Hilbert spaces was formalized by Aronszajn (1950), before finding applications in machine learning (*e.g.,* Scholkopf and Smola, 2001). Minimax rates for kernel ridge regression were achieved by casting the empirical solution estimate as a result of integral operator approximation (Smale and Zhou, 2007; Caponnetto and De Vito, 2006), allowing to control convergence through concentration inequalities in Hilbert spaces (Yurinskii, 1970; Pinelis and Sakhanenko, 1986) and on self-adjoint operating on Hilbert spaces (Minsker, 2017). First derived in L^2-norm, rates were cast in L^∞-norm through interpolation inequalities (*e.g.,* Fischer and Steinwart, 2020; Lin et al., 2020).

Tsybakov margin condition. Learning a mapping into a discrete output space is indeed easier than learning a continuous mapping, as, for binary classification for example, one typically only has to predict the sign of $\mathbb{E}[Y\,|\,X]$ rather than its precise value. As such, calibration inequalities that relate the error on a discrete structured prediction problem to an error on a smooth surrogate problem are often suboptimal. This phenomenon was exploited for density discrimination, a problem consisting of testing if samples were drawn from one or the other of two potential distributions, by Mammen and Tsybakov (1999), and for binary classification by Audibert and Tsybakov (2007). Those works introduce a parameter $\alpha \in [0, \infty)$ characterizing the hardness of the discrete problem, and leverage concentration inequalities to accelerate rates known for regression by a power $\alpha + 1$ (Audibert and Tsybakov, 2007), while rates plugged-in directly through calibration inequalities only present an acceleration by a power ${}^{2(\alpha+1)}/_{(\alpha+2)}$ (*see, e.g.,* Boucheron et al., 2005; Bartlett et al., 2006; Bartlett and Mendelson, 2006; van Erven et al., 2015; Nowak-Vila et al., 2019).

2. Structured Prediction with Surrogate Control

In this section, we introduce the classical supervised learning problem, and a surrogate problem that consists of conditional mean estimation. We recall a calibration inequality relating the original problem to the surrogate one. We mention how empirical estimations of the conditional means usually deviate from the real means following a sub-exponential tail bound, similarly to bounds obtained through Bernstein inequality. We end this section by providing refined surrogate control, that is the key towards "super fast" rates, that is, rates faster than $1/n$.

2.1. Surrogate mean estimation

Consider a classic supervised learning problem, where given an input space $\mathcal{X}$, an observation space $\mathcal{Y}$, a prediction space $\mathcal{Z}$, a joint distribution $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ and a loss function $\ell : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}_+$, one would like to retrieve $f^* : \mathcal{X} \to \mathcal{Z}$ minimizing the risk $\mathcal{R}$.

$$f^* \in \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Z}} \mathcal{R}(f) \qquad \text{with} \qquad \mathcal{R}(f) = \mathbb{E}_{(X,Y)\sim\rho}\left[\ell(f(X), Y)\right].$$

In practice, $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ and ℓ are givens of the problem, while ρ is unknown, yet partially observed thanks to a dataset $\mathcal{D}_n = (X_i, Y_i)_{i\leq n} \sim \rho^{\otimes n}$, with data (X_i, Y_i) sampled independently from ρ. Note that in fully supervised learning, the observation space is the same as the prediction space $\mathcal{Y} = \mathcal{Z}$, yet we distinguish the two for our results to stand in more generic settings, such as instances of weak supervision (Cabannes et al., 2020). In the following, we consider $\mathcal{Z}$ finite. In several cases, solving the supervised learning problem can be done through solving a surrogate problem that is easier to handle. Ciliberto et al. (2016) provide a setup that reduces a wide variety of structured prediction problems (ℓ, ρ) to a problem of mean estimation. It works under the following assumption.

Assumption 1 (Bilinear loss decomposition) *There exists an Hilbert space $\mathcal{H}$ and two mappings $\psi : \mathcal{Z} \to \mathcal{H}$, $\varphi : \mathcal{Y} \to \mathcal{H}$ such that*

$$\ell(z, y) = \langle \psi(z), \varphi(y) \rangle\,.$$

We will also assume that ψ is bounded (in norm) by a constant c_ψ.

This assumption is not really restrictive (Ciliberto et al., 2020). Among others, it works for any losses on finite spaces, usually with spaces $\mathcal{H}$ whose dimensionality is only polylogarithmic with respect to the cardinality of $\mathcal{Z}$ (Nowak-Vila et al., 2019). Under Assumption 1, solving the supervised learning problem can be done through estimating the surrogate conditional mean $g^* : \operatorname{supp} \rho_{\mathcal{X}} \to \mathcal{H}$, defined as

$$g^*(x) = \mathbb{E}_{Y \sim \rho|_x} [\varphi(Y)] \,, \tag{1}$$

where we denote $\rho|_x$ the conditional law of $(Y \mid X)$ under $(X, Y) \sim \rho$.

Lemma 1 (Ciliberto et al. (2016)) *Given an estimate g_n of g^* in Eq. (1), consider the estimate $f_n : \mathcal{X} \to \mathcal{Z}$ of f^*, which is obtained from "decoding" g_n as*

$$f_n(x) = \operatorname*{arg\,min}_{z \in \mathcal{Z}} \langle \psi(z), g_n(x) \rangle \,. \tag{2}$$

Then the excess risk is controlled through the surrogate error as

$$\mathcal{R}(f_n) - \mathcal{R}(f^*) \leq 2c_\psi \left\| g_n - g^* \right\|_{L^1(\mathcal{X}, \mathcal{H}, \rho)} \,. \tag{3}$$

Inequalities relating the original excess risk $\mathcal{R}(f_n) - \mathcal{R}(f^*)$ with a measure of error on a surrogate problem are called *calibration inequalities*. They are useful when the measure of error between g_n and g^* is easier to control than the one between f_n and f^*.

Example 1 (Binary classification) *Binary classification corresponds to $\mathcal{Y} = \mathcal{Z} = \{-1, 1\}$ and $\ell(z, y) = \mathbf{1}_{z \neq y}$ (or equivalently $\ell(z, y) = 2\mathbf{1}_{z \neq y} - 1$). The classical surrogate consists of taking $\mathcal{H} = \mathbb{R}$, with $\varphi = id$ and $\psi = -id$. In this setting, we have $g^*(x) = \mathbb{E}_\rho[Y | X = x]$, and the decoding $f_n(x) := \operatorname{sign} g_n(x)$, for any $g_n(x) \in \mathcal{H}$. In this case $\mathcal{R}(f_n) - \mathcal{R}(f^*) = \mathbb{E}_X \left[\mathbf{1}_{f_n(X) \neq f^*(X)} \left| g^*(X) \right| \right] \leq 2 \left\| g_n - g^* \right\|_{L^1} \leq 2 \left\| g_n - g^* \right\|_{L^2}$. Note that in regression the excess risk reads as the square of the L^2 norm, explaining a loss of a power one half in convergence rates, when going from regression to classification (*e.g. *Chen and Shah, 2018).*

Differences between an empirical estimate and its population version are generally handled through concentration inequalities. In this work, we will leverage concentration on $\|g_n(x) - g(x)\|$ that is uniform for $x \in \operatorname{supp} \rho_{\mathcal{X}}$, motivating the introduction of Assumption 2.

Assumption 2 (Exponential concentration inequality) *Suppose that for $n \in \mathbb{N}$, there exists two reals L_n and M_n, such that the tails of $\|g_n(x) - g(x)\|$ can be controlled for any $t > 0$ as*

$$\sup_{x \in \operatorname{supp} \rho_{\mathcal{X}}} \mathbb{P}_{\mathcal{D}_n} \left(\|g_n(x) - g(x)\| > t \right) \leq \exp\left(-\frac{L_n t^2}{1 + M_n t} \right) . \tag{4}$$

Note that to satisfy Assumption 2, it is sufficient, yet *not necessary*, to have a uniform control on $g_n - g^*$, *i.*e., a control on the tail of $\|g_n - g^*\|_{L^\infty}$, since $\sup_x \mathbb{P}\left(A_x > t\right) \leq \mathbb{P}\left(\cup_x \{A_x > t\}\right) = \mathbb{P}\left(\sup_x A_x > t\right)$, with (A_x) a family of random variables indexed by $x \in \mathcal{X}$.

Usually, in bounds like Eq. (4), M_n is a constant of the problem, while L_n depends on the number of samples, therefore, we would like to give rates depending on L_n. Typically in Bernstein inequalities (see Theorem 32 in Appendix), $L_n = n\sigma^{-2}$ with σ^2 a variance parameter and $M_n = c\sigma^{-2}$ with c a constant of the problem that does not depend on n.

2.2. Refined Calibration

While it is sufficient to control the excess risk through a L^1-norm control on g from Eq. (3), it is not always necessary. In other words, the calibration bound in Lemma 1 is not always tight. Indeed, we do not predict optimally, that is, $\{f_n(x) \neq f^*(x)\}$ only if $g_n(x)$ and $g^*(x)$ do not lead to the same decoding $f_n(x)$ and $f^*(x)$. When $\mathcal{Z}$ is finite, this is characterized by $g_n(x)$ and $g^*(x)$ not falling in the same region R_z of $\mathcal{H}$, where

$$R_z = \left\{\xi \in \mathcal{H} \middle| z \in \operatorname*{arg\,min}_{z' \in \mathcal{Z}} \left\langle \psi(z'), \xi \right\rangle \right\}.$$

To ensure that $g_n(x)$ and $g^*(x)$ fall in the same region, one can ensure that $g_n(x)$ is closer to $g^*(x)$ than $g^*(x)$ is of the frontier of those regions. Those frontiers are defined by points leading to at least two minimizers in Eq. (2):

$$F = \left\{\xi \in \mathcal{H} \,\middle|\, \left|\operatorname*{arg\,min}_{z \in \mathcal{Z}} \left\langle \psi(z), \xi \right\rangle\right| > 1 \right\}.$$

The introduction of F is motivated by the following geometric results.

Lemma 2 (Refined surrogate control) *When $\mathcal{Z}$ is finite, for any $x \in \operatorname{supp} \rho_{\mathcal{X}}$,*

$$\|g_n(x) - g^*(x)\| < d(g^*(x), F) \qquad \Rightarrow \qquad f_n(x) = f^*(x),$$

with d the extension of the norm distance to sets as $d(g^*(x), F) = \inf_{\xi \in F} \|g^*(x) - \xi\|$. This result allows to refine the calibration control from Lemma 1 as

$$\mathcal{R}(f_n) - \mathcal{R}(f^*) \leq 2c_\psi \, \mathbb{E}_X \left[\mathbf{1}_{\|g_n(X) - g^*(X)\| \geq d(g^*(X), F)} \|g_n(X) - g^*(X)\|\right]. \tag{5}$$

Example 2 (Binary classification) *In binary classification* (cf. *Example 1), $F = \{0\}$, and, for any $x \in \operatorname{supp} \rho_{\mathcal{X}}$, $d(g^*(x), F) = |g^*(x)|$. Lemma 2 is based on the fact that $f^*(x) \neq f_n(x)$ implies that* $\operatorname{sign} g^*(x) \neq \operatorname{sign} g_n(x)$ *which itself implies that $|g^*(x) - g_n(x)| = |g^*(x)| + |g_n(x)| \geq |g^*(x)|$.*

To leverage Eq. (5), we need to control $d(g^*(x), F)$ below and $\|g_n(x) - g^*(x)\|$ above. While upper bounds on $\|g_n(x) - g^*(x)\|$ are assumed to have been derived through concentration inequalities, lower bounds on $d(g^*(x), F)$ will be assumed as a given parameter of the problem, see Eqs. (6) and (7).

Remark 3 (Scope of our work) *While we derived the refined calibration inequality Eq. (5) for the surrogate conditional mean g^* and the associated pointwise metric $\|\cdot\|_{\mathcal{H}}$, similar inequality could be obtained for other type of surrogate methods. This suggests that our work could be extended to any smooth surrogate such as the ones considered by Nowak-Vila et al. (2020), as well as Fenchel-Young losses (Blondel et al., 2020).*

2.3. Geometric understanding

In this subsection, we detail how to understand geometrically Lemma 2. While the introduction of φ and ψ could seem arbitrary, it can be thought in a more intrinsic manner by considering the embedding $\varphi(y) = \delta_y$ belonging to the Banach space $\mathcal{H}$ of signed measured, $g^*(x) = \rho|_x$, with the bracket operator, for $\mu \in \mathcal{H}$ and $z \in \mathcal{Z}$, $\langle z, \mu \rangle = \int_{\mathcal{Y}} \ell(z, y)\mu(\mathrm{d}y)$, and the distance between signed measures being $d(\mu_1, \mu_2) = \sup_{z \in \mathcal{Z}} \langle z, \mu_1 - \mu_2 \rangle$. Note that Lemma 2 is a pointwise result, holding for any $x \in \mathcal{X}$, that is integrated over $\mathcal{X}$ afterwards. Therefore, it is enough to consider $\mathcal{X} = \{x\}$ and remove the dependency in $\mathcal{X}$ to understand it. The simplex $\Delta_{\mathcal{Y}}$ naturally splits into decision region R_z for $z \in \mathcal{Z}$ as illustrated on Figure 1. The main idea of Lemma 2 is that one does not have to precisely estimate $g^*(x) = \rho|_x$ but only has to make sure that $g_n(x)$ falls in the same region on Figure 1.

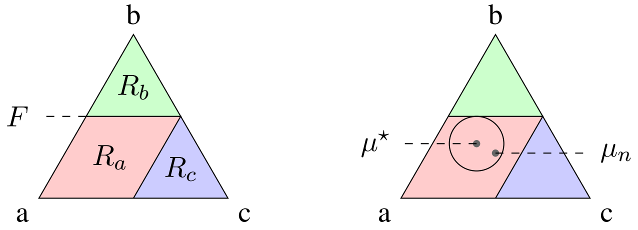

Figure 1: Illustration of Lemma 2. Simplex $\Delta_{\mathcal{Y}}$, for $\mathcal{Y} = \mathcal{Z} = \{a, b, c\}$ and ℓ a symmetric loss defined as $\ell(a, b) = \ell(a, c) = 1$ and $\ell(b, c) = 2$, while $\ell(z, z) = 0$. This leads to the decision regions R_z represented in colors. Given $x \in \mathcal{X}$, if $g^*(x)$ corresponds to a distribution $\mu^* := \rho|_x$ falling in R_a, and if $g_n(x)$ represented by $\hat{\mu}$ falls closer to μ^* than the distance between μ^* and the decision frontier F (represented by a circle on the right figure), then $\hat{\mu}$ is also in R_a, and therefore $f^*(x) = f_n(x) = a$.

3. Rate acceleration under margin condition

In this section, we introduce a condition that g^* is not too often close to the decision frontier F. It generalizes the so-called "Tsybakov margin condition" known for classification. Under this condition, we proves rates that generalize the results of Audibert and Tsybakov (2007) from binary classification to generic structured prediction problems, which opens the way to "super fast" rates in structured prediction.

3.1. No density separation

To get fast convergence rates, one has to make assumptions on the problem. A classical assumption is that g^* is smooth enough in order to get concentration bounds similar to Assumption 2 when considering a specific class of estimates g_n. In our decoding setting (Lemma 1), learning is made easy when it is easy to estimate in which region R_z the optimal g^* will fall in. This is in particular the case, when there is a margin $t_0 > 0$, for which, for no point $x \in \operatorname{supp} \rho_{\mathcal{X}}$, $g^*(x)$ falls at distance t_0 of the decision frontier F, motivating the following definition.

Assumption 3 (No-density separation) *A surrogate solution g^* will be said to satisfy the* no-density separation*, if there exists a $t_0 > 0$, such that*

$$\mathbb{P}_X(d(g^*(X), F) < t_0) = 0. \tag{6}$$

This condition is alternatively called the hard margin condition*, or sometimes "Massart's noise condition" for binary classification (Massart and Nédélec, 2006).*

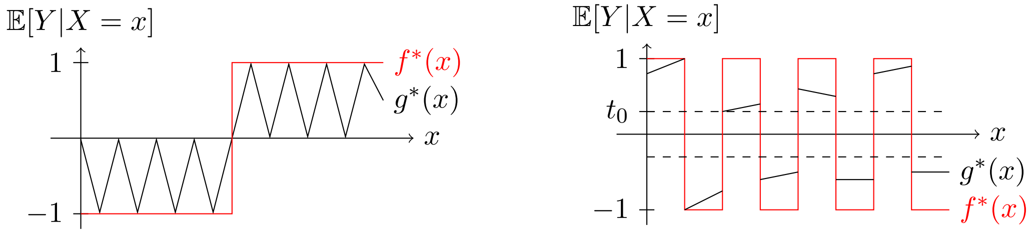

Figure 2: Illustration of Remark 4. We represent two instances of binary classification (see Examples 1 and 2). On the left example, when $\rho_{\mathcal{X}}$ is such that there is no mass where the sign of g^* changes, classes are separated in $\mathcal{X}$, yet the no-density separation is not verified. On the right, classes are not separated in $\mathcal{X}$, but the problem satisfies the no-density separation as there is no x such that $d(g^*(x), F) = |g^*(x)| < t_0$. Note that when $\rho_{\mathcal{X}}$ is uniform, the left problem satisfies a milder separation condition, introduced thereafter and called the 1-low-density separation.

Remark 4 (Separation in $\mathcal{Y}$ and separation in $\mathcal{X}$) *It is important to realize that Eq. (6) is a condition of separation in $\Delta_{\mathcal{Y}}$ that should hold for all $x \in \mathcal{X}$, but it does not state any separation between classes in $\mathcal{X}$ for $f^* : \mathcal{X} \to \mathcal{Z}$. To visualize it, consider the classification problem where $\mathcal{X} = [-1, 1]$, $\mathcal{Y} = \mathcal{Z} = \{-1, 1\}$ and $\ell(z, y) = \mathbf{1}_{z \neq y}$.*

- *A situation where $\rho_{\mathcal{X}}$ is uniform on $\mathcal{X}$ and $\mathbb{E}[Y|X = x] = 2 \cdot \mathbf{1}_{x \in p\mathbb{N} + \{a \,|\, |a| < p/4\}} - 1$, for $p = 1/50$, satisfies separation in $\Delta_{\mathcal{Y}}$ (Eq. (6)), but classes are not separated in $\mathcal{X}$.*
- *A situation where ρ is uniform on $[-1, -.5] \cup [.5, 1]$, with $\mathbb{E}[Y|X = x] = \operatorname{sign}(x)(1 - |x|)^p$, for $p > 0$, satisfies a separation of classes in $\mathcal{X}$ but does not satisfy Eq. (6).*

Note that continuity of g^* and the no-density separation in Eq. (6) imply separation of classes in $\mathcal{X}$. Note also that to get concentration inequality such as Eq. (4), one usually supposes that g^* is smooth. We refer the curious reader to Section 2.4 in Steinwart and Scovel (2007) for separation in $\mathcal{X}$.

The introduction of Assumption 3 is motivated by the following result.

Theorem 5 (Rates under no-density separation) *When ℓ is bounded by ℓ_∞* (i.e., $\ell(z, y) \leq \ell_\infty$ *for any $(z, y) \in \mathcal{Z} \times \mathcal{Y}$) and satisfies Assumption 1, and $\mathcal{Z}$ is finite, under the no-density separation Assumption 3, and the concentration Assumption 2, the excess risk is controlled*

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \ell_\infty \exp\left(-\frac{L_n t_0^2}{1 + M_n t_0}\right).$$

Proof Because we make a mistake only when $d(g^*(x), F) \geq \|g_n(x) - g^*(x)\|$, we make no mistake when $\|g_n(x) - g^*(x)\| < t_0$; otherwise we can consider the worse error we are going to pay, that is ℓ_∞, leading to

$$\mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \ell_\infty \mathbb{P}_X(\|g_n(x) - g^*(x)\| > t_0).$$

Taking the expectation with respect to $\mathcal{D}_n$ and using the fact that $\mathbb{E}_A \mathbb{P}_B(Z) = \mathbb{E}_A \mathbb{E}_B[\mathbf{1}_Z] = \mathbb{E}_B \mathbb{E}_A[\mathbf{1}_Z] = \mathbb{E}_B \mathbb{P}_A(Z)$, and plug-in the concentration inequality Eq. (4), we get the result. ■

Example 3 (Image classification) *In image classification, one can arguably assume that the class of an image is a deterministic function of this image. With the 0-1 loss, it implies that the image classification problem verifies the no-density separation. The same holds for any discrete problem where the label is a deterministic function of the input. Based on Theorem 5 and Eq. (4) in which M is generally a constant when L is proportional to the number of data, it is reasonable to ask for exponential convergences rates on such problems.*

3.2. Low density separation

While we presented the no-density separation first for readability, it is a strong assumption. Recall our example, Remark 4, with $\mathbb{E}[Y|X=x] = \operatorname{sign}(x)(1-|x|)^p$, only around $x=1$ and $x=-1$ is $d(g^*(x), F)$ not bounded away from zero. While the neighborhood of those points should be studied carefully, the error on all other points $x \in [-1+t, 1-t]$ can be controlled with exponential rates. The low-density separation, also known as the Tsybakov margin condition in binary classification, will allow a refined control to get fast rates in such a setting.

Assumption 4 (Low-density separation) *A surrogate solution g^* is said to satisfy the* low-density separation, *if there exists $c_\alpha > 0$, and $\alpha > 0$, such that for any $t > 0$*

$$\mathbb{P}_X(d(g^*(X), F) < t) \leq c_\alpha t^\alpha. \tag{7}$$

This condition is alternatively called the margin condition.

The low-density separation spans all situations from the hard margin condition, that can be seen as $\alpha = +\infty$, to situations without any margin assumption corresponding to $\alpha = 0$. The coefficient α is an intrinsic measure of the easiness of finding f^* in the problem (ℓ, ρ). For example, the setting described in the last paragraph corresponds to the case $\alpha = 1/p$. We discuss the equivalence of Assumption 4 to definitions appearing in the literature in Remark 7.

Theorem 6 (Optimal rates under low density separation) *Under refined calibration in Eq. (5), concentration (Assumption 2), and low-density separation (Assumption 4), the risk is controlled as*

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq 2c_\psi c_\alpha c \left(M_n^{\alpha+1} L_n^{-(\alpha+1)} + L_n^{-\frac{\alpha+1}{2}} \right),$$

for c a constant that only depends on α, that can be expressed through the Gamma function evaluated in quantity depending on α, meaning that when α is big, c behaves like α^α. Note that it is not possible to derive a better bound only given Eqs. (4), (5) and (7).

Proof [Sketch for Theorem 6, details in Appendix A.5] Based on the refined calibration inequality in Eq. (5), and using that $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)\,\mathrm{d}t$, it is possible to show that the expectation of the excess risk behave like

$$\int_0^\infty \mathbb{P}_X(d(g^*(x), F) < t) \sup_x \mathbb{P}_{\mathcal{D}_n}(\|g_n(x) - g^*(x)\| > t)\,\mathrm{d}t.$$

Based on Assumptions 2 and 4, the integrand behaves like $t^\alpha \exp(-L_n t^2/(1+M_n t))$. A change of variable and the study of the Gamma function leads to the result. We provide all the details in Appendix A.5. Note that while we stated Theorem 6 under an exponential inequality of Bernstein type (Assumption 2), similar theorems can be derived for any type of exponential concentration inequality, as stated in Lemma 19 in Appendix A.6. ■

Theorem 6 is to put in perspective with the work of Nowak-Vila et al. (2019) which considers the same setup as ours, yet only succeeds to derive acceleration by a power $2(\alpha+1)/(\alpha+2)$, while we got an acceleration by a power $(\alpha+1)/2$ as already mentioned in the related work section. This gain will appear more clearly in Theorem 15.

Remark 7 (Independence to the decomposition of ℓ) *While we have stated results based on the quantity $d(g^*(x), F)$, generalization of the Tsybakov margin condition has also been expressed through the quantity $\inf_{z \neq z^*} \mathbb{E}_{Y\sim\rho|_x} \ell(z, Y) - \mathbb{E}_{Y\sim\rho|_x} \ell(z^*, Y)$ instead of $d(g^*(x), F)$ (Nowak-Vila et al., 2019). We show in Appendix A.3 that the two definitions of the margin condition are equivalent.*

Remark 8 (Scope of our work) *Our work relies on pointwise exponential concentration inequalities (Assumption 2) which are specially designed to work well with the Tsybakov margin condition. It is natural for localized averaging method such as nearest neighbors, or for surrogate methods leading to L^∞ concentration. For surrogate methods leading to concentration of other quantities, it is possible to use similar tricks under different "margin" conditions (*e.g. *Steinwart and Scovel (2007) for a margin condition designed for the Hinge loss). Note that L^2 concentration on g_n towards g^* (such as the one derived by Marteau-Ferey et al. (2019) for logistic regression) could also be turned into fast convergence of f_n towards f^*, since, in essence, for points $x \in \mathcal{X}$ where $\rho(\mathrm{d}x)$ is high, the quantity $g^*(x) - g_n(x)$ will have a non-negligible contribution to $\|g^* - g_n\|_{L^2}$ – allowing to cast concentration in L^2 to concentration pointwise in x – and for points $x \in \mathcal{X}$ where $\rho(\mathrm{d}x)$ is negligible, it is acceptable to pay the worst error, since it will have a small contribution on the excess of risk. Finally, note that it is also possible to let the right hand-side term in Eq.* (4) *depends on x, and to modify Theorem 5 with $L = \mathbb{E}[L(X)]$.*

3.3. The importance of constants

In this subsection, we discuss on the importance of constants when providing learning rates. Assumption 3 corresponds to asking for $g^*(x)$ never to enter a neighborhood of F defined through t_0. Similarly, when $\mathcal{X}$ is parameterized such that $\rho_\mathcal{X}$ is uniform, the parameter α in Assumption 4 corresponds to the speed at which $g^*(x)$ "get through" the decision frontier F. In order to have a higher α and optimize the dependency in n in the bound of Theorem 6, it is natural to think of infinitesimal perturbations of g^* to make it cross the boundary orthogonally (or even jump over it and satisfy the no-density seperation). To give a precise example, in binary classification, let us artificially add smoothness to the function $g^*(x) = x^q$ when approaching zero. Consider $g^* : [0, 1] \to [-1, 1], x \to c^{q-p} x^p \mathbf{1}_{x<c} + x^q \mathbf{1}_{x\geq c}$, and x uniform, and $p < q$. In this setting, α can be taken anywhere in $[0, p^{-1})$. Naively, we could ask for the biggest possible α in order to have the best dependency in n in the learning rates given by Theorem 6. While this approach will higher α, it will also higher c_α, compensating the gain one could expect from such a strategy. Indeed, for $\alpha \in [0, p^{-1}]$, at best, we can take $c_\alpha = \mathbf{1}_{\alpha<q^{-1}} + c^{1-q\alpha}\mathbf{1}_{\alpha\geq q^{-1}}$. This shows the importance to optimize both α and $c_\alpha c$ to minimize the lower bound appearing in Theorem 6 when given a fixed number of sample n.

In a word, while we only give results that are optimized in n, when n is fixed, better bounds could be given by optimizing parameters and constants simultaneously. For example, when $\mathcal{X} = \mathbb{R}^d$ and g^* belongs to the Sobolev space H^m for all $m \in [0, m_*]$, and satisfies Assumption 4 for all $\alpha \in [0, \alpha_*]$, we expect the best bound, that could be derived from our proof technique, to be of form

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \min_{m\leq m_*, \alpha<\alpha_*} \alpha^\alpha c_\alpha c_\psi \|g^*\|_{H^m} n^{-\frac{m(\alpha+1)}{2d}}.$$

Yet, for simplicity, we will express those bounds as $bn^{-\frac{m_*(\alpha_*+1)}{2d}}$, for b a big constant.

4. Application to nearest neighbors

In this section, we consider the Bayes approximate risk estimator proposed by Stone (1977), with weights given by nearest neighbors (Cover and Hart, 1967). We prove, under regularity assumptions, concentration inequalities similar to Eq. (4), which allow us to derive exponential and polynomial rates. Given samples $(X_i, Y_i) \sim \rho^{\otimes n}$, $k \in \mathbb{N}$ and a metric d on $\mathcal{X}$, the estimator is

$$g_n(x) = \sum_{i=1}^{n} \alpha_i(x)\varphi(Y_i), \quad \text{with} \quad \alpha_i(x) = \begin{cases} k^{-1} & \text{if} \quad \sum_{j=1}^{n} \mathbf{1}_{d(x,X_j)\leq d(x,X_i)} < k \\ 0 & \text{if} \quad \sum_{j=1}^{n} \mathbf{1}_{d(x,X_j)< d(x,X_i)} \geq k \\ (pk)^{-1} & \text{else, with } p = \sum_{j=1}^{n} \mathbf{1}_{d(x,X_j)= d(x,X_i)}. \end{cases} \tag{8}$$

To study the convergence of g_n, we introduce the noise free estimator $g_n^* = \sum_{i=1}^n \alpha_i(x) g^*(X_i)$. This will allow to separate the error due to the randomness of the labels $Y_i \sim \rho|_{X_i}$, and the error due to the difference between $g^*(x)$ and the averaging of g^* on the neighbors of x defining g_n. To control the fist error, we need a bounded moment condition on $\varphi(Y)$. We reuse an assumption from Bernstein (1924), that is classic in machine learning (*e.g.*, Caponnetto and De Vito, 2006; Lin et al., 2020).

Assumption 5 (Sub-exponential moment of $\rho|_x$) *Suppose that there exists $\sigma^2, M > 0$ such that for any $x \in \operatorname{supp}\rho_{\mathcal{X}}$, for any $m \geq 2$, we have*

$$\mathbb{E}_{Y\sim\rho|_x}\left[\|\varphi(Y) - g^*(x)\|^m\right] \leq \frac{1}{2} m! \sigma^2 M^{m-2}.$$

Example 4 (Moment bound on $\varphi(Y)$) *Assumption 5 is a classical assumption that is notably satisfied when $\varphi(Y)$ is bounded by M, with σ^2 its variance, or when $(\varphi(Y) \mid X)$ is Gaussian with covariance bounded by a constant independent of X (see a proof of this standard result by Fischer and Steinwart, 2020).*

To control the second error, we notice, for $x \in \operatorname{supp}\rho_{\mathcal{X}}$, that the quantity $\|g^*(x) - g_n^*(x)\|$ behaves like $\sup_{x'\in\mathcal{B}(x,r)} \|g^*(x) - g^*(x')\|$, with r such that $\rho_{\mathcal{X}}(\mathcal{B}(x,r)) \approx k/n$, such a r modeling the distance between x and its k-th neighbor. This motivates the following assumption.

Assumption 6 (Modified Lipschitz condition (Chaudhuri and Dasgupta, 2014)) *g^* is said to verify the β-Modified Lispchitz condition if there exists $c_\beta > 0$ such that for any $x, x' \in \operatorname{supp}\rho_{\mathcal{X}}$*

$$\|g^*(x) - g^*(x')\| \leq c_\beta \rho_{\mathcal{X}}(\mathcal{B}(x, d(x,x')))^\beta,$$

where d is the distance on $\mathcal{X}$, and $\mathcal{B}(x,t) \subset \mathcal{X}$ the ball of center x and radius t.

Typically the β that appears in Assumption 6 is linked with the dimension of a subset of $\mathcal{X}$ containing most the mass of $\rho_{\mathcal{X}}$ (see below). This will slow the rates accordingly to this dimension parameter, a property referred to as the curse of dimensionality.

Example 5 (Classical assumptions) *When $\mathcal{X} = \mathbb{R}^d$, if g is β'-Hölder continuous, and $\rho_{\mathcal{X}}$ is regular in the sense that, there exists a constant c and $t^* > 0$ such that for $x \in \operatorname{supp}\rho_{\mathcal{X}}$ and any $t \in [0, t^*]$, $\rho_{\mathcal{X}}(\mathcal{B}(x,t)) \geq c\lambda(\mathcal{B}(x,t))$, with λ the Lebesgue measure on $\mathcal{X}$, then g satisfies the modified Lipschitz condition with $\beta = \beta'/d$. The condition on $\rho_{\mathcal{X}}$ is usually split in a condition of minimal mass of $\rho_{\mathcal{X}}$, and a condition of regular boundaries of* $\operatorname{supp}\rho_{\mathcal{X}}$ *(*e.g.*, Audibert and Tsybakov, 2007). We provide more details in Appendix B.1.*

We now state convergence results, respectively proven in Appendices B.2, B.3 and B.4, in which the constant values b_1 to b_6 appear explicitly. Note that results provided by Lemma 9 are already known in the literature (Györfi et al., 2002), while Theorems 10 and 11 were only known in binary classification, but we generalize them to any discrete structured prediction problem. It should be noted that rates in Theorem 11 match the minimax rates derived by Audibert and Tsybakov (2007) in the case of binary classification.

Lemma 9 (Nearest neighbors concentration) *Under Assumptions 5 and 6, there exist constants $b_1, b_2, b_3 > 0$, such that for any $x \in \operatorname{supp} \rho_{\mathcal{X}}$ and any $t > 0$,*

$$\mathbb{P}_{\mathcal{D}_n}\left(\|g_n(x) - g_n^*(x)\| > t\right) \leq 2\exp\left(-\frac{b_1 k t^2}{1 + b_2 t}\right).$$

And for $t > (k/2n)^\beta$, when $\rho_{\mathcal{X}}$ is continuous[1]

$$\mathbb{P}_{\mathcal{D}_n}\left(\|g_n^*(x) - g^*(x)\| > t\right) \leq \exp\left(-b_3 n t^{\frac{1}{\beta}}\right).$$

Theorem 10 (Nearest neighbors fast rates under no-density assumption) *When ℓ is bounded by ℓ_∞, satisfies Assumption 1, and $\mathcal{Z}$ is finite, under the no-density separation, Assumption 3, and Assumptions 5 and 6, there exist two constants $b_4, b_5 > 0$ that do not depend on n or k such that for any $n \in \mathbb{N}^*$ and any k such that $(k/2n)^\beta < t_0$, we have*

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq 2\ell_\infty \exp(-b_4 k) + \ell_\infty \exp(-b_5 n). \tag{9}$$

Theorem 11 (Nearest neighbors fast rates under low-density assumption) *When ℓ satisfies Assumption 1, and $\mathcal{Z}$ is finite, under the low-density separation, Assumption 4, and Assumptions 5 and 6, considering the scheme $k_n = \left\lfloor k_0 n^{\frac{2\beta}{2\beta+1}} \right\rfloor$, for any $k_0 > 0$, there exists a constant $b_6 > 0$ that does not depend on n such that for any $n \in \mathbb{N}^*$,*

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq b_6 n^{-\frac{\beta(\alpha+1)}{2\beta+1}}. \tag{10}$$

Remark 12 (Scope of our work) *The same type of argument works for other local averaging methods, such as Nadaraya-Watson (Nadaraya, 1964; Watson, 1962), local polynomials (Cleveland, 1979; Audibert and Tsybakov, 2007) or decision trees (Breiman et al., 1984).*

5. Application to reproducing kernel ridge regression

In this section, we consider the kernel ridge regression estimate g_n of g^* first proposed by Ciliberto et al. (2016), and we prove, under regularity assumptions, uniform concentration inequalities similar to Eq. (4), which allow us to derive super fast rates at the end of the section. Given a symmetric, positive semi-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, the kernel ridge regression estimation g_n of g^* is defined similarly to Eq. (8) yet with weights $\alpha(x) \in \mathbb{R}^n$ defined as

$$\alpha(x) = (\hat{K} + \lambda)^{-1}\hat{K}_x, \quad \hat{K} = \left(\frac{1}{n}k(X_i, X_j)\right)_{i,j \leq n} \in \mathbb{R}^{n \times n}, \ \hat{K}_x = \left(\frac{1}{n}k(x, X_i)\right)_{i \leq n} \in \mathbb{R}^n.$$

[1] Note that this topological assumption ease derivations but is not fundamental for such non-asymptotic results.

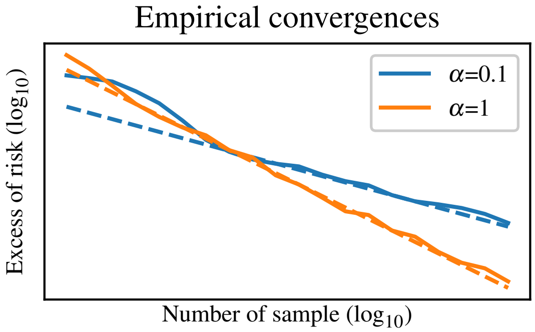

Figure 3: Empirical convergence rates. We consider binary classification, with $\mathcal{X} = [-1, 1]$, $g^*(x) = \operatorname{sign}(x) * |x|^{\frac{1}{\alpha}}$, for $\alpha \in \{.1, 1\}$ and $\rho_{\mathcal{X}}$ uniform. We plot in solid the logarithm of the excess risk averaged over 100 trials against the logarithm of the number of samples for $n \in [10, 10^6]$, and plot in dashed the expected slope of those curves due to Theorem 11 (*i.e.*, we fit the constant C in the rate $Cn^{-\gamma}$ with γ obtained from the bound in Eq. (10)).

To state regularity assumptions, we introduce a minimal setup linked to the reproducing kernel k. To keep the exposition clear, we relegate technicalities in Appendix C. We define the operator K operating on functions $f \in L^2(\mathcal{X}, \mathcal{H}, \rho_{\mathcal{X}})$ and $K_{\mathcal{X}}$ operating on $f \in L^2(\mathcal{X}, \mathbb{R}, \rho_{\mathcal{X}})$, both defined as

$$(Kf)(x') = \int_{\mathcal{X}} k(x', x) f(x) \, \mathrm{d}\rho_{\mathcal{X}}(x).$$

Inheriting from the symmetry and positive semi-definiteness of k, K is self-adjoint with spectrum in $\mathbb{R}_+$. To study the convergence of g_n to g^*, it is useful to introduce the approximate orthogonal projection on $\operatorname{im} K^{\frac{1}{2}}$, defined for $\lambda > 0$ as

$$g_\lambda = K(K + \lambda)^{-1} g^*.$$

We introduce three assumptions linked with the regularity of the problem, referred to as the capacity condition, interpolation inequality and source condition. Those are classical assumption to prove uniform rates of the kernel ridge regression estimates. They could be found, in particular, by Fischer and Steinwart (2020) under the respective names of (EVD), (EMB) and (SRC), but also by Pillaud-Vivien et al. (2018); Lin et al. (2020). Our assumptions differ in that they are expressed for vector-valued functions, which usually generate compactness issues (Caponnetto and De Vito, 2006). However, when $\mathcal{Z}$ is finite, $\mathcal{H}$ is finite dimensional, and K can be shown to be a compact operator, thus allowing to consider fractional power without definition issues.

Assumption 7 (Capacity condition) *Suppose* $\operatorname{Tr}(K_{\mathcal{X}}^{\sigma}) < +\infty$ *for* $\sigma \in [0, 1]$.

Assumption 8 (Interpolation inequality) *Assume the existence of* $p \in [0, \frac{1}{2}]$, $c_p > 0$ *such that*

$$\forall\, g \in (\ker K)^{\perp}, \qquad \|K^p g\|_{L^\infty} \leq c_p \|g\|_{L^2}.$$

Assumption 9 (Source condition) *Suppose* $g^* \in \operatorname{im} K^q$ *for* $q \in (p, 1]$.

When $q = 1/2$, the source condition is often expressed as g^* belonging to the reproducing kernel Hilbert space associated to the kernel k. Note that when k is bounded, Assumptions 7 and 8 hold with $\sigma = 1$ and $p = 1/2$. In those assumptions, for p and σ the smaller, and for q the bigger, the faster the convergence rates will be.

Example 6 (Classical assumptions) *For Assumption 8 to hold, minimal mass and regular support of ρ, similarly to Example 5, are often assumed, as well as regularity of functions in $\operatorname{im} K^p$, in coherence with Remark 8. For Assumption 9 to hold, it is classical to assume regularity of g^*, matching the regularity of function spaces derived from the kernel k. The value of σ in Assumption 7 often comes has a bonus of regularity assumptions on ρ and specificity of the RKHS implied by k. See Example 2 by Pillaud-Vivien et al. (2018) and Section 4 by Fischer and Steinwart (2020) as well as references therein for concrete examples.*

We now state convergence results respectively proven in Appendices C.5 and C.6, C.7, and C.8. Lemma 13 is a generalization to vector-valued functions of kernel ridge regression uniform convergence rates known for real-valued function (see Fischer and Steinwart, 2020). Note that a similar result to Theorem 14 was provided for binary classification by Koltchinskii and Beznosova (2005), but we generalize exponential rates with kernel ridge regression to any discrete structured prediction problem. Theorem 15 is new, even in the context of binary classification. It states that, while, up to now, only rates in $n^{-1/4}$ were known for f_n (Ciliberto et al., 2020), one can indeed hope for arbitrarily fast rates, depending on the hardness of the problem, read in the value of $\alpha \in [0, \infty)$.

Lemma 13 (Reproducing kernel concentration) *Under Assumptions 7, 8 and 9, for any $\lambda > 0$,*

$$\|g_\lambda - g^*\|_{L^\infty} \leq b_1 \lambda^{q-p}.$$

With $b_1 = c_p \|K^{-q} g^*\|_{L^2}$. Moreover, when the kernel k is bounded and under Assumption 5, there exists three constants $b_2, b_3, b_4, b_5 > 0$ that does not depend nor on λ nor on n such that

$$\mathbb{P}\left(\|g_n - g_\lambda\|_\infty > t\right) \leq b_2 \lambda^{-\sigma} \exp\left(-b_3 n \lambda^{2p}\right) + 4 \exp\left(-\frac{n\lambda^{2p+\sigma} t^2}{b_4 + b_5 t}\right).$$

As long as $b_3 n \geq \lambda^{-p}$, and $\lambda \leq \min\left(\|K\|_{\text{op}}, 1\right)$.

Theorem 14 (Kernel ridge regression fast rates under no-density assumption) *When the loss ℓ is bounded, satisfies Assumption 1 and $\mathcal{Z}$ is finite, under the t_0-no-density separation condition, and Assumptions 5, when k is bounded, if $\lambda_n = \lambda$, for any $\lambda > 0$ such that $\|g^* - g_\lambda\|_{L^\infty} < t_0$, then there exist two constants $b_6, b_7 > 0$ such that, for any $n \in \mathbb{N}^*$,*

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq b_6 \exp(-b_7 n), \tag{11}$$

with f_n given by the kernel ridge regression surrogate estimate.

Theorem 15 (Kernel ridge regression fast rates under low-density assumption) *When ℓ satisfies Assumption 1, is bounded and $\mathcal{Z}$ is finite, under the α-low-density separation condition, and Assumptions 5, 7, 8 and 9, if $\lambda_n = \lambda_0 n^{-\frac{1}{2q+\sigma}}$, for any $\lambda_0 > 0$, there exists $b_8 > 0$, such that for any $n \in \mathbb{N}^*$,*

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq b_8 n^{-\frac{(q-p)(1+\alpha)}{2q+\sigma}}. \tag{12}$$

6. Conclusion

In this paper, we have shown how, for discrete problems, to leverage exponential concentration inequalities derived on continuous surrogate problems, in order to derive faster rates than rates directly obtained through calibration inequalities. Those rates are arbitrarily fast, depending on a parameter characterizing the hardness of the discrete problem. We have shown how this method directly applies to local averaging methods and to kernel ridge regression, which allowed us to derive "super fast" rates for any discrete structured prediction problem.

This opens the way to several follow-up, such as

- Applicative follow-up, consisting of tackling concrete problem instances, such as predicting properties of DNA-sequence (Jaakkola et al., 2000), *e.g.*, gene mutations responsible for diseases, with well-designed kernels on DNA in order to higher the exponent appearing in Theorem 15.
- Computational follow-up, pushing our analysis further to understand how to design better algorithms on discrete problems. For example, by adding a regularization pushing g_n away from the decision frontier F, and adding a term in $\mathbf{1}_{\|g_n(x)-g^*(x)\|>d(g_n(x),F)}$ in Eq. (5) for the analysis.
- Theoretical follow-up, to widen our analysis to other types of smooth surrogates, and to parametric methods, such as deep learning models, assuming that functions are parameterized by a parameter θ, that some analysis gives concentration on $\theta_n - \theta^*$ similar to Eq. (4) and that calibration inequalities relate the error on θ with the error between $f_n = f_{\theta_n}$ and $f^* = f_{\theta^*}$.

Acknowledgments

We would like to thanks Alex Nowak for useful discussions, as well as anonymous referees for helpful comments. This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). We also acknowledge support from the European Research Council (grants SEQUOIA 724063 and REAL 947908).

References

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 1950.

Jean-Yves Audibert and Alexander Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 2007.

Peter Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 2006.

Peter Bartlett, Michael Jordan, and Jon McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 2006.

Serguei Bernstein. Sur une modification de l'inéqualite de Tchebychev. *Annals Science Institute Sav. Ukraine*, 1924.

Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer International Publishing, 2015.

Mathieu Blondel, André Martins, and Vlad Niculae. Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 2020.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classificatoin: A survey of some recent advances. *ESAIM: Probability and Statistic*, 2005.

Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. *Classification and Regression Trees*. Chapman & Hall, 1984.

Vivien Cabannes, Alessandro Rudi, and Francis Bach. Structured prediction with partial labelling through the infimum loss. In *International Conference on Machine Learning*, 2020.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 2006.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In *Neural Information Processing Systems*, 2014.

George Chen and Devavrat Shah. *Explaining the Success of Nearest Neighbor Methods in Prediction*. Foundations and Trends in Machine Learning, 2018.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. In *Neural Information Processing Systems*, 2016.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A general framework for consistent structured prediction with implicit loss embeddings. *Journal of Machine Learning Research*, 2020.

William Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 1979.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 1967.

Luc Devroye and Terry Wagner. Distribution-free consistency results in nonparametric discrimination and regression function estimation. *The Annals of Statistics*, 1980.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 2020.

Evelyn Fix and Joseph Hodges. Discriminatory analysis. Nonparametric discrimination: Consistency properties. Technical report, School of Aviation Medicine, Randolph Field, Texas, 1951.

Jerome Friedman. Flexible metric nearest neighbor classification. Technical report, Department of Statistics, Stanford University, 1994.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag New York, 2002.

Tommi Jaakkola, Mark Diekhans, and David Haussle. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 2000.

Vladimir Koltchinskii and Olexandra Beznosova. Exponential convergence rates in classification. In *International Conference on Computational Learning Theory*, 2005.

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Applied and Computational Harmonic Analysis*, 2020.

Enno Mammen and Alexander Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 1999.

Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In *Conference on Learning Theory*, 2019.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 2006.

James Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society*, 1909.

Stanislav Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 2017.

Èlizbar Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 1964.

Alex Nowak-Vila, Francis Bach, and Alessandro Rudi. Sharp analysis of learning with discrete losses. In *Artificial Intelligence and Statistics*, 2019.

Alex Nowak-Vila, Francis Bach, and Alessandro Rudi. A general theory for structured prediction with smooth convex surrogates. In *arXiv*, 2020.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 1962.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Neural Information Processing Systems*, 2018.

Iosif Pinelis and Aleksandr Sakhanenko. Remarks on inequalities for large deviation probabilities. *Theory of Probability and Its Applications*, 1986.

Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 1956.

Bernhard Scholkopf and Alexander J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press, 2001.

Laurent Schwartz. Sous-espaces Hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *Journal d'Analyse Mathématique*, 1964.

Eric Slud. Distribution inequalities for the binomial law. *Annals of Probability*, 1977.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 2007.

Clifford Spiegelman and Jerome Sacks. Consistent window estimation in nonparametric regression. *The Annals of Statistics*, 1980.

Ingo Steinwart and Clint Scovel. Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, 2007.

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 2012.

Charles Stone. Consistent nonparametric regression. *The Annals of Statistics*, 1977.

Charles Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 1980.

Alexander Tsybakov. *Introduction to Nonparametric Estimation*. Springer-Verlag New York, 2009.

Tim van Erven, Peter Grünwald, Nishant Mehta, Mark Reid, and Robert Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 2015.

Geoffrey Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics*, 1962.

Yuhong Yang. Minimax nonparametric classification. I. Rates of convergence. *IEEE Transactions on Information Theory*, 1999.

Vadim Vladimirovich Yurinskii. On an infinite-dimensional version of S. N. Bernstein's inequalities. *Theory of Probability and Its Applications*, 1970.

Appendix A. Fast rates

In the following, we consider $\mathcal{X}$ and $\mathcal{Y}$ to be Polish spaces, *i.e.*, separable completely metrizable topological spaces, in order to define the distribution ρ. We also consider $\mathcal{Z}$ endowed with a topology that makes it compact, and that makes $z \to \mathbb{E}_{Y\sim\mu}\,\ell(z, Y)$ continuous for any $\mu \in \Delta_{\mathcal{Y}}$, in order to have minimizer well defined. For a Polish space $\mathcal{A}$, we denote by $\Delta_{\mathcal{A}}$ the simplex formed by the set of Borel probability measures on this space. For $\rho \in \Delta_{\mathcal{X}\times\mathcal{Y}}$, we denote by $\rho|_x$ the conditional distribution of Y given x, and by $\rho_{\mathcal{X}}$ the marginal distribution over $\mathcal{X}$. We suppose $\mathcal{H}$ separable Hilbert and that the mapping φ is measurable in order to define the pushforward measure $\varphi_*\rho|_x$. We assume that, for $\rho_{\mathcal{X}}$-almost every x, $(\varphi(Y)|X = x)$ has a second moment, in order to consider the conditional mean $g^*(x)$ as the solution of the well defined problem consisting of minimizing $\|\xi - \varphi(Y)\|^2$ for $\xi \in \mathcal{H}$. We consider ψ to be continuous, in order to have the decoding problem well posed.

A.1. Proof of Lemma 1

With the notation of Lemma 1, for $x \in \operatorname{supp}\rho_{\mathcal{X}}$

$$\begin{aligned}
\mathbb{E}_{Y\sim\rho|_x}\left[\ell(f_n(x), Y) - \ell(f^*(x), Y)\right] &= \langle \psi(f_n(x)) - \psi(f^*(x)), g^*(x)\rangle_{\mathcal{H}} \\
&= \langle \psi(f_n(x)), g_n(x)\rangle + \langle \psi(f_n(x)), g^*(x) - g_n(x)\rangle - \langle \psi(f^*(x)), g^*(x)\rangle \\
&\leq \langle \psi(f^*(x)), g_n(x)\rangle + \langle \psi(f_n(x)), g^*(x) - g_n(x)\rangle - \langle \psi(f^*(x)), g^*(x)\rangle \\
&= \langle \psi(f_n(x)) - \psi(f^*(x)), g^*(x) - g_n(x)\rangle \\
&\leq \|\psi(f_n(x)) - \psi(f^*(x))\|_{\mathcal{H}}\,\|g^*(x) - g_n(x)\|_{\mathcal{H}} \\
&\leq 2c_\psi\,\|g^*(x) - g_n(x)\|_{\mathcal{H}}\,,
\end{aligned}$$

where the inequality $\langle \psi(f_n(x)), g_n(x)\rangle \leq \langle \psi(f^*(x)), g_n(x)\rangle$ is due to the fact that $f_n(x)$ minimizes the functional $z \to \langle \psi(z), g_n(x)\rangle$. Integrating over $\mathcal{X}$ leads to the results in Lemma 1.

A.2. Proof of Lemma 2

The first part of the lemma is a geometrical result stating that to go from two elements ξ_1 and ξ_2 in $\Delta_{\varphi(\mathcal{Y})}$, leading to two different decoding, one has to pass by a point $\xi_{1/2} \in F$, where there is at least two possible decodings. Let make it clearer. Consider $x \in \operatorname{supp}\rho_{\mathcal{X}}$ and suppose that $f_n(x) \neq f^*(x)$, define the path

$$\begin{array}{rccc}
\zeta: & [0,1] & \to & \Delta_{\varphi(\mathcal{Y})} \\
 & \lambda & \to & \lambda g_n(x) + (1-\lambda) g^*(x).
\end{array}$$

Consider $d : \Delta_{\varphi(\mathcal{Y})} \to \mathcal{Z}$ the decoding function used to retrieve f^* and f_n, from g^* and g_n, satisfying $d(\xi) \in \arg\min_{z\in\mathcal{Z}} \langle \psi(z), \xi\rangle$. Consider the path $d \circ \zeta : [0,1] \to \mathcal{Z}$, it goes from $\zeta(0) = f^*(x)$ to $\zeta(1) = f_n(x)$. Consider λ_∞ the supremum of $(d\circ\zeta)^{-1}(f^*(x))$. We will show that $\zeta(\lambda_\infty) \in F$, this will lead to

$$\|g_n(x) - g^*(x)\| = \|g_n(x) - \zeta(\lambda_\infty)\| + \|\zeta(\lambda_\infty) - g^*(x)\| \geq \|\zeta(\lambda_\infty) - g^*(x)\| \geq d(g^*(x), F),$$

and to Lemma 2 by contraposition.

To show that $\zeta(\lambda_\infty) \in F$, we will show that $f^*(x) \in \arg\min_z \langle \psi(z), \zeta(\lambda_\infty) \rangle \not\subset \{f^*(x)\}$. By definition of the supremum, there exists a sequence $(\lambda_p)_{p\in\mathbb{N}}$ converging to λ_∞ such that

$$f^*(x) \in \arg\min_z \langle \psi(z), \lambda_p g_n(x) + (1-\lambda_p)g^*(x) \rangle \,,$$

meaning that for all $z \neq f^*(x)$

$$\langle \psi(f^*(x)), \lambda_p g_n(x) + (1-\lambda_p)g^*(x) \rangle \leq \langle \psi(z), \lambda_p g_n(x) + (1-\lambda_p)g^*(x) \rangle \,.$$

By continuity of the scalar product, it means that it holds for $p = \infty$, which means $f^*(x) \in \arg\min_z \langle \psi(z), \zeta(\lambda_\infty) \rangle$. Now, suppose that $\arg\min_z \langle \psi(z), \zeta(\lambda_\infty) \rangle = \{f^*(x)\}$, this means that for all $z \neq f^*(x)$,

$$\langle \psi(f^*(x)), \lambda_\infty g_n(x) + (1-\lambda_\infty)g^*(x) \rangle < \langle \psi(z), \lambda_\infty g_n(x) + (1-\lambda_\infty)g^*(x) \rangle \,.$$

By continuity of this function accordingly to λ, this means that this still holds for $\lambda_\infty + \varepsilon_z$ for $\varepsilon_z > 0$. Taking $\varepsilon = \inf_{z\in\mathcal{Z}} \varepsilon_z$, it means that $\lambda_\infty + \varepsilon \in (d \circ \zeta)^{-1}(f^*(x))$. When $\mathcal{Z}$ is finite, $\varepsilon > 0$, which contradict the definition of λ_∞. Therefore $\zeta(\lambda_\infty) \in F$.

The second part of Lemma 2 follows from derivations in Appendix A.1.

Remark 16 (Extension to discrete cases) *Note that the same argument can be generalized to discrete problems – which could be defined as $\mathcal{Z}$ endowed with a topology that makes $z \to \mathbb{E}_{Y\sim\mu}[\ell(z,Y)]$ continuous with respect to z, and $\mathcal{Z}\setminus\{z\}$ locally compact for any $z \in \mathcal{Z}$ – that are not degenerate, in the sense that $\rho_{\mathcal{X}}$ almost all $x \in \mathcal{X}$, there exists $t > 0$ such that the cardinality of the set defined as $\left\{z \,\middle|\, \mathbb{E}_{Y\sim\rho|_x}[\ell(z,Y)] - \inf_{z'\in\mathcal{Y}} \mathbb{E}_{Y\sim\rho|_x}[\ell(z',Y)] < t\right\}$ if finite. This holds for classification with infinite countable classes, but it does not for regression on the set of rational numbers.*

Remark 17 (Extension to general cases) *To remove the condition $\mathcal{Z}$ finite, one can change the definition of $d(g^*(x), F)$ to $\inf_{\xi\in\mathcal{H};\{f^*(x)\}\neq\arg\min\langle\psi(z),\xi\rangle} \|\xi - g^*(x)\|$, in order to make Lemma 2 hold for any $\mathcal{Z}$.*

A.3. Equivalence between generalizations of the Tsybakov margin condition

While we state the margin condition with $d(g^*(x), F)$, it could also be stated with $d(g^*(x), F \cap \mathrm{Conv}(\varphi(\mathcal{Y})))$ or with, which is the quantity considered by (Nowak-Vila et al., 2019),

$$\gamma(x) = \inf_{z\neq z^*} \mathbb{E}_{Y\sim\rho|_x} \ell(z,Y) - \mathbb{E}_{Y\sim\rho|_x} \ell(z^*,Y) = \inf_{z\neq z^*} \langle \psi(z) - \psi(z^*), g^*(x) \rangle \,.$$

Indeed, when $\mathcal{Z}$ is finite and ℓ is proper in the sense that $\ell(\cdot,y) = \ell(\cdot,z)$ implies $z = y$, and that there is no z that minimizes a linear combination of $(\ell(\cdot,y))_{y\in\mathcal{Y}}$ without minimizing a convex combination of the same family, we have the existence of two constants such that

$$c\gamma(x) \leq d(g^*(x), F \cap \mathrm{Conv}(\varphi(\mathcal{Y}))) \leq d(g^*(x), F) \leq c'\gamma(x).$$

A.3.1. Mildness of our condition

Let z' be the argmin defining γ, geometric properties of the scalar product imply the existence of a $\xi \in (\varphi(z') - \varphi(z^*))^\perp$ such that

$$\left\langle \varphi(z') - \varphi(z^*), g^*(x) \right\rangle = \left\| \varphi(z') - \varphi(z^*) \right\| \left\| g^*(x) - \xi \right\|.$$

Therefore

$$\left\langle \varphi(z') - \varphi(z^*), g^*(x) \right\rangle \geq \min_{y,y'} \left\| \varphi(y) - \varphi(y') \right\| \left\| g^*(x) - \xi \right\|.$$

Note that, by definition of ξ, $\langle \xi, \varphi(z') \rangle = \langle \xi, \varphi(z^*) \rangle$. If $\xi \in R_{z^*}$ then $\xi \in F$, otherwise $\xi \notin R_{z^*}$ and then, there exists a point between ξ and $g^*(x)$ that belongs to the decision frontier (see Appendix A.2 for a proof - for which we need some regularity assumption such as $\mathcal{Z}$ finite). In every case,

$$\left\| g^*(x) - \xi \right\| \geq d(g^*(x), F).$$

This implies the existence of c'.

A.3.2. Strength of our condition

For any g_n such that $f_n(x) = z$, we have

$$\begin{aligned} \langle \psi(z) - \psi(z^*), g^*(x) \rangle &= \langle \psi(z), g^*(x) - g_n(x) \rangle + \langle \psi(z), g_n(x) \rangle - \langle \psi(z^*), g^*(x) \rangle \\ &\leq \langle \psi(z), g^*(x) - g_n(x) \rangle + \langle \psi(z^*), g_n(x) \rangle - \langle \psi(z^*), g^*(x) \rangle \\ &\leq 2c_\psi \left\| g^*(x) - g_n(x) \right\|. \end{aligned}$$

If we take the infimum on both sides we have

$$d(g^*(x), F) = \inf_{g_n(x) \notin R_{f^*(x)}} \left\| g_n(x) - g^*(x) \right\| \geq \frac{1}{2c_\psi} \inf_{z \neq z^*} \langle \psi(z) - \psi(z^*), g^*(x) \rangle,$$

where the left equality is provided, when $\mathcal{Z}$ is finite, by a similar reasoning to the one in Appendix A.2. This implies the existence of c. Note also that if the loss is proper in the sense that if z minimizes $\langle \psi(z), \xi \rangle$ for a $\xi \in \mathcal{H}$, there exists a $\xi \in \operatorname{Conv} \varphi(\mathcal{Y})$ such that z minimizes $\langle \psi(z), \xi \rangle$, we can consider $g_n(x) \in \operatorname{Conv}(\varphi(\mathcal{Y}))$, and therefore restrict F to $F \cap \operatorname{Conv} \varphi(\mathcal{Y})$. Finally we have shown that, when $\mathcal{Z}$ finite and ℓ proper

$$c\gamma(x) \leq d(g^*(x), F \cap \operatorname{Conv}(\varphi(\mathcal{Y}))) \leq d(g^*(x), F) \leq c'\gamma(x).$$

This explains why we would consider $\gamma(x)$, $d(g^*(x), F \cap \operatorname{Conv}(\varphi(\mathcal{Y})))$ or $d(g^*(x), F)$ to define the margin condition, it will only change the value of constants in Assumptions 3 and 4.

A.4. Refinement of Theorem 5

It is possible to refine Theorem 5 to remove the condition that the loss ℓ is bounded. In the following, we omit the dependency of L_n and M_n to n.

Lemma 18 (Refinement of Theorem 5) *Under refined calibration (5), concentration, Assumption 2, and no-density separation, Assumption 3, the risk is controlled as*

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq 4c_\psi L^{-1/2} \exp\left(-\frac{t_0^2 L}{2}\right)^{1/2} + 4c_\psi M L^{-1} \exp\left(-\frac{t_0 L}{2M}\right).$$

Note that it is not possible to derive a better bound only given Eqs. (4), (5) and (6). Yet when ℓ is bounded by ℓ_∞, we have

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \ell_\infty \exp\left(-\frac{Lt_0^2}{1 + Mt_0}\right).$$

Proof Using the calibration inequality along with the no-density separation one, we get

$$\begin{aligned}\mathcal{R}(f_n) - \mathcal{R}(f^*) &\leq 2c_\psi \mathbb{E}_X \left[\mathbf{1}_{\|g_n(X) - g^*(X)\| \geq t_0} \|g_n(X) - g^*(X)\|\right] \\ &= 2c_\psi \int_{t_0}^\infty \mathbb{P}_X \left(\|g_n(X) - g^*(X)\| \geq t\right) \mathrm{d}t.\end{aligned}$$

Taking the expectation over $\mathcal{D}_n$ and using concentration inequality we have

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) &\leq 2c_\psi \int_{t_0}^\infty \mathbb{P}_{X,\mathcal{D}_n} \left(\|g_n(X) - g^*(X)\| \geq t\right) \mathrm{d}t \\ &\leq 2c_\psi \int_{t_0}^\infty \exp\left(-\frac{Lt^2}{1 + Mt}\right) \mathrm{d}t.\end{aligned}$$

We only need to study the integral $\int_{t_0}^\infty \exp\left(-\frac{Lt^2}{1+Mt}\right) \mathrm{d}t$. We first clean the dependency on t insider the exponential using that

$$\frac{1}{2}\left(\exp(-Lt^2) + \exp\left(-\frac{Lt}{M}\right)\right) \leq \exp\left(-\frac{Lt^2}{1 + Mt}\right) \leq \exp\left(-\frac{Lt^2}{2}\right) + \exp\left(-\frac{Lt}{2M}\right).$$

We are left with the study of $\int_{t_0}^\infty \exp(-At^p) \mathrm{d}t$, for $p \in \{1, 2\}$ and $A > 0$. The case $p = 1$, directly leads to $A^{-1} \exp(-At_0)$, explaining the part in L/M. The case $p = 2$ is similar to the Gaussian integral, and can be handle with the following tricks

$$\begin{aligned}\int_{t_0}^\infty \exp(-At^2) \mathrm{d}t &= \frac{1}{2} \int_{(-\infty,-t_0]\cup[t_0,\infty)} \exp(-At^2) \mathrm{d}t \\ &= \frac{1}{2}\left(\int_{((-\infty,-t_0]\cup[t_0,\infty))^2} \exp(-A\|x\|^2)) \mathrm{d}x\right)^{1/2}.\end{aligned}$$

This last integral corresponds to integrate the function $x \to \exp(-A\|x\|^2)$ for $x \in \mathbb{R}^2$ on the domain $((-\infty, -t_0] \cup [t_0, \infty))^2$. This function being positive and the domain being included in the domain $\{\|x\| \geq t_0\}$ and containing the domain $\{\|x\| \geq \sqrt{2}t_0\}$, we get

$$\int_{\{\|x\| \geq \sqrt{2}t_0\}}^\infty \exp(-A\|x\|^2) \mathrm{d}x \leq \left(2\int_{t_0}^\infty \exp(-At^2) \mathrm{d}t\right)^2 \leq \int_{\{\|x\| \geq t_0\}}^\infty \exp(-A\|x\|^2) \mathrm{d}x.$$

Using polar coordinate we get

$$\int_{\{\|x\|\geq t_0\}}^{\infty} \exp(-A\|x\|^2)\,\mathrm{d}x = 2\pi\int_{t_0}^{\infty} r\exp(-Ar^2)\,\mathrm{d}r = \pi A^{-1}\exp(-At_0^2).$$

Therefore

$$2^{-1}\sqrt{\pi}A^{-1/2}\exp(-A2t_0^2)^{1/2} \leq \int_{t_0}^{\infty}\exp(-At^2)\,\mathrm{d}t \leq 2^{-1}\sqrt{\pi}A^{-1/2}\exp(-At_0^2)^{1/2}.$$

This explain the rates in L. ■

A.5. Proof of Theorem 6

Using the calibration and Bernstein inequalities we get, omitting the dependency of L_n and M_n to n,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}\mathcal{R}(f_n) - \mathcal{R}(f^*) &\leq 2c_\psi\,\mathbb{E}_{\mathcal{D}_n,X}\left[\mathbf{1}_{d(g_n(X),g^*(X))\geq d(g^*(X),F)}\|g_n(X)-g^*(X)\|\right] \\
&= 2c_\psi\int_0^\infty \mathbb{P}_{\mathcal{D}_n,X}\left(\mathbf{1}_{d(g_n(X),g^*(X))\geq d(g^*(X),F)}\|g_n(X)-g^*(X)\|\geq t\right)\mathrm{d}t \\
&= 2c_\psi\int_0^\infty \mathbb{E}_X\,\mathbb{P}_{\mathcal{D}_n}\left(\|g_n(X)-g^*(X)\|\geq \max\{t,d(g^*(X),F)\}\right)\mathrm{d}t \\
&\leq 2c_\psi\int_0^\infty \mathbb{E}_X\exp\left(-\frac{L\max\{t,d(g^*(X),F)\}^2}{1+M\max\{t,d(g^*(X),F)\}^2}\right)\mathrm{d}t \\
&= 2c_\psi\int_0^\infty \mathbb{E}_X\left[\mathbf{1}_{d(g^*(X),F)<t}\exp\left(-\frac{Lt^2}{1+Mt}\right)\right]\mathrm{d}t \\
&\qquad + 2c_\psi\int_0^\infty \mathbb{E}_X\left[\mathbf{1}_{d(g^*(X),F)\geq t}\exp\left(-L\frac{d(g^*(X),F)^2}{1+Md(g^*(X),F)}\right)\right]\mathrm{d}t \\
&= 2c_\psi\int_0^\infty \mathbb{P}_X\left(d(g^*(X),F)<t\right)\exp\left(-\frac{Lt^2}{1+Mt}\right)\mathrm{d}t \\
&\qquad + 2c_\psi\,\mathbb{E}_X\left[d(g^*(X),F)\exp\left(-\frac{Ld(g^*(X),F)^2}{1+Md(g^*(X),F)}\right)\right].
\end{aligned}$$

Let begin by working on the first term. We have, using the low-density separation hypothesis

$$\int_0^\infty \mathbb{P}_X\left(d(g^*(X),F)<t\right)\exp\left(-\frac{Lt^2}{1+Mt}\right)\mathrm{d}t \leq c_\alpha\int_0^\infty t^\alpha\exp(-\frac{Lt^2}{1+Mt})\,\mathrm{d}t.$$

Recall the expression of the Gamma integral

$$\int_0^\infty t^\alpha\exp(-Lt)\,\mathrm{d}t = \frac{\Gamma(\alpha+1)}{L^{\alpha+1}} \qquad \text{and} \qquad \int_0^\infty t^\alpha\exp(-Lt^2)\,\mathrm{d}t = \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{2L^{\frac{\alpha+1}{2}}}.$$

Let briefly talk about optimality. Up to now, we have only used three inequality: calibration, concentration exponential inequality and low-density separation. Therefore, when those inequalities hold as equalities, we get an lower bound of order on the excess of risk as

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n}\mathcal{R}(f_n)-\mathcal{R}(f^*) &\geq 2c_\psi c_\alpha\int_0^\infty t^\alpha\exp(-\frac{Lt^2}{1+Mt})\,\mathrm{d}t\\ &\geq 2c_\psi c_\alpha\int_0^\infty \frac{1}{2}t^\alpha\left(\exp(-Lt^2)+\exp\left(-\frac{Lt}{M}\right)\right)\mathrm{d}t\\ &= 2c_\psi c_\alpha\left(\frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{4}L^{-\frac{\alpha+1}{2}}+\frac{\Gamma(\alpha+1)}{2}M^{\alpha+1}L^{-(\alpha+1)}\right).\end{aligned}$$

For the upper bound, using that $\exp(-a/1+b)\leq\exp(-a/2)+\exp(-a/2b)$, we get

$$\begin{aligned}\int_0^\infty t^\alpha\exp(-\frac{Lt^2}{1+Mt})\,\mathrm{d}t &\leq \int_0^\infty t^\alpha\exp(-\frac{Lt^2}{2})\,\mathrm{d}t+\int_0^\infty t^\alpha\exp(-\frac{Lt}{2M})\,\mathrm{d}t\\ &= 2^{\frac{\alpha-1}{2}}\Gamma\left(\frac{\alpha+1}{2}\right)L^{-\frac{\alpha+1}{2}}+2^{\alpha+1}\Gamma(\alpha+1)M^{\alpha+1}L^{-(\alpha+1)}.\end{aligned}$$

Let study the second term in the excess of risk inequality. To enhance readability, write $\eta(X)=d(g^*(X),F)$. We will first dissociate the two parts in the exponential with

$$\mathbb{E}_X\left[\eta(X)\exp\left(-\frac{L\eta(X)^2}{1+M\eta(X)}\right)\right]\leq\mathbb{E}_X\left[\eta(X)\left(\exp\left(-\frac{L\eta(X)^2}{2}\right)+\exp\left(-\frac{L\eta(X)}{2M}\right)\right).\right]$$

We are left with studying $\mathbb{E}[\eta(X)\exp(-A\eta(X)^p)]$, for $A>0$ and $p\in\{1,2\}$. The function $t\to\ t\exp(-At^p)$ achieves its maximum in $t_0=(pA)^{-1/p}$, it is increasing before and decreasing after. Notice that the quantity

$$\begin{aligned}\mathbb{P}(\eta(X)<t_0)\,\mathbb{E}_X\left[\eta(X)\exp(-A\eta(X)^p)\,|\,\eta(X)<t_0\right] &\leq c_\alpha t_0^{\alpha+1}\exp(-At_0^p)\\ &= c_\alpha p^{-\frac{\alpha+1}{p}}\exp(-p^{-1/p})A^{-\frac{\alpha+1}{p}},\end{aligned}$$

is exactly of the same order as the control we had on the first term in the excess of risk decomposition. This suggests to consider the following decomposition

$$\begin{aligned}&\mathbb{E}_X\left[\eta(X)\exp(-A\eta(X)^p)\right]=\mathbb{P}(\eta(X)<t_0)\,\mathbb{E}_X\left[\eta(X)\exp(-A\eta(X)^p)\,|\,\eta(X)<t_0\right]\\ &\quad+\sum_{i=0}^\infty\mathbb{P}(2^it_0\leq\eta(X)<2^{i+1}t_0)\,\mathbb{E}_X\left[\eta(X)\exp(-A\eta(X)^p)\,\middle|\,2^it_0\leq\eta(X)<2^{i+1}t_0\right]\\ &\leq c_\alpha t_o^{\alpha+1}\exp(-At_0^p)+\sum_{i=0}^\infty c_\alpha2^\alpha(2^it_0)^{\alpha+1}\exp(-At_0^p(2^i)^p)\\ &= c_\alpha t_o^{\alpha+1}\left(\exp(-p^{-1/p})+\sum_{i=0}^\infty 2^\alpha2^{i(\alpha+1)}\exp(-p^{-1/p}2^{ip})\right).\end{aligned}$$

The convergence of the last series, ensures the existence of a constant c such that

$$\mathbb{E}_X\left[\eta(X)\exp\left(-\frac{L\eta(X)^2}{1+M\eta(X)}\right)\right]\leq c\left(\left(\frac{L}{2M}\right)^{-(\alpha+1)}+\left(\frac{L}{2}\right)^{-\frac{\alpha+1}{2}}\right).$$

Adding everything together, we get the existence of two constants c',c'', such that

$$\mathbb{E}_{\mathcal{D}_n}\mathcal{R}(f_n)-\mathcal{R}(f^*)\leq 2c_\psi c_\alpha\left(c'M^{\alpha+1}L^{-(\alpha+1)}+c''L^{-\frac{\alpha+1}{2}}\right).$$

This ends the proof by considering $c=\max(c',c'')$.

A.6. Refinement of Theorem 6

Some convergence analyses lead to exponential inequalities that are not of Bernstein type, indeed, our result still holds in those settings, as mentioned by the following lemma. In the following, we omit the dependency of L_n and M_n to n.

Lemma 19 (Refinement of Theorem 6) *Under the assumptions of Theorem 6, if the concentration is not given by Assumption 2 but given, for some positive constants $(a_i, b_i, p_i)_{i\leq m}$, by, for all $x \in \operatorname{supp} \rho_{\mathcal{X}}$ and $t > 0$,*

$$\mathbb{P}_{\mathcal{D}_n}(\|g_n(x) - g^*(x)\| > t) \leq \sum_{i=1}^{n} a_i \exp(-b_i t^{p_i}).$$

Then the excess of risk is controlled by

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq c \sum_{i=1}^{n} a_i b_i^{-\frac{\alpha+1}{p_i}},$$

for a constant c that does not depend on $(a_i, b_i)_{i\leq m}$.

Proof First of all, remark that the proof of Theorem 6 is linear in $\mathbb{P}_{\mathcal{D}_n}(\|g_n(x) - g^*(x)\| > t)$, therefore we only need to prove this lemma for (a, b, p), for which we proceed as in Theorem 6

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) &\leq 2c_\psi \, \mathbb{E}_{\mathcal{D}_n, X} \left[\mathbf{1}_{\|g_n(X) - g(X)\| \geq d(g(X), F)} \|g_n(X) - g(X)\|\right] \\
&= 2c_\psi \int_0^\infty \mathbb{P}_{\mathcal{D}_n, X}\left(\mathbf{1}_{\|g_n(X) - g(X)\| \geq d(g(X), F)} \|g_n(X) - g(X)\| \geq t\right) \mathrm{d}t \\
&= 2c_\psi \int_0^\infty \mathbb{E}_X \, \mathbb{P}_{\mathcal{D}_n}\left(\|g_n(X) - g(X)\| \geq \max\left\{t, d(g(X), F)\right\}\right) \mathrm{d}t \\
&\leq 2c_\psi a \int_0^\infty \mathbb{E}_X \exp\left(-b \max\left\{t, d(g(X), F)\right\}^p\right) \mathrm{d}t \\
&= 2c_\psi a \int_0^\infty \mathbb{E}_X \left[\mathbf{1}_{d(g(X), F) < t} \exp\left(-bt^p\right)\right] \mathrm{d}t \\
&\qquad + 2c_\psi a \int_0^\infty \mathbb{E}_X \left[\mathbf{1}_{d(g(X), F) \geq t} \exp\left(-bd(g(X), F)^p\right)\right] \mathrm{d}t \\
&= 2c_\psi a \int_0^\infty \mathbb{P}_X\left(d(g(X), F) < t\right) \exp\left(-bt^p\right) \mathrm{d}t \\
&\qquad + 2c_\psi a \, \mathbb{E}_X \left[d(g(X), F) \exp\left(-bd(g(X), F)^p\right)\right].
\end{aligned}$$

Let begin by working on the first term. We have, using the low-density separation hypothesis

$$\begin{aligned}
\int_0^\infty \mathbb{P}_X\left(d(g(X), F) < t\right) \exp\left(-bt^2\right) \mathrm{d}t &\leq c_\alpha \int_0^\infty t^\beta \exp(-bt^p) \, \mathrm{d}t. \\
&= b^{-\frac{1+\beta}{p}} c_\alpha \int_0^\infty (b^{1/p}t)^\beta \exp(-(b^{1/p}t)^p) \, \mathrm{d}(b^{1/p}t). \\
&= b^{-\frac{1+\beta}{p}} c_\alpha \int_0^\infty t^\beta \exp(-t^p) \, \mathrm{d}t = c_\alpha \Gamma(\beta, p) b^{-\frac{1+\beta}{p}}.
\end{aligned}$$

Let study the second term in the excess of risk inequality. To enhance readability, write $\eta(X) = d(g(X), F)$. We are left with studying $\mathbb{E}[\eta(X) \exp(-b\eta(X)^p)]$. The function $t \to\ t \exp(-bt^p)$ achieves it maximum in $t_0 = (pb)^{-1/p}$, it is increasing before and decreasing after. Notice that the quantity

$$\begin{aligned}\mathbb{P}(\eta(X) < t_0)\, \mathbb{E}_X\left[\eta(X) \exp(-b\eta(X)^p) \,|\, \eta(X) < t_0\right] &\leq c_\alpha t_0^{\beta+1} \exp(-bt_0^p) \\ &= c_\alpha p^{-\frac{\beta+1}{p}} \exp(-p^{-1/p}) b^{-\frac{\beta+1}{p}},\end{aligned}$$

is exactly of the same order as the control we had on the first term in the excess of risk decomposition. This suggests to consider the following decomposition

$$\begin{aligned}\mathbb{E}_X\left[\eta(X) \exp(-b\eta(X)^p)\right] &= \mathbb{P}(\eta(X) < t_0)\, \mathbb{E}_X\left[\eta(X) \exp(-b\eta(X)^p) \,|\, \eta(X) < t_0\right] \\ &\quad + \sum_{i=0}^{\infty} \mathbb{P}(2^i t_0 \leq \eta(X) < 2^{i+1} t_0)\, \mathbb{E}_X\left[\eta(X) \exp(-b\eta(X)^p) \,\middle|\, 2^i t_0 \leq \eta(X) < 2^{i+1} t_0\right] \\ &\leq c_\alpha t_o^{\beta+1} \exp(-bt_0^p) + \sum_{i=0}^{\infty} c_\alpha 2^\beta (2^i t_0)^{\beta+1} \exp(-bt_0^p (2^i)^p) \\ &= c_\alpha t_o^{\beta+1} \left(\exp(-p^{-1/p}) + \sum_{i=0}^{\infty} 2^\beta 2^{i(\beta+1)} \exp(-p^{-1/p} 2^{ip})\right).\end{aligned}$$

The convergence of the last series ensures the existence of a constant c' such that

$$\mathbb{E}_X\left[\eta(X) \exp(-b\eta(X)^p)\right] \leq c' b^{-\frac{\beta+1}{p}}.$$

Adding everything together ends the proof of this lemma. Note that we have the same type of optimality as the one stated in Theorem 6. ■

Because we use concentration inequalities for terms that are not necessarily centered, we usually get that Eq. (4) only holds for $t > \varepsilon_0$ where, typically $\varepsilon_0 = \|\mathbb{E}_{\mathcal{D}_n} g_n(x) - g^*(x)\|$, we can bypass this problem by adding in $\mathbf{1}_{t<\varepsilon_0}$ in the probability, motivating the study leading to the following lemma.

Lemma 20 (Handling bias in concentration inequality) *Under the assumptions of Theorem 6, if the concentration is not given by Assumption 2 but given, for a $\varepsilon_0 > 0$, by, for all $x \in \operatorname{supp} \rho_\mathcal{X}$ and $t > 0$,*

$$\mathbb{P}_{\mathcal{D}_n}(\|g_n(x) - g^*(x)\| > t) \leq \mathbf{1}_{t<\varepsilon_0}.$$

Then the excess of risk is controlled by

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq 2c_\psi c_\alpha \varepsilon_0^{\alpha+1}.$$

Proof We retake the beginning of the proof of Theorem 6, and change its ending with

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) &\leq 2c_\psi\, \mathbb{E}_{\mathcal{D}_n,X}\left[\mathbf{1}_{\|g_n(X)-g(X)\| \geq d(g(X),F)} \|g_n(X) - g(X)\|\right] \\ &= 2c_\psi \int_0^\infty \mathbb{P}_{\mathcal{D}_n,X}\left(\mathbf{1}_{\|g_n(X)-g(X)\| \geq d(g(X),F)} \|g_n(X) - g(X)\| \geq t\right) \mathrm{d}t \\ &= 2c_\psi \int_0^\infty \mathbb{E}_X\, \mathbb{P}_{\mathcal{D}_n}\left(\|g_n(X) - g(X)\| \geq \max\left\{t, d(g(X), F)\right\}\right) \mathrm{d}t \\ &\leq 2c_\psi \int_0^\infty \mathbb{E}_X\, \mathbf{1}_{t<\varepsilon_0} \mathbf{1}_{d(g(X),F)<\varepsilon_0}\, \mathrm{d}t = 2c_\psi \varepsilon_0\, \mathbb{P}_X\left(d(g(X), F) < \varepsilon_0\right) \mathrm{d}t.\end{aligned}$$

This leads to the result after applying the α-margin condition. ■

Appendix B. Nearest neighbors

B.1. Usual assumptions to derive nearest neighbors convergence rates

Assumption 6 can be seen as the backbone that allow to control $\|g_n^*(x) - g^*(x)\|$ in a uniform manner. This assumption that relates the regularity of g^* with the density of $\rho_{\mathcal{X}}$ has been historically approached in the following manner. Assume that g^* is β'-Hölder, that is, for any $x, x' \in \operatorname{supp} \rho_{\mathcal{X}}$

$$\left\|g^*(x) - g^*(x')\right\| \leq a_1 d(x, x')^{\beta'}.$$

Suppose that $\mathcal{X} = \mathbb{R}^d$, that $\rho_{\mathcal{X}}$ is continuous against λ, the Lebesgue measure, with minimal mass in the sense that there exists a $p_{\min} > 0$ such that $\frac{\mathrm{d}\rho_{\mathcal{X}}}{\mathrm{d}\lambda}(\mathcal{X})$ does not intersect $(0, p_{\min})$, and that $\operatorname{supp} \rho_{\mathcal{X}}$ has regular boundaries in the sense that there exist $a_2, t_0 > 0$ such that for any $x \in \operatorname{supp} \rho_{\mathcal{X}}$ and $t \in (0, t_0)$

$$\lambda\left(\mathcal{B}(x, t) \cap \operatorname{supp} \rho_{\mathcal{X}}\right) \geq a_2 \lambda\left(\mathcal{B}(x, t)\right).$$

For example an orthant satisfies this property with $a_2 = 2^{-d}$ and $t_0 = \infty$, and $\mathcal{B}(0, 1)$ satisfies this property with $a_2 = \lambda\left(\mathcal{B}(0, 1) \cap \mathcal{B}(1, 1)\right) / \lambda\left(\mathcal{B}(0, 1)\right)$ and $t_0 = 1$. In such a setting, we get

$$\left\|g^*(x) - g^*(x')\right\| \leq a_1 d(x, x')^{\beta'} = a_1 \left(\frac{\lambda(\mathcal{B}(x, d(x, x')))}{\lambda(\mathcal{B}(0, 1))}\right)^{\frac{\beta'}{d}}.$$

Where $d(x, x') < t_0$, we have

$$\lambda(\mathcal{B}(x, d(x, x'))) \leq a_2^{-1} \lambda\left(\mathcal{B}(x, d(x, x')) \cap \operatorname{supp} \rho_{\mathcal{X}}\right) \leq a_2^{-1} p_{\min}^{-1} \rho_{\mathcal{X}}(\mathcal{B}(x, d(x, x')))^{\beta}.$$

This means that for any $x \in \operatorname{supp} \rho_{\mathcal{X}}$ and $x' \in \mathcal{B}(x, t_0)$ we have, with $\beta = \frac{\beta'}{d}$ and the constant $a_3 = a_1 a_2^{-\beta} p_{\min}^{-\beta} \lambda(\mathcal{B}(0, 1))^{-\beta}$

$$\left\|g^*(x) - g^*(x')\right\| \leq a_3 \rho_{\mathcal{X}}(\mathcal{B}(x, d(x, x')))^{\beta}.$$

While, we actually do not need the bound to hold for $d(x, x') > t_0$ in the following proof, to check the veracity of our remark on Assumption 6, one can verify that under our assumptions on $\rho_{\mathcal{X}}$, $\operatorname{supp} \rho_{\mathcal{X}}$ is bounded, and therefore g^* is too. And if g^* is bounded by c_{φ}, by considering $a_3' = \max\left(2 c_{\varphi} a_2^{-\beta} p_{\min}^{-\beta} t_0^{-\beta'}, a_3\right)$, this bound holds for any $x, x' \in \operatorname{supp} \rho_{\mathcal{X}}$.

B.2. Proof of Lemma 9

Control of the variance term. For $x \in \rho_{\mathcal{X}}$, the variance term can be written

$$\|g_n(x) - g_n^*(x)\| = \left\|\frac{1}{k} \sum_{i=1}^{k} \varphi(Y_{(i)}) - \mathbb{E}\left[Y_{(i)} \,\middle|\, X_{(i)}\right]\right\|.$$

Where the index (i) is such that $X_{(i)}$ is the i-th nearest neighbor of x in $(X_i)_{i \leq n}$. Since, given $(X_i)_{i \leq n}$, the $(Y_i)_{i \neq n}$ are independent, distributed according to $\otimes_{i \leq n} \rho|_{X_i}$, we can use a concentration inequality to control it. We recall Bernstein concentration inequality in such spaces, derived by Yurinskii (1970), we will use the formulation of Corollary 1 from Pinelis and Sakhanenko (1986).

Theorem 21 (Concentration in Hilbert space (Pinelis and Sakhanenko, 1986)) *Let denote by $\mathcal{A}$ a Hilbert space and by (ξ_i) a sequence of independent random vectors on $\mathcal{A}$ such that $\mathbb{E}[\xi_i] = 0$, and that there exists $M, \sigma^2 > 0$ such that for any $m \geq 2$*

$$\sum_{i=1}^n \mathbb{E}\left[\|\xi_i\|^m\right] \leq \frac{1}{2} m! \sigma^2 M^{m-2}.$$

Then for any $t > 0$

$$\mathbb{P}(\left\|\sum_{i=1}^n \xi_i\right\| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 + 2tM}\right).$$

This explain Assumption 5, allowing, because there is only k ξ_i active in $\sum_{i=1}^n \alpha_i(x)\xi_i$, to get

$$\mathbb{P}_{\mathcal{D}_n}\left(\|g_n(x) - g_n^*(x)\| > t\right) \leq 2\exp\left(-\frac{kt^2}{2\sigma^2 + 2Mt}\right).$$

Control of the bias term. Under the Modified Lipschitz condition, Assumption 6,

$$\begin{aligned}\|g_n^*(x) - g_n(x)\| &= \left\|\sum_{i=1}^n \alpha_i(x)\left(g_n(x) - g^*(X_i)\right)\right\| \leq \sum_{i=1}^n \alpha_i(x)\left\|g_n(x) - g^*(X_i)\right\| \\ &\leq c_\beta \sum_{i=1}^n \alpha_i(x)\rho_{\mathcal{X}}\left(\mathcal{B}(x, d(x, X_i))\right)^\beta \leq c_\beta \rho_{\mathcal{X}}\left(\mathcal{B}(x, d(x, X_k(x)))\right)^\beta.\end{aligned}$$

When $\rho_{\mathcal{X}}$ is continuous, it follows from the probability integral transform (also known as universality of the uniform) that $\rho_{\mathcal{X}}\left(\mathcal{B}(x, d(x, X_k(x)))\right)$ behaves like the k-th order statistics of a sample $(U_i)_{i\leq n}$ of n uniform distributions on $[0, 1]$. Therefore, for any $s \in [0, 1]$

$$\mathbb{P}_{\mathcal{D}_n}\left(\rho_{\mathcal{X}}\left(\mathcal{B}(x, d(x, X_k(x)))\right) > s\right) = \mathbb{P}\left(\sum_{i=1}^n \mathbf{1}_{U_i<s} \leq k\right).$$

Recall the multiplicative Chernoff bound, stating that for $(Z_i)_{i\leq n}$ n independent random variables in $\{0, 1\}$, if $Z = \sum_{i=1}^n Z_i$, and $\mu = \mathbb{E}[Z]$, for any $\delta > 0$

$$\mathbb{P}\left(Z \leq (1-\delta)\mu\right) \leq \exp\left(-\frac{\delta^2\mu}{2}\right).$$

Since, for $s \in [0, 1]$, $\mathbb{E}[\mathbf{1}_{U_i<s}] = \mathbb{P}(U_i < s) = s$, we get, when $k \leq ns/2$

$$\mathbb{P}\left(\sum_{i=1}^n \mathbf{1}_{U_i<s} \leq k\right) \leq \exp\left(-\frac{(ns-k)^2}{2ns}\right) \leq \exp\left(-\frac{ns}{8}\right).$$

With $s = c_\beta^{-1} t^{\frac{1}{\beta}}$, we get

$$\mathbb{P}_{\mathcal{D}_n}\left(\|g_n^*(x) - g_n(x)\| > t\right) \leq \exp\left(-\frac{nt^{\frac{1}{\beta}}}{8c_\beta}\right).$$

Remark that when g^* is β' Hölder, we get the same result with $\rho_{\mathcal{X}}(\mathcal{B}(x, t))$ instead of $t^{\frac{1}{\beta}}$ by considering $\mathbf{1}_{X_i\in\mathcal{B}(x,t)}$ instead of $\mathbf{1}_{U_i\leq t}$. Note that there is way to bound a Binomial distribution with a Gaussian for t smaller than the mean of the binomial distribution, which would allow to get a bound that holds for any $t > 0$ (Slud, 1977).

B.3. Proof of Theorem 10

Using the proof of Theorem 5, we get

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \ell_\infty \, \mathbb{P}_{\mathcal{D}_n} \left(\|g(x) - g_n^*(x)\| > t_0\right).$$

Because $\|g_n(x) - g^*(x)\| > t_0$ implies that either $\|g_n(x) - g_n^*(x)\| > t_0/2$ or $\|g_n^*(x) - g^*(x)\| > t_0/2$, we get using Lemma 9

$$\mathbb{P}_{\mathcal{D}_n} \left(\|g(x) - g_n^*(x)\| > t_0\right) \leq 2\exp\left(-\frac{b_1 k t_0^2}{4 + 2b_2 t_0}\right) + \exp\left(-2^{-\frac{1}{\beta}} b_3 n t_0^{\frac{1}{\beta}}\right) + \mathbf{1}_{t_0 > (k/2n)^\beta}.$$

This explains the result of Theorem 10.

B.4. Proof of Theorem 11

First of all for $t > 0$, and $x \in \operatorname{supp} \rho_{\mathcal{X}}$, because $\|g_n(x) - g^*(x)\| > t$ implies that either $\|g_n(x) - g_n^*(x)\| > t/2$ or $\|g_n^*(x) - g^*(x)\| > t/2$, we have the inclusion of events:

$$\{\mathcal{D}_n \,|\, \|g_n(x) - g^*(x)\| > t\} \subset \{\mathcal{D}_n \,|\, \|g_n(x) - g^*(x)\| > t/2\} \cup \{\mathcal{D}_n \,|\, \|g_n(x) - g^*(x)\| > t/2\},$$

which translates in term of probability as

$$\begin{aligned}\mathbb{P}_{\mathcal{D}_n} \left(\|g_n(x) - g^*(x)\| > 2t\right) &\leq \mathbb{P}_{\mathcal{D}_n} \left(\|g_n(x) - g^*(x)\| > t\right) + \mathbb{P}_{\mathcal{D}_n} \left(\|g_n(x) - g^*(x)\| > t\right). \\ &\leq 2\exp\left(-\frac{b_1 k t^2}{1 + b_2 t}\right) + \exp\left(-b_3 n t^{\frac{1}{\beta}}\right) + \mathbf{1}_{t < \left(\frac{k}{2n}\right)^\beta}.\end{aligned}$$

Using the refinements of Theorem 6 exposed in Appendix A.6, we get that there exists a constant $c > 0$ that does not depend on k or n such that

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R} f_n - \mathcal{R}(f^*) \leq c\left(k^{-\frac{\alpha+1}{2}} + n^{-\beta(\alpha+1)} + (nk^{-1})^{\beta(\alpha+1)}\right).$$

We optimize this last quantity with respect to k by taking $k = n^\gamma$, and choosing γ such that $\gamma = 2(1-\gamma)\beta$ leading to $\gamma = 2\beta/(2\beta+1)$ and to rates in n to the power minus $\beta(\alpha+1)/(2\beta+1)$.

B.5. Numerical Experiments

Interestingly, on numerical simulations such as the one presented on Figure 3, we observed two regimes. A first regime where bound are meaningless because of constants being too big, and where the error decreases independently of the exponent expected through Theorem 11, and a final regime where rates corresponds to the bond given by the theorem. Note that when $\alpha >> 1$, with our computation parameter, we do not get to really illustrate convergence rates, as this final regime get place for bigger n than what we have considered ($n_{\max} = 10^6$), this being partly due to the constant c_β in Assumption 6 being big for the g^* we considered. Furthermore, note that, for example, if a problem satisfied Assumption 3 with a really small t_0, we expect that exponential convergence rates are only going to be observed for $n > N$, with N really big, and for which the excess of risk is already really small.

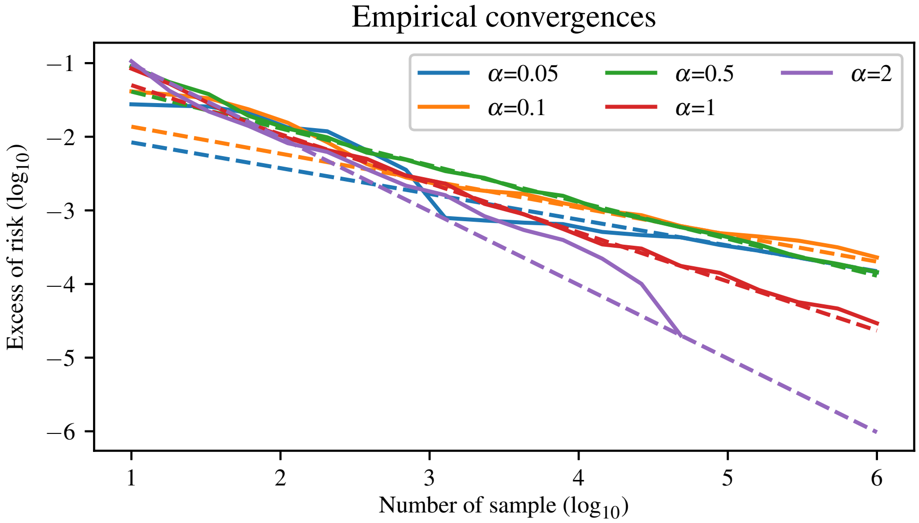

Figure 4: Supplement to Figure 3. We precise the error is evaluated on 100 points forming a regular partition of $\mathcal{X} = [-1, 1]$, and the expectation $\mathbb{E}_{\mathcal{D}_n}$ is approximated by considering 100 datasets. The violet curve is cropped at $n \approx 10^5$, because the error was null afterwards with our evaluation parameters (only 100 points to evaluate the error), forbidding us to consider the logarithm of the excess of risk.

Appendix C. Kernel proofs

In this section, we study L^∞ convergence rates of the kernel ridge regression estimate. We use the L^2-proof scheme of Caponnetto and De Vito (2006) with the remark of Ciliberto et al. (2016) to factorize the action of K on $L^2(\mathcal{X}, \mathcal{H}, \rho_{\mathcal{X}})$ through its action on $L^2(\mathcal{X}, \mathbb{R}, \rho_{\mathcal{X}})$. We retake the work of Pillaud-Vivien et al. (2018) to relax the source condition, and use Fischer and Steinwart (2020) to cast in L^∞ thanks to interpolation inequality. While those results, leading to Lemma 13, are not new, we present them entirely to provide the reader with self-contained materials.

C.1. Construction of reproducing kernel Hilbert space (RKHS)

In the following, we suppose k bounded by κ^2.

Vector-valued RKHS. To study the estimator g_n, it is useful to introduce the reproducing kernel Hilbert space $\mathcal{G}$ associated with k and $\mathcal{H}$ (Aronszajn, 1950). To define $\mathcal{G}$, define the atoms $k_x : \mathcal{H} \to \mathcal{G}$ and the scalar product, for $x, x' \in \mathcal{X}$ and $\xi, \xi' \in \mathcal{H}$ as

$$\left\langle k_x\xi, k_{x'}\xi' \right\rangle_{\mathcal{G}} = \left\langle \xi, \Gamma(x, x')\xi' \right\rangle = k(x, x') \left\langle \xi, \xi' \right\rangle_{\mathcal{H}}.$$

Where Γ is the vector valued kernel inherit from k as $\Gamma(x, x') = k(x, x')I_{\mathcal{H}}$ (Schwartz, 1964). $\mathcal{G}$ is defined as the closure, under the metric induced by this scalar product, of the span of the atoms $k_x\xi$ for $x \in \mathcal{X}$ and $\xi \in \mathcal{H}$. Note that k_x is linear, and continuous of norm $\|k_x\|_{\text{op}} = \sqrt{k(x, x)}$. When $k(\cdot, x)$ is square integrable for all $x \in \operatorname{supp} \rho_{\mathcal{X}}$, $\mathcal{G}$ is homomorphic to a functional space in $L^2(\mathcal{X}, \mathcal{H}, \rho_{\mathcal{X}})$ through the linear mapping S that associates the atom $k_x\xi$ in $\mathcal{G}$ to the function $k(\cdot, x)\xi$

in L^2, defined formally as

$$\begin{array}{rccc} S: & \mathcal{G} & \to & L^2 \\ & \gamma & \to & x \to k_x^\star \gamma. \end{array}$$

While intrinsically similar, it is useful to distinguish between $\mathcal{G}$ and $\operatorname{im} S \subset L^2$. Note that S is continuous, since on atom $k_x\xi$, $\|Sk_x\xi\|_{L^2} \leq \|k_x(\cdot)\|_{L^2} \|\xi\|_{\mathcal{H}} \leq k(x,x) \|\xi\|_{\mathcal{H}} = \|k_x\xi\|_{\mathcal{G}}$. The fact that S is a bounded operator justifies the introduction of the following operators.

Central operators. In the following, we will make an extensive use of $S^\star : L^2(\mathcal{X}, \mathcal{H}, \rho_\mathcal{X}) \to \mathcal{G}$ the adjoint of S, defined as $S^\star g = \mathbb{E}_{\rho_\mathcal{X}}[k_X g(X)]$; the covariance operator $\Sigma : \mathcal{G} \to \mathcal{G}$, defined as $\Sigma := S^\star S = \mathbb{E}_{\rho_\mathcal{X}}[k_X k_X^\star]$; and its action on L^2, $K : L^2(\mathcal{X}, \mathcal{H}, \rho_\mathcal{X}) \to L^2(\mathcal{X}, \mathcal{H}, \rho_\mathcal{X})$, defined as $Kg := SS^\star g = \mathbb{E}_X[k(\cdot, X)g(X)]$. Finally, we have define the four central operators

$$\begin{aligned} S\gamma &= k_{(\cdot)}^\star \gamma, & S^\star g &= \mathbb{E}_{\rho_\mathcal{X}}[k_X g(X)] \\ \Sigma &:= S^\star S = \mathbb{E}_{\rho_\mathcal{X}}[k_X k_X^\star], & Kg &:= SS^\star g = \mathbb{E}_X[k(\cdot, X)g(X)]. \end{aligned} \tag{13}$$

It should be noted that this construction is usually avoided since, based on the fact that the Frobenius norm of K behave like $\dim(\mathcal{H})$, meaning that when $\mathcal{H}$ is infinite dimensional, K is not a compact operator. However, since we consider $\mathcal{Z}$ finite, we can always consider $\mathcal{H} = \mathbb{R}^{\#\mathcal{Z}}$ with $\varphi(y) = (\ell(z,y)_{z\in\mathcal{Z}}$ and $\psi(z) = (\mathbf{1}_{z=z'})_{z'\in\mathcal{Z}}$, and moreover, we will see that a way can be worked out, even when $\mathcal{H}$ is infinite dimensional, which was already shown by Ciliberto et al. (2016).

RELATION BETWEEN REAL-VALUED VERSUS VECTOR-VALUED RKHS.

Usually convergence in RKHS are studied for real-valued function. We need convergence results for vector-valued function. As mentioned above, we only need the results for Euclidean space, however, we will do it for function that are maps going into potentially infinite dimensional Hilbert space. Indeed, this does not lead to major complication. We provide here one way to get around this issue. An alternative formal way to proceed can be found (Ciliberto et al., 2016).

Real-valued RKHS. We build the real-valued RKHS $\mathcal{G}_\mathcal{X}$ as the closure of the span of the atoms $\bar{k}_x$ for $x \in \mathcal{X}$, under the metric induced by the scalar product $\langle \bar{k}_x, \bar{k}_{x'} \rangle = k(x, x')$. Similarly, we build $\bar{S}$, $\bar{S}^\star$, $\bar{\Sigma}$ and $\bar{K}$. We shall see that the action of Σ on $\mathcal{G}$ can be factorized through its actions $\bar{\Sigma}$ on $\mathcal{G}_\mathcal{X}$.

Algebraic equivalences. Based on the fact that $\|\bar{k}_x\|_{\mathcal{G}_\mathcal{X}} = \|k_x\|_{\text{op}} = \sqrt{k(x,x)}$, it is possible to build an isometry that match $\bar{k}_x$ in $\mathcal{G}_\mathcal{X}$ to k_x in the space of continuous linear operator from $\mathcal{H}$ to $\mathcal{G}$. With $(\bar{e}_i)_{i\in\mathbb{N}}$ an orthogonal basis of $\mathcal{G}_\mathcal{X}$, and $(f_j)_{j\in\mathbb{N}}$ an orthogonal basis of $\mathcal{H}$, we get an orthogonal basis $(e_i f_j)_{i,j\in\mathbb{N}}$ of $\mathcal{G}$. This is exaclty the construction $\mathcal{G} = \mathcal{G}_\mathcal{X} \otimes \mathcal{H}$ of (Ciliberto et al., 2016).

Note that for μ_1, μ_2 two measures on $\mathcal{X}$, we can check that

$$\begin{aligned} \|\mathbb{E}_{X\sim\mu_1}[k_X k_X^\star]\, \mathbb{E}_{X_0\sim\mu_2}[k_{X_0}]\|_{\text{op}}^2 &= \left\|\mathbb{E}_{X\sim\mu_1}[\bar{k}_X \bar{k}_X^\star]\, \mathbb{E}_{X_0\sim\mu_2}[\bar{k}_{X_0}]\right\|_{\mathcal{G}_\mathcal{X}}^2 \\ &= \mathbb{E}_{X,X'\sim\mu_1; X,X'\sim\mu_2}[k(X_0, X)k(X, X')k(X', X_0')]. \end{aligned}$$

This explains that we will allow ourselves to write derivations of the type

$$\left\|(\Sigma + \lambda)^{-\frac{1}{2}} k_x g_n(x)\right\|_{\mathcal{G}} \leq \left\|(\bar{\Sigma} + \lambda)^{-\frac{1}{2}} \bar{k}_x\right\|_{\mathcal{G}_\mathcal{X}} \|g_n(x)\|_{\mathcal{H}}.$$

Note also that for $g := \sum_{ij} c_{ij} e_i f_j \in \mathcal{G}$, with $\sum_{ij} c_{ij}^2 = 1$, $\bar{c}_i := (c_{ij})_{j\in\mathbb{N}} \in \ell^2$, $\bar{A}$ an self-adjoint operator on $\mathcal{G}_{\mathcal{X}}$ and A its version on $\mathcal{G}$, we have

$$\begin{aligned}\|Ag\|_{\mathcal{G}}^2 &= \sum_{ijk} c_{ij}c_{kj} \left\langle \bar{A}\bar{e}_i, \bar{A}\bar{e}_k \right\rangle_{\mathcal{G}} = \sum_{ij} \left\langle \bar{c}_i, \bar{c}_j \right\rangle_{\ell^2} \left\langle \bar{A}\bar{e}_i, \bar{A}\bar{e}_k \right\rangle_{\mathcal{G}_{\mathcal{X}}} \\ &\leq \sum_{ij} \|\bar{c}_i\|_{\ell^2} \|\bar{c}_j\|_{\ell^2} \left\langle \bar{A}\bar{e}_i, \bar{A}\bar{e}_k \right\rangle_{\mathcal{G}_{\mathcal{X}}} = \left\| \bar{A} \sum_i \|\bar{c}_i\|_{\ell^2} \bar{e}_i \right\|_{\mathcal{G}_{\mathcal{X}}}^2 \leq \|\bar{A}\|_{\mathrm{op}}^2 ,\end{aligned}$$

which explains why we will consider derivations of the type

$$\left\| (\Sigma + \lambda)^{\frac{1}{2}} (\hat{\Sigma} + \lambda)^{-1} (\Sigma + \lambda)^{\frac{1}{2}} \right\|_{\mathrm{op}} \leq \left\| (\bar{\Sigma} + \lambda)^{\frac{1}{2}} (\hat{\bar{\Sigma}} + \lambda)^{-1} (\bar{\Sigma} + \lambda)^{\frac{1}{2}} \right\|_{\mathrm{op}} .$$

Finally, notice that because of the same consideration, if $(\bar{u}_i)_{i\in\mathbb{N}} \in \mathcal{G}_{\mathcal{X}}^{\mathbb{N}}$ diagonalize $\bar{A}$, $(u_i f_j)_{i,j\leq\mathbb{N}} \in \mathcal{G}^{\mathbb{N}\times\mathbb{N}} \approx \mathcal{G}^{\mathbb{N}}$ diagonalize A in $\mathcal{G}$. This justifies the consideration of fractional operators A^p for $p \in \mathbb{R}_+$, such as in Assumptions 8 and 9. Based on those equivalence, we will forget the bar notations, we incite the careful and attentive reader to recover them.

C.2. Estimate g_n as an empirical approximate projection on RKHS

To obtain bounds like Eq. (4), it is sufficient to control the convergence of g_n to g^* in L^∞. Assumption 8 allow us to cast in L^2 the study of the convergence in L^∞. The convergence of g_n towards g^* can be split in two terms, a term expressing the convergence of g_λ towards g^* that is based on geometrical properties and a term expressing the convergence of g_n towards g_λ, that is based on concentration inequalities in $\mathcal{G}$, such as the ones given by Pinelis and Sakhanenko (1986); Minsker (2017). For this last term, we need to characterize g_n and g_λ with the following lemma.

Lemma 22 (Approximation of integral operators) *g_n can be understood as the empirical approximation of g_λ since*

$$g_n = S(\mathbb{E}_{\hat{\rho}}[k_X k_X^\star] + \lambda)^{-1}\, \mathbb{E}_{\hat{\rho}}[k_X \varphi(Y)], \qquad g_\lambda = S(\mathbb{E}_{\rho}[k_X k_X^\star] + \lambda)^{-1}\, \mathbb{E}_{\rho}[k_X \varphi(Y)],$$

with $\hat{\rho} = \frac{1}{n}\sum_{i=1}^n \delta_{X_i} \otimes \delta_{Y_i}$,

Proof Indeed, the expression of g_n and its convergences towards g^* will be understood thanks to the operator S and its derivatives. When $\operatorname{im} S$ is closed in L^2, on can defined the orthogonal projection of g^* to $\operatorname{im} S$, with the L^2 metric as $\pi_{\operatorname{im} S}(g^*) = S(S^\star S)^\dagger S^\star g^*$. When $\operatorname{im} S$ is not closed, or equivalently when Σ has positive eigen values converging to zero, one can define approximate orthogonal projection, through eigen value thresholding or Tikhonov regularization. This last choice leads to the estimate

$$g_\lambda = S(\Sigma + \lambda)^{-1} S^\star g^* = S(S^\star S + \lambda)^{-1} S^\star g^* = SS^\star (SS^\star + \lambda)^{-1} g^* = K(K + \lambda)^{-1} g^*.$$

Note that, because of the Bayes optimum characterization of g^*, $S^\star g^* = \mathbb{E}_\rho[k_X \varphi(Y)]$. This explains the characterization of g_λ.

Interestingly, the approximation of ρ by $\hat{\rho}$ can be thought with the approximation of $L^2(\mathcal{X}, \mathcal{H}, \rho_{\mathcal{X}})$ by $\ell^2(\mathcal{H}^n) \simeq L^2(\mathcal{X}, \mathcal{H}, \hat{\rho}_{\mathcal{X}})$ where for $\Xi = (\xi_i)$, $Z = (\zeta_i) \in \mathcal{H}^n$,

$$\langle \Xi, Z \rangle_{\ell^2} = \frac{1}{n} \sum_{i=1}^n \langle \xi_i, \zeta_i \rangle_{\mathcal{H}} ,$$

and with the empirical probability measure $\hat{\rho} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \otimes \delta_{y_i}$. We redefine the natural homomorphism of $\mathcal{G}$ into ℓ^2 with

$$\begin{array}{rccc} \hat{S}: & \mathcal{G} & \to & \ell^2 \\ & \gamma & \to & \left(k_{x_i}^\star \gamma\right)_{i\leq n}. \end{array}$$

We check that its adjoint is, for $\Xi \in \mathcal{H}^n$ and $\gamma \in \mathcal{G}$

$$\left\langle \hat{S}^\star \Xi, \gamma \right\rangle_{\mathcal{G}} = \left\langle \Xi, \hat{S}\gamma \right\rangle_{\ell^2} = \frac{1}{n}\sum_{i=1}^{n} \left\langle \xi_i, k_{x_i}^\star \gamma \right\rangle_{\mathcal{H}} = \left\langle \frac{1}{n}\sum_{i=1}^{n} k_{x_i}\xi_i, \gamma \right\rangle_{\mathcal{G}}.$$

Similarly we define $\hat{K} : \mathcal{H}^n \to \mathcal{H}^n$ and $\hat{\Sigma} : \mathcal{G} \to \mathcal{G}$, with

$$\hat{K}\Xi = \hat{S}\hat{S}^\star \Xi = \left(\frac{1}{n}\sum_{i=1}^{n} k(x_j, x_i)\xi_i\right)_{j\leq n}, \qquad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} k_{x_i} \otimes k_{x_i} = \mathbb{E}_{\hat{\rho}_{\mathcal{X}}}[k_X k_X^\star].$$

Finally we define $\hat{\Phi} = (\varphi(y)_i)_{i\leq n} \in \mathcal{H}^n$, so that

$$\widehat{S^\star g^*} := \mathbb{E}_{\hat{\rho}}[\varphi(Y) \cdot k_X] = \hat{S}^\star \hat{\Phi}.$$

Finally we can express g_n as

$$g_n = S(\hat{\Sigma} + \lambda)^{-1}\hat{S}^\star\hat{\Phi} = S(\hat{S}^\star\hat{S} + \lambda)^{-1}\hat{S}^\star\hat{\Phi} = S\hat{S}^\star(\hat{S}\hat{S}^\star + \lambda)^{-1}\hat{\Phi} = S\hat{S}^\star(\hat{K} + \lambda)^{-1}\hat{\Phi}.$$

This explains the equivalence between g_n defined at the beginning of Section 5 and the g_n expressed in the lemma, that will be used for derivations of theorems. ■

C.3. Linear algebra and equivalent assumptions to Assumptions 7, 8

To proceed with the study of the convergence of g_n towards g_λ in L^2, it is helpful to pass by $\mathcal{G}$. To do so, we need to express Assumptions 7 and 8 in $\mathcal{G}$, which we can do using the following linear algebra property.

Lemma 23 (Linear algebra on compact operators) *There exist $(u_i)_{i\in\mathbb{N}}$ an orthogonal basis of $\mathcal{G}_{\mathcal{X}}$, $(v_i)_{i\in\mathbb{N}}$ an orthogonal basis of $L^2(\mathcal{X}, \mathbb{R}, \rho_{\mathcal{X}})$, and $(\lambda_i)_{i\in\mathbb{N}}$ a decreasing sequence of positive real number such that*

$$S = \sum_{i\in\mathbb{N}} \lambda_i^{1/2} u_i v_i^\star, \qquad S^\star = \sum_{i\in\mathbb{N}} \lambda_i^{1/2} v_i u_i^\star, \qquad \Sigma = \sum_{i\in\mathbb{N}} \lambda_i u_i u_i^\star, \qquad K = \sum_{i\in\mathbb{N}} \lambda_i v_i v_i^\star, \tag{14}$$

where the convergence of series as to be understood with the operator norms. Moreover, we have that, if the kernel k is bounded by κ^2,

$$\sum_{i\in\mathbb{N}} \lambda_i \leq \kappa^2 < +\infty.$$

Therefore, K and Σ are trace-class, and S and $S^\star$ are Hilbert-Schmidt.

Proof First of all notice that $\Sigma = \mathbb{E}_X[k_X \otimes k_X]$ and that $\|k_x \otimes k_x\|_{\mathrm{op}(\mathcal{G}_\mathcal{X})} = \|k_x\|_{\mathcal{G}_\mathcal{X}} = k(x,x) \leq \kappa^2$. Therefore Σ is a nuclear operator, so it is trace class and so it is compact.

The first point results from diagonalization of kernel operator, known as Mercer's Theorem (Mercer, 1909; Steinwart and Scovel, 2012). Σ is a compact operator, therefore, the Spectral Theorem gives the existence of a sequence $(\lambda_i) \in \mathbb{R}^\mathbb{N}$ and a orthonormal basis $(u_i) \in \mathcal{G}_\mathcal{X}^\mathbb{N}$ of $\mathcal{G}_\mathcal{X}$ such that

$$\Sigma = \sum_{i\in\mathbb{N}} \lambda_i u_i u_i^\star,$$

where the convergence has to be understood with the operator norm. Because Σ is of the form $S^\star S$, one can consider (λ_i) a decreasing sequence of positive eigen value. Then, by defining, for all $i \in \mathbb{N}$ with $\lambda_i > 0$,

$$v_i = \lambda_i^{-1/2} S u_i$$

we check that (v_i) are orthonormal, and we complete them to form an orthonormal basis of $(L^2(\mathcal{X}, \mathbb{R}, \rho_\mathcal{X}))$. Finally we check that

$$S = \sum_{i\in\mathbb{N}} \lambda_i^{1/2} v_i u_i^\star,$$

and that the other equalities hold too.

To check the second assertion, we use that $k_x k_x^\star$ is rank one when operating on $\mathcal{G}_\mathcal{X}$ and therefore

$$\begin{aligned}\operatorname{Tr}\Sigma &= \operatorname{Tr}\left(\mathbb{E}_X[k_X k_X^\star]\right) = \mathbb{E}_X\left[\operatorname{Tr}\left(k_X k_X^\star\right)\right] = \mathbb{E}_X\left[\|k_X k_X^\star\|_{\mathrm{op}(\mathcal{G}_\mathcal{X})}\right] \\ &= \mathbb{E}_X\left[\|k_X\|_{\mathcal{G}_\mathcal{X}}\right] = \mathbb{E}_X[k(x,x)] \leq \kappa^2.\end{aligned}$$

This shows that S and $S^\star$ are Hilbert-Schmidt operators and that K is also trace class. ■

This allow us to cast in $\mathcal{G}_\mathcal{X}$ the assumptions expressed in L^2.

Lemma 24 (Equivalence of capacity condition) *For $\sigma \in (0,1]$, it is equivalent to suppose that*

- $\operatorname{Tr}_{L^2(\mathcal{X},\mathcal{H},\rho_\mathcal{X})}(K^\sigma) < +\infty$.
- $\operatorname{Tr}_{\mathcal{G}_\mathcal{X}}(\Sigma^\sigma) < +\infty$.
- $\sum_{i\in\mathbb{N}} \lambda_i^\sigma < +\infty$.

In Assumption 7, the smaller σ, the faster the λ_i decrease, the easier is will be to approximate Σ based on approximation of ρ. This appears explicitly in Theorem 31. Indeed, for $\sigma = 0$, the condition should be defined as Σ of finite rank. Note that when k is bounded, we know that Σ is trace class, and therefore, Assumption 7 holds with $\sigma = 1$.

Lemma 25 (Interpolation inequality in RKHS) *Assumption 8 implies that*

$$\forall\, \gamma \in \mathcal{G}, \qquad \|S\gamma\|_{L^\infty} \leq c_p \left\|\Sigma^{\frac{1}{2}-p}\gamma\right\|_{\mathcal{G}}. \tag{15}$$

Proof We begin by showing the property for $\gamma \in \mathcal{G}_\mathcal{X}$. When $\gamma = \sum_{i\in\mathbb{N}} c_i v_i$ with $\sum_{i\in\mathbb{N}} c_i^2 < +\infty$, denote $g = \sum_{i\in\mathbb{N}} \lambda_i^{\frac{1}{2}-p} c_i u_i$, we have $g \in L^2$, therefore, using Assumption 8,

$$\|S\gamma\|_{L^\infty} = \|K^p g\|_{L^\infty} \leq c_p \|g\|_{L^2} = c_p \left\|\Sigma^{\frac{1}{2}-p}\gamma\right\|_{\mathcal{G}_\mathcal{X}}.$$

This ends the proof for $\mathcal{G}_{\mathcal{X}}$. Note also that when the result of the Lemma holds, then Assumption 8 holds for any $g \in \operatorname{im}_{L^2(\mathcal{X},\mathbb{R},\rho_{\mathcal{X}})} K^{\frac{1}{2}-p}$.

Let switch to $\mathcal{G}$ now. Let $\gamma \in \mathcal{G}$, and denote $g = S\gamma$. Suppose that g achieve it maximum in x_∞, define the direction $\xi = {}^{g(x_\infty)}/_{\|g(x_\infty)\|_{\mathcal{H}}}$, and define $g_\xi : x \to \langle g(x), \xi \rangle_{\mathcal{H}} \in L^2(\mathcal{X}, \mathbb{R}, \rho_{\mathcal{X}})$, and $\gamma_\xi = \sum_{j\in\mathbb{N}} \langle g_\xi, v_i \rangle_{L^2} u_i \in \mathcal{G}_{\mathcal{X}}$. We have

$$\|S\gamma\|_{L^\infty} = \|S\gamma_\xi\|_{L^\infty} \leq c_p \left\|\Sigma^{\frac{1}{2}-p}\gamma_\xi\right\|_{\mathcal{G}_{\mathcal{X}}} \leq c_p \left\|\Sigma^{\frac{1}{2}-p}\gamma\right\|_{\mathcal{G}}.$$

When g does not achieve its maximum, one can do a similar reasoning by considering a basis $(f_i)_{i\in\mathbb{N}}$ of $\mathcal{H}$ and decomposition γ on the basis $(u_i f_j)_{i,j\in\mathbb{N}}$, before summing the directions. ■

In Assumption 8, the bigger ${}^1/_2 - p$ the more we are able to control our problem in $\mathcal{G}$, the better. Note that this reformulation of the interpolation inequality allow to generalized it for p smaller than zero. Note that when k is bounded, $\|(S\gamma)(x)\|_{\mathcal{H}} = \|k_X^\star \gamma\|_{\mathcal{H}} \leq \|k_X\|_{\mathrm{op}} \|\gamma\|_{\mathcal{G}} = \sqrt{k(x,x)}\, \|\gamma\|_{\mathcal{G}}$, hence Assumption 8 holds with $p = {}^1/_2$.

C.4. Linear algebra with atoms k_x and useful inequalities

From the study of the convergence of g_n to g_λ will emerge two quantities linked to eigen values of Σ and the position of k_x regarding eigen spaces, that are

$$\mathcal{N}(\lambda) = \operatorname{Tr}\left((\Sigma + \lambda)^{-1}\Sigma\right), \qquad \mathcal{N}_\infty(\lambda) = \sup_{x \in \operatorname{supp}\rho_{\mathcal{X}}} \left\|(\Sigma + \lambda)^{-\frac{1}{2}} k_x\right\|_{\mathrm{op}}. \tag{16}$$

While those quantity could be bounded with brute force consideration, Assumptions 7 and 8 will help to control them more subtly.

Proposition 26 (Characterization of capacity condition) *The property $\sum_{i\in\mathbb{N}} \lambda_i^\sigma < +\infty$, can be rephrased in term of eigen values of Σ as the existence of a $a_1 > 0$ such that, for all $i > 0$,*

$$\lambda_i \leq a_1 (i+1)^{-\frac{1}{\sigma}}. \tag{17}$$

Proof Denote by u_i and S_n the respective quantities λ_i^σ and $\sum_{i=1}^n u_i$. Because S_n converge, it is a Cauchy sequence, so there exits N such that for any $p > q > N$, $S_p - S_q = \sum_{i=q+1}^p u_i \leq 1$. In particular, considering $p = 2q$, and because (λ_i) is decreasing, we have $q u_{2q} \leq \sum_{i=q+1}^{2q} u_i \leq 1$. Therefore, we have that for all $i > 2N$, $u_i \leq 3(i+1)^{-1}$, considering $(a_1)^\sigma = 3 + \max_{i\leq 2N}\{(i+1)u_i\}$, we get the desired result. ■

Proposition 27 (Characterization of $\mathcal{N}$) *When $\operatorname{Tr}(K^\sigma) < +\infty$, with $a_2 = \int_0^\infty \frac{a_1}{a_1 + t^{\frac{1}{\sigma}}}\,\mathrm{d}t$,*

$$\forall\, \lambda > 0, \qquad \mathcal{N}(\lambda, r) \leq a_2 \lambda^{-\sigma}. \tag{18}$$

Proof Expressed with eigenvalues, we have

$$\mathcal{N}(\lambda) = \operatorname{Tr}\left((\Sigma + \lambda)^{-1}\Sigma\right) = \sum_{i\in\mathbb{N}} \frac{\lambda_i}{\lambda_i + \lambda}.$$

Using that $\lambda_i \leq a_1(i+1)^{-\frac{1}{\sigma}}$, that $x \to \frac{x}{x+a}$ is increasing with respect to x for any $a > 0$ and the series-integral comparison, we get for $\sigma \in (0, 1]$

$$\mathcal{N}(\lambda) \leq \sum_{i\in\mathbb{N}} \frac{a_1(i+1)^{-\frac{1}{\sigma}}}{a_1(i+1)^{-\sigma} + \lambda} \leq \int_0^\infty \frac{a_1 t^{-\frac{1}{\sigma}}}{a_1 t^{-\frac{1}{\sigma}} + \lambda}\,\mathrm{d}t = \int_0^\infty \frac{a_1}{a_1 + \lambda t^{\frac{1}{\sigma}}}\,\mathrm{d}t$$
$$= \lambda^{-\sigma} \int_0^\infty \frac{a_1}{a_1 + (\lambda^\sigma t)^{\frac{1}{\sigma}}}\,\mathrm{d}(\lambda^\sigma t) = a_2 \lambda^{-\sigma},$$

where we check the convergence of the integral. ■

Indeed, Assumption 8 has a profound linear algebra meaning, it is a condition on $\rho_{\mathcal{X}}$-almost all the vector $k_x \in \mathcal{G}_{\mathcal{X}}$ not to be excessively supported on the eigenvector corresponding to small eigenvalue of Σ.

Proposition 28 (Characterization of interpolation condition) *The interpolation Assumption 8 implies that, for all $i \in \mathbb{N}$*

$$\sup_{x\in\rho_{\mathcal{X}}} \left|\langle k_x, u_i\rangle_{\mathcal{G}_{\mathcal{X}}}\right| \leq c_p \lambda_i^{\frac{1}{2}-p}. \tag{19}$$

Proof Consider the decomposition of $k_x \in \mathcal{G}_{\mathcal{X}}$ according to the eigen vectors of Σ, with $a_i(x) = \langle k_x, u_i\rangle$. The interpolation condition Assumption 8, expressed in $\mathcal{G}_{\mathcal{X}}$ with Lemma 25, leads to for any $\gamma_{\mathcal{X}} \in \mathcal{G}_{\mathcal{X}}$, and $S\gamma_{\mathcal{X}} : \mathcal{X} \to \mathbb{R}$,

$$|(S\gamma_{\mathcal{X}})(x)| = \left|\langle k_x, \gamma_{\mathcal{X}}\rangle_{\mathcal{G}_{\mathcal{X}}}\right| \leq \|S\gamma_{\mathcal{X}}\|_{L^\infty} \leq c_p \left\|\Sigma^{\frac{1}{2}-p}\gamma_{\mathcal{X}}\right\|_{\mathcal{G}_{\mathcal{X}}}$$

This implies that

$$\langle k_x, \gamma_{\mathcal{X}}\rangle_{\mathcal{G}_{\mathcal{X}}}^2 = \left(\sum_{i\in\mathbb{N}} \langle k_x, u_i\rangle \langle \gamma_{\mathcal{X}}, u_i\rangle\right)^2 \leq c_p^2 \left\|\Sigma^{\frac{1}{2}-p}\gamma_{\mathcal{X}}\right\|_{\mathcal{G}_{\mathcal{X}}}^2 = c_p^2 \sum_{i\in\mathbb{N}} \lambda_i^{1-2p} \langle \gamma_{\mathcal{X}}, u_i\rangle^2.$$

Taking $\gamma_{\mathcal{X}} = u_i$, we get that

$$|\langle k_x, u_i\rangle| \leq c_p \lambda_i^{\frac{1}{2}-p}.$$

This result relates the interpolation condition to the fact that k_x is not excessively supported on the eigenvectors corresponding to vanishing eigenvalues of Σ. ■

Proposition 29 (Characterization of $\mathcal{N}_\infty(\lambda, r)$) *Under the interpolation condition, Assumption 8, we have with $a_3 = c_p(2p)^{-p}(1-2p)^{\frac{1}{2}-p}$, or $a_3 = c_p$ when $p = 1/2$,*

$$\mathcal{N}_\infty(\lambda) \leq a_3 \lambda^{-p}. \tag{20}$$

Proof First of all, notice that

$$\left\|(\Sigma+\lambda)^{-\frac{1}{2}} k_x\right\|_{\mathcal{G}_{\mathcal{X}}} = \sup_{\|\gamma_{\mathcal{X}}\|_{\mathcal{X}}=1} \left\langle \gamma_{\mathcal{X}}, (\Sigma+\lambda)^{-\frac{1}{2}} k_x\right\rangle_{\mathcal{G}_{\mathcal{X}}} = \sup_{c;\sum_{i\in\mathbb{N}} c_i^2=1} \sum_{i\in\mathbb{N}} \frac{c_i \langle k_x, u_i\rangle}{(\lambda+\lambda_i)^{\frac{1}{2}}}$$
$$\leq c_p \sup_{c;\sum_{i\in\mathbb{N}c_i^2}=1} \sum_{i\in\mathbb{N}} \frac{c_i \lambda_i^{\frac{1}{2}-p}}{(\lambda+\lambda_i)^{\frac{1}{2}}} \leq \sup_{t\in\mathbb{R}_+} c_p \frac{t^{\frac{1}{2}-p}}{(\lambda+t)^{\frac{1}{2}}}.$$

When $p \in (0, 1/2)$, this last function is zero in zero and in infinity, therefore its maximum t_0 verifies, taking the derivative of its logarithm,

$$\frac{1/2 - p}{t_0} = \frac{1}{2(t_0 + \lambda)} \qquad \Rightarrow \qquad t_0 = \frac{(1-2p)\lambda}{2p} \qquad \Rightarrow \qquad \sup_{t \in \mathbb{R}_+} \frac{t^{\frac{1}{2}-p}}{(\lambda + t)^{\frac{1}{2}}} = (2p)^{-p}(1-2p)^{\frac{1}{2}-p}\lambda^{-p}.$$

The cases $p \in \{0, 1\}$ are easy to treat. ■

In the previous analysis, one fact does not appear, it is that Σ and k_x are linked to one another, since $\Sigma = \mathbb{E}_X[k_X k_X^\star]$. The following remark builds on it to relates $\mathcal{N}$ and $\mathcal{N}_\infty$.

Remark 30 (Relation between interpolation and capacity condition) *The capacity and interpolation condition are related by the fact that it unreasonable not to consider that $p \leq \sigma/2$.*

Proof Because $k_x k_x^\star$ is of rank one in $\mathcal{G}_\mathcal{X}$, we have

$$\begin{aligned}\mathcal{N}(\lambda) &= \operatorname{Tr}\left((\Sigma+\lambda)^{-1}\Sigma\right) = \mathbb{E}_X\left[\operatorname{Tr}\left((\Sigma+\lambda)^{-1}k_X k_X^\star\right)\right] = \mathbb{E}_X\left[\operatorname{Tr}\left(k_X^\star(\Sigma+\lambda)^{-1}k_X\right)\right] \\ &= \mathbb{E}_X\left[\left\|k_X^\star(\Sigma+\lambda)^{-1}k_X\right\|_{\mathrm{op}}\right] = \mathbb{E}_X\left[\left\|(\Sigma+\lambda)^{-\frac{1}{2}}k_X\right\|_{\mathcal{G}_\mathcal{X}}^2\right].\end{aligned}$$

So indeed, $\mathcal{N}(\lambda)$ is the expectation of the square $\left\|(\Sigma+\lambda)^{-\frac{1}{2}}k_X\right\|_{\mathcal{G}_\mathcal{X}}$, when $\mathcal{N}_\infty(\lambda)$ is the supremum of this last quantity. Therefore

$$\mathcal{N}(\lambda) \leq \mathcal{N}_\infty(\lambda)^2$$

Supposing that the dependency in λ proved above are tight, we should have $\sigma \geq 2p$, which is the statement of this remark. We refer the reader to Lemma 6.2. of Fischer and Steinwart (2020) for more consideration to relates σ and p (reading p and $\alpha/2$ with their notations) ■

C.5. Geometrical control of the residual $\|g_\lambda - g^*\|_{L^\infty}$

The proof of the first assertion in Lemma 13 follows from, using Assumption 9, with $g_0 \in K^{-q}g^*$,

$$\begin{aligned}g_\lambda - g^* &= (K(K+\lambda)^{-1} - I)g^* = -\lambda(K+\lambda)^{-1}g^* = -\lambda(K+\lambda)^{-1}K^q g_0 \\ &= -\lambda K^p(K+\lambda)^{-1}K^{q-p}g_0.\end{aligned}$$

Then using Assumption 8,

$$\begin{aligned}\|g^* - g_\lambda\|_\infty &\leq c_p\lambda\left\|K^{q-p}(K+\lambda)^{-1}\right\|_{\mathrm{op}}\|g_0\|_{L^2} \\ &\leq c_p\lambda\left\|K(K+\lambda)^{-1}\right\|_{\mathrm{op}}^{q-p}\left\|(K+\lambda)^{-1}\right\|_{\mathrm{op}}^{1+p-q}\|g_0\|_{L^2} \\ &\leq c_p\lambda 1^{q-p}\lambda^{-(1+p-q)}\|g_0\|_{L^2} = b_1\lambda^{q-p},\end{aligned}$$

where we have used that $\left\|K(K+\lambda)^{-1}\right\|_{\mathrm{op}} = \|K\|_{\mathrm{op}}/(\|K\|_{\mathrm{op}}+\lambda) \leq 1$ and that $\left\|(K+\lambda)^{-1}\right\| \leq \lambda^{-1}$.

C.6. Convergence of $\|g_n - g_\lambda\|$ through concentration inequality

For the proof of the second assertion in Lemma 13, we will put ourselves in $\mathcal{G}$. For this, we define in $\mathcal{G}$

$$\gamma = \mathbb{E}_\rho[k_X \varphi(Y)], \qquad \gamma_\lambda = (\Sigma + \lambda)^{-1}\gamma, \qquad \hat\gamma = \mathbb{E}_{\hat\rho}[k_X\varphi(Y)], \tag{21}$$

so that $g_\lambda = S\gamma_\lambda$, and $g_n = S(\hat\Sigma + \lambda)^{-1}\hat\gamma$.

C.6.1. Decomposition into a matrix and a vector term

We begin by expressing $g_n - g_\lambda$ in $\mathcal{G}$ with

$$\begin{aligned}
g_n - g_\lambda &= S\left((\hat\Sigma + \lambda)^{-1}\hat\gamma - (\Sigma + \lambda)^{-1}\gamma\right) \\
&= S\left((\hat\Sigma + \lambda)^{-1}(\hat\gamma - \gamma) + ((\hat\Sigma + \lambda)^{-1} - (\Sigma + \lambda)^{-1})\gamma)\right) \\
&= S\left((\hat\Sigma + \lambda)^{-1}(\hat\gamma - \gamma) + (\hat\Sigma + \lambda)^{-1}(\Sigma - \hat\Sigma)(\Sigma + \lambda)^{-1}\gamma)\right) \\
&= S\left((\hat\Sigma + \lambda)^{-1}((\hat\gamma - \hat\Sigma\gamma_\lambda) - (\gamma - \Sigma\gamma_\lambda))\right),
\end{aligned}$$

where we have used that $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$. Therefore, using the expression, Lemma 25, of Assumption 8 in $\mathcal{G}$, we get

$$\begin{aligned}
\|g_n - g_\lambda\|_{L^\infty} \leq c_p \left\|\Sigma^{\frac{1}{2}-p}(\hat\Sigma + \lambda)^{-1}(\Sigma + \lambda)^{\frac{1}{2}+p}\right\|_{\mathrm{op}} \times \cdots \\
\left\|(\Sigma + \lambda)^{-(\frac{1}{2}+p)}((\hat\gamma - \hat\Sigma\gamma_\lambda) - (\gamma - \Sigma\gamma_\lambda))\right\|_{\mathcal{G}}.
\end{aligned}$$

On the one hand, we have concentration of matrix term towards $\Sigma^{\frac{1}{2}-p}(\Sigma + \lambda)^{-(\frac{1}{2}-p)} \preceq I$. On the other hand, we have concentration of the vector $\hat\gamma - \hat\Sigma\gamma_\lambda$ towards $\gamma - \Sigma\gamma_\lambda$. Indeed the concentration of the matrix term is hard to prove (it is only a conjecture), therefore we will go for an other decomposition, that will result in similar rates when $p \geq 0$, that is

$$\begin{aligned}
\|g_n - g_\lambda\|_{L^\infty} &\leq c_p \left\|\Sigma^{\frac{1}{2}-p}(\Sigma + \lambda)^{-\frac{1}{2}}\right\|_{\mathrm{op}} \mathcal{A}(\lambda)\mathcal{B}(\lambda) \\
\mathcal{A}(\lambda) &= \left\|(\Sigma + \lambda)^{\frac{1}{2}}(\hat\Sigma + \lambda)^{-1}(\Sigma + \lambda)^{\frac{1}{2}}\right\|_{\mathrm{op}}, \\
\mathcal{B}(\lambda) &= \left\|(\Sigma + \lambda)^{-\frac{1}{2}}((\hat\gamma - \hat\Sigma\gamma_\lambda) - (\gamma - \Sigma\gamma_\lambda))\right\|_{\mathcal{G}}.
\end{aligned} \tag{22}$$

Recall the definition of the following important quantity that are going to pop up from the analysis

$$\mathcal{N}(\lambda) = \operatorname{Tr}\left((\Sigma + \lambda)^{-1}\Sigma\right), \qquad \mathcal{N}_\infty(\lambda) = \sup_{x \in \operatorname{supp}\rho_{\mathcal{X}}} \left\|(\Sigma + \lambda)^{-\frac{1}{2}}k_x\right\|_{\mathrm{op}}. \tag{16}$$

C.6.2. Extra matrix term

We control the extra matrix term with

$$\left\|\Sigma^{\frac{1}{2}-p}(\Sigma + \lambda)^{-\frac{1}{2}}\right\|_{\mathrm{op}} = \left\|\Sigma^{\frac{1}{2}-p}(\Sigma + \lambda)^{-(\frac{1}{2}-p)}\right\|_{\mathrm{op}} \left\|(\Sigma + \lambda)^{-p}\right\|_{\mathrm{op}} \leq \lambda^{-p}.$$

Using that $\left\|(\Sigma + \lambda)^{-1}\right\|_{\mathrm{op}} \leq \lambda^{-1}$ and that $\left\|(\Sigma + \lambda)^{-1}\Sigma\right\|_{\mathrm{op}} \leq \|\Sigma\|_{\mathrm{op}}/(\|\Sigma\|_{\mathrm{op}} + \lambda) \leq 1$.

C.6.3. MATRIX CONCENTRATION

Let us make explicit the concentration in the matrix term with

$$\begin{aligned}(\Sigma+\lambda)^{\frac{1}{2}}(\hat\Sigma+\lambda)^{-1}(\Sigma+\lambda)^{\frac{1}{2}} &= I + (\Sigma+\lambda)^{\frac{1}{2}}\left((\hat\Sigma+\lambda)^{-1}-(\Sigma+\lambda)^{-1}\right)(\Sigma+\lambda)^{\frac{1}{2}} \\ &= I + (\Sigma+\lambda)^{\frac{1}{2}}(\hat\Sigma+\lambda)^{-1}\left(\Sigma-\hat\Sigma\right)(\Sigma+\lambda)^{-1}(\Sigma+\lambda)^{\frac{1}{2}}.\end{aligned}$$

From here, notice the following implications (that are actually equivalence)

$$\begin{aligned}\Sigma-\hat\Sigma \preceq t(\Sigma+\lambda) \quad &\Rightarrow \quad \hat\Sigma+\lambda \succeq (1-t)(\Sigma+\lambda) \\ &\Rightarrow \quad (\hat\Sigma+\lambda)^{-1} \preceq (1-t)^{-1}(\Sigma+\lambda)^{-1}. \\ &\Rightarrow \quad (\hat\Sigma+\lambda)^{-1}-(\Sigma+\lambda)^{-1} \preceq t(1-t)^{-1}(\Sigma+\lambda)^{-1}. \\ &\Rightarrow \quad (\Sigma+\lambda)^{\frac{1}{2}}\left((\hat\Sigma+\lambda)^{-1}-(\Sigma+\lambda)^{-1}\right)(\Sigma+\lambda)^{\frac{1}{2}} \preceq t(1-t)^{-1} \\ &\Rightarrow \quad (\Sigma+\lambda)^{\frac{1}{2}}(\hat\Sigma+\lambda)^{-1}(\Sigma+\lambda)^{\frac{1}{2}} \preceq (1-t)^{-1}.\end{aligned}$$

The probability of the event $\Sigma-\hat\Sigma \preceq t(\Sigma+\lambda)$, can be studied through the probability of the event $(\Sigma+\lambda)^{-\frac{1}{2}}(\Sigma-\hat\Sigma)(\Sigma+\lambda)^{-\frac{1}{2}} \preceq t$, which can be studied through concentration of self adjoint operators. Finally, we have shown that

$$\left\|(\Sigma+\lambda)^{-\frac{1}{2}}(\Sigma-\hat\Sigma)(\Sigma+\lambda)^{-\frac{1}{2}}\right\|_{\mathrm{op}} \leq t \quad \Rightarrow \quad \mathcal{A}(\lambda) \leq \frac{1}{1-t}. \tag{23}$$

The best result that we are aware of, for covariance matrix inequality, is the extension to self-adjoint Hilbert-Schmidt operators provided by Minsker (2017) in Section 3.2 of its concentration inequality on random matrices Theorem 3.1. It can be formulated as the following.

Theorem 31 (Concentration of self-adjoint operators (Minsker, 2017)) *Let denote by $(\xi_i)_{i\leq n}$ a sequence of independent self-adjoint operator acting on an separable Hilbert space $\mathcal{A}$, such that $\ker(\mathbb{E}[\xi_i]) = \mathcal{A}$, that are bounded by a constant $M \in \mathbb{R}$, in the sense $\|\xi_i\|_{\mathrm{op}} \leq M$, with finite variance $\sigma^2 = \left\|\mathbb{E}\sum_{i=1}^n \xi_i^2\right\|_{\mathrm{op}}$. For any $t > 0$ such that $6t^2 \geq (\sigma^2 + {}^{Mt}/3)$,*

$$\mathbb{P}\left(\left\|\sum_{i=1}^n \xi_i\right\|_{\mathrm{op}} > t\right) \leq 14\, r\left(\sum_{i=1}^n \mathbb{E}\,\xi_i^2\right)\exp\left(-\frac{t^2}{2\sigma^2+2tM/3}\right),$$

with $r(\xi) = {}^{\operatorname{Tr}\xi}/\|\xi\|_{\mathrm{op}}$.

Let us define ξ that goes from $\mathcal{X}$ to the space of self-adjoint operator action on $\mathcal{G}_{\mathcal{X}}$ as

$$\xi(x) = (\Sigma+\lambda)^{-\frac{1}{2}}k_x k_x^\star(\Sigma+\lambda)^{-\frac{1}{2}}. \tag{24}$$

We have that $(\Sigma+\lambda)^{-\frac{1}{2}}(\Sigma-\hat\Sigma)(\Sigma+\lambda)^{-\frac{1}{2}} = \mathbb{E}_\rho[\xi(X)] - \frac{1}{n}\sum_{i=1}^n \xi(x_i)$. To apply operator concentration, we need to bound ξ and its variance.

Bound on ξ. To bound ξ we proceed with, because $k_x k_x^\star$ is of rank one,

$$\begin{aligned}\|\xi(x)\|_{\mathrm{op}} &= \left\|(\Sigma+\lambda)^{-\frac{1}{2}} k_x k_x^\star (\Sigma+\lambda)^{-\frac{1}{2}}\right\|_{\mathrm{op}} = \operatorname{Tr}\left((\Sigma+\lambda)^{-\frac{1}{2}} k_x k_x^\star (\Sigma+\lambda)^{-\frac{1}{2}}\right) \\ &= \operatorname{Tr}\left(k_x^\star(\Sigma+\lambda)^{-1} k_x\right) = \left\|(\Sigma+\lambda)^{-\frac{1}{2}} k_x\right\|_{\mathcal{G}_{\mathcal{X}}}^2 \leq \mathcal{N}_\infty(\lambda)^2.\end{aligned}$$

Variance of ξ. For the variance of ξ we proceed by noticing that

$$\begin{aligned}\mathbb{E}\,\xi(X) &= \mathbb{E}_X (\Sigma+\lambda)^{-\frac{1}{2}} k_X k_X^\star (\Sigma+\lambda)^{-\frac{1}{2}} = (\Sigma+\lambda)^{-\frac{1}{2}}\,\mathbb{E}_X\left[k_X k_X^\star\right](\Sigma+\lambda)^{-\frac{1}{2}} \\ &= (\Sigma+\lambda)^{-\frac{1}{2}}\Sigma(\Sigma+\lambda)^{-\frac{1}{2}} = (\Sigma+\lambda)^{-1}\Sigma.\end{aligned}$$

Hence

$$\mathbb{E}\,\xi(X)^2 \preceq \sup_{x\in\mathcal{X}} \|\xi(x)\|_{\mathrm{op}}\,\mathbb{E}[\xi(X)] \preceq \mathcal{N}_\infty(\lambda)^2(\Sigma+\lambda)^{-1}\Sigma.$$

And as a consequence

$$\left\|\mathbb{E}\,\xi(x)^2\right\| \leq \mathcal{N}_\infty(\lambda)^2,$$

where we have used that $\left\|(\Sigma+\lambda)^{-1}\Sigma\right\|_{\mathrm{op}} = \|\Sigma\|_{\mathrm{op}}/(\|\Sigma\|_{\mathrm{op}}+\lambda) \leq 1$.

Concentration bound on ξ. Using the self-adjoint concentration theorem, we get for any $t > 0$, such that $6nt^2 \geq \mathcal{N}_\infty(\lambda)^2(1+t/3)$,

$$\mathbb{P}_{\mathcal{D}_n}\left(\|\mathbb{E}_{\hat\rho}[\xi] - \mathbb{E}_\rho[\xi]\|_{\mathrm{op}} > t\right) \leq 14\,\frac{\|\Sigma\|_{\mathrm{op}}+\lambda}{\|\Sigma\|_{\mathrm{op}}}\mathcal{N}(\lambda)\exp\left(-\frac{nt^2}{2\mathcal{N}_\infty(\lambda)^2(1+t/3)}\right).$$

Therefore, using the contraposition of the prior implication, we get

$$\mathbb{P}_{\mathcal{D}_n}\left(\mathcal{A}(\lambda) > \frac{1}{1-t}\right) \leq 14\,\frac{\|\Sigma\|_{\mathrm{op}}+\lambda}{\|\Sigma\|_{\mathrm{op}}}\mathcal{N}(\lambda)\exp\left(-\frac{nt^2}{2\mathcal{N}_\infty(\lambda)^2(1+t/3)}\right). \tag{25}$$

C.6.4. Decomposition of vector term in a variance and a bias term

Let switch to the vector term, consider $\xi : \mathcal{X}\times\mathcal{Y} \to \mathcal{G}$, defined as

$$\xi = (\Sigma+\lambda)^{-\frac{1}{2}} k_x(\varphi(y) - k_x^\star\gamma_\lambda).$$

It allows to express in simple form the vector term as

$$\mathcal{B}(\lambda) = \left\|\frac{1}{n}\sum_{i=1}^n \xi(X_i, Y_i) - \mathbb{E}_{(X,Y)\sim\rho}[\xi(X,Y)]\right\|.$$

We can study this term through concentration inequality in $\mathcal{G}$. To proceed we will dissociate the variability due to Y to the one due to X, recalling that $g_\lambda(x) = k_x^\star\gamma_\lambda$ and going for the following decomposition

$$\begin{aligned}&\xi(x,y) = \xi_v(x,y) + \xi_b(x) \\ &\xi_v(x,y) = (\Sigma+\lambda)^{-\frac{1}{2}} k_x(\varphi(y) - g^*(x)), \\ &\xi_b(x) = (\Sigma+\lambda)^{-\frac{1}{2}} k_x(g^*(x) - g_\lambda(x)),\end{aligned} \tag{26}$$

which corresponds to the decomposition

$$\begin{aligned}&\mathcal{B}(\lambda) \leq \mathcal{B}_v(\lambda) + \mathcal{B}_b(\lambda)\\ &\mathcal{B}_v(\lambda) = \|\mathbb{E}_{\hat\rho}[\xi_v(X,Y)] - \mathbb{E}_{\rho}[\xi_v(X,Y)]\|\\ &\mathcal{B}_b(\lambda) = \|\mathbb{E}_{\hat\rho}[\xi_b(X,Y)] - \mathbb{E}_{\rho}[\xi_b(X,Y)]\|\,.\end{aligned} \tag{27}$$

The first term is due to the error because of having observed $\varphi(y)$ rather than $g^*(x)$, often called "variance", and the second term is due to the aiming for g_λ instead of g^* often called "bias".

C.6.5. Control of the variance

To control the variance term, we will use the Bernstein inequality stated Theorem 21.

Bound on the moment of ξ_v. First of all notice that

$$\|\xi_v(x,y)\|_{\mathcal{G}} \leq \left\|(\Sigma+\lambda)^{-\frac{1}{2}}k_x\right\|_{\text{op}} \|\varphi(y) - g^*(x)\|_{\mathcal{H}}\,.$$

Therefore, under Assumption 5, for $m \geq 2$:

$$\begin{aligned}\mathbb{E}_{(X,Y)\sim\rho}\left[\|\xi_v(X,Y)\|^m\right] &\leq \mathbb{E}_{X\sim\rho_{\mathcal{X}}}\left[\left\|(\Sigma+\lambda)^{-\frac{1}{2}}k_x\right\|_{\text{op}}^m \mathbb{E}_{Y\sim\rho|_X}\left[\|\varphi(y)-g^*(x)\|_{\mathcal{H}}^m\right]\right]\\ &\leq \frac{1}{2}m!\sigma^2M^{m-2}\,\mathbb{E}_{X\sim\rho_{\mathcal{X}}}\left[\left\|(\Sigma+\lambda)^{-\frac{1}{2}}k_x\right\|_{\text{op}}^m\right].\end{aligned}$$

We bound the last term with

$$\begin{aligned}\mathbb{E}_{X\sim\rho_{\mathcal{X}}}\left[\left\|(\Sigma+\lambda)^{-\frac{1}{2}}k_x\right\|_{\text{op}}^m\right] &\leq \sup_{x\in\operatorname{supp}\rho_{\mathcal{X}}}\left\|(\Sigma+\lambda)^{-\frac{1}{2}}k_x\right\|_{\text{op}}^{m-2}\mathbb{E}_{X\sim\rho_{\mathcal{X}}}\left[\left\|(\Sigma+\lambda)^{-\frac{1}{2}}k_x\right\|_{\text{op}}^2\right]\\ &= \mathcal{N}_\infty(\lambda)^{(m-2)}\mathcal{N}(\lambda).\end{aligned}$$

Concentration on ξ_v. Applying Theorem 21, we get, for any $t > 0$, that

$$\mathbb{P}\left(\mathcal{B}_v(\lambda) > t\right) \leq 2\exp\left(-\frac{nt^2}{2\sigma^2\mathcal{N}(\lambda) + 2M\mathcal{N}_\infty(\lambda)t}\right). \tag{28}$$

C.6.6. Control of the bias

To control the bias, we recall a simpler version of Bernstein concentration inequality, that is a corollary of Theorem 21.

Theorem 32 (Concentration in Hilbert space (Pinelis and Sakhanenko, 1986)) *Let denote by $\mathcal{A}$ a Hilbert space and by (ξ_i) a sequence of independent random vectors on $\mathcal{A}$ such that $\mathbb{E}[\xi_i] = 0$, that are bounded by a constant M, with finite variance $\sigma^2 = \mathbb{E}[\sum_{i=1}^n \|\xi_i\|^2]$. For any $t > 0$,*

$$\mathbb{P}(\left\|\sum_{i=1}^n \xi_i\right\| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 + 2tM/3}\right).$$

Bound on ξ_b. We have

$$\|\xi_b(x)\|_{\mathcal{G}} \leq \sup_{x \in \operatorname{supp}\rho_{\mathcal{X}}} \left\|(\Sigma + \lambda)^{-\frac{1}{2}} k_x\right\|_{\text{op}} \|g_\lambda(x) - g^*(x)\|_{\mathcal{H}} \leq \mathcal{N}_\infty(\lambda) \|g_\lambda - g^*\|_\infty .$$

Therefore, with Appendix C.5, we get

$$\|\xi_b(x)\|_{\mathcal{G}} \leq b_1 \lambda^{q-p} \mathcal{N}_\infty(\lambda).$$

Variance of ξ_b. For the variance we proceed with

$$\|\xi_b(x)\|_{\mathcal{G}}^2 \leq \mathcal{N}_\infty(\lambda)^2 \|g_\lambda(x) - g^*(x)\|_{\mathcal{H}}^2 .$$

Therefore

$$\mathbb{E}[\|\xi_b(X)\|^2] \leq \mathcal{N}_\infty(\lambda)^2 \|g_\lambda - g^*\|_{L^2}^2 .$$

Using the derivations made in Appendix C.5, we have, using that $q \leq 1$,

$$\begin{aligned}\|g_\lambda - g^*\|_{L^2} &= \lambda \left\|(K+\lambda)^{-1} K^q g_0\right\|_{L^2} \leq \lambda \left\|(K+\lambda)^{-(1-q)}\right\|_{\text{op}} \left\|(K+\lambda)^{-q} K^q\right\|_{\text{op}} \|g_0\|_{L^2} \\ &\leq \lambda^q \|g_0\|_{L^2} .\end{aligned}$$

Concentration on ξ_b. Adding everything together, we get

$$\mathbb{P}\left(\mathcal{B}_b(\lambda) > t\right) \leq 2 \exp\left(-\frac{nt^2}{2\left(\lambda^{2q}\mathcal{N}_\infty(\lambda)^2 \|g_0\|_{L^2}^2 + b_1 \lambda^{q-p} \mathcal{N}_\infty(\lambda) t/3\right)}\right). \qquad (29)$$

Note that based on the bound on the variance, we would like $\mathcal{N}_\infty(\lambda)^2 \lambda^{2q} \approx \lambda^{2(q-p)}$ to be smaller than $\mathcal{N}(\lambda) \approx \lambda^{-\sigma}$. It is the case since $q > p$.

C.6.7. Union bound

To control $\|g_n - g_\lambda\|_{L^\infty} \leq c_p \lambda^{-p} \mathcal{A}(\lambda)(\mathcal{B}_v(\lambda) + \mathcal{B}_b(\lambda))$, we need to perform a union bound on the control of $\mathcal{A}$ and the control of $\mathcal{B} := \mathcal{B}_v + \mathcal{B}_b$, we use that for any $t > 0$ and $0 < s < 1$, $c_p \lambda^{-p} \mathcal{A}\mathcal{B} > t$ implies $\mathcal{A} > 1/(1-s)$ or $\mathcal{B} > (1-s)t\lambda^p/c_p$. Similarly $\mathcal{B}_v + \mathcal{B}_b > t$, implies that either $\mathcal{B}_v > t/2$, either $\mathcal{B}_b > t/2$. Therefore, we have, the following inclusion of events (with respect to $\mathcal{D}_n$)

$$\{\|g_n - g_\lambda\|_{L^\infty} > t\} \subset \left\{\mathcal{A} > \frac{1}{1-s}\right\} \cup \left\{\mathcal{B}_v > \frac{(1-s)t\lambda^p}{2c_p}\right\} \cup \left\{\mathcal{B}_b > \frac{(1-s)t\lambda^p}{2c_p}\right\}.$$

In term of probability this leads to

$$\mathbb{P}_{\mathcal{D}_n}\left(\|g_n - g_\lambda\|_{L^\infty} > t\right) \leq \mathbb{P}_{\mathcal{D}_n}\left(\mathcal{A} > \frac{1}{1-s}\right) + \mathbb{P}_{\mathcal{D}_n}\left(\mathcal{B} > \frac{(1-s)t\lambda^p}{c_p}\right). \qquad (30)$$

Looking closer it is the term in $\mathcal{B}$ that will be the more problematic, therefore we would like s to be small. It we take s to be a constant with respect to t, we will get something that behaves like $\mathbb{P}(B > t\lambda^p)$, which is the best we can hope for (this also explain why we divide $\mathcal{B} > t$ in $\mathcal{B}_v > t/2$ or

$\mathcal{B}_b > {}^t\!/\!_2$). We will consider $s = {}^1\!/\!_2$. We express concentration based on the expression of $\mathcal{N}$ and $\mathcal{N}_\infty$, assuming $\lambda \leq \|\Sigma\|_{\mathrm{op}}$, and $n > a_3^2 \lambda^{-2p}$

$$\mathbb{P}_{\mathcal{D}_n}(\mathcal{A} > 2) \leq 28 a_2 \lambda^{-\sigma} \exp(-\frac{n\lambda^{2p}}{10 a_3^2}).$$

Similarly we get, when $\lambda \leq 1$, using that $\lambda^{-\sigma} \geq 1$

$$\mathbb{P}_{\mathcal{D}_n}(\mathcal{B}_v > {}^t\!/\!_4) \leq 2 \exp\left(-\frac{n\lambda^\sigma t^2}{32\sigma^2 a_2 + 8 M a_3 \lambda^{-p} t}\right).$$

For the bias term, we can proceed at a brutal bounding, based on the fact that for $\lambda \leq 1$, $\lambda^{q-p} \leq 1 \leq \lambda^{-\sigma}$, to get

$$\mathbb{P}_{\mathcal{D}_n}(\mathcal{B}_b > {}^t\!/\!_4) \leq 2 \exp\left(-\frac{n\lambda^\sigma t^2}{32 a_3^2 \|g_0\|_{L^2} + 8 b_1 a_3 \lambda^{-p} t/3}\right).$$

With $b_4 = \max(32\sigma^2 a_2, 32 a_3^2 \|g_0\|_{L^2})$ and $b_5 = \max(8Ma_3, 8b_1 a_3/3)$, we get the following union bound

$$\mathbb{P}_{\mathcal{D}_n}\left(\mathcal{B} > \frac{t\lambda^p}{2}\right) \leq 4 \exp\left(-\frac{n\lambda^{2p+\sigma} t^2}{b_4 + b_5 t}\right).$$

We proceed with the union bound on $\|g_n - g_\lambda\|_{L^\infty}$ as

$$\mathbb{P}_{\mathcal{D}_n}(\|g_n - g_\lambda\|_{L^\infty} > t) \leq b_2 \lambda^{-\sigma} \exp(-b_3 n \lambda^{2p}) + 4 \exp\left(-\frac{n\lambda^{2p+\sigma} t^2}{b_4 + b_5 t}\right),$$

with $b_2 = 28 a_2$ and $b_3^{-1} = 10 a_3^2$, as long as $b_3 n > \lambda^{-2p}$, and $\lambda \leq \max(1, \|K\|_{\mathrm{op}})$.

C.6.8. Refinement of Lemma 13

Remark that the uniform control in Lemma 13 is more than we need, we only need control for each x as described in Assumption 2. Indeed, if $p(x)$ is such that there exists a constant $\tilde{c_p}$ (that does not depend on x or λ), such that for any $i \in \mathbb{N}$

$$\langle k_x, u_i \rangle_{\mathcal{G}_\mathcal{X}} \leq \tilde{c_p} \lambda_i^{p(x)},$$

then considering that

$$g_n(x) - g_\lambda(x) = k_x^\star(\gamma_n - \gamma_\lambda) = k_x^\star (\Sigma + \lambda)^{-\frac{1}{2}} (\Sigma + \lambda)^{\frac{1}{2}} (\gamma_n - \gamma_\lambda),$$

we can get improve the results of Lemma 13 by replacing p by $p(x)$. While we considered $p = \sup_{x \in \rho_\mathcal{X}} p(x)$ as a consequence of our proof scheme, one can expect to end up with the $\mathbb{E}_X[\lambda^{p(X)}]$ instead of λ^p when deriving the proof of Theorems 14 and 15 (for which one has to refine Theorem Theorem 6 in order to integrate dependency of L to x, similarly to what is done in Lemma 18), which will lead to better rates. Yet, because of complexity of expressing a quantity of the type $\mathbb{E}_X[\varphi(p(X))]$, for some function φ, we decided not to present this improved version in the paper.

C.7. Proof of Theorem 14

Based on on the proof of Theorem 5, we know that

$$\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \ell_\infty \mathbb{P}_{\mathcal{D}_n} \left(\left\| g_n - g^* \right\|_\infty > t_0 \right).$$

Now we use that

$$\mathbb{P}_{\mathcal{D}_n} \left(\left\| g_n - g^* \right\|_\infty > t_0 \right) \leq \mathbb{P}_{\mathcal{D}_n} \left(\left\| g_n - g_\lambda \right\|_\infty > t_0 - \left\| g_\lambda - g^* \right\|_\infty \right).$$

The result follows from derivations in Appendix C.6, where we used that when k is bounded, Assumptions 7 and 8 are verified with $\sigma = 1$ and $p = 1/2$. Note that we do not need the source assumption, since we can bound directly $\left\| g_\lambda - g^* \right\|_{L^2} \leq \left\| g_\lambda - g^* \right\|_{L^\infty} < t_0$ while retaking the proof in Appendix C.6. Moreover, the results of this last proof holds under the condition $n\lambda b_3 > 1$, but, since $\mathbb{E}_{\mathcal{D}_n} \mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \ell_\infty$, we can augment the constant b_6 so that the result in Theorem 14 still holds for any $n \in \mathbb{N}^*$.

C.8. Proof of Theorem 15

We can rephrase Lemma 13, using a union bound

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}_n}(\|g_n - g^*\| > t) &\leq \mathbb{P}_{\mathcal{D}_n}(\|g_n - g_\lambda\| > t/2) + \mathbb{P}_{\mathcal{D}_n}(\|g_\lambda - g^*\| > t/2) \\
&\leq b_2 \lambda^{-\sigma} \exp\left(-b_3 n \lambda^{2p}\right) + 4 \exp\left(-\frac{n\lambda^{2p+\sigma} t^2}{4b_4 + 2b_5 t}\right) + \mathbf{1}_{t \leq 2\lambda^{q-p}}.
\end{aligned}$$

Using variant of Theorem 6 presented in Appendix A.6, we get

$$\begin{aligned}
\mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \ell_\infty b_2 \lambda^{-\sigma} \exp\left(-b_3 n \lambda^{2p}\right) &+ 2c_\psi c_\alpha 2^{\alpha+1} \lambda^{(q-p)(\alpha+1)} \\
&+ 2c_\psi c_\alpha c \left(b_4^{\frac{\alpha+1}{2}} (n\lambda^{2p+\sigma})^{-\frac{\alpha+1}{2}} + b_5^{\alpha+1} (n\lambda^{2p+\sigma})^{-(\alpha+1)} \right).
\end{aligned}$$

As long as $\lambda \leq \max(\|K\|_{\mathrm{op}}, 1)$ and $n \geq (b_3 \lambda^{2p})^{-1}$. We optimize those rates with $\lambda = \lambda_0 n^{-\gamma}$, and γ satisfying

$$2\gamma(q-p) = 1 - \gamma(2p+\sigma) \qquad \Rightarrow \qquad \gamma = (2q+\sigma)^{-1}.$$

This leads to, for n after a certain $N \in \mathbb{N}^*$

$$\begin{aligned}
\mathcal{R}(f_n) - \mathcal{R}(f^*) &\leq \ell_\infty b_2 \lambda_0^{-\sigma} n^{\frac{\sigma}{2q+\sigma}} \exp\left(-b_3 n \lambda_0^{2p} n^{\frac{2(q-p)+\sigma}{2q+\sigma}}\right) \\
&\quad + 2c_\psi c_\alpha 2^{\alpha+1} \lambda_0^{(q-p)(\alpha+1)} n^{-\frac{(q-p)(\alpha+1)}{2q+\sigma}} \\
&\quad + 2c_\psi c_\alpha c \left(b_4^{\frac{\alpha+1}{2}} \lambda_0^{\frac{(2p+\sigma)\alpha+1}{2}} n^{-\frac{(q-p)(\alpha+1)}{2q+\sigma}} + b_5^{\alpha+1} \lambda_0^{(2p+\sigma)\alpha+1} n^{-\frac{2(q-p)(\alpha+1)}{2q+\sigma}} \right) \\
&\leq b_8 n^{-\frac{2(q-p)(\alpha+1)}{2q+\sigma}}.
\end{aligned}$$

Since ℓ is bounded, $\mathcal{R}(f_n) - \mathcal{R}(f^*) \leq \ell_\infty$, and we can always higher b_8, in order to have the inequality for any $n \in \mathbb{N}^*$.